Actas BIREDIAL-ISTEC 2024

CONFERENCIA INTERNACIONAL

BIREDIAL-ISTEC

22-24 de octubre de **2024**

SANTIAGO • CHILE

Coordinación general Meilyn Garro, Universidad de Costa Rica

istec

UNIVERSIDAD DE CHILE

Universidad del Rosario

UNIVERSIDAD DEL NORTE

UNIVERSIDAD NACIONAL DE LA PLATA

UNIVERSIDAD DE COSTA RICA

UFRGS UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

REMERI RED MEXICANA DE REPOSITORIOS INSTITUCIONALES

CC.SIBDI.UCR - CIP/4243

Nombres: Conferencia Internacional sobre Bibliotecas y Repositorios Digitales (13 : 2024 : Santiago, Chile), autor. | Garro, Meylin, editora.

Título: Actas BIREDIAL-ISTEC 2024 : Conferencia Internacional BIREDIAL-ISTEC, 22-24 de octubre de 2024, Santiago, Chile / coordinación general Meilyn Garro.

Descripción: Primera edición digital. | San José, Costa Rica : Universidad de Costa Rica, Vicerrectoría de Investigación, 2025. | Algunos textos en portugués.

Identificadores: **ISBN 978-9968-08-017-0 (PDF)**

Materias: ARMARC: Bibliotecas digitales – Congresos, conferencias, etc. | Repositorios institucionales – Congresos, conferencias, etc. | Ciencia abierta – Congresos, conferencias, etc. | Inteligencia artificial – Congresos, conferencias, etc. | Acceso abierto – Congresos, conferencias, etc.

Clasificación: CDD 027–ed. 23

Coordinación general: Meilyn Garro, Universidad de Costa Rica

Edición, maquetación y corrección de estilo: Meilyn Garro

Diseño de tapa y contratapa: Meilyn Garro

Revisión técnica: Marisa R. De Giusti, Universidad Nacional de la Plata; Marlem Uribe, Universidad del Norte

Primera edición digital (PDF descarga y online): 2025.

Ciudad Universitaria Rodrigo Facio, San José, Costa Rica.

Hecho el depósito de ley.

Edición digital de la Vicerrectoría de Investigación, Universidad de Costa Rica.

Fecha de creación: marzo, 2025.

TABLA DE CONTENIDO

PRESENTACIÓN

La Conferencia Internacional BIREDIAL-ISTEC sobre Bibliotecas y Repositorios Digitales de América Latina conmemoró en 2024 su 13º edición y tuvo como sede la ciudad de Santiago de Chile. La organización estuvo a cargo de la Universidad de Chile, en colaboración con ISTEC – Consorcio Iberoamericano para la Educación en Ciencia y Tecnología, Universidad Nacional de la Plata, Universidad del Norte, Universidad del Rosario, UFRGS – Universidade Federal do Rio Grande do Sul, Universidad de Costa Rica, REMERI-Red Mexicana de Repositorios Institucionales y se llevó a cabo en la semana del 22 al 24 de octubre de 2024 en modalidad presencial y con asistencia abierta y gratuita.

El evento se enfocó en la sinergia de la tecnología y la academia para conocer los avances regionales de la aplicación de Inteligencia Artificial en diferentes procesos relacionados con Ciencia Abierta.

La Conferencia dio lugar a la presentación de ponencias, conferencias magistrales, pósteres y mesas de discusión todos las cuales se reúnen en estas Actas. Los trabajos en su totalidad cubrieron cinco ejes que abarcan aspectos fundamentales de la Ciencia Abierta o interaccionan con ella: inteligencia artificial, comunicación académica y científica, datos abiertos e infraestructura tecnológica, que abarcan aspectos fundamentales de la Ciencia Abierta.

Esta edición contó con la coordinación del siguiente Comité Científico:

- Caterina Groposo, Universidade Federal do Rio Grande do Sul (Brasil)
- Malgorzata Lisowska, Universidad del Rosario (Colombia)
- Marisa R. De Giusti, Universidad Nacional de La Plata (Argentina)
- Marlem Uribe Marenco, Universidad del Norte (Colombia)
- Meilyn Garro, Universidad de Costa Rica (Costa Rica)
- Rosalina Vázquez, REMERI – Red Mexicana de Repositorios Institucionales (México)
- Rafael Castillo, Universidad de Chile (Chile)

El detalle de los Ejes fue el siguiente:

Eje 0: Inteligencia artificial (IA) aplicada a la Ciencia Abierta:

- Evaluación científica basada en inteligencia artificial: integra procesos de evaluación por pares, académica y medición de impacto utilizando aplicaciones basadas en IA.
- Experiencias en la implementación de técnicas y aplicaciones basadas en IA para la optimización de flujos de trabajo y el desarrollo de servicios en el contexto de la Ciencia Abierta.
- Formación de usuarios/as aplicada a servicios basados en inteligencia artificial: considera estrategias de integración y uso de la IA.
- Responsabilidad en la IA: aborda buenas prácticas y consideraciones éticas en el uso de la IA.
- Propiedad intelectual, buenas prácticas y consideraciones legales en el uso de la IA.
- Futuro y desafíos de la IA en la Ciencia Abierta.

Eje 1: Comunicación académica, científica y cultural en abierto:

- Gestión y modelos de sostenibilidad de repositorios: destaca la importancia de la sostenibilidad y poblamiento de repositorios institucionales o temáticos.
- Estrategias para mejorar la calidad, visibilidad y posicionamiento de la producción científica, académica y cultural.
- Estrategias de sensibilización, gestión del cambio, capacitación y formación de competencias sobre comunicación de la ciencia.
- Experiencias y buenas prácticas en la gestión de Recursos Educativos Abiertos (REA).
- Experiencias y buenas prácticas en la gestión de contenido cultural.
- Experiencias y buenas prácticas de proyectos de Ciencia ciudadana: metodologías, financiamiento, reconocimiento, propiedad intelectual.
- Estrategias de integración de resultados de investigación que incluye manejo de objetos digitales complejos.

Eje 2: Datos abiertos:

- Gestión de datos de investigación: incluye la gestión integral de datos de investigación, desde data stewardship, aplicación de los principios FAIR y CARE hasta derechos de autor.
- Gestión de repositorios de datos de investigación: abarca la curaduría de datos, normalización, datos enlazados y gestión de grandes volúmenes de datos.
- Planes o estrategias de sensibilización de la comunidad investigadora sobre la importancia de gestionar datos de investigación.

Eje 3: Evaluación y métricas alternativas:

- Métricas alternativas o de última generación: explora nuevas formas de evaluar la actividad científica y su impacto.
- Evaluación abierta por pares y metodologías abiertas: destaca la importancia de metodologías abiertas en la evaluación.
- Herramientas y metodologías de tratamiento de datos de diversas fuentes para apoyar procesos de evaluación, tales como Google Scholar, Scopus, WoS, AmeliCA, Dimensions, Altmetric, DOAJ, entre otros.

Eje 4: Infraestructura tecnológica:

- Plataformas para la implementación de servicios de Ciencia Abierta: incluye repositorios de datos, Recursos Educativos Abiertos (REA), Sistemas de Gestión de Investigación (CRIS), libros electrónicos y software libre o propietario.
- Interoperabilidad e integración entre sistemas y servicios: reconoce la necesidad de interoperabilidad entre diversos sistemas de Ciencia Abierta, tales como CRIS, portales de revistas, repositorios de datos, identificadores persistentes, repositorios de recursos educativos y de patrimonio cultural, entre otros.
- Tecnología para la preservación digital y extracción automática de datos: aborda aspectos tecnológicos fundamentales para la preservación y análisis de datos, contenidos académicos y culturales.
- Nueva generación de plataformas abiertas.

Inteligencia artificial (IA) aplicada a la Ciencia Abierta

Prototipo para la clasificación de Proyectos de Ciencia Ciudadana utilizando inteligencia artificial: Experiencias en la Universidad del Rosario

Malgorzata Lisowska[1], Blanco Castillo Humberto[2]

Palabras claves

Ciencia ciudadana, inteligencia artificial, procesamiento de lenguaje natural, ciencia abierta

Citizen science, artificial intelligence, Natural Language Processing, Technological Innovation

Eje temático

Inteligencia artificial (IA) aplicada a la Ciencia Abierta

Resumen

Este artículo examina un prototipo para evaluar y clasificar proyectos de ciencia ciudadana, integrando una rúbrica de evaluación y un modelo de Procesamiento de Lenguaje Natural (PLN) personalizado. El enfoque, basado en la inteligencia artificial (IA), subraya su relevancia en la clasificación y abre nuevas perspectivas para la investigación en ciencia abierta.

El análisis comienza revisando el estado global de la ciencia ciudadana y los avances tecnológicos que permiten la evaluación y procesamiento de estos proyectos. Se destaca especialmente la experiencia de la Universidad del Rosario en la implementación de su modelo de ciencia abierta, enfrentando desafíos en la clasificación de proyectos y utilizando la IA para desarrollar un prototipo que determina la pertinencia de los proyectos dentro de la ciencia ciudadana. Este prototipo utiliza una rúbrica complementada con preguntas dinámicas que facilitan una clasificación precisa.

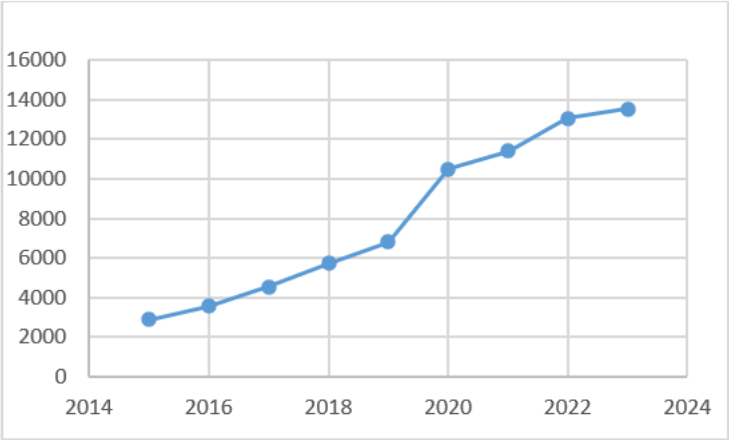
Se concluye discutiendo las ventajas de esta tecnología y su potencial para mejorar la evaluación de proyectos con soporte en inteligencia artificial.

Introducción

La ciencia ciudadana, caracterizada por la participación activa de no científicos en procesos de investigación científica, ha ganado relevancia como un enfoque inclusivo y democratizador para la generación de conocimiento. Además, ha experimentado un crecimiento notable en la última década (Tabla 1), lo cual plantea desafíos significativos para su evaluación y clasificación efectiva.

1 CRAI Universidad del Rosario - Colombia margarita.lisowska@urosario.edu.co
2 CRAI Universidad del Rosario - Colombia humberto.blanco@urosario.edu.co

Tabla 1 - Crecimiento de publicaciones de ciencia ciudadana



Por una parte, el desarrollo de plataformas de ciencia ciudadana ha potenciado la participación masiva de voluntarios en proyectos científicos y mejorado los métodos para la recopilación y clasificación de datos. Plataformas como Zooniverse, la mayor del mundo, con más de 2,000 proyectos y más de 1.6 millones de voluntarios en campos que van desde la astronomía hasta la zoología; iNaturalist, la cual permite mapear y compartir observaciones de biodiversidad, contabilizando más de 50 millones de observaciones hasta 2021; eBird, para registrar observaciones de aves, contribuyendo significativamente al modelado de distribuciones y patrones migratorios globales; Foldit, un juego en línea lanzado en 2008, que a través del juego ha logrado importantes avances en la biología de las proteínas. No solo demuestran el poder de la colaboración masiva y la tecnología para avanzar en la ciencia, sino que también subrayan cómo la participación pública puede transformar la investigación científica y aumentar el conocimiento colectivo de manera significativa y duradera.

Por otro lado, La inteligencia artificial (IA) está revolucionando la ciencia ciudadana al automatizar y mejorar la recopilación y clasificación de grandes volúmenes de datos, especialmente en proyectos relacionados con la identificación de especies y el monitoreo ambiental. Los algoritmos de reconocimiento de imágenes, por ejemplo, facilitan la identificación automática de especies o condiciones ambientales a partir de fotografías cargadas por ciudadanos, aumentando la eficiencia y reduciendo la carga de trabajo manual. Además, los algoritmos de aprendizaje automático están permitiendo el descubrimiento de patrones y correlaciones en los datos que serían difíciles de detectar por humanos, contribuyendo a nuevos descubrimientos científicos y mejorando la comprensión de fenómenos complejos.

Esta sinergia entre la participación humana y la inteligencia artificial no solo potencia la capacidad de campos de conocimiento donde los proyectos de ciencia ciudadana contribuyen a la ciencia formal, sino que también mejora la experiencia y aprendizaje de los participantes, como demuestran estudios de Ceccaroni, L. et al. (2019) y Rafner, J. et al. (2021). También, este potencial para introducir la ciencia ciudadana en más campos del conocimiento, genera un nuevo reto, haciendo más lento el proceso de clasificación y evaluación manual de este tipo de proyectos.

La Universidad del Rosario ha adoptado la ciencia abierta como parte esencial de su quehacer, enfocándose en una investigación de alta calidad que es inclusiva, colaborativa y transparente, destinada a impactar positivamente en la sociedad. Desde hace más de una década, ha participado activamente en el movimiento de acceso abierto, implementando herramientas clave como el repositorio E-docUR y el portal de revistas de acceso abierto, y fue pionera en Colombia al establecer una política institucional de acceso abierto en 2017. En 2020, la universidad intensificó sus esfuerzos creando un modelo de ciencia abierta orientado por ocho pilares fundamentales, enfocados en aspectos como la comunicación académica, los datos de investigación, las métricas de nueva generación, la integridad científica y la ciencia ciudadana. Además, ha desarrollado infraestructura y herramientas adicionales para facilitar la ciencia abierta, culminando en la creación de un Portal Institucional de Ciencia Abierta que refleja su compromiso continuo y promueve la iniciativa a nivel regional.

Luego de establecer el modelo se inició la realización de un diagnóstico en una unidad académica que sirviera como ejemplo para evidenciar la materialización de su modelo la ciencia abierta en la Universidad, con ello se identificaron iniciativas y alcances en los diferentes pilares del modelo incluyendo aquellas iniciativas de investigación que involucran la participación activa de la comunidad. En este proceso de diagnóstico se encontraron algunos desafíos al invitar a la comunidad académica a presentar sus proyectos de ciencia ciudadana, aun cuando se realizó una caracterización de la ciencia ciudadana en la UR.

Para abordar esto desafíos, se propuso un prototipo para automatizar y mejorar el proceso de clasificación y evaluación de proyectos de ciencia ciudadana mediante el uso de tecnologías de procesamiento de lenguaje natural, junto con una rúbrica de evaluación. Su objetivo es proporcionar una herramienta eficaz que permita a evaluar rápidamente si un proyecto cabe dentro de la categoría de ciencia ciudadana basado en una serie de criterios preestablecidos en la caracterización de este eje dentro del modelo de ciencia abierta.

El prototipo propuesto

Aprovechando las ventajas actuales de los modelos de aprendizaje para analizar y comprender el contenido de las propuestas de proyectos de ciencia ciudadana y, aprovechando además su capacidad para identificar objetivos clave, metodologías propuestas, y recursos necesarios descritos en los documentos de proyecto, se incorporó un sistema de clasificación basado en una rubrica predefinida que incluye criterios como innovación, impacto potencial, viabilidad técnica, y alineación con los objetivos de la ciencia ciudadana definidos en el modelo de ciencia abierta de la Universidad del Rosario; además de la clasificación, ofrece retroalimentación y sugerencias de mejora basadas en el análisis realizado, para ayudar en el perfeccionamiento de las propuestas.

Para ello, utilizando ChatGTP se personalizó GTP (RubrikCivitas), con un prompt en el que se define la rúbrica de evaluación basado en 5 criterios principales a los cuales se les otorgó un puntaje:

- Diálogo de saberes entre academia y comunidades (20 puntos)
- Investigación trans e interdisciplinar (20 puntos)
- Fortalecimiento del trabajo investigativo (20 puntos)
- Involucramiento de múltiples fuentes de información (20 puntos)
- Promoción de la ciudadanía activa y la democratización del saber (20 puntos)

Además, se configuró el prompt para que además del resultado de evaluación, el GTP indicara:

- Identificación de las fortalezas y debilidades del proyecto en relación con cada criterio de la rúbrica.
- Ejemplos específicos del proyecto que respalden las evaluaciones del GPT-4.
- Recomendaciones para mejorar el proyecto.

Los resultados preliminares indican una mejora significativa en la eficiencia y calidad de la evaluación de proyectos. Se espera que futuras iteraciones del prototipo integren capacidades de aprendizaje automático más avanzadas y que se expanda su uso a otras facultades y Escuelas.

El prototipo implementado se encuentra en fase de exploración de manera privada, el objetivo final es automatizar el proceso, adicionando otras herramientas de tal manera que pueda ser accesible para cualquier usuario.

Resultados

Para realizar las validaciones se revisaron 17 proyectos (Tabla 2), 12 de los cuales se habían identificado previamente como proyectos de ciencia ciudadana, se contó con un 52% de efectividad respecto a la clasificación de dichos proyectos, en todos los casos por no contar con información suficiente para determinar si los proyectos cumplen o no con los aspectos indicados en la rúbrica de evaluación.

Tabla 2 - Resultados identificación de proyectos de ciencia ciudadana

Número	Proyecto	¿Es proyecto de ciencia ciudadana?
1	El impacto de estrategias de integración asistencial sobre redes integradas de servicios de salud en distintos sistemas de salud de América Latina	Sí
2	Herramientas para la construcción de paz y convivencia en contextos escolares	Sí
3	Educación inicial saludable inclusiva y diversa en el sector El Codito en los cerros orientales de la UPZ 9 (Verbenal) localidad de Usaquén	Sí
4	Modelo de formación en sexualidad y derechos sexuales y reproductivos para personas con discapacidad intelectual	Sí
5	La Colombia imaginada trazos de paz: la literatura infantil como experiencia pedagógica en educación superior	Sí
6	Relación entre contaminación del aire por material particulado (PM10) y morbilidad respiratoria en niños menores de 5 años en la localidad de Kennedy	No
7	Desarrollo de capacidades en jóvenes con discapacidad cognitiva para la gestión del riesgo	No
8	Estrategias de manejo y gestión del riesgo en la localidad de San Cristóbal Sur	No
9	Fortalecimiento de la red de prestadores de servicios de salud en la localidad de Bosa	No
10	Impacto de los espacios de atención ciudadana en la salud mental comunitaria	No

11	Gestión integral de la calidad del agua en la localidad de Chapinero	No
12	Promoción de hábitos saludables en la población escolar de la localidad de Fontibón	No
13	Exploring precipitation toxoplasmosis in Colombia	Si
14	Cuidadores final	No
15	atlas.admon.latam	No
16	Informality	No
17	Replication Data for Network Topology in Decentralized Finance	No

Conclusiones y trabajo futuro

El desarrollo y la implementación del prototipo basado en ChatGPT en la Universidad del Rosario han marca-do un avance significativo en la efectividad con la que se pueden evaluar y clasificar los proyectos de ciencia ciudadana, al facilitar un proceso más rápido y menos subjetivo. Al proporcionar evaluaciones detalladas y sugerencias constructivas basadas en una rúbrica estandarizada, el sistema puede incentivar a los proponen-tes a ajustar y mejorar sus propuestas en fases preliminares de la presentación del proyecto.

Este tipo de iniciativas, se pueden incorporar a algoritmos de aprendizaje profundo de código abierto, que puedan mejorar la capacidad del sistema para entender y procesar lenguaje natural, permitiendo una evaluación aún más precisa de las propuestas.

La evaluación y clasificación de propuestas de ciencia ciudadana, es apenas una muestra del potencial de este tipo de modelos para explorar su efectividad en la evaluación en otros contextos de la ciencia abierta.

Se debe automatizar el proceso y abrirlo a la comunidad a través de la creación de interfaces de usuario adaptadas a las necesidades específicas de diferentes grupos de evaluadores y proponentes, mejorando la accesibilidad y usabilidad del sistema.

En conclusión, la implementación del prototipo basado en ChatGPT en la Universidad del Rosario subraya las ventajas significativas de integrar tecnologías avanzadas en la evaluación de proyectos de ciencia ciudadana. Este enfoque no solo mejora la precisión y la eficiencia en la evaluación de proyectos basados en rúbricas, sino que también demuestra el potencial de la inteligencia artificial para transformar procesos tradicionalmente manuales en sistemas automatizados y altamente eficaces. Según Ceccaroni et al. (2019) y Rafner et al. (2021), el uso de algoritmos avanzados permite una interpretación más profunda y una evaluación más objetiva de los criterios establecidos en las rúbricas, facilitando así decisiones más informadas y justas. Este prototipo representa un paso adelante hacia la optimización de la evaluación de proyectos, ofreciendo un modelo replicable y escalable que puede ser adaptado a diferentes campos y disciplinas científicas, promoviendo una mayor inclusión y democratización en la generación de conocimiento científico.

Bibliografía

Portal institucional de Ciencia Abierta de la Universidad del Rosario. Accedido 29 de abril de 2024. https://cienciaabierta.urosario.edu.co/

Ceccaroni, L., Bibby, J., Roger, E., Flemons, P., Michael, K., Fagan, L., & Oliver, J. (2019). Opportunities and Risks for Citizen Science in the Age of Artificial Intelligence. Citizen Science: Theory and Practice. https://doi.org/10.5334/cstp.241

Rafner, J., Gajdacz, M., Kragh, G., Hjorth, A., Gander, A., Palfi, B., Berditchevskaia, A., Grey, F., Gal, Y., Segal, A., Walmsley, M., Miller, J., Dellerman, D., Haklay, M., Michelucci, P., & Sherson, J. (2021). Revisiting Citizen Science Through the Lens of Hybrid Intelligence. ArXiv, abs/2104.14961.

Malgorzata Lisowska, directora del Centro de Recursos para el Aprendizaje y la Investigación – CRAI, de la Universidad del Rosario en Bogotá.

Magister en Bibliotecología e Información Científica, Universidad Jagiellona de Cracovia, Polonia. Especialista en Administración de Empresas, Universidad del Rosario. Especialista en Gerencia y Gestión Cultural, Universidad del Rosario. Amplia experiencia en bibliotecas públicas y universitarias, con énfasis en gestión y evaluación bibliotecaria y en implementación de nuevas tecnologías. investigadora en el proyecto de la Creación De La Biblioteca Digital Colombiana BDCOL, coordinación de proyectos internacionales como CoLaBoRa (Comunidad Latinoamericana de Bibliotecas y Repositorios Digitales) y en "LA Referencia" patrocinado por de la RedClara y el BID.

Humberto Blanco Castillo, jefe de Innovación y Proyectos del Centro de Recursos para el Aprendizaje y la Investigación – CRAI, de la Universidad del Rosario en Bogotá.

Ingeniero de sistemas, especialista en gerencia de proyectos TIC. Experto en el desarrollo de proyectos enfocados a la implementación, visibilidad e interoperabilidad de repositorios institucionales, así como el desarrollo de soluciones basadas en software libre para la gestión de bibliotecas. Actualmente lidera las estrategias para promover la visibilidad de la producción institucional en acceso abierto, la generación iniciativas y gestión proyectos de base tecnológica que apoyan a los procesos de innovación del CRAI.

Detección de idiomas como tarea de curaduría de datos para repositorios institucionales: desempeño de bibliotecas disponibles y modelos de lenguaje

Carlos Javier Nusch[1], Leticia Cecilia Cagnina[2], Marcelo Luis Errecalde[3], Leandro Antonelli[4], Marisa Raquel De Giusti[5]

Palabras claves

Repositorios Institucionales, tareas de curaduría de datos, herramientas de detección de idiomas, modelos mBERT para detección de idiomas, enfoque zero-shot

Institutional Repositories, Data Curation Tasks, Language Detection Tools, mBERT Models for Language Detection, zero-shot approach

Eje temático

Inteligencia artificial (IA) aplicada a la Ciencia Abierta

Resumen

- Presentación del problema: El enorme volumen de recursos almacenados actualmente en los repositorios digitales representa una gran dificultad a la hora de supervisar y corregir errores o mejorar la calidad de los metadatos. El presente trabajo se enfoca en la corrección del metadato idioma en los registros de resúmenes del repositorio institucional SEDICI.

- Materiales y metodología: A partir de un dataset exportado del repositorio de unos 126.081 ítems se planificó una tarea de detección automática de idiomas utilizando diferentes bibliotecas existentes compatibles con el método zero-shot (langdetect, CLD3, fastText, Polyglot, langid y TextCat). Luego se compararon los resultados obtenidos con los datos de los idiomas registrados por el personal de catalogación del repositorio. Para tratar de mejorar aún más la detección de idiomas se entrenó un modelo mBERT multilenguaje y se comparó su desempeño con el conjunto más pequeño de ítems cuya clasificación por idiomas era diferente entre humanos y la biblioteca Polyglot.

- Resultados: En general, todas las bibliotecas de detección de idiomas mostraron alrededor de un 95% de coincidencia con los idiomas identificados y catalogados por los humanos. En el caso de los modelos mBERT entrenados las coincidencias obtenidas son bajas tanto para los idiomas detectados automáticamente por Polyglot como los catalogados por humanos (78,7% y 19,6% respectivamente). Se encontraron errores de catalogación atribuibles a humanos, pero también errores de las bibliotecas o de los modelos de lenguaje en la tarea de detección.

1 Universidad Nacional de La Plata, PREBI-SEDICI carlosnusch@prebi.unlp.edu.ar
2 Consejo Nacional de Investigaciones Científicas y Técnicas
3 Consejo Nacional de Investigaciones Científicas y Técnicas
4 Universidad Nacional de La Plata, LIFIA
5 Universidad Nacional de La Plata, PREBI-SEDICI

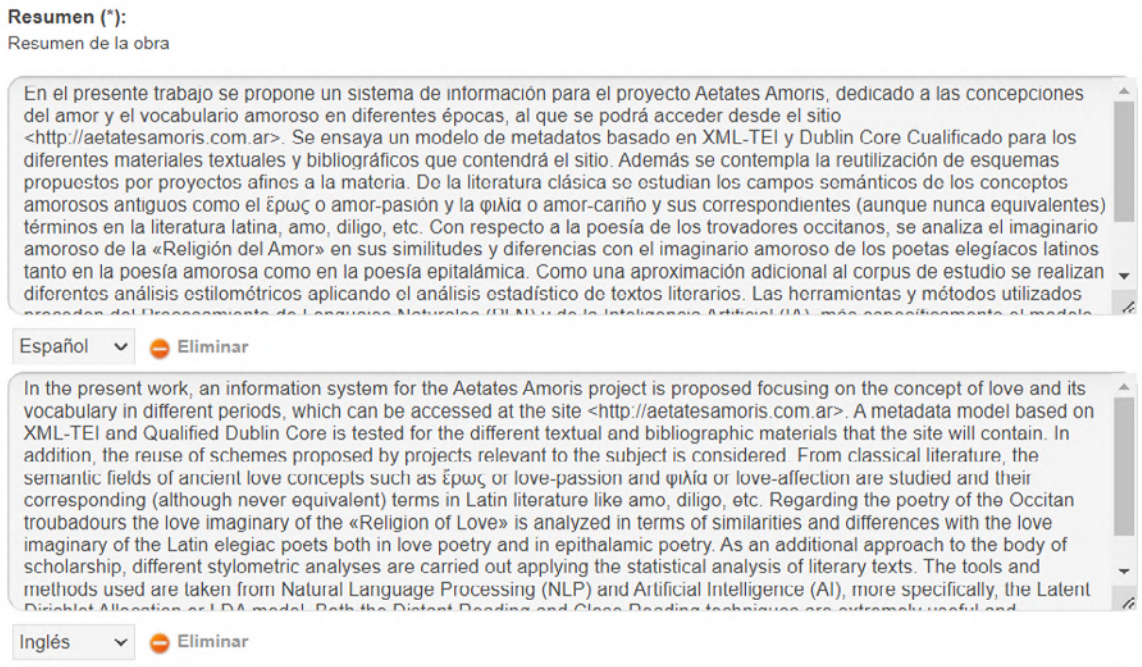
Introducción

Desde los inicios del movimiento de acceso abierto los repositorios institucionales han crecido enormemente en número y volumen de publicaciones. Tal es el caso de SEDICI, el repositorio central de la Universidad Nacional de La Plata, que ha pasado de tener 50 ítems a un año de su creación, 39.000 en 2014 y 156.299 recursos en la actualidad[6]. Entre las diferentes tareas de catalogación llevadas adelante dentro del repositorio está la de asignación del metadato idioma, tanto para el texto completo del material en cuestión como para el o los campos destinados al resumen del artículo, que puede presentarse en varios idiomas diferentes.

Dada la cantidad de campos que el personal a cargo de la catalogación de materiales debe revisar y ajustar en atención a las buenas prácticas, normas y directrices del repositorio, y que dichos campos deben revisarse en cada uno de los ítems que se procesan a diario existe una alta probabilidad de que se cometan diferentes tipos de errores. El riesgo de cometer errores, además, se ha visto acrecentado porque el volumen de ítems que ingestan en el repositorio en tareas automáticas de importación se ha incrementado enormemente. Si bien se ha intentado simplificar y optimizar todas las tareas para llegar a reducir al mínimo estos errores, resulta imposible eliminarlos totalmente.

En las pantallas de control de datos y catalogación del software DSpace existen múltiples campos y uno de ellos es el que se destina a indicar cuál es el idioma de los resúmenes que se están registrando para cada ítem. No es algo tan extraño que se pueda pasar por alto este pequeño campo (ver Figura 1), generalmente situado debajo del campo de resumen, o que se cometa un error de cliqueo al escoger el idioma con el mouse.

Figura 1 - Vista de los campos de resumen para un catalogador en DSpace.



6 Datos accesibles desde: http://sedici.unlp.edu.ar/pages/estadisticasContenidoRepositorio

Con la finalidad de explorar el grado de corrección con el que se estaba catalogando el idioma del campo resumen se exportó un dataset en formato csv el 7 de abril de 2022. El conjunto de datos incluía información de 126.081 ítems, todos los presentes en el repositorio a esa fecha. El objetivo original era llevar a cabo una tarea de curaduría automática aprovechando las diferentes herramientas de detección de idiomas disponibles en la actualidad.

El marco general de las tareas llevadas a cabo puede inscribirse dentro de lo que se conoce como Descubrimiento de Conocimiento en Bases de Datos (KDD, del inglés Knowledge Discovery in Databases) (Fayyad et al., 1996); más comúnmente asociado con la Minería de Datos o extracción de conocimiento e información útiles desde datos crudos. En el caso de la extracción de nueva información y patrones desde de datos de texto se suele denominar Descubrimiento de Conocimiento en Texto (KDT) (Feldman & Dagan, 1995).

Bibliotecas para la detección automática de idiomas

En las tareas de detección automática de idiomas se utilizó el lenguaje Python salvo por el caso de TextCat que se ejecutó en R. Del dataset utilizado solo se analizaron, por obvias razones, los campos de textos de los resúmenes de los diferentes ítems y las etiquetas de idioma aplicadas sobre esos campos. Se utilizaron las bibliotecas langdetect, CLD3, fastText, Polyglot, langid y TextCat con un enfoque zero-shot, esto quiere decir que no se modificaron ni re entrenaron los parámetros del modelo original de la biblioteca. Simplemente se utilizó cada uno de ellos para predecir el idioma de los textos sin necesidad de entrenamiento adicional para el conjunto de datos específico con el que se trabajó. A continuación, se detallan someramente algunas de las características de las bibliotecas de detección automática de idiomas utilizadas.

Langdetect

La biblioteca langdetect[7] es una herramienta de detección de idiomas para Python, inspirada en la biblioteca de Google Language Detection (Compact Language Detector 2) (Shuyo, 2010). Utiliza algoritmos de aprendizaje automático para predecir el idioma de un fragmento de texto. Funciona con textos de diversos dominios y tiene soporte para múltiples idiomas (más de 55). Se trata de una herramienta relativamente ligera, que no requiere una gran cantidad de recursos para funcionar y ofrece resultados confiables en la detección de idiomas.

CLD3

La biblioteca CLD3[8] (Compact Language Detector 3, sucesora de CLD1 y CLD2) es una herramienta de software desarrollada por Google que también emplea modelos de aprendizaje automático para predecir el idioma de un texto (Ooms & Google Inc, 2023). Posee soporte para más de 100 idiomas y puede procesar grandes volúmenes de texto rápidamente. Presenta una alta precisión en la detección de idiomas, inclusive con textos cortos. Puede requerir recursos computacionales mayores.

7 Disponible en: https://pypi.org/project/langdetect/
8 Disponible en: https://github.com/ropensci/cld3

Polyglot

Polyglot[9] es una biblioteca que soporta una amplia gama de tareas y lenguajes (Lui et al., 2014). Puede manejar más de 100 idiomas y posee soporte para una serie de tareas de PLN que exceden la mera detección (como tokenización, reconocimiento de entidades nombradas, análisis de sentimiento, traducción de palabras, etc.). Posee soporte integrado para embeddings de palabras y una serie de modelos pre entrenados lo que permite su uso inmediato sin la necesidad de entrenar modelos desde cero. Una de sus desventajas es que depende de varias bibliotecas y herramientas externas, lo que hace más ardua su instalación y configuración.

Langid

Langid[10] es una herramienta de software libre y de código abierto que puede identificar entre 97 y más de 100 idiomas diferentes (Lui & Baldwin, 2011). Está optimizada para ser rápida y eficiente en términos de uso de memoria y tiempo de procesamiento, inclusive en tareas de procesamiento de texto en tiempo real. Es autocontenida, no depende de servicios externos ni de bases de datos de idiomas, lo que la hace fácilmente instalable y desplegable en cualquier entorno.

TextCat

Textcat[11] es un paquete en R diseñado para la clasificación automática de textos (Hornik et al., 2013). Utiliza patrones de n-gramas para identificar la lengua en la que está escrito un texto, basándose en características estadísticas derivadas de los n-gramas que son únicos o predominantes en idiomas específicos. Se suele utilizar en tareas de procesamiento de lenguaje natural (NLP) que requieren la identificación del idioma antes de realizar análisis más profundos.

FastText

FastText[12] es una biblioteca de aprendizaje automático desarrollada por Facebook AI Research (FAIR) diseñada para la clasificación de textos y la representación de palabras (Bojanowski et al., 2017; Joulin, Grave, Bojanowski, & Mikolov, 2016; Joulin, Grave, Bojanowski, Douze, et al., 2016; Mannes, 2016, 2017). Utiliza modelos de redes neuronales para comprender la representación de las palabras en grandes conjuntos de datos de texto. Una de sus características más sobresalientes es el tratamiento de las palabras como n-gramas de caracteres por lo que puede capturar mejor el significado de palabras cortas, prefijos y sufijos, sobre todo con idiomas de morfología más rica y versátil. Posee una alta precisión en la detección de idiomas, incluso en muestras cortas.

FastText puede ser menos efectivo para algunas tareas de PLN avanzadas comparado con modelos de PLN basados en transformers, como BERT (Devlin et al., 2019), sin embargo suele desempeñarse muy eficientemente en tareas de detección de idiomas.

9 Disponible en: https://github.com/saffsd/polyglot
10 Disponible en: https://github.com/saffsd/langid.py
11 Disponible en: https://cran.r-project.org/web/packages/textcat
12 Disponible en: https://fasttext.cc/

Modelo mBERT entrenado para la detección de idiomas con el dataset de SEDICI

El modelo mBERT[13], o multilingual BERT (BERT multilingüe), es una variante del modelo BERT (Bidirectional Encoder Representations from Transformers) diseñado por Google. BERT marcó un hito en el área de procesamiento del lenguaje natural (NLP) por su capacidad para comprender mejor el contexto de las palabras en un texto, comparado con los modelos anteriores. mBERT está pre entrenado en los textos de Wikipedia de 104 idiomas y es capaz procesar y entender múltiples idiomas sin necesidad de entrenamiento específico del idioma. Al utilizar tecnología de transformers requiere una cantidad de recursos computacionales considerable. Este modelo no se utilizó con el enfoque zero-shot ni tampoco se aplicó a la detección de idiomas de todo el dataset. Se lo entrenó con los datos detectados correctamente por la biblioteca Polyglot para examinar la posibilidad de detectar correctamente idiomas en los casos en los que las otras bibliotecas parecían no responder de la mejor manera.

Resultados preliminares

El desempeño de las diferentes bibliotecas con las que se aplicó el enfoque zero-shot fue relativamente similar en cuanto a la coincidencia del idioma detectado respecto del idioma catalogado por los administradores humanos. Como en algunos casos, las tareas de PLN pueden requerir el uso de recursos importantes, se evaluó además el tiempo requerido para el procesamiento de los datos y la detección de idiomas (Tabla 1). En el caso de las bibliotecas langdetect, CLD3, fastText, Polyglot y langid, se ejecutaron en un entorno de CPU provisto por Google Colab salvo para el caso de TextCat que se ejecutó localmente utilizando los recursos de una notebook. La biblioteca que mayor coincidencia tuvo en la detección de idiomas con los catalogadores humanos fue langid y la de menor tiempo de procesamiento FastText, aunque se trató de la que peores resultados obtuvo.

Tabla 1 - Porcentaje de coincidencias en la detección de idiomas y desempeño de diferentes bibliotecas

Biblioteca	Igual al catalogador humano	Diferente al catalogador humano	Tiempo de ejecución
langdetect	95.3	4.7	25 mins 9.53 secs
CLD3	95.3	4.7	3 mins 56.60 secs
fastText	64.8	35.2	2 mins 5.02 secs
Polyglot	94.7	5.3	2 mins 37.24 secs
langid	95.6	4.4	13 mins 42.24 secs
TextCat	94.3	5.7	2 hours, 2 mins 39 secs[14]

13 Disponible en: https://github.com/google-research/bert/blob/master/multilingual.md

14 La discrepancia entre los tiempos de las otras bibliotecas y TextCat puede deberse a que fue ejecutada en una computadora local en R Studio mientras que las anteriores se corrieron en Google Colab con el lenguaje Python.

Particularidades del dataset

En las primeras pruebas de detección de idiomas con un modelo mBERT el número de predicciones correctas para los idiomas detectados por el modelo eran muy bajas. El español era confundido con el inglés y con el francés en muchos casos. El italiano no tenía predicciones correctas y tanto el francés como el alemán poseían sólo una predicción correcta cada uno. El modelo tenía serias dificultades para clasificar correctamente estas clases ya que el conjunto de datos poseía muy pocos ejemplos para el portugués, francés, alemán e italiano.

Para mejorar el rendimiento del modelo se decidió ajustar la estratificación de los datos de entrenamiento y realizar tareas de aumento de datos para las clases minoritarias. El objetivo de estas tareas era reducir el desbalance en número de ejemplos para cada clase. Además, no todos los resúmenes contaban con el metadato *idioma* (1164 no lo tenían) y por lo tanto no podía corroborarse si el idioma detectado automáticamente era o no correcto. Curiosamente, la ausencia del metadato idioma se dio en muchos de los casos en los que el lenguaje del resumen no era ninguno de los más comunes en el repositorio (español, inglés, portugués, francés, italiano o alemán).

Resultados posteriores al aumento de datos con Marian MT Model

El aumento de datos es una técnica utilizada para generar datos adicionales a partir de datos existentes. En las tareas de PLN se suele partir de textos del dataset y mediante transformaciones que generalmente buscan mantener el mismo significado del texto original, como el uso de sinónimos, por ejemplo, se generan nuevos textos. Al aumentar el conjunto de datos, se puede reducir el sobreajuste y mejorar la capacidad que presenta un modelo a la hora de generalizar con nuevos conjuntos de datos. Otro de los recursos que se suele utilizar es la traducción de textos a otros idiomas. En la tarea de aumento de datos se utilizó MarianMTModel[15] para incrementar el número de ejemplos de las clases minoritarias (francés, portugués, italiano y alemán) a partir de traducciones de ejemplos de las clases mayoritarias (español e inglés).

MarianMTModel forma parte de la familia de modelos de traducción automática neuronal desarrollada por el equipo de Marian NMT (Han et al., 2022; Junczys-Dowmunt et al., 2018; Tiedemann, 2012). Se trata de un modelo diseñado para ser eficiente y liviano, optimizado para aplicaciones en tiempo real y en dispositivos con recursos limitados. Es un proyecto de código abierto compatible con múltiples pares de idiomas.

15 Disponible en: https://huggingface.co/docs/transformers/model_doc/marian

Tabla 2 – Comparación de la distribución de idiomas del dataset original y las nuevas distribuciones generadas con Marian MT Model

Distribución Original			**Distribución luego del Aumento de Datos**		
Idioma	**Ejemplos**	**Porcentaje**	**Idioma**	**Ejemplos**	**Porcentaje**
es	102792	70.41	es	102789	62.69
en	39387	26.98	en	39384	24.02
pt	3346	2.29	pt	6325	3.86
fr	327	0.22	fr	6084	3.71
it	83	0.06	it	6052	3.69
de	52	0.04	de	3343	2.04

Se realizó una tarea de traducción con el modelo Marian MT incrementando las clases minoritarias a un porcentaje de alrededor del 3%. Lamentablemente, para el caso del portugués no se consiguió un modelo de traducción desde el español o el inglés que fuera compatible con la biblioteca.

Resultados luego del primer aumento de datos

Luego de obtener un mayor número de ejemplos de los idiomas de las clases minoritarias se procedió a entrenar un mBERT para clasificación de lenguajes. Con la idea de evitar el sesgo debido al desbalance de clases se redujo el número de ejemplos al número de la clase minoritaria, que luego del aumento de datos resultó ser el portugués. Se creó entonces una nueva muestra con un número igual de ejemplos para cada clase (español, inglés, francés e italiano). Luego se dividió el conjunto de datos balanceado en conjuntos de entrenamiento (12.034 ejemplos) con un porcentaje para pruebas de entrenamiento y validación. Las divisiones realizadas fueron estratificadas según la columna *idioma* manteniendo la misma proporción de clases en cada subconjunto que en el conjunto original.

Se utilizó *BertTokenizer* y *BertForSequenceClassification* para manejar la tokenización y clasificación de textos en múltiples idiomas. Se obtuvieron matrices de confusión para los conjuntos de validación y testeo. También se graficaron las Curvas de Pérdida (Loss) de entrenamiento y validación para evaluar el progreso y el rendimiento del modelo a lo largo de las diferentes épocas.

El número de épocas para el entrenamiento fue de 3 (una época completa significa que cada muestra en el conjunto de datos ha sido presentada una vez al modelo para realizar el aprendizaje). El tamaño del lote (instantaneous batch size per device), es decir, el número de muestras de datos sobre las cuales el modelo calcula la pérdida y actualiza los parámetros en una sola iteración fue de 8.

Resultados del entrenamiento del modelo mBERT

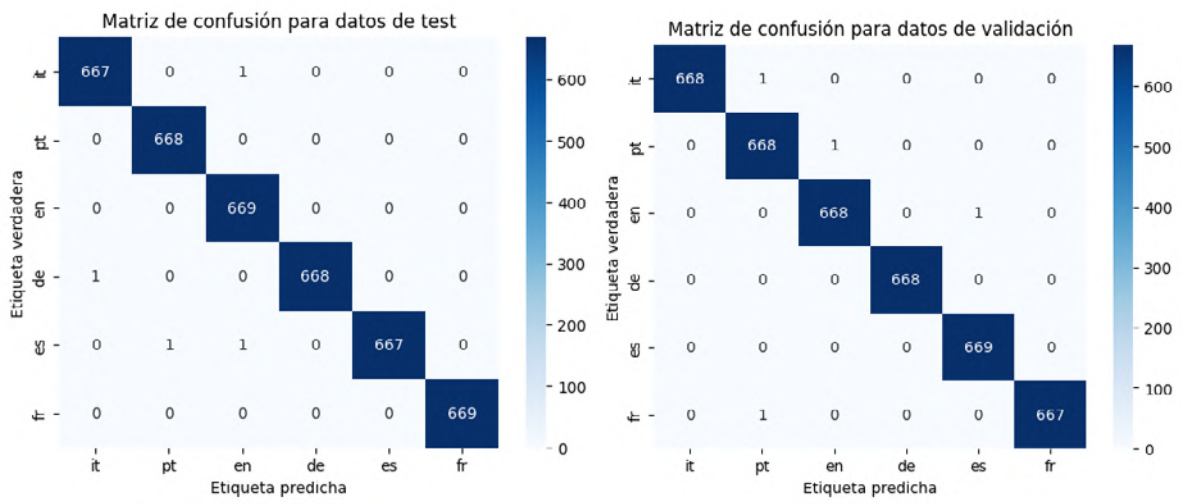
Se utilizaron varias métricas para evaluar el desempeño del modelo que se detallan a continuación:

- *Precision*: para distinguir el número de ítems correctamente identificados como pertenecientes a una clase o proporción de verdaderos positivos entre todos los ítems etiquetados como pertenecientes a esa clase.

- *Recall*: como métrica de sensibilidad del modelo para encontrar todas las instancias pertenecientes a una clase. Es la proporción de verdaderos positivos entre la suma de verdaderos positivos y falsos negativos.
- *F1-Score:* se utiliza como medida de precisión de un test y representa la media armónica de la precisión y el recall. Su valor es de 1 para precisión y recall perfectos y 0 para el peor de los desempeños.
- *Support*: es el número de ocurrencias reales de la clase en el conjunto de datos especificado.
- *Accuracy*: mide la proporción de predicciones correctas (tanto verdaderos positivos como verdaderos negativos) entre el total de casos examinados.

Tanto para los datos de prueba como para los de validación del modelo se obtuvieron precisiones muy altas en todas las clases (el modelo es muy bueno evitando falsos positivos) y los recalls fueron también altos (el modelo es efectivo en identificar todos los verdaderos positivos). El *F1-score* cercano a 1 para todas las clases indicó un buen equilibrio entre precisión y recall. La precisión general (*Accuracy*) fue de 0.999 (casi todas las predicciones del modelo fueron correctas). La consistencia entre los datos de prueba y de validación probó que el modelo generalizaba bien y no mostraba signos de sobreajuste o subajuste significativos[16].

Figura 1 - Matrices de confusión generadas luego del entrenamiento del modelo mBERT con datos aumentados al 3% para las clases minoritarias



16 En aprendizaje automático, el sobreajuste ocurre cuando un modelo aprende a identificar los datos de entrenamiento con demasiada precisión, capturando ruido o detalles irrelevantes. Esto perjudica su capacidad de generalizar a nuevos datos. El subajuste ocurre cuando un modelo es demasiado simple y no puede aprender suficientemente de la estructura subyacente de los datos de entrenamiento como para realizar buenas generalizaciones con nuevos datos.

Tabla 3 - Métricas de evaluación del modelo mBERT para los datos de validación

Reporte de clasificación				
Datos de validación				
Idioma	Precision	Recall	F1-score	Support
it	1	0.999	0.999	669
pt	0.997	0.999	0.999	669
en	0.999	0.999	0.999	669
de	1	1	0.999	668
es	0.999	1	0.999	669
fr	1	0.999	0.999	668
Accuracy			0.999	4012

Tabla 4 - Métricas de evaluación del modelo mBERT para los datos de testeo

Reporte de clasificación				
Datos de testeo				
Idioma	Precisión	Recall	F1-score	Support
it	0.999	0.999	0.999	668
pt	0.999	1	0.999	668
en	0.997	1	0.999	669
de	1	0.999	0.999	669
es	1	0.997	0.999	669
fr	1	1	1	669
Accuracy			0.999	4012

Como métrica adicional del desempeño del modelo se calculó la Pérdida de Entrenamiento (*Training Loss*) una medida que permite evaluar qué tan bien el modelo se ajusta a los datos de entrenamiento (un número más bajo indica un mejor ajuste) y la Pérdida de Validación (*Validation Loss*), una medida de qué tan bien el modelo se generaliza a nuevos datos del conjunto de validación. Durante las tres épocas de entrenamiento del modelo, la Pérdida de Entrenamiento fue consistentemente baja, lo que indica un buen ajuste a los datos de entrenamiento. Entre la primera y la segunda época, se obtuvo una notable mejora en la Pérdida de Validación (de 0.0145 a 0.0086), señal de que el modelo estaba mejorando su capacidad

de generalización. En la tercera época, la Pérdida de Validación continuó disminuyendo ligeramente (de 0.008625 a 0.008507), lo que sugiere una buena generalización sin evidencia de sobreajuste. La Pérdida de Entrenamiento alcanzó un valor extremadamente bajo (0.0001) en esta última época, lo que indica que el modelo ha aprendido casi perfectamente los datos de entrenamiento. La ligera disminución en la Pérdida de Validación entre la segunda y tercera época podría indicar que el modelo está cerca de alcanzar su mejor capacidad de generalización.

Tabla 5 - Pérdida de entrenamiento y validación a través de seis épocas durante el entrenamiento de un modelo de aprendizaje automático

Epoch	Training Loss	Validation Loss
1	0.0024	0.014507
2	0.0086	0.008625
3	0.0001	0.008507

Figura 2 - Curva Loss durante el entrenamiento y la validación

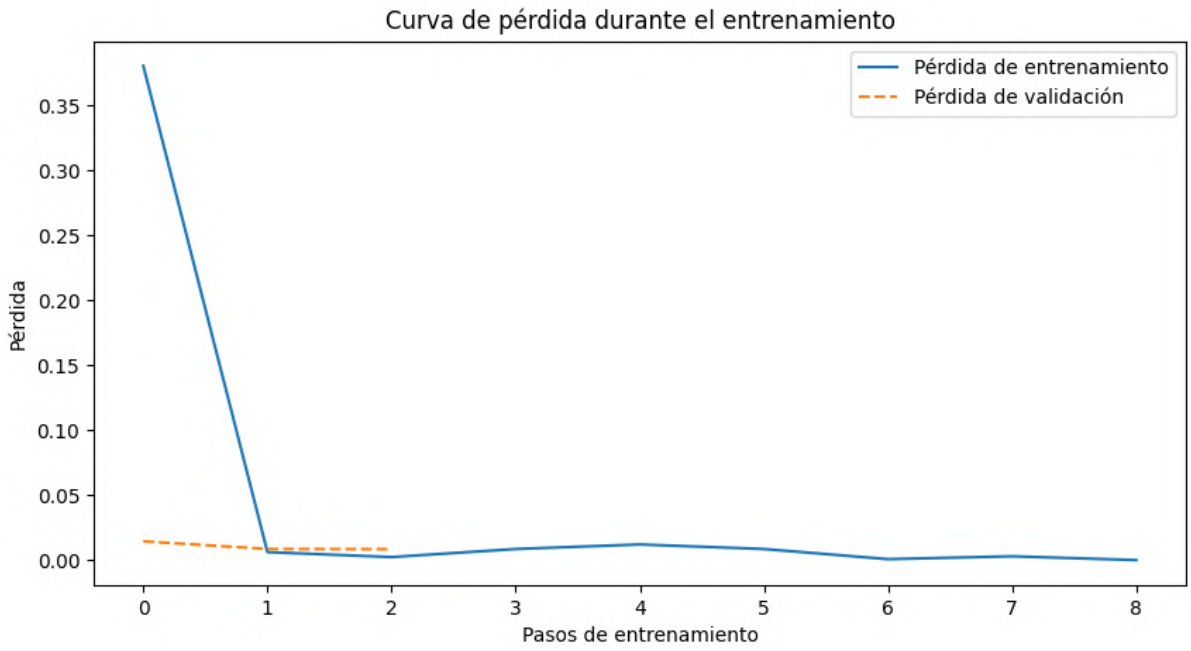


Figura 3 - Gráficos de torta con el porcentaje de coincidencia de los idiomas detectados con cada biblioteca comparado con los idiomas catalogados por humanos

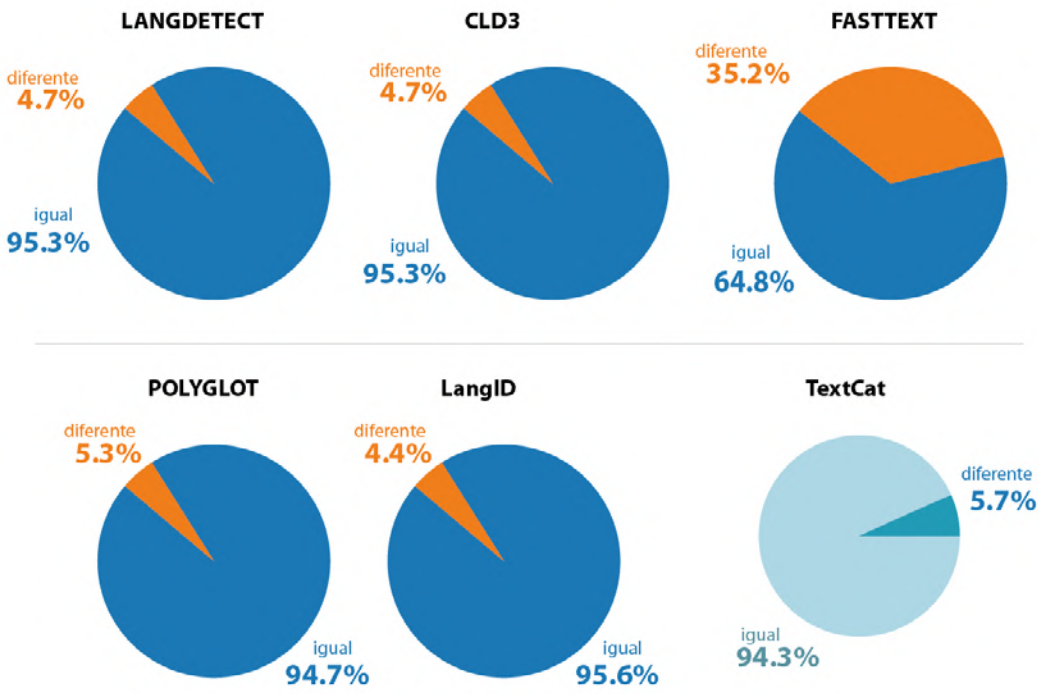
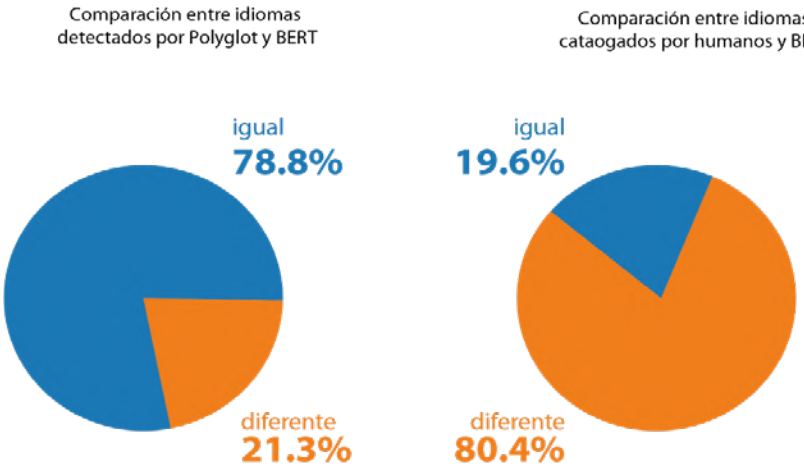


Figura 4 - Gráficos de torta con el porcentaje de coincidencia de los idiomas por mBERT comparado con los idiomas detectados por Polyglot y con los idiomas catalogados por humanos



Conclusiones

En este trabajo se presentaron diferentes resultados de tareas de detección de idiomas utilizando diferentes bibliotecas disponibles para Python y R. En su mayor parte las bibliotecas utilizadas dieron un porcentaje de coincidencia alto (alrededor del 95%) salvo por el caso de FastText. Es muy probable que el trabajo con esta biblioteca requiera entrenar modelos específicos para el conjunto de datos utilizado y también mejorar los parámetros e hiperparámetros de entrenamiento. Lo mismo ocurre con la tarea de detección de idioma que se desarrolló utilizando el modelo entrenado mBERT. Si bien el modelo mostró un excelente desempeño con los datos de entrenamiento y validación, su comparación con los datos en los cuales la detección de la biblioteca Polyglot no coincidía con lo catalogado con humanos arrojó resultados mucho menores. Esto no quiere decir, sin embargo, que el modelo funcione mal, sino que no ha sido entrenado con todos los idiomas presentes en el dataset. Una mejora en el aumento de datos o inclusive la utilización de resúmenes obtenidos de otros repositorios en diferentes lenguajes pueda mejorar el desempeño del modelo.

Otras razones también pueden explicar las fallas constantes de las diferentes bibliotecas y modelos en la detección:

1. En el conjunto de datos utilizado muchos de los resúmenes catalogados por humanos no tenían la etiqueta idioma (por motivos que se ignoran, quizá alguna falla en la migración de versiones de DSpace). Este pequeño porcentaje de idiomas como el latín, el sueco, el holandés, etc. no se encuentran representados explícitamente en las etiquetas con las que se entrenó el modelo mBERT y por lo tanto hubiera sido imposible detectarlos.
2. Algunos textos de los resúmenes simplemente tienen datos insuficientes, es decir, son pocas palabras que no alcanzan para constituir una muestra mínima para las diferentes bibliotecas y modelos.
3. En algunos casos, y con la finalidad de mejorar la visualización de los usuarios del repositorio se optó por incluir código html o LaTex (destinado a visualizar correctamente fórmulas matemáticas) en los textos de los resúmenes. Estos bloques de código seguramente introducen ruido en la detección y dificultan la tarea. Deberán ser eliminados en futuras tareas de detección para mejorar el desempeño de los modelos y bibliotecas.
4. Muchas de las bibliotecas han demostrado fallar en la detección, inclusive de los idiomas mayoritarios, cuando el texto del resumen está compuesto por un listado de palabras o frases.

En trabajos futuros se considerará también la posibilidad de utilizar y evaluar el desempeño de otros modelos de lenguaje como XLM-RoBERTa (XLM-R), Sentence-BERT (SBERT), DistilBERT o ERNIE. Una tarea importante que resta realizar pero que requerirá la intervención de etiquetadores humanos es la de re etiquetar el porcentaje de resúmenes que no cuentan con el campo de idioma y definir, cuál es la opción correcta en los casos en los que las bibliotecas y modelos no coincidieron con el idioma catalogado. Para ello, será necesario desarrollar una herramienta de interacción con catalogadores (probablemente se requiera de más de un humano para controlar los datos) que permita volver a clasificar alrededor del 5% de los ejemplos que conforman el subconjunto de datos en los que la catalogación y la detección no coincidieron. Solo una vez que se tenga la etiqueta de idioma correcta en todos los resúmenes se podrá evaluar con total certeza el desempeño de las herramientas utilizadas. Tal es el caso del modelo BERT entrenado con los datos de Polyglot, que logró un impresionante 78.7 % de coincidencia para los datos en los que las bibliotecas anteriores no coincidían con humanos y un 19,6% de coincidencia con la catalogación humana

de esos mismos datos, lo cual augura un muy buen pronóstico para el uso del modelo en tareas de detección de idiomas en el repositorio. Resta saber si para este subconjunto del dataset, fueron los humanos o las bibliotecas las que reconocieron los idiomas de mejor manera. La finalización de esta tarea que acabamos de iniciar redundará en una mucho mejor calidad de datos para el repositorio.

Bibliografía

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). *Enriching Word Vectors with Subword Information* (arXiv:1607.04606). arXiv. https://doi.org/10.48550/arXiv.1607.04606

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), Article 3. https://doi.org/10.1609/aimag.v17i3.1230

Feldman, R., & Dagan, I. (1995). Knowledge discovery in Textual Databases (KDT). *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, 112-117.

Han, L., Erofeev, G., Sorokina, I., Gladkoff, S., & Nenadic, G. (2022). Examining Large Pre-Trained Language Models for Machine Translation: What You Don't Know about It. En P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, … M. Zampieri (Eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 908-919). Association for Computational Linguistics. https://aclanthology.org/2022.wmt-1.84

Hornik, K., Mair, P., Rauch, J., Geiger, W., Buchta, C., & Feinerer, I. (2013). The textcat Package for n-Gram Based Text Categorization in R. *Journal of Statistical Software*, *52*, 1-17. https://doi.org/10.18637/jss.v052.i06

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). *FastText.zip: Compressing text classification models* (arXiv:1612.03651). arXiv. https://doi.org/10.48550/arXiv.1612.03651

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). *Bag of Tricks for Efficient Text Classification* (arXiv:1607.01759). arXiv. https://doi.org/10.48550/arXiv.1607.01759

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., & Birch, A. (2018). Marian: Fast Neural Machine Translation in C++. En F. Liu & T. Solorio (Eds.), *Proceedings of ACL 2018, System Demonstrations* (pp. 116-121). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-4020

Lui, M., & Baldwin, T. (2011). Cross-domain Feature Selection for Language Identification. En H. Wang & D. Yarowsky (Eds.), *Proceedings of 5th International Joint Conference on Natural Language Processing* (pp. 553-561). Asian Federation of Natural Language Processing. https://aclanthology.org/I11-1062

Lui, M., Lau, J. H., & Baldwin, T. (2014). Automatic Detection and Language Identification of Multilingual Documents. *Transactions of the Association for Computational Linguistics*, *2*, 27-40. https://transacl.org/ojs/index.php/tacl/article/view/86

Mannes, J. (2016, agosto 18). Facebook's Artificial Intelligence Research lab releases open source fastText on GitHub. *TechCrunch*. https://techcrunch.com/2016/08/18/facebooks-artificial-intelligence-research-lab-releases-open-source-fasttext-on-github/

Mannes, J. (2017, mayo 2). Facebook's fastText library is now optimized for mobile. *TechCrunch*. https://techcrunch.com/2017/05/02/facebooks-fasttext-library-is-now-optimized-for-mobile/

Ooms, J. & Google Inc. (2023). *cld3: Google's Compact Language Detector 3* (1.6.0) [Software]. https://cran.r-project.org/web/packages/cld3/

Shuyo, N. (2010). Language detection library for java.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. En N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2214-2218). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

Carlos Javier Nusch es Profesor y Licenciado en Letras por la Universidad Nacional de La Plata y Máster en Humanidades Digitales por la Universidad de Educación a Distancia de España. Ha publicado varios artículos sobre trabajo académico colaborativo, repositorios digitales, digitalización de patrimonio cultural, análisis del discurso político y literatura clásica, medieval y moderna. Trabaja en el Servicio de Difusión de la Creación Intelectual (SEDICI) de la UNLP, en el Proyecto de Enlace de Bibliotecas (PREBI) y en el repositorio CIC-Digital (CICPBA). Es miembro del Comité Asesor del Centro de Servicios en Gestión de Información (CESGI) y personal del Observatorio Medioambiental La Plata (UNLP - CICPBA - CONICET). Coordina la Oficina de Relaciones Institucionales del Consorcio Iberoamericano para la Educación en Ciencia y Tecnología (ISTEC). Participa como docente colaborador ad honorem en el curso de posgrado "Bibliotecas y Repositorios Digitales. Tecnología y aplicaciones" de la Facultad de Informática de la UNLP. Ha participado en proyectos sobre Oralidad, Escritura, Humanidades Digitales Recursos Académicos, Harvesting, OAI-PMH, Visibilidad Web, Repositorios Abiertos, Producción Académica y Científica, Accesibilidad financiados por la UNLP, la CICPBA y el ISTEC.

ORCID: https://orcid.org/0000-0003-1715-4228

Leticia Cecilia Cagnina es Doctora en Ciencias de la Computación, Magíster en Ciencias de la Computación y Licenciada en Ciencias de la Computación. Se desempeña como docente investigadora en la Universidad Nacional de San Luis (UNSL). Es Profesora Adjunta en el Departamento de Informática de la Facultad de Ciencias Físico-Matemáticas y Naturales de la UNSL. Además, es Investigadora Categoría Adjunto en la Carrera de Investigador Científico y Tecnológico del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Su experiencia profesional se enfoca en el campo de la Informática e Inteligencia Artificial, con especialidad en Procesamiento del Lenguaje Natural (PLN). Ha realizado importantes avances en el desarrollo y aplicación de técnicas de PLN en la bioinformática y la detección automática de riesgo en la Web. Su trayectoria académica incluye la dirección y participación en proyectos de investigación en instituciones nacionales e internacionales. Es co-directora del proyecto "Aprendizaje automático y toma de decisiones en sistemas inteligentes para la Web" y ha sido parte del proyecto "Web Information Quality Evaluation Initiative" financiado por la Unión Europea. Además, ha contribuido a proyectos relacionados con la detección de depredadores sexuales en conversaciones de chat y la evaluación de la calidad de contenido web.

ORCID: https://orcid.org/0000-0001-7825-2927

Marcelo Luis Errecalde es Profesor Exclusivo en la Universidad Nacional de San Luis, (Argentina) y dirige el Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC) de la Facultad de Cs. Físico, Matemáticas y Naturales. Trabaja desde hace más de 20 años en temáticas vinculadas a la Inteligencia Artificial, el aprendizaje automático, la minería de textos y la Web y el Procesamiento del Lenguaje Natural. Colabora con diferentes grupos líderes de España, México, Alemania, Austria y Grecia en áreas como la calidad de la información en la web, detección de plagio, detección de depredadores sexuales en la web y determinación del perfil del autor (DPA). Actualmente, el foco de atención en la DPA se centra en la determinación del género, la edad, la orientación política y los rasgos de personalidad de los autores de documentos en la Web. Como resultado de estos trabajos de investigación se han desarrollado sistemas que son actualmente los más efectivos a nivel mundial para la detección de fallas de calidad en Wikipedia y

la detección anticipada de casos de depresión y anorexia en la Web. En la actualidad, sus direcciones de tesis de postgrado se centran en la detección anticipada de riesgos en la Web (depresión, suicidio, anorexia, entre otros), integración de conocimiento externo en los modelos de aprendizaje automático y transparencia e interpretabilidad de los grandes modelos del lenguaje.

ORCID: https://orcid.org/0000-0001-5605-8963

Leandro Antonelli obtuvo el título de Licenciado en Informática en el año 1998 momento en el cual ingresó al Laboratorio de Investigación e Informática Avanzada. En el año 2003 obtuvo el título de Magíster en Ingeniería de Software y en el 2012 el de Doctor en Ciencias Informáticas. Todos los títulos otorgados por la Universidad Nacional de La Plata. Leandro Antonelli se ha desempeñado tanto en la academia como en la industria. En la academia ha atravesado distintas instancias de la docencia, comenzando como ayudante allá por el año 1996. Actualmente se desempeña como Jefe de Trabajos Prácticos en materias de grado y como profesor en materia de posgrado. También realizó investigación principalmente en ingeniería de requerimientos, con publicaciones en conferencias nacionales e internacionales, como así también en revistas. En la industria ha trabajado en reparticiones públicas como así también en ámbitos privados (para clientes nacionales e internacionales). Se ha desempeñado en distintos roles, comenzando como desarrollador en el año 1993 y actualmente se desempeña como ingeniero de software, especializándose tanto en la gestión de requerimientos como en la gestión de proyectos en general (tanto ágiles – es Scrum Master certificado-, como tradicionales).

ORCID: https://orcid.org/0000-0003-1388-0337

Marisa Raquel De Giusti es doctora en Ciencias Informáticas, Ingeniera en Telecomunicaciones y Profesora en Letras de la Universidad Nacional de La Plata (UNLP). Es Profesora de Posgrado en la Facultad de Informática de la UNLP, Directora del Proyecto de Enlace de Bibliotecas (PREBI, 1997) y directora del Servicio de Difusión de la Creación Intelectual (SEDICI, 2002). Impulsó la creación y fue directora hasta el año 2023 del Centro de Servicios en Gestión de Información (CESGI) de la Comisión de Investigaciones Científicas (CIC), donde actualmente reviste como Investigador Emérito. Es presidenta del Consorcio Iberoamericano para Educación en Ciencia y Tecnología (ISTEC) y Directora de la Iniciativa Library linkage (LibLink) de dicho consorcio. Integra el Comité de Expertos del Sistema Nacional de Repositorios Digitales (SNRD) y el Comité Asesor en ciencia abierta y ciudadana. Cuenta con más de 400 trabajos en áreas diversas entre las que se incluyen la gestión de la información, preservación digital, rankings y visibilidad institucional.

ORCID: https://orcid.org/0000-0003-2422-6322

BENANCIB: coletando, organizado, curando e preservando a memória do ENANCIB

Rosa Helena Cunha Vidal[1], *Rene Faustino Gabriel Junior*[2]

Palavras-chave

Benancib; Enancib; Anais de eventosKeywords

Benancib; Enancib; Conference proceedingsEixo temático

Inteligência artificial (IA) aplicada à Ciência Aberta

Resumo

As bases de dados temáticas fora das grandes bases internacionais são importantes para a análise e produção de conhecimentos regionais e nacionais. O objetivo deste trabalho é evidenciar a importância da manutenção, atualização e divulgação da preservação dos anais dos eventos do ENANCIB, de forma a ressaltar a importância de bases de dados curadas, atuando muito além de um repositório, mas como um observatório da área. A BENANCIB já conseguiu realizar alguns de seus objetivos como a coleta e organização de todos os 23 eventos do Enancib realizados no Brasil, agregando mais de 5.900 trabalhos. As metodologias desenvolvidas estão empregando o uso de inteligência artificial para melhoria e curadoria dos dados. A base ainda está em processo de ajustas, mas espera-se em breve a incorporação das citados dos trabalhos.

Introdução

A publicação dos anais dos eventos é uma parte essencial do processo de comunicação científica, pois permite a disseminação ampla e acessível das contribuições apresentadas, promovendo o avanço do conhecimento e facilitando a interação entre os pesquisadores. Conforme mencionado por Silveira, Bufrem e Caregnato (2015), os eventos científicos destacam-se como ferramentas fundamentais para o desenvolvimento profissional, promovendo a interação entre cientistas por meio de uma gama diversificada de formatos, como palestras, mesas redondas, exposições de trabalhos, debates, workshops e cursos breves, entre outras atividades disponíveis.

Além disso, a presença em eventos científicos oferece aos pesquisadores a oportunidade de compartilhar suas pesquisas e ideias com um público especializado, receber feedback construtivo e estabelecer conexões significativas com colegas de todo o mundo. Essa interação direta e o intercâmbio de conhecimentos contribuem significativamente para o desenvolvimento e a evolução contínua das diversas áreas da Ciência da Informação (Alvarez & Caregnato, 2017).

1 Universidade Federal do Rio Grande do Sul, rosadeflor@hotmail.com
2 Universidade Federal do Rio Grande do Sul, rene.gabriel@ufrgs.br

Na área de Ciência da Informação no Brasil o Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB), promovido pela Associação Nacional de Pesquisa em Ciência da Informação (ANCIB), é o principal evento de pesquisa e de pós-graduação brasileiro, fomentando uma discussão ampla e profunda entre sua comunidade (ANCIB, 2024).

O primeiro ENANCIB aconteceu em 1994, organizado pela Universidade Federal de Minas Gerais (UFMG), em Belo Horizonte, e até 2005 não tinha uma periodicidade regular. A partir de 2005 o evento ganhou caráter regular, sendo realizado anualmente no segundo semestre, todos os eventos foram presenciais, com exceção de 2020 que não foi realizado devido à pandemia e em 2021 foi virtual no Rio de Janeiro. Já com 23 edições realizadas, o próximo evento está marcado para ocorrer em novembro de 2024, em Vitória, ES, organizado pela Universidade Federal do Espírito Santo (IBICT, 2021).

Como forma de garantir a preservação e manutenção dos anais desse evento, em 2012, por iniciativa da Universidade Federal Fluminense (UFF), com o apoio da Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), foi lançada a Base de Dados do Encontro Nacional de Pesquisa em Ciência da Informação (BENANCIB) (Gabriel Junior & Vogel, 2022). Utilizando uma plataforma DSpace, e a catalogação era realizada por estudantes do curso de biblioteconomia da UFF. Neste contexto, o BENANCIB buscava resgatar a memória e divulgação do principal evento da área da Ciência da Informação, além de fomentar estudos sobre os GTs e o evento, com disponibilização de dados organizados e curados.

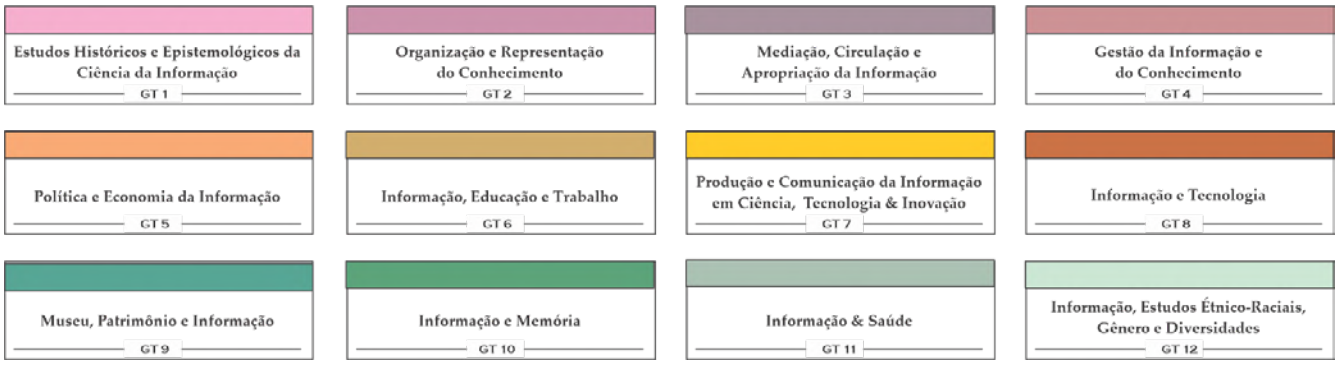
O objetivo deste trabalho é evidenciar a importância da manutenção, atualização e divulgação da preservação dos anais dos eventos do ENANCIB, de forma a ressaltar a importância de bases de dados curadas, atuando muito além de um repositório, mas como um observatório da área.

Sobre o Encontro Nacional de Pesquisa em Ciência da Informação

O ENANCIB destaca-se como um evento significativo em ambiente acadêmico para os programas de pós-graduação no Brasil, estabelecendo-se como um centro de debates e trocas de conhecimento no âmbito da Ciência da Informação. Caracteriza-se como um evento itinerante, tornando-se dinâmico, alcançando diferentes localidades e favorecendo a inclusão de pesquisadores e estudantes de pós-graduação de diversas regiões, isso não apenas diversifica as perspectivas discutidas, mas também fortalece a troca de ideias e estimula parcerias entre instituições acadêmicas e de pesquisa de diferentes localidades (IBICT, 2021).

O evento é organizado por eixos temáticos, chamados de grupos de trabalho (GT). Na primeira edição o evento apresentou os trabalhos de sete GTs, variando esse número até chegar em 2023 com 12 grupos de trabalhos. É importante destacar que da 12ª edição até a 21ª edição permaneceram com 11 GTs. A partir da 4ª edição o evento começou a contar com um tema norteador, o que evidencia a investigação de tópicos inovadores relacionados à Ciência da Informação (Quadro 1).

Quadro 1 – Distribuições temáticas nos GTs do Enancib



Fonte: elaborado pelos autores.

Como já foi citado, o Enancib tem a proposta de reunir as pesquisas de Ciência da Informação, fomentando as discussões e novas pesquisa na área. O Quadro 2 é o resultado do esforço de curadoria dos pesquisa reunindo em um único repositório todas as publicações do evento. Pode-se observar que o primeiro evento ocorreu em 1994, como um encontro dos principais pesquisadores da área no Brasil. Neste evento tem-se o registro de 22 trabalhos apresentados e um total de 35 autores. Destaca-se ainda que nesse evento a modalidade de apresentação foi de shortpapers.

O Quadro 2 evidencia alguns dados relevantes acerca dos eventos do ENANCIB, onde mostra os doze estados brasileiros que já sediaram o evento: Bahia, Belo Horizonte, Brasília, Florianópolis, João Pessoa, Londrina, Marília, Porto Alegre, Rio de Janeiro, São Cristóvão, São Paulo e Valinhos. Essa diversidade nas regiões (Sul, Sudeste, Centro-Oeste e Nordeste) de realização dos encontros incentiva a integração e a consolidação da comunidade científica, destacando também a representatividade e a diversidade geográfica presentes no evento e oferece uma perspectiva de amplitude em eventos futuros na região Norte.

Quadro 2 – Informações gerais acerca de cada ENANCIB na BENANCIB

Edição	Ano	Cidade, Estado	Instituição	Nº GTs	Tema	Nº trabalhos*	Nº autores
1ª	1994	Belo Horizonte, MG	UFMG	7	-	22	35
2ª	1995	Valinhos, SP	PUC Campinas	6	-	56	34
3ª	1997	Rio de Janeiro, RJ	IBICT	6	-	134	69
4ª	2000	Brasília, DF	UnB	8	Conhecimento para o Século XXI: a Pesquisa na Construção da Sociedade da Informação	247	192
5ª	2003	Belo Horizonte, MG	UFMG	8	Informação, conhecimento e transdisciplinaridade	139	224

6ª	2005	Florianópolis, SC	UFSC	7	A política científica e os desafios da sociedade da informação	126	189
7ª	2006	Marília, SP	Unesp	7	A dimensão epistemológica Ciência da Informação e suas interfaces técnicas, políticas e institucionais nos processos de produção, acesso e disseminação da informação	106	171
8ª	2007	Salvador, BA	UFBA	7+1	Promovendo a inserção internacional da pesquisa brasileira em Ciência da Informação	187	292
9ª	2008	São Paulo, SP	USP	8	Diversidade cultural e políticas de informação	151	278
10ª	2009	João Pessoa, PB	UFPB	9	A responsabilidade social da Ciência da Informação	198	320
11ª	2010	Rio de Janeiro, RJ	IBICT/ UFRJ	10	Inovação e inclusão social: questões contemporâneas da informação	248	389
12ª	2011	Brasília, DF	UnB	11	Políticas de informação para a sociedade	259	432
13ª	2012	Rio de Janeiro, RJ	Fiocruz	11	A sociedade em rede para a inovação e o desenvolvimento humano	309	519
14ª	2013	Florianópolis, SC	UFSC	11	Informação e interação: ampliando perspectivas para o desenvolvimento humano	317	523
15ª	2014	Belo Horizonte, MG	UFMG	11	Além das "nuvens": expandindo as fronteiras da Ciência da Informação	333	580
16ª	2015	João Pessoa, PB	UFPB	11	Informação, Memória e Patrimônio: do documento às redes	296	497
17ª	2016	Salvador, BA	UFBA	11	Descobrimentos da Ciência da Informação: desafios da Multi, Inter e Transdisciplinaridade (MIT)	387	640
18ª	2017	Marília, SP	Unesp	11	Informação, sociedade, complexidade	404	672
19ª	2018	Londrina, PR	UEL	11	Sujeito informacional e as perspectivas atuais em Ciência da Informação	448	762
20ª	2019	Florianópolis, SC	UFSC	11	A CI na era da Ciência de Dados	500	850
21ª	2021	Rio de Janeiro, RJ	IBICT	11	50 anos de Ciência da Informação no Brasil: saberes, diversidade e transformação social	356	663
22ª	2022	Porto Alegre, RS	UFRGS	12	O papel da Ciência e da informação em tempos de desinformação	323	606
23ª	2023	São Cristóvão, SE	UFS	12	Das mediações às práticas informacionais: contribuições da Ciência da Informação	418	779

Nota: dados coletados na base de dados BENANCIB, em 03 de março de 2024 (BENANCIB, 2024).

