

4. Desafíos de los repositorios institucionales como fuente de indicadores para monitoreo de las políticas de la Ciencia Abierta y evaluación de la investigación

Lautaro Matas

Secretario ejecutivo de la Red Latinoamericana de Repositorios de Ciencia Abierta - LA Referencia

Andrea Mora-Campos

Presidenta de LA Referencia

Rodolfo Barrere

Coordinador del Observatorio Iberoamericano de la Ciencia, la Tecnología y la Sociedad de la Organización de Estados Iberoamericanos (OEI)

Federico Cetrangolo

Gerente de Administración de LA Referencia

1. Introducción

En 2021, como resultado de la Conferencia General de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) se aprobaron las *Recomendaciones sobre Ciencia Abierta de la UNESCO*, en las que se insta a agencias de financiación e instituciones académicas a revisar y mejorar la forma en que evalúan la investigación científica (UNESCO, 2021).

La agenda propuesta por el Foro Latinoamericano de Evaluación sobre Ciencia (CLACSO, 2023) en Latinoamérica y por la Coalición Internacional para el Avance de la Evaluación de la Investigación (Coalition for Advancing Research Assessment [COARA], 2023), desde Europa, ha marcado una ruta de cambios en los procesos de evaluación, en la cual los repositorios tienen el rol de aportar insumos como componentes del ecosistema de Ciencia Abierta. En este contexto, es necesario analizar sus fortalezas y debilidades como fuente para el cálculo de indicadores, de tal manera que se considera tanto la cobertura y validez estadística de las colecciones como la calidad de la información bibliográfica recogida. Para llevar a cabo este diagnóstico fue preciso contar con una colección sistematizada y representativa de los repositorios latinoamericanos que ofrezca protocolos homogéneos para el procesamiento de los datos.

La Red Latinoamericana de Repositorio de Acceso Abierto (LA Referencia) es una iniciativa de cooperación regional iniciada en 2012, conformada por las autoridades de Ciencia y Tecnología de 11 países latinoamericanos más España. Su principal objetivo es apoyar las estrategias nacionales de acceso abierto a partir de estándares de interoperabilidad. Comparte y visibiliza la producción científica generada en las instituciones de educación superior y de investigación científica. La información recogida por LA Referencia ofrece una muestra representativa y validada de la realidad de repositorios latinoamericanos para los efectos del mencionado diagnóstico (LA Referencia, 2012 y 2023).

Este texto tiene como objetivo discutir el potencial de los repositorios como fuente complementaria de indicadores, comparándola con otras actualmente utilizadas, en términos de cobertura y validez estadística. Asimismo, presenta un diagnóstico sobre la calidad de la información que es posible recolectar en forma sistematizada mediante agregadores como LA Referencia. Finalmente, concluye con una serie de recomendaciones que establecen las posibles acciones tendientes para resolver algunas de las falencias descubiertas en la etapa de diagnóstico.

2. Los repositorios institucionales en comparación con otras fuentes de datos para monitoreo y evaluación

La utilización de documentos científicos como objeto de medición para el monitoreo de la actividad científica se basa en que constituyen productos representativos del trabajo de investigación. Al mismo tiempo, el sistema de evaluación de pares funciona como un filtro que busca garantizar un mínimo de calidad y pertinencia en cada resultado de investigación. Este sistema de doble garantía hace que contar resultados científicos constituya una medida robusta de una de las tantas dimensiones evaluables de la actividad académica.

Por consiguiente, es esperable que los indicadores bibliométricos sean consistentes con otras variables del sistema de ciencia y tecnología; dentro de un determinado rango sería lógico encontrar cierta relación entre indicadores como la inversión en I+D, la cantidad de personas investigadoras y el resultado de su trabajo: la producción científica.

Sin embargo, para que esto ocurra, las fuentes de producción científica utilizadas deben tener una buena cobertura en términos estadísticos. Eso no solo implica un volumen dado de documentos, sino una adecuada cobertura en términos geográficos, disciplinares y temporales, entre otros.