Outro dado revelado é que das 15 instituições que organizaram e sediaram os 23 eventos, a Universidade Federal de Minas Gerais e o Instituto Brasileiro de Informação em Ciência e Tecnologia (cooperando em parceria com a Universidade Federal do Estado do Rio de Janeiro) receberam o evento cada instituição três (3) vezes. A Universidade de Brasília, Universidade Estadual Paulista, Universidade Federal da Paraíba e a Universidade Federal de Santa Catarina acolheram duas (2) vezes cada instituição.

Por ser um evento itinerante, cada sede até então, cada instituições que sedia o evento é responsável pelo site de divulgação do evento, e pelo sistema de submissão, usando o OCS e mais recente o OJS 2. Desde 2021, com o evento no Rio de Janeiro, a Ancib começou custodiar o site de submissão e publicação, sendo a custodiadora dos trabalhos submetidos e apresentados no evento, bem como a publicação dos anais. Com essa dispersão de locais, instituições, muitas das informações acabaram sendo perdidas ou disponíveis aos organizadores de cada evento.

A agregação de todos esses eventos e a disponibilização de acesso ao texto completo está sendo realizado na base de dados do BENANCIB.

BENANCIB

A BENANCIB foi idealizada para reunir todas os trabalhos apresentados nos Enancibs, Porém muitas vezes há dificuldades na recuperação confiável e completa dos anais de cada evento. Isso ocorre devido à dispersão dessas informações em diferentes páginas da internet que sediaram as edições, além das frequentes lacunas, insuficientes metadados ou metadados não padronizados. Essa realidade também se reflete nos anais dos ENANCIBs, onde as informações podem estar ausentes ou apresentar problemas de acesso. Mesmo no site da ANCIB, os anais não estão disponíveis integralmente, havendo edições faltantes ou dados inconsistentes.

Por tudo isso, a criação da BENANCIB, ainda em 2012, foi uma tentativa de reunir os anais do evento e garantir sua preservação digital. A base de dados, seguindo uma abordagem otimizada, procura elevar o padrão dos metadados e alinhar os pontos de acesso dos autores, mesmo quando identificados de maneiras alternativas, com o propósito de atenuar as discrepâncias identificadas (Gabriel Junior & Vogel, 2022). Por ter demandas muito grandes em pequenos espações de tempo, os metadados das publicações normalmente são pouco detalhados, e com poucos critérios de padronização e qualidade.

A ideia do BENANCIB foi concebida na UFF e lançada em formato beta em 2012, com o apoio da FAPERJ e do CNPq. A partir de 2016, a base sofreu com a falta de atualização e a ausência de novos trabalhos, alternando com períodos de reabastecimento e problemas técnicos no servidor. Em 2021, por meio de um convênio firmado entre a UFF e a Universidade Federal do Rio Grande do Sul (UFRGS), os anais dos eventos do ENANCIB passaram a ser replicados e organizados na BENANCIB, que agora está hospedada na BRAPCI, a Base de Dados em Ciência da Informação. Esta última disponibiliza suas ferramentas para o gerenciamento dos anais dos eventos. Dessa forma, a BENANCIB é a guardiã da memória desse evento e se fortalece como valiosa fonte de informação aos pesquisadores (Gabriel Junior; Vogel, 2022).

Procedimento metodológico

Este estudo está estruturado na vertente de organização da informação em repositório de documentos. Porém para realizar a organização foi necessário a realizar uma pesquisa documental, resgatando publicações relacionada ao evento, recorrendo a documentos publicados, sites na internet e fontes pessoais de informação.

Com grande parte do trabalho já realizado pela equipe da Universidade Federal Fluminense, que coletou e organizou os metadados de título, resumo e palavras-chaves dos anais do Enancib até 2016, que foram incorporados a nova base de dados, foi necessário realizar a coleta dos dados compreendidos aos eventos de 2017 até 2023. Para este fim, foram desenvolvidos robôs de coleta de dados em python, utilizando o protocolo OAI-PMH, nos sites que hospedaram o evento. Este processamento possibilitou a incorporação de todos os anais publicados.

Neste procedimento, observou-se alguns problemas, como a disponibilização de alguns trabalhos que não foram aprovados, mas o software disponibilizava os metadados como se o trabalho tivesse sido publicado, problema esses detectado na versão os OCS/OJS utilizado. Desta forma é necessário realizar a curadoria dos metadados. Percebeu-se também que em alguns trabalhos existe uma divergência do título publicado e no metadado (título e nome e ordem dos autores) disponibilizado. Parte-se da hipótese que essas informações foram alteradas pelo autor por indicação dos avaliadores, ou reorganização de apresentação no documento final, porém o autor não ajustou os metadados do sistema, e a equipe de publicação não verificou esses dados.

Com o objetivo de ser também um repositório dos trabalhos, houve a necessidade de coletar os PDF's (trabalhos completos) de todos os trabalhos publicados. Desta forma foi necessário a criação de outro robô que fizesse o download do arquivo e incorporasse a base de dados associando ao registro coletado. Foi nessa fase que foi possível identificar os trabalhos não apresentados e não publicados. Na base, todos os trabalhos sem o PDF ficam na cor vermelha, avisando a equipe de curadoria que tem algo errado com o registro.

Outro grande problema encontrado na organização refere-se a padronização dos nomes dos autores, pois os Enancib ainda não utilizam identificadores persistentes para autores, a entrada é realizada diretamente pelo nome do autor. Porém muitas vezes a submissão é realizada por alunos, fazendo com que não seja inserido o nome completo do autor, trocando letras em alguns nomes, ou ainda trocando a posição do sobrenome na inserção.

Na Benancib, como na Brapci está se aplicando uma metodologia desenvolvia pela própria equipe da Brapci para utilizar recursos de inteligência artificial, algoritmos e linkedData objetivando identificar os autores e associar suas variantes, possibilitando a incorporação de identificadores persistes como os fornecidos pelo BrCris, VIAF, Lattes, ISNI, OrcID. Alguns testes foram realizados com o cálculo de distância de Levenshtein (Ruberto; Antoniazzi, 2017), porém o método não se demonstrou eficaz, pois acabava unificando pessoas diferentes com nome parecidos, como por exemplo Thiago Sales Silva e Tiago Silva Sales[3], que são pessoas diferentes.

O método desenvolvido para possibilita a identifica mais precisa foi baseado no agrupamento de autores, ou seja, reúne-se em uma lista todos os autores e seus coautores nos diversos trabalhos publicados, tanto nos anais do evento como em outras publicações da base Brapci. Desta forma é possível ter uma lista

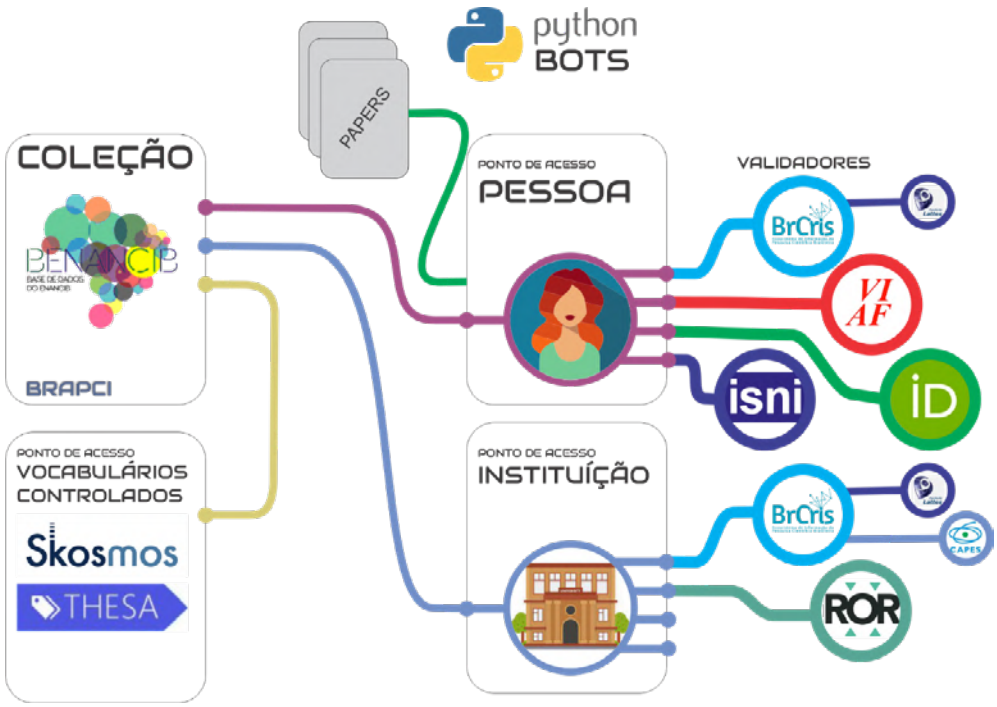
3 Os nomes são fictícios para demonstrar o problema do algoritmo.

de autores e identificar variações de um mesmo autor, como no exemplo: Maria da Silva Costa e Maria S. Costa. O robô com base na aproximação neste cluster de autores consegue ter maior precisão e identificar que é o mesmo autor. Ressalta-se que toda a unificação dos nomes é realizada por seres humano, sendo o algoritmo somente uma ferramenta facilitadora. Uma vez identifico nome do autor e suas variantes, parte-se para consulta nos fornecedores de identificadores persistentes como o OrcID, BrCris, Lattes entre outros, buscando reunir mais dados sobre os autores.

Está em fase de desenvolvimento alguns robôs que analisam o texto completo (PDF) e extraem algumas informações importantes, como afiliação dos autores, e-mail, entre outros dados. Essas informações são aplicadas em modelos de inteligência artificial para possibilita maior agregação de dados. A Figura 1 apresenta de forma resumida esse modelo de agregação.

Com os dados coletados dos PDF também é possível realizar a inferência da instituição de afiliação dos autores, comparando com outros elementos extraídos de outras fontes. Para dados de pesquisadores brasileiros, a metodologia está se demonstrado bastante eficaz. Porém ela ainda está em fase de testes.

Figura 1 – Modelo de integração de dados do BENANCIB



Fonte: elaborado pelos autores.

Os nomes dos autores são padronizados, a fim de eliminar prenomes e/ou nomes faltantes, erros ortográficos ou qualquer outra inconsistência, assim como são acrescentados os títulos, as palavras-chave e os resumos nos idiomas faltantes em cada trabalho para que todos apresentem esses dados em português,

inglês e espanhol, visto que nem todos os documentos possuem esses dados. Com a proposta da BRAPCI em disponibilizar os metadados nesses três idiomas, amplia o alcance tanto no Brasil quanto na América Latina: a base faz a tradução valendo-se da API do Google, completando, então, os metadados.

Outro problema encontrado nos metadados dos eventos é a falta de informações sobre o título em outros idiomas, em muitos casos a falta do resumo (em vários idiomas) e das palavras-chave. A proposta da base é ser indexada por ferramentas de descoberta e aplicação de técnicas de SEO para indexação pelos motores de busca como o Google e o Bing. Desta forma foi necessário desenvolver robôs que identificam no texto elementos como resumo, palavras-chave nos diversos idiomas, novamente aplicando algoritmos de inteligência artificial e modelos de treinamento. Uma vez identificado esses dados, elas são incorporadas a base de dados, e no caso de não ter a tradução em todos os idiomas da base (português, inglês e espanhol) um robô de IA utiliza API de tradução do google para disponibiliza essas informações incorporando aos metadados dos trabalhos.

Ainda no processo de organização, estão sendo construído um tesauro com todos os conceitos utilizados pelos autores, gerando uma padronização dentro da base de dados, melhorando os processos de precisão e revocação do sistema (Fujita; Santos, 2016). Para a construção desse tesauro está sendo utilizada a metodologia de microtesauros temáticos com o uso do Thesa (Gabriel Junior; Laipelt, 2017).

Desde 2023 a BENANCIB, agora subsidiada pelas ferramentas da BRAPCI incorporadas após o novo convênio entre a UFF e UFRGS, vem sendo alimentada com os anais dos eventos do ENANCIB. A base incorpora os trabalhos desde o primeiro evento, em 1994, até o último tão logo estejam disponíveis pela organização do evento.

Resultados preliminares

Com a incorporação de todos os eventos que ocorreram dos Enancib, tens atualmente (dados de abril de 2024) um total de 23 eventos e 5.986 trabalhos catalogados. O processo de curadoria ainda está sendo realizado com ajuda dos robôs. Destaca-se que todos os algoritmos estão sendo desenvolvidos e testados de forma a possibilita de forma mais eficaz a curadoria dos dados.

Dentro da Brapci se criou uma coleção especial que pode ser consultada individualmente, por evento ou incorporada na busca federada dentro da base.

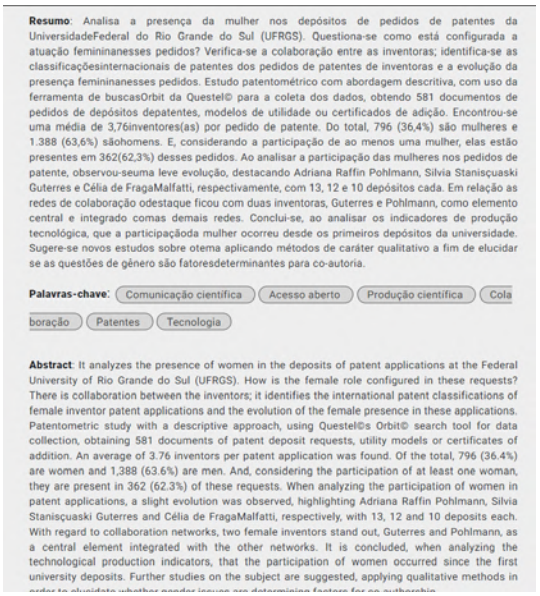
Figura 2 – Interface dentro do Benancib dos eventos distribuídos por ano

Fonte: elaborado pelos autores

O evento mais expressivo em termos de número de trabalhos ocorreu em Santa Catarina em 2019, com a apresentação de 500 trabalhos nas modalidades de trabalho completo e pôster, reunindo aproximadamente 850 autores distintos. Nota-se que o número de autores pode diminuir à medida que se aprofunda a curadoria dos nomes. Naquele evento, a média de autores por trabalho foi de 1,7. Observa-se, ainda, que a relação autores/trabalho vem crescendo nos eventos pós-pandemia, atingindo 1,87 em 2022 e 1,83 em 2023, comparado a 1,67 em 2021 e 1,28 em 2000. Esses dados sugerem uma crescente tendência para publicações colaborativas, em detrimento das pesquisas de autoria única.

A Figura 3 representa a aplicação de robôs de IA para completar os metadados faltantes no registro coletado, tarefa esta antes realizada manualmente, passou a ser realizada por aplicações automatizadas. As palavras-chaves apresentadas na figura 3 foram retiradas do texto completo em PDF, que foi convertido em TXT e analisado por ferramentas de IA, no caso o ChatGPT4.

Figura 3 – Complementação dos metadados faltantes por Robôs - Exemplo



Fonte: elaborado pelos autores

Considerando a qualidade dos metadados disponibilizados pela base tem melhorado significativamente com cada ampliação das metodologias de organização da informação. No entanto, ainda há um longo caminho a ser percorrido. A Benancib propõe não apenas facilitar algumas análises diretamente na interface, mas também possibilitar a exportação de dados em diversos formatos, como CSV, XLS e DOC, que já estão implementados. Além disso, estão em desenvolvimento módulos para a exportação em outros formatos, como o RIS. Essas exportações habilitam metodologias como mineração de texto e análise de tópicos, permitindo explorar tendências temáticas ao longo do tempo e identificar áreas emergentes e em declínio dentro da Ciência da Informação. Considerando que a base concentra os principais pesquisadores de Ciência da Informação no Brasil, o desenvolvimento de indicadores científicos e acadêmicos robustos pode ser crucial para auxiliar instituições acadêmicas e órgãos de fomento a tomar decisões baseadas em evidências.

Perspectivas futuras

O BENANCIB, como uma base de dados na área de Ciência da Informação, abre várias perspectivas para estudos futuros e metodologias a serem exploradas, especialmente no contexto da bibliometria, cientometria e informetria. A base atualmente (abril de 2024) suporta uma variedade de análises bibliométricas e cientométricas, produzindo indicadores de produção e colaboração. Está em desenvolvimento um projeto para a criação de robôs capazes de identificar, coletar e integrar as citações de cada trabalho. Isso permitirá, no futuro, a realização de estudos de citações, análises de tendências teóricas e pesquisas sobre acoplamento bibliográfico, seja de palavras-chave ou de autores citados.

Considerações finais

O papel do BENANCIB na preservação e organização da memória científica do ENANCIB é incontestável, oferecendo uma plataforma vital para a disseminação e análise da produção científica em Ciência da Informação no Brasil. O projeto não só resgata, mas também valoriza a produção intelectual, fornecendo ferramentas robustas para a exploração de dados através de metodologias avançadas como mineração de texto e análise bibliométrica.

No entanto, muitos desafios persistem, principalmente relacionados à qualidade e padronização dos metadados. A iniciativa de implementar tecnologias avançadas, como inteligência artificial para aperfeiçoar a curadoria de metadados e a expansão de formatos de exportação, aponta para um futuro em que o acesso e a manipulação de dados podem se tornar mais eficientes e abrangentes.

A evolução contínua do BENANCIB, com a adição de novos módulos e aprimoramento dos existentes, demonstra um compromisso com a melhoria contínua. O envolvimento da comunidade acadêmica e a colaboração entre instituições são cruciais para o desenvolvimento sustentável da base, que se destaca como um recurso indispensável para pesquisadores da área.

Ao olhar para o futuro, a adoção de identificadores persistentes e a integração de novas tecnologias serão fundamentais para superar as barreiras atuais, permitindo que o BENANCIB continue a servir não só como um repositório de documentos, mas também como uma ferramenta de análise e descoberta científica.

Com a continuidade desses esforços, o BENANCIB está bem posicionado para ampliar seu impacto na comunidade científica, promovendo uma maior compreensão das tendências e evolução da Ciência da Informação no Brasil e no mundo.

Referências

Alvarez, G. R., & Caregnato, S. E. (2017). A Ciência da Informação e sua contribuição para a avaliação do conhecimento científico. Biblio: Revista do Instituto de Ciências Humanas e da Informação, 31(1), 9-26.

Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação. (2023, 03 de dezembro). Diretrizes gerais para o Enancib. Disponível em https://ancib.org/diretrizes-gerais/

Base de Dados do ENANCIB. (2024, 03 de março). [Busca]. Disponível em https://cip.brapci.inf.br/benancib

Bottentuit, A. M., Santos, P. L. V. A. C., & Jorente, M. J. (2008). Visualização da Ciência da Informação e seu prêmio científico. In Anais do IX Encontro Nacional de Pesquisa em Ciência da Informação. São Paulo, SP: ANCIB.

Gabriel Junior, R. F., & Vogel, M. J. M. (2022). BRAPCI-BENANCIB: base de dados de texto completo dos Enancib. In Anais do VIII Encontro Brasileiro de Bibliometria e Cientometria. Maceió, AL: UFAL. Disponível em http://hdl.handle.net/10183/257193

Instituto Brasileiro de Informação em Ciência e Tecnologia. Estudos Críticos em Informação, Tecnologia e Organização Social. (2021). Escritos tem presença marcante no XXI Enancib 2021. Disponível em http://escritos.ibict.br/escritos-tem-presenca-marcante-no-xxi-enancib-2021/

Ramalho, R. A. S. (2015). Ontologias e Simple Knowledge Organization System (SKOS): aproximações e diferenças. In J. A. C. Guimarães; V. Dodebei. (Org.), Organização do conhecimento e diversidade cultural (1a. ed., V. 1, p. 100-107). Marília, SP: ISKO-Brasil; FUNDEPE.

Ruberto, D. L. V. G.; Antoniazzi, R. L (2017). Análise e Comparação de Algoritmos de Similaridade Distância entre strings Adaptados ao Português Brasileiro. In: Escola Regional De Banco De Dados (ERBD), 13. , 2017, Passo Fundo. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2017. ISSN 2595-413X.

Gabriel Junior, R. F.; Laipelt, R. C. (2017) Thesa: ferramenta para construção de tesauro semântico aplicado interoperável. Revista P2P & Inovação, Rio de Janeiro, v. 3, n. 2, p.124-145, Mar./Set. 2017.

Rosa Helena Cunha Vidal

Graduação em Biblioteconomia pela Universidade Federal do Rio Grande do Sul (2008). Atualmente é bibliotecária - Goethe-Institut Porto Alegre e mestranda no Programa de Pós-Graduação em Ciência da Informação na Universidade Federal do Rio Grande do Sul. Tem experiência na área de Ciência da Informação, com ênfase em Ciência da Informação.

Rene Faustino Gabriel Junior

Graduado em Biblioteconomia e Documentação pela Pontifícia Universidade Católica do Paraná (2008), com mestrado em Ciência, Gestão e Tecnologia da Informação pela Universidade Federal do Paraná (2011) e doutorado em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho (2014). Atualmente é professor adjunto da Universidade Federal do Rio Grande do Sul e do Programa de Pós-Graduação em Ciência da Informação (PPGCIN) da mesma universidade e chefe do Departamento de Ciências da Informação (DCI) Tem experiência na área de Ciência da Informação, com ênfase em Biblioteconomia, atuando principalmente nos seguintes temas: Ciência da Informação, Estudos Métricos da Informação, Bibliometria, Brapci, Comunicação Científica, Dados de Pesquisa e Produção Científica. Implantou e coordena a Base de Dados de Periódicos em Ciência da Informação (BRAPCI). Membro do Grupo de Pesquisa de Comunicação Científica e do Núcleo de Estudos em Ciência, Inovação e Tecnologia da UFRGS.

e-mail: rene.gabriel@ufrgs.br

Mejora en la Precisión de IA mediante Acceso Optimizado a Datos de OJS: Análisis de Conexión Directa a Base de datos vs. OAI-PMH

Rafael Castillo Guerrero[1], *Francisco Garrido Sandoval*[2]

Palabras claves

Open Journal System, inteligencia artificial, procesamiento de lenguaje natural, OAI-PMH, Bases de datos relacionales

Open Journal System, artificial intelligence, Natural Language Processing, OAI-PMH, Relational databases

Eje temático

Inteligencia artificial (IA)

Experiencias en la implementación de técnicas y aplicaciones basadas en IA para la reducción de alucinaciones, mejorando la precisión en los resultados de búsqueda en revistas que utilizan OJS

Resumen

La decisión entre conectarse directamente a la base de datos de OJS o extraer los registros vía OAI-PMH para generar *word embeddings* depende de varios factores. Conectarse directamente ofrece acceso rápido y consultas flexibles, pero requiere permisos adicionales y puede cargar el servidor. En cambio, OAI-PMH es un protocolo estandarizado que facilita la interoperabilidad y reduce los riesgos de seguridad, aunque puede ser más lento y limitado en detalles específicos. Para optimizar el proceso y mejorar la precisión de la inteligencia artificial, reduciendo las alucinaciones o errores en sus predicciones, se recomienda una combinación de ambos enfoques. Usar OAI-PMH para una extracción inicial de gran volumen y establecer una base de datos local, y luego conectarse directamente a OJS para actualizaciones incrementales y consultas específicas. Esta estrategia aprovecha la seguridad y estandarización del OAI-PMH junto con la flexibilidad y rapidez del acceso directo a la base de datos, mejorando así la eficiencia y precisión en la generación de *word embeddings*. La calidad de los datos es crucial para el éxito del proyecto, asegurando que la IA esté bien entrenada y sea capaz de realizar tareas complejas con un alto grado de precisión, disminuyendo las alucinaciones.

Introducción

En la inteligencia artificial (IA), la calidad y el acceso eficiente a los datos son determinantes para el rendimiento y la precisión de los modelos. En particular, la generación de *word embeddings*, una técnica clave en el procesamiento del lenguaje natural (NLP), depende en gran medida de la disponibilidad de datos textuales ricos y bien estructurados. Open Journal System (OJS) es una plataforma ampliamente uti-

1 SISIB – Universidad de Chile - Chile rafael.castillo@uchile.cl
2 SISIB – Universidad de Chile - Chile francisco.garrido@uchile.cl

lizada para la gestión de revistas académicas y ofrece una fuente valiosa de datos textuales. Sin embargo, la metodología empleada para acceder a estos datos puede influir significativamente en la eficiencia del proceso.

Existen dos enfoques principales para acceder a los datos de OJS: conectarse directamente a la base de datos de OJS o extraer los registros vía OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). Cada uno de estos métodos presenta ventajas y desventajas específicas, que deben ser cuidadosamente consideradas en función de varios factores como la infraestructura disponible, la frecuencia de acceso a los datos, la cantidad de datos y la complejidad de las consultas.

La conexión directa a la base de datos de OJS permite un acceso rápido y consultas flexibles, proporcionando la capacidad de realizar consultas SQL personalizadas para extraer datos específicos en el formato deseado. Sin embargo, este método requiere permisos adicionales y puede presentar riesgos de seguridad, además de imponer una carga significativa en el servidor de la base de datos, afectando potencialmente su rendimiento general. Además, es necesario que los desarrolladores tengan un conocimiento completo de la estructura de datos y las tablas de la base de datos para poder realizar consultas eficaces y mantener la integridad de los datos.

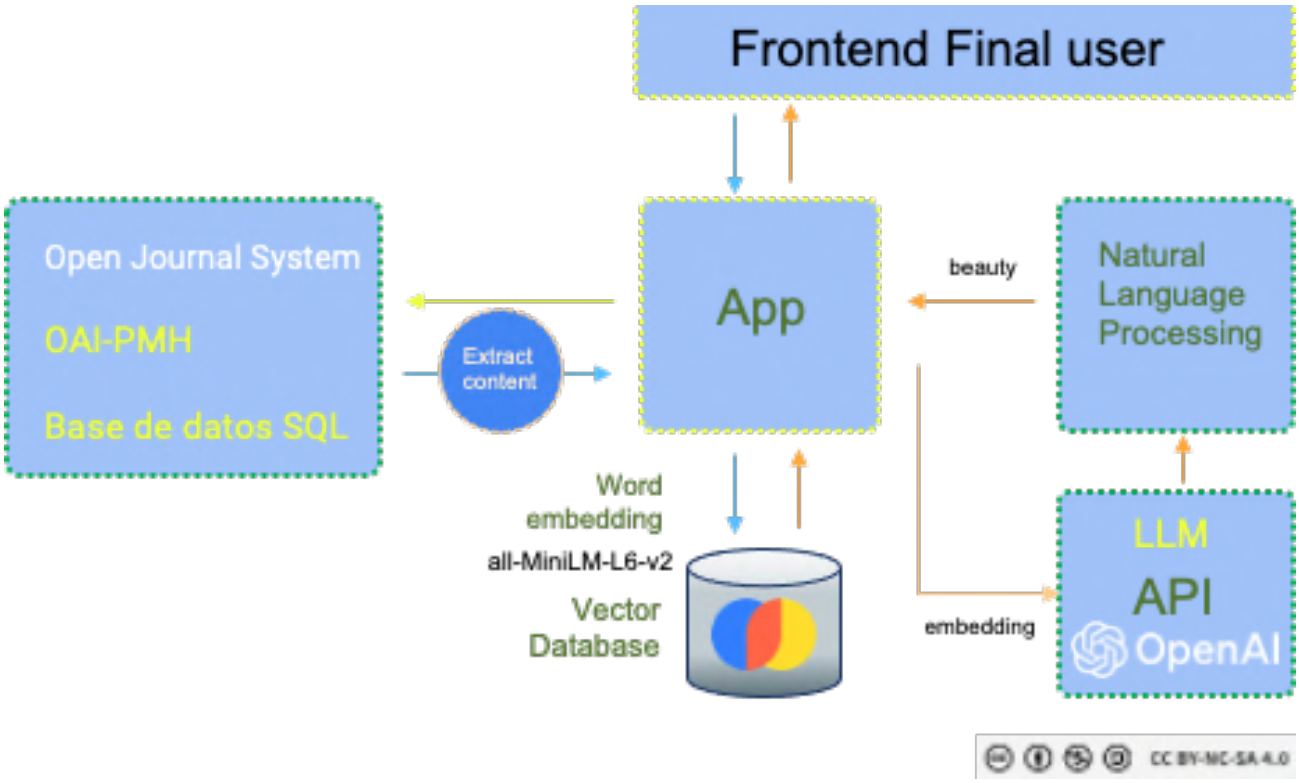
Por otro lado, la extracción de datos vía OAI-PMH es un protocolo estandarizado que facilita la interoperabilidad y reduce los riesgos de seguridad al limitar el acceso directo a la base de datos. No obstante, este enfoque puede ser más lento debido a la sobrecarga del protocolo y las limitaciones de la API, además de no siempre proporcionar acceso a todos los detalles específicos disponibles a través de consultas directas.

La eficiencia y la precisión en la generación de *word embeddings* no solo dependen de la accesibilidad y estructura de los datos, sino también de la capacidad para manejar grandes volúmenes de información y realizar actualizaciones continuas. La combinación de ambos enfoques puede ofrecer una solución óptima: utilizar OAI-PMH para una extracción inicial de gran volumen de datos y establecer una base de datos local, complementada con conexiones directas a OJS para actualizaciones incrementales y consultas específicas. Esta estrategia aprovecha la seguridad y estandarización del OAI-PMH junto con la flexibilidad y rapidez del acceso directo a la base de datos, mejorando así la eficiencia del proceso.

Al adoptar esta combinación de enfoques, no sólo se optimiza la generación de *word embeddings*, sino que también se mejora la precisión de los modelos de IA, reduciendo las alucinaciones o errores en sus predicciones. La calidad de los datos es crucial para el éxito del proyecto, asegurando que la IA esté bien entrenada y sea capaz de realizar tareas complejas con un alto grado de precisión. Este estudio aborda estas consideraciones y ofrece una guía sobre cómo implementar eficientemente estos métodos para maximizar los beneficios en el contexto del procesamiento de datos textuales provenientes de OJS.

Además, es importante destacar que los *word embeddings* generados a partir de estos datos se utilizan para almacenar información en una base de datos de vectores. Esta base de datos de vectores permite una búsqueda eficiente y rápida de información relevante, facilitando el acceso a contenido semánticamente similar y mejorando significativamente la capacidad de los sistemas de IA para comprender y procesar lenguaje natural en diversas aplicaciones.

Prototipo propuesto



La imagen muestra un esquema de integración de bases de datos relacionales con inteligencia artificial (IA). A continuación, se presenta una explicación detallada de los componentes y el flujo de datos representados en la imagen:

1. Open Journal System (OJS):

- OAI-PMH y Base de datos SQL: En la parte izquierda, se muestra que los datos de Open Journal System pueden ser extraídos utilizando dos métodos: OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) y acceso directo a una base de datos SQL. Ambos métodos proporcionan datos estructurados y necesarios para el análisis y procesamiento posterior.

2. Extracción de Contenidos:

- Un proceso de "Extract content" (extraer contenido) se encarga de recolectar datos tanto desde OAI-PMH como desde la base de datos SQL de OJS. Este proceso es el primer paso para obtener los datos que serán utilizados por el sistema de IA.

3. Aplicación (App):

- Los datos extraídos se envían a una aplicación central (App). Esta aplicación es el núcleo del sistema, donde se gestionan y procesan los datos para diferentes propósitos.
- Vector Database: La aplicación utiliza una base de datos de vectores (Vector Database) para almacenar los *word embeddings* generados. Específicamente, se menciona el modelo `all-MiniLM-L6-v2` para la creación de estos embeddings.

4. Procesamiento de lenguaje natural (NLP):

- La aplicación interactúa con herramientas de procesamiento de lenguaje natural (Natural Language Processing), lo que sugiere que los datos textuales son analizados y procesados para mejorar su calidad y extraer información relevante.
- El término "beauty" indica que los resultados del procesamiento NLP son devueltos a la aplicación, para mejorar la presentación y accesibilidad de la información.

5. API de Modelos de Lenguaje de Gran Escala (LLM API):

- Se muestra una integración con una API de OpenAI para modelos de lenguaje de gran escala (LLM API). Esto sugiere que los embeddings generados se utilizan o mejoran con un modelo de lenguaje avanzado provisto por OpenAI u otro modelo.
- Los embeddings generados por la aplicación se envían a la API de OpenAI u otro modelo para procesamiento adicional, y los resultados son devueltos para ser utilizados en la aplicación.

6. Interfaz de Usuario Final (Frontend Final User):

- Los resultados procesados por la aplicación son finalmente entregados a una interfaz de usuario final (Frontend Final User). Esta interfaz presenta los datos de manera accesible y útil para los usuarios finales.

Finalmente, mencionar que la imagen representa un sistema de integración donde los datos de Open Journal System son extraídos mediante OAI-PMH o acceso a bases de datos SQL, procesados para generar *word embeddings* y almacenados en una base de datos de vectores. Estos embeddings se utilizan en aplicaciones de procesamiento de lenguaje natural y se mejoran con una API de modelos de lenguaje de gran escala, para finalmente ser presentados a los usuarios a través de una interfaz amigable. Este sistema optimiza el acceso y la calidad de los datos para mejorar la precisión y la utilidad de los modelos de IA.

Resultados

Entre los resultados esperados de la implementación del sistema descrito se incluye lo siguiente:

1. Mejora en la Precisión de la IA:

- Generación de *word embeddings* de alta calidad: Al utilizar datos estructurados y enlazados extraídos tanto mediante OAI-PMH como acceso directo a bases de datos SQL, se espera generar *word embeddings* más precisos y representativos, lo cual mejorará la comprensión del lenguaje natural por parte de los modelos de IA sobre los temas que plantean las revistas estudiadas.

2. Reducción de alucinaciones en la IA:

- Datos estructurados y verificados: El uso de datos provenientes de Open Journal System, organizados y verificados mediante procesos estandarizados, debería reducir las alucinaciones o errores en las predicciones de la IA, al proporcionar un contexto más sólido y consistente.

3. Optimización del proceso de extracción y procesamiento de datos:

- Eficiencia en la extracción de datos: La combinación de métodos de extracción vía OAI-PMH y acceso directo a bases de datos SQL asegura un proceso eficiente y flexible, adaptado a diferentes necesidades de actualización y consulta de datos.
- Almacenamiento eficiente en base de datos de vectores: El uso de una base de datos de vectores para almacenar los *word embeddings* permite un acceso rápido y eficiente a los datos procesados, facilitando tareas de búsqueda y recuperación de información relevante.

4. Interoperabilidad y escalabilidad del sistema:

- Interoperabilidad: El uso de estándares como OAI-PMH facilita la interoperabilidad con otros sistemas y fuentes de datos, permitiendo una integración más amplia y flexible.
- Escalabilidad: La arquitectura del sistema permite escalar tanto en volumen de datos como en capacidad de procesamiento, adaptándose a crecientes necesidades de datos y análisis.

5. Acceso Mejorado a Información Relevante:

- Búsqueda y recuperación eficiente: Los usuarios finales pueden acceder a información relevante de manera rápida y eficiente, gracias a la base de datos de vectores y las capacidades de procesamiento de lenguaje natural.
- Presentación de datos: Los resultados del procesamiento son presentados de manera clara y útil, mejorando la toma de decisiones y la utilización de la información disponible.

En conjunto, estos resultados se traducen en un sistema más robusto y efectivo para el manejo y procesamiento de datos textuales, que mejora significativamente la precisión y la fiabilidad de los modelos de inteligencia artificial.

Conclusiones y trabajo futuro

Este trabajo ha demostrado la importancia de la integración eficiente de bases de datos relacionadas y el uso de técnicas avanzadas de procesamiento de lenguaje natural para mejorar la precisión y reducir las alucinaciones en los modelos de inteligencia artificial (IA). Al combinar los métodos de extracción de datos vía OAI-PMH y el acceso directo a bases de datos SQL de Open Journal System (OJS), hemos logrado optimizar el acceso y la calidad de los datos textuales necesarios para generar *word embeddings*.

Las principales conclusiones por mencionar son las siguientes:

1. Mejora en la precisión y reducción de alucinaciones: La calidad de los datos estructurados y verificados provenientes de OJS ha permitido generar *word embeddings* más precisos, mejorando la comprensión del lenguaje natural y reduciendo los errores en las predicciones de la IA.
2. Eficiencia en la extracción y procesamiento de datos: La combinación de métodos de extracción ha permitido un proceso eficiente y flexible, adaptado a las necesidades de actualización continua y consulta específica de datos.
3. Optimización del almacenamiento de datos: El uso de una base de datos de vectores para almacenar *word embeddings* ha facilitado el acceso rápido y eficiente a la información procesada, mejorando las capacidades de búsqueda y recuperación de datos.

4. Interoperabilidad y escalabilidad: El sistema diseñado es interoperable y escalable, lo que permite su integración con otras fuentes de datos y su adaptación a crecientes necesidades de datos y análisis.
5. Mejora en la experiencia de usuario: Los datos procesados y enriquecidos han sido presentados de manera accesible y útil a los usuarios finales, optimizando la toma de decisiones y la utilización de la información.

Bibliografía

Apuntes de Python. (s.f.). Trabajo con bases de datos SQL en Python: Interacción con bases de datos relacionales. Accedido el 26 de mayo de 2024, desde https://apuntes.de/python/trabajo-con-bases-de-datos-sql-en-python-interaccion-con-bases-de-datos-relacionales/#gsc.tab=0

KDnuggets. (2023, mayo 18). Ollama Tutorial: Running LLMs Locally Made Super Simple. Accedido el 24 de mayo desde https://www.kdnuggets.com/ollama-tutorial-running-llms-locally-made-super-simple

LangChain. (s.f.). Graph use cases: Semantic search. Accedido el 25 de mayo de 2024, desde https://python.langchain.com/v0.1/docs/use_cases/graph/semantic/

LangChain. (2023, abril 18). How to build the ultimate AI automation with multi-agent collaboration. LangChain Blog. Accedido el 26 de mayo de 2024, desde https://blog.langchain.dev/how-to-build-the-ultimate-ai-automation-with-multi-agent-collaboration

MyScale. (2023, marzo 15). What is SQL vector databases? Accedido el 12 de abril de 2024 desde https://myscale.com/blog/what-is-sql-vector-databases/

Open Archives Initiative. (s.f.). *The Open Archives Initiative Protocol for Metadata Harvesting*. Accedido el 26 de mayo de 2024, desde https://www.openarchives.org/pmh/

Public Knowledge Project. (s.f.). *Open Journal Systems*. Accedido el 26 de mayo de 2024, desde https://pkp.sfu.ca/software/ojs/

Severance, C. (s.f.). Bases de datos. Python para todos. Accedido el 24 de mayo de 2024, desde https://es.py4e.com/html3/15-database

Urrego, N. (2021, enero 20). Introducción a las bases de datos relacionales: Entendiendo su estructura. Medium. Accedido el 23 de abril de 2024 desde https://nicolasurrego.medium.com/introducci%C3%B3n-a-las-bases-de-datos-relacionales-entendiendo-sus-estructura-57f978be069a

Voita, L. (s.f.). Word embeddings. Accedida el 26 de abril de 2024, desde https://lena-voita.github.io/nlp_course/word_embeddings.html

Vout, T. (2023, March 15). *ChatGPT's 'Snow White' Problem: The Danger of Common Knowledge.*. Oxford Semantic Technologies. Accedido el 27 de mayo de 2024, desde https://www.oxfordsemantic.tech/blog/chatgpts-snow-white-problem-the-danger-of-common-knowledge

Rafael Castillo Guerrero, Jefe de la Unidad de Gestión de Bibliotecas – SISIB, de la Universidad de Chile.

Bibliotecario, Diseñador Gráfico y Máster en Gestión de la Información

Durante la práctica de la bibliotecología, se dedicó al estudio y desarrollo de páginas web, destacándose luego como consultor en el desarrollo de catálogos de bibliotecas en Internet, asesorando a instituciones públicas y privadas. Realizó trabajo académico durante 8 años en Santiago de Chile y ha sido conferencista en numerosas ocasiones. Últimamente, se ha enfocado en el estudio de ontologías para la descripción de la información, así como en el estudio de Bibframe, en el cual ha realizado presentaciones en Santiago, Buenos Aires y Lima. Además, es un entusiasta del estudio de la inteligencia artificial con énfasis en la recuperación y de información. Actualmente, ocupa el cargo de Jefe de la Unidad de Gestión Bibliotecaria en la Universidad de Chile.

Francisco Garrido Sandoval, Ingeniero desarrollador – SISIB, de la Universidad de Chile.

Ingeniero Civil en Informática de la Universidad del Bio-Bio

Mi área de interés es la gestión de información y la coordinación para el desarrollo de proyectos de TI. Tengo una amplia experiencia en el área de análisis y desarrollo de prototipos para sistemas de información relacionados con bibliotecas y centros de documentación. He participado activamente en la definición y captura de requisitos, además de la creación y modificación de código fuente para aplicaciones web y el mantenimiento posterior de estos sistemas en plataformas como: Dspace, Open Journal System, Open Monograph, Dataverse, diversos gestores de contenido como Wordpress y Joomla, entre otros. Tengo experiencia en varios lenguajes de programación como Java, Php, Python, Ruby, Javascript, y también tengo un nivel avanzado en hojas de estilo, XML, Json. Además, poseo experiencia en el desarrollo de portales de la Web Semántica. En cuanto a la gestión de OJS y DSPACE, tengo un nivel experto con 12 años de experiencia en la instalación y personalización de los softwares en sus diferentes versiones. Estoy capacitado para generar y recuperar información directamente desde la base de datos, ya que tengo conocimientos avanzados de sus estructuras.

Comunicación académica, científica y cultural en abierto

Repositório bilíngue em língua de sinais: formação na perspectiva inclusiva

Tania Chalhub[1], Maria José Veloso da Costa Santos[2]

Palabras claves

Objetos bilíngues, Educação de surdos, Repositórios educacionais, Surdos, Biblioteconomia social.

Eje temático

Comunicación académica, científica y cultural en abierto

Resumen

A Ciência da Informação com as conquistas do Open Access e tecnologias da informação e comunicação impactaram significativamente a acessibilidade com compartilhamento de informações acessível. A Língua Brasileira de Sinais, língua para educação e comunicação dos surdos, Lei nº 10.436/2002. A criação dos repositórios possibilitou a comunicação de conteúdo em língua de sinais e criação de canais de comunicação e divulgação destes materiais como o Repositório Huet. Porém a formação de profissionais da informação que tem como parâmetro a informação escrita de língua oral. O objetivo é descrever um projeto de extensão da Biblioteconomia da UFRJ e do Instituto Nacional de Educação de Surdos para formação de profissionais no tratamento de informações em Libras para repositório bilíngue. Pesquisa descritiva com abordagem qualitativa, baseada na observação participante e análise documental dos relatórios do projeto de extensão da UFRJ-INES. Foram realizadas atividades de planejamento do tratamento dos objetos, discutidos os metadados e a inserção com análise dos objetos bilíngues. Cada etapa foi desenvolvida com base discussão em encontros da equipe e com alunos surdos do curso de pedagogia. Foram inseridos diferentes materiais, Programas da TV INES, revistas editadas pelo INES, artigos diversos em acesso aberto e Trabalho de Conclusão de Curso da Pedagogia.

Introdução

No século XXI representa significativo potencial de acessibilidade informacional para diferentes grupos da sociedade, minoria muitas das vezes invisibilizadas pelas suas configurações informacionais diferenciadas como as pessoas com deficiência sensorial, principalmente as pessoas com deficiência auditiva e visual. As tecnologias da informação e comunicação (TIC) que possibilitaram os avanços do Movimento de Acesso Aberto iniciado no século XX, possibilitaram que realidade comunicacional dos sujeitos surdos fosse potencializada uma vez que a base de sua comunicação se estabelece no campo da visualidade, principalmente vídeos para os diálogos em línguas de sinais. A Língua Brasileira de Sinais (Libras) se concretizou enquanto língua reconhecida para a educação e comunicação dos surdos pela Lei nº 10.436 de 24 de abril de 2002 (Brasil, 2002), sendo regulamentada pelo Decreto nº 5.626, 2005 (Brasil, 2005), reconhecendo

1 UFPA e INES, chalhubtania@gmail.com
2 UFRJ, msantos1402@facc.ufrj.br

como a língua de comunicação dos surdos brasileiros e o direito de acesso à educação nesta língua visual. A educação de surdos se beneficiou dos avanços da tecnologia, principalmente neste século com a possibilidade da comunicação de conteúdo acadêmico em língua de sinais.

O Instituto Nacional de Educação de Surdos (INES) foi criado em 1857 por iniciativa do professor surdo francês E Huet (1822-1882) para institucionalizar a educação de surdos no âmbito da educação pública e é um dos atores responsáveis pelo debate e práticas pedagógicas para a educação de surdos. O INES oferece educação bilíngue em todos os segmentos, da educação infantil à formação para profissionais surdos e ouvintes no Curso Bilíngue de Graduação em Licenciatura em Pedagogia, além de cursos de pós-graduação lato e stricto sensu. Com tradição de vasta produção e divulgação de material pedagógico, fonoaudiológico e de vídeos em Libras ao longo dos seus 167 anos dedicados à educação de surdos, desenvolveu o Repositório Digital Huet para agregar e disponibilizar materiais acadêmicos e culturais, com a utilização de sistema de código livre para armazenar e disponibilizar materiais bilíngues em diversos formatos, textos, vídeos, imagens etc.

A nova política educacional de inclusão de pessoas com deficiência (PcD) no ensino regular estabelece que estudantes devem receber instrução em classes inclusivas e que os materiais atendam suas necessidades informacionais, ou seja, pessoas com deficiência visual devem ter materiais acadêmicos em braille, com fonte ampliada por exemplo, enquanto as pessoas com deficiência auditiva devem ter acesso aos materiais em Libras, com imagens etc. Considerando que Libras ainda é uma língua com produção acadêmica incipiente e a dispersão dos materiais disponíveis em diferentes instituições de ensino no Brasil, o INES desenvolveu um repositório temático para e sobre educação de surdos de livre acesso. Utilizou-se no trabalho a nomenclatura surdos para designar as pessoas com deficiência auditiva por ser a escolha dos sujeitos da pesquisa. Esta é uma escolha pautada na perspectiva de surdez como uma questão cultural e não de doença defendida por pesquisadores surdos como Campello (2008), Perlin e Strobel (2014), Silveira e Campello (2015), Stumpf e Wanderley (2016) e ouvintes Quadros (2005) Gesser (2009).

O desenvolvimento de repositórios é considerado de suma importância na literatura de Biblioteconomia e Ciência da Informação (BCI) porque reúne em um único sistema os objetos desenvolvidos por uma especialidade e produzidos por instituições. Para tal, necessário se faz a descrição física e de conteúdo de objetos informacionais compatíveis com os mesmos fundamentos que regem os princípios e instrumentos utilizados para a implementação de catálogos on line e de acesso público em unidades de informação, que nada mais são que Sistemas de Recuperação da Informação (SRI) que permitem a comunicação do usuário com o acervo, avanços já alcançados nas áreas de BCI, associadas aos recursos das Tecnologias da Informação e da Comunicação (TIC) que apresentam soluções inovadoras e bem sucedidas. A Representação da informação, prática denominada nos repositórios como indexação, faz a intermediação entre documentos e usuários, isso porque a representação das características do documento, como a descrição física e de conteúdo é realizada de forma compacta o que vem a propiciar a substituição do documento por registros bibliográficos que, uma vez recuperados, levam o usuário ao texto completo do documento. (Alvarenga, 2003).

A presente pesquisa discute a experiência desenvolvida por meio de parceria do INES com o Curso de Biblioteconomia e Gestão de Unidades de Informação (CBG) da Universidade Federal do Rio de Janeiro (UFRJ), Brasil, fruto do projeto de extensão intitulado Repositório Bilíngue para Educação de Surdos: mecanismo de inclusão, apoiado pela Pró-Reitoria de Extensão da UFRJ com bolsas de extensão para alunos do CBG. O projeto tem como objetivo desenvolver ações de extensão que visam a organização e a representação de objetos digitais bilíngues (Libras e Português) para o povoamento do Repositório Huet,

observando os padrões internacionais de tratamento técnico de objetos digitais segundo técnicas biblioteconômicas e parâmetros de informação dos sujeitos surdos representados por docentes e discentes de Pedagogia. Nessa perspectiva, o projeto garante maior impacto na visibilidade e na recuperação da informação no Repositório Digital Huet, bem como no gerenciamento das interações do público alvo com as redes sociais utilizadas pelo Repositório. O Repositório Digital Huet é assim denominado tendo em vista a criação do INES por iniciativa do surdo francês É. Huet (1822-1882).

O artigo justifica-se uma vez que o acervo de objetos digitais bilíngues é crescente nas instituições especializadas e a reunião e tratamento técnico no Repositório Huet vem possibilitando à comunidade surda o acesso livre a essa fonte de materiais institucionais em variados suportes, procedendo com isso a um trabalho social relevante para a educação, produção e troca de conhecimentos entre a comunidade surda e de ouvintes. A implementação da presente ação de extensão propiciou o emprego de novas práticas pedagógicas com alunos surdos, numa perspectiva de professor, alunos e pesquisadores, produtores e consumidores de conhecimento.

O texto encontra-se estruturada nas seguintes seções: a primeira referente à Introdução. A segunda descreve a acessibilidade à informação na perspectiva da pessoa surda e descreve o Repositório Huet. A terceira seção relata a experiência de parceria do INES com o Curso de Biblioteconomia e Gestão de Unidades de Informação (CBG) da UFRJ, o projeto de extensão e a atuação dos estudantes envolvidos. A quarta é referente aos procedimentos metodológicos, seguida da sexta seção que apresenta e discute os resultados. A última seção, a sete discorre sobre as considerações finais e em seguida a seção referente à Bibliografia.

Objetos digitais para surdos: Acessibilidade via repositório temático

Abordar o tema acessibilidade para surdos por meio de práticas pedagógicas requer reflexão profunda sobre diversos aspectos sociais, políticos e culturais que permeiam esse tema. Um dos elementos principais é a abordagem inclusiva que envolve não apenas o uso de tecnologias digitais, mas também a garantia de que estas tecnologias sejam projetadas e utilizadas de maneira a atender às necessidades específicas da comunidade surda, como o uso de interfaces visuais claras, suporte para língua de sinais e recursos de acessibilidade, como legendas e transcrições automática nos materiais didáticos. Nesse sentido, a pedagogia para alunos surdos deve ser centrada na língua de sinais e na cultura surda, reconhecendo e valorizando suas identidades e experiências, o que pode envolver a criação de materiais educacionais adaptados, o desenvolvimento de estratégias de ensino que promovam a participação ativa dos alunos surdos e a formação de professores capacitados em língua de sinais e educação bilíngue.

O Repositório Huet, objeto deste projeto, foi planejado pelo INES, visando a agregação e compartilhamento de objetos para educação de surdos produzidos no Instituto e em outras instituições comprometidas com a educação deste grupo de cidadãos, cuja identidade linguística demanda uma nova concepção de comunicação. Essa iniciativa foi debatida entre diferentes profissionais das áreas de Pedagogia, Informática, Linguística, Educação de surdos e de BCI e que frutificou na criação de um repositório de acesso aberto – o Repositório Digital Huet - que disponibiliza a produção acadêmico científica e cultural de qualquer área do conhecimento produzidos pelo INES ou por outras instituições de ensino e pesquisa para a comunidade surda, constituindo-se, portanto, de um repositório temático – educação de surdos. Representar objetos informacionais no repositório passa pela capacidade de simbolizá-los e identificá-los a partir de um conjunto de elementos que identifiquem seus atributos de modo a permitir sua recuperação,

focada na "interatividade do usuário com os sistemas de informação e as interfaces de busca [...]" (Rosseto, 2003, p.3), o que vem assegurar a confiabilidade e a precisão na recuperação da informação, acrescido da perspectiva de preservação do conhecimento registrado para as próximas gerações.

O Repositório Huet foi desenvolvido com o software DSpace (http://www.dspace.org), "projeto cooperativo liderado pelas bibliotecas do Massachusetts Institute of Technology (MIT) e pelos laboratórios da corporação Hewlett-Packard (HP), conduzido sob as diretrizes da DSpace Federation". É descrito como "um sistema [...] inovador que captura, armazena, indexa, preserva e redistribui materiais de pesquisa em formato digital produzida por comunidades acadêmicas dentro do contexto de organizações de pesquisa e de universidades". (Sayão e Marcondes, 2009, p. 44). É distribuído pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), que disponibiliza para a comunidade acadêmica gratuitamente uma versão em português. O DSpace foi customizado para atender ao parâmetro de visualidade de surdos falantes de Libras, tendo como um dos principais elementos ser uma ferramenta de armazenamento e recuperação de objetos bilíngue. (DSpace, 2024).

Para a descrição dos objetos digitais e sua inserção (indexação) no repositório, os especialistas do INES e da Fundação Oswaldo Cruz (FIOCRUZ) optaram pelo uso de metadados, que segundo Márdero Arellano (2009) visam cumprir a função básica de prover informação sobre o documento digital e que alimenta os processos de gestão, recuperação e reprodução. Assim, foi selecionado o esquema de metadados Dublin Core (DC), por apresentar simplicidade na descrição dos recursos, interoperabilidade semântica, consenso internacional de uso e extensibilidade. Sua composição geral é de um conjunto mínimo de 15 elementos propostos em consenso por um grupo multidisciplinar internacional de bibliotecários, analistas, museólogos e linguistas, entre outros, para dar conta de descrever os conteúdos com a finalidade de descoberta e navegação, possuindo uma variedade de recursos. Os 15 elementos do DC são: Título, Criador, Assunto, Descrição, Editor, Contribuinte, Data, Tipo, Formato, Identificação, Fonte, Idioma, Relação, Cobertura, Direitos. A esses 15 elementos podem ser incluídos qualificadores adicionais, para atender as particularidades do usuário. (Dublin Core, 2024). As premissas para que o material seja incluído no repositório é a de que o material deva estar preferencialmente em formato bilíngue (Libras e português) e seja produzido por uma instituição. Dessa forma, o Repositório Digital Huet visa atender à "demanda de um grupo que apresenta realidade linguística significativamente distinta […] com desigualdade de acesso à produção científica e educacional.". (Chalhub; Dionysio, 2023, p. 9).

O desenvolvimento do Repositório temático seguiu a metodologia utilizada por Sales (2011) que definiu as seguintes etapas para a construção de repositórios dessa natureza.

Seleção de software de gestão de repositórios (DSpace) além de aquisição de equipamento, instalação;

Planejamento do repositório;

Definição de políticas;

Definição de materiais digitais a serem incluídos;

Definição das coleções;

Definição de serviços oferecidos;

Implantação do repositório;

Capacitação e divulgação do repositório.

Segundo ainda Chalhub e Dionysio (2023, p.20), "foi um longo processo de dois anos em que foi decidido coletivamente desde o layout da página inicial, com acessibilidade em Libras, centralidade no visual e pouca informação textual, até a definição dos materiais e organização das comunidades e das coleções" (2023, p. 20). As autoras apontam para a necessidade de avaliação, aperfeiçoamento e ampliação da comunicação para garantir o acesso a informações educacionais, científicas e culturais.

A parceria UFRJ/INES e o treinamento de profissionais da Biblioteconomia na indexação de materiais em língua de sinais

O projeto Repositório Bilíngue para Educação de Surdos: mecanismo de inclusão tem como foco ações de extensão que integram discentes e docentes em duas áreas do conhecimento, numa abordagem interdisciplinar, Biblioteconomia e Pedagogia da UFRJ e INES, respectivamente, fruto de parceria entre as duas instituições. O trabalho conjunto possibilita ampliar a visibilidade e a recuperação da informação do Repositório Digital Huet, permitindo a democratização da informação acadêmica e cultural em língua de sinais para surdos, sejam eles professores e alunos em nível nacional de forma ampliada, articulando saberes da Biblioteconomia no conhecimento operacional nas áreas de tratamento técnico e acesso livre à informação e da Pedagogia Bilíngue com o conhecimento complexo sobre questões inerentes à comunidade surda e aos objetos educacionais para educação desses sujeitos, o que possibilita a interdisciplinaridade e a interprofissionalidade do projeto. Seguindo essa linha, o objetivo central do projeto é implementar, fomentar e consolidar a descrição de objetos digitais, de forma a garantir a ampliação na disponibilização desses materiais e possibilitar maior impacto na consulta, recuperação e no uso desses objetos digitais de aprendizagem incluídos no Repositório. (Santos, 2022).

O projeto de extensão ora descrito permite a interação dialógica da universidade e suas relações com comunidades sociais, proporcionando a troca de saberes entre esses dois setores, no caso do presente projeto, da UFRJ e a comunidade surda. A troca de expertises entre docentes e discentes das áreas envolvidas acredita-se que potencializa seu impacto em diferentes espaços educacionais, extrapolando os muros institucionais com maior troca de saberes com outras comunidades.

A relação de ensino pesquisa e extensão no projeto concretiza-se por pesquisas orientadas pelos profissionais envolvidos e pelo desenvolvimento de atividades práticas para os estudantes, aliadas aos ensinamentos teóricos apreendidos em sala de aula, o que permite associar à sua aprendizagem ações sobre a realidade do cotidiano de uma comunidade especial, que contribui para a formação cívica, política e cidadã, importantes ao futuro exercício profissional dos alunos envolvidos.

Procedimentos metodológicos

A pesquisa caracteriza-se como descritiva que segundo Braga tem por característica "descrever o comportamento dos fatos e fenômenos" (2007, p. 25) com abordagem quali-quantitativa das ações e reflexões dos atores envolvidos bem como dos objetos tratados. A escolha pela abordagem múltipla se deve por concordarmos com Braga (2007, p. 29) que diz

> [...] a tarefa de escolher, descrever e aplicar uma metodologia adequada [é] uma das fases mais delicadas do planejamento da pesquisa. Metodologias quantitativas e qualitativas não devem ser consideradas concorrentes nem tampouco excludentes, podendo ser aplicadas de maneira concomitante na pesquisa social, desde que respondam adequadamente ao objetivo estabelecido.

A coleta de dados foi baseada na observação participante, documentos, revisão bibliográfica e relatórios estatísticos do sistema e para cumprir seus objetivos de contribuir de forma efetiva com a comunidade surda brasileira, o projeto de extensão Repositório Huet, em continuidade, é composta das seguintes etapas:

Etapa 0 - reuniões semanais para apresentação do projeto, definição das estratégias de atuação dos bolsistas, leituras especializadas na área de Representação da Informação e Comunidade Surda para embasamento teórico dos estudantes;

Etapa 1– desenvolvimento de uma política de desenvolvimento de coleções para entrada no Repositório Huet, principalmente no que diz respeito à seleção de objetos digitais de aprendizagem a serem incluídos no Repositório - será realizada por pessoal especializado do INES, tendo como premissa de que o objeto seja produzido por uma instituição;

Etapa 2 - treinamento dos membros do projeto em Língua Brasileira de Sinais (Libras) - será realizado pelos professores do INES envolvidos;

Etapa 3 – visitas de observação a salas de aula do INES para a percepção da interação entre a comunidade surda e de ouvintes;

Etapa 4 – pesquisa de objetos digitais indexados em outros repositórios – detectar a existência do objeto digital a ser descrito em outro acervo ou repositório de forma a evitar o retrabalho;

Etapa 5 – Tratamento técnico dos objetos digitais de aprendizagem a serem disponibilizados no Repositório – realizado pelos bolsistas com a supervisão dos docentes do CBG e do INES envolvidos no projeto;

Etapa 6 – Acompanhamento – o acompanhamento é realizado por meio de reuniões mensais da equipe de professores e em reuniões ampliadas com todo o grupo de forma presencial e remota.

Etapa 7 – Avaliação do projeto – a avaliação das atividades desenvolvidas é realizada junto ao público-alvo por meio de formulários eletrônicos elaborados pela equipe;

Etapa 8 – Socialização das atividades desenvolvidas - é realizada por meio de comunicações apresentados em eventos e pela publicação de artigos e capítulos de livros;

Etapa 9 – Avaliação de bolsistas – é realizada pela apresentação de relatórios estatísticos mensais, pelo desenvolvimento de atividades propostas no plano de trabalho e pela participação em eventos, como Jornada de Iniciação Científica (JIC) e Semana de Integração Acadêmica (SIAC) da UFRJ

Etapa 10 - Monitoramento das redes sociais do Repositório – esta etapa ainda está em fase de planejamento devido a questões técnicas externas à dinâmica do projeto.

Resultados e discussões

A experiência do projeto de extensão teve início em 2022 com a seleção de uma aluna do Curso de Biblioteconomia como bolsista e a participação de estudante que ficou em segundo lugar na seleção realizada para bolsistas de extensão e que se integrou ao projeto de forma voluntária pela motivação de conhecer mais sobre educação de surdos, comunicação em língua de sinais. Ambos estavam muito envolvidos nas discussões sobre a educação inclusiva, importância da comunicação em língua de sinais para surdos.

Como primeiro movimento os alunos extensionistas visitaram o INES para conhecer os alunos, a dinâmica das aulas curso de Pedagogia na modalidade presencial e produção de materiais bilíngues para o curso de Pedagogia na modalidade EAD. As visitas à instituição possibilitaram a apreensão da rotina e dos atores envolvidos. Essa participação foi fundamental para que pudessem conhecer na prática os conceitos basilares da educação de surdos, a visualidade, a cultura surda e a língua visuoespacial.

Os estudantes ligados ao projeto de extensão interagiram com os estudantes do curso de Pedagogia em atividade de sala de aula, exercitaram seus conhecimentos de Libras básica e foram recebidos com muito entusiasmo pelos alunos surdos que se sentiram respeitados pela comunicação na sua própria língua. Essas atividades estavam relacionadas ao processo de conhecimento sobre a realidade comunicacional singular dos surdos falantes de língua de sinais visando a possibilitar uma atuação como profissionais da informação mais sensível à demanda de acessibilidade informacional. Um dos resultados dessas interações foi a criação de uma proposta visual para a mídia digital.

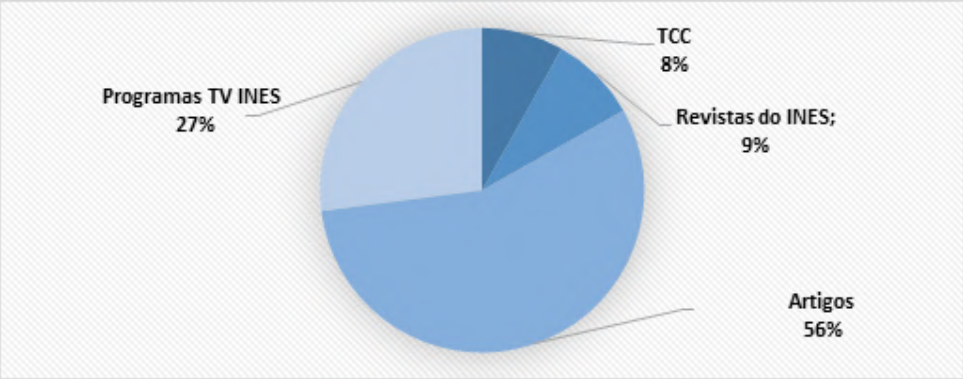
Foram realizadas reuniões presenciais no setor de produção de materiais didáticos digitais que conta com equipe multiprofissional com professores, designer instrucional, designer gráfico, editor de vídeos, profissional de legendagem, videografista, intérpretes de Libras e tecnologia da informação. Nas reuniões os estudantes conheceram a dinâmica de produção, os profissionais e os diferentes tipos de materiais, informações que seriam importantes na hora de procederem o tratamento dos materiais.

É importante destacar como o processo de aprendizagem de tratamento de materiais diferenciados demonstra o compromisso dos estudantes com relação à equidade da informação para a comunidade surda. Essa integração e compromisso podem ser comprovados pela citação de um aluno da UFRJ em seu relatório anual de 2023: "Destacamos a importância da adaptação padronizada e estruturada dos documentos, reconhecendo a necessidade fundamental de atender às especificidades da comunidade surda para otimizar a recuperação de informações. Este projeto não apenas atua como um mecanismo inclusivo, mas também potencializa o impacto na formação dos estudantes, promovendo a troca de saberes entre docentes e discentes envolvidos".

Antes da inclusão dos objetos digitais no Repositório Huet foi realizado um estudo de desenvolvimento de coleções para que os materiais fossem avaliados nos seus aspectos educacionais bem como informacionais, estimulando, de certa forma, o pensamento crítico e a análise de diferentes perspectivas da equipe.

No período de maio de 2022 a abril de 2024 foram inseridos diferentes materiais, Programas da TV INES, Revista Arqueiro e Revista Fórum editadas pelo INES, artigos diversos destas e de outras revistas em acesso aberto e Trabalho de Conclusão de Curso (TCC) do curso de Pedagogia. O gráfico 1, a seguir ilustra o quantitativo de materiais incluídos no Repositório Huet.

Gráfico 1. Inclusão de materiais no Repositório Huet 2022-224



Fonte: Dados da pesquisa

O tipo de material mais inserido foi o artigo de periódico, com 56 %. A inserção desse tipo de material se deve pela demanda de textos acadêmicos sobre educação de surdos e em Libras. Realizou-se as analíticas dos artigos, um a um e por ano de publicação nas três revistas do Instituto: Revista Espaço, Revista Arqueiro e Revista Fórum, de acordo com o capítulo 13 do Código de Catalogação Anglo-Americano, dando-se destaque à Revista Fórum que publica os artigos em formato bilíngue, seja vídeo do artigo em Libras e Resumo em Português ou vice-versa, (Figura 1).

Figura 1. Artigo da Revista Forum em formato bilíngue disponibilizado via QRCode


PROCESSO DE AQUISIÇÃO DA LINGUAGEM DO SUJEITO SURDO: UMA ANÁLISE DO BILINGUISMO COMO PRÁTICA SOCIAL

Letícia Fernandes Alvarenga Monteiro
Alice Caldas de Morais Sodré Coutinho
Juliana Felicidade Teixeira dos Santos
Karen Loami Lima da Silva
Tathianna Prado Dawes

RESUMO EM LIBRAS

RESUMO

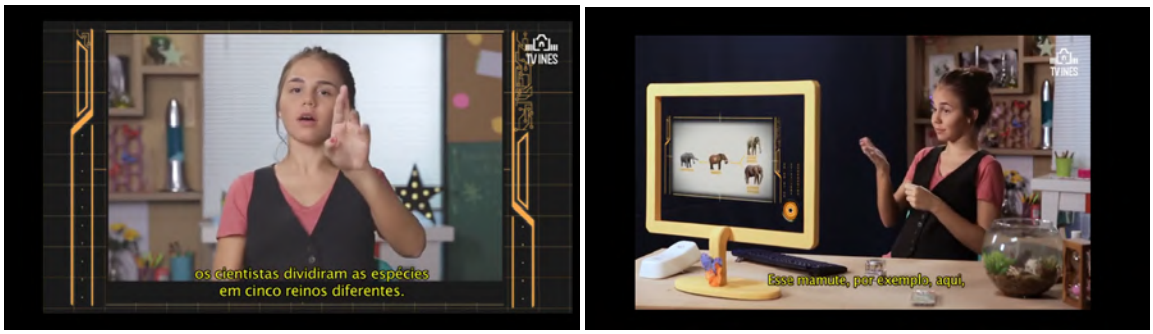
A educação bilíngue de surdos tem sido tema de debate recorrente entre os pesquisadores da área de surdez. Com o intuito de aprofundar esse debate na comunidade acadêmica da Universidade Federal

Fonte: Revista Forum, v. 38, 2018

http://repositorio.ines.gov.br/ilustra/handle/123456789/1116

Os programas da TV INES incluídos no Repositório foram: De olho na Ciência, A Vida em Libras e A História das Coisas, todos vídeos em Libras com áudio e legendas em português. São programas de caráter educativos e culturais que tem como parâmetros os elementos de visualidade da comunidade surda. A figuras 2 A e 2 B, mostram os prints de tela de um episódio do programa De Olho na Ciência.

Figuras 2 A e 2 B. Print de tela do vídeo do programa De Olho na ciência: biodiversidade (Parte 1).



Fonte: Repositório Digital Huet

http://repositorio.ines.gov.br/ilustra/handle/123456789/1240

Este episódio do programa De Olho na Ciência foi inserido na coleção Ciências Biológicas com as palavras-chave Seres vivos, Biodiversidade, Evolução e TV INES, com população-alvo a Educação Básica e Educação Continuada. O material teve a seguinte descrição: "Vídeo com duração de 11 minutos e 58 segundos. Apresentado em Libras com áudio e legendas em Português." A Libras e a visualidade estão presentes no vídeo que tem efeitos visuais para tornar a mensagem mais clara e a língua de sinais é central. Para preenchimento do metadado para registro das línguas foram utilizados dois padrões, para Libras o sgn_BR e para português o pt_BR.

A participação de discentes e docentes da Biblioteconomia na atividade de tratamento dos materiais digitais para a inserção no Repositório se desenvolveu junto com docentes e discentes do curso de Pedagogia bilíngue para surdos. Foi uma experiência acadêmica ímpar com desafios para as duas áreas do conhecimento.

Considerações Parciais

O projeto de extensão está em sua segunda edição dada a importância da temática e interesse dos alunos do Curso de Biblioteconomia da UFRJ.

A utilização dos objetos produzidos com objetivos educacionais específicos quanto daqueles que, apesar de terem sido inicialmente criados para divulgações jornalísticas, documentárias ou culturais, apresentam elementos que podem enriquecer o processo de aprendizagem em todos os segmentos educacionais.

Ao oferecer uma ampla gama de materiais, desde vídeos educacionais até documentários e conteúdo cultural, o sistema proporciona aos educadores e alunos uma variedade de recursos para enriquecer o ambiente de aprendizagem. Essa diversidade estimula a criatividade e permite que os educadores explorem diferentes abordagens pedagógicas, adaptando o conteúdo de acordo com as necessidades e interesses dos alunos.

Em avaliação da equipe que é composta pelos estudantes, coordenadora do projeto e coordenadora do repositório, além de uma bibliotecária da UFRJ e mais dois professores do CBG, o projeto deve evoluir com mais interação dos extensionistas com os usuários, mais visitações ao espaço do INES para aprofundar a compreensão das necessidades informacionais diretamente com o público e a inserção de mais materiais externos a produções do INES, contudo que ainda englobem as prioridades dos estudos da comunidade surda.

Finalmente, considera-se que o projeto de extensão em parceria do Curso de Biblioteconomia da UFRJ com o INES está de acordo com as Diretrizes de Extensão Universitária propostas pela Pró-Reitoria de Extensão da UFRJ, tais como: a Interação dialógica - com ênfase na interação de conhecimentos e experiências entre academia e sociedade e na articulação com organizações de outros setores da sociedade; a Interdisciplinaridade e a Interprofissionalidade - unindo saberes da Biblioteconomia, da Pedagogia e da Tecnologia da Informação; a Indissociabilidade do tripé ensino, pesquisa e extensão – com o impacto do projeto na formação dos estudantes envolvidos, e, por fim, o Impacto na transformação social - verificada potencial contribuição para a formulação, implantação e acompanhamento de políticas públicas prioritárias para a comunidade surda brasileira.

Bibliografia

Brasil. Lei nº 10.436 de 24 de abril de 2002. Dispõe sobre a Língua Brasileira de Sinais - Libras e dá outras providências. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/2002/l10436.htm

Brasil. Decreto nº 5.626 de 22 de dezembro de 2005. Regulamenta a Lei nº 10.436, de 24 de abril de 2002, que dispõe sobre a Língua Brasileira de Sinais - Libras, e o art. 18 da Lei nº 10.098, de 19 de dezembro de 2000. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/decreto/d5626.htm

Campello, A. R. S. (2008). Aspectos da visualidade na educação de surdos. 228 f. Tese (Doutorado em Educação) – Universidade Federal de Santa Catarina, Florianópolis. Disponível em: http://repositorio.ines.gov.br/ilustra/handle/123456789/277

Chalhub, T. & Dionysio, R. (2023). O protagonismo de uma linguagem visual: a construção de um repositório para educação de surdos. In: Miranda, A.C.D.; Oliveira, A.; Queiroz, C.F. de & Araujo, L. D. de. (2023). *Repositórios:* visão e experiência. Rio de Janeiro: Fiocruz; Rio Grande, RS: Ed. da FURG, 2023.

Gesser, A. (2009). Do patológico ao cultural na surdez: para além de um e de outro ou para uma reflexão crítica dos paradigmas. In Quadros, R. M. de & STUMPF, M. R. *Estudos surdos IV*. Petrópolis, RJ: Arara Azul, 2009. Disponível em: http://repositorio.ines.gov.br/ilustra/handle/123456789/672

Márdero Arellano, M. & Leite, F. (2009). Acesso aberto à informação científica e o problema da preservação. *Biblio*, (35), 2009, jan./mar. Disponível em: http://www.realp.unb.br/jspui/handle/10482/4937?locale=en

Perlin, G. & Strobel, K. (2014). História cultural dos surdos: desafio contemporâneo. *Educar em Revista*, Curitiba, Edição Especial (2): p. 17-31. Disponível em: https://www.scielo.br/j/er/a/qR5cDC7tgf5SyMtrSGvSVFC/?format=pdf

Quadros, R. (2008). Estudos Surdos III. Petrópolis, RJ: Arara Azul, 2008. Disponível em: http://repositorio.ines.gov.br/ilustra/handle/123456789/1048

Rosseto, M. (2003). Metadados e recuperação da informação: padrões para bibliotecas digitais. In II Cibernética: Simpósio Internacional de Propriedade Intelectual, Informação e Ética. [*Anais*]. Disponível em: https://www.acbsc.org.br/cursos/ciberetica

Santos, M.J.V.C. *Repositório Bilíngue para Educação de Surdos: mecanismo de inclusão*. Projeto de Extensão Universitário apresentado à Pró-Reitoria de Extensão da UFRJ. Rio de Janeiro, 2022.

Silveira, L. C. & Campelo, A. R. S. (2015). Materiais didáticos em Libras como facilitadores do processo inclusivo. Espaço, Rio de Janeiro, (43), 220-238, jan./jun. Disponível em: http://repositorio.ines.gov.br/ilustra/handle/123456789/905

Strobel, K. (2008). *As imagens do outro sobre a cultura surda*. Florianópolis: Ed. da UFSC.

Stumpf, M. R. & Wanderley, D. C. (2016). Quem fala português, escreve em português. Quem fala inglês, escreve em inglês. Os surdos: escrevem em que língua? *Revista Letras Raras*, 5 (1).

Tania Chalhub

Professora Adjunta do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal do Pará (UFPA), atuando no Mestrado e doutorado, e do Instituto Nacional de Educação de Surdos (INES) atuando na graduação. Possui graduação em Serviço Social pela Universidade Federal Fluminense (1982), mestrado em Serviço Social pela Pontifícia Universidade Católica do Rio de Janeiro (1988) e doutorado em Social Work - University of Minnesota, EUA (1995). Realizou pesquisa de pós-doutoramento no IBICT (2012) com o tema acesso aberto via repositórios. É líder do Grupo de Pesquisa Acessibilidade, Interculturalidade e Educação de Surdos. Atualmente está como Coordenadora do Repositório Digital Huet que tem por objetivo ampliar a acessibilidade a objetos educacionais para surdos. Tem experiência na área de Educação, com ênfase em tecnologias na educação de surdos e acessibilidade comunicacional bem como na Ciência da Informação com disciplinas da pós-graduação com tema acessibilidade informacional e está desenvolvendo projetos de pesquisas intitulados *Produção do conhecimento de pesquisadores surdos no Brasil: comunicação científica de uma minoria linguística* e *Acessibilidade Informacional: Contribuição das Instituições de Ensino e Pesquisa na Região Norte*. Tem interesse nos seguintes temas: comunicação científica, tecnologias e acessibilidade informacional, acesso livre à informação científica em repositórios e acessibilidade. ORCID: 0000-0001-7160-3886

Maria José Veloso da Costa Santos

Professora Adjunta do Departamento de Biblioteconomia, da Faculdade de Administração e Ciências Contábeis (FACC) da Universidade Federal do Rio de Janeiro (UFRJ), onde atua no Curso de Biblioteconomia e Gestão de Unidades de Informação (CBG), nas áreas de pesquisa Organização da Informação e do Conhecimento, Comunicação Científica e Bibliometria. Integra, como segundo líder, a partir de 2014, o Grupo de Pesquisa (Diretório de Pesquisa do CNPq) "Bibliometria e Cientometria, como abordagem teórico-metodológica para a Organização do Conhecimento". Integra o Grupo de Pesquisa: Organização do Conhecimento e Análise (Crítica) do Discurso: dimensões dialógicas, epistemológicas e empíricas. Doutora em História das Ciências pelo Programa de História das Ciências e das Técnicas e Epistemologia (HCTE)/UFRJ. Professora do

Curso de Especialização em Políticas Públicas e Organização do Conhecimento (Convênio Arquivo Nacional/UFRJ), 2012 a 2013. Consultora do Sistema de Bibliotecas e Informação/UFRJ desde 1994 e da Seção de Memória e Arquivo do Museu Nacional da UFRJ desde 2008. Graduada em Biblioteconomia (Universidade Federal do Pará - 1970). Especialista em Documentação Científica pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) - em Convênio com a UFRJ - 1972, Especialista em Redes e Sistemas de Informação pela Universidade de São Paulo (USP) - 1973. Mestre em Ciência da Informação pelo Instituto Brasileiro de Informação em Ciência e Tecnologia em convênio com a Escola de Comunicação (ECO) da UFRJ - 1992. Bibliotecária/Documentalista aposentada em julho de 2008, atuando no Museu Nacional/UFRJ de 1975-2008, exercendo o cargo de chefe da Biblioteca (1989-1994) e chefe da Seção de Memória e Arquivo (1999-2007). ORCID: 0000-0003-0473-5680

Formação Profissional em Repositórios Digitais: um curso criado para melhorar a gestão dos profissionais de Informação

Claudete Fernandes de Queiroz[1], *Leonardo Simonini Ferreira*[2]

Palabras claves

Formação Profissional. Repositórios Digitais. Profissional de Informação.