Como parte de este trabajo, y para conocer las posibilidades de cobertura geográfica de los repositorios más allá de la red LA Referencia, se realizó una consulta a un conjunto extendido de países de la región. El objetivo era conocer el estado de avance en la implementación de infraestructuras de repositorios institucionales de literatura científica, de datos de investigación, de *preprints* y mixtos (literatura y datos). La información fue consultada a las personas representantes de los nodos de los países que integran LA Referencia y en países no miembros, personas líderes de Ciencia Abierta en los Organismos Internacionales de Ciencia y Tecnología (ONCYT) u otros organismos gubernamentales.

TABLA 1. Cantidad de repositorios por país y tipo en América Latina

País	Repositorios de literatura científica (N)	Cantidad de repositorios de datos	Cantidad de repositorios mixtos (literatura científica y datos de investigación)
Argentina*	77	1	12
Brasil*	119	14	3
Colombia*	130	4	#
Costa Rica*	8	0	1
Cuba	16	0	0
Ecuador*	56	1	0
El Salvador*	11	0	0
España*	113	2	37
Guatemala	1	0	1
Honduras	1	0	0
México*	135	5	4
Nicaragua	15	0	0
Panamá*	8	0	0
Perú*	192	0	0
Puerto Rico	1	0	0
Rep. Dominicana	0	0	12
Uruguay*	9	0	0

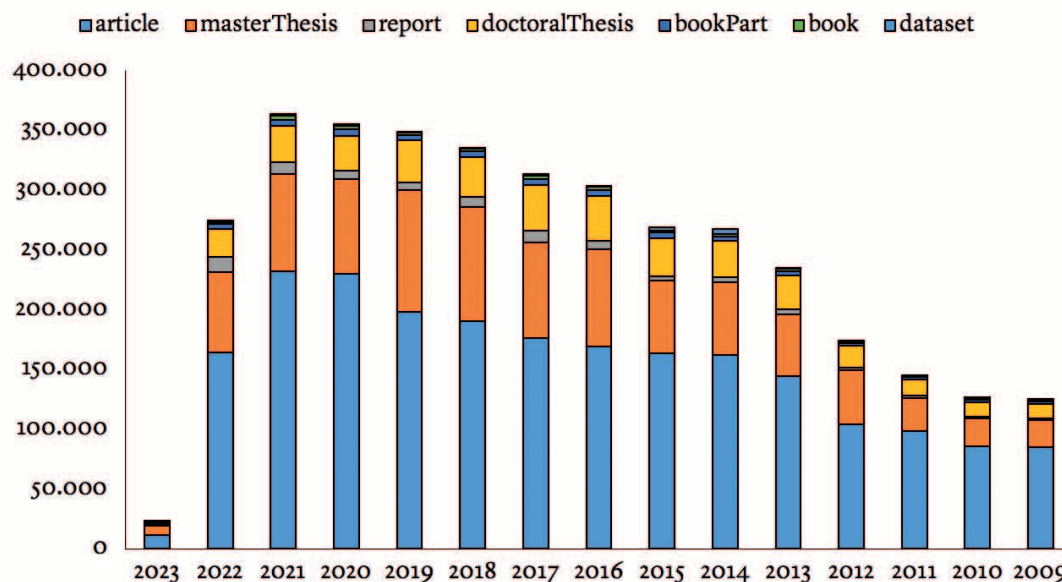
Fuente: Elaboración propia con base en la información brindada por las personas representantes de los 12 nodos de los países que integran LA Referencia y en países no miembros, personas líderes de Ciencia Abierta en los Organismos Internacionales de Ciencia y Tecnología (ONCYT) u otros organismos gubernamentales. Información consultada en agosto de 2023 por LA Referencia.

Según los resultados de la Tabla 1, no se reportan grandes avances en el desarrollo de repositorios de *preprint*; solo Brasil reportó dos casos. Otro hallazgo es la cantidad de repositorios mixtos en algunos países, esto se debe a limitaciones de recursos, puesto que están depositando conjuntos de datos en plataformas originalmente pensadas para literatura científica. Finalmente, se evidencia que la colección de LA Referencia —países marcados con asterisco (*)— puede ser utilizada como una muestra representativa de la región, al menos para este primer diagnóstico.