Keywords

Professional qualification. Digital Repositories. Information Professional.

Eje temático

Comunicación académica, científica y cultural en abierto

Resumen

O texto discorre sobre a criação do curso "Formação Profissional em Repositórios Digitais", que será oferecido pela Fiocruz na modalidade Qualificação para os profissionais de Informação que trabalham com repositorios e participam das Redes de Repositórios Digitais distribuídas no Brasil. O curso foi configurado pela equipe executiva do Repositório Arca e da Secretaria Acadêmica da Fiocruz, que desenvolveram planos de ensino para a montagem de uma disciplina que abrangesse os módulos necessários para cada assunto. O objetivo final do curso será o de oferecer aos membros da Redes um curso criado para melhorar a gestão dos repositorios e assim aprimorar as competências e habilidades ocupacionais promovendo eficiência, eficácia e qualidade na inserção de documentos nos Repositórios.

1 Fundação Oswaldo Cruz. Instituto de Comunicação e Informação Científica e Tecnológica em Saúde. Rio de Janeiro, RJ, Brasil, claudete.queiroz@fiocruz.br

2 Fundação Oswaldo Cruz. Instituto de Comunicação e Informação Científica e Tecnológica em Saúde. Rio de Janeiro, RJ, Brasil, leonardo.simonini@fiocruz.br

A Rede Sudeste de Repositórios Digitais3 criada em 2017 é coordenada pelo Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT) da Fiocruz4, e conta com a participação de 94 instituições de Ensino e Pesquisa, que reúnem esforços em prol da gestão e da visibilidade da produção científica incluída em seus repositórios e tem como objetivo principal promover o compartilhamento de informações através da realização de reuniões, cursos e eventos. A Rede também entende a importância da atuação dos profissionais de informação como gestores desse processo, porque tem uma visão ampla da organização em que estão inseridos, integrados com a missão e os valores institucionais.

O profissional Bibliotecário que atua com Repositórios precisa estar alinhado com outras particularidades, como por exemplo, sua atuação como gestor do sistema, nas comunidades e coleções dentro de um contexto informacional bem abrangente, visando atender a Instituição e os pesquisadores. É fundamental que esse profissional tenha habilidades e competências para atuar em diversos processos, além de ter conhecimentos sobre diferentes tecnologias digitais. A gestão das informações produzidas e inseridas num repositório, são fundamentais porque potencializam e identificam o conteúdo produzido para divulgação num cenário social e organizacional, compreendendo que essa estrutura é realizada a partir de um planejamento que envolva toda equipe de forma a garantir a qualidade, confiabilidade e autenticidade dos registros (Amante, 2014).

Com o objetivo de contribuir para o aprimoramento dos profissionais de informação que trabalham com repositórios digitais, a Rede Sudeste tem desenvolvido desde 2017, diversos projetos que permitem que os profissionais de informação de cada instituição participante possam relatar e trocar experiências, além do compartilhamento de informações relevantes sobre a temática (Sayão, Sales, 2019).

Um dos projetos que merece destaque foi desenvolvido pelo Subgrupo Montagem de Cursos5, coordenado pelas equipes que trabalham com Repositórios Digitais da Marinha do Brasil e da Fiocruz com a colaboração dos membros das instituições participantes da Rede Sudeste, e que discutiu ações necessárias para a capacitação dos profissionais que integram a Rede. Esse projeto teve início em 2019 com a aplicação de um questionário que buscou identificar quais temas os cursos deveriam atender, e após o retorno das respostas, iniciou-se o processo de análise dos dados que indicaram as temáticas que seriam contempladas. Assim, foi iniciada a elaboração da programação preliminar dos cursos, as disciplinas, as ementas, os planos de ensino, o cronograma, os professores e a ferramenta de transmissão, já que estávamos no período da pandemia e todas as atividades estavam sendo realizadas no formato remoto. Outro ponto importante durante a elaboração dos planos foi a necessidade de informar o nível de prioridades para os cursos, determinado no próprio questionário. O resultado deste trabalho foi que a Rede conseguiu realizar entre 2020 e 2022, 26 cursos com um total de 2.591 participantes (Figura 1) (Valentim, 2012).

3 Passou a se denominar Rede Sudeste de Repositórios Digitais a partir de 2022. A Rede Sudeste integra a Rede Brasileira de Repositórios Digitais, coordenada pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). Disponível em: http://rbrd.ibict.br/rede-sudeste-de-repositorios-digitais/

4 Disponível em: https://portal.fiocruz.br/

5 Os subgrupos são formados pelos membros da Rede Sudeste.

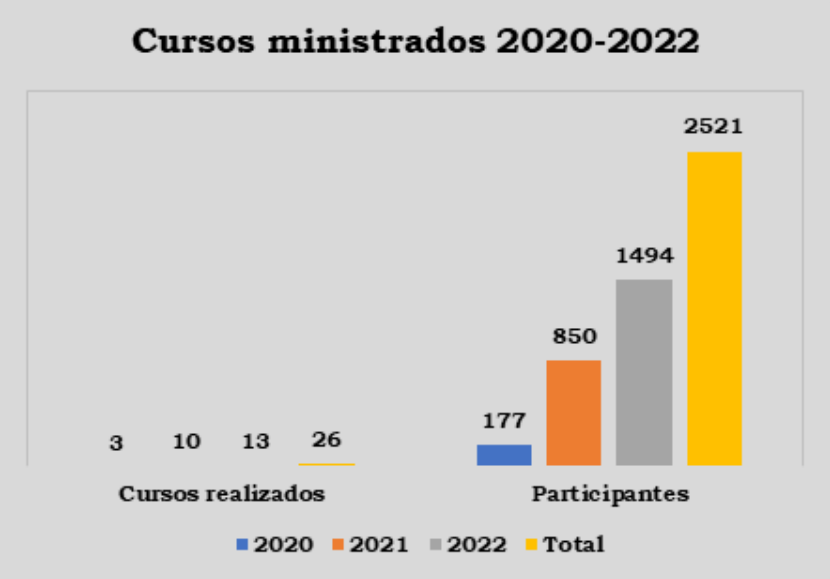


Figura 1: Cursos ministrados pela Rede Sudeste entre 2020-2022

Dentro deste contexto, a equipe executiva do Repositório Arca6, em 2023, entendeu a necessidade a partir dos resultados obtidos, de desenvolver um curso que atendesse as demandas, mas que tivesse uma padronização nos moldes dos cursos oferecidos pela Secretaria acadêmica da Fiocruz7 e que fosse incorporado em um programa regulamentado na área de Educação. A partir dessa constatação, foi elaborado um plano de ensino para a montagem de módulos que integrassem um único curso para apresentação a Secretaria Acadêmica, para ser avaliado visando aprovação dos critérios e requisitos solicitados. Após análise da coordenação de Educação, obtivemos a autorização para implementação do curso intitulado "Formação Profissional em Repositórios Digitais", que se enquadrou na modalidade "Qualificação" e que faz parte da concepção de construção coletiva do conhecimento, em que os membros da Rede Sudeste possam desenvolver saberes, práticas e diálogos através de produções científicas na área da Ciência da Informação e de Tecnologia da Informação (Le Coadic, 2004).

Vale ressaltar que a formação de uma Rede envolve várias ações, mas principalmente o aprimoramento profissional dos seus membros, e para corroborar essa premissa, destacamos Queiroz e Araujo (2020) que afirmam que

"[...] o profissional Bibliotecário que atua com Repositórios precisa estar alinhado com outras particularidades, como por exemplo, sua atuação como gestor do sistema, nas comunidades e coleções dentro de um contexto informacional bem abrangente, visando atender a Instituição e seus pesquisadores. [...] É fundamental que o profissional que gerencia e trabalhe com Repositórios tenha habilidades e competências para atuar em diversos processos. As competências envolvem conhecimentos, aptidões e qualidades decorrentes das experiências acumuladas, como também capacidades nas ferramentas computacionais e trabalho em rede".

6 Disponível em: www.arca.fiocruz.br
7 https://www.icict.fiocruz.br/qualificacao

O objetivo principal do curso será apresentar aos alunos conceitos, métodos, recursos informacionais e ferramentas relacionadas aos Repositórios e sua importância no contexto institucional e informacional, e será oferecido de forma remota, com carga horária total de 75h e disponibilidade de 200 vagas a partir do segundo semestre de 2024, com a emissão de certificados. O perfil dos alunos levantados foram: pesquisadores, estudantes e profissionais que atuam em atividades ligadas aos ambientes digitais, bem como aqueles que atuam no desenvolvimento e apoio à pesquisa em Ciência, Tecnologia e Inovação. A inscrição e divulgação dos cursos será realizada pela própria Fiocruz.

Outro benefício que podemos destacar é que o curso poderá ser oferecido não somente para a Rede Sudeste, mas também para as demais regiões do Brasil, alcançando assim, as Redes de Repositórios Digitais do Norte, Nordeste, Centro-Oeste e Sul (Toutain, 2007).

A estrutura curricular do Curso será oferecida da seguinte forma:

Acesso Aberto e Repositórios Digitais

Proposta: Identificar, gerenciar, monitorar e preservar todos os tipos de ativos digitais para garantir sua disponibilidade fidedigna a longo prazo para o seu público-alvo e promover a adoção de padrões e melhores práticas para curadoria digital no patrimônio da instituição, por meio da adoção de políticas próprias para cada fase da curadoria.

DSpace

Proposta: Conhecer de forma básica as principais funções do software livre DSpace, a nível de usuário e administrador de repositório DSpace. Entender o funcionamento básico do DSpace para que consiga aplicar e adaptar às realidades de cada instituição.

Preservação Digital

Proposta: Compreender a importância da preservação digital no contexto institucional e informacional. Conhecer os processos básicos de preservação digital e documentação dos fluxos de trabalho relacionados. Identificar os responsáveis, as funções e os recursos necessários para a implementação de ações de preservação digital.

Vocabulários Controlados

Proposta: Apresentar o que é um vocabulário controlado e a linguagem utilizada no processo de indexação. Abordar as principais Teorias que envolvem o processo de indexação. Apresentar o vocabulário controlado e sua metodologia de construção.

Certificação em Repositórios

Proposta: Entender o que é uma Certificação e como pode ser aplicada em Repositórios. Identificar o que são Repositórios Confiáveis. Aprender os critérios e procedimentos determinados para Certificação e qualificar o Repositório da Instituição.

Direito Autoral em Repositórios

Proposta: Compreender o conceito de Direito autoral e relacionar às iniciativas de acesso aberto. Elucidar dúvidas em relação aos procedimentos para inclusão de documentos em repositórios digitais garantindo a preservação dos direitos autorais.

Profissional de Informação e os Repositórios

Proposta: Compreender a importância da atuação do profissional bibliotecário junto aos Repositórios. Conhecer a legislação que regulamenta a profissão. Compreender os processos e fluxos de trabalho relacionados. Identificar os responsáveis, as funções e os recursos necessários para o desenvolvimento das atividades do profissional Bibliotecário. Conhecer as competências informacionais utilizadas nos Repositórios.

Métricas em Repositórios

Proposta: Definir o que são métricas para repositórios e sua utilização. Conhecer as principais métricas de repositórios. Identificar os principais mecanismos e ferramentas disponíveis. Apresentar os resultados obtidos.

Processo de Indexação para Repositórios

Proposta: Apresentar o conceito de Indexação e abordar a integração da Indexação com os repositórios. Discorrer sobre a análise temática da indexação e presentar um modelo de política de indexação para repositórios.

Princípios FAIR aplicados a Repositórios

Proposta: Discorrer sobre o movimento GO-FAIR. Apresentar os princípios FAIR e a iniciativa Global Open FAIR. Fundamentar os princípios FAIR em aplicações e serviços de gestão de dados de pesquisa. Explicar a importância dos princípios FAIR para os repositórios.

Inteligência Artificial Documental

Proposta: Dar uma visão geral da área, em particular no contexto que envolva o processamento de linguagem natural (Natural Language Processing – NLP). Apresentar algumas dessas ferramentas/resultados e discutir potencial uso em algumas situações na área documental e discutir potencial aplicações de nossos próprios resultados nos problemas detectados nos Repositórios.

Zotero: Gestão bibliográfica para uso em pesquisa

Proposta: Qualificar os profissionais de informação no uso do software livre Zotero, um gerenciador de referências bibliográficas que visa facilitar a elaboração de trabalhos acadêmicos e científicos, como teses, dissertações, trabalhos de conclusão de curso e artigos científicos.

Noções de Acessibilidade

Proposta: Apresentar os principais conceitos sobre Acessibilidade e sua importância para os profissionais de informação.

Recursos Educacionais Abertos e Bibliotecários

Proposta: Discorrer sobre o que são os Recursos Educacionais Abertos, seus benefícios para a educação, bem como estimular a prática da utilização, criação e compartilhamento desses recursos. Reutilizar, revisar, reter, remixar e redistribuir são os cinco pilares de liberdade de uso dos REA, o que permite também mais independência de alunos e docentes no processo de ensino.

Considerações finais:

o Bibliotecário que atua, gerencia e coordena os Repositórios são responsáveis pelas informações depositadas, bem como pela curadoria dos dados que ficarão disponibilizados. Esse gestor deve estabelecer processos, critérios e estratégias para sistematizar e organizar as informações para viabilizar o acesso e disseminar o conteúdo informacional.

É importante ressaltar, que os repositórios digitais possuem uma característica fundamental que é a de reunir, organizar, disseminar e preservar em um único local a produção intelectual de uma instituição, mas que precisam de profissionais qualificados para a gestão desses sistemas.

A Rede Sudeste nos últimos anos, tem se tornado fundamental para fortalecer a importância da gestão realizada de forma colaborativa, apresentando diversos resultados positivos na condução das atividades de um repositório.

O curso será importante também para aprimorar os conhecimentos sobre Ciência Aberta, Acesso Aberto, dentre outros tópicos, mas corroborar e entender que os Repositórios, como via verde para o depósito da produção intelectual, preservam a memória institucional e promovem a disseminação do conhecimento científico para toda sociedade

Bibliografía

Referencias bibliográficas de los trabajos citados en la ponencia. Usar norma APA 6ta edición.

Amante, M. J. (2014). O bibliotecário como gestor do conhecimento: o caso dos repositórios. ***RECIIS - Revista Eletrônica de Comunicação, Informação e Inovação em Saúde****, Rio de Janeiro, v. 8, n. 2, p. 243-254, jun. 2014. Disponível em: https://www.arca.fiocruz.br/handle/icict/17100.*

Le Coadic, Y.-F. (2004). ***A Ciência da Informação****. Brasília, DF: Briquet de Lemos. 205 p.*

Queiroz, C. F. de; Araujo, L. D. de. (2020). ***Bibliotecário de Repositórios****. In: SILVA, F. C. C. da (org.). O perfil das novas competências na atuação bibliotecária. Florianópolis, SC, Rocha Gráfica e Editora. p. 133-163. Disponível em: https://www.arca.fiocruz.br/handle/icict/45558*

Sayão, L. F., Sales, L. F. (2019). Curadoria de dados de pesquisa em repositórios. In: ENCONTRO DA REDE SUDESTE DE REPOSITÓRIOS INSTITUCIONAIS, 1., 2019, Rio de Janeiro. ***Anais...*** *Rio de Janeiro: Fiocruz/Icict/UFRJ. 80 p.*

Toutain, L. M. B. B. (2007). ***Para entender a Ciência da Informação****. Salvador: EDUFBA. 242 p. Disponível em: https://repositorio.ufba.br/bitstream/ufba/145/1/Para%20entender%20a%20ciencia%20da%20informacao.pdf*

Valentim, M. L. P. (2012). Estrutura de bases de dados: modelos de metadados e a qualidade de resposta. ***Transinformação****, Campinas, v. 13, n. 1, p. 67–80. Disponível em: https://www.scielo.br/j/tinf/a/hncTtMcY-8dBKNbjsG5NhtWs/?lang=pt*

Claudete Fernandes de Queiroz – Fundação Oswaldo Cruz

Doutoranda em História, Política e Bens Culturais pela Fundação Getúlio Vargas-RJ. Mestre em História, Política e Bens Culturais pela Fundação Getúlio Vargas-RJ. Possui especialização em Docência Superior pelo ISEP e graduação em Biblioteconomia pela Universidade Santa Úrsula. Atuou como Bibliotecária nas seguintes instituições: SENAC/Departamento Nacional; SENAI/RJ/Centro de Tecnologia Euvaldo Lodi; Documentar; Conselho Federal de Enfermagem; Ministério da Defesa/Centro Tecnológico do Exército; e atualmente exerce o cargo de Tecnologista em Saúde Pública na Fiocruz, atuando na coordenação técnica do Repositório Institucional Arca, Rede Sudeste de Repositórios, Biblioteca Virtual em Saúde e outros projetos.

Endereço para acessar este CV: http://lattes.cnpq.br/5902232749593657

Leonardo Simonini Ferreira – Fundação Oswaldo Cruz

Mestre em Biblioteconomia pela Universidade Federal do Estado do Rio de Janeiro - UNIRIO (2017). Possui graduação em Sistemas de Informação pela Universidade Estácio de Sá (2007). Tem experiência na área de Ciência da Informação e Saúde. É técnico em saúde pública no Instituto de Comunicação Científica e Tecnológica em Saúde (ICICT) na Fiocruz. Coordenador do "Curso de Acesso à Informação Científica e Tecnológica em Saúde" desde 2022, e docente deste mesmo curso desde 2008. Docente em cursos de Pós-Graduação sobre Gestão de Referências Bibliográficas ZOTERO, para pesquisadores e discentes.

Endereço para acessar este CV: http://lattes.cnpq.br/3674190956855127

Compartir para generar nuevo conocimiento: construcción de una propuesta para el fortalecimiento de las prácticas en ciencia abierta para los grupos de Investigación de la Facultad de Odontología, Universidad de Antioquia

Ana Isabel Correa-Orrego[1]

Palabras claves

Ciencia abierta, comunicación científica, democratización del conocimiento, prácticas abiertas

Open science, science communication, democratization of knowledge, open practices

Eje temático

Comunicación académica, científica y cultural en abierto

Resumen

Introducción: las prácticas en ciencia abierta representan alternativas para la generación de conocimiento, transformando la producción, difusión, evaluación y comunicación científica.

Objetivo: diseñar un plan de fortalecimiento de prácticas en ciencia abierta para los grupos de investigación de la Facultad de Odontología de la Universidad de Antioquia, Colombia, quienes buscan mejorar la salud bucal de las comunidades compartiendo conocimiento, principalmente, a través de publicaciones en revistas de acceso abierto, con poca participación de otros componentes de ciencia abierta. En Colombia se están desarrollando y estableciendo políticas para promover la apertura del conocimiento, lo cual requiere de espacios de sensibilización, capacitación y formación para la comunidad odontológica en relación con prácticas abiertas.

Metodología: se realizó una exploración documental para conocer las prácticas en ciencia abierta en odontología a nivel global, nacional y local, y las implicaciones de la Política Nacional de Ciencia Abierta para los grupos de investigación y sus prácticas de generación de conocimiento.

Resultados: se encontró que las prácticas de ciencia abierta en odontología se desarrollan alrededor del acceso abierto, datos abiertos, medición científica, innovación y apropiación social del conocimiento. Se identificaron 9 metas de la política nacional de ciencia abierta en las que los grupos deben direccionar su quehacer científico y académico.

Introducción

Con el objetivo de diseñar una propuesta para el fortalecimiento de las prácticas en ciencia abierta, o prácticas abiertas, en los grupos de investigación de la Facultad de Odontología de la Universidad de Antioquia, Colombia, y establecer acciones que permitan que la generación de conocimiento sea más accesible, eficiente, transparente y genere impactos a nivel científico y social, se decide emprender este proyecto.

1 Editora técnica, Revista Facultad de Odontología Universidad de Antioquia, *aico1095@gmail.com - aisabel.correa@udea.edu.co*

En el campo de la odontología, la información hace referencia a los datos, hechos y evidencias que se generan a partir de la docencia, la investigación y la relación con las comunidades, siendo crucial para el avance y desarrollo de la disciplina, apoyo a la toma de decisiones y al impacto social. En este contexto, la información se categoriza como una "cosa" que aporta conocimiento acerca de algo (Hjorland, 2021), necesario para la comprensión y problematización de fenómenos odontológicos; también se le incluye dentro de la categoría información-como-conocimiento que busca comunicarlo para que otros aprendan.

Los grupos de investigación de la Facultad, conformados por docentes y estudiantes, direccionan sus planes estratégicos trabajando bajo diferentes líneas y temáticas, con el fin de generar productos de conocimiento que contribuyan al posicionamiento de la institución (Gómez Velásquez, Correa Orrego y Toro Alzate, 2021) y al desarrollo de la investigación en Odontología. La Facultad se debe a la sociedad al pertenecer a una universidad pública, es por esto por lo que los grupos de investigación crean proyectos con miras al mejoramiento de la salud bucal y colectiva, así como la generación de productos de conocimiento de alto impacto, no solo científico sino académico y social. Este conocimiento es compartido a la comunidad académica y científica a través de diferentes formatos, siendo el artículo de revista el más recurrente; estos son publicados, en su mayoría, en revistas internacionales con un alto impacto científico, generando visibilidad a la Facultad y a los grupos de investigación. Los procesos de comunicación de la ciencia en la Facultad se han venido desarrollando a través de canales tradicionales y preestablecidos; sin embargo, en la última década han surgido nuevos componentes que se integran a la comunicación de la ciencia, transformando el modo en cómo se desarrollan los procesos de investigación (García Espinosa, 2019) ampliando el espectro en materia de comunicación de la ciencia para los grupos.

La ciencia abierta, como un elemento importante de la comunicación científica, es un movimiento que busca la apertura de la investigación científica, basado en el trabajo cooperativo y en las nuevas formas de diseminación del conocimiento (García Espinosa, 2019; Comisión Europea, 2018); en este sentido, en los grupos de investigación ha comenzado a despertarse el interés por adaptarse a nuevas formas de producir y compartir el conocimiento de forma abierta, a su vez que se establece la Política Nacional de Ciencia Abierta 2022-2031 del Ministerio de Ciencia, Tecnología e Innovación en Colombia, y se publica la recomendación de la UNESCO sobre ciencia abierta (2021), las cuales buscan promover la apertura y el acceso de la investigación científica, que son tenidas en cuenta a la hora de incluir requerimientos para que las instituciones y los investigadores compartan los datos de investigación, publiquen en revistas de acceso abierto y se fomente la colaboración internacional en proyectos de corte científico. Para las universidades y sus grupos de investigación, ambos documentos impactan de manera significativa la forma en cómo realizan sus actividades investigativas, además de indicar rutas emergentes en cuanto a comunicación científica se trata, esto se evidencia en la medición de grupos de investigación, en tanto que las convocatorias comenzarán a incluir criterios relacionados con la ciencia abierta y productos de apropiación social del conocimiento, siendo necesario que los grupos de investigación comiencen a comunicar sus productos de conocimiento de una forma más accesible, tanto para su público especializado como el general.

La relación de los grupos de investigación de la Facultad con la ciencia abierta se ha presentado de forma parcial. Se adaptan a convocatorias institucionales, tanto internas como externas, que financian los proyectos de investigación y establecen el tipo de producto a generar y bajo qué modelo debe disponerse que, en su mayoría, corresponden a artículos publicados en revistas indexadas de alto impacto internacional. Este tipo de publicación ha representado un interés económico para profesores-investigadores tanto de universidades públicas como de algunas privadas en Colombia, las cuales han desarrollado políticas donde se prioriza el componente de la investigación, ofreciendo incentivos por generar conocimiento y

publicarlo (Méndez Sayado y Vera Azaf, 2015), a su vez que los mayores incentivos en Colombia están dirigidos a la publicación de artículos en revistas de este tipo (Colombia. Ministerio de Ciencia, Tecnología e Innovación, 2022). Así mismo, se expide el Decreto 1279 de 2002[2], cuya finalidad era crear un sistema de estímulos económicos a la producción científica, donde a mayor número de publicaciones por año (en este caso artículos), mejor categoría de la revista y menos cantidad de autores, el salario se incrementa (Wilches Visbal et al., 2022), lo que ha aumentado significativamente la producción de conocimiento que, muchas veces, no cumple con estándares editoriales y su impacto no es significativo.

En la Facultad, el acercamiento que se ha tenido con los componentes de la ciencia abierta, y su formación, ha sido el acceso abierto donde predomina la publicación de artículos de este tipo, generalmente bajo la ruta dorada o de pago por APC (Article Processing Charges), así mismo, se han ofrecido asesorías y acompañamiento en los procesos de publicación en revistas bajo este modelo; además, en la Facultad de Odontología se presentan ciertas debilidades respecto a estas prácticas, evidenciando una preferencia por las formas tradicionales de comunicación científica, cuyo enfoque ha sido compartir conocimiento especializado a un segmento meramente científico. Las prácticas en ciencia abierta buscan ampliar y diversificar los usuarios y productores de conocimiento científico (Meneses-Placeres et al., 2022), siendo necesario comenzar a integrar componentes de la ciencia abierta como —además del acceso abierto—, datos abiertos de investigación, revisión por pares abierta, políticas, prácticas abiertas de investigación, investigación reproducible, software de código abierto, entre otros (FOSTER, 2019). También es importante comenzar a integrar estrategias y herramientas de divulgación científica, métricas responsables y la construcción de conocimiento con comunidades, logrando una apropiación social del conocimiento. Para que lo anterior se logre, es necesario que se construyan estrategias y se ejecuten acciones encaminadas hacia el fortalecimiento de estas prácticas, que contribuyan a la visibilidad de los productos de conocimiento que se generan a partir de la actividad investigativa de los grupos en la Facultad de Odontología.

Metodología

Este proyecto busca responder a la pregunta ¿Cuáles son las estrategias más adecuadas para fortalecer las prácticas en ciencia abierta de los grupos de investigación de la Facultad de Odontología de la Universidad de Antioquia? y el objetivo general bajo el que se pretende responder a dicha pregunta es Diseñar un plan estratégico que busque el fortalecimiento de las prácticas en ciencia abierta para que la generación de conocimiento de los grupos de investigación de la Facultad de Odontología sea más accesible, eficiente y transparente, que genere impacto científico y apropiación social del conocimiento.

El proyecto aún se encuentra en desarrollo como parte del proceso de Maestría en Ciencia de la Información, en la Escuela Interamericana de Bibliotecología de la Universidad de Antioquia. Para esto, se han establecido cinco objetivos específicos de los cuales, hasta la fecha, se han desarrollado los dos primeros, buscando dar cumplimiento al objetivo general, centrándose en:

- Identificar los antecedentes de prácticas en ciencia abierta desarrolladas en el área de la odontología a nivel global, nacional y local
- Analizar las implicaciones de la Política de Ciencia Abierta en Colombia para los grupos de investigación y sus prácticas de generación de conocimiento

2 Colombia. Ministerio de Educación. (2022). Decreto 1279 de Junio 19 de 2002. Disponible en https://www.mineducacion.gov.co/1621/articles-86434_Archivo_pdf.pdf

- Identificar la productividad, visibilidad e impacto de la producción científica en la Facultad de Odontología, 2011-2023
- Identificar las prácticas en ciencia abierta de los grupos de investigación de la Facultad de Odontología
- Establecer líneas estratégicas, objetivos, acciones, prácticas, productos y agentes, que permitan que el conocimiento, los métodos y datos estén en abierto.

La primera parte de este proyecto es de tipo descriptivo porque busca conocer las prácticas en ciencia abierta en el área de la odontología a nivel global, nacional y local, además de identificar las implicaciones de la Política Nacional de Ciencia Abierta respecto a las prácticas de generación y comunicación de conocimiento de los siete grupos de investigación adscritos a la Facultad de Odontología, clasificados y reconocidos por el Ministerio de Ciencia, Tecnología e Innovación. Esta primera parte se realizó a través de una exploración documental apoyada por una bitácora de búsqueda que permitió la elaboración de análisis. A partir de allí, se identifican componentes que se deberán tener en cuenta en el diseño del plan estratégico, que corresponde a la segunda parte que es de tipo proyectivo, haciendo referencia al diseño y planificación de acciones encaminadas hacia la transformación de una situación existente para alcanzar un fin (Hurtado de Barrera, 2000), buscando el fortalecimiento de las prácticas en ciencia abierta en los grupos de investigación.

Resultados

Prácticas abiertas en el área de la odontología: una mirada global, nacional y local

En este primer apartado se presentan los resultados de una revisión bibliográfica que permitiera identificar reflexiones que dentro del área de la odontología se dan en relación con las prácticas abiertas, desde una mirada global, nacional y local, además de indagar por las estrategias implementadas en el área para fortalecer estas prácticas, que sirvan como insumo para la construcción de la propuesta de formación, adaptándose a las necesidades particulares de los grupos de investigación de la Facultad de Odontología.

Se realizaron búsquedas a través de Google Scholar, Scopus, PubMed, LENS y DOAJ. Se tuvieron en cuenta fuentes documentales como artículos, memorias, trabajos de pregrado, tesis de posgrado, libros, proyectos, entre otras, cuyos metadatos se registraron en una bitácora de búsqueda que facilitó el análisis y permitió un acercamiento a las temáticas desarrolladas en las fuentes alrededor de la ciencia abierta. Fue necesario establecer los componentes de la ciencia abierta para delimitar las búsquedas, como: prácticas en ciencia abierta, ciencia abierta, acceso abierto, gestión de datos, métricas abiertas, innovación abierta y apropiación social del conocimiento. Se establecieron estrategias de búsqueda con operadores booleanos, para que la recuperación de las fuentes fuera más eficiente, acertada y pertinente.

Tras la búsqueda bibliográfica, se seleccionaron aquellos documentos relacionados con la odontología y sus diferentes áreas de especialidad, y que también desarrollaran algún componente de la ciencia abierta, además de excluirse aquellos documentos cuyo tema central no estuviera relacionado con el área de la odontología. A partir de esta selección inicial, se obtuvo un total de 250 documentos, sin embargo, al realizar un revisión de registros repetidos, se establece un total de 103 documentos, en los cuales se aplicaron criterios de inclusión y exclusión, como la antigüedad de las publicaciones (sólo aquellos publicados durante el 2019 y el 2024); idioma inglés y español, cobertura nacional e internacional y producción

de conocimiento donde se hiciera una reflexión, o análisis, de los componentes de la ciencia abierta y su influencia dentro del área de la odontología y sus diferentes áreas de especialidad. Como resultado de esto, sólo 20 documentos se consideraron pertinentes dado que cumplían con los criterios mencionados.

Antecedentes de prácticas abiertas en odontología

Como resultado de la revisión bibliográfica, se evidencian diferentes perspectivas desde donde se abordan las prácticas de ciencia abierta, partiendo desde la conciencia por parte de autores en relación con las transformaciones que estas han generado en los procesos de investigación, producción, comunicación y acceso al conocimiento (figura 1). Las prácticas en ciencia abierta comienzan a hacerse presente en procesos de creación de espacios de aprendizaje con el uso de las TIC, así como el desarrollo de herramientas encaminadas hacia la apertura de la ciencia como son los repositorios institucionales o de acceso abierto (Terreros et al., 2017), la gestión editorial de revistas académicas, espacios para la divulgación del conocimiento y la promoción de prácticas abiertas, todo esto para responder a dinámicas de corte internacional que motivan hacia una comunicación pública de la ciencia.

Figura 1 - Prácticas abiertas en odontología identificadas a partir de la revisión bibliográfica

Las reflexiones y discusiones alrededor de la gestión editorial de revistas se presentan con mayor frecuencia en las publicaciones, donde se enfatiza el papel que deben cumplir las revistas en relación con la promoción de prácticas transparentes y abiertas (Santos et al., 2024; Sofi-Mahmudi & Raittio, 2022); así mismo, en las revistas debe haber un mayor compromiso con estas prácticas, en especial con sus políticas editoriales encaminadas hacia la disponibilidad de los datos, el código abierto, declaraciones de ética claros y conflictos de interés definidos.

Se halla diversidad de estudios métricos respecto a temas como acceso abierto, y el uso constante de indicadores tradicionales para medir la productividad y el impacto científico (Martin et al., 2020), no obstante, se observa en los estudios de Elangovan & Allareddy (2019) y Isfandyari-Moghaddam et al. (2019), la implementación de métricas alternativas para medir visibilidad y citación a través de aplicaciones como *Altmetric Explorer, Web of Science Citations* y *Mendeley*.

En cuanto a la gestión de datos abiertos, las opiniones y experiencias de los autores son variadas. Uribe et al. (2022) y Vidal-Infer et al. (2019) concuerdan con la importancia del intercambio de datos, así como las estrategias para promover esta práctica en los procesos de producción de conocimiento, dado que promueve la creación de redes de colaboración que permiten lograr nuevos hallazgos y avances dentro del campo, sin embargo, Spallek et al. (2019) manifiesta su preocupación respecto al uso que se le dan a los datos una vez abiertos, como la protección de datos personales sensibles, lo que ha llevado a que muchos investigadores sean reticentes con la apertura de los datos. Por su parte, para Uribe et al. (2022) una de las preocupaciones se debe a la calidad de los datos que se comparten, dado que la mayoría se presentan en formatos que no son comprensibles por máquina, presentan una baja calidad e incumplen con los principios FAIR[3] (Findable, Accessible, Interoperable, Reusable).

En innovación abierta, Estrada y De la Cruz (2022) presentan una propuesta para implementar un modelo de innovación abierta en una empresa prestadora de servicios odontológicos, donde se hace un énfasis en la importancia de abrir el conocimiento que se genera a nivel interno, así como vincularse con los factores y agentes del entorno. Dentro de este se expone que el ambiente actual, en términos de negocio y transferencia de conocimiento, se ha alejado del modelo tradicional de innovación que estaba orientado hacia un modelo cerrado, dado que al abrirse facilita la generación de ideas dentro la organización que a su vez se conviertan en salidas que el mercado puede aprovechar, donde es importante establecer relaciones con instituciones especializadas, proveedores, clientes y agentes con los cuales generar colaboración que potencien la innovación durante todo el proceso, no solo de investigación, sino en los procesos de atención y tratamiento en pacientes. Así mismo, Estrada y De la Cruz (2022) mencionan que en el área hay una ausencia de estructuras organizadas o profesionales que se dediquen a obtener información sobre tendencias en el mercado, cambios, investigaciones en curso o desarrollos tecnológicos.

Para concluir, la implementación de prácticas abiertas en odontología todavía se encuentra en desarrollo, entendiéndose estas, finalmente, como metodologías o medios para la evaluación de la producción científica, la implementación de las TIC en procesos de enseñanza y aprendizaje y el uso de nuevas tecnologías a nivel empresarial y de atención a pacientes.

Implicaciones de la Política de Ciencia Abierta en Colombia para los grupos de investigación y sus prácticas de generación de conocimiento: el caso de la Facultad de Odontología

En el proceso de comunicación y acceso al conocimiento científico, la mercantilización es un aspecto que abre las brechas y cierra el conocimiento, provocando la creación de movimientos que luchan por su democratización, reclamándolo como un bien común (Bollier, 2016). A raíz de esta mercantilización, promovida por las grandes editoriales científicas, algunos entes gubernamentales e instituciones han construido y establecido políticas para impulsar el acceso libre del conocimiento, como es el caso de Colombia, que en el año 2022 presenta su Política Nacional de Ciencia Abierta, 2022-2031, que busca fomentar la apertura y el acceso al conocimiento científico, tecnológico y de innovación, promoviendo la colaboración, la transparencia y la democratización del conocimiento en beneficio de la sociedad, que influye en las prácticas de generación y comunicación científica en los investigadores a nivel nacional. Así pues, en este apartado se presentan las implicaciones de esta política nacional para los grupos de investigación de la Facultad de Odontología respecto a sus prácticas de generación y comunicación de conocimiento, así como los cambios que se vislumbran sobre estos procesos.

3 *FAIR principles* disponible en https://www.go-fair.org/fair-principles/

La Política Nacional de Ciencia Abierta y sus componentes

La Política Nacional de Ciencia Abierta en Colombia busca "aumentar el acceso, la visibilidad, la reproducibilidad y la utilidad de los datos, recursos, productos y resultados científicos, tecnológicos y de innovación colombianos, ampliando la formación, apropiación, institucionalización y las infraestructuras de Ciencia Abierta del país" (p. 47), a través de acciones encaminadas al fortalecimiento de la gobernanza de este modelo abierto, el impacto social de la producción de conocimiento científico, el fortalecimiento de prácticas abiertas para las instituciones y la inversión en infraestructura que refuerce la comunicación y acceso al conocimiento generado en el país, con el fin de consolidar una cultura que vea en la ciencia la posibilidad de realizar aportes significativos para el desarrollo de la sociedad (Uribe Tirado y Ochoa, 2022). La Política se estructura a partir de seis apartados:

1. Apartado introductorio.
2. Justificación de la política, cuyo fin es que esta se convierta en un instrumento para la inclusión y la democratización del conocimiento.
3. Presentación de antecedentes, donde se muestran los avances y experiencias a nivel mundial en materia de construcción de políticas públicas en ciencia abierta.
4. Marco conceptual y teórico que sirva como base epistemológica de la ciencia abierta.
5. Diagnóstico sobre la ciencia abierta en Colombia.
6. Finalmente, se definen los fundamentos estratégicos para que la Política pueda implementarse y aporte a la construcción de una sociedad basada en el conocimiento y en su democratización (Colombia. Ministerio de Ciencia, Tecnología e Innovación, 2022).

Componentes de la Política Nacional que impactan a los grupos de investigación de la Facultad de Odontología y sus prácticas de generación, acceso y comunicación científica

A partir de los cinco objetivos específicos presentados en la Política, se establecen acciones estratégicas donde se definen metas que buscan impulsar a la ciencia abierta. Conviene aclarar que, no todos los componentes, acciones y metas de la Política son competencia de los investigadores, algunos de estos componentes corresponden al Estado y a las instituciones particularmente, por lo que se presentarán a continuación aquellos que competen a los grupos de investigación de la Facultad de Odontología y por lo cuales deben comenzar a trabajar, reconociendo los retos que la ciencia abierta trae consigo para construir estrategias que les permita hacer frente a lo planteado en la Política nacional. Se presentan en forma de metas, siguiendo la estructura que se presenta en dicho documento (Figura 2).

Figura 2 - Metas identificadas a partir de la Política Nacional de Ciencia Abierta para los grupos de investigación de la Facultad de Odontología

Meta 1. A partir del 2024, la investigación financiada con recursos públicos en el Sistema Nacional de Ciencia, Tecnología e Innovación (SNCTI) en Colombia, así como las publicaciones científicas resultado del proceso de investigación, deben estar en abierto, teniendo en cuenta las particularidades de cada proceso y siempre y cuando sea posible.

Meta 2. Crear espacios de articulación, diálogo e interacción con los diferentes actores generadores de conocimiento, como investigadores de otras áreas del conocimiento, Universidades, empresas, asociaciones y Estado, a través de alianzas, redes colaborativas y en consorcio con organizaciones, tanto públicas como privadas y mixtas, con miras a aumentar la circulación y optimización del uso del conocimiento abierto.

Meta 3. En la Política se plantea el diseño y piloto de un modelo financiero que permita financiar e incentivar los procesos de ciencia abierta, así como aquellos proyectos financiados con recursos públicos que deberán integrar en su formulación el componente de financiación bajo este modelo. Se tiene planteado crear e implementar paquetes de incentivos, de carácter monetario y no monetario, para promover y fomentar la apertura de la ciencia, esto representa una oportunidad para los grupos de la Facultad, en tanto se convertiría en un motivante para adelantar procesos de investigación, desarrollo tecnológico e innovación orientado a lo abierto, que involucre a los diferentes actores generadores de conocimiento.

Meta 4. Crear e implementar estrategias de difusión y divulgación orientadas hacia la comunicación pública de la ciencia, que integre a los actores e instituciones generadoras de conocimiento y a la ciudadanía en sus territorios.

Meta 5. Crear e implementar estrategias de ciencia ciudadana, orientadas a la identificación y priorización de los retos y problemáticas que se evidencian en los territorios, creando espacios donde se integre, participe y sensibilice a las comunidades para que, en conjunto, puedan resolverse, generar, usar y evaluar el impacto social de los productos de conocimiento.

Meta 6. Como afirma el Observatorio de la Universidad Colombiana (2022), una de las problemáticas actuales se debe al insuficiente desarrollo en formación y vinculación de capital humano en Ciencia, Tecnología e Innovación (CTeI), por lo cual es fundamental diseñar y/o participar en estrategias de formación y capacitación permanentes, continuos y focalizados en los componentes y prácticas de la ciencia abierta,

dirigidos hacia la comunidad científica, académica, administrativos, gestores de ciencia abierta, personal de apoyo en investigación e innovación, entre otros actores, que garantice integrar las prácticas abiertas en los procesos de desarrollo científico y tecnológico.

Meta 7. Aquellas investigaciones financiadas con recursos públicos deben difundir en abierto los datos de investigación, teniendo en cuenta las condiciones particulares de cada investigación. En este punto, es deber del Ministerio de Ciencia generar las condiciones técnicas que garanticen la apertura y reutilización de los datos, construyendo directrices y lineamientos para su gestión y preservación, así como la inversión en la infraestructura necesaria para este fin.

Meta 8. En acuerdo con el Ministerio de Educación Nacional, se prevé que para el año 2025, las prácticas de ciencia abierta se comenzarán a contemplar y valorar como indicadores sustanciales en los procesos de autoevaluación, planes de mejoramiento y registros calificados.

Meta 9. Desde el año 2023, los componentes de la ciencia abierta se han integrado en los procesos de evaluación de la investigación, el reconocimiento y clasificación de investigadores y grupos y, próximamente, en los procesos de clasificación de revistas científicas. Esto representa un reto para los investigadores en tanto sus prácticas de generación de conocimiento y productos se deben comenzar a orientar hacia prácticas abiertas.

La ciencia abierta plantea cambios y retos en las formas de generación y comunicación del conocimiento, siendo necesaria la creación de una cultura científica en la que se integren actores dispuestos al cambio, al diálogo, la inclusión, la cooperación y la responsabilidad social, como generadores de conocimiento (Uribe Tirado y Ochoa, 2022; Colombia. Ministerio de Ciencia, Tecnología e Innovación, 2022); para esto es crucial la puesta en marcha de estrategias orientadas a la formación y capacitación con miras a transformar prácticas tradicionales de generación de conocimiento, diversificando e incorporando otros actores que, pocas veces, se ven representados en estos procesos, como el conocimiento ancestral, local, diferentes razas y culturas (Observatorio de la Universidad Colombiana, 2022).

Conclusiones

A nivel de Facultad, es importante generar reflexiones y establecer estrategias que permitan adaptarse a las nuevas prácticas de generación de conocimiento. En un corto plazo, una apertura hacia el cambio, hacia el conocimiento de los componentes de la ciencia abierta y sus retos; en el mediano, estrategias formativas, o planes de formación e investigación, que promuevan y fortalezcan las prácticas abiertas, a través de la colaboración entre la comunidad académica y la ciudadanía; y en un largo plazo, una postura formal, donde se integren los componentes de la ciencia abierta a los planes estratégicos y de trabajo, así como en el currículo, que se ajusten a las necesidades de generación y uso del conocimiento, así como afrontar los retos que a nivel internacional se plantean respecto a la producción, comunicación e impacto científico.

A nivel país, de acuerdo con Uribe Tirado y Ochoa (2022), se debe prever la inversión en recursos tecnológicos, así como en la formación y cultura de los distintos actores que interactúan con el ecosistema de la ciencia abierta, sin esto, el desarrollo de los componentes, y de la ciencia abierta en general, no se daría de la manera esperada. La comunidad generadora de conocimiento debe velar porque el estado y las instituciones dispongan de herramientas, infraestructura e incentivos que los motive a integrar estas

prácticas en su quehacer científico y académico, al mismo tiempo que ellos mismos ejecuten acciones con miras a obtener financiación externa, crear servicios y construir redes de colaboración que permitan integrar distintos actores y latitudes, para que las prácticas de ciencia abierta sean sostenibles en el tiempo.

Colombia continúa trabajando por una ciencia más justa y más equitativa, y avanza hacia una apertura del conocimiento científico como instrumento para su democratización, asumiéndolo como un bien común y público. A partir de la puesta en marcha de esta política, se observa un largo camino y una lista de retos que, como país, se deben asumir, con miras a que el conocimiento trascienda los espacios académicos y llegue a los territorios, buscando una apropiación del conocimiento e impacto social (Colombia. Ministerio de Ciencia, Tecnología e Innovación, 2022).

Bibliografía

Bollier, D (2016). *Pensar desde los comunes: una breve introducción*. España, traficantes de sueños.

Colombia. Política Nacional de Ciencia Abierta 2022-2031. Ministerio de Ciencia, Tecnología e Innovación. Resolución 0777 de 2022.

Comisión Europea. (2018). *Commission Re-commendation of 25 April 2018 on access to and preservation of scientific infor-mation.* Brussels: European Commision. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018H0790

Elangovan, S. & Allareddy, V. (2019). Publication metrics of dental journals: what is the role of self citations in determining the impact factor of journals? *Journal of Evidence Based Dental Practice*, 15(3), 97-104. https://doi.org/10.1016/j.jebdp.2014.12.006

Estrada, M.M, y De la Cruz Torres, L.C. (2022). *Diseño de un modelo de gestión de la innovación para una Institución prestadora de servicios de odontología – SOMECA Ltda.* [Tesis de maestría, Universidad del Norte]. Repositorio Institucional de la Universidad del Norte. https://bit.ly/4cazBJ5

FOSTER. (2019). Manual de Capacitación sobre Ciencia Abierta. https://book.fosteropenscience.eu/es/

García Espinosa, E. (2019). Open science and scientific communication. *Revista Ciencias Médicas, 23(6)*.

Gómez Velásquez, S.N., Correa Orrego, A.I. y Toro Alzate, M. (2021). Trayectorias de la investigación 2011-2020: una mirada reflexiva y de retos. Facultad de Odontología, Universidad de Antioquia. https://bit.ly/3Swlkvk

Hjorland, B. (2021). Información. ISKO. https://www.isko.org/cyclo/information

Hurtado de Barrera, J. (2000). Metodología de la investigación holística. Caracas: Fundación Sypal. https://bit.ly/3v1L8Kb

Isfandyari-Moghaddam, A., Danech, F., Hadji-Azizi, N. (2019). Webometrics as a method for identifying the most accredited free electronic journals: The case of medical sciences. *The Electronic Library,* 33(1). https://doi.org/10.1108/EL-10-2012-0141

Martin, M.A., Lipani, E., Lorenzo, A.A., Aiuto, R., Garcovich, D. (2020).Trending topics in orthodontics research during the last three decades: A longitudinal bibliometric study on the top-cited articles. Orthodontics & Craniofacial Research, 23(4), 462-470. https://doi.org/10.1111/ocr.12396

Méndez Sayado, J.A., y Vera Azaf, L., (2015). Salarios, incentivos y producción intelectual docente en la universidad pública en Colombia. *Apuntes del Cenes, 34(60)*. http://www.scielo.org.co/scielo.php?pid=S0120-30532015000200004&script=sci_arttext

Meneses-Placeres, G., Álvarez Reinaldo, L.A., Machado Rivero, M.O. (2022). Revisión de las Prácticas de Ciencia Abierta en América Latina y el Caribe. *Revista Cubana de Transformación Digital, 3(1),* 1-8. http://portal.amelica.org/ameli/journal/389/3893118003/3893118003.pdf

Observatorio de la Universidad Colombiana. (2022). *Ciencia Abierta en Colombia: Definición, dificultades y oportunidades*. https://bit.ly/4cawmkN

Santos, W.V.O, Dotto, L., Ferreira, T.G.M. & Sarkis-Onofre, R. (2024). Endorsement of open science practices by dental journals: a meta-research study. *Journal of Dentistry.* https://doi-org.udea.lookproxy.com/10.1016/j.jdent.2024.104869

Sofi-Mahmudi, A. & Raittio, E. (2022). Transparency of COVID-19-Related Research in Dental Journals. *Frontiers in Oral Health, 3*, 1-6. https://doi.org/10.3389/froh.2022.871033

Terreros, M.A., Toala Reyes, A., y Salazar Arrata, J. (2017). Odontología Basada en la Evidencia, hacia el manejo de las buenas prácticas Odontológicas [Ponencia]. *1er Congreso Internacional de Investigación y Producción Científica en el Campo de la Estomatología,* Guayaquil, Ecuador. https://bit.ly/4c8Qnlu

UNESCO. (2021). *Recomendación de la UNESCO sobre la Ciencia Abierta*. Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. https://unesdoc.unesco.org/ark:/48223/pf0000379949_spa

Uribe Tirado, A., Ochoa, J. (2022). Perspectivas de la ciencia abierta: un estado de la cuestión para una política nacional en Colombia. *BiD: textos universitaris de biblioteconomia i documentació, (40)*. https://dx.doi.org/10.1344/BiD2018.40.5

Vidal-Infer, A., Tarazona, B., Alonso-Arroyo, A., & Aleixandre-Benavent, R. (2019). Public availability of research data in dentistry journals indexed in Journal Citation Reports. *Clinical oral investigations,* 22(1), 275–280. https://doi.org/10.1007/s00784-017-2108-0

Wilches Visbal, J.H., Castillo Pedraza, M.C., Pérez Anaya, O. (2022). El plagio y las revistas depredadoras: un problema económico y ético en universidades públicas de Colombia. *Revista Cubana de Información en Ciencias de la Salud, 33*. http://scielo.sld.cu/scielo.php?pid=S2307-21132022000100003&script=sci_arttext

Ana Isabel Correa-Orrego

Bibliotecóloga, Maestranda en Ciencia de la Información de la Universidad de Antioquia. Actualmente se desempeña como Editora Técnica en la Revista Facultad de Odontología Universidad de Antioquia. Entre sus intereses profesionales se destacan temas relacionados con la comunicación científica, ciencia abierta, estudios métricos, gestión de procesos editoriales y marketing digital.

La ruta de la Ciencia Abierta en Uruguay: políticas, infraestructuras y actores

Magela Cabrera Castiglioni[1], Carina Patrón[2], Mabel Seroubian[3]

Palabras claves:

Ciencia abierta; Políticas de ciencia abierta; Infraestructuras de ciencia abierta; Uruguay

Keywords:

Open science; Open science policies; Open science infrastructures; Uruguay

Eje temático

Comunicación académica, científica y cultural en abierto

Resumen

Se propone un estudio descriptivo sobre la situación de la ciencia abierta en Uruguay, explorando las políticas públicas, iniciativas académicas y de sociedad civil relacionadas con esta área. Se busca comprender el estado actual de la ciencia abierta en el país, examinando las distintas acciones llevadas adelante por diferentes actores. En primer lugar, se analizan las políticas públicas vigentes sobre ciencia abierta, destacando los principales enfoques y objetivos para promover la apertura y accesibilidad. Se describen los componentes más significativos de los instrumentos legales y normativos que respaldan estas políticas, así como los mecanismos de implementación. Además, se revisan las iniciativas desarrolladas por instituciones académicas y organizaciones de la sociedad civil en este ámbito.

En relación con las políticas públicas mencionadas, se presenta un análisis de las infraestructuras que se han establecido para respaldar la ciencia abierta en Uruguay. Esto incluye plataformas digitales, repositorios bibliográficos y de datos, y otros recursos destinados a facilitar el acceso a la información. Finalmente se obtiene una síntesis donde se destacan los avances, limitaciones, desafíos y oportunidades en este campo. Se espera que este estudio contribuya a fortalecer el ecosistema de la ciencia abierta en el país y fomente el debate sobre el tema.

Introducción

El avance de la ciencia abierta en América Latina es un hecho y cuenta con varios estudios (Alperin y Fischman, 2015; De Filippo y D'Onofrio, 2019; Babini y Rovelli, 2020) que sistematizan y analizan el tema. Allí se pueden ver iniciativas individuales pero también muchas acciones regionales que se encuentran interconectadas y que son las que en muchos casos han contribuido al avance en cada país. Tanto en América Latina como en Iberoamérica, se pueden observar una tendencia de las universidades e instituciones relacionadas a la ciencia, marcada por esfuerzos que buscan reducir las desigualdades existentes a nivel

1 Facultad de Información y Comunicación. Universidad de la República. magela.cabrera@fic.edu.uy
2 Facultad de Odontología. Universidad de la República carina@odon.edu.uy
3 Servicio Central de Informática. Universidad de la República. mabel.seroubian@seciu.edu.uy

internacional en la producción, circulación y acceso al conocimiento. Dichos esfuerzos se enmarcan en el paradigma de la ciencia abierta y buscan contrarrestar las tendencias hacia la comercialización de la ciencia (Babini & Rovelli, 2020).

Dentro de los componentes de la Ciencia Abierta podemos asegurar, basados en estudios precedentes que en América Latina, que el componente que más se ha desarrollado es el acceso abierto:

Los resultados obtenidos nos llevan a plantear que en Latinoamérica las iniciativas para el desarrollo y promoción de la ciencia abierta resultan recientes –se despliegan a lo largo de la última década– y que aún las experiencias más destacadas –que suceden en los países pioneros y más activos de la región en esta cuestión– se concentran fundamentalmente en la promoción del acceso abierto, sus infraestructuras y normativas (De Filippo y D'Onofrio, 2019, p. 45).

Otros temas como los datos abiertos y la ciencia ciudadana aparecen como aspectos en crecimiento aunque en menor representación en las políticas públicas (De Filippo y D'Onofrio, 2019).

También es de destacar la posición varios actores de la región sobre algunas cuestiones clave en el modelo de ciencia abierta, más específicamente en el acceso abierto como lo es la fuerte posición declarada contraria al modelo de APC en la conocida como Declaración de México (Reunión de Consorcios de Iberoamérica y el Caribe, 2017)

A nivel regional resulta relevante mencionar la Declaración de Panamá sobre Ciencia Abierta (2018) donde se hace hincapié en la necesidad de *"avanzar hacia modelos colaborativos de creación, gestión, comunicación, preservación y apropiación entre Academia-Ciudadanía-Estado-Empresa"* así como en la relación entre el desarrollo de *"políticas científicas abiertas como estrategia para mejorar la eficiencia y la productividad de la inversión en ciencia y tecnología"* (párr. 1; 3), siendo estos elementos clave para la construcción de ciudadanía y el fortalecimiento de la democracia.

Desde Uruguay se han seguido estos esfuerzos (Prieto, 2022, Aguirre-Ligüera et al, 2019) y se ha trabajado en forma colaborativa con varias iniciativas pero tal como plantea Chan et al (2019) los procesos de ciencia abierta deben tener un carácter situado y deben pensarse de acuerdo a cada contexto para así aprovechar el potencial de la Ciencia Abierta para contribuir al desarrollo inclusivo y sostenible en cada contexto. Es importante en relación a esto considerar el sistema de evaluación de los investigadores que se apoya en indicadores bibliométricos como único tipo de producción (De Giusti, 2022) y que influye en la elección de difusión de la investigación, en la forma como se almacenan los datos y la relación con investigadores de otras regiones (Prieto, 2022).

En este trabajo se plantea presentar y sistematizar las políticas y acciones desarrolladas en Uruguay, explorando las políticas públicas, iniciativas académicas y de sociedad civil. Una vez reunida y analizada la información generada en los últimos años en torno a políticas, acciones e infraestructuras será posible contar con un estado de situación actualizado y abarcativo de los diferentes componentes y actores sobre el tema.

Método

Se aplica una metodología cualitativa, a través de un estudio exploratorio y descriptivo. Para la recolección de los datos relacionados a las políticas se recurrió al relevamiento de documentos de diferente naturaleza como informes técnicos, artículos académicos, declaraciones y acuerdos de carácter gubernamental, normativa nacional e institucional y portales relacionados a la temática.

Para la realización del relevamiento de infraestructuras se tomó como delimitación del alcance la concepción de infraestructuras que se brinda en la Declaración de Ciencia Abierta de la Unesco de 2021 donde indican que las mismas:

se refieren a las infraestructuras de investigación compartidas (virtuales o físicas, en particular los grandes equipos científicos o conjuntos de instrumentos, los recursos basados en el conocimiento, como las colecciones, las revistas y las plataformas de publicación de acceso abierto, los depósitos, los archivos y los datos científicos, los sistemas de información de investigación actuales, los sistemas bibliométricos y cienciométricos abiertos para evaluar y analizar los ámbitos científicos, las infraestructuras informáticas y de manipulación de datos abiertas que permiten el análisis de datos colaborativo y multidisciplinario y las infraestructuras digitales) que son necesarias para apoyar la ciencia abierta y atender las necesidades de las diferentes comunidades. (Unesco, 2021, p.12).

A partir del relevamiento se realizó una sistematización de la información que se presenta en el siguiente apartado.

Estado de situación

En primer lugar, se analizan las políticas públicas vigentes sobre ciencia abierta, destacando los principales enfoques y objetivos para promover la apertura y accesibilidad. Se describen los componentes más significativos de los instrumentos legales y normativos que respaldan estas políticas, así como los mecanismos de implementación. En este caso la información se agrupa según las iniciativas son de carácter gubernamental, institucional público o de organizaciones de la sociedad civil.

Políticas públicas

A nivel de gobierno

Política de Datos Abiertos (2011). Por medio de la Agenda Digital Uruguay 2011-2015 se establecieron los primeros lineamientos para la Estrategia Nacional de Datos Abiertos. Lanzando también en este año el sitio: Datos.gub.uy, donde se comenzó a recopilar datos de diferentes fuentes públicas (en diferentes formatos, no necesariamente abiertos). Para finales del 2011 ya se contaba con un primer borrador de Plan de Acción Nacional de Datos Abiertos 2011-2015 que proponía tres líneas de acción; Infraestructura, Normativa y Apropiación y Fomento del uso de datos abiertos.

Planes de gobierno abierto. Desde el año 2012 hasta la fecha Uruguay ha elaborado cinco planes de Gobierno Abierto. Desde su inicio en 2012 se planteaba como objetivos: aumentar la integridad Pública; una gestión más eficiente de los recursos públicos y mejorar la prestación de los servicios públicos. Para su elaboración contó con la participación de: la Oficina de Planeamiento y Presupuesto (OPP), el Ministerio de Economía y Finanzas (MEF), la Unidad de Acceso a la Información Pública (UAIP), el Instituto Nacional de Es-

tadística (INE) y AGESIC, además de contar con la colaboración de CAINFO. Actualmente rige el Quinto Plan de Acción Nacional de Gobierno Abierto 2021-2024 el cual se propone "fortalecer la transparencia y rendición de cuentas en diversas temáticas, tales como: las compras públicas, la gestión y políticas de salud, los beneficios entregados al sector empresarial, el uso de la Inteligencia Artificial en el Estado y las políticas ambientales" (Uruguay. Presidencia de la República, 2021, p.3). Entre las líneas de acción propuestas se encuentra la implementación de la Estrategia nacional de Datos Abiertos 2021-2024. En este documento se contempla la participación activa de la academia pero no es mencionado el componente Ciencia abierta.

Sistema nacional de repositorios digitales abiertos en ciencia y tecnología.

En 2018 la Agencia Nacional de investigación e innovación convocó a las principales instituciones de investigación nacional para discutir la conveniencia y oportunidad de crear un nodo nacional de la Red de repositorios de acceso abierto a la ciencia (LA Referencia, hoy Red latinoamericana para la ciencia abierta) (Aguirre-Ligüera et al (2019). En 2019 se crea formalmente el Sistema, que hoy involucra a 13 instituciones, entre las cuales se encuentran las principales productoras de investigación a nivel nacional.

Suscripción a la Recomendación de la Unesco sobre Ciencia Abierta (2021). En su calidad de estado miembro Uruguay es suscriptor a la recomendación de Unesco donde se alienta a los Estados miembros a que generen mecanismos de seguimiento y evaluación para medir la eficacia y eficiencia de las políticas y los incentivos de la ciencia abierta; recopilar y difundir información relativa a los avances y las buenas prácticas en materia de ciencia abierta; elaborar un marco de seguimiento, integrado en planes estratégicos nacionales, que incluya objetivos y medidas para la aplicación de la Recomendación; elaborar estrategias de seguimiento sobre la eficacia y la eficiencia a largo plazo de la ciencia abierta, que incluyan un enfoque multipartito participativo (Unesco, 2021)

Ciencia Abierta en el MERCOSUR: situación y recomendaciones (2022). En el marco de la LXVI Reunión Especializada de Ciencia y Tecnología del MERCOSUR (RECYT) una delegación de Uruguay presentó el documento "Ciencia Abierta en el MERCOSUR: situación y recomendaciones" elaborado por un conjunto de expertos convocados por la Dirección de Nacional de Innovación, Ciencia y Tecnología (DICYT). En este documento se destaca la importancia de la ciencia abierta para mejorar la calidad, reproducibilidad e impacto de la investigación. Se discuten aspectos como el acceso abierto a publicaciones, datos abiertos de investigación, investigación abierta reproducible, recursos educativos abiertos, participación de agentes sociales y evaluación responsable de la investigación. Se enfatiza la importancia de la disponibilidad inmediata de los resultados de investigación para acelerar el desarrollo científico e innovador. Destacando además, que la investigación en la región se financia mayormente con fondos públicos, lo que resalta la relevancia de la transparencia y accesibilidad en la difusión de conocimientos científicos. En cuanto a las publicaciones científicas se plantea la promoción del acceso abierto diamante como una alternativa sostenible y transparente para la publicación científica. También se destaca la importancia de fortalecer los repositorios institucionales para garantizar la disponibilidad y preservación de publicaciones científicas. En relación a los Datos abiertos de investigación se resalta el potencial de estos para la discusión de resultados, reproducibilidad de métodos y colaboración, recomendado estimular el desarrollo de repositorios de datos interdisciplinarios y consolidar marcos normativos para la apertura de datos. Sobre la evaluación responsable de la investigación se plantea promover la apertura del conocimiento científico a través de procedimientos de evaluación y sistemas de incentivos que valoren la apertura de los resultados de investigación. Finalmente se menciona la relación ciencia abierta y propiedad intelectual recomendando analizar

los marcos jurídicos desde la perspectiva de ciencia abierta, de forma que se puedan incorporar limitaciones y excepciones al derecho de autor para la educación y la investigación en entornos digitales. (Dirección Nacional de Innovación Ciencia y Tecnología, Uruguay, 2022). A partir de este documento elevado por Uruguay a la RECyT se formó un grupo de trabajo sobre ciencia abierta en el marco de RECyT, con presencia de los 4 miembros plenos, cuyo primer producto fue una declaración sobre ciencia abierta elevada por RECyT al Grupo Mercado Común, que lo tiene en agenda para su próxima reunión en el mes de mayo.

A nivel institucional

La **Universidad de la República** ha trazado, a través de un conjunto de disposiciones generadas en los últimos años, un camino hacia una política institucional de ciencia abierta. Entre las principales reglamentaciones en este sentido se encuentran: la reglamentación del repositorio institucional y el Estatuto del personal docente. La reglamentación de repositorio determina en 2013, junto con su creación, que allí deberán alojarse todas las publicaciones producidas por docentes de la institución, así como las versiones electrónicas de las tesis de grado y posgrado (Res. CDC, No.5, 2013). En consonancia con esto en 2014, se formula la Ordenanza que lo regula, donde se determinan aspectos clave como: definición y dependencia, cometidos, contenidos y formatos, sujetos obligados, organización, acceso y uso, licencias (una de las 6 licencias Creative Commons a elección del depositante) del Repositorio Abierto de la Universidad de la República (Universidad de la República, 2014). Por su parte, el Estatuto del Personal Docente (2019) establece como deberes de los docentes el depósito de la producción en el repositorio institucional, estableciendo también que si los documentos no están en castellano se debe incluir un resumen en este idioma. En cuanto al formato de las publicaciones se indica que alcanzan toda aquella forma de difusión de la actividad académica como artículos, libros, ponencias, informes, composiciones musicales, registro de obras de arte visual o escénico y otras que existan o puedan crearse (Universidad de la República, 2019).

Como institución, la Universidad de la República, ha participado en instancias regionales e internacionales donde la temática Ciencia Abierta ha sido el foco de atención. Destacamos en este sentido la participación en el **Plenario de Rectores y Rectoras, la Asociación de Universidades Grupo Montevideo** (AUGM) Debate virtual «Política de Ciencia Abierta en América Latina. Una mirada desde los países de la AUGM» en el que participaron rectores de las universidades integrantes y la directora de la Oficina de Ciencia para América Latina y Caribe de la UNESCO (octubre 2020). Todos los panelistas destacaron la importancia de la difusión abierta del conocimiento científico para la sociedad, que quedó en evidencia con la pandemia, y la necesidad de investigadores y universidades de colaborar y compartir en el quehacer científico.

Por su parte la **Agencia Nacional de Investigación e Innovación**, en 2019 comienza a implementar una política institucional que incluye: por un lado el impulso y coordinación del Sistema nacional de repositorios digitales abiertos en ciencia y tecnología (SNRD), en forma concomitante genera un Reglamento de acceso abierto propio que establece la obligación de depositar en repositorios nacionales adheridos al SNRD los productos de investigación resultantes de proyectos de I+D y becas de posgrado financiados por ANII, bajo determinadas condiciones y sujeto a excepciones y embargos justificados.

En este contexto también se lleva adelante la creación de un repositorio gestionado por ANII, donde se puede depositar productos de investigación generados por investigadores e instituciones que no cuentan con repositorios institucionales, que a la fecha brinda servicios a instituciones sin repositorios propios, que pueden generar y gestionar allí colecciones institucionales.

En cuanto a la apertura de datos, en 2023 la ANII define una estrategia para impulsar la apertura de datos de investigación a través de un incentivo positivo para proyectos de I+D financiados por uno de sus principales fondos horizontales para investigación (Fondo Clemente Estable), que pueden optar por una financiación del 10% adicional sobre el monto total del proyecto si presentan en la postulación un Plan de gestión de datos sujeto a principios FAIR y se comprometen a la publicación de datos resultantes de la investigación bajo licencias que permitan su reutilización. Este incentivo se concibe como un piloto, que se evaluará extender a otros instrumentos. Esta iniciativa va de la mano con la creación de un repositorio de datos específico para proyectos de la institución y un plan de capacitación, orientado en una primera etapa a beneficiarios y evaluadores del Fondo Clemente Estable, próximas etapas en etapa de diseño.

Infraestructuras

En relación con las políticas públicas mencionadas, se presenta un análisis de las infraestructuras que se han establecido en los últimos años para respaldar la ciencia abierta en Uruguay. Esto incluye plataformas digitales, repositorios bibliográficos y de datos, y otros recursos destinados a facilitar el acceso abierto a la información y promover la colaboración.

A los efectos de organizar la presentación de las infraestructuras disponibles en el país se agrupan según el material que reúnen, generando dos categorías, las vinculadas al acceso abierto a publicaciones y las vinculadas al acceso a datos e imágenes.

Infraestructuras de Acceso abierto a publicaciones

Colibrí (Conocimiento Libre. Repositorio Institucional) [https://www.colibri.udelar.edu.uy/jspui/] 2014. La creación y desarrollo del repositorio COLIBRI es parte de una política de Acceso Abierto adoptada por la Udelar en el año 2013. Como principal institución generadora de conocimiento del país, decide crear un repositorio de acceso libre y gratuito a sus publicaciones, adhiriendo al principio de que la producción académica financiada con fondos públicos debe ser accesible al conjunto de la sociedad y se debe constituir en un patrimonio de la institución, y, en consecuencia, es necesario crear las condiciones para su adecuado mantenimiento, acceso y utilización (Proyecto Repositorio Institucional COLIBRI, 2014). El 7 de octubre de 2014 se aprueba la Ordenanza del Repositorio (Resolución n.o 15 del Consejo Directivo Central) y se determina el uso de las licencias Creative Commons 4.0 (Resolución n.o 16 del Consejo Directivo Central). Se define como "Colección digital de acceso abierto que agrupa y resguarda la producción de la Universidad de la República, con la finalidad de preservar su memoria, poner dicha producción a disposición de toda la sociedad y contribuir a incrementar su difusión y visibilidad, así como potenciar y facilitar nuevas producciones" El Repositorio Institucional COLIBRI integra el Sistema Nacional de Repositorios Abiertos de Ciencia y Tecnología (SILO) y a través de SILO tiene presencia en LA REFERENCIA y OpenAIRE Explore (Seroubian M, 2022).

Biblioteca digital y accesible (BIDYA) 2019 [https://www.colibri.udelar.edu.uy/jspui/handle/20.500.12008/9313] La creación de esta colección tiene como marco la ratificación de Uruguay del Tratado de Marrakech (primer país latinoamericano). Dicho tratado surge en el ámbito de la Organización Mundial de la Propiedad Intelectual (OMPI) y establece excepciones y limitaciones a la ley de derecho de autor en favor de las personas con discapacidad visual, motriz o con otra discapacidad para el acceso al texto impreso. El Decreto del Poder Ejecutivo 295/2017 reglamenta y hace posible la aplicación en Uruguay

del tratado. El decreto define las obras que se consideran comprendidas y los requisitos que debe cumplir la producción de una obra bajo esta excepción, las personas beneficiarias de la excepción y quienes pueden producir obras, las instituciones autorizadas y las personas beneficiarias. Las obras deberán estar en un formato accesible que permita un acceso igualitario como el de otras personas sin discapacidad. Bajo la órbita de esta reglamentación es que se crea la Biblioteca Digital y Accesible, BIDYA constituyendo una de las primeras bibliotecas digitales accesibles en la región. La misma fue presentada por el Núcleo de Recursos Educativos Abiertos y Accesibles de la Udelar y la Unión Nacional de Ciegos del Uruguay, al Programa FRIDA, en 2016. El proyecto fue financiado y seleccionado como uno de los diez innovadores en América Latina y el Caribe. A través del mismo se generó una colección de 500 textos y materiales accesibles para personas ciegas o con baja visión, orientados a los planes de estudio de enseñanza primaria y media de Uruguay, colecciones alojadas en el Repositorio Institucional COLIBRI de la Udelar. En una segunda etapa, buscando ampliar el impacto y los alcances de la Biblioteca Digital Accesible, a todo el Sistema Nacional de Educación, se incorporaron textos y otros materiales de nivel terciario. Actualmente se disponen materiales accesibles para estudiantes de las Facultades de Psicología y Ciencias Sociales de la Udelar.

Silo [https://silo.uy/] es el Portal del Sistema nacional de repositorios de acceso abierto en ciencia y tecnología, que reúne en un catálogo centralizado la producción disponible en los repositorios de acceso abierto de las instituciones adheridas. Utiliza tecnología desarrollada por LA Referencia para la cosecha y normalización de los registros, a la fecha cosecha, estandariza y expone los metadatos de más de 30.000 productos de investigación de 13 instituciones.

SciELO Uruguay [https://www.scielo.edu.uy/]**.** SciELO Uruguay es una biblioteca electrónica que abarca una colección de algunas revistas científicas uruguayas, pertenece a la Red SciELO y fue inaugurada en 2002 y certificada en 2017. Uno de los objetivos de la biblioteca electrónica es dar acceso completo y gratuito a una colección de revistas y sus artículos. SciELO exige a las revistas un conjunto de requisitos básicos de edición de revistas. Luego de la Declaración de la Unesco sobre Ciencia Abierta estos requisitos han ido incrementando y haciéndose obligatorios como por ej: publicación continua, reconocimiento anual de pares revisores, obligación de los autores de cada artículo de expresar qué función cumplieron en la elaboración y publicación del trabajo y que se exprese qué Editor de la revista revisó el artículo. En el último encuentro de SciELO 25 años se explicitó que se empezará a exigir que los datos provenientes de las investigaciones se publiquen en un repositorio de datos y se exprese donde exactamente así como también pensar en la revisión abierta de pares, inclusión de pre-prints en el flujo de publicación de la investigación científica (Zetter Patiño, 2023).

Latindex Uy. [https://latindex.org/] Es el nodo uruguayo de la Red Latindex que colabora en el asesoramiento, selección y evaluación de las publicaciones periódicas uruguayas con el fin de integrarlas a la red iberoamericana. Entre sus tareas se han incluido dar formación a los editores, responder consultas y luego realizar la evaluación de la publicación para visibilizarlas y contribuir a la mejora de la calidad de las mismas. La participación de Uruguay en este nodo comenzó en 2010.

Asociación Uruguaya de Revistas Académicas (AURA). [https://aura.edu.uy/] Es la asociación de editores científicos uruguayos creada en 2015 y donde se procura formar, acompañar y representar a los editores uruguayos de revistas académicas de acceso abierto. Se realizan continuos cursos de formación y se responde a los editores de sus dudas con respecto a buenas prácticas en la edición científica, sitios donde postular y formación en herramientas de gestión editorial. Las revistas cumplen con algunos preceptos de la línea diamante del acceso abierto como no cobro de APC, utilización de software libre para la gestión

de la revista como OJS que permite la interoperabilidad con otros sistemas como ORCid, OPENAire y DOAJ, y de a poco se están implementando algunas como puesta a disposición de los datos y revisión por pares abierta.

REA Ceibal. [https://rea.ceibal.edu.uy/] Uruguay cuenta desde 2007 con el Plan Ceibal, un programa de innovación educativa con tecnologías digitales, que tiene como uno de sus mayores componentes al programa internacional *One Laptop per Child*. En este marco es que se desarrolla el repositorio de recursos educativos abiertos REA Ceibal. Para proporcionar soporte a este repositorio se trabaja con un equipo de contenidistas, formado por docentes de educación primaria y media, especializados en el tema. Su principal objetivo es promover una comunidad educativa de docentes y estudiantes para crear y compartir recursos bajo la cultura de lo abierto (Ceibal, 2024).

Otros repositorios de publicaciones. A nivel institucional pueden encontrarse otras iniciativas ya sea públicas o privadas como los repositorios institucionales de: Instituto de Investigaciones Biológicas Clemente Estable, Fundación Ceibal, Consejo de Formación en Educación, LATU, Instituto Universitario Asociación Cristiana de Jóvenes, Instituto Pasteur de Montevideo, Instituto Nacional de Investigación Agropecuaria, Universidad ORT del Uruguay, Universidad de Montevideo, Universidad Católica del Uruguay y Universidad Tecnológica. Todos ellos se encuentran reunidos en SILO.

Repositorios de datos e imágenes abiertas

Redata (ANII) [https://redata.anii.org.uy/] Es un repositorio específico para conjuntos de datos de investigación (redata.anii.org.uy) lanzado en 2024, donde cualquier investigador puede publicar sus datos y abierto a grupos o instituciones que quieran alojar allí colecciones específicas. Este repositorio también forma parte de SILO (repositorio de carácter nacional mencionado antes) lo que permite una recuperación integrada de datos y publicaciones. Entre los objetivos de Redata se encuentran: incentivar la publicación y reutilización de los datos resultantes de investigaciones; facilitar la identificación, verificación y reutilización de datos; incrementar la reproducibilidad, transparencia y evitar la duplicación de esfuerzos en la recolección o generación de datos (ANII, 2024). Si bien REDATA es desarrollado por ANII, pretende extender su alcance y que todas las instituciones, grupos de investigación o comunidades académicas puedan gestionar allí sus propias colecciones.

Microscopio Virtual (2021-2023) [https://microscopiovirtual.udelar.edu.uy/] Se trata de un Proyecto de la Red de Macrouniversidades y que surge como un producto del análisis colectivo realizado en 2021 por un grupo de especialistas en tecnología y educación conformado por: Universidad de la República de Uruguay (UdelaR); Universidad de Buenos Aires (UBA); Universidad Nacional de Córdoba (UNC); Universidad Nacional de La Plata (UNLP); Universidad Veracruzana (UV) y Universidad Nacional Autónoma de México (UNAM). El equipo de trabajo se conformó con personal de diferentes disciplinas: médicos, veterinarios, farmacobiólogos, especialistas en Centros de datos, diseñadores de sistemas y repositorios, especialistas en educación y tecnología y profesionales de la información. El proyecto de "Microscopía virtual y recursos para estudios de la salud" se propone como herramienta para la docencia, de carácter abierto, que pueda ser utilizada por cualquier Universidad. Tiene por objetivo crear un grupo de desarrollo tecnológico, dentro de la Red de Macrouniversidades, con el propósito de trabajar de forma conjunta para el desarrollo de herramientas tecnológicas. Concretando esta idea con la construcción de un portal web con imágenes

digitales de microscopía que aporten las diferentes universidades participantes. La Udelar, como integrante del proyecto, desarrolló su propio Repositorio Microscopio Virtual y alimenta con recursos al "Portal de Microscopía virtual y recursos para estudios de la salud".

***Catálogo Nacional de Datos Abiertos. Agesic* (2012)** [https://catalogodatos.gub.uy/]. Esta base, gestionada por el equipo de Gobierno Abierto de AGESIC permite acceder a datos abiertos de organismos públicos, academia, organizaciones de sociedad civil y empresas privadas. Al 2024 cuenta con 2.482 conjuntos de datos publicados pertenecientes a 62 instituciones.

Mirador de gobierno abierto y ley de acceso a la información pública (2022) [https://miradordegobiernoabierto.agesic.gub.uy/] Este espacio es una herramienta de monitoreo que reúne las iniciativas contenidas en los Planes de acción de Gobierno Abierto en Uruguay. A través del mismo se busca aumentar la transparencia y brindar posibilidades de control a la ciudadanía.

Otras iniciativas

Sin ánimo de ser exhaustivos se mencionan algunas de las iniciativas desarrolladas por instituciones académicas y organizaciones de la sociedad civil en el ámbito de la ciencia abierta.

En ese ámbito podemos destacar a *Datysoc* [https://datysoc.org/] que es una organización que trabaja en Gobierno Abierto, Datos Abiertos, Tecnología Cívica, Transparencia y Acceso a la Información. Esta organización civil ha desarrollado proyectos y programas que apuntan a la información e inclusión ciudadana a través del uso de los datos abiertos.

También a *Wikimedistas de Uruguay* [https://wikimedistas.uy/] que es un grupo de usuarios y usuarias de Wikipedia y Wikimedia que promueven la circulación del conocimiento libre de Uruguay hacia el mundo utilizando las herramientas que brinda Wikimedia. Realizan cursos, vínculos con Instituciones afines y jornadas donde se editan los artículos y proyectos asociados a Wikipedia con material libre y abierto. Especialmente, utilizan Wikidata ya que es una de los proyectos que compendian datos a ser utilizadas por las Wikipedias de todos los idiomas.

Vinculado al ámbito académico se ha consolidado el proyecto de Ciencia Ciudadana *Biodiversidata* [https://biodiversidata.org/], es un proyecto colaborativo que procura poner a la vista de la comunidad los datos abiertos generados de su investigación acerca de la biodiversidad en Uruguay. Su aporte a la ciencia abierta es poner a disposición datos, códigos y artículos de forma colaborativa para que puedan ser utilizados en la difusión, educación e investigación científica además de que sean herramientas para la toma de decisiones y elaboración de políticas públicas ambientales.

También desde la academia se ha promovido el trabajo en Recursos Educativos Abiertos a través del Núcleo Interdisciplinario de Recursos Educativos Abiertos y Accesibles (Núcleo REAA) siendo este un espacio para la producción interdisciplinaria, que ha convocando a diversos actores y ha generando una red entre los espacios académicos y no académicos, en el ámbito nacional e internacional durante el período 2015- 2017 (Rodés y Motz, 2022). Otro actor relevante en el ámbito académico es el Grupo Educación Digital Abierta (ciEDA) quien integra actores académicos de la Universidad de la República, agentes de la sociedad civil y del exterior.

Sistemas de evaluación de la ciencia

La discusión sobre los mecanismos de evaluación de investigadores (un aspecto clave para el impulso de la ciencia abierta) está siendo objeto de reflexión y revisión en distintos ámbitos institucionales, como ejemplo el proceso de discusión a la interna de UdelaR que se materializa en 2024 con el ciclo *"Evaluar es necesario: encuentros sobre prácticas, tradiciones y necesarias renovaciones en la evaluación académica"*. Esta iniciativa es impulsada por el Prorrectorado de Investigación junto a la Unidad Académica de la CSIC y el Núcleo Ciencia, Tecnología e Innovación para un Nuevo Desarrollo (CiTINDe). Se trata de una serie de eventos que convoca a personas de diversos ámbitos del Sistema de Ciencia, Tecnología e Innovación del país y que busca reflexionar sobre qué evaluar, qué mejoras son necesarias y los desafíos que eso supone. Entre las diferentes visiones y perspectivas que se han dado lugar se encuentran los cuestionamientos a la medición y formas de evaluación basadas estrictamente en sistemas de comunicación que no contemplan dimensiones de la Ciencia Abierta.

Por otra parte desde el CONICyT se ha promovido la realización de una consultoría, encargada a Fernanda Beiguel, cuyo objetivo es realizar un "análisis y sistematización de las experiencias nacionales e internacionales en materia de evaluación de la carrera del investigador/a, así como la elaboración de recomendaciones para aplicar en los diferentes instrumentos de promoción de la investigación" (DICYT, 2024). En el informe final de dicha consultoría se presenta un trabajo integral que contempla varias dimensiones vinculadas a la evaluación y entre ellas algunas menciones relacionadas a la ciencia abierta como los incentivos al acceso abierto, la promoción de la publicación por la vía de acceso abierto diamante, la inclusión en el Curriculum institucional (Cv.uy) de servicios de indexación que permitan una clara identificación sobre si las publicaciones son de acceso abierto y la promoción de ejercicios exploratorios de evaluación abierta. El informe destaca como debilidades la ausencia de acciones directas en la evaluación académica que sean de incentivo para la publicación en los repositorios de acceso abierto y para la publicación de los datos de investigación (Beigel, 2024).

En diálogo con la discusión sobre evaluaciones abiertas también se ha trabajado desde la perspectiva de la generación de agendas abiertas de investigación. En este sentido se destaca la figura de la tradición extensionista de las universidades públicas latinoamericanas quienes han promovido desde sus inicios el trabajo con la comunidad. Entre los aportes al tema se encuentra un estudio de Natalia Gras y Claudia Cohanoff, donde se analizan las formas en que las políticas universitarias de estímulo a la investigación contribuyen al desarrollo de procesos abiertos de producción y uso socialmente valioso del conocimiento. Al mismo tiempo que reflexiona sobre los procesos de evaluación involucrados en la selección de las propuestas de investigación abierta y sobre la posibilidad de fomentar agendas abiertas de investigación en la Universidad (Gras & Cohanoff, 2022).

Conclusiones

Luego de haber ordenado y sistematizado las acciones, iniciativas e infraestructuras vinculadas a la ciencia abierta en Uruguay lo primero que identificamos es la ausencia, a nivel de gobierno, de normativa o política explícita que se posicione sobre el tema y/o impulse alguno de sus componentes. Esto se hace más evidente si lo comparamos con otros países de la región que desde hace varios años han avanzado en materia de legislación.

Esto no quiere decir que a nivel de gobierno no se hayan realizado avances en el tema o no se hayan dado señales de que hay una intención de recorrer el camino de la ciencia abierta. Tal es el caso de las iniciativas de gobierno abierto y de las propuestas a nivel de MERCOSUR que han llevado adelante organismos como la DICyT.

Según los diferentes componentes de la Ciencia Abierta podemos distinguir que en cuestiones de acceso abierto es donde se ha avanzado con mayor fuerza y concreción de políticas y acciones. Tal es el caso de los repositorios institucionales y el nodo nacional, así como las reglamentaciones internas que regulan los depósitos de los archivos. También en el área del acceso abierto a publicaciones numerosas revistas académicas han puesto en práctica algunos de sus preceptos como el acceso gratuito a los textos y lecturas, el uso de licencias creative commons y programas de acceso abierto para la gestión de la edición pero aún falta implementar la gestión de datos abiertos y la revisión abierta por pares.

En cuanto a la presencia de la ciencia ciudadana, se observa una participación activa de organizaciones e instituciones que generan, difunden y a su vez acceden a datos abiertos para proporcionar conocimiento y formación a la comunidad. Algunos ejemplos son Biodivesidata y Datysoc.

El componente de evaluación de la ciencia se encuentra en plena etapa de discusión en diferentes instituciones del país, algunos indicios permiten pensar en un cambio en el sistema de evaluación de los investigadores y en este sentido es de esperarse un viraje hacía métodos de evaluación que contemplen aspectos de la ciencia abierta y adaptados a mediciones acordes al contexto local.

En términos de oportunidades se puede visualizar la presencia del país y sus instituciones en proyectos regionales de alto impacto, como es el caso de SciElo, La Referencia y Latindex. América Latina se posiciona como un modelo pionero en muchos aspectos vinculados a la Ciencia Abierta y poder ser partícipe activo de las diferentes iniciativas, proyectos y programas se vuelve un factor clave para países como Uruguay, que aún se encuentran en etapa de desarrollo sobre el tema. Es por esto que también se señalan los proyectos cooperativos y las alianzas regionales como oportunidades para consolidar y principalmente poder sostener en el tiempo diferentes infraestructuras y programas de colaboración que promuevan y puedan dar sostenibilidad a las iniciativas de ciencia abierta.

En síntesis, podemos identificar una serie de iniciativas y recursos que promueven algunos de los componentes de la ciencia abierta en el país, pero no se evidencia una política nacional de forma explícita, más bien los esfuerzos se encuentran arraigados en diferentes instituciones. A nivel de latinoamérica, si bien se acompañan varios proyectos regionales aún queda mucho camino por andar para alcanzar algunos de los avances que otros países han logrado en el tema.

Agradecimientos:

A Juan Maldini por sus comentarios al texto.

Bibliografía

Aguirre-Ligüera, N., Maldini, J., & Fontans, E. (2019). Acceso abierto a la producción científica de Uruguay: poca historia en 10 años (2009-2018). *Palabra Clave (La Plata)*, *9*(1), e079. https://doi.org/10.24215/18539912e079

Alperin, J. P. y Fischman, G. E. (Eds.) (2015). *Hecho en Latinoamérica: acceso* abierto, revistas académicas e innovaciones regionales. Buenos Aires: CLACSO. http://biblioteca.clacso.edu.ar/clacso/se/20150722110704/HechoEnLatinoamerica.pdf

ANII. Agencia Nacional de Investigación e Innovación. (2019) *Reglamento de acceso abierto.* https://www.anii.org.uy/upcms/files/listado-documentos/documentos/anii-reglamento-de-acceso-abierto.pdf

Babini, D., & Rovelli, L. I. (2020). *Tendencias recientes en las políticas científicas de ciencia abierta y acceso abierto en Iberoamérica*. https://www.memoria.fahce.unlp.edu.ar/libros/pm.5293/pm.5293.pdf

Beigel, F. (2024) *Asesoría Conicyt. Un estudio de la evaluación académica en Uruguay en perspectiva reflexiva: Informe ejecutivo*. Conicyt. https://www.conicyt.gub.uy/sites/default/files/2024-04/Informe%20Ejecutivo%20CONICYT%20final.pdf

Chan, L.; Okune, A.; Hillyer, R.; Albornoz, D. y Posada, A. (Eds.) (2019). *Contextualizing Openness: Situating Open Science*. Ottawa: University of Ottawa Press - IDRC. https://www.idrc.ca/en/book/contextualizing-openness-situating-open-science

Declaración de Panamá para la Ciencia Abierta (2018). https://hiperderecho.org/wp-content/uploads/2018/11/declaracion_panama_ciencia_abierta.pdf

De Filippo, D. y D'Onofrio, M. G. (2019). Alcances y limitaciones de la ciencia abierta en Latinoamérica: análisis de las políticas públicas y publicaciones científicas de la región. *Hipertext.net*, 19, 32-48. https://www.raco.cat/index.php/Hipertext/article/view/360106

De Giusti, Marisa R.. (2022). Ciencia abierta: el corazón del problema. Informatio, 27(1), 309-335. Epub 01 de junio de 2022.https://doi.org/10.35643/info.27.1.3

Gras, N., & Cohanoff, C. (2022). Agendas abiertas de investigación y el abordaje de problemas en interacción social: la experiencia de la Universidad de la República de Uruguay. *Informatio. Revista del Instituto de Información de la Facultad de Información y Comunicación*, *27*(1), 168-198. https://doi.org/10.35643/Info.27.1.2

Gualdron Guerrero, O. E. (2017). Política pública de ciencia abierta en Colombia: presente y futuro. https://repository.urosario.edu.co/items/a6ed1ce8-d2ee-4b1a-aeeb-80f2f757dc3f

Prieto, D. (2022). Ciencia Abierta: desafíos y oportunidades para Uruguay y el Sur Global. *Informatio*, *27*(1), 253-283. Epub 01 de junio de 2022.https://doi.org/10.35643/info.27.1.5

Reunión de Consorcios de Iberoamérica y el Caribe (2017). Declaración de la Primera Reunión de Consorcios de Iberoamérica y el Caribe.

Rico Castro, P. (2023). La política nacional de ciencia abierta. Mesa redonda" Nuevas estrategias en investigación: ciencia abierta, evaluación, reconocimiento"(2023).

Rodés Paragarino, V., & Motz Carrano, R. (2022). Construcciones interdisciplinarias en Educación y Ciencia Abiertas. *Informatio*, *27*(1), 146-167. https://doi.org/10.35643/Info.27.1.10

Seroubian, M. (2022). Acceso abierto y ciencia abierta. Experiencia desde la gestión del repositorio institucional COLIBRI de la Universidad de la República. *Informatio*, *27*(1), 284-308. Epub 01 de junio de 2022.https://doi.org/10.35643/info.27.1.6

Unesco. *Recomendación de la Unesco sobre la Ciencia Abierta* (2021) https://unesdoc.unesco.org/ark:/48223/pf0000379949_spa

Universidad de la República (Uruguay). Comisión Sectorial de Investigación Científica (2024) *Evaluar es necesario: Encuentros interinstitucionales sobre evaluación académica.* https://www.csic.edu.uy/content/evaluar-es-necesario-encuentros-interinstitucionales-sobre-evaluaci%C3%B3n-acad%C3%A9mica

Universidad de la República (Uruguay). Dirección General de Jurídica (2019). *Estatuto del Personal Docente*. https://dgjuridica.udelar.edu.uy/estatuto-del-personal-docente-aplicar-a-partir-del-ano-2021/

Universidad de la República (Uruguay) (2014). *Ordenanza del Repositorio Abierto de la Universidad de la República*. https://www.colibri.udelar.edu.uy/jspui/Documentos/Ordenanza-aprobada.pdf

Universidad de la República (Uruguay). Resolución 5 del Consejo Directivo Central. (2013) https://www.expe.edu.uy/expe/resoluci.nsf/4e1fd2c2a317193a03256dcc003b902f/24c4bde88b32fb6103257b66005ce05c?OpenDocument&Highlight=0,acceso,abierto

Uruguay. Ceibal. (2014) *Repositorios de recursos abiertos*. https://rea.ceibal.edu.uy/acerca-de

Uruguay. Ministerio de Educación y Cultura. Dirección Nacional de Innovación Ciencia y Tecnología, (2022) *Ciencia abierta en el MERCOSUR: situación y recomendaciones.* https://www.gub.uy/ministerio-educacion-cultura/sites/ministerio-educacion-cultura/files/documentos/noticias/Ciencia_Abierta_MERCOSUR-DICYT-MEC_20220920.pdf

Uruguay. Ministerio de Educación y Cultura. Dirección Nacional de Innovación Ciencia y Tecnología. (2024) *Revisión del sistema de evaluación académica en Uruguay*. https://www.conicyt.gub.uy/node/574

Uruguay. Presidencia de la República. Agesic (2021) *Plan de Acción Nacional de Gobierno Abierto 2021-2024, 5to*. https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/gobierno-abierto

Zetter Patiño, J. (2023). Reseña ejecutiva: "Reunión SciELO 25 años" Ciencia Abierta con: Impacto, Diversidad, Equidad, Inclusión y Accesibilidad (IDEIA). https://b35957d240.imgdist.com/public/users/Integrators/BeeProAgency/965042_949606/Rese%C3%B1a%20SciELO%2025.pdf

Magela Cabrera Castiglioni. Magíster en Información y Comunicación por la Facultad de Información y Comunicación, Universidad de la República (UdelaR), Uruguay. Licenciada en Bibliotecología por la Escuela Universitaria de Bibliotecología y Ciencias Afines(UdelaR). Docente Grado 3 de la Facultad de Información y Comunicación, responsable de las unidades curriculares Fuentes de Información Especializada y Alfabetización en información. ORCID: 0000-0002-6289-5538

Carina Patrón. Licenciada en Bibliotecología. Directora del Departamento de Documentación y Biblioteca. Facultad de Odontología, Universidad de la República. VicePresidenta de AURA. ORCID: https://orcid.org/0000-0002-8662-9437

Mabel Seroubian. Magíster en Dirección Estratégica de Tecnologías de la Información por la Universidad Europea Miguel de Cervantes y la Universidad Europea del Atlántico, España. Licenciada en Bibliotecología, Universidad de la República (Uruguay). Directora del Departamento de Sistemas Documentales del Servicio Central de Informática (Udelar), Coordinadora del Repositorio Institucional COLIBRI. ORCID: http://orcid.org/0000-0001-9722-8500

Evaluación de estrategias de servicios de marcación y de publicación para artículos científicos

Santiago Soler[1], Dolores García[2], Gonzalo Luján Villarreal[3], Adela Ruiz[4]

Palabras clave

marcación, interoperabilidad, publicaciones científicas, XML JATS

Eje temático

Comunicación académica, científica y cultural en abierto

Resumen

Este trabajo pone el foco en la evaluación de las distintas herramientas, disponibles actualmente, para el marcado y la generación de XML JATS que puedan ser utilizadas por las instituciones editoras de revistas científicas. Para esto, se experimentará con herramientas que estén dentro del contexto de Open Journal Systems (OJS), sistema de gestión de revistas científicas, con el objetivo de tener un reconocimiento de las herramientas y sus funcionalidades. Se busca optimizar la funcionalidad del proceso editorial, de ser posible, o plantear qué posibles trabajos futuros permitirían que las revistas de la Universidad Nacional de La Plata hagan uso de estas herramientas.

En conclusión, existe la posibilidad de implementar un flujo editorial completo dentro de OJS con un conversor a XML JATS, con extracción automática de metadatos y, adicionalmente con herramientas de presentación a varios formatos a partir del XML, como PDF, HTML, etc. Sin embargo, las múltiples dificultades que presenta su implementación determinan que su futura aplicación en los procesos editoriales requiera de importantes cambios en la manera en la que se desarrollan los flujos actuales de trabajo.

Introducción

En la actualidad, las publicaciones científicas implementan diversas prácticas orientadas a optimizar el circuito editorial y a incrementar la visibilidad de sus artículos. Entre ellas, la marcación de artículos es uno de los procesos que permite potenciar la gestión editorial, aumentar la visibilidad, generar datos desagregados sobre la ciencia publicada (para analizar qué, quién, cómo, dónde, con quiénes, entre otros), crear grandes sistemas de información y defender la soberanía del conocimiento (Rozemblum, 2021).

1 Universidad Nacional de La Plata, PREBI-SEDICI, santiago.soler@sedici.unlp.edu.ar

2 Comisión de Investigaciones Científicas, CESGI, dolores.garcia@sedici.unlp.edu.ar

3 Universidad Nacional de La Plata, PREBI-SEDICI, y Comisión de Investigaciones Científicas, CESGI, gonetil@prebi.unlp.edu.ar

4 Universidad Nacional de La Plata, Coordinación de Revistas Científicas, adelaruiz@perio.unlp.edu.ar

En línea con esta perspectiva, el principal objetivo de este trabajo es evaluar las distintas herramientas (*plugins*) que podrían instalarse y estar disponibles en OJS para realizar el proceso de marcación y para generar documentos en formato XML JATS que puedan ser utilizados por las instituciones editoras de revistas científicas. Por un lado, se busca desarrollar herramientas que les permitan a los equipos responsables de las revistas alcanzar la independencia editorial sin depender de terceros o estar sujetos a las restricciones que establecen sistemas externos a las instituciones editoras o empresas proveedoras de este tipo de servicios. Por otro lado, se pretende contribuir a las políticas de fortalecimiento editorial que llevan adelante espacios como la Coordinación de Revistas Científicas de la UNLP, que regularmente asisten y brindan capacitaciones para que los equipos editoriales adquieran los conocimientos que les permitan utilizar las herramientas que demanda el etiquetado XML, en pos de ganar autonomía en estas tareas y hacer de la marcación un proceso sostenible en el tiempo (Ruiz et al., 2022).

Las herramientas evaluadas fueron, en gran medida, las que están dentro del contexto de Open Journal Systems (OJS), el sistema de gestión editorial que utiliza en la actualidad gran parte de las publicaciones científicas, y que abordan el problema de su integración al flujo editorial. Además, se consideraron algunos instrumentos externos que no estaban integrados en OJS, pero que consideraban el problema de la marcación y la generación de XML JATS.

En lo que refiere a la metodología de trabajo, se llevaron a cabo varias pruebas exploratorias en un OJS instalado de forma local. A esa plataforma se le instalaron los distintos *plugins* y herramientas que se consideraron evaluar para abordar el problema planteado y se hicieron las pruebas, mediante la utilización de varios artículos reales.

Marco teórico

En la era digital, OJS es la plataforma de código abierto que usan gran parte de las revistas científicas y académicas para gestionar y publicar sus contenidos. Resulta flexible y se ajusta a las necesidades de cada equipo editorial, además, se puede descargar de forma gratuita e instalar en un servidor web local. Como plantea García (2018):

Esta herramienta se utiliza para la creación y configuración de una revista y de su sitio web, así como también para realizar las tareas operativas propias de una revista: el proceso de envío del autor, la revisión por pares, la edición, la publicación, el archivo y la indexación (p. 29).

Las principales características de OJS son: se instala y se controla localmente; los editores configuran requisitos, secciones, proceso de revisión, etc.; presentación en línea, revisión doble ciego y gestión de todo el contenido; indización completa del contenido; puede ser responsiva; permite notificaciones vía mail para lectores; ayuda para envío y procesamiento de artículos; asistencia en línea; soporte multilingüe; entre otras.

En lo que respecta a los formatos HTML, PDF, EPUB, fueron incorporados por el mundo editorial a partir del traspaso de lo impreso a lo digital, con el fin de facilitar la lectura en diferentes dispositivos.

El formato más difundido para la publicación de textos es el formato de documento portátil o PDF, pero no es raro encontrar autoarchivos de materiales hechos en formatos .doc o .docx u otro tipo de formatos de texto editable como .odt. En estos casos siempre es recomendable la transformación del material al formato PDF. (...) A los fines de la preservación digital el formato recomendado es el PDF/A. El PDF/A es

el estándar más común para los documentos de texto con formato, pero muchas entidades que ofrecen contenidos en formatos de texto electrónico en formato EPUB. Ambos formatos están basados en XML. (De Giusti, 2016)

Con el paso del tiempo y con la preponderancia de temas como la interoperabilidad, las métricas y la preservación de los recursos científicos, es que surge la necesidad de implementar esquemas de metadatos estandarizados para la publicación de los contenidos. Este proceso de normalización se realiza mediante la utilización de un lenguaje de marcas para la confección de los documentos a ser publicados, se asignan metaetiquetas que transforman los elementos constitutivos de un artículo en un conjunto de elementos interoperables, generando una estructura de comunicación (Rozemblum, 2021).

El estándar que se emplea es XML (Extensible Markup Language)[5] y está compuesto por etiquetas y atributos con una estructura y una semántica particular para la descripción de metadatos que dotan de significado al contenido de un texto marcado (Redalyc, 2016a). A esto se le suma el estándar técnico JATS (Journal Article Tag Suite), que es un conjunto de elementos y atributos XML para la descripción del contenido gráfico y textual de artículos científicos. Se estructura en Front –con la información bibliográfica del artículo y la revista que lo publica–, Body –el contenido propiamente dicho– y Back –referencias, apéndices, glosarios– (Redalyc, 2016b; National Information Standards Organization, 2019; McGlone, 2013; Böschen, 2021).

Etapas del flujo editorial en OJS

En el contexto de las revistas científicas, se reconocen seis etapas, que podrían variar según para qué fueran destinadas, pero este trabajo se centrará en el contexto de las etapas que ofrece el flujo editorial de OJS y la que actualmente se emplea en el contexto de las revistas de la Universidad Nacional de La Plata. Tener un buen reconocimiento de estas etapas resulta fundamental para poder separar el problema en tres partes y entender en qué momento del flujo editorial convendrá abordar cada problema.

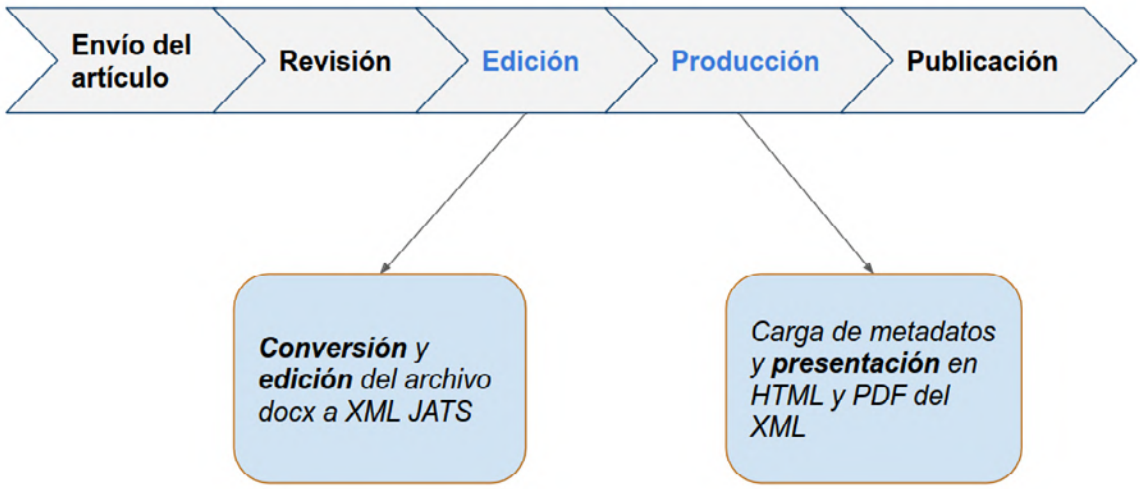
En consecuencia, se propondrán tres ejes: **1) conversión, 2) edición de XML y 3) presentación**. La **conversión** se refiere a la transformación automática del documento a XML JATS; en lo que respecta a **edición**, el eje se plantea para atacar los problemas que se arrastran de la transformación, donde puede haber errores o correcciones que precise el XML construido; y, en cuanto a la **presentación**, una vez que se obtiene un XML JATS correcto, que represente el documento original, es posible exportar a distintos formatos como PDF, HTML, ePub, etc. (Cho, 2022; McGlone, 2013). Por lo que, la conversión de XML a estos formatos será abordado en esta etapa.

Una vez que son reconocidas las etapas, se procederá a situar, dentro del flujo editorial, dónde se abordará cada problemática. Dentro de OJS se encuentran las siguientes etapas del flujo editorial: **1) envío del artículo, 2) revisión, 3) editorial, 4) producción y 5) publicación** (ver figura 1). En consonancia con el objetivo de este trabajo, que es adobar el problema de obtener el JATS XML a partir de la versión final de documento, se plantea la necesidad de llevar a cabo el procedimiento luego de que haya sido completada la etapa de revisión, es decir en la etapa tres, de edición, que es la que cuenta con el documento definitivo. A partir de allí, podrá trabajarse en la conversión a JATS XML y, luego, se podrá proceder a la edición del

5 Es un metalenguaje derivado de la norma ISO 8879, diseñado para la publicación electrónica y el intercambio de información en la web (https://www.w3.org/XML/).

XML obtenido. Posteriormente, en la etapa de producción se planteará la tarea de mejorar la presentación del documento, donde se aplicarán distintas técnicas sobre el XML para lograr la exportación de los distintos tipos de documentos.

Figura 1. Ejes dentro del flujo editorial en OJS



Fuente: Soler et al., 2023.

Eje 1: Conversión a XML JATS

Una vez que se concluye la revisión por pares y se obtiene la versión lista para ser publicada el equipo editorial de la revista se encarga de formar la publicación (Becerril-García et al., 2023; Becerril-García y Aguado-López, 2018). Es por eso que en la etapa de edición del flujo editorial ya se cuenta con el artículo terminado, por eso se busca obtener en esta etapa el XML JATS. Si se puede obtener el XML en esta etapa, la de producción, estará preparado para la exportación de los distintos formatos que se generan a partir de un XML JATS.

A continuación, se enumeran las distintas herramientas que se encontraron y que se indagaron en profundidad, mediante pruebas en un OJS instalado de forma local y, al mismo tiempo, integrarlas para hacer evaluaciones experimentales. Los instrumentos que se analizarán para la etapa de conversión son (ver tabla 1): docxToJats,[6] desarrollado en el lenguaje de programación php; docxConverted,[7] plugin para OJS 3.X, que integra la herramienta anteriormente mencionada; DOCX2JATS,[8] desarrollado en el lenguaje de programación JAVA.

docxToJats: Es una biblioteca desarrollada en PHP que tiene como objetivo convertir archivos en formato DOCX, que cumplen con el estándar de OOXML, a formato XML JATS. Surge como propuesta de uno de los desarrolladores de Public Knowledge Project (PKP), iniciativa de investigación, sin fin de lucro,

6 https://github.com/Vitaliy-1/docxToJats/tree/main
7 https://github.com/Vitaliy-1/docxConverter
8 https://github.com/Vitaliy-1/DOCX2JATS

de la Facultad de Educación de la Universidad de British Columbia, que se encuentra detrás del desarrollo de OJS. La herramienta busca abordar el problema planteado en esta investigación; la idea es crear una estructura básica de XML JATS a partir de un documento dado en formato DOCX. El plugin en sí mismo era de difícil uso, por lo que el mismo desarrollador planteó una integración de esta herramienta al flujo editorial de OJS, que resolvería el problema del grado de dificultad que le implica a los usuarios sin conocimientos, y que no usan herramientas externas (ver tabla 1).

docxConverted: Como solución al problema que se planteó antes, es que se desarrolló esta extensión para OJS 3.1, que permite convertir artículos en formato DOCX a XML JATS. En resumen, es la misma herramienta pero integrada al flujo de trabajo editorial de OJS. Este plugin tiene como objetivo ayudar a los editores que se encuentran utilizando XML JATS como formato fundamental para su flujo de trabajo de publicación. La idea detrás del conversor, es la misma que la anteriormente planteada, crear una estructura básica a partir del documento dado, en DOCX (ver tabla 1).

DOCX2JATS: Proyecto Java, destinado a facilitar DOCX a la transformación XML JATS para artículos científicos. Este proyecto surgió como una versión anterior de las herramientas que se venían analizando y es una biblioteca de JAVA, implementada por el mismo desarrollador. Las principales dificultades que empiezan a surgir de este tipo de herramientas que, en un principio, no están integradas al flujo de trabajo de OJS. En este caso, es una herramienta poco accesible para un usuario que no tenga conocimiento de desarrollador, o de manejo de consolas, por lo que dificulta aún más la integración a un flujo de trabajo en funcionamiento (ver tabla 1).

Por cómo están planteadas estas herramientas, deben recibir un documento con formato DOCX que esté bien estructurado, es decir que cumpla con ciertos requerimientos obligatorios planteados previamente para que la conversión se realice correctamente y sea lo más precisa posible. Esto planteó la primera dificultad, ya que demandaría que los autores estructuren el documento con ciertas reglas a cumplir, lo que implicaría un cambio importante en las directrices para los autores de una revista. Otra alternativa es que la tarea de estructurar correctamente el documento a convertir sea delegada, parcial o totalmente, a los editores de la revista, lo que implica trabajo adicional. Si alguna de las reglas no se cumple, hay más probabilidades de que la herramienta no funcione correctamente. En algunos casos, podría llegar a hacerlo, pero esto presentaría problemas a posteriori, como, por ejemplo, que el plugin de edición de XML JATS en OJS (Texture) no sea capaz de editarlo correctamente.

Una vez que se tiene el documento correctamente estructurado, y donde se haya aplicado bien la herramienta, se presentan otras dificultades, ya que no ofrece un XML JATS completo, que sea una representación correcta y fidedigna del documento original, sino que es una representación parcial que en rangos generales cumple.

Al hablar de docxConverted, una de las limitaciones más importantes está relacionada con las referencias bibliográficas (ver tabla 2), ya que la herramienta solo permite la conversión de las referencias a XML JATS cuando son agregadas por Zotero.[9] Debido a que este gestor no presenta un uso generalizado, esto limita considerablemente la cantidad de documentos de los cuales se pueden obtener las referencias bibliográficas.

9 Gestor de referencias bibliográficas de uso libre.

En cuanto a DOCX2JATS, se analizó esta herramienta como una excepción, ya que no está integrada a OJS, pero es creada por el mismo que desarrolló docxConverted. Además, se destacaron algunas funcionalidades (que la anterior no tiene), entre ellas, permite realizar una conversión de las referencias bibliográficas (ver tabla 2) del documento docx sin necesidad de utilizar Zotero, u otras herramientas externas; el problema es que sólo acepta un único tipo de estándar, AMA (Vancouver citations style), que para el contexto de las revistas de ciencias sociales, y en particular de la UNLP, este tipo de referencias no suele utilizarse. Por lo que, para la implementación de esta solución implicaría un cambio en la directrices y en el estilo de cada revista, es decir, deberían modificar el formato de referencias, sólo por una cuestión de limitaciones técnicas. Por lo tanto no resulta conveniente su implementación, pero lo que sí se puede obtener de esta es analizar cómo funciona la extracción de citas para poder implementarlas en la otra herramienta que resuelve la parte de integración en OJS, pero no la de referencias.

En este sentido, se podría plantear la creación de expresiones regulares necesarias para poder descomponer las partes que integran a una referencia bibliografía en estándar APA 7, la utilizada en la UNLP, o la que sea necesaria,e insertarla en el complemento docxConverted, que ya está integrado en OJS. Por lo que, esto podría ser solución para una de las principales dificultades que presenta el plugin durante el análisis de su implementación. En cuanto al análisis de imágenes por las distintas pruebas que se han realizado podemos determinar que la herramienta no logra extraer correctamente las imágenes.

Estas herramientas ofrecen la posibilidad de trabajar, entre otras cosas, con tablas e imágenes (ver tabla 2), ya que permiten extraer las distintas imágenes del documento original e insertarlas en el XML JATS. Ahora tras evaluar ambas herramientas podemos decir que, en general, docxConverted resuelve el problema y, además, lo hace dentro del flujo editorial de OJS consiguiendo extraer las imágenes, insertarlas dentro del OJS y referenciarlas desde el XML JATS. En cambio, DOXC2JATS presenta dificultades para la extracción de imágenes ya que arroja algunos errores.

En el uso de las herramientas se encontraron otras dificultades, como la extracción de metadatos (ver tabla 2), que al no poder ser extraídos en su totalidad deben ser cargados en forma manual.

Eje 2: Edición de XML JATS

Una vez realizada la conversión del documento con formato DOCX al XML JATS, se deben realizar nuevas validaciones para poder corroborar la correctitud de la conversión. Se tendrá que validar que la información contenida en el XML JATS representa el documento original, que no haya errores o que se reconozca la información faltante. Trabajar sobre un XML JATS directamente es muy tedioso, ya que, al menos, requiere conocer cómo se estructura un XML y, en profundidad, cuál es el estándar de JATS. En esta etapa, se buscarán herramientas que permitan trabajar sobre los XML JATS, sin requerir tener un conocimiento técnico de estos conceptos.

Tal como se mencionó anteriormente, esta etapa busca dar por finalizados los problemas que se generan durante la salida de la etapa anterior, es decir de la conversión, por lo tanto, está enfocada en la intervención humana sobre el resultado obtenido. Preferentemente, siempre es necesario hacer validaciones manuales, principalmente, cuando se están probando herramientas con muchas falencias y/o problemas a resolver. Además, puede haber casos en los que se necesite agregar al XML JATS información que no se encuentre en el documento original.

Dado que el objetivo de este trabajo es analizar y buscar herramientas que permitan editar XML JATS sin tener conocimientos técnicos, lo que se busca es poder insertarlo dentro de un flujo editorial en funcionamiento, y para esto se debe atacar, al mismo tiempo, el problema de la integración a OJS. La herramienta que cumple estos requerimientos es Texture[10] (ver tabla 1).

Texture-plugin de edición de XML JATS: Texture es un desarrollo de PKP, que tiene el objetivo de integrar una herramienta de edición directa de documentos XML JATS al flujo de trabajo de OJS. Funciona como un editor de texto, similar a un word. Resulta amigable para el usuario y permite trabajar directamente sobre el XML JATS, pudiendo visualizar la información contenida en este. El usuario no necesita saber sobre el lenguaje XML para utilizar Texture. La herramienta ofrece distintas funcionalidades que facilitan la tarea de editar el documento, a grandes rasgos, cómo agregar, eliminar o modificar títulos y subtítulos de secciones; trabajar sobre las imágenes pudiendo agregar, eliminar o modificar cualquier imagen, además para cada imagen permite modificar sus metadatos como títulos, subtítulos y descripción. Además, ofrece la posibilidad de editar las referencias bibliográficas, permitiendo su modificación o eliminación; a través de una interfaz amigable, el usuario puede agregar toda la información correspondiente de la referencia.

A pesar de las distintas limitaciones que tiene esta herramienta, hoy es la que mejor resuelve el problema de la edición y de la integración en OJS (ver tabla 2).

Uno de los principales problemas que presenta la herramienta es el soporte técnico para otros idiomas, tanto en la interfaz del usuario, como en la documentación, ya que todo se encuentra en inglés y se dificulta, en gran medida, su implementación en lo inmediato al flujo editorial de cualquier revista de habla hispana. Otro de los problemas que presenta la herramienta se desprende de los resultados obtenidos en la etapa anterior, ya que debido a las falencias que puedan generarse en la conversión es probable que Texture no funcione correctamente; por ejemplo, que la herramienta no reconozca la estructura del XML JATS, haciendo imposible abrirlo, o que al abrirlo haya errores que se puedan visualizar desde la interfaz gráfica del usuario.

Eje 3: Presentación de XML JATS

En este eje se trabaja sobre el XML JATS bien estructurado. El objetivo es aprovechar el XML para generar los múltiples formatos de salida, como HMTL, PDF, ePub, etc. Además, se evaluaron herramientas de visualización para poder ver los distintos formatos generados.

Las herramientas se dividirán en: las que permiten la generación de distintos formatos a partir del JATS XML y las que dejan visualizarlos. Estas son (ver tabla 1): *JATSParserPlugin*,[11] un plugin para OJS que permite obtener el HTML o PDF a partir de un XML JATS y visualizar su contenido; *lensGalley*,[12] un plugin que ofrece una interfaz para la visualización de XML JATS, con una interface amigable y con la posibilidad de ver por separado el artículo y de los metadatos; *PDFs PDF.JS*,[13] para poder visualizar los PDF.

JATSParserPlugin: La función que ofrece este plugin es dar la opción, en la etapa de producción, de seleccionar un XML JATS cargado en el flujo de trabajo, a partir del que se genera un PDF usando una plantilla de estilo por defecto y queda cargado en la galerada. Además, una vez seleccionado el XML JATS con

10 https://github.com/pkp/texture
11 https://github.com/Vitaliy-1/JATSParserPlugin
12 https://github.com/asmecher/lensGalley/
13 https://github.com/pkp/pdfJsViewer

el que se desea trabajar, se visualiza automáticamente en la pantalla de inicio del artículo en cuestión el contenido del JATS, lo que le permite al lector, por un lado, ver el contenido del artículo y, por otro, extraer del XML JATS las referencias bibliográficas e insertarlas en la sección de metadatos.

La principal limitación que presenta esta herramienta se relaciona con el estilo del PDF, ya que solo ofrece una plantilla de estilo por defecto (ver tabla 2). Para las revistas, esto supone un limitante importante dado que muchas tienen sus propias maquetas diseñadas, lo que les permite mantener una identidad. Por lo tanto, implementar esta herramienta pone en riesgo la identidad de las revistas.

A continuación, se evaluarán las herramientas que se centran en la visualización de documentos, por ejemplo, *lensGalley* y *PDFs PDF.JS*.

leansGalley: Es una extensión que busca integrar *eLife Lens* en el flujo editorial de OJS. Este último proporciona una forma novedosa de ver el contenido en la web. Está diseñado para hacer más fácil la vida a investigadores, revisores, autores y lectores. La razón de esto es que la mayoría de los artículos de investigación en línea se publican en una versión digital fija del artículo original. Con *eLife Lens*, puede aprovecharse al máximo la naturaleza dinámica de HTML, combinado con javascript (Grubisic et al., s.f.). Esta es la herramienta que se utilizará para visualizar el XML JATS, independientemente de cómo se generó.

PDF.JS PDF ViewerPDF.js: Es un "lector" que permite visualizar documentos PDF directamente en un navegador web, sin tener que descargarlos ni usar programas adicionales. Es como abrir un libro digital en el navegador: se le puede hacer una lectura diagonal, buscar palabras y ver imágenes. Esto hace que sea más conveniente para visualizar y compartir documentos PDF en línea, sin necesidad de software especializado.

Tabla 1. ¿En qué eje funciona cada herramienta?

Eje 1: Conversión a XML JATS	Eje 2: Edición de XML JATS	Eje 3: Presentación
DocxToJats	Texture Plugin	eLife Lens
DOCX 2 JATS		JATSParserPlugin
docxConvertd		PDF.JS PDF Viewer

Fuente: elaboración propia.

Tabla 2. Limitaciones y soluciones de las herramientas relevadas

Herramienta	Eje	Solución	Problema
docxConverter	Conversión	Integración de docxToJats al flujo editorial de OJS	Solo funciona con referencias bibliográficas en Zotero. Consta de distintas limitaciones para extracción de metadatos.
DOCX2JATS	Conversión	Conversión a JATS XML. Extracción de referencias bibliográficas	No está integrado a OJS. No se puede implementar dentro del flujo editorial.

Texture Plugin	Edición	Edición de XML JATS	Posee un único formato de referencias. Solo tiene soporte para un idioma: inlges
JATSParserPlugin	Presentación	Genera PDF y HTML como formatos de salida a partir de un XML JATS.	Solo posee una única plantilla de estilos para el PDF.
eLife Lens Article Viewer	Presentación	permite visualizar el XML JATS. Separando contenido de metadatos.	
PDF.JS PDF Viewer	Presentación	permite visualizar el PDF, sin necesidad de degargarlo	

Fuente: elaboración propia.

Conclusiones

A modo de corolario, se puede decir que a partir de la evaluación exhaustiva de las herramientas mencionadas disponibles para la marcación y la generación de XML JATS en OJS plantea posibilidades y limitaciones para su implementación en el flujo editorial de revistas científicas en OJS. Si bien estas herramientas ofrecen soluciones para la conversión, la edición y la presentación de documentos científicos, se han identificado varios obstáculos que deben superarse para su adopción generalizada.

Uno de los desafíos principales reside en la necesidad de garantizar la interoperabilidad y la calidad de los resultados obtenidos. Las herramientas de conversión, si bien pueden automatizar ciertos procesos, requieren de documentos bien estructurados y del cumplimiento de ciertas normas, lo que puede implicar cambios en las prácticas de los autores y editores. Además, las limitaciones en la extracción de metadatos y de referencias bibliográficas representan una barrera significativa para la adopción de estas herramientas, en tanto que "las citas son un componente fundamental de la publicación científica, ya que vinculan los resultados de la investigación a lo largo del tiempo" (Nicholson et al., 2021).

En cuanto a la edición de XML JATS, por un lado, el plugin *Texture* ofrece una solución accesible para los editores, pero presenta desafíos en cuanto al soporte multilingüe y a la dependencia de la calidad del XML generado previamente, entre otras limitaciones. Por otro lado, las herramientas de presentación ofrecen opciones para visualizar y para compartir el contenido de manera eficiente, pero carecen de flexibilidad en la personalización de los formatos de salida (Cuculovic et al., 2022).

En consecuencia, se hace evidente la necesidad de continuar desarrollando y mejorando estas herramientas para lograr un mejor ajuste a las necesidades y a las prácticas editoriales.

Referencias

Becerril-García, A., Aguado López, E. y Macedo García, A. (2023). Marcalyc: software para la marcación XML JATS para las revistas científicas de acceso abierto diamante. *Palabra Clave, 12*(2), e179. https://doi.org/10.24215/18539912e179

Becerril-García, A. y Aguado-López, E. (2018). The end of a Centralized Open Access Project and the beginning of a community-based sustainable infrastructure for Latin America: Redalyc.org after fifteen years the open access ecosystem in Latin America. *ELPUB 2018*, Junio. https://dx.doi.org/10.4000/proceedings.elpub.2018.27

Böschen, I. (2021). Software review: The JATSdecoder package—extract metadata, abstract and sectioned text from NISO-JATS coded XML documents; Insights to PubMed central's open access database. *Scientometrics, 126*, 9585–9601. https://doi.org/10.1007/s11192-021-04162-z

Cho, Y. (2022). Open-source code to convert Journal Article Tag Suite Extensible Markup Language (JATS XML) to various viewers and other XML types for scholarly journal publishing. *Science Editing, 9*. https://doi.org/10.6087/kcse.284

Cuculovic, M., Fondement, F., Devanne, M., Weber, J. y Hassenforder, M. (2022). A JATS XML comparison algorithm for scientific literature. *Journal Article Tag Suite Conference (JATS-Con) Proceedings 2022-NCBI Bookshelf*. https://www.ncbi.nlm.nih.gov/books/NBK579687/

De Giusti, M. R. (2016). *Las dificultades de la preservación digital: problemas, desafíos y propuestas para los repositorios* [presentación en congreso]. Conferencia Internacional BIREDIAL-ISTEC (San Luis Potosí, México, 17 al 19 de octubre de 2016). http://sedici.unlp.edu.ar/handle/10915/56288

García, D. (2018). *Revistas científicas electrónicas sobre comunicación* [Tesis de grado, Universidad Nacional de La Plata]. http://sedici.unlp.edu.ar/handle/10915/70486

Grubisic, I., Aufreiter, M., Buchtala, O., Nott, G., Close, R., Korosec, S., Hamilton, I. y Mulvany, I. (s.f.). *eLife Lens: A novel way of seeing content*. Elifesciences. Recuperado el 20 de abril de 2024 de https://lens.elifesciences.org/

McGlone, J. (2013). Preserving and publishing digital content using XML workflows. En A. P. Brown (Ed.), *The Library Publishing Toolkit* (pp. 97-108). IDS Project Press. http://hdl.handle.net/2027.42/99563

National Information Standards Organization. (2019). *ANSI/NISO Z39.96-2019, JATS: Journal Article Tag Suite, version 1.2*. https://www.niso.org/standards/z3996-2019-jats

Nicholson, J. M., Mordaunt, M., López, P., Uppala, A., Rosati, D., Rodrigues, N.P., Grabitz, P. y Rife, S. C. (2021). scite: un índice de cita inteligente que muestra el contexto de las citas y clasifica su intención usando el aprendizaje profundo. *Quantitative Science Studies, 2*(3), 882-88. https://doi.org/10.1162/qss-a-00146

Redalyc. (2016a). ¿Qué es XML? [Entrada de Blog]. Recuperado de https://xmljatsredalyc.org/2016/07/29/que-es-xml/

Redalyc. (2016b). ¿Qué es JATS? [Entrada de Blog]. Recuperado de https://xmljatsredalyc.org/xml-jats-en-redalyc/

Rozemblum, V. (2021). *Propuesta de implementación de marcado XML-JATS para revistas científicas sostenidas por la Universidad Nacional de La Plata* [Tesis de grado, Universidad Nacional de La Plata]. http://sedici.unlp.edu.ar/handle/10915/115124

Ruiz, A., Correa, L., Bárcena, L. y Cristina, L. (2022). Preparación y envío de artículos para marcado. Dirección de Visibilización de la Producción Científica y Académica. http://sedici.unlp.edu.ar/handle/10915/139737

Soler S., Villarreal, G. L. y García, D. (2023). *Marcación y generación de XML JATS en OJS*. Proyecto de Enlace de Bibliotecas, Servicio de Difusión de la Creación Intelectual. http://sedici.unlp.edu.ar/handle/10915/159355

Santiago Soler es estudiante avanzado de la Licenciatura en Informática de la UNLP. Es desarrollador en PREBI-SEDICI, con enfoque en Open Journal Systems (OJS), y participa en proyectos de personalización y creación de plugins para optimizar la gestión de revistas científicas. Además, es investigador en nuevas tecnologías para mejorar los flujos de trabajo editorial. Está abocado a la promoción del avance científico a través del desarrollo de software de código abierto. ORCID: https://orcid.org/0009-0003-0594-9581

Dolores García es Licenciada en Comunicación Social, egresada de la Facultad de Periodismo y Comunicación Social de la UNLP en 2018. Fue becaria y pasante de la Comisión de Investigaciones Científicas de la provincia de Buenos Aires. Desde 2016 desarrolla tareas de gestión, técnicas y capacitaciones a equipos editoriales y a comités en el Portal de Revistas de la UNLP y en el Portal de Congresos de la UNLP. Se desempeña como asistente técnica en la Revista Argentina de Antropología Biológica y en la Asociación Argentina de Antropología Biológica. Forma parte del equipo de RevPsi como asistente editorial. En abril de 2023, asumió el cargo de Profesional Adjunto en CESGI, otorgado por la Comisión de Investigaciones Científicas de la provincia de Buenos Aires. ORCID: https://orcid.org/0000-0002-6686-3138

Gonzalo Luján Villarreal es Doctor en Ciencias Informáticas, forma parte de PREBI-SEDICI desde el año 2004 y es coordinador del Portal de Revistas (2008), del Portal de Congresos (2009), del Proyecto de Visibilidad Web Institucional (2012) y del Portal de Libros (2015). Es también director del Centro de Servicios en Gestión de Información (CESGI, 2016) de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, coordinador informático de revistas científicas de la Universidad Nacional de La Plata y profesor de la Facultad de Informática de la misma universidad. ORCID: https://orcid.org/0000-0002-3602-8211

Adela Ruiz es Licenciada en Comunicación Social por la Universidad Nacional de La Plata (UNLP) y Diplomada en Políticas Editoriales y Proyecto Cultural por la Universidad de Buenos Aires (UBA). Desde 2021, se desempeña como Coordinadora General de Revistas Científicas de la UNLP y, desde 2014, como Directora de Publicaciones Científicas de la Facultad de Periodismo y Comunicación Social (FPyCS) y como asesora en Sistemas de Evaluación de Revistas de la Facultad de Artes (FA). Es profesora titular del Taller de Edición Técnica y docente de la Especialización en Edición de la FPyCS. ORCID: https://orcid.org/0000-0002-2873-006X

Consideraciones y buenas prácticas en la aplicación de Inteligencia artificial en revistas diamante: caso de la revista Tecnología en marcha

Alexa Ramírez-Vega[1]

Palabras claves

Inteligencia artificial; revistas diamante; acceso abierto; ChatGPT; políticas editoriales.

Artificial intelligence; diamond journals; open access; ChatGPT; editorial policies.

Eje temático

Inteligencia artificial (IA) aplicada a la Ciencia Abierta

Resumen

El artículo se analiza la proliferación de la inteligencia artificial (IA) en el ámbito académico y sus implicaciones en los procesos de investigación, escritura y revisión de manuscritos académicos. Se destaca la popularización de herramientas como ChatGPT, que facilitan el trabajo de los autores, pero también plantean desafíos para los editores y revisores de revistas científicas. Por su parte, el Instituto Tecnológico de Costa Rica (TEC) implementó estrategias para regular el uso de IA en la revista *Tecnología en marcha* con el objetivo de asegurar la transparencia, calidad y ética en el proceso editorial. Se revisaron y ajustaron las políticas internas, instrucciones para autores y métodos de comunicación con las partes involucradas. El principal aporte está en la importancia de establecer lineamientos claros sobre el uso de IA para mantener la integridad y calidad de la publicación académica. La transparencia y responsabilidad ética en el uso de IA fortalecen la confianza en la revista y su reputación. Además, adaptarse a los avances tecnológicos y mantener políticas actualizadas muestra un compromiso con la excelencia académica.

Introducción

La proliferación de la Inteligencia Artificial (IA) en los últimos años ha cambiado la forma en cómo hacer las cosas en diversos ámbitos. De esto no se escapa el ámbito académico, y mayor aún con la popularización de ChatGPT y otras herramientas de IA especializadas en áreas científico-académicas.

En este mismo sentido, si bien la IA puede ayudar en muchos procesos involucrados en la investigación y fases de la escritura de artículos, ensayos, tesis y cualquier documento académico, esto implica un gran desafío para los involucrados en los procesos revisión y edición de estos manuscritos. Esto repercute directamente en los editores y encargados de revistas científicas, los cuales deben tomar acciones concretas para regular el uso de IA en el proceso de elaboración de los artículos que se reciben. Así mismo, (Mollaki, 2024),. destaca la necesidad de transparencia en el uso de IA, la importancia de la supervisión humana en la toma de decisiones editoriales, y la responsabilidad de los autores y editores en garantizar que los resultados generados por la IA sean precisos y confiables.

1 alramirez@itcr.ac.cr, Instituto Tecnológico de Costa Rica

De igual forma, las labores de los autores, editores de textos científicos y sus revisores se han visto influenciadas por el uso de IA en los procesos de revisión por pares y algunos procesos editoriales, lo cual conlleva a desafíos éticos y cómo esto puede afectar la imparcialidad y objetividad del proceso. Esto ha llevado a la creación de una serie de manuales y recomendaciones sobre el uso ético transparente de IA en el flujo editorial (López-Martín, 2023).

En este sentido, el papel de la IA en el proceso de revisión por pares en la publicación académica permite agilizar la revisión, mejora la calidad de las evaluaciones y la identificación de posibles conflictos de interés entre revisores y autores. Sin embargo, también surgen desafíos asociados, como la necesidad de supervisión humana para garantizar la imparcialidad y precisión de los resultados generados por la IA, así como los problemas éticos que pueden surgir. Por lo tanto, es necesario regular su uso de manera ética y responsable en el proceso de revisión por pares, destacando la importancia de mantener altos estándares de calidad y transparencia (Arthur Tang, 2023).

Por su parte, los autores de los artículos científicos son los más influenciados, actualmente, ante la proliferación de IA en las etapas de escritura de un manuscrito, ya que además de *ChatGPT*, existen diversas herramientas que ayudan y facilitan la elaboración de un documento científico. En este sentido los autores tienen acceso a herramientas como *Semantic Scholar* para búsqueda de fuentes bibliográficas; *Grammarly* para corrección gramatical y ortográfica; también *Scite* que ayuda a los autores a encontrar citas relevantes para sus artículos y verificar el contexto de esas citas; entre otras.

Además, durante el año 2023 se recibieron diversas consultas por parte de los autores y revisores sobre los lineamientos de la revista sobre el uso de IA en los procesos de elaboración y revisión de los artículos. En concreto se consultó sobre la generación de imágenes de apoyo para el artículo y como citar las fuentes, también sobre el uso de herramientas de IA en el procesamiento de datos y la traducción de textos.

Debido a loa anterior, se indagó sobre la adopción de políticas de uso de IA en otras revistas científicas, en este sentido como se menciona en (Ulloa Valenzuela, 2023), la revista *Science* se pronunció al respecto, indicando que no admite el uso de IA para la generación de artículos, por su parte la revista *Nature* ha adoptado dentro de sus políticas editoriales la regulación de uso de herramientas de IA en la elaboración de artículos, donde se indica que no se permite designar autoría a una herramienta de IA, porque esto implica una responsabilidad que solo personas reales pueden asumir; ahora bien, se indica además, que si se usan herramientas en alguna de las etapas de la elaboración del artículo, éstas deben indicarse en las secciones de metodología o agradecimientos. De igual forma, el editor de la revista Colombiana de Obstetricia y Ginecología indica que para regular el uso de herramientas de IA en la escritura de los artículos sometidos a la revista se basan en las iniciativas del Comité Internacional de editores de revistas médicas (ICMJE por sus siglas en inglés) (Gaitán-Duarte, 2023). Dichas recomendaciones exponen que los autores deben indicar si utilizaron tecnologías de IA y explicar cómo se usaron en la carta de presentación y en la sección correspondiente del artículo. Los chatbots, como ChatGPT, no deben ser listados como autores ya que no pueden asumir la responsabilidad por la exactitud y originalidad del documento. Los autores son responsables de revisar y editar cuidadosamente el contenido generado por cualquier herramienta de IA, asegurando que no haya plagio y que se realicen las atribuciones adecuadas de todo el material citado (International Committee of Medical Journal Editors, 2024).

De esta manera, el Instituto Tecnológico de Costa Rica (TEC), como parte de los procesos de mejora continua de las revistas científicas, inició con la implementación de diversas estrategias sobre el uso regulado de IA en las diferentes etapas editoriales de la revista *Tecnología en marcha*, con el objetivo de atender las consultas de los autores y revisores, así como estar alineados con las tendencias internacionales, sin dejar de lado la calidad, ética y transparencia de la revista.

Metodología

Dadas a las consultas recibidas por autores y revisores de la revista, y como parte de la mejora continua en los procesos editoriales de la revista, surge la necesidad de adoptar una política para la regulación del uso de IA en el proceso editorial de los artículos sometidos a la revista Tecnología en marcha.

En este sentido, el objetivo fue identificar los procesos en los cuales se puede involucrar el uso de IA por parte de los diferentes actores del flujo editorial (autores, revisores y editores) y con base en esto, indagar sobre guías, políticas, regulaciones y lineamientos implementados por otras revistas y entidades relacionadas. Entre las cuales se destacan las siguientes:

- Recomendación sobre la ética de la inteligencia artificial de la UNESCO (UNESCO, 2022).
- Declaración de los editores sobre el uso responsable de tecnologías de IA generativa en la publicación de revistas académicas (Kaebnick, y otros, 2023).
- Guía para la regulación del uso de Inteligencia Artificial en revistas científico-académicas de ciencias sociales, en procesos de arbitraje y para reportar su uso en textos científicos. (Penabad-Camacho, Morera-Castro, & Penabad-Camacho, 2024)
- Recomendaciones de la Asociación Mundial de editores médicos (WAME, por sus siglas en inglés), (International Committee of Medical Journal Editors (ICMJE), 2023)

La aplicación de las regulaciones sobre IA se basó en las guías y recomendaciones anteriores, permitiendo realizar cambios pertinentes en políticas, metodologías y lineamientos de la revista *Tecnología en marcha*, para regular su uso y garantizar la transparencia y ética en todos los procesos de edición.

Resultados

La aplicación de diversas estrategias y mejoras en la regulación del uso de IA en la revista *Tecnología en marcha* se llevaron a cabo en tres etapas:

1. *Actualización de reglamento interno (editores y revisores)*

La primera etapa para regulación de uso de IA en la revista Tecnología en marcha fue a través del reglamento interno, donde se normó y definió por parte del Consejo editorial de la revista la manera en que las herramientas de IA podrían apoyar algunas funciones editoriales.

Dentro del reglamento se incluyen los siguientes enunciados:

- El uso de IA no está prohibido en ninguna de las etapas del proceso editorial de la revista, pero debe seguir lo indicado en las políticas de ética y buenas prácticas editoriales.

- Los editores pueden hacer uso de herramientas de IA como apoyo en la verificación de requisitos, traducciones, búsqueda de referencias o información sobre los autores o el artículo.
- Todo uso de IA debe estar debidamente justificado, mediado y revisado por seres humanos.
- La escogencia, asignación de revisores, revisión de evaluaciones y análisis debe ser realizada por el editor a cargo.
- Se realiza control de similitud y control de generación de textos por medio de herramientas de IA. Los informes ahí mostrados no serán concluyentes y deben ser analizados detalladamente antes de emitir un criterio al respecto.
- Es deber del editor declarar e informar a los involucrados (autores, revisores) el uso de herramientas de IA en cualquier parte del proceso.

En el siguiente enlace se puede consultar la política de IA de la revista:

https://revistas.tec.ac.cr/index.php/tec_marcha/libraryFiles/downloadPublic/115

Por su parte, los expertos evaluadores de la revista también se les solicita la declaración de uso de alguna herramienta IA en la revisión de los textos. Como se muestra en la figura 1, esto se incluye en el formulario de revisión y como un requisito para completar la evaluación del artículo.

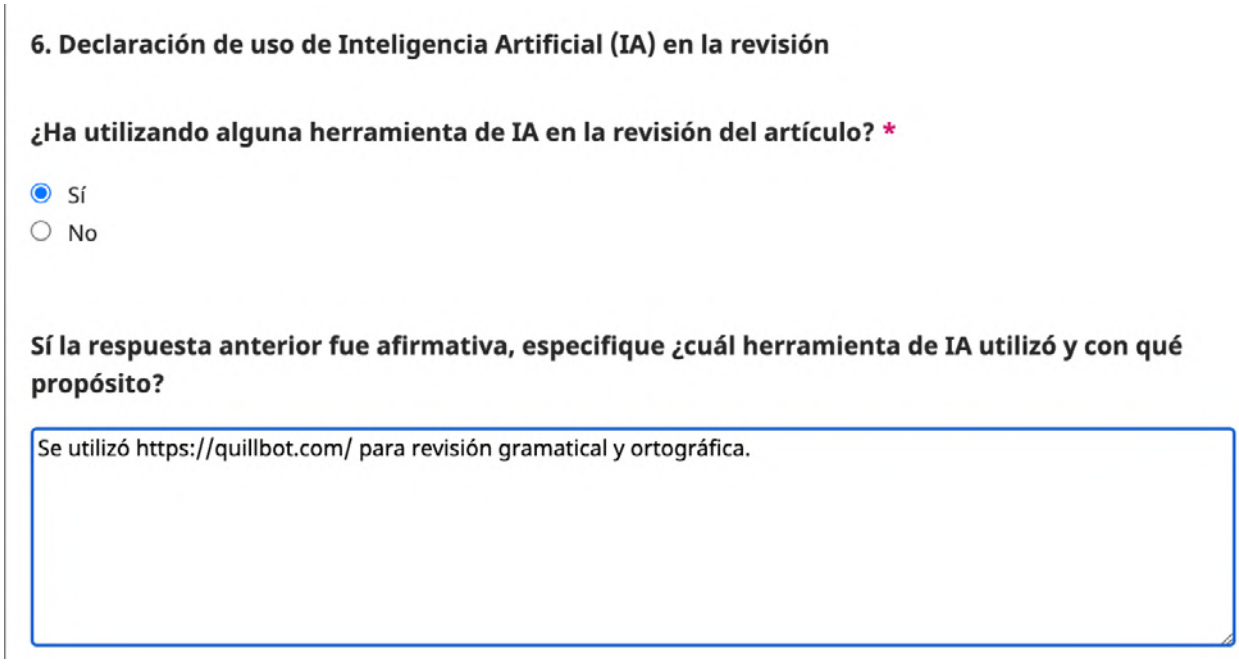


Figura 1. Declaración de uso de IA en formulario de revisión de experto.

2. *Actualización de instrucciones para autores*

Dentro de los lineamientos de la revista *Tecnología en marcha* se solicita a los autores la declaración de uso de IA, en cualquier parte del proceso de elaboración del artículo científico que están sometiendo a la revista. O en su defecto, indicar que no se hizo uso de herramientas de IA en ninguna de las etapas. Este apartado se incluye en la plantilla de artículos al final del documento, junto con la declaración de distribución de autores y de disponibilidad de datos (ver figura 2).

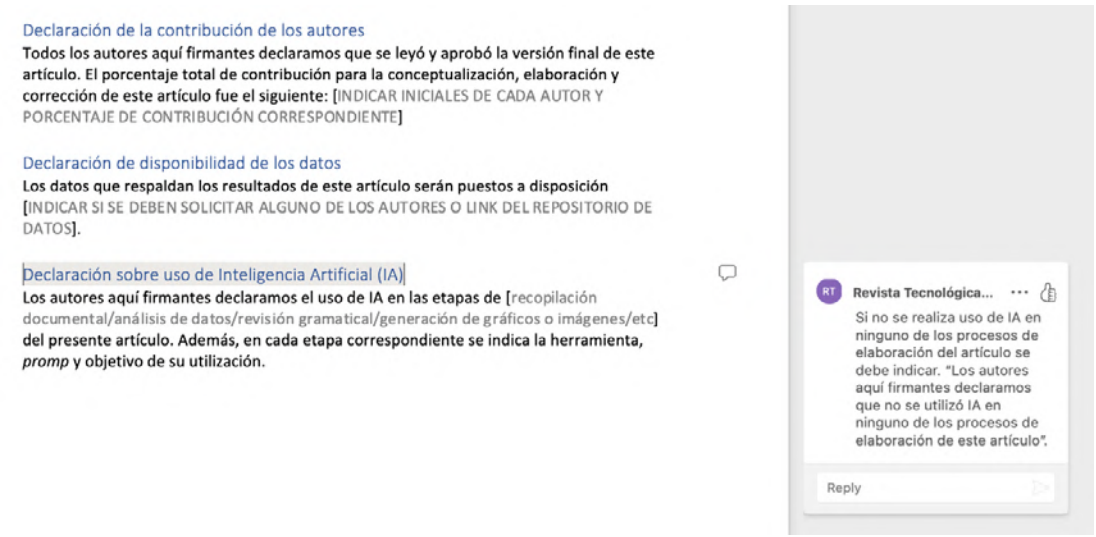


Figura 2. Declaración sobre uso de IA en plantilla de formato de artículos.

Además, en la página de la revista se incluye en las instrucciones para publicar se tiene un enlace a las políticas de IA de la revista, donde se muestran ejemplos (ver cuadro 1) de cómo declarar el uso de herramientas de IA según sea el caso: imágenes, gráficos, recopilación de fuentes bibliográficas y otros.

Cuadro 1. Ejemplos de declaración de uso de IA.

Tipo de uso	**Texto sugerido**
No uso de IA	Los autores declaramos que no se utilizó IA para la conceptualización o redacción de este artículo.
Redacción	Los autores declaramos que hemos utilizado una herramienta de inteligencia artificial [NOMBRE DE LA HERRAMIENTA O SITIO WEB] para asistirnos en la redacción de este artículo. Esta herramienta nos ayudó a mejorar la estructura y la claridad del texto. Los contenidos generados por la IA fueron revisados minuciosamente por nosotros para asegurar su precisión y coherencia con el objetivo del estudio.

Análisis de datos	En este estudio, empleamos un algoritmo de aprendizaje automático [NOMBRE DE LA HERRAMIENTA O SITIO WEB] para analizar los datos recopilados. Esta herramienta nos permitió identificar patrones y tendencias que de otro modo podrían haber pasado desapercibidos. Nos aseguramos de validar los resultados obtenidos con otros métodos de análisis para evitar sesgos.
Traducción	Utilizamos la herramienta de inteligencia artificial [NOMBRE DE LA HERRAMIENTA O SITIO WEB] para traducir partes de este artículo del inglés al español. La herramienta nos ayudó a agilizar el proceso de traducción, pero realizamos una revisión exhaustiva para asegurar la calidad y precisión de las traducciones.
Revisión gramatical	Para la revisión gramatical y ortográfica de este artículo, empleamos la herramienta de IA [NOMBRE DE LA HERRAMIENTA O SITIO WEB]. Esta nos permitió identificar errores y mejorar la fluidez del texto. No obstante, realizamos una revisión final para garantizar que el artículo cumpliera con los estándares de calidad de la revista.
Generación de imágenes	En este estudio, generamos las imágenes [INDICAR CUALES] utilizando una herramienta de IA [NOMBRE DE LA HERRAMIENTA O SITIO WEB]. Las imágenes se emplearon para ilustrar ciertos conceptos del estudio. Hemos verificado que las imágenes sean precisas y representativas de los datos y teorías discutidos en el artículo.

Comunicación a los involucrados

La tercera etapa consistió en comunicar a las partes involucradas y por todos los medios posibles los cambios sobre uso de IA en la revista:

- Publicación en sitio web de la revista.
- Publicación en redes sociales de la revista y entidad editora.
- Comunicación por correo electrónico a lista de distribución de autores y lectores de la revista.
- Comunicación por correo electrónico institucional a interesados.

Conclusiones

La implementación de lineamientos claros sobre el uso de herramientas de IA en la revista *Tecnología en marcha* ha demostrado ser fundamentales para garantizar la integridad, transparencia y calidad del proceso editorial. Estos lineamientos equilibran la necesidad de innovación tecnológica con los principios éticos de la publicación académica, permitiendo que los autores hagan uso responsable de las herramientas de IA en la elaboración de sus artículos.

Además, el establecimiento de regulaciones específicas para la aplicación de IA, han permitido responder a las necesidades expuestas por parte de autores, revisores y equipo editorial, así como a directrices institucionales que recomiendan la regulación de estas herramientas. De esta manera, al proporcionar directrices precisas y exigir la declaración explícita del uso de IA, se fomenta la transparencia en todas las etapas del proceso editorial, lo que a su vez contribuye a la credibilidad y reputación de la publicación.

El esfuerzo continuo para adaptar la revista a los avances tecnológicos y mantener políticas actualizadas muestra un compromiso con la excelencia académica y la evolución de las normas éticas en el ámbito científico. A medida que las tecnologías de IA continúan desarrollándose, es crucial que las revistas diamante se mantengan a la vanguardia, adoptando enfoques proactivos para equilibrar la innovación con la responsabilidad ética. Esto no solo asegura la calidad de las publicaciones, sino que también facilita la colaboración y el avance en el campo científico-académico.

Bibliografía

Mollaki, V. (2024). Death of a reviewer or death of peer review integrity? the challenges of using AI tools in peer reviewing and the need to go beyond publishing policies. *Research Ethics, 20*(2), 239-250.

López-Martín, E. (2023). The role of generative artificial intelligence in scientific publishing. *Educación XX1, 27*(1), 9-15.

Arthur Tang, K.-K. L. (2023). The importance of transparency: Declaring the use of generative artificial intelligence (AI) in academic writing. *Journal of Nursing Scholarship, 56*(2), 314-318.

UNESCO. (2022). *Recomendación sobre la ética de la inteligencia artificial.* UNESCO.

Kaebnick, G. E., Magnus, D. C., Kao, A., Hosseini, M., Resnik, D., Dubljević, V., . . . Cherry, M. J. (2023). Editors' Statement on the Responsible Use of Generative AI Technologies in Scholarly Journal Publishing. *Hastings Center Report, 53*(5), 3-6.

Penabad-Camacho, L., Morera-Castro, M., & Penabad-Camacho, M. A. (2024). *Guía para la regulación del uso de Inteligencia Artificial en revistas científicoacadémicas de ciencias sociales, en procesos de arbitraje y para reportar su uso en textos científicos.*

Ulloa Valenzuela, G. (2023). El desafío del uso de inteligencia artificial para la elaboración de la literatura ciéntífica: el caso de ChatGPT, un debate abierto. *Cuadernos Médico Sociales, 63*(1), 27–31.

International Committee of Medical Journal Editors (ICMJE). (2023, February 20). *ICMJE Authorship Guidelines and Acknowledging Non-author Contributions*. Retrieved from American Medical Writers Association: https://blog.amwa.org/icmje-authorship-guidelines-and-acknowledging-non-author-contributions

Gaitán-Duarte, H. (2023). The use of artificial intelligence and scientific papers published in the Colombian Journal of Obstetrics and Gynecology. *Colombian Journal of Obstetrics and Gynecology, 74*(3), 199-201.

International Committee of Medical Journal Editors. (2024). *Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals*. Retrieved from https://www.icmje.org/icmje-recommendations.pdf

Alexa Ramírez-Vega

Bachiller en ingeniería en computación y licenciada en enseñanza de la matemática del Instituto Tecnológico de Costa Rica (TEC). Además, es máster en inteligencia artificial avanzada de la UNED de España. Actualmente se desempeña como editora técnica de revistas científicas del TEC, donde tiene a cargo la revista Tecnología en marcha y el Portal de revistas de dicha institución. Es miembro de la subcomisión de Ciencia Abierta de CONARE. Además, ha impartido múltiples charlas y cursos sobre edición de revistas, escritura de artículos científicos, acceso abierto y ciencia abierta, tanto a nivel nacional como internacional.

Datos abiertos

Análise das propostas de certificação de repositórios ao Core Trust Seal: o que podemos aprender com elas?

Samile Andrea de Souza Vanz[1], Rene Faustino Gabriel Junior[2], Marcel Garcia de Souza[3], Washington Segundo[4], Caterina Groposo Pavão[5]

Palabras claves

Certificação, Repositório confiável, Core Trust Seal.

Certification, Trustworthy data repositories, Core Trust Seal.

Eje temático

Datos abiertos

Resumen

Observa-se o crescimento do movimento da ciência aberta e maior adesão da comunidade científica à prática de compartilhamento e reuso de dados de pesquisa. Paulatinamente, a infraestrutura para possibilitar estas práticas vem sendo desenvolvida no Brasil, e a partir de seu estabelecimento, surge a necessidade de aprimorar sua qualidade. Neste sentido, inúmeras instituições têm discutido as características e os requisitos de um repositório confiável. Este estudo tem como objetivo analisar as respostas de três repositórios certificados em 2024 pelo Core Trust Seal, com o intuito de explorar possibilidades de atendimento aos 16 requisitos da instituição certificadora. Observou-se que a possibilidade de respostas é bastante ampla e respeita as características do repositório em si, da área de pesquisa e da tipologia do dado.

Introdução

A maior adesão da comunidade científica à ciência aberta é paulatina à disponibilidade de repositórios confiáveis. Assim como o acesso aberto às revistas científicas despertou desconfiança e inúmeras críticas nos anos iniciais, a abertura dos dados de pesquisa tem gerado as mesmas incertezas. Soma-se a isso o desconhecimento acerca do significado do compartilhamento e reuso de dados, como parte do movimento da ciência aberta (Caregnato et al., 2019). Nesse contexto, a existência de repositórios certificados por instituições reconhecidas pode despertar maior confiança dos pesquisadores e fortalecer as iniciativas de compartilhamento.

1 UFRGS, samile.vanz@ufrgs.br
2 UFRGS, rene.gabriel@ufrgs.br
3 IBICT, marcelsouza@ibict.br
4 IBICT, washingtonsegundo@ibict.br
5 UFRGS, caterina@ufrgs.br

Um dos argumentos em prol do compartilhamento de dados recai sobre o custeio da pesquisa científica. No ecossistema da ciência, os financiadores desempenham papel fundamental ao fornecer recursos à pesquisa básica e aplicada. No Brasil, estudos relatam que quase 70% das publicações são financiadas por dez agências públicas, principalmente Capes, CNPq e FAPESP, e oitenta empresas (principalmente industriais e farmacêuticas) também foram identificadas por financiar pesquisas no Brasil, nenhuma delas nacional (McManus & Baeta Neves, 2021). Considerando que a ciência abrange grande parcela de investimento público, é mister garantir a perenidade e a otimização desse investimento. Nesse sentido, o compartilhamento possibilita o aproveitamento e utilização do dado de pesquisa por mais tempo, por diferentes cientistas e grupos de pesquisa, e permite que o mesmo seja analisado a partir da aplicação de diferentes métodos de pesquisa.

Do ponto de vista do pesquisador responsável pela coleta de dados, um repositório confiável garante que seus dados estão em local seguro, permanecendo acessíveis, utilizáveis e inteligíveis ao longo do tempo. Do ponto de vista do pesquisador que deseja utilizar novamente dados coletados em uma pesquisa anterior, o interesse é por dados de alta qualidade que vêm sendo preservados adequadamente (Rezende, 2021). A infraestrutura para arquivamento, preservação e disseminação dos dados de pesquisa vêm sendo desenvolvida há alguns anos, estruturada em repositórios criados em softwares próprios para esta finalidade (Rocha et al., 2021).

No contexto da América Latina, a criação de repositórios vem acontecendo no interior das instituições de pesquisa e universidades (Gabriel Junior et al., 2019; Silveira, Pavão & Vanz, 2023). No caso do Brasil, a exigência do depósito de dados e da apresentação de um Plano de Gestão de Dados começou a ser feita pelas agências de financiamento há alguns anos. A Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) estão entre as fundações que tornaram o planejamento do compartilhamento de dados mandatório. Esta demanda se tornou um estímulo ao desenvolvimento dos repositórios nacionais e, ultrapassado o momento inicial de desenvolvimento destes repositórios, torna-se premente buscar critérios de qualidade para o estabelecimento desta infraestrutura.

Dada a importância dos repositórios confiáveis, inúmeras instituições vêm desenvolvendo critérios e metodologias de certificação (Santos & Vanz, 2023). Especificamente, a certificação promete que o atendimento a esses critérios e padrões tornará as partes interessadas dos repositórios mais confiantes de que os dados que eles contêm serão protegidos, gerenciados adequadamente e estarão disponíveis para reutilização futura. Até o momento, o Brasil ainda não possui nenhum repositório de dados certificado. Nesse sentido, observar a experiência de repositórios estrangeiros pode ser considerada uma excelente oportunidade para apoiar a certificação dos repositórios nacionais. Este estudo tem como objetivo analisar as respostas de três repositórios certificados em 2024 pelo Core Trust Seal, com o intuito de explorar possibilidades de atendimento aos 16 requisitos da instituição certificadora.

Certificação de repositório digital confiável

Um repositório digital confiável é aquele capaz de manter a autenticidade, preservar e fornecer acesso a materiais digitais ao longo do tempo. Conforme o relatório **Trusted Digital Repositories: attributes and responsibilities** (RLG/OCLC, 2002), os repositórios digitais confiáveis devem assumir a responsabilidade pela manutenção dos documentos digitais em nome de seus depositantes. Para honrar este compro-

misso, é necessário que possuam uma estrutura organizacional que garanta não apenas a viabilidade do próprio repositório a longo prazo, mas também a preservação dos materiais digitais sob sua responsabilidade.

Cada vez mais as instituições e as bibliotecas sentem a necessidade de mostrar aos pesquisadores a qualidade dos repositórios nos quais eles depositam os seus dados de pesquisa. Nos Estados Unidos, agências federais como os National Institutes for Health (NIH) estão trabalhando em políticas e orientações para guiar a seleção de repositórios confiáveis para o depósito de dados de pesquisa quando estas são financiadas com recursos públicos (Key, Llebot & Boock, 2023).

O resultado da certificação, de acordo com Donaldson (2020), é que ela fornece validação externa de que os repositórios que afirmam preservar e fornecer acesso a recursos digitais, tanto agora como no futuro, estão realmente à altura do desafio. Ou seja, o repositório, foi avaliado por outra pessoa, para determinar se ele pode preservar, gerenciar e fornecer acesso a diversos tipos de materiais digitais em vários formatos, fazer a curadoria desses materiais para permitir a pesquisa, a descoberta e a reutilização e controlar para que o material digital seja autêntico, confiável, acessível e utilizável continuamente. Os méritos da certificação do repositório devem ser divulgados para financiadores, produtores de dados, depositantes e consumidores e os repositórios devem utilizar selos e marcas de certificação para comunicar a sua certificação e passar segurança para a comunidade designada.

Um repositório digital confiável é "[...] aquele cuja missão é fornecer acesso confiável e de longo prazo a recursos digitais gerenciados para sua comunidade designada, agora e no futuro" (RLG/OCLC, 2002). Há quase 30 anos, um grupo internacional de representantes de arquivos e bibliotecas nacionais, faculdades e universidades, agências de financiamento, indústria, editoras, e outras organizações governamentais e do setor privado articularam pela primeira vez pela necessidade de repositórios digitais confiáveis (Donaldson, 2020). Desde então, vários padrões para certificação de repositórios foram desenvolvidos em todo o mundo. O cenário global de certificação de repositórios está em constante mudança e adaptação, no Quadro 1 pode-se verificar um resumo das principais certificações para repositórios de dados.