Al analizar los repositorios a través de la colección de LA Referencia (datos de julio de 2023) se encuentran más de 4,5 millones de metadatos de producción científica (artículos, reportes, tesis de doctorado y maestría, libros, capítulos de libro y conjuntos de datos) cosechados por el agregador regional, pro-

venientes de sus 12 nodos nacionales (Argentina, Brasil, Chile, Colombia, Costa Rica, Ecuador, España, El Salvador, México, Panamá, Perú, Uruguay). En la Figura 1, se observa la evolución temporal de los metadatos colectados segmentados por tipo de documento.

FIGURA 1. Distribución de los documentos científicos agregados en LA Referencia, según tipología por año



Fuente: Elaboración propia. LA Referencia agrega siete tipologías documentales: artículo científico, tesis doctoral, tesis de maestría, libro, capítulo de libro, conjunto de datos. Existe un rezago en el depósito de documentos en repositorios que se refleja en los últimos años (2022 y 2023).

Si se compara la colección de LA Referencia con otras fuentes bibliográficas, actualmente utilizadas, en 2020 la base de datos Scopus registró un total de 187.209 documentos con participación de población autora latinoamericana. En paralelo, los repositorios de la región acumularon en ese año un total de 354.038 registros, de los cuales alrededor de 230.000 fueron artículos. Esto muestra una cantidad mayor de artículos depositados en repositorios respecto a los de Scopus. En ese contexto vale analizar si poseen una representatividad estadística similar.

Respecto al problema de la representatividad, es relevante mencionar que ninguna de las bases de datos utilizadas para estudios bibliométricos captura el total de la producción científica. Se trata de colecciones construidas en función de los niveles de citación en cada una de las disciplinas que se busca cubrir. La idea de que un conjunto relativamente pequeño de revistas concentra los trabajos más citados de cada campo de estudio, basada en los estudios de Samuel Bradford (1934), guía la construcción de colecciones de revistas que resultan representativas de un colectivo mucho mayor. Las bases de datos bibliométricas son diseñadas con este criterio. En resumen, en términos estadísticos, tener más elementos no garantiza la representatividad. El análisis de la producción científica no escapa de este principio.

En ese sentido, es esperable que un conjunto de documentos con adecuada cobertura y representatividad indique resultados estadísticos consistentes con la evidencia obtenida por otros caminos. Por ejemplo, para ser una herramienta adecuada para el monitoreo y la evaluación, es esperable que la

producción científica siga patrones y tendencias similares a los de la inversión en I+D o a la cantidad de población investigadora. Se trata de un sistema en el cual ciertos recursos permiten producir conocimiento que deja huella en la producción de documentos. Si la medida de recursos y resultados no es mínimamente consistente, el sistema de monitoreo puede estar midiendo fenómenos diferentes: es decir, no son representativos del mismo universo o, en otras palabras, no tienen adecuada cobertura.

De acuerdo con la muestra analizada, los repositorios latinoamericanos tienen, aproximadamente, la misma cantidad de artículos que Scopus con la participación de autoría de la región, pero no está claro si cubren el mismo universo. En ese sentido, existe una primera distinción sobre el universo que buscan cubrir: por construcción, los repositorios recolectados por LA Referencia solo cuentan con artículos en acceso abierto, mientras que las bases de datos como Scopus cubren revistas con distintos tipos de acceso.

En 2020, según datos de SJR World Report, el 57% de los artículos con participación de población autora latinoamericana en Scopus fueron publicados en acceso abierto. Esto equivale a 106.709 registros. Visto de esa forma, se trata de un volumen significativamente menor al de los más de 230.000 artículos recogidos en repositorios. Esto puede representar una fortaleza, dado que un mayor volumen presenta una plataforma privilegiada desde la cual analizar la representatividad estadística de los repositorios como una herramienta de monitoreo y evaluación. Sin embargo, en ciertas disciplinas, principalmente las experimentales, las revistas más valoradas en muchos sistemas actuales de evaluación no son de acceso abierto y por tanto quedan fuera del ámbito de cobertura de los repositorios.

Un ejemplo de este fenómeno puede verse dentro de Scopus. Siguiendo en el análisis de la producción de población autora latinoamericana en esa fuente, en 2020 el 70% de los documentos de ciencias sociales fueron publicados en acceso abierto, mientras que en química, por citar un ejemplo, fue solo el 45%.