Quadro 1 – Principais certificadores de repositórios

NOME	DESCRIÇÃO	ANO	CORRENTE
Open Archival Information System (OAIS). ISO 14721:2012	Criado pelo Consultative Committee for Space Data Systems (CCSDS). É um framework, não uma certificação. A maioria das certificações, atuais ou não, são baseadas nele.	2012; versão obsoleta de 2003	Sim
NESTOR, DIN 31644	Adaptado ao contexto alemão, publicado pelo Working Group Trusted Repositories – Certification. Consiste em um catálogo de critérios que são destinados, principalmente, para as organizações que têm o compromisso de preservar a memória, como, por exemplo, arquivos, bibliotecas e museus. Orienta a elaboração, o planejamento e a implementação de um repositório digital confiável a longo prazo.	2007	Sim

NOME	DESCRIÇÃO	ANO	CORRENTE
Trustworthy Repositories Audit and Certification (TRAC)	Criado pelo Center for Research Libraries (CRL), of Research Libraries Group (RLG) e o National Archives and Records Administration (NARA). Baseado em OAIS, consistia em produzir e esquematizar um processo genérico para auditar e certificar repositórios digitais. Substituído pela ISO 16363.	2007	Não
Data Seal of Approval (DSA)	Desenvolvido pela Dutch Data Archiving and Networked Services (DANS) para repositórios de dados. Fundido com a Certificação WDS para se tornar o Core Trust Seal.	2008	Não
Audit and Certification of Trustworthy Digital Repositories (ACTDR)	Baseada em OAIS. Desenvolvido pelo Consultative Committee for Space Data Systems (CCSDS). Destina-se a administradores de repositórios e certificadores. Avalia a infraestrutura organizacional, sustentabilidade financeira, gerenciamento dos objetos digitais e gestão de riscos dos repositórios. Tornou-se a norma ISO 16363:2012	2011	Sim
ISO 16363 - Audit and Certification of Trustworthy Digital Repositories	Define práticas recomendadas para avaliar a confiabilidade dos repositórios digitais. É aplicável a toda a gama de repositórios digitais, não apenas para repositórios de dados.	2012	Sim
WDS Certification	O World Data System (WDS) é um órgão afiliado do International Science Council (ISC) para repositórios de dados. Sua missão é melhorar as capacidades, o impacto e a sustentabilidade dos repositórios de dados e serviços de dados dos seus membros. Fundiu-se com o Data Seal of Approval (DAS) para tornar-se o Core Trust Seal.	2011	Não
Core Trust Seal	Mescla a certificação WDS e DAS. Baseia-se numa autoavaliação que é revista por pares e em evidências publicamente disponíveis. Para repositórios de dados em todas as disciplinas. Abrange aspectos importantes como governança, infraestrutura técnica, modos de acesso e licenciamento, gestão de metadados, preservação digital e sustentabilidade financeira.	2017	Sim

Fonte: Adaptado de (Key, Llebot & Boock, 2023; Santos, 2018).

Neste estudo aborda-se o Core Trust Seal, uma certificação baseada em autoavaliação e projetada especificamente para repositórios de dados de todas as áreas do conhecimento. O Core Trust Seal (CTS) é um certificado internacional que atesta a confiabilidade e a qualidade dos repositórios de dados de pesquisa. O CTS fornece critérios de avaliação para verificar se um repositório emprega boas práticas de gestão e preservação de dados, que abrangem aspectos como governança, infraestrutura técnica, modos de aces-

so e licenciamento, gestão de metadados, preservação digital e sustentabilidade financeira (Santos & Vanz, 2023). A terminologia utilizada pelo CRS é baseada no Modelo OAIS e os Princípios FAIR estão implícitos nos requisitos.

O certificado Core Trust Seal é concedido após uma avaliação criteriosa por meio do preenchimento de um formulário online específico onde o repositório deve provar que cumpre 16 requisitos. A equipe gestora do repositório deve indicar um nível de conformidade para cada um dos 16 requisitos: 0 – Não se aplica; 1 – O repositório ainda não considerou este requisito; 2 – O repositório possui um conceito teórico sobre este requisito; 3 – O repositório está em fase de implementação deste requisito; 4 – Este requisito foi totalmente implementado no repositório.

Ao longo do formulário de submissão, todo em língua inglesa, o responsável deve informar os níveis de conformidade, que são indicadores do processo auto avaliativo do repositório. A avaliação de conformidade declarada nas respostas é feita pelos avaliadores a partir das evidências comprobatórias apresentadas, como links de documentos formais, websites, entre outros.

Exige-se justificativa detalhada para resposta "não se aplica" conferida a algum requisito. A certificação pode ser concedida se alguns dos requisitos estiverem em fase de implementação (3) e o restante implementado (4). No entanto, níveis de conformidade na fase (1) ou (2) não são suficientes para a obtenção da certificação. Todos os 16 requisitos são mandatórios pois refletem as características desejáveis de um repositório confiável, por isso possuem pesos iguais no processo avaliativo. Algumas sobreposições e/ou informações duplicadas são inevitáveis.

A certificação é válida por 3 anos e desde fevereiro de 2024 a taxa para submissão é de EUR 3.000, para cobrir custos com avaliação por pares e a administração da organização sem fins lucrativos. A taxa administrativa dá direito a cinco revisões.

Procedimentos metodológicos

O Core Trust Seal disponibiliza um arquivo PDF contendo as respostas e evidências fornecidas pelos repositórios acreditados, bem como comentários dos avaliadores para cada item avaliado. Foram analisadas as respostas concedidas por três repositórios certificados em fevereiro de 2024, momento da coleta de dados:

1. GAMS (https://gams.uni-graz.at), repositório específico para Ciências Humanas coordenado pelo Centre for Information Modelling at the University of Graz (Alemanha); https://doi.org/10.34894/BYIJIP
2. Level-1 Atmosphere Archive & Distribution System (LAADS) Level-1 and Atmosphere Archive & Distribution System Distributed Active Archive Center - LAADS DAAC (nasa.gov), Distributed Active Archive Center (DAAC), vinculado à NASA; https://doi.org/10.34894/ZMSMZI
3. IPSL Computing and Data Center Institut Pierre-Simon Laplace – Sciences du climat (ipsl.fr), dedicado ao gerenciamento, coleta, distribuição e serviços de dados em ciências climáticas para observação e modelagem de dados, vinculado ao Institute Pierre Simon Laplace (IPSL); https://doi.org/10.34894/C4OTAF

O LAADS DAAC submeteu proposta conforme o CoreTrustSeal Requirements 2023-2025. O GAMS e o IPSL utilizaram a versão CoreTrustSeal Requirements 2020-2022. As versões dos requisitos apresentam pequenas diferenças que foram compatibilizadas, e apresentadas em um quadro de acordo com a versão mais atual, qual seja, CoreTrustSeal Requirements 2023-2025.

Além da análise das respostas, todos os documentos indicados como evidência comprobatória para os 16 requisitos foram acessados a partir do link fornecido no documento PDF. Uma síntese de seu conteúdo foi registrada em um quadro comparativo.

Resultados

A leitura dos formulários preenchidos pelos três repositórios revelou inúmeras possibilidades de preenchimento das informações para atender aos 16 requisitos do Core Trust Seal. As respostas foram sintetizadas e organizadas em um quadro, conforme os requisitos. Observou-se que as particularidades dos repositórios são respeitadas. Os três casos analisados apresentam características bastante distintas, tanto em relação à área de pesquisa quanto em relação ao tipo e formato de dados.

Quadro 2 – Resposta dos repositórios analisados e organizados conforme Requisitos CTS

	Requisito e Escopo	GAMS	LAADS DAAC	IPSL
	R0 Contexto do repositório Informar: tipo de repositório; breve descrição do repositório e comunidade; nível de curadoria;	Informaram o identificador Re3data, tipo de repositório, vínculo institucional do repositório e comunidade depositante, infraestrutura e equipe responsável.	Informaram o identificador Re3data, tipo de repositório, vínculo institucional do repositório, dados estatísticos como volume de arquivos depositados, usuários e visitas; instituições parceiras.	Informaram e apresentaram as diversas comunidades científicas que utilizam o repositório, e apresentaram a rede de parceiros.
Infraestrutura organizacional	R1 Missão/escopo Descrever a missão da organização em preservar e fornecer acesso aos dados.	Informaram a missão, o âmbito da coleção e a comunidade designada. Reconheceram a responsabilidade pela criação do repositório o que garante as perspectivas de continuidade e preservação da infra-estrutura.	Informaram a missão do repositório e encaminharam por meio de links para Requisitos para arquivamento, distribuição e serviços ao usuário no sistema de informações e dados, assim como para informações adicionais.	Informaram a missão e os vários campos relativos a aspectos do ciclo de vida dos dados, produtos de dados de diferentes interesses para a comunidade das ciências da terra e apresentaram links para diversos sites e plataformas.
	R2 Licenças Informar as licenças aplicáveis	Informaram que o Instituto promove o acesso aberto e a disponibilidade gratuita de dados de pesquisa; tipo de licença e a regulamentação sobre dados pessoais ou dados confidenciais, assim como os procedimentos em casos de violação de propriedade intelectual ou de direitos pessoais.	Informaram que o repositório está em conformidade com a Política de Dados e Informações de Ciências da Terra da NASA, com acesso gratuito e aberto às suas coleções de dados. Indicaram site da Política de dados, Direitos de dados e instruções de citação.	A maioria das informações dizem respeito a licenças, que os dados não estão sujeitos a qualquer regulamentação de direitos autorais, patentes, marcas registradas ou segredos comerciais. Informaram que não dispõem de uma Política de Dados geral, que é elaborado um contrato com cada fornecedor de dados.
	R3 Continuidade do serviço Ter um plano para garantir o acesso contínuo e a preservação de seus dados e metadados	Apresentaram a política institucional da Universidade de Graz que se responsabiliza pela manutenção do repositório, com Plano de preservação para 10 anos. Em caso de descontinuidade, está prevista devolução aos proprietários dos dados ou entrega a repositórios temáticos ou institucionais.	Apresentaram declaração da NASA, dizendo acreditar que a administração de longo prazo dos dados de sensoriamento remoto e de campanha de campo coletados pela NASA é essencial.	A ampla rede de parceiros garante a sucessão e disponibilidade dos dados. O texto informa as responsabilidades de cada parceiro e a correspondente previsão de anos de preservação.
	R4 Questões éticas e legais Informar as disposições éticas e de privacidade que afetam a criação, curadoria e uso dos dados	Indicaram que o depositante é responsável por seguir normas éticas e legais nacionais e internacionais e específicas de cada área; dados pessoais e sensíveis são de responsabilidade do depositante e podem ser excluídos do repositório a qualquer momento; requerem a assinatura do termo de depósito.	Os dados arquivados se referem a Terra e seu ambiente e, portanto, não são suscetíveis a risco, são disponíveis gratuitamente e o depositante assina o termo de depósito.	Os dados arquivados se referem a observações e simulações sobre clima, por isso não há preocupação com questões éticas. As questões que exigem atenção se referem a datasets protegidos por embargo e dados questionados pela comunidade científica. O repositório mantén uma equipe preparada para estes casos.
	R5 Governança e recursos Ter financiamento adequado e número suficiente de funcionários gerenciados por meio de um sistema claro de governança para realizar a missão com eficácia	Informaram o número de funcionários em tempo integral que cuidam da infra-estrutura, realizam tarefas técnicas, e curadoria de conteúdo; não há riscos de financiamento, mas a perda de membros-chave do pessoal é um risco, que é minimizado com documentação e transferência de conhecimento.	Informaram o gasto médio anual que permite cumprir a sua missão, incluindo apoio ao pessoal, recursos de TI, formação e viagens; número de funcionários; treinamentos anuais exigidos pela NASA para a equipe lidar com dados.	Detalharam a quantidade de profissionais, sua formação, suas atividades e regime de trabalho; orçamento anual e o valor utilizado para equipamentos; em termos de governança esclareceram a composição e o trabalho de comitês e comissões.
	R6 Orientação de especialista Adotar mecanismos para garantir conhecimento, orientação e feedback contínuos de especialista	Informaram como é realizada a comunicação interna, por meio de reuniões de pesquisa e grupos de trabalho. Os comentários e solicitações dos usuários são geralmente direcionados ao endereço de e-mail do departamento e distribuídos à equipe técnica.	O repositório mantém um Escritório de Atendimento ao Usuário para interagir com a comunidade. Informaram que participam anualmente de pesquisa que fornece um índice de satisfação do cliente. Esclareceram que participam de uma comunidade em rede que reúne profissionais de ciência, dados e tecnologia da informação de mais de 120 organizações, incluindo agências federais dos EUA, universidades e entidades comerciais.	Busca feedback da comunidade, em apresentações, reuniões e sessões práticas organizadas durante os treinamentos. Produzem tutoriais que incluem melhores práticas e uso passo a passo dos serviços e ferramentas para seus próprios projetos de pesquisa. Acrescentam que mantém a comunidade informada por meio de seu site. Possui um link "Contato" que permite aos usuários fazer perguntas ou fornecer comentários ao suporte ao usuário.

	Requisito e Escopo	GAMS	LAADS DAAC	IPSL
Gerenciamento de objetos digitais	R7 Proveniência e autenticidade garantir a integridade e autenticidade dos dados	Informaram que o Fedora suporta versionamento de todos os aspectos do recurso digital, porém às vezes falta transparência desses processos para o usuário final. Por esse motivo, até não conseguirem a implementação completa, o nível de conformidade do requisito foi definido como 3.	Informaram os procedimentos de Integridade do arquivo de dados, garantia de qualidade, validação do produto, gerenciamento de versões, documentação do produto e descoberta e acesso a dados.	Informaram que, devido à origem dos dados, não é necessário verificar sua autenticidade. Utilizam histórico de versões e acesso privilegiado. Acrescentam informações sobre rastreabilidade e identificadores persistentes.
	R8 Depósito e avaliação aceitar dados e metadados baseados em critérios previamente definidos garantindo relevância e inteligibilidade aos usuários	Informaram que na sua maioria os dados serão depositados como parte de um projeto de pesquisa, complementados por um plano de gestão de dados. Cada projeto é apoiado por um gestor de metadados que auxilia no fluxo de trabalho, modelagem de dados, processo de depósito e publicação. O controle de qualidade é apoiado pelo uso de vocabulários controlados e arquivos de autoridade. Informaram que utilizam o padrão de Dublin Core para recuperação de informações e identificação de recursos e operam com um conjunto limitado de formatos e tipos de dados para manter o repositório sustentável por longo período de tempo.	Informaram que os produtores são obrigados a enviar dados que estejam em conformidade com os padrões da comunidade no que se refere a formatos de dados, metadados e interfaces e estes são verificados para garantir que atendem padrões suficientes para auxiliar na sua interpretação, utilização e garantir que todos os requisitos de metadados sejam atendidos durante o ciclo de vida de dados e a preservação a longo prazo. Disponibilizam o link com uma lista de formatos de dados e metadados aceitáveis.	Esclareceram que os dados são validados pelas equipes científicas responsáveis pelo conjunto de dados antes da divulgação, verificando compressão dos dados através dos seus metadados e a conformidade com as sintaxes de dados que definem formatos de arquivos, nomes de arquivos, estruturas de diretórios, etc. Informaram sobre as diretrizes para auxiliar a formatar os dados e completar os metadados e que o depósito de dados em formato não preferencial ou não suportado é permitido apenas para dados de cauda longa, pois para permitir a descoberta dos dados é dada especial atenção ao formato dos dados.
	R9 Plano de preservação Assumir a responsabilidade pela preservação a longo prazo e gerência essa função de forma planejada e documentada	Apresentaram uma síntese dos aspectos principais e diversos links para documentos comprobatórios do plano de preservação.	Apresentaram uma síntese dos aspectos principais do plano de preservação e o link para o texto integral.	O ESPRI arquiva datasets de diferentes naturezas e provenientes de diversas fontes de referência, que requerem diferentes níveis de arquivo a longo prazo. Apresentaram link para todos esses documentos.
	R10 Qualidade dos dados possuir experiência apropriada para lidar com dados técnicos e qualidade de metadados e garantir que informações suficientes estejam disponíveis para os usuários finais fazerem avaliações relacionadas à qualidade	Detalharam procedimentos para garantir a qualidade dos dados e dos metadados, baseado principalmente no rigoroso trabalho da equipe com formação na área de Humanas e Ciência da Informação, o que possibilita um diálogo entre equipe do repositório e pesquisadores.	Informaram que o repositório utiliza mecanismos automatizados para controle da qualidade dos dados e metadados. Contato por telefone e e-mail é utilizado para dirimir dúvidas dos usuários.	Informaram que as ferramentas desenvolvidas permitem o controle dos metadados e garantem autodescrição dos metadados (cada variável no arquivo tem uma descrição associada, incluindo unidades físicas, localização espacial, proveniência, citação, etc.). O repositório delega ao depositante o controle da qualidade, e não se responsabiliza por controle adicional.
	R11 Workflows (Fluxos de trabalho) A gestão de objetos digitais ocorre de acordo com fluxos de trabalho definidos, desde o depósito até o acesso	Apresentaram o link para o workflow do repositório, e sintetizaram os procedimentos principais.	Mencionaram que os processos de ingesta, curadoria, exportação e arquivamento possuem workflows.	Detalharam os processos de ingesta, curadoria e disseminação dos dados.
	R12 Identificação e descoberta de dados Permitir a descoberta dos dados e consulta de forma persistente por meio de citações adequadas	Detalharam todas as possibilidades de pesquisa e acesso aos dados, indicando os respectivos sites. Detalhes sobre identificador persistente também foram fornecidos.	Detalharam todas as possibilidades de pesquisa e acesso aos dados, indicando os respectivos sites. Detalhes sobre identificador persistente também foram fornecidos.	Detalharam procedimentos de harvesting e catálogos interoperáveis; o repositório arquiva diversos tipos de dados que demandam ferramentas distintas para descoberta e acesso.

	Requisito e Escopo	GAMS	LAADS DAAC	IPSL
Tecnologia da Informação e Segurança	R13 Reuso dos dados Permitir a reutilização dos dados ao longo do tempo, garantindo que os metadados apropriados estejam disponíveis para apoiar a compreensão e o uso dos dados	Informaram que o primeiro ponto de acesso geralmente será a interface gráfica do usuário criada especificamente para o Repositório. Esclareceram sobre o padrão de metadados, os Elementos Semânticos e o Modelo de Dados da Europeana utilizados.	Esclareceram que fornecem metadados em nível de coleção e de arquivo para facilitar a pesquisa e descoberta de dados. Detalharam os formatos de dados e o fornecimento de links para documentos mais detalhados.	Esclareceram que um objetivo da descoberta de dados é permitir que o repositório seja coletado por outros catálogos de dados interoperáveis. Detalharam as pesquisas do catálogo para dados geoespaciais e forneceram uma lista de links para auxiliar no entendimento deste requisito.
Tecnologia da Informação e Segurança	R14 Armazenamento e integridade Permitir a reutilização dos dados ao longo do tempo, garantindo que os metadados apropriados estejam disponíveis para apoiar a compreensão e o uso dos dados	Esclareceram que dados e metadados serão migrados, caso haja necessidade no planejamento da preservação.	Informaram que existem vários documentos de Controle de Interface, Documentos de Requisitos e Acordos de Operações que regem a dinâmica do fluxo de dados e especificam requisitos para segurança da informação.	Esclareceram que aplicam os princípios FAIR na gestão dos seus dados e infraestruturas e, em particular, no princípio da reutilização de dados. Especificaram a utilização do formato formato netCDF.
	R15 Infraestrutura técnica O repositório é gerenciado em sistemas operacionais bem suportados e outros núcleos software e hardware de infraestrutura apropriados aos serviços que presta aos seus Comunidade Designada	Informaram os servidores existentes e localização dos mesmos, a lista de softwares e tecnologias envolvidas.	Informaram a infraestrutura técnica e localização dos servidores.	Forneceram todo o detalhamento dos equipamentos e da infraestrutura disponível, informando tamanhos dos servidores e das redes disponíveis.
	R16 Segurança O repositório protege a instalação e seus dados, metadados, produtos, serviços e usuários	Informaram volume de servidores e locais, volume de backups, quem faz a gestão do repositório, planos de substituição de pessoas ao longo dos anos.	Detalharam a segurança física e de software.	Informaram em detalhes todos os procedimentos técnicos para manutenção do repositório em segurança.

Fonte: dados da pesquisa.

Os Repositórios Digitais Confiáveis devem dispor de governança sustentável e estruturas organizacionais, infraestrutura confiável e políticas abrangentes de apoio às práticas acordadas pela comunidade; devem demonstrar capacidades essenciais e duradouras para permitir o acesso e reutilização de dados ao longo do tempo para as comunidades que atendem (Lin et al., 2020). Os 16 requisitos do Core Trust Seal estão organizados no sentido de avaliar as capacidades do repositório em relação às características de confiabilidade exigidas. A confiabilidade é demonstrada por meio de evidências que dependem de transparência, desta forma, os repositórios devem fornecer evidências transparentes, honestas e verificáveis de sua prática. No caso dos três repositorios analisados, as evidências consistiram no link para documentos importantes, como o Preservation Plan GAMSRepository (https://gams.uni-graz.at/o:gams.preservationplan); NASA Data and Information Policy (https://www.earthdata.nasa.gov/engage/open-data-services-and-software/data-and-information-policy); e fluxos de trabalho mapeados no ESPRI (https://cloud.ipsl.fr/index.php/s/ZiZWeeetRYw8CSG).

Observou-se que no caso do IPSL Computing and Data Center - ESPRI (Ensemble de Services Pour la Recherche à l'IPSL), considerando ser um repositório dedicado a dados observados e de modelagem nas ciências climáticas, a proposta envolve inúmeros parceiros institucionais, todos eles evidenciados através de documentos como convênios e contratos de parceria. É possível perceber que a equipe reuniu documentos que vão além daqueles relacionados de forma direta ao repositório, abrangendo a documentação de origem dos convênios e contratos institucionais de colaboração científica e parceria. Sem dúvida, ao evidenciar o compromisso institucional com os dados e o próprio repositório, a equipe avaliadora do Core Trust Seal pode confiar na garantía de integridade, autenticidade, precisão, confiabilidade e acessibilidade dos dados por longo prazo.

A questão da segurança é tema recorrente e perpassa vários dos 16 requisitos do Core Trust Seal (CORETRUSTSEAL, 2023). Para atender tais requisitos, o repositório deve analisar ameaças potenciais, avaliar riscos e criar um sistema de segurança consistente. Deve descrever cenários de danos com base em ações maliciosas, erro humano ou falha técnica que representam uma ameaça ao repositório e seus dados, produtos, serviços e usuários. A probabilidade e o impacto de tais cenários devem ser mensurados, para possibilitar a avaliação acerca de quais níveis de risco são aceitáveis e determinar quais medidas devem ser tomadas para combater as ameaças ao repositório e sua comunidade designada. Para este requisito deve ser descrito: o sistema de segurança de TI, funcionários com funções relacionadas à segurança e quaisquer ferramentas de análise de risco utilizados; os níveis de segurança que são exigidos e como eles são suportados; quaisquer procedimentos de autenticação e autorização empregados para gerenciar com segurança o acesso aos sistemas em uso.

Conclusão

O processo de certificação serve como catalisador para reunir grupos de pessoas em torno da biblioteca com interesses comuns. O grupo responsável pela preservação digital deve responder aos requisitos da estratégia de preservação, e grupos diversos de usuários dos repositórios devem ser consultados para tomar decisões relativas a serviços do repositório. Além disso, a equipe que trabalha diretamente na customização do repositório e a equipe que coleta, prepara a documentação deve contribuir para o compartilhamento de conhecimento permitindo solucionar inconsistências e problemas antes e depois da submissão da certificação.

Na certificação de um repositório destaca-se a importância da documentação disponibilizada de forma pública e aplicação das boas práticas, assim como a sua documentação. Além disso, manter uma equipe dedicada para os serviços do repositório torna-se essencial para satisfazer a maioria dos requisitos. O efeito combinado desses requisitos se traduz no aumento significativo na confiança e na qualidade dos repositórios.

No estudo desenvolvido por Donaldson (2020) foram levantados alguns questionamentos sobre a certificação de repositórios. A pesquisa mostrou que os repositórios podem preservar efetivamente a informação digital sem necessariamente seguir as melhores práticas e que aqueles que tentam seguir práticas recomendadas específicas muitas vezes têm dificuldade em fazê-lo. O autor coloca não saber se a auditoria e a certificação dos repositórios digitais confiáveis realmente são importantes, por exemplo, se os repositórios digitais são na verdade, melhores na preservação de informações digitais após a certificação do que eram antes e se os repositórios digitais confiáveis preservam melhor a informação digital do que os seus homólogos, embora os padrões dos repositórios digitais confiáveis promulguem esta suposição. Uma forma de avaliar se a auditoria e a certificação dos repositórios digitais confiáveis são importantes consistiria em examinar o seu impacto nas partes interessadas dos repositórios digitais confiáveis.

Porém, os instrumentos de certificação estão melhorando ao longo do tempo e estes, por sua vez, estão sendo utilizados para melhorar políticas, procedimentos, competências, ferramentas e serviços dos repositórios de dados para permitir a reutilização de dados de forma contínua. O processo de certificação permite refletir sobre as melhorias necessárias no repositório e nos serviços de dados de pesquisa, incluindo eficiência de fluxo de trabalho, melhorias nos recursos de software e na preservação dos conjuntos de dados. A partir da primeira certificação, a literatura consultada mostra que, as bibliotecas percebem a necessidade de elaborar novos manuais como por exemplo, sobre os princípios FAIR e as características desejáveis para os dados da pesquisa financiada pela instituição, visando aumentar a conformidade com os requisitos na renovação da sua certificação.

Este estudo ajudou a entender não apenas o processo de submissão da proposta de certificação de repositório digital confiável ao Core Trust Seal, mas principalmente, verificar o tipo de resposta que os avaliadores esperam dos gestores do repositório, o nível, a quantidade e a disponibilidade da documentação relativa ao repositório e suas boas práticas à qual os certificadores devem ter acesso para confirmar as respostas dos 16 requisitos. Verificou-se que em alguns casos as respostas eram bastante sucintas e foram consideradas satisfatórias mesmo não remetendo para links de documentação adicional. Também, foi possível avaliar o nível de conformidade e os comentários.

Bibliografía

Caregnato, S. E., Vanz, S. A. De S., Pavão, C. M. G., Passos, P. C. S. J., Borges, E. N., Gabriel Junior, R. F., Azambuja, L. A. B. & Rocha, R. P. (2019). Práticas e percepções dos pesquisadores brasileiros sobre serviços de acesso aberto a dados de pesquisa. LIINC EM REVISTA, 15, 121-141. https://doi.org/10.18617/liinc.v15i2.4771

CORETRUSTSEAL. Coretrustseal Trustworthy Data Repositories Requirements 2023-2025. 2023. https://zenodo.org/records/7051012

Donaldson, D. R. (2020). Certification information on trustworthy digital repository websites: A content analysis. PLoS ONE, 15(12): e0242525. https://doi.org/10.1371/journal. pone.0242525

Gabriel Junior, R. F., Rocha, R. P., Caregnato, S. E., Pavão, C. M. G., Passos, P. C. S. J., Borges, E. N., Vanz, S. A. de S. & Azambuja, L. A. B. (2019). Acesso aberto a dados de pesquisa no Brasil: mapeamento de repositórios, práticas e percepções dos pesquisadores e tecnologias. Ciência da Informação (Online), 48, 87-101. https://doi.org/10.18225/ci.inf.v48i3.4958

Key, C., Llebot, C. & Boock, M. (2023). Building a Trustworthy Data Repository: CoreTrustSeal Certification as a Lens for Service Improvements. Journal of eScience Librarianship, 12(3): e761. https://doi.org/10.7191/jeslib.761

Lin, D., Crabtree, J., Dillo, I. et al. The TRUST Principles for digital repositories. Sci Data 7, 144 (2020). https://doi.org/10.1038/s41597-020-0486-7

McManus, C. & Baeta Neves, A.A. (2021). Funding research in Brazil. Scientometrics 126, 801–823. https://doi.org/10.1007/s11192-020-03762-5

Rezende, L. V. R. (2021). Preservação e certificação de repositórios Dataverse. Revista Brasileira de Preservação Digital, 2(00): e021001. https://doi.org/10.20396/rebpred.v2i00.15810

RLG/OCLC Working Group on Digital Archive Attributes. (2002). Trusted digital repositories: Attributes and responsibilities. Mountain View, CA: Research Libraries Group. http://www.oclc.org/research/ activities/past/rlg/trustedrep/repositories.pdf

Rocha, R. P., Gabriel Junior, R. F., Vanz, S. A. de S., Borges, E. N., Azambuja, L. A. B., Caregnato, S. E., Pavão, C. M. G., Passos, P. C. S. J. & Felicissimo, C. H. (2021). Análise dos sistemas DSpace e Dataverse para repositórios de dados de pesquisa com acesso aberto. Revista Brasileira de Biblioteconomia e Documentação (Online), 17, 1-25. https://rbbd.febab.org.br/rbbd/article/view/1572

SANTOS, H. M. dos. (2019). Auditoria de repositórios arquivísticos digitais confiáveis. Informação em Pauta, 4(2), 156-172. https://doi.org/10.32810/2525-3468.ip.v4i2.2019.41787.156-172

Santos, D. B. & Vanz, S. A. de S. (2023). Repositórios de dados de pesquisa: confrontação dos princípios, critérios e requisitos internacionais de avaliação da confiabilidade. Revista Brasileira de Preservação Digital, 4: e023003. https://doi.org/10.20396/rebpred.v4i00.17355

SILVEIRA, J. I. & Vanz, S. A. de S. (2023). Diretrizes para políticas de depósito, acesso e uso de dados de pesquisa: proposta a partir da análise de repositórios de dados universitários internacionais. ATOZ: Novas Práticas em Informação e Conhecimento, 12. https://doi.org/10.5380/atoz.v12i0.87331

Silveira, Pavão e Vanz, 2023, ANAIS DO BIREDIAL ainda não publicado.

Vanz, S. A. de S., Pavão, C. M. G., Caregnato, S. E., Passos, P. C. S. J., Moura, A. M. M. de, Borges, E. N., Gabriel Junior, R. F. & Rocha, R. P. (2021). Diretrizes para o estabelecimento de um checklist para curadoria de dados de pesquisa. Informação em Pauta, 6, 1-18. https://doi.org/10.36517/2525-3468.ip.v6i00.2021.68088.1-18

Dra. Samile Andréa de Souza Vanz

Professora associada do Departamento de Ciências da Informação e do Programa de Pós-graduação em Comunicação da Universidade Federal do Rio Grande do Sul (PPGCOM UFRGS). Graduada em Biblioteconomia pela Universidade Federal do Rio Grande do Sul (1999), mestre e doutora em Comunicação e Informação pelo PPGCOM UFRGS (2004 e 2009), com estágio sanduíche na Dalian University of Technology (China, 2007-2008). Pós-doutorado pela Universidad Carlos III de Madrid (Madrid, 2016). Editora da revista Em Questão (2014 –). Desenvolve pesquisas na área de Comunicação Cientifica, com ênfase na produção de indicadores científicos, bibliometria, colaboração cientifica, análise de citação, análise de co-citação e rankings universitários.Tem experiência acadêmica e profissional na área de Planejamento, gestão e arquitetura de Bibliotecas e Unidades de Informação.

Dr. Rene Faustino Gabriel Junior

Graduado em Biblioteconomia e Documentação pela Pontifícia Universidade Católica do Paraná (2008), com mestrado em Ciência, Gestão e Tecnologia da Informação pela Universidade Federal do Paraná (2011) e doutorado em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho (2014). Atualmente é professor adjunto da Universidade Federal do Rio Grande do Sul e do Programa de Pós-Graduação em Ciência da Informação (PPGCIN) da mesma universidade e chefe do Departamento de Ciências da Informação (DCI). Tem experiência na área de Ciência da Informação, com ênfase em Biblioteconomia, atuando principalmente nos seguintes temas: Ciência da Informação, Estudos Métricos da Informação, Bibliometria, Brapci, Comunicação Científica, Dados de Pesquisa e Produção Científica. Implantou e coordena a Base de Dados de Periódicos em Ciência da Informação (BRAPCI). Membro do Grupo de Pesquisa de Comunicação Científica e do Núcleo de Estudos em Ciência, Inovação e Tecnologia da UFRGS.

Me. Marcel Garcia de Souza

Doutorando em Ciência da Informação pela Universidade de Brasília. Mestre em Educação em Ciências pela Universidade Federal do Rio Grande do Sul (2016). Graduado em Psicologia pela Universidade Católica de Brasília (2005). Servidor público federal; Analista em Ciência e Tecnologia no Instituto Brasileiro de Informação em Ciência e Tecnologia atuando como Coordenador de Tratamento, Análise e Disseminação da Informação Científica, além de coordenar pesquisas aplicadas voltadas à Ciência da Informação, Informação para Sustentabilidade, Avaliação do Ciclo de Vida, Informação Tecnológica.

Dr. Washington Luís Ribeiro de Carvalho Segundo

Doutor e Mestre em Informática pela Universidade de Brasília, com Estágio de Doutorado Sanduíche no Kings College London. Possui graduação em Matemática (Bacharelado e Licenciatura) também pela Universidade de Brasília. É Coordenador-geral de Informação Científica e Técnica no Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict / MCTI). É membro e coordena projetos, comitês nas áreas de Ciência Aberta e Ciência de Dados. É líder do Grupo de Pesquisa e Laboratório do Ecossistema da Pesquisa Científica Brasileira (LaEPeCBr) (http://dgp.cnpq.br/dgp/espelhogrupo/9750187028652303, https://pnipe.mctic.gov.br/laboratory/3911). Áreas de interesse em pesquisa: Métodos Formais, Repositórios Digitais Abertos, Repositórios de Dados Científicos, Interoperabilidade entre Sistemas de Informação Abertos, Ciência Aberta e Ciência de Dados.

Dra. Caterina Marta Groposo Pavão

Bacharel em Biblioteconomia pela Universidade Federal do Rio Grande do Sul, mestrado e doutorado em Comunicação e Informação pelo Programa de Pós-Graduação em Comunicação e Informação da mesma Universidade e doutorado sanduíche na Universidad Complutense de Madrid. É docente do Departamento de Ciência da Informação da Faculdade de Biblioteconomia e Comunicação da Universidade Federal do Rio Grande do Sul e do Programa de Pós-Graduação em Ciência da Informação da Faculdade de Biblioteconomia e Comunicação da mesma Universidade. Exerceu suas atividades profissionais de 1994-2018 no Centro de Processamento de Dados da Universidade Federal do Rio Grande do Sul, onde dedicou-se à gerência e administração do sistema de automação do SBUFRGS do Lume, repositório institucional da Universidade. Integra a equipe de pesquisadores do Grupo de Pesquisa Comunicação Científica da Faculdade de Biblioteconomia e Comunicação da UFRGS, do Grupo de Estudos e Práticas de Preservação Digital da Rede Cariniana e do grupo de pesquisadores da NUAWEB – Núcleo de Pesquisa em Arquivamento da Web e Preservação Digital. Participa de projetos do Centro de Documentação e Acervo Digital da Pesquisa (CEDAP/UFRGS) e do Centro de Processamento de Dados da UFRGS para implementação do Projeto Repositório de dados científicos da Universidade.

Dados de pesquisa: percepções e práticas de compartilhamento de cientistas da Pequena Ciência

Rosane Teles Lins Castilho[1]

Palavras chave

Dados de pesquisa; Compartilhamento de dados de pesquisa; Comportamento dos pesquisadores (dados de pesquisa); Comunidades da "Pequena Ciência".

Research data; Sharing research data; Researchers' behavior (research data); "Small Science" Communities.

Eixo temático

Dados abertos

Resumo

O compartilhamento de dados dos pesquisadores da "Pequena Ciência" é analisado no presente estudo e tem por objetivo apresentar um panorama dos resultados de pesquisas realizadas em instituições científicas, nacionais e internacionais, sobre o tema, visando identificar os motivos que concorrem para o não compartilhamento e a consequente invisibilidade dos dados. A metodologia adotada é de abordagem qualitativa e, visando identificar estudos que analisassem as práticas e percepções dos pesquisadores de ambas as comunidades sobre compartilhamento de dados de pesquisa, foi realizada pesquisa bibliográfica nas bases de dados brasileiras BRAPCI e BENANCIB e internacionais do Google Acadêmico e Scholar. Do conjunto de artigos recuperados das quatro bases de dados, foram selecionados trabalhos representativos do tema e considerados relevantes aos objetivos da pesquisa. Os resultados dos estudos analisados revelam a existência de barreiras de natureza institucionais, de infraestrutura e pessoais que impedem o compartilhamento, embora atitudes positivas em favor desse compartilhamento já sejam observadas. O estudo conclui que há semelhanças entre as práticas e percepções dos pesquisadores sobre a questão, entre as diversas comunidades estudadas.

Introdução

Os séculos XX e XXI, sobretudo a partir do pós-Segunda Guerra Mundial, assistiram a uma ruptura paradigmática na comunicação científica, com o advento e a popularização da Internet e das tecnologias digitais de informação e comunicação (Córdula & Araújo, 2019). O periódico científico impulsiona a disseminação da produção científica e desempenha um papel fundamental no meio acadêmico, promovendo avanços e destacando autores e editores. A partir da década de 1990, com a propagação da informação digital e o uso da Internet, essa ruptura influenciou o modo de editorar e de disseminar informações, principalmente da produção científica (Fachin, 2002). Com os avanços da computação e das tecnologias da informação os dados de pesquisa tornaram-se as principais fontes de pesquisa digitais e, como consequência

1 PPGCI IBICT/UFRJ, ORCID https://orcid.org/000-002-7142-6813 rosanetlcastilho@gmail.com

dessa nova ordem, os dados gerados, obtidos, coletados e utilizados na produção da ciência e de artigos científicos passaram a ser requisitados e disponibilizados em repositórios digitais visando, principalmente, ao compartilhamento (Sayão & Sales, 2015).

O compartilhamento de dados está na cerne do sucesso dessa nova ordem, pois a coleta, o armazenamento e a utilizaçao de dados só podem ocorrer depois que os meios de compartilhamento estiverem em vigor. Nas ciências servidas por disciplinas ou iniciativas de infraestrutura de âmbito nacional, o compartilhamento de dados de pesquisa é considerado ser uma tendência inevitável No entanto, ao contrário do que ocorre na "Gande Ciência", na "Pequena Ciência" não é comum ou esperado, funciona em grande parte como uma indústria caseira na qual os dados são trocados com base nas relações profissionais e na comunicação pessoal dos cientistas ou por meio dos *websites* dos projetos ou laboratórios, e não confere maior atenção à uniformização dos dados gerados. (Cragin et al., 2010).

Grande Ciência *(Big Science*) e Pequena Ciência (*Little Science)* são termos diferenciados por Derek de Solla Price (1963), em sua obra homônima. O autor refere-se à grande ciência como aquela realizada com equipamentos caros e complexos e grandes equipes de pesquisa, onde identificam-se os grandes projetos e programas concentrados em grandes laboratórios, como os de Astronomia e Física, que contam com grandes investimentos e uma complexa ciberinfraestrutura, mantêm um volume imenso de dados, oferecem facilidades de descoberta e acesso e são muito ágeis no atendimento às demandas de compartilhamento de dados (Sales & Sayão, 2019). A pequena ciência refere-se à pesquisa de baixo custo realizada por um indivíduo ou um grupo pequeno deles, em geral por um orientador e seus orientados, em universidades e institutos de pesquisa, onde bastam auxílios relativamente pequenos para agregar uma nova técnica de análise a um laboratório já consolidado numa instituição de pesquisa (Cragin et al., 2010).

O comportamento da grande e da pequena ciência é estudado por um modelo específico de distribuição estatística, denominado "a cauda longa", popularizado por Chris Anderson (2004), em que uma pequena parte da população tem muitas ocorrências – a chamada "cabeça"; enquanto outra parte dessa população tem poucas ocorrências isoladamente – a chamada "cauda longa" da distribuição. No contexto da geração de dados da pesquisa, é referida como a "cauda longa da ciência", assim, na "cabeça" identificam-se os grandes projetos e programas mais próximos dos padrões da grande ciência. Por sua vez, na "cauda longa", ou pequena ciência, identificam-se grandes quantidades de projetos liderados por cientistas individualmente ou por pequenas equipes, cujos projetos isoladamente geram poucos dados, mas, se considerados coletivamente, produzem um número gigantesco de coleções de dados, igualmente valiosos, que superam em muito as coleções geradas pela cabeça (Sales & Sayão, 2020).

O compartilhamento de dados pode acelerar a descoberta científica e, ao mesmo tempo, aumentar o retorno do investimento para além do pesquisador ou grupo que os produziu, porém, os avancos experimentados na comunicação científica, acesso e compartilhamento, propiciados pelas tecnologias digitais e ciência aberta, nem sempre estão acessiveis à pequena ciência, o que contribui para sua invisibilidade e, portanto, não trazem benefícios para a sociedade, nem para os pares, no cenário e dinâmica de uma comunidade científica (Sales & Sayão, 2020). Os repositórios de dados institucionais (RIs) têm permitido o compartilhamento e preservação de dados a longo prazo, mas pouco se sabe ainda sobre as percepções que os cientistas desse segmento têm sobre eles e as suas perspectivas sobre práticas de depósito, gestão e barreiras que impedem o compatilhamento de dados (Cragin et al., 2010).

O objetivo deste trabalho é apresentar um panorama dos resultados de estudos sobre o comportamento dos pesquisadores no contexto da pequena ciência, suas atitudes e percepções sobre dados de pesquisa, abrangendo as dificuldades e barreiras ao compartilhamento, bem como as motivações para adotar as suas práticas. Como contribuição, este trabalho espera trazer reflexões aos cientistas do segmento da pequena ciência e aos órgãos que a acolhem, constituindo os cenários onde ela se desenrola, de modo a fazerem um esforço conjunto visando ao desenvolvimento de uma ciência realmente aberta.

Metodologia

A metodologia adotada é de abordagem qualitativa e visou identificar estudos que analisassem as práticas e percepções de pesquisadores, integrantes de comunidades científicas, nacionais e internacionais, identificadas como integrantes do segmento da pequena ciência, sobre o tema compartilhamento de dados de pesquisa, de modo a obter um quadro atual do movimento, ainda que não exaustivo. Assim, no mês de outubro de 2023, foi realizada pesquisa bibliográfica nas bases de dados brasileiras BRAPCI e BENANCIB e internacionais do Google Acadêmico e Scholar, e feitas buscas nos metadados título e resumo com os termos "dados de pesquisa", "compartilhamento de dados de pesquisa", "pequena ciência", "universidades", "institutos de pesquisa", "invisibilidade de dados de pesquisa" e seus correspondentes no idioma inglês. Do conjunto de artigos recuperados das quatro bases de dados, foram selecionados os trabalhos representativos do tema e relevantes aos objetivos da pesquisa que são analisados nas próximas seções.

Referencial teórico

A pequena ciência tem sido tradicionalmente caracterizada como pesquisa baseada em hipóteses liderada por um único pesquisador principal, no qual o progresso e a recompensa são contingentes na geração e análise dos próprios dados. O financiamento da investigação pode ser limitado, e a condução diária da pesquisa depende muitas vezes de alguns alunos graduados que realizam grande parte da coleta de dados e gerenciam e processam esses dados durante o curso de um projeto. Nesse segmento da comunidade científica, os sistemas de gerenciamento de dados tendem a ser *ad hoc* e, se existirem padrões de dados, eles são raramente aplicados No entanto, estes arranjos nem sempre são estáticos, e segundo Craigin et al. (2010) a pequena ciência está se expandindo de duas maneiras marcantes: para alguns laboratórios, a configuração organizacional tradicional está ampliando as redes de pesquisa mais orientadas para a comunidade, a outra mudança é o surgimento de tecnologias intensivas na ciência de dados em algumas subdisciplinas que continuam a exibir a estrutura tradicional.

Outras características da pequena ciência residem na gestão diversificada dos dados, que é apoiada financeiramente pelas verbas concedidas pelas agências de fomento destinadas aos projetos e programas do pesquisador, as ferramentas mais comuns utilizadas são os softwares genéricos que já se encontram prontos no mercado, portanto, não fornecem as necessárias funcionalidades para um eficiente uso e sustentabilidade dos dados. Além dos projetos serem liderados por cientistas individualmente, podem sê-lo por pequenas equipes, em laboratórios independentes que desenvolvem um grande número de projetos científicos, em várias e diferentes universidades e institutos de pesquisa e em diversos domínios disciplinares, no entanto, são raramente arquivados para o compartilhamento e reuso (Sales & Sayão, 2019, Sales & Sayão, 2020).

A longo prazo, os cientístas da pequena ciência, abrangem muitos campos e produzem muitas formas diferentes de dados, variando em termos de formato de arquivo, padrões decorrentes da diversidade de instrumentos e das tecnologias adotadas, complexidade dos objetos digitais, podendo ter diferentes versões e variar com o tempo. Compreendem um número imenso de coleções, ou seja, mais dados até do que em grandes áreas científicas, como as da grande ciência. Grande parte das descobertas permanece nos computadores pessoais dos pesquisadores, não aparecem na literatura como artigos publicados em periódicos ou em outras mídias, o que concorre para sua inviisibilidade. Esses cientistas vem recorrendo cada vez mais às suas bibliotecas universitárias e repositórios institucionais para assistência com seus problemas de dados. Em resposta, os RIs de muitas bibliotecas acadêmicas vem fornecendo apoio à pesquisa primária, aos dados com arquitetura variada e implementações de modelos de serviço disciplinares, apoiados pelas equipes das bibliotecas que exercem também importante papel de curadoria dos dados (Craigin et al., 2010, Sayão & Sales 2019).

São em grande número as dificuldades e as barreiras alegadas pelos pesquisadores do segmento para o não compartilhamento dos dados de suas pesquisas e que contribuem para a sua invisibilidade. Sales e Sayão, (2020), distribuem-nas entre dois tipos, individuais e sistémicas, citando as seguintes:

a) dificuldade de acesso a infraestruturas tecnológicas e gerenciais e políticas institucionais que assegurem a estabilidade, persistência e interoperabilidade dos dados;
b) b) dificuldades relativas ao controle de qualidade e a padronização, uma vez que a natureza heterogênea e fragmentada dessas coleções exige estratégias diversificadas para a sua gestão;
c) c) dificuldades quanto à ausência de políticas voltadas para a publicação de dados, incluindo a inexistência de esquemas de reconhecimento da autoria e políticas de recompensa pela organização e disseminação dos dados;
d) d) falta de interesse dos pesquisadores em divulgar dados além dos limites profissionais mais próximos, dados sobre hipóteses não confirmadas e resultados negativos e dados considerados auxiliares de estudos publicados em artigos.

Para enfrentar essas dificuldades e transpor as barreiras do não compartilhamento, Sales e Sayão (2020) elencam algumas soluções:

a) desenvolver infraestruturas e práticas que tornem esses dados úteis para a sociedade através_da adoção de uma política científica nacional e seguida pelas instituições de pesquisa, difundindo os repositórios digitais disciplinares ou temáticos;
b) atender as necessidades de gestão, de curadoria e de controles personalizados que devem ser realizados, idealmente, em ambientes orientados por disciplina, apoiando o desenvolvimento de planos de gestão de dados de pesquisa;
c) apoiar a superação das barreiras relacionadas a percepção e motivações dos pesquisadores, tais como mecanismos de reconhecimento e recompensa, a disponibilidade de ferramentas apropriadas e de uma infraestrutura tecnológica que torne o compartilhamento possível,
d) apoiar a publicação dos dados de pesquisa como objeto de informação independente, em repositórios de dados ou centros de dados, a publicação de documentação textual em data jornal, na forma de data papers, com o propósito de descrever as coleções de dados por meio de links que passam a ter valor semântico, nas chamadas publicações ampliadas para contextualizar e relacionar todos os produtos de pesquisa em torno de um artigo;

e) apoiar a publicação de dados de pesquisa negativos em periódicos voltados para essa condição.

Para ilustrar o movimento em torno do compartilhamento de dados, resultados de estudos considerados representativos das dificuldades e barreiras, nos âmbitos nacional e internacional, são apresentados a seguir. Alguns desses estudos já apontam para uma maior motivação dos pesquisadores a efetuarem o compartilhamento dos seus dados de pesquisa.

Resultados de estudos sobre o compartilhamento de dados de pesquisa na pequena ciência

Os estudos realizados nos âmbitos nacional e internacional, a seguir, apresentam elementos considerados pelos cientistas como obstáculos para adotar atitudes de compartilhamento.

Estudos realizados na comunidade nacional

Veiga (2017) relata pesquisa realizada na Fundação Oswaldo Cruz (FioCruz), na área de Neurociência, para obter um diagnóstico da percepção dos pesquisadores sobre o compartilhamento de dados, quais barreiras têm afastado o pesquisador do compartilhamento e da abertura de dados de pesquisa, bem como os benefícios, ao compartilhar seus dados. Como fatores motivadores, defendem a abertura dos dados, pois contribui para a transparência na conduta da pesquisa, a otimização do avanço científico para a reprodutibilidade, o aumento da credibilidade da própria pesquisa, uma maior integridade e para a visibilidade, acessibilidade o compartilhamento das pesquisas, ao garantir o acesso e a sua estabilidade por longos períodos nos repositórios. Por outro lado, tendem a resistir ao compartilhamento como imposição de políticas mandatórias. Como principais barreiras ao compartilhamento e abertura de dados alguns citam não saber como fazê-lo ou desconhecer repositórios para fazer esse compartilhamento, outros temem pelos dados a serem divulgados conterem informação sigilosa, ou possam ser mal utilizados ou mal interpretados por outros pesquisadores; temem, ainda, a perda de oportunidades de publicação ao abrir os dados e preocupam-se com dados que possam gerar patente; têm receio de que as ideias de pesquisa sejam roubadas por outros pesquisadores e, na sua área, desconhecem quem faça esse compartilhamento, portanto não veem motivo para fazê-lo. Conclui a autora que o compartilhamento de dados de pesquisa é algo recente na comunicação cientifica contemporânea, portanto, o tema precisa ser mais bem estudado e envolver assuntos correlatos que analisados conjuntamente podem "construir caminhos" para apoiar o compartilhamento.

Veiga, Silva e Borges (2021), nessa mesma linha, analisaram custos, benefícios e os fatores contextuais para o compartilhamento de dados. Segundo as autoras, ao identificar as barreiras percebidas pelo pesquisador é possível tentar minimizá-las, conhecer os benefícios compreendidos por ele (pesquisador) e, assim, ser possível elaborar serviços e produtos que o estimulem ao compartilhamento. Por fim, ao discriminar os fatores contextuais é possível verificar quais elementos desses contextos são favoráveis e quais são desfavoráveis ao compartilhamento na percepção do pesquisador. Para tal, as autoras desenvolveram um modelo que foi aplicado na área de Neurociências em instituições de pesquisa, no Brasil e em Portugal, e que está sendo aplicado em outras instituições do campo da saúde, de modo a compreender a percepção dos pesquisadores quanto ao compartilhamento de seus dados e, assim, promover estímulos ao compartilhamento.

Utilizando a entrevista semiestruturada como técnica de coleta de dados aplicada a bolsistas de produtividade da Fundação Carlos Chagas, Santos (2022) relata os resultados da pesquisa sobre o que motiva ou inibe o compartilhamento de dados de pesquisadores no campo da Genética e, para tal, foram estabelecidos como objetivo Identificar as práticas de compartilhamento de dados de pesquisa nesse campo; é relatado que o compartilhamento acontece mais facilmente entre os pesquisadores mais jovens, se dá em subáreas mais genéricas, contrariamente ao que se dá em subáreas mais restritas, ou mais competitivas; como estímulo ao compartilhamento consideram a discussão dos resultados entre seus pares muito profícua para construir uma Ciência com mais qualidade, contribuir para seus avanços, bem como devolver à população os investimentos públicos recebidos, como forma de reconhecimento. Como fatores inibidores, citam o receio de revelar seus achados, diante da alta competitividade em nível internacional, do uso indevido dos dados por pesquisadores e laboratórios, das questões dos dados sensíveis e dos direitos de uso, como patentes e propriedade intelectual; em nível institucional há resistência quanto à obrigatoriedade ao compartilhamento principalmente entre os pesquisadores que não recebem incentivos da instituição como reconhecimento. Concluindo, a autora observa que trata-se de um tema recente, com poucas iniciativas, mas é necessário que se crie políticas nacionais e diretrizes institucionais para a adoção de práticas adequadas de produção, compartilhamento e reuso dos dados.

Estudos realizados na comunidade internacional

No âmbito internacional, seguem os resultados de pesquisas bem extensas: uma pesquisa usando grupos focais com cientistas de cinco disciplinas (ciências atmosféricas e da terra, ciência da computação, química, ecologia e neurociência) e de duas outras pesquisas, complementares, sobre práticas e compartilhamento de dados de cientistas em vários países, realizadas no intervalo de dez anos, onde observam-se mudanças ocorridas no comportamento de dados dos cientista no período.

Donaldson e Koepke (2020) conduziram uma pesquisa usando a metodologia de grupos focais com cientistas de cinco disciplinas (ciências atmosféricas e da terra, ciência da computação, química, ecologia e neurociência), sobre práticas de compatilhamento de dados. Os resultados revelam que muitos cientistas ainda não compartilham os dados de suas pesquisas, com exceção de pesquisas associadas a financiamento governamental, revelam também, que a maioria das disciplinas opera sem diretrizes estabelecidas de compartilhamento ou gerenciamento de dados, contam com soluções individuais ou institucionais para esse fim, por isso defendem a existência de serviços de repositório e de planos de gestão de dados (PDGs) para ajudá-los a implementar as partes de compartilhamento e preservação de seu dados. No entanto, os serviços de repositórios ainda levantam-lhes algumas incertezas quanto à sua adesão e elencam vários recursos desejados nos repositórios, incluindo: controle de qualidade dos metadados, rastreabilidade de dados, segurança, infraestrutura estável, restrições de uso de dados e a biblioteconomia de dados. Uma outra opção aos RIs consideradas são os sistemas proprietários de armazenamento em nuvem (por exemplo, DropBox, GitHub e GoogleDrive), vez que os usuários de dados são preocupados com as limitações de tamanho dos arquivos, custos, preservação a longo prazo, mineração de dados pelos provedores de serviços e o número de soluções de armazenamento, tornando-os onerosos.

No quesito ratreabilidade**,** desejam saber como o seus dados estão sendo usados e que fossem rastreados após o depósito em repositórios, quantos pesquisadores os visualizam, citam e publicam com base nos dados que depositam. Os cientistas reinvidicam também que os repositórios forneçam sistemas de notificação para os depositantes de dados e avisos quando novas versões ou trabalhos derivados ba-

seados em seus dados estiverem disponíveis, bem como notificações para os depositantes sobre quando seus dados forem visualizados, citados ou incluídos em uma publicação. Sobre metadados, os cientitas expressam o desejo de metadados de alta qualidade (ricos) nos repositórios, alguns defendem a criação automatizada de metadados ao carregar seus dados em repositórios para economizar tempo e fornecer pelo menos algum nível de descrição de seus dados e, ainda, metadados expandidos para dados dos sistemas de informações geográficas (GIS).

Quanto a restrições de uso de dados**,** participantes dos cinco grupos focais concordam que os repositórios precisam explicar claramente o que um pesquisador pode ou não fazer com um conjunto de dados, isto é, devem indicar claramente em cada conjunto de dados se os investigadores podem basear novas pesquisas nos dados, publicar com base neles e utilizar os dados para fins comerciais.

Quanto à demada sobre infraestrutura estável**,** os participantes descrevem essa infraestrutura em termos de atualização de arquivos de dados (ou seja, controle de versão) e formatos ao longo de longo prazo e garantia de sua usabilidade. Esse temor pela estabilidade leva os cientistas a procurar e utilizar soluções alternativas de armazenamento. Quanto à segurança, temem que a falta dela possa comprometer seus dados, especificamente permitir a sua exploração antes que os depositantes dos dados possam fazer uso deles através da publicação. Aqueles que lidam com dados com informações confidenciais, sensíveis ou de identificação pessoal expressam maior preocupação com possíveis violações de segurança, porque isso poderia resultar em perda de confiança dos participantes atuais e futuros do estudo, tornando mais difícil para eles próprios e para futuros pesquisadores recrutar participantes para o estudo no longo prazo.

A necessidade de treinamento em gerenciamento de dados é defendida pelos participantes, bem como a necessidade de informação sobre repositórios específicos de disciplinas que estão atualmente disponíveis, mas que encontram-se de posse apenas dos seus pares ou bibliotecários. A formação do próprios pesquisadores e de novos estudantes em ferramentas mais simples ou a realização de formação "fragmentada" em ferramentas avançadas de gestão de dados é tembem reivindicada considerando que a falta de ambas limitaam a produtividade do projecto.

O papel dos bibliotecários no gerenciamento de dados é também discutido, alguns grupos não crêem que os bibliotecários devam exercer um papel na gestão de seus dados por duas razões: primeiro, pensam que os seus dados são demasiado técnicos ou especializados para que os bibliotecários possam contribuir significativamente para a sua gestão; em segundo lugar, presumem que os bibliotecários são muto ocupados com suas próprias atividades e não teriam disponibilidade de tempo para tratar efetivamente de seus dados; em contraste, outros participantes defendem que os bibliotecários podem desempenhar um papel na gestão e compartilhamento de dados científicos, bem como fornecer assistência na publicação dos dados, são valiosos para pesquisas bibliográficas, assistência na pesquisa e busca de informações e junto a patentes, pesquisas de direitos de autor, gestão de mandatos de dados, aplicação de embargos, literaciamento da informação e padronização de metadados.

A pesquisa de Tenopir et al. (2011) relata estudo envolvendo 1329 cientistas de diversas áreas disciplinares, 75% dos quais na América do Norte e os 25% restantes na União Europeia, objetivando investigar as práticas dos pesquisadores quanto a tornar seus dados visíveis, bem como as barreiras e os facilitadores do compartilhamento de dados. Como principais descobertas, o estudo revela que: os cientistas não disponibilizam seus dados por exigir tempo e por não obterem retorno financeiro ou outro tipo de reconhecimento que incentive a prática; queixam-se das organizações não fornecerem suporte a seus pesquisadores para o gerenciamento de dados, tanto de curto quanto de longo prazo; concordam que estão dispostos a

compartilhar seus dados em publicações, se certas condições forem atendidas, por exemplo, se receberem citação formal e compartilhamento de reimpressões; a maioria está satisfeita com seus processos atuais de curto prazo do ciclo de vida da pesquisa, mas não com a preservação de dados de longo prazo. Na pesquisa também foram encontradas diferenças significativas de abordagens nas práticas de gerenciamento de dados, por exemplo, a não adoção de planos formais de gestão de dados e o tratamento da questão por parte das agências de financiamento, por área da disciplina.

Passados dez anos da primeira pesquisa, Tenopir et al. (2020) retornam ao tema para analisar as mudanças nas atitudes e práticas de compartilhamento de dados dos cientistas, concentrando-se em estágios específicos do ciclo de vida dos dados, como descrição, preservação e descoberta, e constatam uma sensível mudança positiva. Também examinam onde os cientistas armazenam dados, tanto a curto como a longo prazo, quais padrões de metadados os cientistas usam (se houver) para descrever os dados e quais barreiras enfrentam para encontrá-los e reutilizá-los, bem como os incentivos que os levam a compartilhar seus dados.

A pesquisa online, postada em links no Twitter, foi realizada entre 2017 e 2018, em duas etapas: na primeira contou com a participação de membros da American Society for Geophysics, da segunda participaram as comunidades de 18 disciplinas temáticas, de 116 países, compreendendo Europa e Rússia, Américas Central e Sul, Ásia, África e Oriente Médio, obtendo um total de 2184 respostas que foram processadas e analisadas pelo o software para análise de dados IBM SPSS 25. Os autores observam que ao longo da última década, nos Estados Unidos e em países da União Europeia, desde 2010, vêm sendo implementadas políticas abrangentes de dados abertos mais rigorosas pelos governos e agências de financiamento e vêm estabelecendo mandatos. Esses ógãos vêm passando a exigir um plano de gestão de dados, como parte do pedido de financiamento, e incentivam o compartilhamento de dados com base na declaração de que os dados nas suas diversas formas (desde dados brutos a publicações científicas) precisam ser armazenados, mantidos disponibilizados e acessíveis abertamente a todas as comunidades científicas (European Commission, 2019 como citado em Tenopir, 2020). Entre os achados, constataram que a maior parte dos dados abertos da Europa e Estados Unidos são do setor público, o que significa que são amplamente acessíveis e estão disponíveis para reutilização, por vezes sem condições restritivas. Dados aberto também são disponibilizados por fundações privadas e outras grandes organizações de financiamento que exigem o compartilhamento de dados por parte dos pesquisadores, bem como por muitas revistas de sociedades científicas, por exemplo, a American Geophysical Union, que exigem o depósito de todos os dados utilizados nos estudos relatados em suas publicações em repositórios públicos. Apesar de toda esse movivento de incentivo aos dados abertos, autores citam fatores individuais que afetam o compartilhamento de dados.

A pesquisa de Tenopir et al. (2011) revelou que a maioria dos cientistas estaria disposta a compartilhar seus dados "se certas condições fossem atendidas, tais como citação formal e compartilhamento de reimpressões". A pesquisa atual ampliada de Tenopir et al. (2020) demonstra "maior aceitação e disposição [dos cientistas] para se envolver no compartilhamento de dados, bem como observaram um aumento real nos comportamentos de compartilhamento" conforme relatado a seguir.

Quanto â reutilização dos dados**,** os setores governamental e comercial uma parcela expressiva dos entrevistados tem uma atitude positiva, usam regularmente dados coletados de terceiros em suas pesquisas e continuariam usando-os se pudessem facilmente ser acessados. De acordo com os entrevistados, detalhes escritos sobre métodos de coleta e garantia de qualidade, padrões de metadados explicitamente declarados, e informações detalhadas sobre a proveniência são os critérios mais importantes que influen-

ciam sua confiança no uso de dados coletados de outros. A ideia do compartilhamento de seus dados com um amplo grupo de pesquisadores é vista de forma positiva pelos os cientistas, porém estariam mais dispostos a fazê-lo se pudessem impor algumas condições de utilização àqueles que reutilizam os seus dados, por exemplo, citação, como requisito quase universal para receberem crédito de citação. A necessidade de publicar primeiramente em veículos formais da comunicação científica, em geral, mais lentos no seu processo de publicação, foi relatada como a principal barreira ao compartilhamento de dados, seguida pela falta resguardo de direitos aos seus dados, de tempo para preparar adequadamente os dados para compartilhamento e de restrições de financiamento que limitam a sua capacidade de preparar e depositar os dados. A grande maioria dos entrevistados considera que a falta de acesso aos dados gerados por outros investigadores ou instituições é um grande impedimento ao progresso da ciência, embora apenas cerca da metade deles pense que isso restringe a sua própria capacidade de responder a questões científicas.

Como conclusão, Tenopir et al. (2020) identificam entre os entrevistados um movimento geral de mudanças nas atitudes e comportamentos da comunidade científica global. Segundo os autores, o progresso na transição para a ciência aberta reflete-se na crescente aceitação dos conceitos cruciais para a ciência aberta de compartilhamento e reutilização de dados e de boas práticas de dados, incluindo a utilização e compreensão dos metadados. Da mesma forma, um número de entrevistados reconhece que várias barreiras que os impedem de partilhar os seus dados têm diminuído constantemente, ao longo do tempo decorrido entre as duas análises, pois há uma dinâmica positiva observada em termos de envolvimento das organizações em favor do compartilhamento, uso e reutilização de dados. Observam que o aumento da exigência de planos de gestão de dados por parte das organizações e maior oferta de treinamento e assistência em questões relacionadas ao gerenciamento de dados, são atitudes que têm um efeito positivo sobre as organizações que estão mais envolvidas com a gestão de dados.

Os respondentes que indicaram seu vínculo de trabalho estão assim distribuidos: vinculados ao setor acadêmico (72,8%), seguido pelos vinculados ao governo (16,6%), ao setor comercial (3,6%), a entidades sem fins lucrativos (4,3%) e a outros setores (2,7%). É importante reconhecer as diferenças no compartilhamento, uso e reutilização de dados entre respondentes vinculados aos quatro setores de trabalho. O setor governamental emergiu como um líder em atitudes positivas, como aceitação e disposição a acompartilhar e reutilizar dados, exibindo boas práticas e aumento o no envolvimento organizacional na gestão e treinamento em dados. De um modo geral, além do aumento observado um aumento no número de cientistas que têm atitudes positivas em relação compartilhamento e reutilização de dados, mais entrevistados compreendem e fazem uso de metadados, outro indicador do progresso que uma comunidade científica está fazendo em aceitar e seguir essas boas práticas de dados. Outro achado importante relaciona-se ao número de entrevistados que reconhecem que as várias barreiras que os impedem de compartilhar seus dados têm diminuído constantemente ao longo do tempo examinado.

Segundo os autores, os resultados desta pesquisa podem fornecer subsídios para a implementação dos princípios FAIR, pois dados abertos requerem mais atenção e esforço extras das partes interessadas, por exemplo, garantir as melhores práticas de dados através de planos de gestão de dados, no entanto, apenas cerca de metade dos entrevistados indicam que a sua principal agência de financiamento exige um plano de gestão de dados. Em síntese, observam que variações são constatadas sobre as práticas e atitudes de gerenciamento de dados dos cientistas, com base no setor de trabalho, na disciplina e, às vezes, na idade dos pesquisadores e, embora haja um progresso notável na transição para dados abertos e ciência aberta, ainda há uma discrepância entre atitudes positivas e a implementação real desses princípios pela

comunidade científica, para tal os autores consideram que a assistência de gestores de dados ou bibliotecários de dados, repositórios de dados prontamente disponíveis para armazenamento de longo e curto prazo e programas educacionais para gerar boas práticas de dados são necessários.

Discussão

Os resultados das pesquisas sobre práticas de compartilhamento e reutilização de dados, relatadas neste trabalho, demonstram que governos, a comunidade científica e a comunidade editorial, internacional, já vêem adotando estas práticas de uma forma cada vez mais favorável, e parece que os mandatos dos financiadores, em países estrangeiros, podem ser o motivador mais importante. Embora ainda existam barreiras significativas à adoção generalizada do compartilhamento de dados os resultados demonstram que os cientistas, de comunidades nacionais e internacionais, incluindo aqueles do segmento da pequena ciência, podem estar mais dispostos a compartilhar dados, sobretudo, se tiverem a garantia de receber uma citação formal pelo seu trabalho e se tiverem conhecimento e acesso à pesquisas que utilizam os seus dados, bem como reconhecimento por sua cooperação na empreitada de torná-los aptos para o compartilhamento. Alguns dos obstáculos citados que impedem a partilha de dados nesse segmento, em ambas as comunidades, incluem tempo insuficiente para executar os procedimentos de preparo dos dados para compartilhamento e divulgação, a necessidade de publicar primeiramente em veículos formais os resultados dos seus dados, para que sejam citados e interpretados corretamente e a falta de financiamento para dar suporte a todas ações que envolvem o compartilhamento, no entento, mesmo assim, os cientistas geralmente estão dispostos a compartilhar os seus dados, desde que os seus direitos como pesquisadores e autores sejam protegidos. A conveniência é frequentemente um fator nas decisões dos cientistas sobre onde e quando compartilhar os seus conjuntos de dados e é mais provável que os cientistas os compartilhem se os processos forem padronizados, simples e se lhes for prestada assistência, como nos RIs. Melhorar os padrões dos metadados, tornando-os mais ricos, pode aumentar o nível de confiança que os cientistas têm em outros membros da comunidade, bem como aumentar a capacidade de descoberta, acessibilidade e validade dos conjuntos de dados e garantir que possam reutilizar os dados sem descobrir posteriormente que eles contêm erros ou foram falsificados. Conclui-se que é necessário um esforço conjunto por parte das comunidades científicas, editorial e de elaboração de políticas para trabalharem em conjunto de modo a aumentar as práticas de compartilhamento e reutilização de dados da comunidade científica da pequena ciência para remover barreiras a um ambiente científico mais aberto.

Considerações finais

Os estudos aquí relatados são resultados de pesquisas que visaram obter um quadro do comportamento dos pesquisadores identificados no segmento da pequeña ciência, suas atitudes e percepções sobre o compatilhamento de dados de pesquisa científica.

Três estudos reprentativos da comunidade brasileira foram selecionados e relatados: uma pesquisa realizada junto à Fundação Carlos Chagas, na área de Genética, os outros dois realizados junto à Fundação Oswaldo Cruz, na área de Neurociências, um deles, indo além da proposta inicial**,** ao se aprofundar nas questões sobre as barreiras percebidas pelos pesquisadore e tentar minimizá-las, ao oferecer-lhes serviços e produtos que os estimulem ao compartilhamento.

Três estudos considerados representativos da comunidade internacional também foram selecionados e são apresentados: uma pesquisa adotando a metodologia de grupos focais com cientistas de cinco disciplinas (ciências atmosféricas e da terra, ciência da computação, química, ecologia e neurociência), afiliados a universidades, sobre práticas de compartilhamento de dados; as outras duas, envolvendo pesquisadores de diversos países, que são complementares, no sentido de que foram conduzidas pela mesma equipe, após um intervalo de dez anos, visando descobrir avanços no comportamento dos pesquisadores quanto às práticas e percepções sobre dados de pesquisa no período - ressalte-se que nesta pesquisa são incluídos outros setores, além do acadêmico, porém, são analisados isoladamente.

De um modo geral, os resultados revelam que grande parte docientistas do setor acadêmico ainda não compartilham os dados de suas pesquisas pelos diversos motivos aqui descritos. Exceção é observada no setor governamental, líder em atitudes positivas e boas práticas de dados, onde são constatados os maiores avanços. Revelam, também, que a maioria dos cientistas opera sem diretrizes estabelecidas de compartilhamento, e, entre as principais demandas, reivindicam treinamento em gerenciamento de dados, infraestrutura estável, desejam saber cómo os seus dados estão sendo usados e contar com o apoio do bibliotecário de dados. Finalmente, percebe-se que as questões relacionadas à motivação ao compartilhamento de dados científicos são comuns a ambas comunidades, nacionais e internacionais, bem como o são as barreiras que dificultam o compartilhamento. No Brasil, cujas práticas de dados são bem mais recentes, lições podem ser aprendidas com a experiência internacional, aqui abordadas, de modo a adotar estratégias que possibilitem alcançar níveis de compartilhamento comparáveis e com maior celeridade. No entanto, uma diferença sensível observada entre as comunidades, internacional e nacional, é relacionada à morosidade na implementação de políticas de dados no País, assim, entende-se que uma diretriz nacional para a gestão de dados de pesquisa deva ser uma ação prioritária.

Bibliografía

Anderson, C. (2004). The long tail. Wired, oct. https://www.wired.com/2004/10/tail/

Córdula, F. L. & Araújo, W. J. (2019). O compartilhamento de dados científicos na era do e-science. In: G. A. Dias, B. M. J. F. Oliveira (Orgs.), *Dados científicos*: *perspectivas e desafíos* (v. 6, pp. 189-207). Editora UFPB. https://figshare.com/articles/book/O_COMPARTILHAMENTO_DE_DADOS_CIENT_FICOS_NA_ERA_DO_E-SCIENCE/11876691

Cragin M. H., Palmer, C. L., Carlson, J. R. & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 368*(1926): 4023–4038. https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2010.0165

De Solla Price, D. (1963). *Little Science, Big Science*. Columbia University Press.

Donaldson, D. R. & Koepke, J. W. (2022) A focus groups study on data sharing and research data management. *Nature Scientific Data, v.9* (345), jun. https://www.nature.com/articles/s41597-022-01428-w

Fachin, G. R. B. (2002). *Modelo de avaliação para periódicos científicos online: proposta de indicadores bibliográficos e telemáticos*. (Dissertação de Mestrado. Universidade Federal de Santa Catarina. https://repositorio.ufsc.br/bitstream/handle/123456789/83088/185438.pdf?sequence=%3E

Sales, L., Sayão, L. F. (2019) A grande e a pequena ciência: análise das diferenças na gestão de dados de pesquisa. *Informação & Sociedade: Estudos*, *v. 29* (3). https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/47615

Sales, L.& Sayão, L. F. (2020). A ciência invisível: compartilhamento de dados na cauda longa de pesquisa. In: L. V. R. Pinheiro, P. M. Valerio (Orgs.). *Da gênese à contemporaneidade da comunicação e divulgação científicas* **(**pp. 289-304). Editora UFPB. https://brapci.inf.br/index.php/res/v/103678

Santos, P. R. (2022). *Práticas de comunicação de dados de pesquisa: a percepção de bolsistas de produtividade brasileiros do campo da genética* (Dissertação de Mestrado, Universidade de Brasília). https://bdtd.ibict.br/vufind/Record/UNB_6f40a85967dfc7315b84e8b7092c018e

Tenopir, C., Allard, S., Douglas, K., Aydinoglu, A. U., … & Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PLOS ONE*, *v. 6,* (6) e21101, june**.** https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021101

Tenopir, C., Rice, M. M., Allard, S., Baird, L., ... & Sandusky, R. J. (2020). Data sharing, management, use, and reuse: practices and perceptions of scientists worldwide. *PLOS ONE*, v. 15, no. 3. https://www.researchgate.net/publication/339869356_Data_sharing_management_use_and_reuse_Practices_and_perceptions_of_scientists_worldwide

Veiga, V. S. O. (2017). A percepção dos pesquisadores portugueses e brasileiros da área de neurociências quanto ao compartilhamento de artigos científicos e dados de pesquisa no acesso aberto verde: custos, benefícios e fatores contextuais (Tese de Doutorado. Instituto de Comunicação e Informação em Saúde (ICICT). https://www.arca.fiocruz.br/bitstream/handle/icict/26842/Viviane_Veiga_Tese_ICICT_2018.pdf?sequence=4&isAllowed=y

Veiga, V. S. O.; Silva, C. H. & Borges, M. M. (2021). Modelo de fatores que Influenciam no comportamento de compartilhamento de dados de pesquisa (pp. 153-188). In: M. M. Borges & E. S. Casado. *Sob a lente da ciência aberta olhares de Portugal, Espanha e Brasil*. Imprensa da Universidade de Coimbra. https://www.arca.fiocruz.br/bitstream/handle/icict/46243/Sob_lente_Ciencia_Aberta_olhares_PT_ES_BR.pdf?sequence=2&isAllowed=y

Rosane Teles Lins Castilho, ORCID https://orcid.org/0000-0002-7142-6813, é graduada em Biblioteconomia e Documentação pela Universidade Federal do Estado do Rio de Janeiro (UNIRIO), Mestre e Doutoranda em Ciência da Informação pelo Programa de Pós-Graduação em Ciência da Informação (PPGCI), do Instituto Brasileiro de Informação em Ciência e Tecnología (IBICT), em convênio com a Universidade Federal do Rio de Janeiro (UFRJ). Por mais de trinta anos, atuou no Departamento de Informática, da Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), como gestora da Biblioteca de Informática, integrada ao Sistema de Bibliotecas da PUC-Rio, e à frente dos serviços técnicos; como gestora da Assessoria de Biblioteca, Documentação e Informação, como editora técnica dos relatórios de pesquisa, série Monografias em Ciência da Computação; na gestão do Repertório da Produção Acadêmica do Departamento de Informática. Integra o Grupo de Estudos em Metrias (GEM) da Comunicação Científica e o Grupo Comunicação |Científica: percursos, desafios e perspectivas, ambps da UFRJ. Publicou artigos em periódicos, anais de congressos e capítulo de livro.

Evaluación y métricas alternativas

HERA 2.0: Más Funcionalidad para la Evaluación de Recursos Académicos

Ezequiel Carletti[1], Enzo Rucci[2], Gonzalo Luján Villarreal[3]

Palabras claves

Artículo científico, Revista científica, Paper, Journal, Bibliometría, Bases de datos académicas, Bases de dato bibliográficas, Cienciometría, Recuperación de información, Evaluación bibliográfica.

Eje temático

Evaluación y métricas alternativas

Resumen

En el marco de un escenario académico-científico donde la producción de información crece exponencialmente, la necesidad de herramientas que asistan en la evaluación de la calidad e impacto de los recursos disponibles se torna esencial. En esta línea, la aplicación HERA se erige como un recurso que busca agilizar y respaldar el proceso de valoración de artículos y revistas académicas. Sin embargo, reconociendo que las oportunidades de mejora siempre están presentes, este artículo tiene como objetivo presentar el desarrollo de una segunda versión de HERA, la cual contempla un conjunto de mejoras y extensiones. HERA 2.0 representa una versión sofisticada y extendida de su predecesora, al mejorar su rendimiento, escalabilidad, alcance y soporte. Considerando las características de esta nueva versión, se espera que los miembros de la comunidad académico-científica la encuentren de mayor utilidad para evaluar la calidad y el impacto de los recursos académicos y que contribuya a facilitar y acelerar dicha tarea.

Introducción

En la actualidad, los investigadores enfrentan un verdadero desafío al momento de tener que determinar la calidad y el impacto de los recursos académicos, debido a la combinación de dos factores. En primer lugar, el sostenido crecimiento de publicaciones científicas procedentes del desarrollo tecnológico (Smith, 2013; Sumit, 2024). En segundo lugar, la carencia de estándares y la disponibilidad de múltiples sistemas de evaluación y métricas, que aún cuando comparten objetivos, no siempre utilizan las mismas metodologías (Porto, 2021).

En este contexto, HERA surgió como una respuesta ante estos desafíos, siendo una aplicación web que enriquece recursos académicos (artículos o revistas) al integrar información de diferentes bases de datos académicas (Porto et al., 2022a, 2022b). Esta herramienta automatizada está dirigida principalmente a miembros de la comunidad académico-científica, y busca simplificar, agilizar y respaldar el proceso de determinar la calidad y el impacto de un recurso académico. Como parte de su funcionamiento, HERA con-

1 Facultad de Informática, Universidad Nacional de La Plata, Argentina, carlettieze@gmail.com

2 III-LIDI, Facultad de Informática, Universidad Nacional de La Plata y Comisión de Investigaciones Científicas, Argentina, erucci@lidi.info.unlp.edu.ar

3 PREBI-SEDICI Universidad Nacional de La Plata y CESGI Comisión de Investigaciones Científicas, Argentina, gonzalo@prebi.unlp.edu.ar

sulta múltiples fuentes en tiempo real para luego ofrecer información de un recurso determinado, como ser sus metadatos, su pertenencia a índices y bases de datos, sus indicadores de citas y menciones, e información de la publicación donde figura dicho recurso en caso que corresponda. En particular, HERA se nutre de múltiples bases de datos como CrossRef[4], DOAJ[5], Scopus[6], REDIB[7], Dimensions[8], entre otras; y se encuentra disponible para la comunidad en forma abierta[9].

En su primera versión, HERA resultó funcional y de gran utilidad para los usuarios. Recientemente, se ha lanzado una segunda versión, la cual contempla un conjunto de mejoras y extensiones. A partir de estas, HERA 2.0 busca optimizar su rendimiento, escalabilidad, alcance y soporte.

HERA 1.0

Propósito

HERA es una herramienta diseñada para simplificar, agilizar y apoyar el proceso de determinar la calidad y el impacto de un recurso académico, como un artículo o una revista. Sus características principales son:

Permite al usuario visualizar de manera sencilla información relevante proveniente de múltiples bases de datos y criterios de expertos. Esto facilita la comprensión y análisis de los datos de una manera rápida y eficiente.

Permite automatizar el proceso de recopilación de información, lo que evita que el usuario tenga que buscar en cada sitio de forma individual. Esto ahorra tiempo y esfuerzo, permitiendo enfocarse en la interpretación de los datos recabados.

Permite perfilar rápidamente los datos obtenidos, como el número de citas, factores de impacto, información sobre licencias, ubicación de la publicación y más. Esto facilita la toma de decisiones y permite al usuario realizar un análisis personalizado según sus objetivos e intereses.

En resumen, HERA 1.0 brindaba a los usuarios la posibilidad de centrarse en el análisis y la interpretación de los datos, ahorrando tiempo y esfuerzo en la búsqueda de información.

Diseño y Desarrollo

En la implementación de HERA 1.0, se conformó en una aplicación web organizada en dos desarrollos: una aplicación front-end desarrollada con ReactJS[10] y una aplicación back-end desarrollada con NodeJS[11]. El propósito de la aplicación backend es actuar como intermediaria entre la aplicación frontend y los diferentes sistemas que suministran datos y métricas de recursos académicos, asumiendo así la mayor parte de la carga de trabajo.

4 https://www.crossref.org/
5 https://doaj.org/
6 https://www.scopus.com/
7 https://www.redib.org/
8 https://www.dimensions.ai/
9 http://hera.sedici.unlp.edu.ar
10 https://react.dev/
11 https://nodejs.org/en

La aplicación frontend, por otro lado, solicita información a la aplicación backend a través de una API REST, recopila los resultados y se encarga de generar visualizaciones en un sitio web.

Funcionamiento

El uso de HERA 1.0 es bastante sencillo: el usuario dispone de una barra de búsqueda en la que debe ingresar un identificador, que puede ser el ISSN para analizar una publicación o un DOI para analizar un artículo (ver Figura 1).

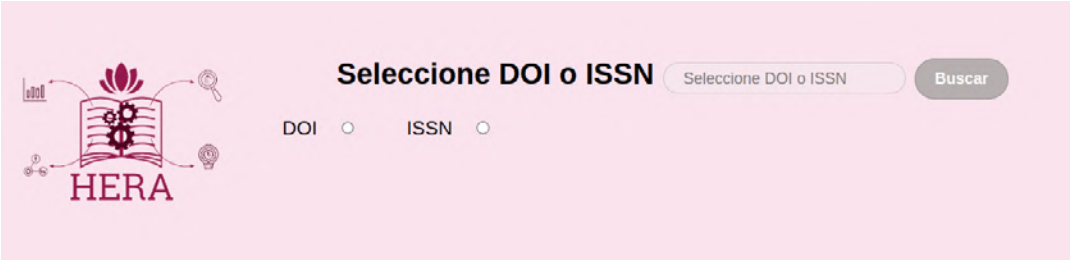


Figura 1 - Interfaz de búsqueda de la herramienta con input de búsqueda y selectores de tipo de identificador (DOI e ISSN)

A continuación, se debe hacer clic en el botón "Buscar" para indicarle a HERA que comience la búsqueda de información del recurso asociado a dicho identificador. Luego, HERA 1.0 establece conexiones concurrentes con un conjunto de servicios en línea y sitios web (bases de datos en adelante) para intentar obtener información del recurso de interés. Además, HERA 1.0 determina la mejor estrategia para extraer información de cada base de datos (por ejemplo, una API REST o web scraping), así como también qué tipo de información puede obtener de cada una de ellas: pertenencia o no a la base de datos, citas, gráficos, altmetrics, etc. En (Porto et al., 2022b) se puede consultar el flujo detallado para las diferentes búsquedas posibles.

HERA 2.0: Mejoras y Extensiones

HERA 2.0 mantiene el propósito de su versión previa. Esta nueva versión busca extender y mejorar la funcionalidad, además de optimizar el rendimiento. Estas mejoras y extensiones tienen como meta principal ofrecer una experiencia mejorada tanto para los usuarios finales como para los desarrolladores involucrados en el proyecto.

Actualización del Banco de Bases de Datos

Es fundamental para el valor y utilidad de HERA mantener actualizado el repositorio de bases de datos. Este proceso consta de un análisis meticuloso de las fuentes de datos en uso, la identificación de nuevas fuentes de datos potenciales, y la implementación de medidas que aseguren un servicio continuo.

Evaluación de las Fuentes de Datos de HERA 1.0

Se realizaron revisiones de todos los servicios que emplea HERA 1.0, para asegurar que estos siguen proporcionando datos pertinentes, manteniendo sus API accesibles y que cualquier cambio en su estructura de datos o términos de uso no afecta adversamente la funcionalidad de la herramienta. En particular, se revisaron los servicios provistos por Crossref, DOAJ, Dimensions, Altmetrics, SemanticScholar, Scopus, WoS, SJR, Microsoft Academic y Google Scholar.

Durante el proceso anterior, se constató que Microsoft Academic dejó de estar operativo, tal como se anunció en un comunicado oficial en mayo de 2021[12], indicando que su sitio web y las API subyacentes se retirarían el 31 de diciembre de 2021. Por tanto, Microsoft Academic fue eliminado de la lista de fuentes de datos utilizadas en HERA 2.0.

Adicionalmente, se identificó un problema temporal con la fuente de datos REDIB, la cual, según un comunicado oficial, podría tener interrupciones temporales o retrasos en la actualización de contenidos debido a cambios en la administración y la financiación del proyecto[13]. A pesar de estos inconvenientes, en HERA 2.0 se decidió mantener REDIB como fuente de datos, aunque actualmente no está respondiendo.

Identificación de Nuevas Fuentes de Datos

Además de la revisión de las fuentes de datos actuales, se identificó y añadió una nueva fuente de datos para mejorar la funcionalidad y cobertura de HERA 2.0. En particular, se trata de OpenAlex (Priem et al., 2022), la cual es una plataforma de datos académicos de acceso abierto desarrollada por la organización sin fines de lucro OurResearc[14]. OpenAlex fue creada como una alternativa a Microsoft Academic Graph (MAG) y ofrece una base de datos amplia y actualizada de información académica que incluye publicaciones, autores, afiliaciones, citas y otros datos relacionados con la investigación.

Esta nueva versión de HERA emplea OpenAlex como su principal fuente de datos, en lugar de basarse principalmente en Crossref como en la versión 1.0, considerando que OpenAlex ya contiene la información de Crossref. Además, este cambio permitió ampliar las capacidades de HERA para incluir información y enlaces a PubMed[15] y PubMed Central[16].

Por otra parte, Google Scholar (GS) reúne una vasta cantidad de información académica, facilitando el acceso a una diversidad de artículos científicos. Sin embargo, la integración de GS plantea una serie de desafíos técnicos, legales y éticos que tuvieron que ser considerados en el desarrollo de esta nueva versión de la herramienta. Debido a esta limitaciones, se decidió agregar un botón en la interfaz de usuario que redirige al usuario a la página de GS para el DOI buscado (ver Figura 4).

Por último, si bien Scopus estaba presente para revistas y congresos, se decidió también incluirla en la búsqueda de DOI. Aunque Scopus provee una API para esta funcionalidad, se decidió seguir el mismo enfoque que con GS por cuestiones operativas y de seguridad (Carletti, 2023).

12 Comunicado Microsof Academic
13 Comunicado Oficial de REDIB
14 http://ourresearch.org/
15 PubMed - Sitio Web
16 Sitio web de PubMed Central https://pmc.ncbi.nlm.nih.gov

Nueva Interfaz

Para la implementación de la nueva interfaz en HERA 2.0, se optó por incorporar el uso de React-Bootstrap[17], una combinación de Bootstrap y ReactJS. Bootstrap es un marco de diseño popular y ampliamente utilizado que proporciona una serie de estilos y componentes predefinidos para facilitar el desarrollo de interfaces de usuario ricas y atractivas[18]. ReactJS, por otro lado, es una biblioteca de JavaScript para la construcción de interfaces de usuario de manera eficiente y flexible[19].

La incorporación del paquete React-Bootstrap en la nueva versión de HERA ha producido resultados significativos en términos de diseño y usabilidad. React-Bootstrap, ha permitido una interfaz más intuitiva y atractiva. Mediante este cambio se mejora la apariencia de la aplicación, facilitando su uso en diferentes dispositivos y resoluciones de pantalla. Se considera que la integración de React-Bootstrap ha simplificado el proceso de desarrollo, proporcionando un conjunto coherente y estandarizado de componentes estilizados y responsivos. Esto ha mejorado la cohesión visual y la consistencia de diseño en toda la aplicación, proporcionando una experiencia de usuario más coherente.

En la aplicación, se presenta una interfaz de usuario única que se caracteriza por la presencia de una destacada barra de búsqueda. En esta barra, los usuarios deben ingresar el identificador del recurso que desean buscar, tal como se ilustra en la Figura 2.

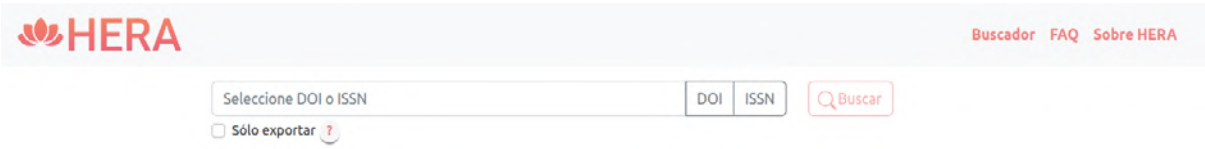


Figura 2: Interfaz de búsqueda de la herramienta con input de búsqueda y selectores de tipo de identificador (DOI e ISSN)

Dentro de la interfaz, los usuarios deben elegir el tipo de búsqueda que desean realizar, ya sea mediante el uso del DOI o del ISSN. Una vez seleccionado el tipo de búsqueda, deberán ingresar el identificador correspondiente en la barra designada. Más allá de la nueva interfaz, el flujo de búsqueda es el mismo que en el caso de HERA 1.0 en general. El único cambio consiste en la posibilidad de introducir múltiples identificadores en la barra de búsqueda (se deben separar por comas) para permitir su búsqueda simultánea.

La Figura 3 exhibe los resultados para un artículo de ejemplo. Entre los resultados, se pueden observar varios metadatos del recurso, como el título, el tipo de recurso, el resumen, entre otros. Además, se muestra el nombre de la revista en la que se publicó el artículo, junto con un enlace que dirige a la página original del mismo. En caso de ser posible, la aplicación también intentará recopilar métricas del journal en el que se encuentra el artículo, y las mostrará en conjunto, como se muestra en la Figura 4. Para esto, se

17 https://react-bootstrap.netlify.app/
18 https://getbootstrap.com/
19 https://es.react.dev/

extrae el ISSN del journal de los metadatos recuperados y se realiza automáticamente una búsqueda por ISSN en segundo plano. De esta manera, se amplía el contexto de evaluación de un artículo, permitiendo ver sus métricas de forma aislada y también acceder a las métricas del journal, lo que ayuda a comprender qué influencia puede tener en las métricas del artículo. Al hacer clic en el botón "Ver más" que se muestra en esta última figura, se genera una vista expandida de las métricas, donde se puede obtener información más detallada (en los casos en que las bases de datos proporcionen datos adicionales), como se ilustra en las Figuras 5, 6 y 7.

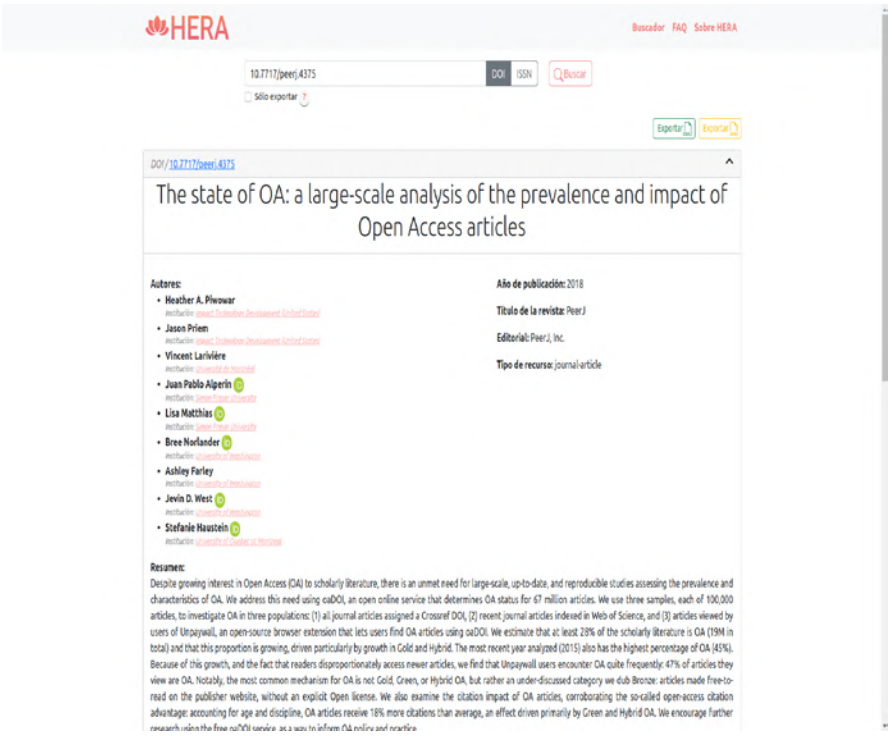


Figura 3. Resultados generales de la búsqueda de un *paper* (vista resumida)

Figura 4. Métricas correspondientes al *paper* y *journal* donde se encuentra el *paper* (vista resumida en tarjetas)

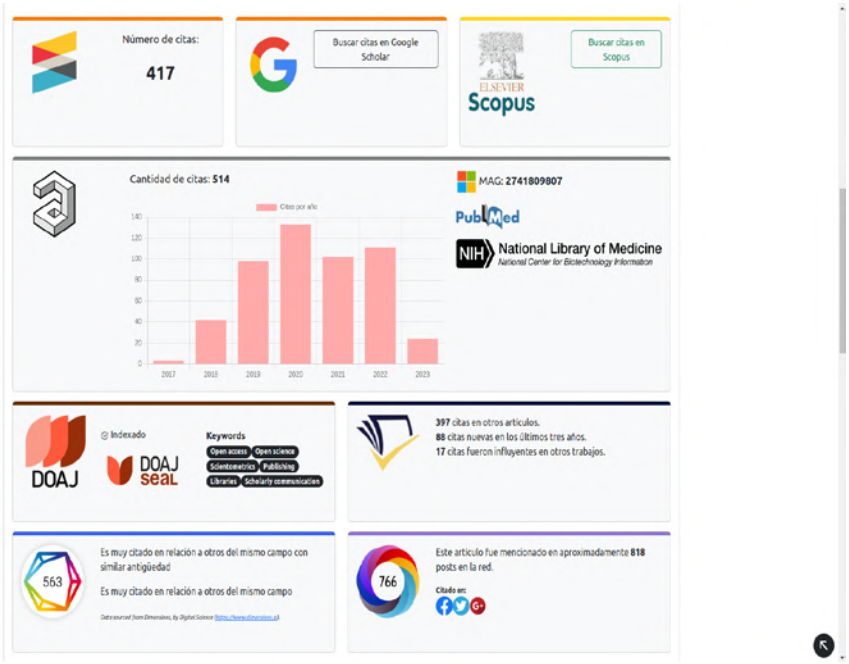


Figura 5. Vista expandida de métricas de un paper

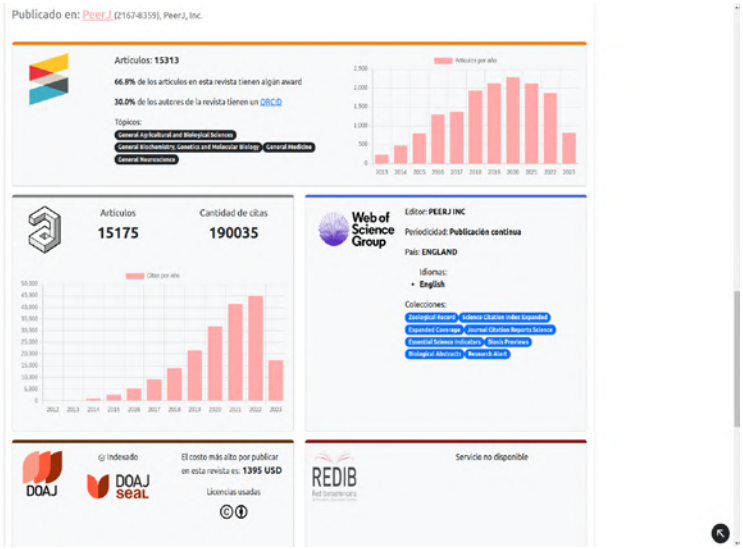


Figura 6. Vista expandida de métricas de un journal (parte 1)

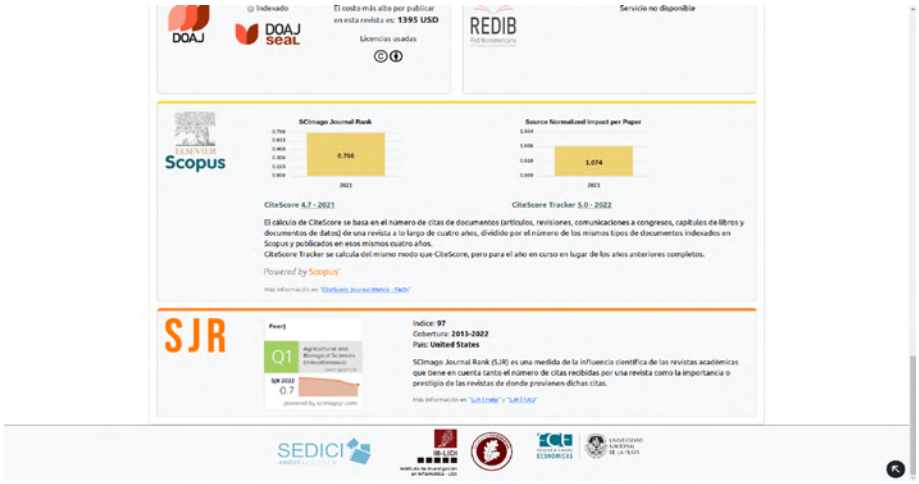


Figura 7. Vista expandida de métricas de un *journal* (parte 2)

Integración de Información de Autores e Instituciones

Para integrar una mayor cantidad de información sobre los autores de una publicación en la aplicación HERA, se contempló la utilización de dos fuentes de datos adicionales: Research Organization Registry (ROR)[20] y Open Researcher and Contributor ID (ORCID)[21] . Por un lado, con la integración de la información de ROR, HERA ahora puede mostrar datos detallados de la organización o institución a la que pertenecen los autores. Esto ofrece a los usuarios un contexto más completo y detallado sobre el entorno de investigación del autor. Por otro lado, al integrar la información de ORCID, HERA ahora puede proporcionar un perfil más completo de cada autor, incluyendo sus publicaciones anteriores, afiliaciones y contribuciones. Este nivel de detalle adicional proporciona a los usuarios una visión más profunda del trabajo y la trayectoria de cada autor, lo que mejora la utilidad y la riqueza de la información que HERA 2.0 puede proporcionar.

La Figura 8 ilustra los beneficios de la integración mencionada, aunque es importante destacar que esta información adicional solo se muestra cuando está disponible para un autor específico.

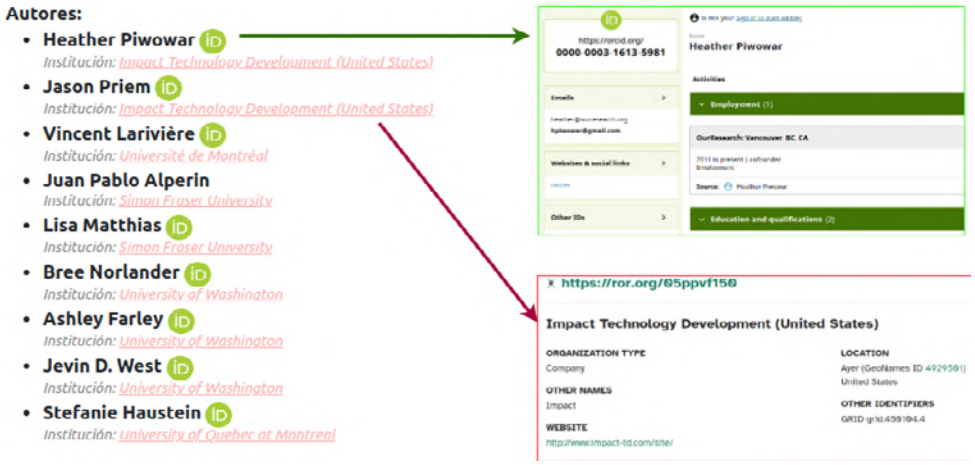


Figura 8: Ejemplo de vista de autores en HERA 2.0.

Exportación a Formatos Procesables

La implementación de esta funcionalidad proporciona una considerable ventaja para los usuarios que necesiten realizar análisis de datos, investigaciones académicas, auditorías bibliográficas, entre otras aplicaciones. En particular, se consideraron dos formatos que son ampliamente usados en la actualidad, como CSV y JSON.

20 ROR es un registro global y comunitario que proporciona identificadores únicos para las organizaciones que participan en la producción de investigación.

21 ORCID proporciona un identificador digital persistente que distingue a un investigador de otros.

La exportación de datos permite a los usuarios obtener los principales indicadores y métricas para un recurso académico (tanto para DOI como para ISSN) de forma conjunta y estructurada. Cuando se realizan búsquedas simultáneas, la información recolectada de cada DOI o ISSN se presenta en filas separadas, facilitando su lectura y posterior procesamiento. En la Figura 9, se muestra un ejemplo de una descarga en formato CSV correspondiente a 10 DOIs.

	doi	type	title	authors	abstract	publication_year
1	10.3389/fenvs.2020.581591	journal-article	Analysis of Water Pollution Us	Rohit Sharma,Raghvendra Kuma	The Yamuna river has bec	2020
2	10.1109/LCOMM.2019.2898944	journal-article	Deep Learning-Based Channe	Mehran Soltani,Vahid Pourahmad	In this letter, we present a	2019
3	10.1016/j.telpol.2021.102261	journal-article	Internet of things and the ecor	Günter Knieps,Johannes M. Baue	Fifth generation (5G) netw	2022
4	10.24215/26838559e28	journal-article	Diabetes Link: innovación tec	Enzo Rucci,Lisandro Nahuel Deli	La Diabetes Mellitus (DM)	2021
5	10.1007/978-3-031-07802-6_9	book-chapter	Migrating CUDA to oneAPI: A	Manuel Costanzo,Enzo Rucci,Ca	In order to tackle the prog	2022
6	10.1109/CLEI53233.2021.9640225	proceedings-article	Performance vs Programming	Manuel Costanzo,Enzo Rucci,Ma	Historically, Fortran and C	2021
7	10.1007/978-3-030-61702-8_25	book-chapter	Diabetes Link: Platform for Se	Enzo Rucci,Lisandro Nahuel Deli	Diabetes Mellitus (DM) is	2020
8	10.7717/peerj.4375	journal-article	The state of OA: a large-scale	Heather A. Piwowar,Jason Priem	Despite growing interest i	2018
9	10.24215/18522971e087	journal-article	Desafíos de la videovigilancia	Andrés Herrera Esquivel	La videovigilancia automa	2021
10	10.5710/PEAPA.21.05.2020.298	journal-article	NEUROANATOMÍA DEL SAU	Ariana Paulina Carabajal,Leonar	this taxon has uncertain p	2020

	crossref_cites	openalex_doi_cites	doaj_doi_presence	semanticscholar_cites	dimensions_cites	altmetric_cites
1	33	33	Si (Si)	19	51	1
2	261	281	No	278	278	2
3	7	6	No	8	11	1
4	0	0	No	0	0	2
5	2	2	No	5	6	17
6	1	1	No	1	2	
7	0	0	No	0	0	
8	434	541	Si (Si)	411	588	817
9	0	0	No		0	
10	1	1	No	1	2	1

Figura 9: Datos DOI en formato CSV sin la información de sus correspondientes ISSN (modificados para facilitar su visualización en una sola imagen)

Manejo de Errores y Recursos no Encontrados

Un aspecto crítico de cualquier sistema de software es cómo se manejan los errores y las situaciones en las que no se encuentran los recursos solicitados. En HERA 2.0 se ha robustecido la gestión de errores en combinación con una comunicación más clara hacia los usuarios cuando ocurren situaciones excepcionales.

Existen diferentes casos en los que puede surgir un error o no encontrarse un recurso en el contexto de la aplicación. Algunos ejemplos comunes son:

Recurso no encontrado: ocurre cuando el usuario realiza una búsqueda de un DOI o un ISSN que no existe en ninguna de las bases de datos consultadas. En este caso, HERA 2.0 devuelve un mensaje claro al usuario indicando que el recurso solicitado no se encontró. En la Figura 10, la fuente de información DOAJ responde con el mensaje de recurso no encontrado:

Figura 10: Métricas de DOAJ no encontradas

Servidor no disponible: surge en situaciones en las que se intenta hacer una solicitud a una base de datos y el servidor de dicho banco de datos no responde o no está disponible en ese momento. En este caso, HERA 2.0 informa al usuario que el servidor está actualmente no disponible (ver Figura 11).

Figura 11: Tarjeta de REDIB informando: Servidor no disponible.

Límite de solicitudes alcanzado: sucede en circunstancias donde se han realizado demasiadas solicitudes a una base de datos en un período corto de tiempo y se ha alcanzado el límite establecido por esa API externa. Este es un caso común con las API que tienen un límite en el número de solicitudes que se pueden hacer en un cierto período de tiempo. En este caso, HERA 2.0 informa al usuario que se ha alcanzado el límite de solicitudes para esa base de datos y que debe esperar un cierto tiempo antes de realizar más solicitudes (ver Figura 12).

Figura 12: Tarjeta de OpenAlex informa: Límite de solicitudes alcanzado

Este sistema de manejo de errores y notificaciones no sólo mejora la experiencia del usuario, sino que también aumenta la robustez de la aplicación, ya que puede manejar situaciones excepcionales y proporcionar información útil al usuario en caso de que ocurran.

Búsqueda simultánea de múltiples recursos

El traslado de la lógica de negocio del *frontend* al *backend* en HERA se considera un avance en términos de rendimiento y tiempo de respuesta (ver sección siguiente). En la versión 1.0, para un recurso como un DOI, se debían realizar hasta siete consultas, y siete adicionales si el DOI contaba con un ISSN correspondiente. Estas consultas, se realizaban de manera secuencial por efectuarse en el frontend, lo que implicaba un consumo considerable de tiempo.

En HERA 2.0, estas peticiones se gestionan mediante el framework de aplicaciones web Express en el backend, permitiendo su ejecución concurrente (paralela si el hardware subyacente lo permite) en lugar de secuencial. Este cambio en la estrategia de ejecución ha permitido una optimización en la eficiencia y velocidad de las consultas (Carletti, 2023). Al realizar múltiples solicitudes en concurrente, se reduce el tiempo de espera asociado con las consultas secuenciales, mejorando así la experiencia del usuario y el rendimiento global de la herramienta.

Pese a que la búsqueda simultánea de recursos se optimizó a través de consultas concurrentes entre los DOI/ISSN, se enfrentó una limitación: cada API externa impone restricciones en cuanto a la cantidad de consultas que pueden realizarse en un período de tiempo determinado. Para superar este desafío, se implementó un sistema de búfer para regular la cantidad de consultas simultáneas. Por ejemplo, en lugar de buscar 30 artículos a la vez, se estableció un búfer para realizar las consultas en lotes de 5 a la vez. Este enfoque, a pesar de incorporar un grado de secuencialidad, permite mantener un alto rendimiento sin exceder las limitaciones impuestas por las API externas.

Refactorización de código

La refactorización de código es una técnica esencial para mantener la calidad del software, permitiendo mejorar la estructura interna del código sin cambiar su comportamiento externo (Méndez, 2010; Team, 2021). En HERA 1.0 se encontraron varios problemas o posibles oportunidades de optimización, incluyendo:

Falta de modularización: la estructura del código de HERA 1.0 carece de modularización, lo cual puede complicar la legibilidad y mantenimiento del sistema.

Manejo de múltiples fuentes de datos: la gestión de múltiples fuentes de datos presenta margen para optimización, mejorando así la eficiencia en su procesamiento.

Lógica de negocio: la implementación actual de la lógica de negocio en HERA 1.0 muestra áreas de mejora, con posibles optimizaciones para incrementar la funcionalidad de la aplicación.

Es por ello que llevó a cabo una refactorización integral del código de HERA para mejorar su eficiencia, mantenibilidad y escalabilidad. Más detalles pueden ser consultados en (Carletti, 2023).

Implementación de API REST

La construcción de la API REST en HERA constituye una actualización estratégicamente significativa, la cual contribuye a una mayor integración con aplicaciones y servicios externos (ver Figura 13), a la optimización del rendimiento en las solicitudes y al aumento de la escalabilidad de la plataforma.

Interoperabilidad y flexibilidad: la adopción de una API REST ofrece a HERA 2.0 la capacidad de interactuar con mayor eficiencia con las diferentes fuentes de datos que utiliza, como OpenAlex, Crossref y DOAJ. Este enfoque facilita la incorporación de nuevas fuentes de datos en el futuro y la flexibilidad requerida para adaptarse a los cambios y evolución en el ecosistema de datos. Al adherirse a los principios de la arquitectura REST, la API permite la interacción de otras aplicaciones con HERA, potencialmente generando nuevas vías para la utilización y crecimiento de la plataforma.

Eficiencia: la implementación de una API REST mejora la eficiencia de la plataforma. Las API REST emplean métodos HTTP estándar (GET, POST, PUT, DELETE), los cuales son ampliamente adoptados y optimizados en la industria. Adicionalmente, una API REST es capaz de manejar múltiples tipos de datos (como JSON o CSV), lo que permite a los clientes solicitar los datos en el formato que mejor se adapte a sus requerimientos y a la naturaleza de sus operaciones.

Escalabilidad: en términos de escalabilidad, una API REST es capaz de manejar una gran cantidad de solicitudes y sustentar el crecimiento de la plataforma sin afectar su rendimiento. Las API REST son stateless[22], lo que significa que cada solicitud es autónoma y no depende de la información de solicitudes previas. Esta característica facilita la distribución de las solicitudes entre múltiples servidores, mejorando así el rendimiento y la capacidad de respuesta de la plataforma.

Gestión de múltiples consultas: una característica adicional de la API REST de HERA es su capacidad de generar múltiples búsquedas y proporcionar múltiples resultados simultáneamente. Esta funcionalidad es fundamental para el perfilamiento eficiente de los datos, permitiendo obtener una visión amplia y detallada de los conjuntos de datos con los que se trabaja.

22 Protocolo sin estado.

La incorporación de la API REST permitió el diseño del frontend y de un complemento para el navegador Google Chrome como las primeras aplicaciones clientes livianas y externas al core de HERA.

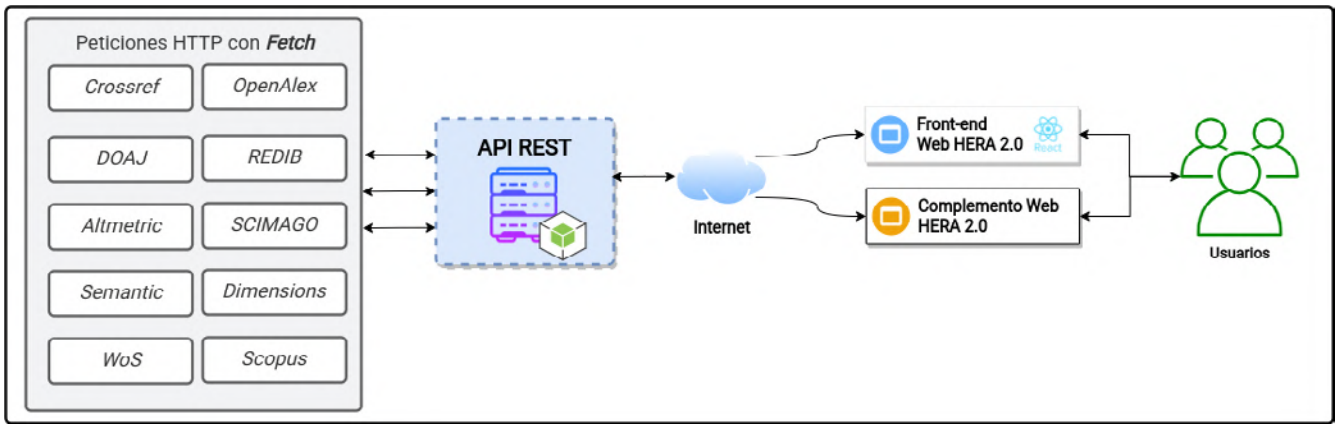


Figura 13: Esquema de comunicación de la API REST

Gestión de las fuentes de información

En el proceso de reestructuración y mejora del código de HERA, una cuestión de relevancia era garantizar la flexibilidad del sistema frente a posibles cambios en las fuentes de información. Con este fin, se diseñó un archivo de configuración que almacena información detallada de cada fuente de datos externa, entre la que se incluye un identificador único de cada fuente (por ejemplo, "CrossrefDOI"), su correspondiente URL y un indicador booleano "enabled" para señalar si la fuente está actualmente activa o no.

En la Tabla 1, se presenta un ejemplo simplificado del archivo de configuración. Dicha tabla sirve como guía para la interacción del sistema con las diferentes fuentes de datos y componentes, especificando los parámetros necesarios para su correcta operación.

Fuente de datos	URL	enabled
OpenAlexDOI	https://api.openalex.org/works?filter=doi:https://doi.org/	true
CrossrefDOI	https://api.crossref.org/works/	true
DoajDOI	https://doaj.org/api/v2/search/articles/doi:	true
Microsoft Academic	https://www.microsoft.com/en-us/research/project/academic/	false
ScimagoISSN	https://www.scimagojr.com/journalsearch.php?q=	true

Tabla 1: Representación resumida del archivo de configuración.

En primer lugar, este enfoque externaliza los detalles específicos de cada fuente de datos del código base, proporcionando un alto grado de modularidad. Así, la inclusión o exclusión de fuentes de datos se simplifica enormemente, requiriendo solo la modificación del valor del campo "enabled" en la configu-

ración correspondiente. Esta flexibilidad probó su valor en situaciones como fue el caso con Microsoft Academic, que cesó su funcionamiento durante esta etapa de desarrollo de HERA. En lugar de requerir una revisión exhaustiva del código para eliminar referencias a dicha fuente, fue suficiente con cambiar su estado a "inhabilitado" en el archivo de configuración.

En segundo lugar, esta implementación aumenta la robustez del sistema. Si una API deja de estar disponible o experimenta interrupciones, se puede desactivar fácilmente en la configuración sin que esto afecte a la funcionalidad general de HERA. En lugar de causar errores o problemas de rendimiento, la falta de disponibilidad de una API se traduce simplemente en que esa fuente de datos no se incluye en el contexto generado.

Conclusiones y trabajos futuros

HERA proporciona una herramienta de gran utilidad para simplificar y agilizar el proceso de recolección de indicadores actualizados sobre la calidad de los recursos académicos y el impacto que estos están generando en la comunidad académica internacional. HERA 2.0 representa una versión sofisticada y extendida de su predecesora, al mejorar su rendimiento, escalabilidad, alcance y soporte. Considerando las características de esta nueva versión, se espera que los miembros de la comunidad académico-científica la encuentren de mayor utilidad para evaluar la calidad y el impacto de los recursos académicos y que contribuya a facilitar y acelerar dicha tarea.

Las tareas de refactorización de código, el rediseño de la arquitectura de la aplicación y la incorporación de la API REST, realizadas en la última versión, sientan las bases para considerar nuevos desarrollos que permitan ampliar el alcance y brindar mayor escalabilidad para soportar un mayor número de usuarios. Entre los desarrollos a futuro se están considerando extensiones similares a la ya desarrollada para Google Chrome, para los navegadores web Firefox, Edge y Safari. Asimismo, los indicadores recuperados por HERA cobran una gran relevancia en contextos donde se gestionan documentos académicos, como por ejemplo repositorios institucionales, portales de revistas científicas o sitios web institucionales. En la siguiente etapa del desarrollo se propondrán integraciones para sistemas de gestión de contenidos, como por ejemplo Wordpress o Drupal, de gestión editorial, como por ejemplo Open Journal Systems, y plataformas para repositorios digitales, como por ejemplo Dspace. Finalmente, a fin de soportar una mayor carga en cuanto a la cantidad de consultas que deberá atender HERA en el futuro, así como también de evitar o minimizar las penalizaciones por superar los límites de consultas a las fuentes de datos, se evaluará la incorporación de un módulo de almacenamiento de indicadores en una caché propia. Este módulo permitirá a HERA enviar indicadores de recursos para los que ya había recibido consultas, sin tener que volver a consultar las fuentes de datos primarias. La implementación de este módulo requerirá evaluar los tiempos de expiración de los datos en caché para cada fuente de datos, así como también la manera de almacenar y gestionar el volumen de información que será necesario alojar.

Referencias

Carletti, E. (2023). *HERA 2.0: Extensión de alcance y funcionalidad* [Tesina de Licenciatura en Informática, Universidad Nacional de La Plata]. http://sedici.unlp.edu.ar/handle/10915/157417

Méndez, M. (2010). *Refactoring de código estructurado* [Tesis, Universidad Nacional de La Plata]. http://sedici.unlp.edu.ar/handle/10915/4184

Porto, J. F. (2021). *HERA: Herramienta para Enriquecimiento de Recursos Académicos* [Tesina de Licenciatura en Sistemas, Universidad Nacional de La Plata]. http://sedici.unlp.edu.ar/handle/10915/129874

Porto, J. F., Rucci, E., & Villarreal, G. L. (2022a). HERA, una Herramienta para la Evaluación de Recursos Académicos. *Actas del XXVIII Congreso Argentino de Ciencias de la Computación (CACIC 2022)*, 546-557. http://sedici.unlp.edu.ar/handle/10915/149623

Porto, J. F., Rucci, E., & Villarreal, G. L. (2022b). HERA-Herramienta para Enriquecimiento de Recursos Académicos. *Actas de la XI Conferencia Internacional de Bibliotecas y Repositorios Digitales (BIREDIAL-ISTEC)*, 251-266. http://sedici.unlp.edu.ar/handle/10915/148922

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts* (arXiv:2205.01833). arXiv. https://doi.org/10.48550/arXiv.2205.01833

Smith, R. (2013, julio 1). *How Technology is Changing Academic Research | WIRED*. https://www.wired.com/insights/2013/07/how-technology-is-changing-academic-research/

Sumit. (2024, marzo 22). *How Does Technology Help In Doing Academic Research? | Scoop Byte*. https://www.scoopbyte.com/how-does-technology-help-in-doing-academic-research/

Team, L. (2021, diciembre 3). *What Is Code Refactoring? Definition, Benefits and Why It's Important*. Lvivity. https://lvivity.com/what-is-code-refactoring

Ezequiel Carletti es Analista Programador Universitario y estudiante de Licenciatura en Informática por la Facultad de Informática de la Universidad Nacional de La Plata. En el año 2023 presentó su tesis de grado "HERA 2.0: Extensión de alcance y funcionalidad", proyecto realizado bajo la dirección del Dr. Enzo Rucci y la codirección del Dr. Gonzalo Villarreal, y en colaboración con la Facultad de Ciencias Económicas, la Facultad de Informática y PREBI-SEDICI, dependientes de la UNLP.

Enzo Rucci (https://orcid.org/0000-0001-6736-7358) es Doctor en Ciencias Informáticas por la Facultad de Informática de la Universidad Nacional de La Plata, docente-investigador de la UNLP en las áreas relacionadas con procesamiento concurrente y paralelo, e investigador adjunto de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires. Forma parte del Instituto de Investigación en Informática LIDI (III-LIDI, UNLP-CIC), donde realiza sus actividades de investigación en temáticas vinculadas a Cómputo de Alto Rendimiento y Aplicación de TICs a Ciencias de la Vida y Bibliometría. Es co-editor del Journal of Computer Science and Technology.

Gonzalo Luján Villarreal (https://orcid.org/0000-0002-3602-8211) es Doctor en Ciencias Informáticas por la Facultad de Informática de la Universidad Nacional de La Plata, es docente-investigador de la UNLP y director del Centro de Servicios en Gestión de Información (CESGI) de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires. Desarrolla su actividad docente en grado y posgrado de la Facultad de Informática de la UNLP. Trabaja con la plataforma OJS desde el año 2008, cuando en el marco de PREBI-SEDICI la UNLP lanzó el Portal de Revistas de la UNLP. En la actualidad es coordinador técnico de revistas científicas de la UNLP, y brinda asesoramiento a organizaciones y equipos editoriales temas vinculados a publicaciones científicas, circuitos editoriales y gestión de OJS.

Infraestructura tecnológica

Panorama dos repositórios de dados de pesquisa brasileiros

Carla Beatriz Marques Felipe[1], *Raimunda Fernanda dos Santos*[2]

Palabras claves

Repositório de dados de pesquisa. Panorama dos repositórios. Representação da informação.

Eje temático

Infraestructura tecnológica

Resumen

Estuda questões concernentes aos repositórios de dados de pesquisa no Brasil. Tem como objetivo geral apresentar um panorama dos Repositórios de Dados de Pesquisa brasileiros, em especial no que diz respeito à representação da informação. Tem como objetivos específicos: identificar o perfil dos repositórios de dados de pesquisa brasileiros; verificar como são realizadas as práticas de representação da informação nesses ambientes; averiguar as ferramentas que são utilizadas como infraestrutura para as práticas de representação da informação. Para atender o objetivo geral desta investigação, foram utilizadas como metodologia as pesquisas bibliográfica, documental, exploratória e descritiva com abordagem qualiquantitativa. A busca dos repositórios brasileiros foi realizada no buscador internacional Re3data em março de 2024, sistema que engloba repositórios de dados de pesquisa do mundo todo e das mais variadas disciplinas científicas. Como resultados para a seleção do *corpus*, a busca executada em novembro de 2023, foram identificados 19 Repositórios de Dados de Dados de Pesquisa em âmbito nacional, todos eles com acesso aberto aos dados armazenados em sua ambiência. Constata que os repositórios brasileiros ainda se encontram em fase de desenvolvimento, sobretudo no que diz respeito aos processos e instrumentos relacionados à indexação.

Introdução

Os dados de pesquisa são fundamentais para o avanço da ciência, sendo insumos básicos no fazer científico. Nesse contexto, com o avanço da ciência em termos de tecnologia e o surgimento do movimento acesso à informação, o compartilhamento, uso e reuso dos dados são práticas emergentes e necessárias. Alguns dados têm sido armazenados e representados em ambientes como os Repositórios de Dados de Pesquisa. Esses ambientes devem proporcionar ao usuário informações claras para que o reuso seja realmente efetivo.

1 *UFRJ, felipecarla12@gmail.com*
2 UFRN, *raimunda.fernanda@ufrn.br*

Os repositórios de dados podem ser definidos como "... bases de dados digitais, onde são armazenados, disseminados e preservados os dados de pesquisa em formato digital" (Costa, 2017, p. 46). Tais ambientes se constituem como parte essencial do fazer científico, auxiliando os pesquisadores na verificação da veracidade e qualidade de dados para uso e reuso em pesquisas futuras.

Para que seja possível o uso e reuso dos dados nesses ambientes, a representação da informação se configura como uma atividade essencial, tanto em relação ao tratamento descritivo (físico) como no tratamento temático (conteúdo) dos dados, haja vista que os repositórios de dados de pesquisa integram metadados que necessitam ser bem estruturados para atender à recuperação, uso e reuso de dados de pesquisa (Felipe & Santos, 2022).

Nessa seara, a presente investigação tem como objetivo geral apresentar um panorama geral dos Repositórios de Dados de Pesquisa brasileiros. Para tanto, objetiva-se especificamente: identificar o perfil dos repositórios de dados de pesquisa brasileiros; verificar como são realizadas as práticas de representação da informação nesses ambientes; averiguar as ferramentas que são utilizadas como infraestrutura para as práticas de representação da informação.

A relevância desta pesquisa decorre, em linhas gerais, da importância dos repositórios de dados de pesquisa brasileiros para o compartilhamento, uso e reuso dos dados. Considera-se necessário também o perfil desses ambientes, bem como identificar como são realizadas as práticas de organização e representação da informação, as quais podem viabilizar ou não a recuperação dos dados de pesquisa nesse contexto.

Para dar seguimento a essas considerações, a seguir são apresentados aspectos concernentes aos repositórios de dados de pesquisa.

Repositórios de dados de pesquisa

Para Pampel e Kindling (2017), os debates acerca de *data publishing*, estruturas para o compartilhamento dos dados, ganharam força nos últimos anos por causa do movimento Acesso Aberto. Estes autores dissertam que: "Para criar incentivos para os pesquisadores tornarem seus dados acessíveis, estratégias de publicação foram estabelecidas entre bibliotecários, editores e os próprios cientistas nos últimos anos, publicando estratégias que garantem o reconhecimento para quem disponibiliza os dados por outros pesquisadores"[3] (Pampel & Kindling, 2017, p. 18). Portanto, as publicações de dados surgem como alternativas para quem quer compartilhar os dados e para quem quer ter acesso a eles. Todas as suas estruturas são pensadas nas formas como os dados são desenvolvidos.

Assim, sendo surgem repositórios de dados de pesquisa, repositórios digitais cuja função é armazenar e disseminar dados de pesquisa, garantindo assim a conservação dos dados e permitindo o reuso por parte de quem o desejar. Os repositórios de dados podem estar ligados às instituições ou a grupos de pesquisa. Acerca dos repositórios de dados Sayão e Sales (2016, p. 96) declaram:

3 Texto original: "Um Forschenden Anreize zur Zug¨anglichmachung ihrer Daten zu schaffen, haben sich im Zusammenspiel von Wissenschaft, Bibliotheken und Verlagen in den vergangen Jahren Publikationsstrategien etabliert, die den Forschenden, die Forschungsdaten Dritten bereitstellen, entsprechende".

São infraestruturas de base de dados desenvolvidas para apoiar todo o ciclo da gestão de dados de pesquisa, incluindo as ações mais dinâmicas e contundentes sobre os dados, que coletivamente são chamadas de curadoria de dados de pesquisa, que visam adicionar valor aos dados, avaliando, formatando, agregando e derivando novos dados.

Nesse contexto, os repositórios podem auxiliar o pesquisador desde o início da pesquisa até o final, no qual ele decide compartilhar os dados. Além disso, garantem a preservação dos dados com o auxílio da curadoria, cuja prática pode trazer benefícios para quem compartilha e reutiliza os dados.

Por sua vez, o OpenAIRE (2018, para. 6, tradução nossa) define repositório de dados como "um arquivo digital que coleta e exibe conjuntos de dados e seus metadados[4]". A utilização dos metadados são fundamentais para a compreensão dos conjuntos de dados. Isso porque os dados podem ser planilhas, figuras, metodologia, algoritmos, espécies, áudios e outras formas que se diferem de documentos bibliográficos e assim, merecem atenção na descrição facilitando o seu reuso. Dessa forma, se faz necessário verificar como os repositórios de dados de pesquisa brasileiros estão sendo estruturados e remodelados com vistas a recuperação, acesso e uso dos dados de pesquisa.

Procedimentos metodológicos

Para atender o objetivo geral desta investigação, foram utilizadas como metodologia as pesquisas bibliográfica, documental, exploratória e descritiva com abordagem qualiquantitativa. A busca dos repositórios brasileiros foi realizada no buscador internacional Re3data em março de 2024, sistema que engloba repositórios de dados de pesquisa do mundo todo e das mais variadas disciplinas científicas.

Os indicadores observadores na pesquisa no âmbito mais geral foram: ano de criação, instituições aos quais os repositórios estão ligados, temáticas englobadas. Acerca da representação e disponibilização da informação os indicadores foram: apresentação de informações objetivas, formatos de registros bibliográficos, diretrizes para preenchimento de metadados, diretrizes para indexação e orientação para uso de vocabulários controlados.

A seguir serão apresentados os resultados como estabelecidos nos procedimentos metodológicos.

Resultados

Como resultados da busca, foram identificados 19 Repositórios de Dados de Dados de Pesquisa em âmbito nacional, todos eles com acesso aberto aos dados armazenados em sua ambiência. Desse total, 9 repositórios apresentam colaborações de países como Estados Unidos, Alemanha, Japão, China, Austrália e Índia.

Identificou-se ainda que os Repositórios de Dados de Pesquisa Brasileiros estão vinculados às seguintes instituições: Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT/UFRJ); Universidade Federal do Paraná (UFPR); Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio); Universidade Federal do Rio Grande do Sul (UFRGS); Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP); Universidade Estadual de Campinas (UNICAMP); Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP); Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA); Fundação Oswaldo Cruz (FIOCRUZ).

4 Texto original: "a digital archive collecting and displaying datasets and their metadata".

A tabela 1, mostra os nomes dos repositórios e as respectivas instituições em que se encontram vinculados.

Tabela 1: Instituições vinculadas aos repositórios de dados

Instituições vinculadas aos repositórios de dados	**Repositorios de dados vinculado**
Centro de Referência em Informação Ambiental	WorldClim - Global Climate Data
Biblioteca Eletrônica Científica Online	SciELO Data
Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES	Programa Internacional de Descoberta do Oceano
Instituto Nacional de Pesquisa Espacial, Projeto Global de Terras (Alemanha)	Globe
INCT Centro de Estudos Integrados da Biodiversidade Amazônica	Repositório de Dados do PPBio
FAPESP	Compartilhamento de dados FAPESP COVID-19/BR
Instituto Fleury	
Hospital Israelita Albert Einstein	
Hospital Sírio-Libanês	
Universidade de São Paulo	
Universidade Federal do Paraná	Base de Dados Científicos da Universidade Federal do Paraná
Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)	Aleia
Instituto Brasileiro de Informação em Ciência e Tecnologia, Rede Cariniana	Rede IBICT Cariniana Dataverse
Ministério da Ciência, Tecnologia e Inovações	
Empresa Brasileira de Pesquisa Agropecuária	REDAPE
Universidade de Campinas	REDU - Repositório de Dados de Pesquisa Institucional da Unicamp
Repositório de Dados de Pesquisas do Instituto Federal Goiano – Campus Urutaí	Dados Abertos De Pesquisas
Agência Nacional do Petróleo, Gás Natural e Biocombustíveis	Banco de dados de exploração e produção
Pontifícia Universidade Católica do Rio de Janeiro	Dados Abertos de Pesquisa @PUC-Rio
Canada Foudation For Innovation	Sons de peixe
Projeto Dataverso/ Fundação Oswaldo Cruz	Arca Dados
Universidade Estadual Paulista	Repositório Institucional da UNESP

Universidade Federal do Rio Grande do Sul	Repositório de Dados de Pesquisa CEDAP - dados de pesquisa
Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)	Depósito de dados

Fonte: Dados da pesquisa (2024).

A tabela 1 mostra que as mais variadas instituições possuem repositórios de dados, tais como instituições de pesquisa nacionais e internacionais, Universidades, hospitais, bibliotecas e fundações ligadas à ciência. Isso demonstra o interesse de instituições relacionadas à ciência e o seu compromisso na preservação e compartilhamento dos dados de pesquisa. Em relação ao período de criação desses repositórios, o gráfico 1 apresenta de maneira detalhada.

Gráfico 1. Data de criação dos repositórios no Brasil.

Fonte: Dados da pesquisa (2024).

Como se pode observar no gráfico acima, a criação dos repositórios no Brasil é recente, concentrando a criação deste em uma década 2013/2023, onde em 2013 foi criado o primeiro repositório, o Programa Internacional de Descoberta do Oceano. Os anos com alta incidência de produção de repositórios de da-

dos foram: 2016, 2020 e 2022, sendo esse último ano o período em que surgiram 4 repositórios de dados de pesquisa no Brasil. Isso demonstra que, mesmo recente, o surgimento dos repositórios de dados está em ascensão no Brasil, aspecto importante para o avanço da ciência no país.

Em relação às temáticas em que esses repositórios de dados de pesquisa se encontram associados, verificou-se uma variedade de temáticas, porém as áreas de maior destaque foram as ciências naturais, ciências da vida e ciências da engenharia, como pode ser observado na figura 1.

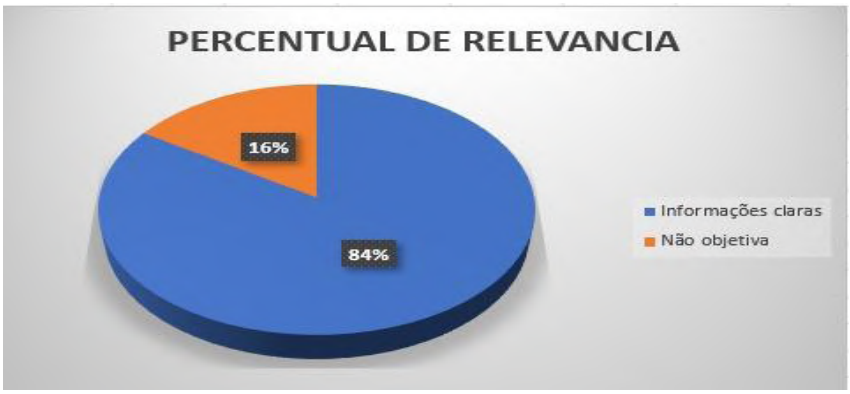
Figura 1. Áreas do conhecimento dos repositórios.

Fonte: Dados da pesquisa (2024).

Assim como os repositórios estão ligadas à variadas instituições, estes também englobam vários enfoques como as ciências naturais, ciências da vida, humanidades e ciência sociais, ciências das engenharias, ciências da computação, etc. Isso demonstra que no Brasil, várias áreas do conhecimento estão comprometidas com os aspectos dos dados abertos.

Como citado acima, não basta apenas o compartilhamento dos dados nos repositórios, estes devem estar organizados e representados, para fins do seu uso e reuso em pesquisas futuras. Para a representação dos dados nesses ambientes, faz-se necessário compreender o seu domínio. Com isso, verificou-se inicialmente se as informações relacionadas ao perfil e ao escopo dos repositórios se apresentam de maneira clara e bem definida.

Gráfico 2: Apresentação das informações.



Fonte: Dados da pesquisa (2024).

Apesar de 84,21% dos repositórios serem claros em seus objetivos (Gráfico 2), observa-se a falta de informações sobre os repositórios em seus sites, em alguns deles não foi possível encontrar a própria data de criação do repositório, se fazendo necessário obter informações em fontes externas, por exemplo. O Quadro 1 a seguir apresenta quais são os repositórios de dados de pesquisa que apresentam perfil e escopo claramente definidos:

Tabela 2: Repositórios de dados de pesquisa com perfil e escopo claramente definidos

Repositórios de dados de pesquisa
WorldClim
SciELO Data
Globo
Fapesp Covid-19 Data Sharing/Br
Aleia
Depósito de dados
Base de dados científicos da Universidade do Paraná
Rede IBICT Cariniana Dataverse

Repositório de Dados de Pesquisa CEDAP - dados de pesquisa
Programa Internacional de Descoberta do Oceano
Repositório Institucional Unesp
Arca
Sons de peixe
Dados de Pesquisa Aberta @Puc-Rio
REDU - Repositório de Dados de Pesquisa Institucional da Unicamp
Repositório de Dados do PPBio

Fonte: Dados da pesquisa (2024).

As informações identificadas nessas análises foram relacionadas ao escopo, tipos de dados armazenados e disponibilizados, temáticas encontradas, instituições parceiras. Também foram identificadas informações relacionadas à infraestrutura para a representação dos dados por meio da catalogação e indexação, em alguns dos repositórios listados anteriormente.

No contexto da Representação da informação, foram investigados os padrões de metadados que fornecem a infraestrutura para a representação dos dados. Sabe-se que os dados possuem vários formatos e devem ser descritos de formas variadas de acordo com sua tipologia, necessitando de infraestruturas e metadados específicos para cada tipo de dado. Assim, os formatos padrões utilizados pelos repositórios e identificados nesta pesquisa são apresentados na figura 2.

Figura 2. Padrões de metadados.

Fonte: Dados da pesquisa (2024).

Os padrões de metadados em destaque encontrados são o Dublin core, DDI (*Data Documentation Initiative*) e *Data Cite*. O *Dublin core* surge como um padrão voltado para representação de objetos digitais e é utilizado na representação de dados abertos. Já o *Data Cite* é um padrão voltado para o compartilhamento de dados, porém pode ser aplicado em qualquer domínio. Por sua vez, o DDI é um padrão de metadados

voltado para dados de pesquisa em ciências sociais, comportamentais e econômicas. Outro padrão cujo objetivo é um domínio específico presente nos repositórios é o *Darwin Core*, que é específico para o domínio da Biodiversidade. Pode se notar ainda que nenhum desses padrões utilizados foram desenvolvidos no Brasil.

Em alguns repositórios, o próprio pesquisador deve fazer o depósito dos seus dados. Com base nesse cenário, foi observado se os repositórios apresentam orientações acerca do preenchimento dos metadados para os pesquisadores.

Gráfico 3. Diretrizes para preenchimento dos metadados.

Fonte: Dados da pesquisa (2024).

O gráfico 3 demonstra que 52,63% dos repositórios apresentam orientações sobre o preenchimento dos metadados para o pesquisador depositar os seus dados de pesquisa. Cabe frisar que, sem essas orientações, o pesquisador pode encontrar dificuldade em organizar e descrever os seus dados para fins de recuperação, acesso e reuso por outras pessoas. Os repositórios que apresentam essas orientações são: SciELO Data, Programa Internacional de Descoberta do Oceano, Globe, Repositório de Dados do PPBio, Universidade Federal do Paraná, Aleia, Rede IBICT Cariniana Dataverse, REDAPE, Arca dados, Dados Abertos de Pesquisas.

Também foi investigado aspectos ligados à indexação, como orientação para indexação, uso de vocabulários controlados e orientação para a utilização de vocabulários controlados nos repositórios analisados. O gráfico 4, acerca de diretrizes sobre indexação.

Gráfico 4. Diretrizes para a indexação.

Fonte: Dados da pesquisa (2024).

Conforme foi possível visualizar no gráfico 4, apenas dois repositórios apresentaram diretrizes para indexação, foram eles: o *WorldClim - Global Climate Data* e Repositório de Dados do PPBio. Esses dois repositórios também apresentam orientações acerca do uso de vocabulários controlados, juntamente com o Programa Internacional de Descoberta do Oceano. Esses três repositórios recomendam a utilização de quatro vocabulários controlados para a indexação, a saber: itution_id CMIP6, IODP *Depth Scales Terminology*, *Thesaurus of Geographic Names* (TGN) e o vocabulário da Ppbio, sendo esse último desenvolvido pela própria instituição mantenedora do repositório. Pode-se perceber que assim como alguns padrões de metadados são específicos para as temáticas dos repositórios, os vocabulários encontrados também são, o que pode auxiliar no processo de indexação e recuperação dos dados de pesquisa.

Em linhas gerais, foi possível perceber que os repositórios de dados de pesquisa brasileiro contém infraestrutura adequada para a representação descritiva dos dados a partir dos padrões de metadados, porém no que concerne à indexação dos dados esses ambientes ainda estão em desenvolvimento. A presença de diretrizes e instrumentos para as práticas de indexação, bem como de preenchimento dos metadados nos repositórios permite não só a descrição dos dados, mas também pode auxiliar na integração e na interoperabilidade entre sistemas.

Considerações finais

Os repositórios de dados de pesquisa no Brasil são fontes de informação recentes, surgindo na última década e ainda são poucos quando comparados a países do hemisfério norte. Portanto, considera-se que os repositórios de dados de pesquisa brasileiro ainda se encontram em fase de desenvolvimento, embora esteja em constante crescimento em número no país. Não foi possível encontrar repositórios de todas as disciplinas do conhecimento, mas grandes áreas como ciências da vida e humanas começam a disseminar os dados e permitir o reuso por parte de quem interessar.

Diante disso, observa-se a importância de outras áreas do conhecimento, em integração com a Ciência da Informação, se dedicar aos processos de modelagem, criação, preservação, representação, recuperação e acesso de dados de pesquisa.

Os enfoques relacionados à representação da informação foram analisados na presente pesquisa. No que concerne à representação descritiva, verifica-se que esses repositórios se encontram bem desenvolvidos, porém contata-se que os administradores dos repositórios devem se dedicar a aperfeiçoar as práticas de indexação nesses cenários.

Por fim, conclui-se que a gestão dos repositórios de dados não cabe somente aos profissionais da Ciência da Computação, mas que deve ser desenvolvida em consonância com o profissional Bibliotecário, com domínio de processos, produtos, instrumentos e serviços da organização e representação da informação e do conhecimento para fins de recuperação da informação.

Bibliografía

Costa, M. P. da. (2017). *Fatores que influenciam a comunicação de dados de pesquisa sobre o vírus da zika, na perspectiva de pesquisadores*. (Tese de Doutorado). Universidade de Brasília, Brasília, DF. doi: http://dx.doi.org/10.26512/2017.02.T.23000

Felipe, C. B. M., & Santos, R. F. dos. (2022, Julho-Setembro). Avaliação de metadados em Repositórios de Dados de Pesquisa sobre biodiversidade. *Em Questão (Porto Alegre)*, 28(3), 1-19. doi: https://doi.org/10.19132/1808-5245283.117591

OPENAIRE. *What are repositories?* (2018). Recuperado de https://www.openaire.eu/where-can-i-read-more-about-fp7

Pampel, H., & Kindling, M. (2017). Informationsinfrastrukturangebote fur digitale Forschungsdaten. In B. Jacob, M. Kindling, & U. Müller (Eds.) (2017). *Peter Schirmbacher sei Dank E(hren)-Journal* (pp. 15-33). Berlin: Humboldt-Universität zu Berlin, Philosophische Fakultät I, Institut für Bibliotheks- und Informationswissenschaft. Recuperado de https://edoc.hu-berlin.de/bitstream/handle/18452/2993/3.pdf

Sayão, L. F., & Sales, L. F. (2016, maio-agosto). Algumas considerações sobre os repositórios digitais de dados de pesquisa. *Informação & Informação (Londrina), 21*(2),90-115. Doi: https://doi.org/10.5433/1981-8920.2016v21n2p90

Tecnologias livres utilizadas para construção de Repositórios e Bibliotecas Digitais no Brasil

Diego José Macêdo[1], Ingrid Torres Schiessl[2], Mirele Carolina Souza Ferreira Costa[3], Lucas Ângelo Silveira[4], Fernando de Jesus Pereira[5], Elton Mártires Pinto[6], Milton Shintaku[7]

Palabras claves

Repositório institucionais. Bibliotecas digitais. Software livre. DSpace.

Institutional repositories. Digital libraries. Open-source software.

Eje temático

Infraestructura tecnológica

Resumen

O uso de repositórios tem se tornado cada vez mais comum para gerenciar objetos digitais. Entretanto, nem sempre se tem a noção de quais ferramentas podem ser utilizadas para criação desses sistemas de informação, nos seus diversos contextos, requerendo estudos para contribuir com as suas pesquisas. Por isso, o estudo tem o objetivo de levantar as ferramentas e áreas do conhecimento que utilizam repositórios no Brasil para gerir objetos digitais. Para tanto, faz uso de levantamento bibliográfico com análise cientométrica. Os resultados revelam que o termo repositório ganha uma acepção maior, sendo utilizado em várias áreas de aplicação, da computação a arquivologia, dando ao termo uma abrangência conceitual maior, com uma grande variedade de ferramentas.

1 Introdução

Historicamente as bibliotecas digitais tiveram um grande desenvolvimento no final da década de 1990, por causa do movimento dos arquivos abertos, para disseminar a literatura cinzenta. Os repositórios, por sua vez, aparecem com maior força no movimento do Acesso Aberto, chamados de via verde, para dar acesso à literatura branca de forma livre, antes publicadas em canais restritos. Assim, na sua gênese, as bibliotecas digitais foram criadas para serem primeira fonte de literatura cinzenta (Fox, Eaton e McMIllan, 1996) e repositórios para segunda fonte de literatura branca (Weitzel, 2006). Entretanto, como relatam Shintaku e Vidotti (2016), repositórios e bibliotecas digitais estão cada vez mais atuando como publicadores dos mais diversos tipos de documentos, incluindo os denominados técnicos.

1 Instituto Brasileiro de Informação em Ciência e Tecnologia, diegomacedo@ibict.br
2 Instituto Brasileiro de Informação em Ciência e Tecnologia, ingridschiessl@ibict.br
3 Instituto Brasileiro de Informação em Ciência e Tecnologia, mirelecosta@ibict.br
4 Instituto Brasileiro de Informação em Ciência e Tecnologia, lucasangelo@ibict.br
5 Instituto Brasileiro de Informação em Ciência e Tecnologia, fernandopereira@ibict.br
6 Instituto Brasileiro de Informação em Ciência e Tecnologia, eltonpinto@ibict.br
7 Instituto Brasileiro de Informação em Ciência e Tecnologia, shintaku@ibict.br

No Brasil, grande parte dos repositórios e bibliotecas digitais foram criados com o software DSpace, em parte, pela ação do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), que disseminou essa tecnologia. Esse apoio institucional ao DSpace foi resultado de estudos comparativos entre essa tecnologia e o EPrints, outro software livre voltado a criação de repositórios e bibliotecas digitais, desenvolvido pela Universidade de Southampton, Estados Unidos, no qual o DSpace se mostrou tecnicamente melhor.

Assim, sabe-se que há uma grande quantidade de ferramentas voltadas para criação de repositórios, muitos dos quais são desenvolvidos para esse fim ou adaptados, mesmo que grande parte desses sistemas utilizem o DSpace. Entretanto, com a evolução, o próprio termo repositório tem sido alterado, da semântica inicial, de ser um lugar onde se repõe documentos já publicados, sendo utilizados em diversas disciplinas, como no caso da ciência da computação para iniciativas para distribuição de códigos fontes (Majumdar, Jain e Barthwal, 2017).

Nesse contexto, o presente estudo tem o objetivo de levantar as ferramentas e áreas do conhecimento que utilizam repositórios no Brasil para gerir objetos digitais. Para tanto, entende-se que objeto digital carrega uma definição complexa, como apresentada por Yamaoka e Gauthier (2013), mas que em termo mais simplista, pode ser compreendida como uma construção digital de uma representação do conhecimento, nas suas mais diversas formas. Uma inscrição de sinais binários registrados, que podem ser armazenados em mídias ou circuitos eletrônicos.

2 Repositórios e Bibliotecas Digitais no Brasil

Em 1991 a Web foi criada e, com isso, a possibilidade de criação de sites hipertextuais e a disponibilização de documentos em formato digital, a ponto de que seu criador Tim Berners-Lee defender que a Web ser uma revolução, uma forma de democratização em si própria (Berners-Lee, 2010). Na esteira da web, surgiram diversos movimentos na ciência, como os portais científicos, blogs de pesquisadores e tantas outras iniciativas, muitos em consonância a movimentos de abertura da ciência. Nesse caminho, o surgimento do Movimento dos Arquivos Abertos (Open Archives) impactou muito a ciência.

Sompel e Lagoze (2000) relatam sobre a Convenção de Santa Fé, Novo México nos Estados Unidos da América, e o surgimento desse movimento, muito do qual, foi inspirado no lançamento da iniciativa do ArXives[8], que possibilita a publicação de pré-prints pelo próprio autor, uma grande inovação na época. Os autores elencaram algumas das premissas dos arquivos abertos que foram incluídos nos softwares de biblioteca digital, ofertando novas possibilidades para a ciência, sendo:

- Depósito e gestão de artigos em formato pré-print (originalmente usavam o termo arquivar)
- Sistema de provedores de dados e serviços (ferramentas básicas da interoperabilidade)
- Disponibilização do texto completo com o uso de metadados para descrição completos
- Padrão aberto de metadados
- Uso de protocolos de comunicação

8 Disponível em: https://arxiv.org/. Acesso em: 25 abr. 2024.

No Brasil, Triska e Café relatam sobre o projeto Biblioteca Digital Brasileira (BDB) que implementam os preceitos dos Arquivos Abertos (2001), no âmbito do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), de forma a fomentar a criação de bibliotecas digitais. Possivelmente o maior dos resultados desse projeto foi a criação da Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), tanto o agregador, quanto as primeiras bibliotecas digitais locais, desenvolvidas. Carvalho Segundo et al (2015) e Santos e Souza (2017) relatam sobre a ação do Ibict com a distribuição da Ferramenta Teses E Dissertações Eletrônicas (TEDE) para criação local de bibliotecas digitais, que posteriormente teriam seus metadados coletados pela BDTD.

Destaca-se que na terminologia do movimento, usa-se Biblioteca Digital e seu foco ficou na disseminação de pré-prints, teses e dissertações, entre outros tipos de documentos, considerados como literatura cinzenta e memória técnica. Nesse contexto, Basevi (2005) descreve o uso desses preceitos para a criação da Biblioteca Digital Jurídica (BDJur), vinculado ao Superior Tribunal de Justiça (STJ), para disseminar a produção intelectual do tribunal, tornando-se um modelo para outros órgãos de governo. Tanto que, Macêdo, Shintaku e Brito (2015) relatam que vários órgãos passaram a criar as suas bibliotecas digitais, grande parte de tribunais, seguindo os passos da BDJUR.

Os repositórios, por sua vez, foram alavancados pelo Movimento de Acesso Aberto (Open Access), considerado por Hanard et al (2008) como via verde (green road) para disseminação de publicações em texto completo, publicados anteriormente de forma restrita, ou seja, dá o sinal verde ao acesso aos conteúdos publicados restritamente. Para os autores, os repositórios e as revistas científicas de acesso aberto (via dourada) formam os principais sistemas de disseminação do movimento. Assim, repositórios em sua base no acesso aberto é, como defende Björk (2007) um facilitador de acesso a publicações científicas de forma livre.

Se os Arquivos Abertos tiveram a Convenção de Santa Fé como marco, o Movimento do Acesso Aberto teve a publicação do Budapest Open Access Initiative (BOAI), cujos princípios foram sintetizados por Suber (2002), acalmando a comunidade científica, quanto as desconfianças, com:

- Manter a revisão pelos pares
- fomentar a qualidade profissional dos periódicos
- Aumentar o prestígio das revistas
- Preservar o acesso ao conhecimento
- Compatibilidade com os direitos autorais
- Lucros condizentes com os gastos da revista
- Possibilidade de venda de edições melhoradas
- Preferência pela publicação online

No Brasil, novamente o Ibict capitaneou a implementação de repositórios institucionais em várias instituições de ensino e pesquisa, por meio de um projeto financiado pela Financiadora de Estudos e Projetos (Finep), que distribuiu kits tecnológicos utilizando o edital FINEP/PCAL/XBDB, baseados na tecnologia DSpace. Com isso, o instituto fomentou a criação de repositórios institucionais, de forma a alimentar um

grande agregador nacional de publicações em acesso aberto, o Portal brasileiro de publicações e dados científicos em acesso aberto (OasisBr). Assim, grande parte dos repositórios de universidades federais são oriundos desse projeto.

Nessa trajetória, os repositórios nascem para serem segunda fonte, sendo as revistas a primeira e os agregadores a terceira (Weitzel, 2006) . Entretanto, no Brasil, grande parte das revistas são publicadas com o acesso aberto (dourado ou diamante), desde a criação do Scientific Electronic Library Online (SciELO), que iniciou o movimento de acesso aberto no Brasil com as revistas científicas na área de saúde e depois expandiu para todas as disciplinas. Assim, os repositórios passaram a disseminar outros tipos de documentos de primeira fonte (Shintaku e Vidotti, 2016).

Nesse contexto, no Brasil há uma grande quantidade de repositórios em instituições de ensino e pesquisa, para disseminar documentação científica e técnica, e bibliotecas digitais em órgãos de governo para disseminar a memória técnica, em grande parte utilizado o software DSpace. Entretanto, nota-se que o uso do termo repositório está cada vez mais sendo utilizado em outros contextos, com o uso de muitas ferramentas distintas e propósitos diversos.

3 Metodologia

Conforme o objetivo deste trabalho, de levantar as ferramentas e áreas do conhecimento que utilizam repositórios no Brasil para gerir objetos digitais, a pesquisa tem características exploratórias, de forma a dar maior familiaridade com o fenômeno estudado (Gil, 2008). Com isso, o estudo tem aspectos totalmente qualitativos, buscando atender aos propósitos do estudo.

Com isso, o método de levantamento dos dados se deu por meio da pesquisa bibliográfica, apontada pelo mesmo autor, como aquela que busca informações em fontes acadêmicas. Assim, a busca foi feita no google acadêmico, em língua portuguesa, por se restringir ao contexto brasileiro, com o argumento de busca "repositório" e "software" nas palavras chave, com o operador booleano E (AND), visto a grande quantidade de documentos indexados no texto completo com esse termo.

A escolha da fonte deu-se, pois, conforme Jacsó (2005), Harzing (2017b) e Zientek et al (2018), o Google Acadêmico indexa grande parte das fontes científicas mundiais. No Brasil o Google Acadêmico indexa a BDTD, com as teses e dissertações defendidas no Brasil, o OasisBr, SciELO e Brapci com a maioria das revistas de acesso aberto, para a coleta no Google Acadêmico utilizou-se o software Publish or Perish (PoP) (Harzing, 2024), tal ferramenta é indicada para coleta no Google Acadêmico como afirmam Jacsó (2009), Harzing (2016, 2017a).

O levantamento bibliográfico gerou uma base de dados, exportada em formato Comma-separated Values (CSV), que foi processado por meio da linguagem Python para processar as informações demográficas e também realizar a coleta dos tipos de documentos, que não foram recuperados na busca. Em relação a coleta das informações sobre os tipos de documento, foi desenvolvido o código que buscava a partir do metadado ArticleURL o valor do metadado dc.type. conforme demonstrado no Figura 1.

Figura 1 - Código para buscar o valor cadastrado no metadado dc.type

```
# Carregar o arquivo CSV em um DataFrame
df = pd.read_csv('Biredial - Dados tratados v2.csv')

# Definir a função para buscar o tipo de documento
def get_document_type(url):
    # Ajustando a URL para incluir 'mode=full' de forma condicional
    if '?' in url:
        full_url = url + "&mode=full"
    else:
        full_url = url + "?mode=full"

    try:
        response = requests.get(full_url, timeout=30)  # Fazendo a requisição para a URL ajustada com timeout
        soup = BeautifulSoup(response.text, 'html.parser')

        # Verificar se o metadado 'dc.type' existe de outra forma no HTML
        meta_tag = soup.find('meta', attrs={'name': 'DC.type'})
        if not meta_tag:  # Se não encontrado nos meta tags, tentar buscar no conteúdo da página
            meta_tag = soup.find('div', text=lambda text: text and 'dc.type' in text.lower())

        if meta_tag:
            doc_type = meta_tag['content'] if 'content' in meta_tag.attrs else meta_tag.text
            doc_type = doc_type.split('/')[-1].strip().upper()  # Normalizando a saída
        else:
            doc_type = 'DOCUMENT TYPE NOT FOUND'  # Valor padrão caso não encontre

        print(doc_type)
        return doc_type
    except Exception as e:
        print(f"Erro ao obter o tipo de documento para {url}: {e}")
        return None

# Aplicar a função para obter o tipo de documento para cada URL na coluna 'ArticleURL'
df['DocumentType'] = df['ArticleURL'].apply(get_document_type)
```

Fonte: Elaboração dos autores (2024).

Dos tipos de documentos encontrados, identificou-se termos duplicados e em outros idiomas, por exemplo: TRABALHO DE CONCLUSÃO DE CURSO - GRADUAÇÃO - BACHARELADO; TRABALHO DE CONCLUSÃO DE CURSO; TCC e BACHELORTHESIS. Foram avaliados todos os casos para consolidar em uma única terminologia.

Nesse sentido, pode-se ter uma visão geral das publicações, por meio do uso de variáveis cientométricas, para entendimento da posição do tema nas pesquisas. Posteriormente, selecionou os artigos que tratavam de ferramentas distintas para criação de repositórios, de forma a amparar o estudo, por meio da análise dos resumos. com o corpus dos artigos selecionados, deu-se a leitura dos textos completos.

4 Resultados

Os resultados iniciais apresentaram 857 trabalhos, com busca efetuada em abril de 2024, com publicação entre 2004 a 2024. Entretanto, como relatado por Ceni Coelho (2018), há problemas nos metadados ou na indexação do Google Acadêmico que interferem no resultado da busca. Por isso, 36 registros foram descartados: 20 registros foram considerados "erros de recuperação" por serem documentos que não atendem aos critérios de busca definidos (argumento de busca "repositório" e "software" nas palavras-chave) e 16 correspondem a registros duplicados. Assim, o corpus final consistiu em 820 registros.

Para formação da base de dados foi preciso atuação manual, pois alguns campos vieram sem dados, como o ano de publicação (74 registros) e tipo de documentos (todos os registros) que vieram sem as informações. Assim, foi preciso fazer buscas em sites das revistas, congressos e curriculum lattes dos autores para certificar-se da data com precisão. Em alguns poucos casos, nos trabalhos que apresentavam citação, olhou-se os trabalhos que citam as obras, nas referências, para recuperar informações. No caso específico dos tipos de documentos, a partir do código em Python elaborado, apenas 254 registros não foi possível identificar o tipo de documento, portanto, foi verificado manualmente por meio do acesso aos websites dos trabalhos. Entre a tipologia documental, recuperou-se 35 tipologias, após análise resultou em 20 tipologias, como apresentado no Gráfico 2.

Quanto a série histórica de publicações por ano, nota-se que o interesse pelo tema tem crescido com o tempo (Gráfico 1). O ano de 2023 teve o maior número de publicações, com destaque para os trabalhos apresentados em eventos, com 36 publicações, sendo nove no XXII Seminário Nacional de Bibliotecas Universitárias (SNBU). Em 2004, primeiro ano da coleta das seis publicações, quatro eram de eventos, com destaque ao autor Eloy Rodrigues, da Universidade do Minho, cujo trabalho de 2004 possui 47 citações identificadas pelo google acadêmico (trabalho mais citado da coleta).

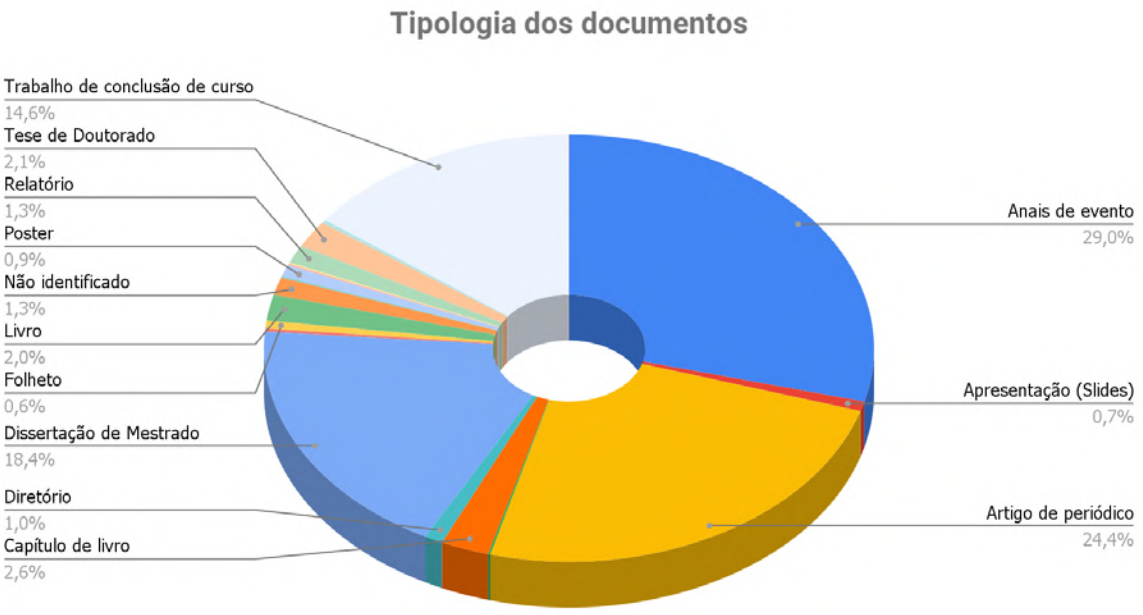
Gráfico 1 - Quantidade de publicações sobre o tema por ano

Fonte: Elaboração dos autores (2024).

Séries históricas de quantidade de publicação servem, entre outros pontos, para verificar o crescimento de interesse sobre o tema, e, como defendem Camargo e Barbosa (2018) são indicadores de atividades, expressando a produtividade e evolução da área. Os dados de 2024 ainda estão incompletos, pois a busca e tratamento dos dados foi feita em abril de 2024. No entanto, mesmo com apenas quatro meses, já foram levantados seis trabalhos, o mesmo de 2004 inteiro, com três artigos de periódicos.

Quanto a tipologia documental levantada, ratificou-se certa preferência pelos anais de eventos, com cerca de 29% dos trabalhos (Gráfico 2). Sabe-se, no entanto, que muitos dos trabalhos apresentados em eventos são decorrentes de estudos de pós-graduação, sendo uma publicação intermediária, como afirma Lievrouw (2009), em seu modelo de comunicação científica. Possivelmente por isso, certa equivalência entre dissertações e teses, trabalhos em anais de eventos e artigos de periódicos.

Gráfico 2 - Tipologia de documentos

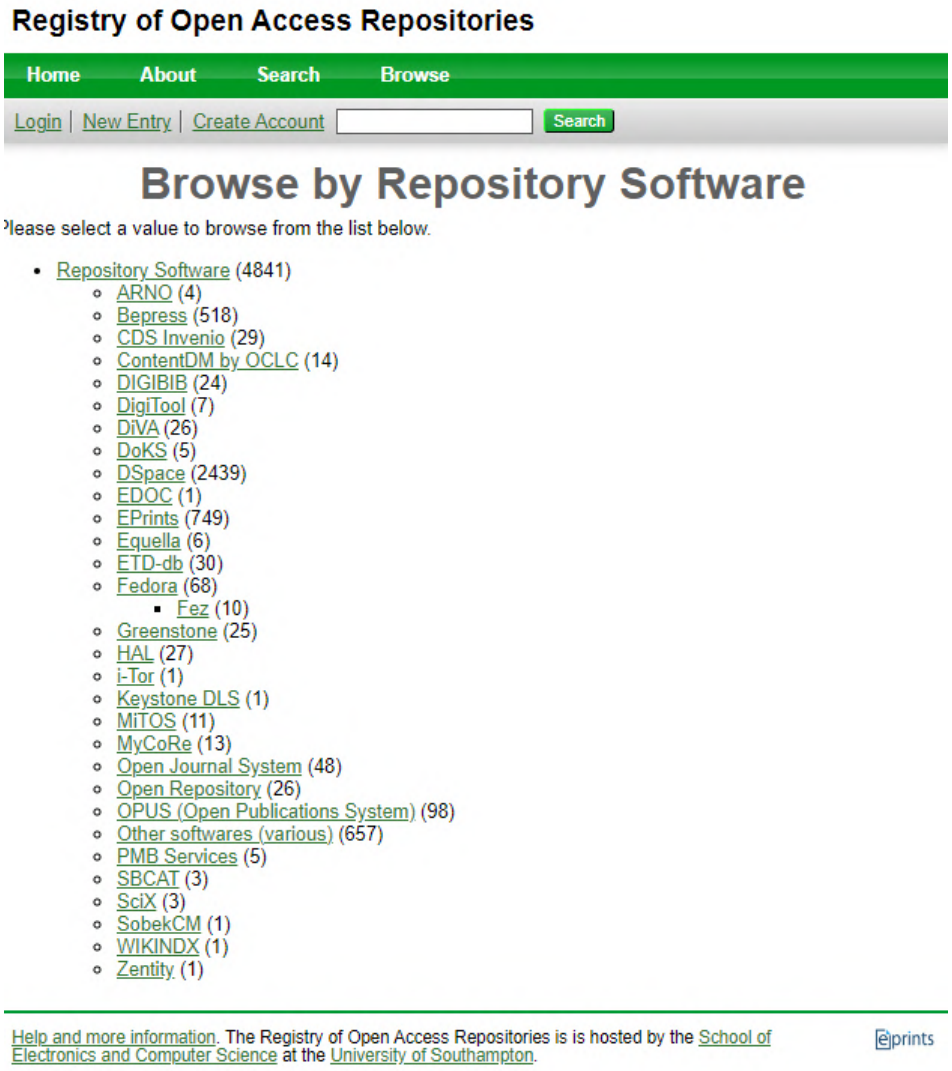


Fonte: Elaboração dos autores (2024).

De certa forma, a grande presença de artigos de eventos e periódicos, mostra certa diversificação nos canais de publicação. Da mesma forma, ratifica o embranquecimento das teses e dissertações, antes considerados literatura cinzenta pela dificuldade de acesso ao texto completo (não publicação e disseminação tradicionalmente). A indexação da Biblioteca Brasileira de Teses e Dissertações (BDTD) pelo google acadêmico deu maior visibilidade a essa literatura, ratificando o que Shintaku e Vidotti (2016) falam sobre a importância dos repositórios e bibliotecas digitais na disseminação de primeira fonte.

Quanto às tecnologias relatadas nos trabalhos, destaca-se o DSpace, em grande parte pela atuação do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) em repasse de conhecimentos. Também, pelo uso do DSpace em órgãos de governo, extrapolando a área acadêmica, como identificado por Macedo, Shintaku e Brito (2015). Esse ponto confirma os indicadores de uso do DSpace como sendo a ferramenta mais utilizada no mundo para criação de repositórios, tal indicador também é confirmado no website Registry of Open Access Repositories (ROAR), que apresenta uma grande quantidade de repositórios que utilizam tal software.

Figura 2 - Website do ROAR com a opção de pesquisa por tipo de software para repositórios



Fonte: Captura de tela (2024).

No entanto, uma nova proposição está em grande voga pelas publicações com um novo termo com base em repositórios, o chamado "repositório digital confiável" de caráter arquivístico, voltado à preservação digital (Santos & Flores, 2015). Em parte, utiliza-se tecnologias como o Access To Memory (AToM) e Archivematica para implementar esses repositórios. Não apenas na preservação arquivística esses repo-

sitórios são utilizados, Sayão (2010) relata sobre a preservação de periódicos eletrônicos a longo prazo. A questão da preservação digital com uso de repositórios digitais confiáveis provocou a criação de um modelo complexo de apoio, denominado de "Modelo Hipátia", desenvolvido pelo Ibict (Braga, 2022).

Outras ferramentas são apresentadas em menor número, como o software Omeka, que pode ser utilizado para criação de repositórios em galerias, bibliotecas, arquivos e museus (Shintaku et al., 2018). Da mesma forma, outras opções com o Open Monograph Press (OMP), criado para criação de portais de editoras digitais, voltado para gerenciamento de E-books, foram utilizados para a criação de bibliotecas digitais (Soares, 2019).

Fora da gestão documental, repositórios de códigos fontes são extremamente comuns na ciência da computação, principalmente no desenvolvimento de softwares livres. Ferramentas baseadas em GIT (github e gitlab). Essas ferramentas têm grandes similaridades com os gestores de documentação, com a diferença de atuarem com códigos fontes e possibilitarem o versionamento controlado. Entretanto, em todos os casos são ferramentas voltadas à gestão de objetos digitais.

5 Considerações Finais

Este estudo forneceu um panorama sobre a evolução e utilização de repositórios e bibliotecas digitais no Brasil, destacando a predominância do software DSpace, fomentado pelo Ibict. A análise demonstrou a diversificação na aplicação de repositórios digitais, não apenas na academia mas também em órgãos governamentais, refletindo uma ampla adoção dessas tecnologias para a gestão de objetos digitais. A pesquisa também revelou o aumento significativo de publicações sobre o tema, indicando um crescente interesse e reconhecimento da importância dos repositórios digitais na disseminação do conhecimento e na preservação digital.

Os dados coletados apontam para uma evolução do termo "repositório", que agora abrange uma variedade de ferramentas e contextos, incluindo repositórios digitais confiáveis para preservação arquivística. Este fenômeno reflete uma mudança na percepção e nos requisitos para a manutenção e acesso ao conhecimento científico e técnico. A entrada de novas tecnologias, como Omeka e OMP, sugere uma futura expansão nos tipos de conteúdos gerenciados e nas plataformas utilizadas.

Por fim, a pesquisa indica a necessidade de continuidade nos estudos sobre o impacto dessas tecnologias no desenvolvimento dos repositórios, de forma a compreender as transformações e desafios associados à gestão de informações digitais. A expansão das capacidades e a adaptação às novas demandas de preservação digital serão essenciais para o desenvolvimento futuro de repositórios e bibliotecas digitais no Brasil e no mundo.

Bibliografía

Basevi, T. (2005). BDJur Consortium: Juridical Digital Library: Implementing DSpace in the Brazilian Judiciary. Proceedings, 127–132. Leuven-Heverlee, Bélgica: Peeters Publishing Leuven. Recuperado de https://elpub.architexturez.net/doc/oai-elpub-id-150elpub2005

Berners-Lee, T. (2010). Long live the web: A call for continued Open Standards and neutrality. Scientific American, p. [online]. Recuperado de https://www.scientificamerican.com/article/long-live-the-web/

Björk, B.-C. (2007). A model of scientific communication as a global distributed information system. Information Research, 12(2). Recuperado de http://informationr.net/ir/12-2/paper307.html

Braga, T. E. N. (2022). O modelo Hipátia: A proposta do Ibict para a preservação digital arquivística. Em T. E. N. Braga & M. Á. Márdero Arellano (Orgs.), Hipátia: Modelo de preservação para repositórios arquivísticos digitais confiáveis (p. 52–65). Brasília: Ibict. doi: 10.22477/9786589167501.cap4

Camargo, L. S. de, & Barbosa, R. R. (2018). Bibliometria, cienciometria e um possível caminho para a construção de indicadores e mapas da produção científica. PontodeAcesso, 12(3), 109–125. doi: 10.9771/rpa.v12i3.28408

Carvalho Segundo, W. L. R. de, Shintaku, M., Oliveira, A. C. de L., & Assis, T. B. de A. (2015). Sistema de publicação eletrônica de teses e dissertações (TEDE): Instalação, migração e configuração. Brasília: Ibict. Recuperado de http://livroaberto.ibict.br/handle/1/1059

Coelho, G. C. (2018). Avaliação de impacto de periódicos brasileiros de extensão universitária. Biblios Journal of Librarianship and Information Science, (71), 81–89. doi: 10.5195/biblios.2018.468

Fox, E. A., Eaton, J. L., McMillan, G., Kipp, N. A., Weiss, L., Arce, E., & Guyer, S. (1996). National Digital Library of Theses and Dissertations. D-lib Magazine. http://www.dlib.org/dlib/september96/theses/09fox.html

Gil, A. C. (2008). Como elaborar projetos de pesquisa (6º ed). São Paulo: Atlas.

Harnad, S., Brody, T., Vallières, F., Carr, L., Hitchcock, S., Gingras, Y., … Hilf, E. R. (2008). The access/impact problem and the green and gold roads to Open Access: An update. Serials Review, 34(1), 36–40. doi: 10.1080/00987913.2008.10765150

Harzing, A.-W. (2016, fevereiro 6). Publish or Perish. Recuperado 25 de abril de 2024, de Harzing.com website: https://harzing.com/resources/publish-or-perish

Harzing, A.-W. (2017a, fevereiro 7). Using Publish or Perish to do a literature review. Recuperado 25 de abril de 2024, de Harzing.com website: https://harzing.com/blog/2017/02/using-publish-or-perish-to-do-a-literature-review

Harzing, A.-W. (2017b, fevereiro 28). Google Scholar is a serious alternative to Web of Science. Recuperado 24 de fevereiro de 2023, de Harzing.com website: https://harzing.com/blog/2017/02/google-scholar-is-a-serious-alternative-to-web-of-science

Harzing, A.-W. (2024). Publish or Perish [Windows]. Recuperado de https://harzing.com/resources/publish-or-perish

Jacsó, P. (2005). Google Scholar: The pros and the cons. Online Information Review, 29(2), 208–214. doi: 10.1108/14684520510598066

Jacsó, P. (2009). Calculating the h-index and other bibliometric and scientometric indicators from Google Scholar with the Publish or Perish software. Online Information Review, 33(6), 1189–1200. doi: 10.1108/14684520911011070

Lievrouw, L. A. (2009). New media, mediation, and communication study. Information, Communication & Society, 12(3), 303–325. doi: 10.1080/13691180802660651

Macêdo, D. J., Shintaku, M., & Brito, R. F. de. (2015). Dublin Core usage for describing documents in Brazilian Government Digital Libraries. Anais, 129–135. São Paulo: DCMI. Recuperado de https://dcpapers.dublincore.org/pubs/article/view/3768

Majumdar, R., Jain, R., Barthwal, S., & Choudhary, C. (2017). Source code management using version control system. Proceedings, 278–281. https://doi.org/10.1109/ICRITO.2017.8342438

Santos, H. M. dos, & Flores, D. (2015). Repositórios digitais confiáveis para documentos arquivísticos: Ponderações sobre a preservação em longo prazo. Perspectivas em Ciência da Informação, 20(2), 198–218. doi: 10.1590/1981-5344/2341

Santos, K. G. dos, & Souza, L. G. S. (2017). A importância do Ibict para a divulgação científica brasileira. Bibliotecas Universitárias: pesquisas, experiências e perspectivas, 3(2), 3–18. Recuperado de https://periodicos.ufmg.br/index.php/revistarbu/article/view/3092

Sayão, L. F. (2010). Repositórios Digitais Confiáveis para a Preservação de Periódicos Eletrônicos Científicos. UFBA. Recuperado de http://eprints.rclis.org/15903/

Shintaku, M., Gomes, R. F., Brito, R., Costa, L., Pereira, V. C., & Oliveira, K. S. de. (2018). Guia do usuário do Omeka. Brasília: Ibict. Recuperado de http://ridi.ibict.br/handle/123456789/1157

Shintaku, M., & Vidotti, S. A. B. G. (2016). Bibliotecas e repositórios no processo de publicação digital. Biblos, 30(1), 60–79. Recuperado de https://periodicos.furg.br/biblos/article/view/5762

Sompel, H. V. de, & Lagoze, C. (2000). The Santa Fe Convention of the Open Archives Initiative. D-Lib Magazine, 6(2). doi: 10.1045/february2000-vandesompel-oai

Suber, P. (2002). Open access to the scientific journal literature. Journal of Biology, 1(1), 3. doi: 10.1186/1475-4924-1-3

Triska, R., & Café, L. (2001). Arquivos abertos: Subprojeto da Biblioteca Digital Brasileira. Ciência da Informação, 30(3), 92–96. doi: 10.18225/ci.inf.v30i3.917

Weitzel, S. da R. (2006). O papel dos repositórios institucionais e temáticos na estrutura da produção científica. Em questão, 12(1), 51–71. Recuperado de https://seer.ufrgs.br/EmQuestao/article/view/19

Yamaoka, E. J., & Gauthier, F. O. (2013). Objetos digitais: Em busca da precisão conceitual. Informação & Informação, 18(2), 77–97. doi: 10.5433/1981-8920.2013v18n2p77

Zientek, L. R., Werner, J. M., Campuzano, M. V., & Nimon, K. (2018). The use of Google Scholar for research and research dissemination. New Horizons in Adult Education and Human Resource Development, 30(1), 39–46. doi: 10.1002/nha3.20209

Diego José Macêdo

http://lattes.cnpq.br/2205539000237712

https://orcid.org/0000-0002-5696-0639

Mestre em Ciência da Informação pela Universidade de Brasília. Bacharel em Sistema de Informação pela Universidade Católica de Brasília. Atualmente é tecnologista do Instituto Brasileiro de Informações em Ciência e Tecnologia - Ibict.

Elton Mártires Pinto

http://lattes.cnpq.br/0079746446660087

https://orcid.org/0000-0002-1348-4185

Doutor em Ciência da Informação e Bacharel em Biblioteconomia pela Universidade de Brasília (UnB). Pesquisador do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict).

Fernando de Jesus Pereira

http://lattes.cnpq.br/5676432086598287

https://orcid.org/0000-0001-5587-4619

Bacharel em Biblioteconomia pela Universidade de Brasília (UnB). Assistente de pesquisa da Coordenação de Tecnologia para Informação (Cotec) do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict).

Ingrid Torres Schiessl

http://lattes.cnpq.br/3155894540549262

https://orcid.org/0000-0001-5815-2574

Mestre em Ciência da Informação e bacharela em Biblioteconomia pela Universidade de Brasília (UnB). Bibliotecária e pesquisadora no Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict).

Lucas Ângelo Silveira

http://lattes.cnpq.br/9490636632029069

https://orcid.org/0000-0002-8107-9659

Mestre em Ciência da Computação pela Universidade de Brasília (UnB), professor adjunto na Faculdade SENAC, desenvolvedor e assistente de pesquisa no Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict).

Milton Shintaku

http://lattes.cnpq.br/8605833104600600

https://orcid.org/0000-0002-6476-4953

Doutor em Ciência da Informação pela Universidade de Brasília. Coordenador de Tecnologia para Informação (Cotec) do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict).

Mirele Carolina Souza Ferreira Costa

http://lattes.cnpq.br/8547303047227327

https://orcid.org/0000-0002-1337-4672

Mirele Carolina Souza Ferreira Costa Doutoranda e Mestre em Informática pela Universidade de Brasília (UnB). Bacharela em Ciência da Computação pela Universidade Federal do Mato Grosso (UFMT). Desenvolvedora e pesquisadora no Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict).

Creación y evaluación de una herramienta para la conversión por lote de archivos PDF/A

Lorenzo Calamante[1], María Marta Vila[2], Mariano Ezequiel Villalba[3], Marisa Raquel De Giusti[4], Carlos Javier Nusch[5], Gonzalo Luján Villarreal[6]

Palabras claves

Preservación digital, procesamiento por lotes, estándar PDF/A

Digital preservation, Batch processing, PDF/A standard

Eje temático

Infraestructura tecnológica

Resumen

Presentación del problema: Los repositorios institucionales realizan cosechas de grandes cantidades de contenidos dispersos en la red, es necesario que esos contenidos cumplan con estándares de preservación digital, sin dejar de lado la eficacia en el uso de los tiempos, por lo que es necesario contar con un método de normalización de lotes de archivos conforme a esos estándares. En este trabajo se presenta una herramienta para el procesamiento por lote de archivos PDF en conformidad con el estándar PDF/A.

Materiales y metodología: Se desarrolló un script escrito en python llamado *PlusUltraPDF*. Consiste en una estructura de control que recorre un directorio padre, sus posibles subdirectorios y archivos PDF y analiza de forma recurrente sucesivos informes de conformidad con el estándar PDF/A (realizados con veraPDF). Luego se invocan dos programas de manipulación de PDF (Ghostscript y OCRmyPDF) que generan nuevos archivos PDF/A-2b derivados de los PDF originales.

Resultados: La evaluación de *PlusUltraPDF* dió buenos resultados: procesó el 97,9% de los archivos y generó un PDF/A-2b válido en el 94,5% de los casos, en comparación con otro script desarrollado en el repositorio que implementa 3-Heights (con mejores posibilidades de conversión), resulta un buen complemento.

1 Universidad Nacional de La Plata, PREBI-SEDICI y Comisión de Investigaciones Científicas, CESGI, lorenzo.calamante@sedici.unlp.edu.ar

2 Universidad Nacional de La Plata, PREBI-SEDICI, vilamm@sedici.unlp.edu.ar

3 Universidad Nacional de La Plata, PREBI-SEDICI, villalba.mariano@prebi.unlp.edu.ar

4 Universidad Nacional de La Plata, PREBI-SEDICI, marisa.degiusti@sedici.unl.edu.ar

5 Universidad Nacional de La Plata, PREBI-SEDICI carlosnusch@sedici.unlp.edu.ar

6 Universidad Nacional de La Plata, PREBI-SEDICI y Comisión de Investigaciones Científicas, CESGI, gonetil@prebi.unlp.edu.ar

Introducción

La digitalización de recursos es una de las tareas esenciales de los repositorios institucionales, encargados de la curatela y difusión de la producción intelectual de las instituciones públicas en el marco de la promoción del acceso abierto. Digitalizar no sólo implica el proceso de migración de objetos en formato físico al digital, sino que requiere un conocimiento y uso de estándares de preservación digital que deben ser aplicados tanto a objetos digitalizados como también a objetos nativos digitales. Estos objetos digitales pueden obtenerse tanto de manera individual como presentarse grandes lotes, principalmente si se tiene en cuenta que:

Los repositorios pueden realizar tareas de recuperación de contenidos que pertenecen a autores de la institución y no han sido autoarchivados. Cuando se identifica un espacio externo al repositorio que cuenta con contenido de la propia institución, es posible realizar una operación de cosecha o recuperación, y posterior ingesta masiva por procesos informáticos. (Soloaga et al., 2020, p.17).

Previo a su ingesta, debe verificarse que los contenidos recuperados cumplan con los estándares de preservación digital; dado que estos procesos pueden involucrar cientos o miles de objetos digitales, es necesario contar con herramientas de evaluación y normalización de esos contenidos por lotes. En este trabajo se presenta una herramienta para el procesamiento por lote de archivos PDF en conformidad con el estándar PDF/A.

Usos del formato PDF y estándar PDF/A

Creado en la década del 90 por Adobe y regulado por la Norma ISO 32000, el formato PDF ha sido ampliamente aceptado. Una de las razones de su éxito se debe a la capacidad de presentar contenidos en distintos formatos de información dentro de un único documento: además del texto, es posible embeber imágenes, videos, código ejecutable, formularios, etc. Sin embargo, con el paso del tiempo han comenzado a surgir problemas de visualización de documentos PDF, algunos de ellos ocasionados por la obsolescencia de las tecnologías y otros debido a la pérdida de información externa al archivo PDF en sí mismo (puede ser la falta de una fuente o del espacio de color). El estándar PDF/A, regulado por las normas ISO 19005-1:2005, promueve un conjunto de normas para evitar estos problemas de visualización en documentos PDF.

Un archivo PDF/A contiene toda la información necesaria para su correcta visualización y se puede utilizar en cualquier plataforma. Existen a su vez aplicaciones comerciales, programas gratuitos y herramientas de código abierto tanto para su creación como para su reproducción. El estándar hace autónomo al archivo de las instancias creadora y visualizadora, pues exige embeber las fuentes y las imágenes y seguir una única especificación de metadatos y un espacio de color independiente de los dispositivos; otra de las ventajas es que impide modificaciones posteriores, así como configuraciones de seguridad que restrinjan el libre uso del archivo. Por otro lado, evita cualquier conflicto de derechos de autor prohibiendo el uso de contenidos cerrados, como es el caso de la compresión LZW para las imágenes o de las fuentes con copyright.

Los cuatro estándares de PDF/A

El estándar PDF/A ha evolucionado hacia diferentes versiones y niveles dentro de estas versiones, las cuales no implican –por definición– la obsolescencia de las versiones anteriores, sino la ampliación de las posibilidades de archivo.

La versión PDF/A-1 tiene dos subniveles: el *b* (básico), que cumple con los requisitos y especificaciones mínimas del estándar y prioriza la visualización del documento y el *a* (accesible), que cumple con los requisitos del estándar, permite embeber etiquetas que describan tanto el orden como la jerarquía de la lectura y exige que cada carácter tenga definido su Unicode. Esto promueve la accesibilidad del contenido del archivo para las personas con discapacidad, ya que permite la correcta reproducción audible del texto y la descripción de los paratextos.

PDF/A-2 permite la inclusión de transparencias, el uso de la compresión JPEG2000 y la incrustación de otros archivos PDF/A, mientras que PDF/A-3 permite incluir cualquier tipo de archivo que no necesariamente se visualice en el archivo, cumpliendo con las restricciones definidas en el estándar ISO 32000-1, razón por la cual si no se desea incluir tales archivos es recomendable utilizar PDF/A-2. Las versiones PDF/A-2 y PDF/A-3 tienen tres subniveles *b* (básico), *a* (accesible) y *u* (unicode): *b* y *a* son análogos a los subniveles de PDF/A-1, mientras que *u* (unicode) agrega el código Unicode que garantiza la indexación y lectura adecuada de los textos.

Finalmente, la versión PDF/A-4 tiene 2 subniveles: PDF/A-4e (*Engineering*), que admite modelos 3D interactivos, medios enriquecidos y anotaciones 3D, así como archivos embebidos y PDF/A-4f, que permite incrustar archivos en cualquier formato.

Elección del estándar

Para la elección del estándar de los archivos PDF/A que serán incorporados a los repositorios SEDICI y CIC-DIGITAL se tuvieron en cuenta las siguientes consideraciones:

Aún no se conocen ni en las normativas de otros repositorios ni en la práctica cotidiana en SEDICI y CIC-DIGITAL casos de uso de PDF/A-4.

A la hora de trabajar con archivos PDF/A, es menester poder controlar los archivos embebidos, razón por la cual se descarta cualquier empleo de PDF/A-3.

Quedan por lo tanto las versiones PDF/A-1 y 2, considerando que si un PDF que se ha de convertir presenta transparencias, se usará PDF/A-2. Sin embargo, puesto que en digitalización de textos digitales se trabaja con imágenes de mapa de bits con capa de texto detrás y que muchas veces es necesario rasterizar (es decir, convertir un gráfico vectorial en una imagen de mapa de bits) algunas imágenes con transparencia y gráficos de vectores por causa de un OCR externo que exige ser corregido, se puede utilizar PDF/A-1.

Ahora bien, elegido un determinado estándar, se debe decidir cuál subnivel utilizar: por ejemplo, el subnivel a requiere de una estructura lógica que en ocasiones los motores de OCR no definen bien, por lo que es necesario realizar ajustes manuales. En tal caso, estructurar el PDF en conformidad con la elección del subnivel *a* depende de la disponibilidad de recursos humanos, ya que es una tarea que insume mucho tiempo. Depende también de la relevancia del trabajo, porque permite la interpretación del documento por la máquina, por ejemplo, para la conversión de texto a audio, que hace accesible el documento para las personas con discapacidad visual.

Pese a las ventajas del subnivel *a*, si se cuenta con un lote grande de documentos que debe ser adecuado al estándar PDF/A en un tiempo razonable, la recomendación será el uso del estándar PDF/A-2b.

Automatización de operaciones

Para la correcta creación de un archivo PDF/A deben embeberse fuentes, caracteres, imágenes, metadatos y espacios de color. Si bien todas estas tareas pueden ser automatizadas, debe observarse que existen ciertas características de la imagen y el texto que no hacen al cumplimiento del estándar, pero que pueden afectar el uso del archivo. Para el caso de un archivo con texto, debe mínimamente revisarse el resultado de un OCR o evitarse el cambio brusco de alguna fuente o carácter que pueda afectar el fluir del texto. Para imágenes, se ha observado que la rasterización pixela las curvas de letras vectorizadas y que, aunque la resolución de imagen no sea buena, el hecho de que esté embebida hace que cumpla con el estándar.

Tampoco es suficiente contar con los metadatos que permitan recuperar la información de la imagen para su preservación, aún sean estos los más adecuados, pues no son garantía del valor y la calidad de una imagen fija en un mapa de bits. Respecto a los metadatos propuestos por el laboratorio de imágenes digitales de la Administración Nacional de Archivos y Registros (NARA) en su guía para la digitalización de materiales de archivo para acceso digital, se afirma que:

Judgments about the quality of an image require a visual inspection of the image, a process that cannot be automated. Quality is influenced by many factors—such as the source material from which the image was scanned, the devices used to create the image, any subsequent processing done to the image, compression, and the overall intended use of the image. (...) The metadata can make no guarantee about the quality of the data. Even if files appear to have a full complement of metadata and meet the recommended technical specifications as outlined in these Technical Guidelines, there may still be problems with the image file that cannot be assessed without some kind of visual inspection. (Puglia et al., 2004, pp.11-12)

A la fecha existen líneas de investigación que buscan mejorar el análisis de documentos utilizando inteligencia artificial para evaluar su calidad visual (Yang et al., 2024). Mientras tanto, sigue siendo recomendable realizar un mínimo control visual de los archivos resultantes que considere siempre la eficacia en el uso de los tiempos. Esto implica definir unas muy pocas variables que pueden surgir del uso cotidiano de los archivos en el repositorio, por ejemplo, la catalogación exige un OCR aceptable en ciertos elementos de la estructura lógica del archivo, como el título, el resumen ó las palabras clave y el público lector que texto e imágenes sean como mínimo legibles.

Programas utilizados en los repositorios SEDICI y CIC-Digital

Existen muchas herramientas informáticas que permiten tanto generar documentos en formato PDF/A como también validar el grado de cumplimiento del estándar de un documento. A continuación se describen brevemente las herramientas utilizadas por los equipos de digitalización de los repositorios institucionales de la Universidad Nacional de La Plata (SEDICI) y de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC-DIgital).

Creación de archivos PDF/A

Para la creación de archivos PDF/A se utilizan:

ABBYY FineReader[7]: normalmente, luego de editar las imágenes se realiza el OCR con *ABBYY FineReader*. Este programa permite seleccionar el contenido según se trate texto, imagen o cuadro y aplicar el estándar PDF/A al archivo. En el momento del guardado, el programa permite modificar la compresión para obtener documentos más pequeños, que pueden ir desde compresiones sin pérdida a compresiones con pérdida de calidad. A partir de la versión 14, el programa permite tareas por lote mediante la herramienta *ABBYY Hot Folder*.

Ghostscript (gs)[8]: es un intérprete de archivos PDF desarrollado por *Artifex Software* y disponible tanto en las licencias *GNU GPL Affero* y de uso comercial. Permite crear archivos PDF/A en cualquiera de las tres primeras versiones con la limitación de que siempre será bajo el subnivel b. Resulta una herramienta adecuada cuando no se quiere intervenir fuertemente sobre el archivo que se ha de convertir, pues no rasteriza las páginas, aunque no siempre los resultados cumplan efectivamente con el estándar.

OCRmyPDF[9]: es un software libre desarrollado por James R. Barlow que combina otros software libres: *unpaper* (edición de imagen)[10], *tesseract* (OCR)[11] y *Ghostscript*. Da por resultado un archivo PDF/A-2b. Se utiliza mediante línea de comandos y los parámetros mínimos son la ruta de los archivos PDF de entrada y de salida. En caso de contar con texto, cuenta con una función que rasteriza las páginas y les hace un OCR. Cuenta con un script (*watcher*) que permite crear un hotfolder que controla la entrada de un nuevo archivo y lo procesa.

3-Heights: este software, desarrollado por *pdf-tools*[12], permite la conversión de archivos PDF a cualquiera de los subestándares de PDF/A-1;2 y 3. En el repositorio se ha desarrollado con anterioridad un script escrito en bash que analiza el estándar de conversión conveniente, convierte por lotes archivos PDF a PDF/A y valida los que ya cumplen con la norma.

Validación de documentos

Es necesario validar el archivo PDF obtenido para comprobar que efectivamente cumpla con el estándar PDF/A, pues aunque cuente con el metadato técnico que permite a los lectores reconocerlo como tal, pueden persistir algunos problemas que ocasionan incumplimientos con la norma.

En los repositorios SEDICI y CIC-DIGITAL se utilizan con los siguientes programas de validación de PDF/A:

Acrobat Reader: cuenta con un validador de estándares de PDF. Si la validación es errónea, se puede corregir desde el mismo *Acrobat*: el propio programa analiza y aplica los cambios necesarios para convertirlo correctamente al estándar PDF/A seleccionado (aplica espacios de color, incrusta fuentes, elimina caracteres no definidos, entre otros).

7 Accesible desde:

8 Accesible desde: https://Ghostscript.readthedocs.io/en/latest/
9 Accesible desde: https://ocrmypdf.readthedocs.io/en/latest/
10 Accesible desde: https://github.com/unpaper/unpaper
11 Accesible desde: https://tesseract-ocr.github.io/tessdoc/Home.html
12 Accesible desde: https://www.pdf-tools.com/

VeraPDF[13]: es un software libre, desarrollado por *Open Preservation Foundation*, que contiene todos los estándares de PDF/A y sus niveles de conformidad necesarios para la preservación. El programa analiza y produce un informe de validación que muestra qué estándares se verifican y si los documentos PDF seleccionados los cumplen. Cuenta con un parámetro (*-x, --extract*) que permite ver de forma detallada sus características (metadatos, fuentes incrustadas, espacios de color, etc.). Cuenta tanto con una interfaz gráfica como con una línea de comandos.

Procesamiento por lotes

A la hora de realizar procesamientos por lote, los repositorios SEDICI y CIC-Digital cuentan con tres alternativas: *ABBYY Hot Folder*, el script que implementa *3-Heights* y el script *watcher* de *OCRmyPDF*. Estas tres alternativas tienen sus inconvenientes: tanto 3-Heights como ABBYY son software propietario, por lo que su uso está limitado a las condiciones de la licencia, a la vez que *ABBYY Hot Folder* y *OCRmyPDF* necesariamente rasterizan las páginas del archivo PDF, con la consiguiente pérdida de calidad, por tales motivos se buscó una solución que evitara el uso de software propietario y la rasterización de las páginas: como primera opción se eligió *Ghostscript*, pero se observó que los PDF/A resultantes no siempre eran válidos, por lo que se convino en recurrir a la rasterización mediante *OCRmyPDF* sí y sólo sí el PDF procesado con *Ghostscript* no era válido.

Era necesario a su vez definir un flujo de procesamiento de los archivos y se propuso el siguiente:

I. Primera iteración

1. Evaluación del lote en bruto
2. Filtrado de los que deben convertirse y los que ya son válidos
3. Conversión de los que no cumplen el estándar

II. Segunda iteración

4. evaluación del lote convertido
5. Filtrado de los válidos y los que deben corregirse
6. Corrección de los que no cumplen el estándar

III. Tercera iteración

7. Evaluación del lote convertido y corregido
8. Informe de errores (si se consideran agotadas las posibilidades de corrección)

Entre las primera y segunda iteración se puede establecer un punto de control para que el usuario decida si procede con la corrección en caso que la considere necesaria, o interrumpe el proceso para continuarlo en otro momento.

13 Accesible desde: https://docs.verapdf.org/

Descripción de las principales funcionalidades

El script, llamado *PlusUltraPDF*, está escrito en python 3. Se importan los módulos *os*[14] para acceder a los directorios e invocar a los programas que utilizará y *time*[15] para calcular los tiempos de ejecución del programa. Mediante la función de python *input* se define el directorio padre en el que se encuentran los archivos a convertir a PDF/A.

Se implementó un conjunto de funciones, *verapdf, primera_pasada_gs, ocr_pdfa* y *control_error* , para las que se ha confeccionado un bucle de control de directorios que recorre de manera recursiva todo el directorio padre en el que se encuentran los archivos a convertir, y ejecuta las funciones de conversión y validación sobre los archivos encontrados en cualquier directorio perteneciente a la estructura jerárquica.

La función *verapdf* es aplicada repetidas veces a lo largo de la ejecución del script: en la primera iteración, para la validación inicial que define el input de la primera iteración con *Ghostscript* luego, para validar el resultado de esa iteración y obtener el input de la corrección con *OCRmyPDF* y, finalmente, para registrar si existe algún PDF que no cumple con el estándar al final del proceso.

```
verapdf --maxfailuresdisplayed 10 "path">"path+/veraLOG.xml"
```

Imagen 1. Comando de veraPDF. Elaboración propia.

Esta función genera, como resultado del procesamiento de los documentos, una salida en formato xml, la cual es almacenada en un archivo de salida (veraLOG.xml en el ejemplo de la Imagen 1) que fungirá como input para los procesos de conversión a PDF/A, corrección del output y cálculo de errores.

La función *primera_pasada_gs* utiliza *Ghostscript* en la primera iteración a fines de que la conversión a PDF/A evite la rasterización de los gráficos de vectores. Respecto a las fuentes, sin embargo, es posible que haya cambios notorios, dado que *Ghostscript* incrustará la fuente que más se asemeje a la del archivo a partir de un conjunto de fuentes propio que puede no coincidir con el original.

Una vez completada la ejecución de la función primera_pasada_gs, se adiciona el prefijo "-PDFA" al nombre del archivo, a fin de indicar que se trata de un archivo adecuado al estándar PDF/A.

La invocación a *Ghostscript* queda parametrizada de acuerdo con el ejemplo en la Imagen 2, donde *name.text* es la ruta completa del archivo entrada, que se obtiene a partir del XML generado por *veraPDF* y *output* es la ruta del archivo de salida, que se genera reemplazando la extensión ".pdf" de *name.text* por "-PDFA.pdf".

```
gs -dPDFA=2 -dBATCH -dNOPAUSE -sColorConversionStrategy=UseDeviceIndependentColor -sDE-
VICE=pdfwrite -dPDFACompatibilityPolicy=1 -sOutputFile="+output+""+name.text+"
```

Imagen 2. Comando de Ghostscript. Elaboración propia.

Entre los parámetros que se utilizar durante la invocación al comando gs se destacan:

-dPDFA=2 define la versión de PDF/A. De acuerdo con lo anteriormente explicado, se elige PDF/A-2.

-dNOPAUSE aplica el mismo proceso a cada página sin esperar otra indicación, y avanza automáticamente a la siguiente página.

14 Accesible desde: https://docs.python.org/3/library/os.html
15 Accesible desde: https://docs.python.org/3/library/time.html

-dBATCH cierra Ghostscript una vez finalizado todo el procesamiento.

-sColorConversionStrategy=UseDeviceIndependentColor: al momento de convertir el espacio de color, la estrategia definida es usar uno independiente del dispositivo. El espacio de color que embebe es un perfil ICC propio de Artifex, desarrolladora de Ghostscript.

-sDEVICE=pdfwrite -sDEVICE define el dispositivo[16] encargado de procesar la salida, que puede ser desde un archivo postscript hasta un comando para una impresora determinada. En este caso pdfwrite crea un archivo PDF.

-dPDFACompatibilityPolicy=1 cuando se identifica una incompatibilidad respecto a PDF/A en el PDF a convertir, se debe definir una política para tratar con dicha situación. El valor 0 imprime el metadato técnico del PDF/A y conserva la incompatibilidad, lo que da por definición un archivo no válido. El valor 2 detiene el procesamiento del PDF, lo que genera un PDF/A "roto". Por último, el valor 1 elimina el elemento que causa la incompatibilidad, lo cual aumenta (pero no garantiza) la probabilidad de que el archivo resultante sea válido, con el consiguiente riesgo de pérdida de información.

-sOutputFile= indica la ruta del archivo que se creará.

Una vez que se ha creado mediante *Ghostscript* un archivo correspondiente a cada PDF que no cumplía con el estándar, una nueva validación informa si cada uno de ellos cumple con el estándar. En caso de no ser así, la función *ocr_pdfa* recurre al nuevo informe e invoca *OCRmyPDF*, que procesa cada PDF/A derivado que no cumple con el estándar. Es un proceso más interventivo, pues rasteriza cada página, realiza un OCR y los convierte a PDF/A-2b.

La invocación a *OCRmyPDF* queda parametrizada como se indica en la Imagen 3, donde "name.text" es la ruta del PDF, tal y como se la obtiene a partir del informe de veraPDF. Se repite a fines de sobreescribir el archivo.

```
ocrmypdf --force-ocr -l spa+por+eng --optimize 02 --max-image-mpixels 900 "name.text" "name.text"
```

Imagen 3. Comando de OCRmyPDF. Elaboración propia.

Se desglosan a continuación algunos de los parámetros:

--force-ocr es un parámetro que pasa por alto el texto del archivo, rasteriza cada página y le realiza un nuevo OCR.

-l (--language): en *PlusUltraPDF* se parametrizan tres idiomas para el OCR: español, portugués e inglés. El parámetro.

--optimize aplica compresiones sobre las imágenes resultantes de la rasterización, que graduadas en el nivel 02 aplican una compresión con la menor pérdida de calidad posible.

El parámetro **--max-image-mpixels** permite definir un máximo de megapíxeles soportado por el proceso (aquí definido en 900) si tiene que recurrir a ese parámetro, evitará la ruptura del proceso, aunque dará como resultado un archivo de gran tamaño.

16 En los sistemas operativos basados en Unix, como por ejemplo GNU/Linux, los dispositivos del sistema son representados por archivos, y son utilizados para encapsular tanto estructuras lógicas (por ejemplo archivos almacenados en disco) como también estructuras físicas (como por ejemplo componentes de hardware: memoria, interfaz de red, etc.)

Finalmente, como el archivo que no cumple el estándar es inaceptable, se lo sobreescribe definiendo como ruta de salida la misma que de entrada.

La función *control_error* repasa el informe final de veraPDF y si encuentra que alguno de los archivos con el sufijo *-PDFA* no es válido, presenta en pantalla y en el informe de cuál se trata y el mensaje de error que devuelve veraPDF.

Una última función, definida como *segunda_pasada*, aplicará las funciones *ocr_pdfa, verapdf y control_error*. Se ejecuta o bien inmediatamente luego de la conversión con *Ghostscript*, o bien en otro momento, gracias a un punto de control que permite al usuario decidir si procede con la corrección (siempre que sea necesaria) o concluye el proceso y lo continúa en otro momento.

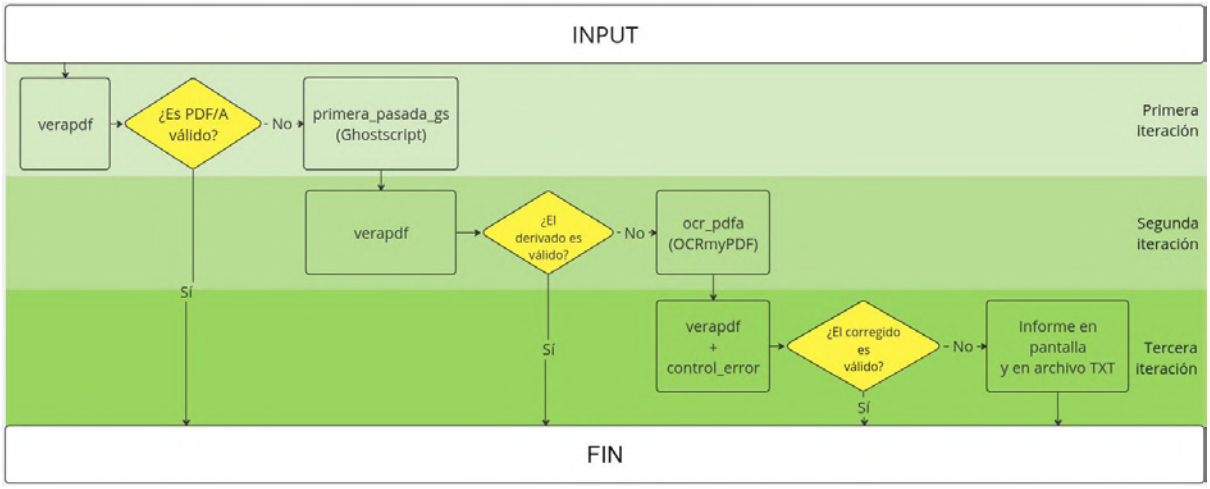


Gráfico 1. Diagrama de decisión de *PlusUltraPDF*. Elaboración propia.

Evaluación

En esta sección se describen las pruebas realizadas sobre este desarrollo: en primera instancia para verificar que PlusUltraPDF resuelve correctamente lo solicitado (es decir, que procesa todos y cada uno de los archivos del lote y que los archivos PDF/A resultantes son válidos) y en segunda instancia para comparar con la solución alternativa también desarrollada por los equipos de los repositorios SEDICI y CIC-Digital, basada en 3-Heights. La evaluación se realiza con los siguientes supuestos:

A fines de proceder con esta evaluación, no resulta relevante saber cuáles son los errores que ocasionan el incumplimiento con el estándar PDF/A, sino hasta que se hayan agotado las posibilidades de conversión, que son primero intentar convertir con el menor cambio posible y luego rasterizar las imágenes de página.

A fin de evitar la introducción de sesgos durante la evaluación, las pruebas se realizan sobre un dataset generado por terceros. Se utilizan los test suite de Isartor (isartor-test-suite)[17] y de BFO (pdfa-testsuite)[18]. En el caso del isartor-test-suite, su prolija estructura de directorios permite evaluar también el bucle de control recursivo.

Los resultados de cada una de las iteraciones deben validarse con otro validador de PDF/A, pues veraPDF ya está incluído en PlusUltraPDF. En este caso, se utiliza el validador de Acrobat DCPro y de su herramienta "Asistente de acciones" para validar por lotes.

Las coincidencias y diferencias entre software de validación, aplicadas a las evaluaciones aquí propuestas, se resuelven de la siguiente forma:

Si es validado por ambos programas, se lo toma por válido.

Si no es validado por ambos programas, no se lo toma por válido.

Si un archivo es validado por un programa y por el otro no, entonces no se lo toma por válido (si se trata de uno de los archivos resultantes del proceso, esto indica un problema, especialmente si sucede al finalizar la ejecución, ya que en este punto se esperaría que fuera válido).

Finalmente, para la evaluación de eficacia se tendrán en cuenta los resultados de cada una de las iteraciones, mientras que para la comparación con el script de *3-Heights*, consideraremos el resultado de la primera y el resultado final.

Resultados de la evaluación de eficacia

A continuación, se presentan los resultados de las pruebas realizadas sobre dos datasets generados por terceros: los test suite de Isartor (isartor-test-suite) y de BFO (pdfa-testsuite).

Isartor test suite

Tanto *veraPDF* como *Acrobat DCPro* confirmaron que los 204 archivos de este dataset de entrada no eran válidos. A partir de la segunda iteración, *veraPDF* (y por lo tanto todo el proceso) arrastra cinco excepciones (es decir validaciones que no pudo realizar y por lo tanto carecen de resultado) que corresponden a cinco archivos derivados creados en la primera iteración con *Ghostscript*. *Acrobat DCPro* encuentra que esos cinco archivos no son válidos. Esos cinco archivos deben descontarse como no procesados en la tercera iteración y, en el conteo de válidos, deben sumarse a otros ocho que entre *veraPDF* y *Acrobat DCPro* son contabilizados como no válidos, es decir, 13 archivos derivados no cumplen con el estándar. Se tiene entonces un 97,5% de procesamiento y un 93,6% de resultados válidos.

17 Accesible desde: https://web.archive.org/web/20161009080014/http://www.pdfa.org/2011/08/isartor-test-suite/

18 Accesible desde: https://github.com/bfosupport/pdfa-testsuite

BFO pdfa-testsuite

Este dataset está compuesto de 34 archivos PDF/A (10 válidos y 24 no válidos) Sin embargo *veraPDF* contabilizó 9 válidos y 25 no válidos y *Acrobat DCPro*, 11 válidos y 23 no válidos. El script filtró 9 archivos válidos (se debe recordar que se basa en veraPDF para realizar esta validación) y procesó los otros 25 archivos. Los dos archivos que hacen la diferencia entre *veraPDF* y *Acrobat DCPro* arrastran esa diferencia a lo largo de las validaciones del proceso (es decir que no hay diferencias con los nuevos archivos creados en el proceso). Con un 100% de procesamiento y 100% de resultados válidos, la prueba ha resultado exitosa.

	Isartor		BFO	
	Procesamiento	Resultados válidos	Procesamiento	Resultados válidos
Primera iteración	100%	0%	100%	26,47058824%
Segunda iteración	100%	84,31372549%	100%	82,35294118%
Tercera iteración	97,5490196%	93,62745098%	100%	100%

Tabla 1. Porcentaje de procesamiento y archivos válidos por iteración para cada dataset. Elaboración propia.

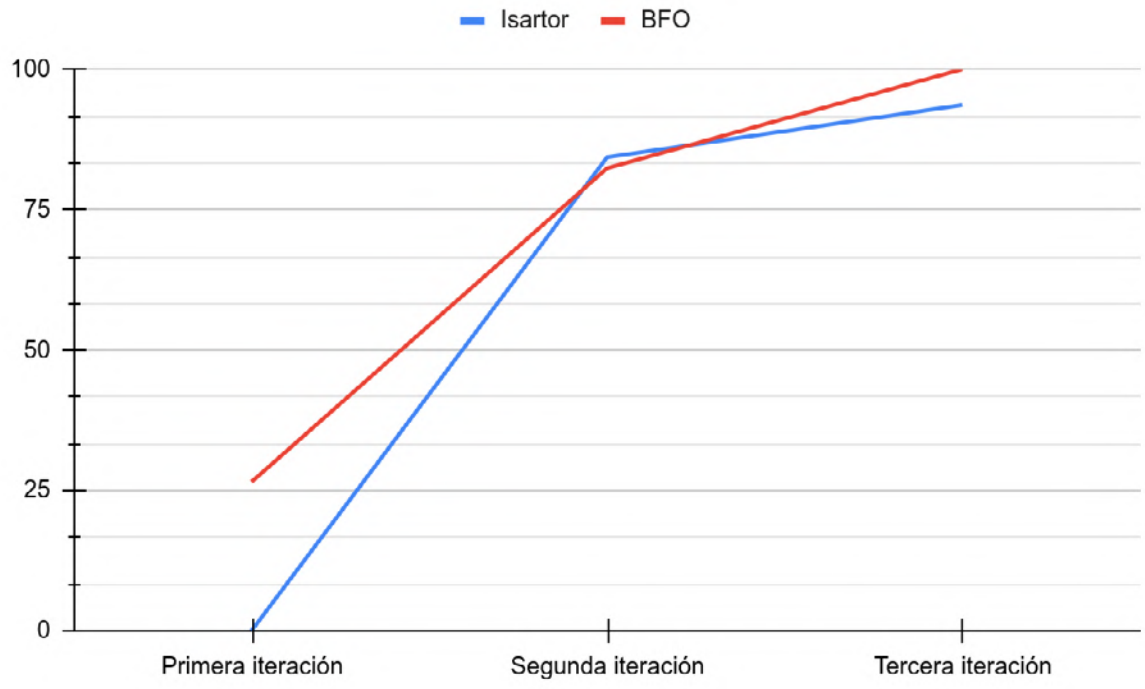


Gráfico 2. Progresión del porcentaje de archivos válidos por iteración para cada dataset. Elaboración propia.

Resultados de la confrontación con el script *3-Heights*

Isartor test suite

Realizados 3 intentos, el script que implementa *3-Heights* no procesó ninguno de los archivos de este dataset.

BFO pdfa-testsuite

De 34 archivos, el script que implementa *3-Heights* procesó 33, cuyos derivados fueron 26 PDF/A-1b y 7 PDF/A-2u. La diferencia entre los no válidos del resultado final del script *3-Heights* y la primera iteración, para ambos validadores, es de 1 archivo, es decir, que hay uno no válido de más; se trata de un caso en el que el archivo original no pudo ser procesado y fue derivado al directorio ocr, donde se generó un archivo derivado no válido.

La diferencia entre los resultados finales del script 3-Heights y *PlusUltraPDF* es de 8 válidos y 1 no válido, lo que significa que hay 9 archivos de diferencia. Estos archivos corresponden a los 9 archivos originales que *PlusUltraPDF* tomó por válidos y que el script que implementa *3-Heights* convirtió. Ese único archivo no válido corresponde al caso anteriormente referido.

	PlusUltraPDF		Script 3-Heights	
	Procesamiento	Válidos	Procesamiento	Válidos
Isartor	97,5490196%	93,60%	0%	0%
BFO	100%	100%	100%	97%
Ponderados	97,8991596%	94,53781513%	14,28571429%	13,86554622%

Tabla 2. Comparación del procesamiento y los resultados válidos entre *PlusUltraPDF* y el script que implementa 3-Heights. Elaboración propia.

Conclusiones

En este trabajo se presentó un script llamado *PlusUltraPDF* para la conversión por lotes de archivos PDF a PDF/A. El mismo consiste en una estructura de control que recorre un directorio en busca de archivos PDF y en el análisis recurrente de sucesivos informes de conformidad con el estándar PDF/A (realizados con *veraPDF*). Para la conversión de los archivos se utilizan dos programas de manipulación de PDF (*Ghostscript* y *OCRmyPDF*) que generan nuevos archivos PDF/A-2b derivados de los PDF originales. Este script se formuló como alternativa a un desarrollo previo basado en 3-Heights para su uso en los repositorios SEDICI y CIC-Digital. Para verificar su funcionamiento, se realizó una evaluación utilizando dos datasets con PDF/a no válidos generados por terceros, a fines de evitar sesgos: Isartor (isartor-test-suite) y BFO (pdfa-testsuite).

La evaluación de *PlusUltraPDF* dió buenos resultados, ya que se procesó el 97,9% de los archivos y se generó un documento PDF/A-2b válido en el 94,5% de los casos.

Ahora bien, no podemos dejar de considerar que en razón de las restricciones del estándar PDF/A, toda normalización de un archivo PDF corre el riesgo de pérdida de información, más aún si se realiza por lotes: la rasterización ocasiona grandes cambios en la calidad de la imagen, pero es necesario también destacar que la alternativa menos interventiva también tiene sus consecuencias. En el uso de *Ghostscript* se definió una tolerancia de conversión que elimina el elemento que causa la incompatibilidad, lo cual aumenta (pero no garantiza) la probabilidad de que el archivo resultante sea válido, pero también pone en riesgo la información no conforme con el estándar. Respecto a las fuentes es posible que haya cambios notorios, dado que *Ghostscript* incrusta la fuente que más se asemeja a la del archivo a partir de un conjunto que puede no coincidir con el original, problema que de todos modos se subsana si se piensa que el riesgo de pérdida es mayor para una fuente no incrustada.

Comparado con el script que implementa *3-Heights*, éste tiene la ventaja de contar con un mayor control en la relación análisis-conversión que le permite obtener el mejor estándar para un archivo. Por el contrario, *PlusUltraPDF* funciona como lecho de Procusto[19], pues ese es el objetivo: tratar de obtener un archivo PDF/A-2b a como dé lugar con lotes muy grandes que no pueden trabajarse caso por caso. En este sentido, dado que el script que implementa *3-Heights* no logró procesar un lote entero, sería recomendable utilizar *PlusUltraPDF* en aquellos casos en los que el script que implementa *3-Heights* no pueda resolver el problema.

19 Procusto era un personaje de la mitología griega que sometía a los viandantes de los caminos del Ática a una tortura que consistía en atarlos a un lecho: a aquellos que eran más bajos que el lecho, les estiraba las extremidades, mientras que a los más altos se las cortaba.

Para futuras mejoras de esta herramienta debería contemplarse que, por un lado, los resultados de OCR pueden ocasionar problemas si a partir de él se realiza la obtención de los metadatos y, por el otro, los idiomas con los que se trabaja presentan limitaciones y sería conveniente brindar soporte al proceso de OCR en idiomas francés, italiano y alemán como mínimo.

Bibliografía

PDF Association. (2010). *Introducción al PDF/A (Español)* Recuperado 4 de enero de 2024, de https://pdfa.org/resource/introduccion-al-pdfa/

Puglia, S., Reed, J., & Rhodes, E. (2004). *Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files—Raster Images*. https://www.archives.gov/files/preservation/technical/guidelines.pdf

Soloaga, I., Fernández, E. C., & De Giusti, M. R. (2020, diciembre 3). *Cómo crear paquetes de información para repositorios digitales*. Mini Curso «Cómo crear paquetes de información para el repositorio» (Brasilia, 2020). http://sedici.unlp.edu.ar/handle/10915/110561

ISO 32000-1:2008(en), Document management—Portable document format—Part 1: PDF 1.7. (s. f.). Recuperado 17 de abril de 2024, de https://www.iso.org/obp/ui/en/#iso:std:iso:32000:-1:ed-1:v1:en

ISO 19005-1:2005(en), Document management—Electronic document file format for long-term preservation—Part 1: Use of PDF 1.4 (PDF/A-1). (s. f.). Recuperado 17 de abril de 2024, de https://www.iso.org/obp/ui/en/#iso:std:iso:19005:-1:ed-1:v2:en

Yang, H.-W., Agrawal, A., Fragkogiannis, P., & Mulay, S. N. (2024). Can AI Models Appreciate Document Aesthetics? An Exploration of Legibility and Layout Quality in Relation to Prediction Confidence (arXiv:2403.18183). arXiv. https://doi.org/10.48550/arXiv.2403.18183

Lorenzo Calamante es profesor en Letras por la Facultad de Humanidades y Ciencias de la Educación de la UNLP y desde 2019 realiza tareas de digitalización en CESGI-CIC y en PREBI-SEDICI. Colaboró en la digitalización de los tomos de la colección Cervantina de la Biblioteca Pública de la UNLP.

ORCID: https://orcid.org/0000-0002-2776-3564

María Marta Vila integra desde el año 2003 el staff permanente del repositorio central digital de la Universidad Nacional La Plata (UNLP). Dentro del equipo de trabajo del Servicio de Difusión de la Creación Intelectual (SeDiCI) realiza tareas de desarrollo, investigación y mantenimiento de software. Los principales lineamientos de su labor se relacionan con el web scraping, harvesting, oai-pmh, visibilidad Web, interoperabilidad de repositorios y sistemas de gestión de contenidos entre otros. Desde el año 2016 participa además del grupo de preservación digital, difusión y análisis de impacto del Centro de Servicios en Gestión de Información (CESGI) de la Comisión de Investigaciones Científicas (CIC) de la provincia de Buenos Aires.

ORCID: https://orcid.org/0000-0001-9341-9510

Mariano Ezequiel Villalba es estudiante de Ingeniería Electromecánica en la Facultad de Ingeniería de la Universidad Nacional de La Plata. A partir de febrero del año 2020 realiza digitalización de material bibliográfico en PREBI-SEDICI.

Marisa Raquel De Giusti es doctora en Ciencias Informáticas, Ingeniera en Telecomunicaciones y Profesora en Letras de la Universidad Nacional de La Plata (UNLP). Es Profesora de Posgrado en la Facultad de Informática de la UNLP, Directora del Proyecto de Enlace de Bibliotecas (PREBI, 1997) y directora del Servicio de Difusión de la Creación Intelectual (SEDICI, 2002). Impulsó la creación y fue directora hasta el año 2023 del Centro de Servicios en Gestión de Información (CESGI) de la Comisión de Investigaciones Científicas (CIC), donde actualmente reviste como Investigador Emérito. Es presidenta del Consorcio Iberoamericano para Educación en Ciencia y Tecnología (ISTEC) y Directora de la Iniciativa Library linkage (LibLink) de dicho consorcio. Integra el Comité de Expertos del Sistema Nacional de Repositorios Digitales (SNRD) y el Comité Asesor en ciencia abierta y ciudadana. Cuenta con más de 400 trabajos en áreas diversas entre las que se incluyen la gestión de la información, preservación digital, rankings y visibilidad institucional.

ORCID: https://orcid.org/0000-0003-2422-6322

Carlos Javier Nusch es Profesor y Licenciado en Letras por la Universidad Nacional de La Plata y Máster en Humanidades Digitales por la Universidad de Educación a Distancia de España. Ha publicado varios artículos sobre trabajo académico colaborativo, repositorios digitales, digitalización de patrimonio cultural, análisis del discurso político y literatura clásica, medieval y moderna. Trabaja en el Servicio de Difusión de la Creación Intelectual (SEDICI) de la UNLP, en el Proyecto de Enlace de Bibliotecas (PREBI) y en el repositorio CIC-Digital (CICPBA). Es miembro del Comité Asesor del Centro de Servicios en Gestión de Información (CESGI) y personal del Observatorio Medioambiental La Plata (UNLP - CICPBA - CONICET). Coordina la Oficina de Relaciones Institucionales del Consorcio Iberoamericano para la Educación en Ciencia y Tecnología (ISTEC). Participa como docente colaborador ad honorem en el curso de posgrado "Bibliotecas y Repositorios Digitales. Tecnología y aplicaciones" de la Facultad de Informática de la UNLP. Ha participado en proyectos sobre Oralidad, Escritura, Humanidades Digitales Recursos Académicos, Harvesting, OAI-PMH, Visibilidad Web, Repositorios Abiertos, Producción Académica y Científica, Accesibilidad financiados por la UNLP, la CICPBA y el ISTEC.

ORCID: https://orcid.org/0000-0003-1715-4228

Gonzalo Luján Villarreal es Doctor en Ciencias Informáticas, forma parte de PREBI-SEDICI desde el año 2004 y es coordinador del Portal de Revistas (2008), del Portal de Congresos (2009), del Proyecto de Visibilidad Web Institucional (2012) y del Portal de Libros (2015). Es también director del Centro de Servicios en Gestión de Información (CESGI, 2016) de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, coordinador informático de revistas científicas de la Universidad Nacional de La Plata y profesor de la Facultad de Informática de la misma universidad.

ORCID: https://orcid.org/0000-0002-3602-8211

Integración de HERA con Aplicaciones de Terceros. Oportunidades y Beneficios

Lautaro Josin Saller[1], Pablo Gabriel Terrone[2], Ezequiel Carletti[3], Enzo Rucci[4], Gonzalo Luján Villarreal[5]

Palabras claves

Evaluación científica, indicadores cienciométricos, revista científica, artículo científico

Keywords

Scientific evaluation, scientometric indicators, scientific journal, scientific paper

Eje temático

Infraestructura tecnológica

Resumen

Este trabajo analiza las posibilidades de integración de aplicaciones y sistemas web con la plataforma HERA 2.0 a través de su API REST, a fin de que los primeros puedan incorporar métricas e indicadores sobre calidad e impacto de recursos académicos. Para ello, se analizan 3 contextos de integración concretos: el navegador web Google Chrome, el gestor de contenidos Wordpress y el sistema de gestión editorial Open Journal System. Para cada contexto se analizan los objetivos buscados en la integración y las posibilidades que ofrece cada uno de los sistemas informáticos para agregar funcionalidad. Luego de una primera prueba de concepto, se implementó un conector genérico de HERA para su integración con aplicaciones externas. En particular, se detalla su diseño y la forma en que fue integrado en cada uno de los contextos. Finalmente, se mencionan otras posibilidades de integración de HERA en diferentes contextos de aplicaciones del sistema científico.

1 PREBI-SEDICI Universidad Nacional de La Plata, Argentina, lautaro.josin@sedici.unlp.edu.ar

2 PREBI-SEDICI Universidad Nacional de La Plata, Argentina, pabloterrone@sedici.unlp.edu.ar

3 Facultad de Informática, Universidad Nacional de La Plata, Argentina, carlettieze@gmail.com

4 III-LIDI, Facultad de Informática, Universidad Nacional de La Plata y Comisión de Investigaciones Científicas, Argentina, erucci@lidi.info.unlp.edu.ar

5 PREBI-SEDICI Universidad Nacional de La Plata y CESGI Comisión de Investigaciones Científicas, Argentina, gonzalo@prebi.unlp.edu.ar

Abstract

This work analyzes the possibilities of integrating applications and web systems with the HERA 2.0 platform through its REST API, so that the former can incorporate metrics and indicators about the quality and impact of academic resources. To achieve this, three specific integration contexts are analyzed: the web browser Google Chrome, the content management system Wordpress, and the editorial management software Open Journal System. For each context, the integration objectives are examined along with the possibilities offered by each of the computer systems to add functionality. After an initial proof of concept, a generic HERA connector was implemented for integration with external applications. In particular, its design and the manner in which it was integrated into each of the contexts are detailed. Finally, other possibilities for integrating HERA into different contexts of scientific system applications are mentioned.

Introducción

En las últimas décadas ha proliferado el uso de plataformas en línea y sitios web como espacios para la comunicación y difusión de artículos científicos, tesis, libros, datos de investigación y cualquier producto resultante de actividades de investigación. Estos espacios pertenecen a instituciones académicas y científicas, a grupos editoriales, a sociedades científicas e instituciones académicas que gestionan publicaciones periódicas, y a investigadores, becarios o grupos de estudio que realizan investigaciones, publican sus resultados en distintos formatos, y desean dar a conocer dichos resultados. La calidad de los resultados publicados, así como también el impacto que han generado en la comunidad académica y científica, son aspectos determinantes a la hora de hacer uso de estos productos por parte de terceros (Lindsey, 1989). Por ejemplo, artículos publicados en revistas que se encuentran en determinada base de datos, o que han obtenido mayor cantidad de citas según determinado servicio, son por lo general mejor vistos por la comunidad científica que artículos que no han recibido citas o que han sido publicados en revistas que carecen de avales externos. Se observa entonces que el número de citas de un artículo o la pertenencia de una revista a ciertas bases de datos son indicadores que correlacionan positivamente con la calidad y el impacto de tales recursos científicos (Repiso, 2015).

Si bien la evaluación de un recurso científico, como por ejemplo un artículo, a partir de un conjunto de indicadores parece ser una tarea relativamente simple, esto dista mucho de ser así. La multiplicidad de indicadores disponibles, sumado a la diversidad de metodologías que existen para su cálculo (Kim y Chung, 2018), llevan a que la evaluación de calidad e impacto de un artículo científico que considere todos los indicadores (o al menos muchos de ellos) sea una tarea realmente compleja (Kavic y Satava, 2021). Asimismo, deben considerarse dificultades adicionales como que estos indicadores evolucionan constantemente, que permanentemente se crean nuevos indicadores, y que por lo general las organizaciones responsables de calcular los distintos indicadores no exponen correlaciones de sus datos con los datos que generan otras organizaciones. Por lo tanto, se torna aún más compleja la obtención de métricas integradas o al menos la comparación de indicadores similares generados por distintos organizaciones. Cabe destacar aquí que no es objetivo de los autores de este trabajo posicionarse a favor o en contra de ninguna organización ni de ningún indicador, sino destacar la heterogeneidad intrínseca en los sistemas de generación de indicadores, y proponer herramientas que permitan lidiar con la complejidad y el dinamismo propios de los sistemas de evaluación científica.

La Herramienta para Enriquecimiento de Recursos Académicos (HERA) es un proyecto que recopila e integra indicadores de revistas científicas y de artículos publicados en revistas científicas en un único espacio (Porto et al., 2022). Desde su versión 2.0, HERA incorpora una Interfaz de Programación de Aplicaciones o API REST (Prayogi et al., 2020) que permite utilizar los datos recopilados por HERA desde otros sistemas, y de este modo promover la incorporación de indicadores de calidad e impacto de revistas y artículos desde cualquier sistema o servicio que tenga como objetivo gestionar este tipo de recursos (Carletti, 2023). Este tipo de sistemas incluye a portales de revistas, sitios web institucionales, repositorios digitales, sitios web personales de investigadores, sitios web de grupos de investigación, sistemas agregadores de recursos cosechados desde fuentes externas, sistemas CRIS, entre otros.

En este trabajo se describen un conjunto de herramientas informáticas desarrolladas para utilizar la API de HERA 2.0, y se introducen algunas de las estrategias utilizadas para la integración de datos obtenidos desde HERA con distintos tipos de sistemas de gestión o publicación científica utilizando dichas herramientas.

HERA

El enriquecimiento de datos es el proceso de incorporar actualizaciones y nueva información en datos existentes que ya se encuentran consolidados. Este proceso es aplicado sobre sistemas y bases de datos de múltiples organizaciones y empresas (Gutierrez 2016; Rettore 2020; Djiroun 2023), y resulta de gran utilidad para una mejor toma de decisiones. En el contexto de las publicaciones científicas, pueden considerarse como datos existentes y consolidados tanto a los artículos científicos publicados en revistas, como a las mismas revistas que publican dichos artículos. Sobre estos datos consolidados pueden aplicarse procesos de enriquecimiento para obtener información que permite comprender mejor cómo han influido hasta el momento en el área de la ciencia al que pertenecen, o con qué credenciales y avales cuentan las revistas donde han sido publicados. Una vez enriquecidos, estos datos pueden servir, por ejemplo, para mejorar los sistemas de evaluación de las instituciones, para optimizar las estrategias de visibilidad de los equipos editoriales, para evaluar el grado de confianza de artículos científicos o para decidir cuál será la revista a la que se enviará un manuscrito que se desea publicar.

Existen muchos indicadores y formas de calcularlos, tanto vinculados a la calidad como también al impacto de las publicaciones científicas. Para evaluar el impacto de un artículo, muchos indicadores están basados en la cantidad de veces que dicho artículo ha sido citado. Muchas organizaciones y servicios ofrecen métricas basadas en el número de citas, como por ejemplo Google Scholar, Scopus Semantic Scholar o Dimensions. Sin embargo, dado que cada uno de ellos utiliza diferentes fuentes de datos para realizar estos cálculos y aplica distintos sistemas de control y limpieza sobre dichos datos, los resultados que arrojan para un mismo artículo pueden ser muy dispares (Falagas 2008). Para evaluaciones sobre la calidad de un recurso, una estrategia utilizada es la observación del espacio donde el mismo fue publicado, o dicho de otro modo, se asocia la calidad de un artículo con la calidad de la revista que lo ha publicado. En este sentido, entre las principales fuentes de información de calidad de una revista se encuentran los índices y directorios que aplican procesos de evaluación de revistas y, cuando las revistas superan dichos procesos, las incorporan dentro de sus bases de datos. Existen aquí también muchos índices y directorios, que evalúan distintos aspectos de las revistas: licencias, transparencia, citaciones recibidas, procesos editoriales, entre otros. Algunos ejemplos de estos índices son Latindex, Scopus, DOAJ, Sherpa-ROMEO, entre otros.

Dada la heterogeneidad de indicadores descritos previamente, y por lo tanto la complejidad a la hora de evaluar un artículo o una revista científica, en el año 2021 se creó el proyecto HERA con el propósito de simplificar, agilizar y apoyar el proceso de determinar la calidad y el impacto de un recurso académico, como un artículo o una revista. Para ello, HERA recupera información en tiempo real sobre artículos y revistas proveniente desde múltiples fuentes y las integra en un espacio único. Esto no sólo evita tener que acceder a cada fuente de manera individual, sino que le permite al usuario poder visualizar los datos de manera unificada y en pocos segundos , realizar comparaciones entre indicadores similares obtenidos desde diferentes fuentes, y tomar decisiones a partir de los datos recolectados.

El funcionamiento de HERA es relativamente sencillo. Para obtener indicadores de artículos científicos, HERA requiere que se ingrese el DOI de dicho artículo[6] (Carletti , 2023) , siendo capaz de obtener datos provenientes de servicios como CrossRef, OpenAlex, DOAJ, Scopus, SemanticScholar, Dimensions, Altmetric y Google Scholar. Por otra parte, para obtener indicadores de revistas científicas, HERA requiere que se ingrese el ISSN de la revista, siendo capaz de obtener datos provenientes de CrossRef, OpenAlex, DOAJ, REDIB, Web of Science, Scopus y Scimago Journal Ranking (SJR)[7].

En su primera versión, HERA ofrecía a los usuarios el servicio de recolección y exposición de indicadores a través de su sitio web https://hera.sedici.unlp.edu.ar , en donde debían ingresar un DOI o un ISSN, y esperar unos segundos hasta que los datos fueran recopilados e integrados. A partir de la versión 2.0, HERA ofrece nuevas opciones para sus usuarios, incluyendo la posibilidad de realizar búsquedas múltiples (por ejemplo, enviando varios DOI a la vez), de descargar los datos obtenidos en archivos CSV o JSON, y de realizar solicitudes a través de una API REST. Esta última funcionalidad es de las más destacables, ya que abre un universo de posibilidades de integración de indicadores en cualquier sistema que gestione o exponga información de artículos científicos (o de cualquier recurso que posea un DOI, como por ejemplo un libro, una tesis o un dataset) y de revistas científicas. En este trabajo se mostrarán algunas de las posibilidades que brinda dicha API REST a partir de ejemplos reales en los cuáles ya ha sido integrada. Para ello, se describirán un conjunto de herramientas informáticas que permite aprovechar los servicios de HERA desde sistemas externos, y se mostrarán integraciones con el navegador web Google Chrome y los sistemas Wordpress y Open Journal Systems (OJS).

Breve Descripción de la API REST

La API REST de HERA en la versión 2.0 contribuyó a una mayor integración con aplicaciones y servicios externos, a la optimización del rendimiento en las solicitudes hechas a la aplicación y al aumento de la escalabilidad de la plataforma. Esta API REST puede ser utilizada a través de un único endpoint, sobre el cual es necesario enviar el identificador del recurso a buscar y el tipo de búsqueda que se quiere realizar (por DOI o por ISSN).

Como respuesta a la solicitud, HERA devuelve un archivo en formato JSON[8] . Cada clave representa un dato o indicador de un artículo o revista, por ejemplo "crossref_cites" que representa el número de citas que este artículo ha recibido según la base de datos Crossref; "abstract" que representa el resumen del artículo, o "sjr_all-time_best_quartile" que es una medida que determina la calidad e influencia de una revista

6 En el futuro HERA podría soportar otros indicadores, como ser HANDLE o ARK

7 Futuras versiones de HERA podrían incorporar otras bases de datos, tanto de artículos como de revistas)

8 Estandar ECM4 (Json) https://ecma-international.org/publications-and-standards/standards/ecma-404/. Accedido el 16/04/2024

académica. Además, si una fuente externa consultada por HERA devuelve un error, se muestra información sobre el mismo en el campo de la fuente correspondiente (p.e. en los casos que el recurso no sea encontrado o el servidor no esté disponible). La Figura 1 exhibe un ejemplo de respuesta en formato JSON para el requerimiento asociado al artículo DOI:10.1088/0034-4885/61/2/002

Oportunidades de Integración de HERA con Aplicaciones de Terceros

Prueba de concepto: Extensión de HERA para Google Chrome

Google Chrome es un navegador web desarrollado por Google, con la principal cuota del mercado de los navegadores, que ronda el 65% según datos de statista.com[9] así como también de StatCounter[10]. El desarrollo de Google Chrome está basado en el framework ElectronJS, que también ha servido como base para la implementación de otras aplicaciones como el navegador web Chromium, el entorno de desarrollo VS Code y el sistema de comunicación Discord, entre otros.

9 Cuota del mercado de los principales navegadores de internet https://es.statista.com/estadisticas/600249/cuota-de-mercado-mensual-de-los-principales-navegadores-de-internet/ . Accedido el 12/04/24

10 Browser Market Share Worldwide https://gs.statcounter.com/browser-market-share , Accedido el 15/4/24

```
[
 {
  "doi": "10.1088/0034-4885/61/2/002",
  "type": "article",
  "title": "Quantum computing",
  "authors": "Andrew Steane",
  "abstract": "...",
  "publication_year": "1998",
  "crossref_cites": 919,
  "openalex_doi_cites": 1065,
  "doaj_doi_presence": "No",
  "semanticscholar_cites": 250,
  "dimensions_cites": 983,
  "altmetric_cites": 27,
  "issn": "0034-4885",
  "journal_title": "Reports on progress in physics",
  "publisher": "IOP Publishing",
  "openalex_issn_cites": 373671,
  "crossref_dois": 2485,
  "doaj_issn_presence": "No",
  "redib_presence": "No",
  "wos_collections": "Essential Science Indicators,Science Citation Index Expanded,Research Alert,Journal Citation Reports Science,CC/Physical, Chemical & Earth Sciences",
  "scopus_citescore": "32.3 (2022)",
  "scopus_citescoretracker": "31.8 (2023)",
  "sjr_h-index": "247",
  "sjr_all-time_best_quartile": "Q1 (2023)"
 }
]
```


Figura 1. Ejemplo de documento JSON para el artículo 10.1088/0034-4885/61/2/002 - Quantum computing (el abstract fue eliminado por cuestiones de visibilidad)

Google Chrome posee un sistema de extensiones que permite ampliar su funcionalidad básica mediante la incorporación de pequeños programas o complementos. Las extensiones son diseñadas para añadir nuevas características al navegador, mejorar la productividad, proporcionar herramientas de seguridad, entre otras. Incluso, las extensiones permiten integrar al navegador con otras aplicaciones. Teniendo esto en cuenta, se creó la extensión de HERA para Google Chrome[11], que permite hacer búsquedas por DOI o ISSN sin la necesidad de ingresar al sitio web de HERA. Para ello, es necesario estar en una página que cuente con alguno de los identificadores antes mencionados y al hacer click en el ícono de la extensión, HERA obtiene el identificador del recurso y comienza a recolectar la información para luego mostrarla. La Figura 2 ilustra un ejemplo de uso de la extensión de HERA para Chrome para el artículo identificado como DOI: 10.1088/0034-4885/61/2/002.

Figura 2. Ejemplo de uso de la extensión de HERA para Chrome. Se recuperan indicadores y métricas del artículo 10.1088/0034-4885/61/2/002 - Quantum computing

En caso de que la página actual no cuente con ningún DOI o ISSN, entonces HERA notificará la situación al usuario. Es importante mencionar que la integración de HERA con un navegador como Chrome sirve como prueba de concepto de que es factible implementar extensiones similares para otros navegadores (p.e., Firefox o Safari).

11 https://chrome.google.com/webstore/detail/hera-browser-extension/cdmmeadgfiafkgpdiccepbmngbbfghcm/related?hl=es&authuser=0

Hacia la generalización: Conector genérico de HERA

La integración de HERA con sistemas web presenta desafíos particulares propios de cada sistema, como ser el uso de un lenguaje de programación particular o la implementación de una estructura particular compatible con el esquema de extensión que utiliza cada sistema (plugins, extensiones, módulos, componentes, etc.). Sin embargo, existen una serie de requerimientos funcionales comunes que cualquiera de estos sistemas web deberá implementar. Es por lo que se decidió diseñar un conector genérico de HERA para aplicaciones web. En esta sección se describe brevemente dicho conector, para luego detallar cómo ha sido integrado en desarrollos ad-hoc para Wordpress y para Open Journal System.

La motivación detrás de este desarrollo radica en la necesidad de simplificar y ampliar el acceso a las métricas proporcionadas por HERA en una amplia variedad de entornos y sistemas web. Uno de los principales requisitos que se plantearon para el conector, era que este sea liviano y altamente adaptable, a fin de permitir su incorporación en una variedad de aplicaciones web sin requerir una cantidad significativa de recursos de cómputo o humanos para su adopción. Para ello, se tomó un modelo de integración similar al utilizado por otros sistemas, que utilizan el lenguaje Javascript para embeber porciones de código ejecutable dentro del documento XHTML que las aplicaciones presentan a los usuarios finales. Ejemplos típicos es este modelo son los códigos de seguimiento para registro de analíticas web utilizados por sistemas como Matomo o Google Analytics, o para la incorporación de badges en servicios como Altmetric o Dimensions. Al igual que estas herramientas, el conector genérico de HERA también fue desarrollado en el lenguaje Javascript, y realiza las tareas de conexión con el servidor de HERA, recuperación de datos, procesamiento y presentación en forma de widget interactivo. Dada la diversidad de sistemas desde los cuales es posible utilizar este conector, se implementó una función especial, llamada *FetchHeraData,* que recibe el identificador de un recurso académico, su tipo (ISSN / DOI) a ser buscado por HERA, y un selector de un nodo del HTML donde se incorporará el widget interactivo; la figura 4 muestra el uso del widget interactivo en el contexto de una artículo y en el contexto de una revista. Todos los parámetros requeridos por esta función pueden indicarse como atributos del elemento HTML donde se insertará el widget interactivo, de la siguiente forma:

```
<div class="HeraConnector" DOI="10.4431/273922" ></div>
<div class="HeraConnector" ISSN="1111-2222"></div>
<div class="HeraConnector" ISSN="3322-2233" callback="someFunction"></div>
```

Como puede observarse en el ejemplo anterior, en caso que el usuario utilice el atributo ISSN o el atributo DOI, el conector reconocerá el tipo de identificador a utilizar y el valor del mismo. Además, en caso que el usuario desee implementar una función propia para procesar los resultados en lugar del widget interactivo, podrá especificarlo a través del atributo *callback*. Finalmente, los usuarios que deseen integrar el conector de HERA con sus aplicaciones, sólo deberán implementar las funciones de recuperación de los datos (particulares de cada aplicación) y generar el código HTML como en los ejemplos previos. La Figura 3 recrea el funcionamiento del conector de HERA para 2 aplicaciones externas diferentes.

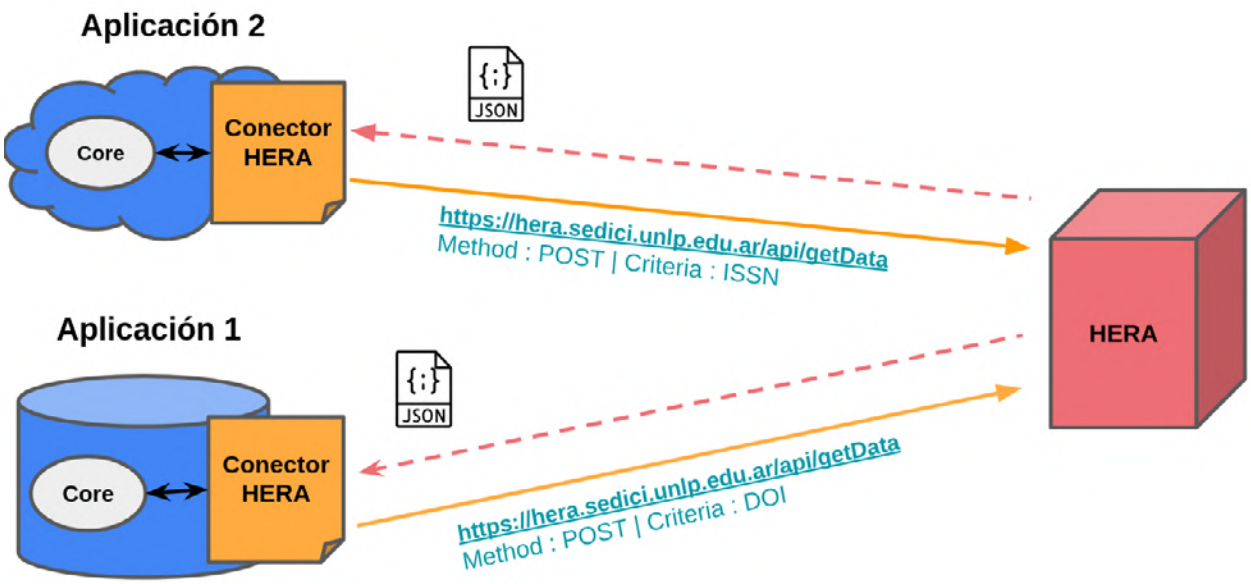


Figura 3. Recreación del funcionamiento del conector de HERA. La imagen muestra cómo el conector se integra a la aplicación usada y desde allí realiza consultas a la API de HERA.

A continuación, se presentan dos casos específicos de integración del conector genérico de HERA con los sistemas web Wordpress y OJS.

Caso de aplicación #1: Integración con Wordpress

Wordpress es un Sistema de Gestión de Contenido (CMS) de código abierto usado para crear sitios web, blogs o incluso aplicaciones. Según W3Tech[12], Wordpress es el CMS más utilizado a nivel mundial, principalmente por su facilidad de uso y su amplia comunidad de desarrolladores y usuarios. Ofrece una gran variedad de funcionalidades y características para que incluso un usuario con poca experiencia pueda desarrollar sus proyectos. En particular, WordPress cuenta con una gran cantidad de temas y plugins disponibles. Los plugins son herramientas adicionales que se integran con WordPress para añadir funcionalidades específicas al sitio web, como formularios de contacto, galerías de imágenes, herramientas de SEO, entre otros. Estos pequeños programas permiten a los usuarios personalizar y ampliar las capacidades de su sitio de manera sencilla y sin necesidad de tener conocimientos avanzados de programación. Por otro lado, los temas permiten a los usuarios modificar el diseño visual de su sitio web, incluyendo aspectos como la disposición de los elementos, los colores, las fuentes y la estructura general del sitio. La combinación de temas y plugins hace que WordPress sea una plataforma altamente flexible y adaptable.

12 Market share yearly trends for content management systems. https://w3techs.com/technologies/history_overview/content_management/ms/y . Visto el 12/4/24

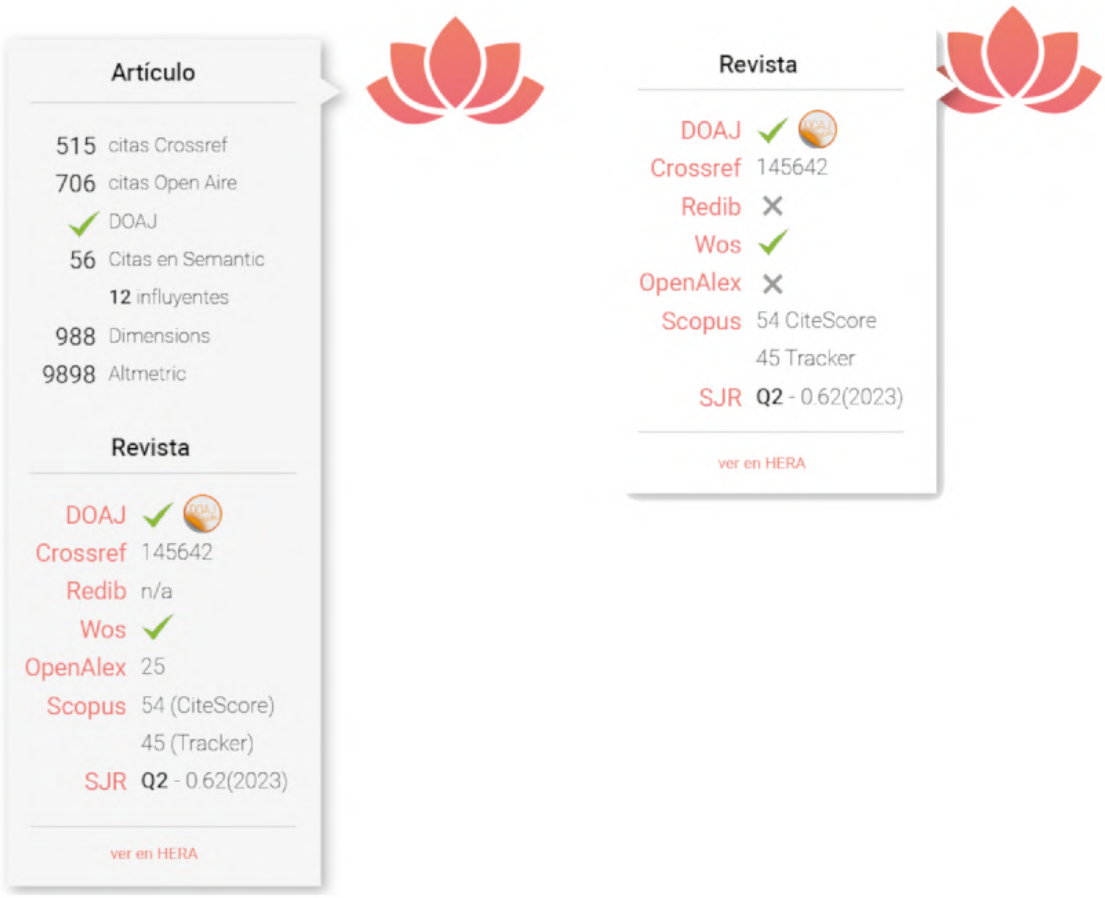


Figura 4. Indicadores y métricas presentados en el widget interactivo en el contexto de un artículo y de una revista respectivamente

Muchas instituciones académicas, como universidades, facultades o laboratorios, adoptan WordPress como su plataforma para gestionar sus sitios web, en los que presentan sus proyectos, describen sus actividades y difunden los resultados obtenidos. En este sentido, ya existen proyectos dedicados a mostrar recursos académicos y científicos en sitios web desarrollados con Wordpress (Villarreal, 2017; Villarreal, 2023). A modo de ejemplo, se tomará el Portal de Revistas de la UNLP, accesible desde https://portalderevistas.unlp.edu.ar/, para mostrar la integración entre Wordpress y HERA. Cabe aquí aclarar que la UNLP tiene varias instalaciones de OJS, y este portal web sirve como espacio centralizado para mostrar todas las revistas editadas por la UNLP en cualquiera de sus ámbitos; para cada revista el portal muestra información sobre el equipo editorial, bases de datos donde se encuentra indizada, año de inicio, ISSN y enlace a su sitio web. Esta información resulta suficiente para brindar un servicio de búsqueda simple, y para mostrar la cantidad y la variedad de publicaciones periódicas que se generan desde la UNLP, pero no incluye indicadores actualizados que podrían ser de gran utilidad para comprender el impacto de cada revista, como por ejemplo la cantidad de artículos publicados según OpenAlex, la clasificación Q1 a Q4 según SJR, o si ha obtenido el Sello DOAJ. Es por eso que se decidió integrar el sitio con la funcionalidad que ofrece HERA, a fin de adicionar a la información de cada revista indicadores sobre calidad e impacto de la misma.

En base a lo explicado anteriormente, se desarrolló un plugin de WordPress, denominado WP-Hera, que incorpora una nueva entidad llamada Revista (Custom Post Type[13] en términos de Wordpress), con varios campos (Custom Fields[14]) entre los que se incluye un campo que almacena el ISSN de la revista. El plugin brinda dos servicios esenciales al portal web: por un lado obtener y mostrar los metadatos de cada revista a los visitantes del sitio web, y por el otro lado mostrar indicadores asociados a cada revista. Para ello, se obtiene el valor del custom field ISSN de la revista que el usuario está visualizando, el cual se reenvía al endpoint de HERA. Esto genera como resultado un texto en formato JSON, el cual es procesado por una función Javascript en el navegador web que incorpora dinámicamente los datos contenidos en el JSON a la información de la revista. Asimismo, además de mostrar indicadores detallados en la página de cada revista, el plugin es capaz de mostrar indicadores resumidos en los listados de revistas (por ejemplo, luego de realizar una búsqueda); para ello, se hace uso de la capacidad de HERA de realizar búsquedas múltiples, a partir de un listado de identificadores (en este caso, un listado de ISSN).

Cabe destacar que este es sólo un ejemplo de integración con Wordpress, pero que podría fácilmente adaptarse a otros sitios web realizados con Wordpress, o incluso con cualquier CMS (como por ejemplo Drupal o Joomla!), gracias al uso de Javascript para la interacción con HERA, lo que brinda total independencia del lenguaje de programación con el cual funciona el CMS: PHP, Java, Python, etc.

Caso de aplicación #2: Integración con OJS

OJS es un sistema de gestión y publicación de revistas online, que fue lanzado en el año 2002 como un software de código abierto. OJS fue diseñado para gestionar el flujo de trabajo de una revista, desde el envío del manuscrito a través de la revisión hasta el trabajo editorial y finalmente la publicación. Al mismo tiempo, ofrece una forma sencilla de publicar una edición en línea y gestionar mejor los costos operativos de la revista. Para poder extender su utilidad, OJS utiliza un modelo basado en plugins al igual que Wordpress.

Una de las características más destacadas de OJS es su capacidad para mostrar mucha información relevante tanto para los artículos publicados como para las revistas que los alojan. En cuanto a los artículos, proporciona metadatos detallados, archivos del artículo, licencias y derechos de autor e incluso información sobre los autores. Además, ofrece información crucial sobre las revistas en sí, incluyendo detalles sobre los responsables editoriales, el alcance temático, los números publicados, las bases de datos donde ha sido indexada, etcétera.

En base a lo anterior, un usuario que acceda al sitio web público de una revista estará visualizando, por lo tanto, o bien la página de un artículo específico, o bien cualquier otra página de la revista. Esto genera dos posibles contextos (la revista o un artículo), lo que brinda la oportunidad de incorporar dos tipos de integraciones con HERA, una para cada contexto.

13 "WordPress comes with five default post types: post, page, attachment, revision, and menu. While developing your plugin, you may need to create your own specific content type…" (https://developer.wordpress.org/plugins/post-types/registering-custom-post-types/)

14 "WordPress has the ability to allow post authors to assign custom fields to a post. This arbitrary extra information is known as metadata." (https://wordpress.org/documentation/article/assign-custom-fields/)

Para el contexto *revista* se expone la cantidad total de artículos publicados, el CiteScore de Scopus, si se encuentra indexado en la Web Of Science, la clasificación Q1 a Q4 según SJR, y cualquier otro indicador a nivel de revista obtenido por HERA. Esta información podrá integrarse debajo de cada página que se presenta al usuario, en la barra lateral, como parte del pie de página o sólo en determinadas páginas.

Para el contexto *artículo* se expone el número total de citas según OpenAlex, el número de citas influyentes según Semantic Scholar, el valor de Altmetric, y cualquier otro indicador a nivel de artículo obtenido por HERA. Esta información puede mostrarse o bien de manera detallada dentro de la página del artículo (por ejemplo, debajo del espacio dedicado al resumen del artículo), o bien de manera resumida en la barra lateral.

Al momento de escribir este trabajo, la presentación de datos obtenidos por HERA dentro de OJS se encuentra en desarrollo, con lo cual aún deben tomarse decisiones sobre la localización y el formato en que se incorporarán los distintos indicadores en ambos contextos.

Conclusiones y Trabajos Futuros

En este trabajo se describieron algunos ejemplos de integración de HERA con diferentes aplicaciones y sistemas web. En primer lugar se detalló el funcionamiento de la extensión para Google Chrome, lo que sirve de base para implementar extensiones similares con otros navegadores como Firefox, Safari o Edge, y a su vez muestra la viabilidad de implementar integraciones de HERA con otras aplicaciones de escritorio que gestionen recursos académicos. Por ejemplo, podría diseñarse un módulo para la aplicación Mendeley Desktop[15], que incorpore métricas en vivo de los artículos que cada usuario almacenó en su propia base de datos bibliográfica, o implementar bots para Discord[16] , Slack[17] o Telegram[18], a los cuales se le pueden solicitar indicadores en vivo indicando simplemente el DOI o ISSN del recurso en cuestión.

En segundo lugar se describió un caso de uso específico, el Portal de Revistas de la UNLP, que funciona sobre el CMS Wordpress, y se detalló la integración de HERA con este CMS. Este ejemplo muestra las posibilidades de adaptar el conector genérico de HERA a aplicaciones web que no están necesariamente vinculadas al ámbito científico, como ser Wordpress. Podría tomarse la misma base para implementar integraciones con otros CMS, como por ejemplo Drupal o Joomla!, e incluso diseñar integraciones con otras familias de aplicaciones web más allá de los CMS. Una oportunidad interesante podría ser, por ejemplo, una integración con LMS (Learning Management Systems), como por ejemplo Moodle o Canvas, y así mostrar a quienes realizan capacitaciones o cursos a través de estos sistemas, indicadores vinculados a los materiales de lectura propuestos por los docentes o capacitadores.

Finalmente, se detalló el desarrollo que se está implementando para la integración de indicadores obtenidos por HERA en OJS. Más allá de las particularidades de OJS, lo interesante de este ejemplo es mostrar las posibilidades que ofrece el conector genérico de HERA para incorporar métricas en tiempo real en aplicaciones web específicas del ámbito académico y científico. Aquí las posibilidades son muy amplias, y dado que los recursos alojados en este tipo de aplicaciones poseen en muchos casos metadatos normalizados que incluyen identificadores (ISSN, DOI, HANDLE, ARK, PMCID, etc.), la integración con HERA podría generalizarse a cualquier instancia de una misma aplicación. Por ejemplo, podría diseñarse una extensión

15 https://www.mendeley.com/guides/desktop/
16 https://discord.com/
17 https://slack.com/
18 https://telegram.org/

al software para repositorios digitales DSpace que incorpore métricas en vivo para los recursos que aloja, y esta extensión podría utilizarse sobre cualquier instalación de DSpace. Esto también puede aplicarse a la familia de sistemas de información actualizada sobre investigación (Current research information system, CRIS), como por ejemplo DSpace-CRIS, a sistemas cosechadores y/o agregadores de recursos, como por ejemplo VIVO, VuFind o PKP Harvester, e incluso integrarse con sistemas basados en desarrollos propios como por ejemplo los repositorios de preprints arXiv, PubMed y RePEC.

Referencias

Carletti, E. (2023). HERA 2.0: Extensión de alcance y funcionalidad. [Tesis de grado]. Facultad de Informática, Universidad Nacional de La Plata. https://hdl.handle.net/10915/157417

Djiroun, R.; Lachachi, L. Y.; Eddine Azzouni, N. F.; Guessoum, M. A.; Boukhalfa, K. and Benkhelifa, E. H, "Search Approach for External Data Sources for Data Warehouse Enrichment in Business Intelligence Context," 2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA), Giza, Egypt, 2023, pp. 1-8, doi: 10.1109/AICCSA59173.2023.10479350.

Falagas, Matthew & Pitsouni, Eleni & Malietzis, George & Pappas, Georgios. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. FASEB journal : official publication of the Federation of American Societies for Experimental Biology. 22. 338-42. 10.1096/fj.07-9492LSF.

Gutiérrez, Yoan; Vázquez, Sonia; Montoyo, Andrés (2016). A semantic framework for textual data enrichment. Expert Systems with Applications, Volume 57, Sept. 2016.

Kavic, M. S., & Satava, R. M. (2021). Scientific literature and evaluation metrics: Impact factor, usage metrics, and altmetrics. *Journal of the Society of Laparoendoscopic Surgeons*, *25*(3), e2021.00010. https://doi.org/10.4293/jsls.2021.00010

Kim, K., & Chung, Y. (2018). Overview of journal metrics. *Science Editing*, *5*(1), 16–20. https://doi.org/10.6087/kcse.112

Lindsey, D. (1989). Using citation counts as a measure of quality in science measuring what's measurable rather than what's valid. Scientometrics, 15(3–4), 189–203.

Porto, J. F., Rucci, E. y Villarreal, G. L. (3-7 de octubre de 2022) .*HERA - Herramienta para Enriquecimiento de Recursos Académicos*. Actas de la XI Conferencia Internacional de Bibliotecas y Repositorios Digitales. Ibero-American Science and Technology Education Consortium. https://hdl.handle.net/10915/148922

Prayogi, A. A., Niswar, M., Indrabayu y Rijal, M. (2020). Design and Implementation of REST API for Academic Information System. *IOP Conference Series: Materials Science and Engineering,875*(1), 012047. https://dx.doi.org/10.1088/1757-899X/875/1/012047

Repiso, Rafael Cómo identificar una revista de calidad. Cardiocore [en línea]. 2015, 50(2), 46-48 [fecha de Consulta 6 de Agosto de 2021]. ISSN: 1889-898X. Disponible en: https://www.redalyc.org/articulo.oa?id=277041630002

Rettore, P. H. L.; Santos, B. P.; Rigolin R.; Lopes, F.; Maia, G.; Villas, L. A.and Loureiro A. A. F. "Road Data Enrichment Framework Based on Heterogeneous Data Fusion for ITS," in IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 4, pp. 1751-1766, April 2020, doi: 10.1109/TITS.2020.2971111.

Villarreal, G. L., Manzur, E., Vila, M. M. y De Giusti, M. R. (2017). *Interoperabilidad con repositorios digitales: uso de OpenSearch en sitios web institucionales*. En VII Conferencia Internacional sobre Bibliotecas y Repositorios Digitales de América Latina (BIREDIAL-ISTEC'17) y XII Simposio Internacional de Biblioteca Digitales (SIBD'17) (La Plata, 2017). https://hdl.handle.net/10915/63566

Villarreal, Gonzalo Luján; Terrone, Pablo Gabriel; de Albuquerque, Pablo César; De Giusti, Marisa Raquel (2023). Análisis de escenarios y protocolos para integración de repositorios digitales y sitios web institucionales. En XII Conferencia Internacional sobre Bibliotecas y Repositorios Digitales, octubre 2023. https://hdl.handle.net/10915/161920

Lautaro Josin Saller (https://orcid.org/0009-0005-7820-7962) es estudiante avanzado de Licenciatura de Informática por la Facultad de Informática de la Universidad Nacional de La Plata. Miembro del grupo de desarrollo y gestión de la red de sitios web de Unidades de Investigación y Desarrollo de la Universidad Nacional de La Plata.

Pablo G. Terrone (https://orcid.org/0009-0007-7700-827X) , Analista en Computación y estudiante avanzado de Licenciatura de Sistemas. Miembro del grupo de desarrollo y gestión de la red de sitios web de Unidades de Investigación y Desarrollo de la Universidad Nacional de La Plata, y docente de la Facultad de Informática de la Universidad Nacional de La Plata

Ezequiel Carletti es Analista Programador Universitario y estudiante de Licenciatura en Informática por la Facultad de Informática de la Universidad Nacional de La Plata. En el año 2023 presentó su tesis de grado "HERA 2.0: Extensión de alcance y funcionalidad", proyecto realizado bajo la dirección del Dr. Enzo Rucci y la codirección del Dr. Gonzalo Villarreal, y en colaboración con la Facultad de Ciencias Económicas, la Facultad de Informática y PREBI-SEDICI, dependientes de la UNLP.

Enzo Rucci (https://orcid.org/0000-0001-6736-7358) es Doctor en Ciencias Informáticas por la Facultad de Informática de la Universidad Nacional de La Plata, docente-investigador de la UNLP en las áreas relacionadas con procesamiento concurrente y paralelo, e investigador adjunto de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires. Forma parte del Instituto de Investigación en Informática LIDI (III-LIDI, UNLP-CIC), donde realiza sus actividades de investigación en temáticas vinculadas a Cómputo de Alto Rendimiento y Aplicación de TICs a Ciencias de la Vida y Bibliometría. Es co-editor del Journal of Computer Science and Technology.

Gonzalo Luján Villarreal (https://orcid.org/0000-0002-3602-8211) es Doctor en Ciencias Informáticas por la Facultad de Informática de la Universidad Nacional de La Plata, es docente-investigador de la UNLP y director del Centro de Servicios en Gestión de Información (CESGI) de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires. Desarrolla su actividad docente en grado y posgrado de la Facultad de Informática de la UNLP. Trabaja con la plataforma OJS desde el año 2008, cuando en el marco de PREBI-SEDICI la UNLP lanzó el Portal de Revistas de la UNLP. En la actualidad es coordinador técnico de revistas científicas de la UNLP, y brinda asesoramiento a organizaciones y equipos editoriales temas vinculados a publicaciones científicas, circuitos editoriales y gestión de OJS.

El desarrollo de Sistemas de Gestión de la Investigación (CRIS) en América Latina y el Caribe: Estudio 2021-2024

Rosalina Vázquez Tapia[1]

Resumen

Los Sistemas de Gestión de la Investigación, conocidos por sus siglas en inglés como RIM (Research Managment System) o CRIS (Current Research Information System), son herramientas informáticas que permiten integrar, organizar, analizar y gestionar la información científica de una institución u organismo de investigación. Con el propósito de establecer un estado de la cuestión sobre el desarrollo de Sistemas CRIS/RIM en países de América Latina y el Caribe, durante el periodo 2021-2024 se llevó a cabo un estudio diagnóstico que comprendió dos fases, la primera, mediante una encuesta en línea para identificar sistemas CRIS institucionales, y la segunda, utilizando el método de estudio de casos con la finalidad de analizar y describir los sistemas CRIS/RIM de alcance nacional, incluyendo plataformas similares. Entre los resultados, se identificaron y analizaron un total de 62 sistemas CRIS Institucionales en 14 países de la región, 5 sistemas CRIS/RIM nacionales y 2 plataformas para la gestión y evaluación de la investigación y/o producción científica nacional. Este trabajo forma parte de una investigación doctoral y tiene como objetivo presentar la metodología y resultados globales del estudio diagnóstico, identificando las características generales de los diferentes sistemas y plataformas, de acuerdo con un conjunto de criterios establecidos.

Abstract

Research Management Systems (RIM) or CRIS (Current Research Information System) are computer tools that allow the integration, organization, analysis, and management of the scientific information of a research institution or organization. In order to establish a state of the art on the development of CRIS/RIM systems in Latin American and Caribbean countries, a diagnostic study was carried out that comprised two phases during the period 2021-2024, the first through an online survey to identify institutional CRIS systems, and the second, using the case study method in order to analyze and describe CRIS/RIM systems of national scope, including similar platforms. Among the results, a total of 62 institutional CRIS systems were identified and analyzed in 14 countries of the region, 5 national CRIS/RIM systems, and 2 platforms for the management and evaluation of national research and/or scientific production. This paper is part of a doctoral research and aims to present the methodology and global results of the diagnostic study, identifying the general characteristics of the different systems and platforms, according to a set of established criteria.

Palabras claves/Keywords

Sistema de Gestión de la Investigación, CRIS, RIM, Sistema de Producción Científica Nacional, Infraestructuras abierta, Encuesta, Estudio de caso, Latinoamérica.

Current Research Information System, Research Information Manager, CRIS, RIM, National Science Production System, Open Infrastructures, Survey, Case Study, Latin-America.

1 Universidad Autónoma de San Luis Potosí, México, alinavn@uaslp.mx

Introducción

La adecuada gestión de la actividad científica y de sus productos derivados, facilita y promueve el desarrollo y divulgación de la ciencia. En décadas pasadas, donde las herramientas informáticas eran escasas o muy básicas, la información institucional sobre las actividades, proyectos, resultados y productos de la investigación, era almacenada y procesada de manera desarticulada y fragmentada, en diferentes sistemas, archivos y bases de datos; lo que dificultaba la recopilación, análisis y compartición de los datos e información científica para efecto de financiamiento, optimización de recursos, visibilidad y toma de decisiones.

En este sentido, los RIM (Research Managment System), más comúnmente conocidos como CRIS (Current Research Information System), surgen como una herramienta para integrar, organizar, analizar y gestionar la información científica de una institución u organismo de investigación.

Un RIM es la agregación, curación y utilización de metadatos sobre actividades y productos de investigación; así como también, de sus investigadores e instituciones de afiliación; publicaciones, conjuntos de datos y patentes; proyectos y fondos de investigación; premios y distinciones académicas (Bryant et al., 2017)

Un CRIS es un sistema de información para almacenar, administrar, gestionar e intercambiar datos e información, sobre actividades, resultados y productos de investigación generados por las instituciones y sus comunidades de investigadores. Entre los ámbitos de aplicación y/o funciones principales de un RIM o CRIS, se pueden mencionar las siguientes: Gestión de proyectos e identificación de oportunidades de adjudicación; gestión de publicaciones; gestión y publicación de perfiles de investigadores; generación de reportes e indicadores de producción científica; cumplimiento de mandatos de Acceso Abierto; integración de datos de diferentes sistemas internos y externos (Dempsey, 2014).

Una característica fundamental de un sistema CRIS es su interoperabilidad con otras fuentes de información internas y externas, para vincular (enlazar), intercambiar, importar o exportar metadatos, e inclusive archivos, entre diferentes sistemas, aplicaciones y bases de datos. Particularmente, se busca vincular (enlazar) los perfiles de los investigadores con sus publicaciones y productos de investigación (patentes, proyectos, informes, etc.) depositados en los repositorios institucionales y/o los portales de revistas y libros, para evitar duplicidad, promover el autoarchivo y facilitar los flujos de trabajo entre las diferentes plataformas institucionales.

Para el intercambio de información entre sistemas CRIS y con otras infraestructuras abiertas de investigación, la organización EuroCRIS , una asociación internacional sin fines de lucro fundada en 2002 que reúne a expertos en investigación y sistemas CRIS, desarrolló el estándar CERIF (Common European Research Information Format), un modelo relacional descriptivo creado originalmente bajo el auspicio de la Comunidad Europea, que define un formato en XML para la descripción e intercambio de metadatos e información de investigación entre y con sistemas CRIS (The International Organization for Research Information [EuroCRIS], s.f.-a)

Además, EuroCRIS proporciona el servicio de DRIS (Directory of Research Information System), un directorio mundial de sistemas CRIS de instituciones, agencias de investigación y otros organismos (EuroCRIS, s.f.-b) y el Repositorio EuroCRIS, implementado con el software open source DSpace-CRIS, donde se depositan las publicaciones de sus eventos, reuniones anuales y de la conferencia bianual CRIS Conference (EuroCRIS, s.f.-c)

Adicionalmente, OpenAIRE[2] incluye como parte de sus directrices de interoperabilidad, la Guía para administradores de sistemas CRIS (OpenAIRE Guidelines for CRIS Managers v1.2.0) para la recolección (harvesting) e importación de metadatos de sistemas CRIS. Por su parte, COAR[3] promueve la interoperabilidad de los Sistemas CRIS con los Repositorios Institucionales como una buena práctica para el poblamiento y sostenibilidad de los repositorios.

Para la implementación de un sistema CRIS/RIM, entre las soluciones de software comerciales, una de las primeras fue SYMPLETIC, desarrollada en 2004 y que ahora forma parte de Digital Science; PURE desarrollado por Atira y adquirido en 2012 por Elsevier; y CONVERIS comercializado por Thomson Reuters (ahora Clarivate Analytics) desde 2013. Dentro de las soluciones de open source, se encuentran DSpace-CRIS, desarrollado por 4Science y VIVO del DuraSpace de LYRASIS.

En cuanto a su alcance, los sistemas CRIS se pueden clasificar de cinco tipos: 1) Institucional, cuando organizan la información científica de una institución; 2) Nacional, cuando gestionan la información científica de instituciones y fuentes de información de un país; 3) Regional, que gestionan información o integran sistemas de varias instituciones pertenecientes a una región geográfica; 4) Internacional, que integran y gestionan información y/o sistemas de diferentes países; 5) Financiador, creados por las agencias u organismos de financiamiento.

Justificación y objetivo de la propuesta

El objetivo de este trabajo propuesto para ponencia, es presentar los objetivos de investigación, metodología, desarrollo y resultados generales de un estudio diagnóstico sobre el desarrollo de Sistemas CRIS en países de América Latina y el Caribe, efectuado en el periodo 2021-2024, el cual forma parte de los objetivos de investigación establecidos para la elaboración de la Tesis Doctoral "Modelo para el desarrollo de un Sistema de Gestión de la Investigación (CRIS) interoperable con un Repositorio Institucional de Acceso Abierto" del Programa de Doctorado en Formación en la Sociedad del Conocimiento de la Universidad de Salamanca, España. Por tanto, se trata de una investigación doctoral.

Uno de los objetivos de la investigación doctoral, es establecer un estado de la cuestión sobre el desarrollo de Sistemas de Gestión e Investigación (CRIS/RIM) institucionales y nacionales en países de América Latina y el Caribe, con la finalidad de identificar las características de los sistemas CRIS y plataformas similares, en aspectos técnicos, de gestión y contenidos, así como también, conocer los diferentes modelos de implementación, experiencias y buenas prácticas.

Para ello, se llevó a cabo un estudio diagnóstico que comprendió dos fases, la primera, enfocada al desarrollo de sistemas CRIS institucionales y llevada a cabo durante el periodo 2021-2022 mediante una encuesta en línea, y la segunda efectuada en 2023-2024, utilizando el método de estudio de casos con la finalidad de analizar y describir los sistemas CRIS/RIM de alcance nacional, incluyendo plataformas similares para la gestión y evaluación de la investigación y/o producción científica.

2 OpenAIRE es una organización sin fines cuya misión es promover la investigación abierta y mejorar el descubrimiento, la accesibilidad, compartición, reutilización, reproducibilidad y el seguimiento de los resultados de la investigación basada en datos, a escala mundial. https://www.openaire.eu/

3 COAR (Confederation of Open Access Repositories) es una asociación internacional con más de 130 miembros y socios de todo el mundo, que trabajan para crear capacidad, alinear políticas y prácticas, y actuar como una voz global para la comunidad de repositorios y redes de repositorios. https://coar-repositories.org/

En los siguientes apartados, se describen los objetivos de investigación de cada fase del estudio diagnóstico referido con anterioridad, así como la metodología, desarrollo y resultados globales obtenidos.

Encuesta diagnóstica 2021-2022

El objetivo general de la Encuesta diagnóstica fue identificar que instituciones contaban con un sistema de gestión de la investigación (CRIS/RIM) y sus características en determinados aspectos, así como también, identificar a las instituciones que no lo tenían y a las que estuvieran considerando su su desarrollo en el mediano o largo plazo. Por tanto, estuvo dirigida a las instituciones de educación superior, centros de investigación, redes u organismos nacionales de investigación de 21 países de América Latina y el Caribe, que tuvieran o no un sistema CRIS/RIM.

Objetivos

1. De manera específica, los objetivos de la encuesta diagnóstica fueron los siguientes:
2. Identificar a las instituciones de educación superior o centros de investigación que hayan implementado sistemas CRIS.
3. Identificar a las instituciones que no cuentan con un sistema CRIS, y de ellas, cuantas tienen planeado su implementación o cuentan con un proyecto en desarrollo.
4. Determinar las características de los sistemas CRIS implementados en las instituciones, en aspectos de contenidos, componentes y funcionalidad.
5. Identificar las soluciones de software, estándares de metadatos, directrices de interoperabilidad e identificadores digitales persistentes, que más son utilizadas para la implementación de los sistemas CRIS.
6. Determinar los tipos de interacción o interoperabilidad de los sistemas CRIS con fuentes de información internas y externas, así como los protocolos de intercambio de información empleados para ello.
7. Identificar cuáles son las instancias y perfiles del personal a cargo de la gestión y administración de los sistemas CRIS, así como los tipos de financiamiento y si cuentan con políticas y procedimientos institucionales.

Metodología

Para determinar el grupo meta y seleccionar una muestra por país, se diseñó un perfil de institución que debía cumplir con al menos dos de los siguientes criterios:

a) Sólo instituciones de educación superior, centros de investigación u organismos de ciencia y tecnología.

b) Ser Miembro de alguna de las Redes Nacionales de Investigación y Educación de la Red Latinoamericana de Educación e Investigación Red CLARA[4], la cual cuenta con actualmente con 10 miembros (Redes Nacionales): Brasil-RNP, Colombia-RENATA, Costa Rica-Red CONARE, Chile-REUNA, Ecuador-CEDIA, Guatemala-RAGIE, Honduras-RedNESAH, México-CUDI, Nicaragua-Red RUNBA, Uruguay-RAU.

4 RedCLARA - National Research and Education Networks of the Latin American Cooperation in Advanced Networks. https://www.redclara.net/index.php/en/

c) Contar con un Repositorio Institucional cosechado, a través del nodo nacional, por la Red Latinoamérica de Repositorios de Producción Científica LA Referencia[5]

d) Tener registrado un Repositorio Institucional en el Directorio OpenDOAR

e) Tener registrado un Sistema CRIS en el Directorio DRIS (Directory of Research Information System) de EuroCRIS

f) Contar con una instalación del software VIVO en el registro de DuraSpace de LYRASIS[6]

g) Contar con una instalación del software DSpace-CRIS[7] en el wiki de 4Science

h) Contar con una instalación del software Pure de Elsevier [8]

Estos criterios de selección fueron establecidos, partiendo de la hipótesis de que las instituciones de educación superior e investigación con un repositorio institucional tendrían mayor probabilidad de contar con un sistema CRIS, que aquellas que fueran de nivel básico o medio superior, o bien, que no contaran con un repositorio institucional o disciplinar. Con base en estos criterios, se obtuvo un grupo meta inicial.

Para determinar la muestra por país, se construyó una tabla con las listas de instituciones que cumplían con cada uno de los criterios, partiendo de las que pertenecen a una RNIE de Red CLARA o bien, a un nodo o repositorio nacional cosechado por LA Referencia. Enseguida, se agregaron las instituciones que tenían un Repositorio Institucional o disciplinar registrado en OpenDOAR. Posteriormente, se aplicaron el resto de los criterios, haciendo una búsqueda de la institución en el DRIS y en los sitios de las diferentes soluciones de software. Como resultado, se obtuvo una tabla por país con la relación de instituciones que cumplen con los diferentes criterios, de las cuales se generaron dos versiones, una con los datos de contacto localizados en el sitio web de la institución o del repositorio, y la otra con la indicación si había respondido o no la encuesta. En la Tabla 1 se muestra el número de instituciones por país que cumplieron con cada uno de los criterios.

Tabla 1. Muestra total final del número de instituciones por país que cumplen con los criterios de selección (abril 2022)

No.	Country	Red CLARA	Repositorio Nacional	Open DOAR	LA Referencia	Euro CRIS	DRIS	VIVO	DSpace-CRIS	Elsevier PURE	Total Ins./IR
1	Argentina		41	43	33		2		3		53
2	Bolivia			3						1	3
3	Brasil	52	55	90	20	3	1	3			116
4	Chile	18		13	1	1	2	2		4	27
5	Colombia	32		65	36	1	4	2		4	80
6	Costa Rica	7		8	8		2	2			9
7	Cuba			8		1		1	1		11

5 Federated Network of Institutional Repositories of Scientific Publications. https://www.lareferencia.info/en/

6 DuraSpace Registry - Duraspace.org. https://duraspace.org/registry/

7 DSpace-CRIS Users - DSpace-CRIS – LYRASIS. https://wiki.lyrasis.org/display/DSPACECRIS/DSpace-CRIS+Users

8 Elsevier Pure Client portals. https://www.elsevier.com/solutions/pure/pure-in-action

8	Ecuador	41		34	22	2	1			1	54
9	El Salvador		13	8	7						16
10	Guatemala	3		1							3
11	Honduras			4				1			4
12	Jamaica			3							3
13	México	96		34	69	1	7	4	2	5	116
14	Nicaragua	9		9							12
15	Panamá			5							5
16	Paraguay			1							1
17	Perú		144	126	52	3	7		2	8	144
18	Puerto Rico			1							1
19	República Dominicana			2					2		4
20	Uruguay	8		5	5						8
21	Venezuela			6							6
	Totales	266	253	469	253	12	26	15	10	23	676

Nota: Adaptado de "Development and characterisation of CRIS systems in Latin America: Preliminary results of diagnostic survey" (p. 270-271), por R. Vázquez, 2022, *Procedia Computer Science, 211*(2022).

La metodología en extenso, así como los resultados preliminares de este primer estudio, fueron presentados en la *15th International Conference on Current Research Information System CRIS2022* (Vázquez, 2022-a) y en la edición especial de la revista científica *Procedia Computer Science* (Vázquez, 2022-b).

Diseño del instrumento diagnóstico y aplicación de la encuesta

El diseño del instrumento se llevó a cabo en dos fases. La primera fue una prueba piloto con la finalidad de validar el instrumento con un grupo seleccionado de 23 expertos y responsables de sistemas CRIS de nueve países de Iberoamérica: Argentina, Brasil, Ecuador, España, Colombia, Costa Rica, Perú, Portugal y México. Con base en los resultados y retroalimentación de los participantes, se realizaron modificaciones menores que contribuyeron a la mejora del instrumento.

El formulario final comprendió 32 preguntas distribuidas en 7 dimensiones o secciones:

a. Sección I: Datos de la Institución. Comprendió 6 preguntas sobre datos generales de la institución y de la persona de contacto a cargo de responder la encuesta.
b. Sección II: Datos del Sistema de Gestión de la Investigación (CRIS/RIM/RMA). Comprendió 5 preguntas, una de ellas de control, para discriminar a las instituciones que no contaban con un sistema CRIS, en cuyo caso se transfería hacia la última sección de la encuesta.
c. Sección III: Características del Sistema de Gestión de la Investigación (CRIS/RIM/RMA). Comprendió 6 preguntas acerca del alcance, etapa de desarrollo, tipos de contenido y funciones principales del sistema CRIS.

d. Sección IV: Infraestructura tecnológica del Sistema de Gestión de la Investigación (CRIS/RIM/RMA). Consistió en 10 preguntas enfocadas a identificar la solución de software implementada, el uso de estándares de metadatos, directrices de interoperabilidad, identificadores digitales persistentes y los tipos de interacción o interoperabilidad del sistema CRIS con fuentes de información internas o externas.
e. Sección V: Gestión y administración del Sistema de Gestión de la Investigación (CRIS/RIM/RMA). Comprendió 6 preguntas acerca de la unidad o departamento responsable del sistema CRIS, los perfiles y número de personas dedicadas a la gestión del sistema, los tipos de políticas de operación y fuentes de financiamiento para el desarrollo y mantenimiento del sistema.
f. Sección VI: Colaboración interinstitucional. Comprendió 3 preguntas. Se les preguntó si la institución participaba en un proyecto nacional o regional y cuáles eran las siguientes etapas de desarrollo o acciones de mejora de su sistema CRIS.
g. Sección VII: Aviso de confidencialidad. Consistió en 2 preguntas para autorizar el uso de los datos para fines académicos y una pregunta abierta para comentarios o sugerencias

El formulario fue implementado utilizando la herramienta Forms de Microsoft Office. La convocatoria para responder a la encuesta estuvo abierta del 1º de julio al 20 de noviembre de 2021. Para la difusión fueron utilizados diferentes medios y estrategias, entre ellos, redes sociales, eventos virtuales, listas de correo de diferentes redes de acceso abierto y repositorios, así como también, el envío de invitaciones a los contactos de los sitios web de las instituciones, repositorios o sistemas del grupo meta (más de 500 correos electrónicos enviados).

Resultados

De acuerdo con Vázquez (2022-b), se recibieron 140 respuestas, de las cuales fueron eliminadas 7 duplicadas, quedando un total de 133 respuestas válidas. Sólo el 23% de las instituciones encuestadas, que equivale a un total de 31 instituciones, respondieron que Sí cuentan con un sistema CRIS. Del 77% restante, el 56% tienen planeado en el mediano o largo plazo, el desarrollo de un sistema CRIS en su institución.

Por otro lado, se encontraron instituciones que cuentan con una instalación de algún software diseñado para la gestión de sistemas CRIS, pero que no respondieron la encuesta. Con la finalidad de establecer un panorama más completo sobre el desarrollo de sistemas CRIS en cada uno de los países, se elaboró una tabla de instituciones por país con sistemas CRIS, conjuntando los resultados de la encuesta con los criterios de selección; es decir, se sumaron las instituciones que habían respondido que sí tienen un sistema CRIS con las instituciones que tienen un sistema registrado en DRIS, o bien, una instalación del software VIVO, DSpace-CRIS o PURE (Vázquez, 2022-b).

Como resultado, se identificaron un total de 65 sistemas CRIS (tres nacionales y el resto institucionales), de los cuales el 34% utilizan el software comercial PURE, el 22% el software open source VIV0, seguido de DSpace-CRIS con un 18% y un 26% utilizan otras soluciones (desarrollo propio, hibrido o por convenio). Los países con mayor número de sistemas CRIS fueron Perú y México (Vázquez, 2022-b). En la Tabla 2 se muestra una síntesis de los resultados, considerando tanto los resultados de la encuesta, como la aplicación de los criterios de selección.

Tabla 2. Número de encuestas respondidas y sistemas CRIS identificados por país (abril 2022)

No.	Country	Respuesta Encuesta	CRIS Instit.	Euro CRIS	DRIS	VIVO	DSpace-CRIS	Elsevier PURE	Otras soluciones
1	Argentina	8	6		2		3		2
2	Bolivia	0	1					1	
3	Brasil	13	5	1	1	3			2
4	Chile	9	7		2	2		4	1
5	Colombia	11	6	1	4	2		4	
6	Costa Rica	4	4		2	2			3
7	Cuba	5	3	1		1	1		2
8	Ecuador	11	5	1	1			1	4
9	El Salvador	3	0						
10	Honduras	0	1			1			
No.	Country	Respuesta Encuesta	CRIS Instit.	Euro CRIS	DRIS	VIVO	DSpace-CRIS	Elsevier PURE	Otras soluciones
11	México	34	12	1	7	4	3	5	2
12	Nicaragua	5	0						
13	Panamá	1	0						
14	Paraguay	1	1				1		
15	Perú	24	11	1	7		2	8	
16	República Dominicana	1	2				2		
17	Uruguay	2	1						1
18	Venezuela	1	0						
	Totales	133	65	6	26	15	12	23	17

Estudio de casos 2023-2024

El objetivo general del Estudio de casos fue establecer un panorama general sobre el desarrollo de Sistemas CRIS/RIM de alcance nacional en países de América Latina y el Caribe, identificando quienes cuentan con un sistema de este tipo y cuáles son sus principales características, o bien, que hayan desarrollado una plataforma similar para la gestión y evaluación de la investigación y/o producción científica nacional. El estudio fue llevado a cabo en 2023 y actualizada a la fecha (abril de 2024) y estuvo dirigido a los consejos o agencias nacionales de ciencia y tecnología y redes **nacionales de investigación.**

Objetivos

Los objetivos específicos del estudio de casos fueron los siguientes:

1. Identificar a los países de América Latina que tienen un proyecto o sistema de gestión de la investigación o portal de producción científica a nivel nacional y en que consiste, su alcance, objetivos, ejes de desarrollo, servicios y beneficios.

2. Determinar las características generales de los proyectos o sistemas CRIS en términos operativos, considerando aspectos técnicos, normativos, metodológicos y de gestión de información.
3. Realizar un análisis y descripción de cada caso de acuerdo con las variables e indicadores establecidos, identificando las similitudes o aspectos en común y los contrastes entre los diferentes casos.
4. Determinar los procesos de implementación, resultados, áreas de oportunidad y buenas prácticas de cada estudio de caso, que aporten a una propuesta de modelo para el diseño **e implementación de un sistema CRIS.**

Metodología

Para realizar el estudio diagnóstico sobre el desarrollo de sistemas CRIS nacionales en América Latina se utilizó el método de estudio de caso de tipo cualitativo descriptivo.

Martínez (2006) plantea cinco pasos para el diseño del estudio de caso:

a) Establecer las preguntas de investigación y las proposiciones teóricas que servirán de referencia o punto de partida. Ambos elementos constituyen los constructos (conceptos, dimensiones, factores o variables) de los cuales se obtendrá la información.
b) Explicitar las diferentes fuentes de información para la recolección y análisis de los datos.
c) Diseñar los instrumentos que han de utilizarse para la recolección de los datos. En este punto se recomienda utilizar "el protocolo de estudio de caso" que contiene los siguientes elementos: Semblanza del estudio de caso, preguntas del estudio de caso, procedimientos a ser realizados y guía del reporte de estudio de caso.
d) Analizar los datos obtenidos y vincularlos con las proposiciones teóricas y objetivos de investigación.
e) Presentar los resultados de la investigación a través de una serie de conclusiones.

Aplicando esta metodología, se llevaron a cabo las siguientes actividades:

a) Validación de los casos de sistemas CRIS nacionales y proyectos similares, que fueron identificados durante la primera fase del estudio diagnóstico (2021-22), mediante comunicaciones personales con los líderes o responsables en cada caso. Como resultado, se confirmó la continuidad de los proyectos de Argentina, Brasil, Chile, Ecuador y Perú; pero se descartaron las iniciativas de Colombia y Costa Rica que no prosperaron.
b) Recopilación de datos y análisis de información mediante la exploración en los sitios web de los organismos nacionales de ciencia y tecnología del resto de los países miembros de LA Referencia y de Red CLARA; así como también, de DRIS y soluciones de software (Pure, VIVO, DSpace-CRIS y Sympletic). Como resultado, se identificaron dos nuevos sistemas CRIS/RIM nacionales, en el caso de Panamá y Puerto Rico.
c) Diseño de una guía para la recopilación de información que fue enviada por correo electrónico a las personas responsables o de contacto de cada uno de los proyectos identificados. Como resultado, 5 de los 6 participantes completaron la guía y enviaron su información.
d) Compilación de datos, documentos, fuentes de información y referencias bibliográficas de cada caso, con base en la guía.

e) Análisis documental y descriptivo de cada estudio de caso, considerando un conjunto de factores e indicadores, en relación con la descripción y alcance del proyecto, servicios y beneficios, normativa y gestión de información, y plataforma tecnológica.

Resultados

Como resultado de la metodología anteriormente descrita, se determinaron 7 casos de estudio, 5 de ellos corresponden a Sistemas CRIS Nacionales y 2 plataformas para la gestión y evaluación de la investigación y/o producción científica nacional.

Los sistemas CRIS nacionales con mayor alcance y desarrollo son #PerúCRIS de Perú y BrCRIS de Brasil. Ambos están implementados con plataformas open source, tecnologías abiertas e interoperables con sus repositorios, bases de datos, portales de revistas y fuentes externas. Están diseñados con la visión de construir ecosistemas de información científica; CONCYTEC e IBICT son socios estratégicos de EuroCRIS y de otros importantes organismos internacionales.

A excepción de Puerto Rico, el resto de los países con sistemas CRIS nacionales utilizan soluciones de software open source (DSpace-CRIS y VIVO), tecnologías abiertas y el estándar CERIF como parte de su arquitectura.

En la Tabla 3 se describe los nombres, sitios y tipos de sistema, así como el organismo responsable en cada caso.

Tabla 3. Casos de estudio de Sistemas de gestión de la investigación Nacionales en América Latina

No.	País	Tipo	Nombre y sitio del CRIS/RIM/ Plataforma nacional	Organismo responsable
1	Argentina	Plataforma	SIGEVA - Sistema Integral de Gestión y Evaluación https://sigeva.conicet.gov.ar/	Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
2	Brasil	CRIS	BrCRIS – Ecosistema do Pesquisa Brasileira. https://brcris.ibict.br/	Instituto Brasileño de Información en Ciencia y Tecnología (IBICT)
3	Chile	Plataforma	DATACIENCIAS – Dimensiones de la Producción Científica Nacional. https://dataciencia.anid.cl/	Agencia Nacional de Investigación y Desarrollo (ANID)
4	Ecuador	CRIS	REDI - Repositorio Semántico de Investigadores del Ecuador. https://redi.cedia.edu.ec/	Corporación Ecuatoriana para el Desarrollo de la Investigación y la Academia (CEDIA)
5	Panamá	CRIS	CONECTO – Perfiles de la Ciencia y Tecnología de Panamá. http://conecto.senacyt.gob.pa/conecto/	Secretaria Nacional de Ciencia, Tecnología e Innovación (SENACYT)
No.	País	Tipo	Nombre y sitio del CRIS/RIM/ Plataforma nacional	Organismo responsable

6	Perú	CRIS	#PerúCRIS https://perucris.concytec.gob.pe/perucris/presentacion	Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC)
7	Puerto Rico	RIM	BEACON – Sistema de Gestión de Perfiles de Investigadores de Puerto Rico https://prsciencetrust.org/beaconplatform/	Fideicomiso de Ciencia, Tecnología e Investigación de Puerto Rico

En este trabajo se presentan los resultados preliminares. La metodología en extenso, así como el análisis y descripción de los casos de estudio, serán presentados en la *16th International Conference on Current Research Information System CRIS2024* (Vázquez, 2024).

Conclusiones

El desarrollo de sistemas CRIS/RIM a nivel institucional y nacional en países de América Latina y el Caribe, ha tenido un crecimiento importante en los últimos años. En mayo de 2022, había un total de 26 sistemas CRIS registrados en DRIS mientras que, hasta abril de 2024, se encuentran registrados un total de 39 sistemas, lo que representa un incremento del 50% en dos años. El número de sistemas de alcance nacional se duplico y en el caso de los institucionales hubo un incremento del 30.64%.

Perú encabeza la lista de los países de América Latina con mayor crecimiento y número de sistemas CRIS institucionales implementados, de 10 que se tenían en 2022 pasó a 23, lo cual representa un incremento del 56.5%. Cabe destacar que los 13 sistemas adicionales están implementados con la plataforma PURE.

Los tres países con un mayor número de sistemas CRIS son: Perú con un sistema nacional y 24 institucionales; Chile con 12 institucionales y una plataforma nacional; y, México con 11 institucionales.

En cuanto a las instalaciones de software también hubo incrementos, el uso de DSpace-CRIS aumentó un 20%, VIVO un 13.3% y PURE un 56.52%. En este aspecto, la solución de software más utilizada por los países de la región es PURE, con un total de 36 instalaciones, seguido de VIVO con 17, otras soluciones con 14, DSpace-CRIS con 12 y Sympletic con 2.

Finalmente, como trabajo futuro, para concretar el estado de la cuestión sobre el desarrollo de Sistemas de Gestión e Investigación (CRIS) a nivel institucional y nacional, en países de América Latina y el Caribe, será necesario hacer un seguimiento por institución/país con base en los resultados de la encuesta diagnóstica (primera fase del estudio) y del estudio de casos (segunda fase), para confirmar los sistemas que se encuentren operando, los nuevos registrados en directorios o instalaciones de software, incluyendo otras soluciones comerciales u open source, y validar y complementar la información documentada en cada estudio de caso, a través de entrevistas con el personal responsable y staff técnico de cada uno de los siete sistemas/plataformas nacionales.

Referencias bibliográficas

Bryant, R., Clements, A., Feltes, C., Groenewegen, D., Huggard, S., Mercer, H., Missingham, R., Oxnam, M., Rauh, A., y Wright, J. (2017). *Research Information Management: Defining RIM and the Library's Role*. Dublin, OH: OCLC Research. https://doi.org/10.25333/C3NK88

Dempsey, L (26 de octubre de 2014). *Research information management systems: a new service category?*. LorcanDempsey.net. https://www.lorcandempsey.net/research-information-management-systems-a-new-service-category/

Martínez Carazo, P. C., (2006). El método de estudio de caso: estrategia metodológica de la

investigación científica. Pensamiento & Gestión, (20), 165-193. https://www.redalyc.org/articulo.oa?id=64602005

The International Organization for Research Information (s.f.-a). *Common European Research Information Format (CERIF)*. Recuperado el 25 de abril de 2024 de https://eurocris.org/services/main-features-cerif

The International Organization for Research Information (s.f.-b). *Directory of Research Information Systems (DRIS)*. Recuperado el 28 de abril de 2024 de https://eurocris.org/services/dris

The International Organization for Research Information (s.f.-c). *EuroCRIS DSpace CRIS digital repository.* Recuperado el 26 de abril de 2024 de https://dspacecris.eurocris.org/

Vázquez, T.R. (12-14 de mayo de 2022-a). *Development and characterisation of CRIS systems in Latin America: Preliminary results of diagnostic survey.* 15th International Conference on Current Research Information Systems (CRIS2022), Dubrovnik, Croatia. http://hdl.handle.net/11366/2010

Vázquez, T.R. (2022-b). Development and characterisation of CRIS systems in Latin America: Preliminary results of diagnostic survey. *Procedia Computer Science, 211*(2022), 267-276.

https://www.sciencedirect.com/science/article/pii/S1877050922016660?via%3Dihub

https://doi.org/10.1016/j.procs.2022.10.201

Vázquez, T.R. (15-17 de mayo de 2024). Development of National CRIS Systems in Latin America and the Caribbean: Case Studies 2023. 16th International Conference on Current Research Information Systems (CRIS2024), Vienna, Austria. http://hdl.handle.net/11366/2549

MTE ROSALINA VÁZQUEZ TAPIA

Fundadora y Coordinadora General de la Red Mexicana de Repositorios Institucionales – REMERI
Funcionaria e Investigadora de la Universidad Autónoma de San Luis Potosí – UASLP, México.
alinavn@uaslp.mx, alinavt@cudi.edu.mx

Es Maestra en Tecnología Educativa por el Instituto Tecnológico de Estudios Superiores de Monterrey, Ingeniero en Sistemas Computacionales por el Instituto Tecnológico de San Luis Potosí, y actualmente candidata a Doctora en el Programa de Doctorado en Formación de la Sociedad del Conocimiento de la Universidad de Salamanca, España. Desde 1992 es Funcionaria y Jefa de Departamento de la Administración Central de la Universidad Autónoma de San Luis Potosí, ocupando diferentes puestos directivos.

Ha tenido a su cargo el desarrollo de 26 proyectos institucionales e interinstitucionales y la responsabilidad técnica de 7 proyectos de investigación con financiamiento de la Secretaría de Educación Pública (SEP) y del Consejo Nacional de Ciencia y Tecnología (CONACYT) de México.

Ha colaborado como evaluadora experta de proyectos de investigación en materia de Repositorios de Ciencia Abierta y de Sistemas de Gestión de la Investigación (CRIS), en convocatorias públicas de financiamiento de CONACYT y de la Fundación Española para la Ciencia y la Tecnología (FECYT) de España.

Es investigadora y experta en Repositorios Institucionales, Acceso Abierto, Ciencia Abierta, Sistemas de Gestión de la Investigación (CRIS), Bibliotecas Digitales y Tecnologías Educativas. De igual forma, ha publicado artículos arbitrados en congresos y revistas científicas e impartido múltiples conferencias magistrales, ponencias y talleres en congresos y eventos académicos nacionales e internacionales.

Es Fundadora y Coordinadora General de la Red Mexicana de Repositorios Institucionales- REMERI y Responsable Técnica del nodo técnico de México en la Red Federada de Repositorios Institucionales de Producción Científica de América Latina- LA Referencia desde el año 2012. Además, colabora con las siguientes organizaciones nacionales e internacionales: Miembro del Comité permanente de la Conferencia Internacional BIREDIAL-ISTEC; Socio Técnico del Proyecto Los Primeros Libros de las Américas - PLA; Miembro de la Confederación de Repositorios - COAR y Organización Internacional EuroCRIS; Coordinadora de la Comisión de Repositorios y Recursos Educativos Digitales Idel Grupo de Tecnología Educativa de la Asociación Nacional de Universidades e Instituciones de Educación Superior – ANUIES-TIC; y, Coordinadora del Eje de Tecnología Educativa y Servicios digitales del Equipo Colaborativo para la Transformación Digital de la Educación- EduTraDi.

Póster

Lista de póster presentados en el evento:

1. Guía de regulación de uso y reporte de Inteligencia Artificial en publicaciones científico-académicas en los roles de autoría, edición y revisión por pares. Una perspectiva desde la Ciencia Abierta. *Liana Penabad-Camacho, María Morera-Castro, María Amalia Penabad-Camacho*
2. Política de cambio de nombre de autoría para identidad de género. *Enrique Muriel-Torrado, Lúcia da Silveira, Juliana Aparecida Gulka*
3. Apoio técnico editorial a periódicos científicos: a atuação do Laboratório de Periódicos Científicos da UFSC. *Enrique Muriel-Torrado, Patricia da Silva Neubert, Rosângela Schwarz Rodrigues, Edgar Edgar Bisset-Alvarez, Luiz Roberto Curtinaz Schifini*
4. Situación actual de las revistas científicas nacionales en el proyecto SciELO Uruguay. *Laura Machado*
5. Relevamiento de publicaciones digitales y acervo documental de los centros de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires. *Dolores García, Lorenzo Calamante, Gonzalo L. Villarreal, Lucas Eduardo Folegotto*
6. Construcción de sitios web institucionales integrados con sistemas externos. *Gonzalo L. Villarreal, Pablo G. Terrone, Lautaro Josin Saller*
7. Ecossistema da educação aberta brasileira: mapeamento das tendências atuais e de seus elementos constituintes. *Eva Priscila Vieira Dann, Caterina Groposo Pavão*
8. Ciência aberta e o papel do Repositório Institucional Ninho. *Kátia Simões, Robson Martins, Camila Belo, Mariana Teles*
9. Acervo digital da Biblioteca de Obras Raras Fausto Castilho da Unicamp: estudo preliminar de conservação de livros raros e especiais. *Danielle Thiago Ferreira, Isabella Nascimento Pereira*
10. Impacto del uso de redes sociales para comunicar desde el Repositorio de Datos Académicos RDA-UNR. *Paola Bongiovani*
11. Ciência à vista no Repositório Institucional da UFSC. *Sandra Sobrera Abella, Denise Machado, Marli Dias de Souza Pinto*
12. Gestão de conteúdo em repositórios institucionais de universidades estrangeiras: análise de diretrizes a partir de boas práticas internacionais. *Denise Machado, Marli Dias de Souza Pinto*
13. Avaliação dos repositórios de dados em biodiversidade: uma análise com base nos princípios FAIR. *Carla Marques Felipe*
14. Modelo de depósito de dados assistido realizado por equipe multidisciplinar da área da Saúde: a experiência do Arca Dados (Fiocruz). *Vanessa de Arruda*
15. ¿En quién pienso cuando comparto mis datos de investigación? *María Hidalgo, Meilyn Garro*
16. Rede Moara para compartilhamento de códigos fonte no âmbito da Ciência Aberta. *Diego José Macêdo, Bernardo Dionízio Vechi, Rebeca dos Santos de Moura, Lucas Rodrigues Costa, Ingrid Torres Schiessl, Milton Shintaku*

Conferencias magistrales y mesas de discusión

Conferencia: Inteligencia Artificial, una revolución a plena marcha

Álvaro Soto

Esta conferencia estuvo a cargo del Dr. Álvaro Soto, Director del Centro de Nacional de Inteligencia Artificial (CENIA).

En su presentación, el profesor Soto expuso sobre el auge explosivo que ha tenido la Inteligencia Artificial (IA) y sus aplicaciones prácticas, y complementó su disertación con un detalle minucioso de las bases tecnológicas de la IA.

En primer lugar, el disertante compartió su visión acerca de dónde están las claves de esta importante revolución tecnológica que está basada en dos tecnologías que hicieron su aparición en la última década. También dio su punto de vista con relación a las implicancias del uso de la IA. Realizó un decurso desde las épocas de la revolución industrial para deslindar las capacidades, la inteligencia especial de los seres humanos y lo esperable de la IA. El disertante describió el inicio de la IA a partir de la curiosidad del ser humano por conocer los mecanismos de su propia inteligencia y transferirlos a una máquina para dotarla de razonamientos y respuestas similares.

El Prof. Soto mostró ejemplos diversos de la evolución de la IA desde sus comienzos hasta la actualidad, se valió de distintos videos para ejemplificar las primitivas formas de dominio de la IA hasta el actual *deep learning*, mostrando los avances desde la primera tecnología que comienza a masificarse en el año 2012 hasta la actualidad.

All final de su charla, el disertante presentó el Centro Nacional de Chile de Inteligencia Artificial, cuyo sitio web se encuentra en: https://cenia.cl/

Conferencia: Ciencias Sociales para Chile, una red de colaboración en Ciencia Abierta

Antonieta Urquieta

Esta Conferencia estuvo a cargo de la Dra. Antonieta Urquieta, Directora Académica de la Facultad de Ciencias Sociales de la Universidad de Chile.

La Profesora Urquieta enfatizó la relevancia de las ciencias sociales para el desarrollo científico del país y la necesidad de promover una red de colaboración que fortalezca la ciencia abierta. También destacó la preocupación por la reducción sistemática del financiamiento para la investigación en este campo, lo que representa un desafío importante para visibilizar y apoyar el quehacer de las ciencias sociales a nivel nacional.

A continuación, presentó al Co-Laboratorio como una respuesta concreta a la problemática antes mencionada en las ciencias sociales dado que abre la posibilidad de crear nuevos espacios para la investigación colaborativa y la transferencia de conocimiento, un paso crucial para que las ciencias sociales ocupen un lugar prominente en la agenda científica de Chile. El Co-Laboratorio tiene como objetivo divulgar la producción científica de investigadores en el campo de las ciencias sociales, destacando sus trayectorias académicas y promoviendo una gestión del conocimiento en lógica colaborativa, abierta y transdisciplinar.

En la exposición también se mostró el sitio web y las facilidades de la plataforma de acceso abierto *Co-Laboratorio de Investigación en Ciencias Sociales*, desarrollada por la Facultad de Ciencias Sociales de la Universidad de Chile. Para ver en detalle las facilidades que se ofrecen, se sugiere acceder al sitio web https://cits.uchile.cl/co-laboratorio

La iniciativa disputa con la lógica de productivismo académico individual, pues apuesta por la conformación de una red colaborativa de producción de conocimiento científico que pone a las ciencias sociales al servicio de los desafíos sociales que las interpelan.

El Co-Laboratorio fue presentado oficialmente en la Universidad de Chile el jueves 11 de julio de 2024, se ha extendido fuera del ámbito universitario buscando nuevos actores y ha logrado la incipiente participación de otras universidades chilenas.

Conferencia: Open Alex: Abordando las desigualdades en las fuentes bibliográficas

Juan Pablo Alperin

Esta conferencia estuvo a cargo de Juan Pablo Alperin, profesor de la Universidad Simon Fraser en Canadá y Director Científico del Public Knowledge Project.

El planteamiento de esta conferencia se basa en que la evaluación de la investigación en el mundo académico actual se basa en gran medida en las citas y el impacto de las publicaciones indexadas en unos pocos espacios editoriales comerciales; sin embargo, estas fuentes bibliográficas tradicionales suelen tener un sesgo hacia las revistas de alto impacto y las instituciones de renombre, lo que fomenta la desigualdad. En este contexto, Open Alex, el eje de esta conferencia, surge como una alternativa inclusiva que contribuye a democratizar el acceso a la información.

El Prof. Alperín se enfocó en la idea de que usar bases de datos selectivas que solo tienen una cantidad limitada de revistas simplemente no es congruente con el tiempo actual; hay un cambio de paradigma que ya está sucediendo y se puede visualizar en una tendencia que empuja al uso de infraestructuras abiertas. Esto en consonancia con la declaración de la UNESCO que señala a las infraestructuras abiertas como uno de los cuatro pilares para el avance de la ciencia abierta; es decir, hay un movimiento relevante hacia las infraestructuras abiertas que involucra a instituciones e incluso países en el uso de fuentes abiertas para hacer sus evaluaciones. El disertante puso varios ejemplos; entre otros, el Ministerio de Ciencias de Francia ha hecho realidad este paradigma al adoptar solo fuentes abiertas para hacer su evaluación.

A continuación, Alperín planteó la pregunta de por qué se sigue procediendo según la manera tradicional, considerando que tanto el contexto como las tecnologías de base han cambiado. Así, pasó a presentar las alternativas que se han tratado de aprovechar: Google Scholar y Microsoft Academic. Estas dos iniciativas aprovechan el contexto de la web para cambiar el paradigma hacia uno no basado en papel, sino en todo lo que ya está en digital. No obstante, solo una de ellas es realmente una fuente abierta.

Luego, el disertante expuso que dos empresas están utilizando infraestructuras abiertas y empezando a crear conexiones; a su vez, existen dos grupos que están utilizando esas infraestructuras abiertas como base: Dimensions de Digital Science y OpenAlex. Expuso que la primera es una empresa comercial que opera con fines de lucro y vende los datos, mientras que la segunda es una ONG que dispone datos y procesos en abierto, por lo que su exposición continuó con OpenAlex.

El disertante describió a OpenAlex como una fuente bibliográfica más inclusiva que cualquiera de las alternativas cerradas y habló de la importancia de su política de indexación, ya que abre la posibilidad de una evaluación de la investigación que considera toda la investigación existente, independientemente de dónde se publique.

El disertante incluyó en su conferencia una demostración de la hipótesis de mayor inclusividad de OpenAlex, para la cual se basó en una investigación que analizó más de 45000 revistas publicadas utilizando OJS, y puso de manifiesto que aún existen desigualdades estructurales en el acceso a las infraestructuras académicas que mantienen a miles de revistas fuera de estas fuentes.

La presentación mostró que OpenAlex ya puede ser utilizado para diversos tipos de análisis, pero que aún es necesario hacer un esfuerzo para mejorar la disponibilidad de información de investigación de comunidades que han sido históricamente excluidas.

Conferencia: Peace Engineering – Ingeniería para la Paz

Ramiro Jordan

Durante su conferencia el Dr. Jordán definió Ingeniería para la Paz (Peace Engineering) como la aplicación intencional del pensamiento sistémico de los principios de la ciencia, tecnología y la ingeniería para promover y apoyar directamente las condiciones para la paz.

Especificando las bases con las que trabaja "Peace Engineering" y sus destinatarios: un mundo donde la prosperidad, la sostenibilidad, la equidad social, el espíritu empresarial, la transparencia, la participación de la comunidad, la ética y la cultura de calidad prosperen.

En el núcleo de Peace Engineering, sostuvo el Dr. Jordán se encuentra el futuro sostenible del planeta, que está llamando a los líderes a actuar en concierto con una mentalidad sistémica. El Prof. Jordán sostuvo la necesidad de cultivar juntos el desarrollo de líderes de la próxima generación para continuar impulsando el cambio, sostuvo asimismo que la ingeniería para la paz es una nueva mentalidad disruptiva en todas las disciplinas existentes y nuevas a crearse para enfrentar los desafíos globales.

La conferencia hizo hincapié en ejemplos de desarrollo dentro de los Estados Unidos, en particular, sobre los llamados urgentes a la acción por parte de la NASEM (Academia Nacional de Ciencias de USA) , la Cumbre de Premios Nobel, la ONU y los científicos de todo el mundo para abordar y resolver desafíos globales cruciales y ampliamente reconocidos para la paz y la seguridad antes de que se vuelvan más complejos y más ambientalmente, financieramente y socialmente costosos; antes que lleguemos al punto sin retorno (2030).

Conferencia: ¿La inteligencia artificial, es realmente inteligencia?

Jorge Solís Tovar

En esta ponencia se sostiene que los términos de Inteligencia Artificial o Inteligencia Artificial Generativa, no son mas que nombres comerciales para vender servicios que ahora son posibles de brindar gracias al avance en el desarrollo de la tecnología informática digital, pues la verdadera inteligencia corresponde a los profesionales que desarrollan los algoritmos que permiten extraer de la inmensa cantidad de información en la nube la respuesta al pedido de los usuarios.

Sin negar la gran utilidad que tiene esta herramienta en los distintos campos de la educación, el trabajo y las comunicaciones, cuyo logro ha sido posible gracias al abaratamiento de los dispositivos de almacenamiento de información, al desarrollo de las comunicaciones inalámbricas y al incremento en la velocidad de proceso de los computadores, se advierte sobre los peligros que representa esta herramienta cuando se confía plenamente en los resultados que ofrece, sobre todo si no tenemos forma de verificar o por lo menos intuír su pertinencia.

Mg. Jorge Solís Tovar es Ingeniero Civil por la Pontificia Universidad Católica de Perú (PUCP). Magister en Políticas y Gestión Universitaria por la Universidad de Barcelona., Con más de 60 años de experiencia profesional. Asesor Técnico del Rectorado de la PUCP y responsable de su Repositorio Institucional.

Mesa de discusión: IA y sistemas de descubrimientos e interfaces

Esta mesa estuvi a cargo de Claudio Escobar, Universidad Alberto Hurtado, Chile; Gabriela Arriagada Bruneau, Instituto de Éticas Aplicadas y el Instituto de Ingeniería Matemática Computacional, Chile; y Pamela Cariceo Rivera, Banco de Crédito e inversiones, Chile. Moderador: Rafael Castillo, Universidad de Chile

La Mesa tuvo como modalidad preguntas disparadoras para la intervención de los integrantes de la Mesa. Inició con una explicación de qué son los algoritmos entendidos como listas de instrucciones que permiten entrenar la capacidad de un sistema en diferentes actividades tales como las de identificación o la predicción, separando aquellos algoritmos supervisados de los que no lo son y trabajan a partir de estímulos de percepción de patrones. Se usó como ejemplo el chat GPT que se basa en una estructura de transformers dado que es un algoritmo de aprendizaje profundo y utiliza, además de una ingente cantidad de datos, capas de redes neuronales. Se habló del entrenamiento de grandes modelos de lenguaje para realizar tareas muy específicas en diferentes disciplinas.

A continuación, se realizó un deslinde sobre Inteligencia Artificial Generativa y la importancia de la selección de los conjuntos de datos de manera que los mismos resulten insesgados. La próxima pregunta buscó dar respuesta al tema de las interfaces y la necesidad de contar con un universo de usuarios representativo para lograr el diseño de un producto o servicio que cumpla con las necesidades y expectativas, para lo cual, se resaltó la necesidad de conversar con personas para conocer en detalle las necesidades y mejorar el diseño.

Los disertantes abordaron luego el tema de la antropomorfización de la IA por ejemplo el uso de intervenciones de IA generativa en educación, particularmente se habló, desde una experiencia en el área de psicología.

La Mesa retornó al tema de los datos que alimentan a las inteligencias artificiales y la necesidad de ofrecer un conocimiento situado ya que muchas veces acontece que las decisiones deben programarse en tanto legislaciones que son privativas de los países.

Finalmente se habló de la revalorización de las funciones que han tenido históricamente los sistemas de información y las bibliotecas y que es en esos ámbitos entre otros donde debe pensarse el uso de inteligencia artificial para agilizar la gestión y el tratamiento de los datos.

Los integrantes de la Mesa resaltaron sobre la necesidad de un trabajo colaborativo para pensar en conjunto las aplicaciones de IA por ejemplo en las gestiones de material bibliográfico y en la identificación de usuarios.

Los investigadores culminaron con el relato del estado de situación actual y las miradas hacia el futuro.

Mesa de discusión: Nuevas propuestas de evaluación de la actividad científica

Esta mesa estuvo a cargo de Marisa De Giusti, Universidad Nacional de la Plata, Argentina; René Faustino Gabriel Junior, Universidade Federal do Rio Grande do Sul, Brasil; Caterina Groposso, Universidade Federal do Rio Grande do Sul, Brasil; y Soledad Bravo-Marchant, Unidad de Acceso a la Información Científica de la Agencia Nacional de Investigación y Desarrollo, Chile. Moderador de la mesa: Juan Pablo Alperin

La mesa se planteó en torno a una dinámica, previamente acordada, para el tratamiento de la discusión actual sobre la forma en que se evalúa la actividad científica en toda su dimensión así como las carreras profesionales. Tal cuestionamiento parece apropiado a la luz del descrédito de los indicadores tradicionales tales como el factor de impacto y, fundamentalmente ante un avance de la ciencia abierta que propone cambios profundos en el proceso de investigación y en la inclusión de productos que se consideran importantes más allá del artículo científico, así como del proceso de comunicación de la ciencia que requiere profundos cambios y nuevas habilidades o el proceso en particular de intervención -y apropiación- en el caso de la ciencia ciudadana que exigen un trabajo colaborativo que requiere otros modos de actuación.

La discusión precedente se pensó en relación a la **gobernanza** de la ciencia el cual impone además de la evaluación políticas, incentivos y financiamiento y tanto desde Europa cuanto desde América Latina se han planteado numerosas propuestas alternativas aunque sin demasiado éxito a la hora de plasmarlas en la evaluación.

La Mesa trató aspectos concretos para cambiar la evaluación a través de:

1) Un resumen de nuevas propuestas e indicadores que devienen de distintos espacios geográficos, una problematización sobre las prácticas de evaluación en particular de AL con atención a las prácticas de los distintos campos de conocimiento.
2) Una problematización sobre la coherencia entre las distintas políticas de AA/CA en AL y las formas de evaluación en los países que estuvieron representados en la mesa. Un relato particularizado de las nuevas políticas de evaluación de CAPES en Brasil así como la situación en Chile.
3) Una puesta a la luz de las incoherencias que surgen también de los espacios editoriales consagrados y una forma de propuesta de indicadores y herramientas alternativas.
4) Una actualización sobre cómo la IA está disparando algunas prácticas de superproducción de artículos y cómo va a tratarse este asunto.
5) Un cierre propositivo.