No obstante, existe en toda la región un fuerte movimiento hacia el acceso abierto, con el respaldo de organismos de política, gestión y promoción de la ciencia y la tecnología. Es necesario que ese impulso permee los mecanismos de evaluación para impulsar la publicación en acceso abierto en todas las disciplinas. Dado que el universo que cubren los repositorios es el de la publicación en acceso abierto, es relevante comparar el volumen de documentos que recopilan en comparación con bases bibliográficas de acceso abierto como SciELO y Redalyc.

Un trabajo reciente del equipo del CECIC-Universidad Nacional de Cuyo, CONICET, Argentina muestra resultados estadísticos de la base de datos OLIVA, que integra registros de SciELO y Redalyc. En el último año disponible (2017), esta base de datos integrada reunió alrededor de 50.000 documentos de población autora latinoamericana. En ese mismo año, los repositorios de acceso abierto de la región acumularon casi 130.000 documentos. Se trata de una nueva señal alentadora del potencial de los repositorios de la región (Gallardo, 2022).

Con base en estos datos positivos sería notable analizar si, más allá de las cantidades, los repositorios latinoamericanos cuentan con una adecuada representatividad estadística a nivel de país. Como se ha mencionado, los indicadores deben responder a una idea sistémica y de medición coherente de insumos, procesos y resultados. Como verificación de coherencia se pueden realizar comparaciones con indicadores de otras variables del sistema. A modo de ejemplo, puede estudiarse la inversión en I+D de los países de la región en relación con los documentos obtenidos de repositorios de acceso abierto.

Según datos de la Red Iberoamericana de Indicadores de Ciencia y Tecnología (RICYT), en 2020 Brasil era responsable del 63% de la inversión regional en I+D, seguido por México con el 12,5%, Argentina con el 8,4%, y Chile y Colombia con alrededor del 2,5% cada uno. Entre los cinco países acumulaban

casi el 90% de la inversión regional. Los repositorios, en cambio, muestran un panorama diferente. El primer punto a destacar es la muy baja presencia de México, uno de los países más importantes de la región, con menos del 3% de los documentos. Brasil aporta el 59% de los documentos disponibles, un valor consistente con los datos de inversión en I+D, pero si se contara con información más completa de México ese porcentaje de participación, necesariamente, debería descender. El segundo lugar lo ocupa Colombia con el 14% de los registros, muy por encima de su participación en la I+D latinoamericana, y un fenómeno parecido ocurre con Perú.

Las diferencias en la representación de los países en los documentos disponibles son el reflejo de una desigual cobertura de la producción de las instituciones de cada país. Por ejemplo, en el caso de Brasil, la Universidade Federal do Rio de Janeiro aparece en 2020 en el lugar 18 en cantidad de documentos, cuando es una de las principales instituciones del país y la cuarta en producción científica registrada en Scopus.

Existen en todo caso datos alentadores en torno a la cobertura y representatividad de los repositorios como fuente para el monitoreo y la evaluación, principalmente relacionados con el volumen de datos existentes. LA Referencia, como muestra de la red de repositorios, es la fuente con mayor cantidad de registros en acceso abierto disponible en la región. Sin embargo, es necesario trabajar para mejorar su cobertura disciplinaria, regional e institucional; sin eso, las posibilidades de que se convierta en una herramienta clave para el monitoreo y la evaluación se ven muy disminuidas.

3. Diagnóstico de la calidad de metadatos

Además de considerar la cobertura y representatividad de los repositorios como fuente, es imprescindible analizar la calidad de la información que los registros aportan a las distintas dimensiones y categorías relevantes para el cálculo de indicadores. Estas informaciones, en el caso de los repositorios, se denominan metadatos y son recolectados mediante protocolos de interoperabilidad.

Los metadatos facilitan la recuperación, acceso, visibilidad y preservación de los objetos digitales depositados en los repositorios, por lo que están diseñados como estándares de descripción precisa que proporcionan un contexto a los objetos digitales preservados en los repositorios. Para que los repositorios de acceso abierto cumplan con su función es necesario que sean interoperables entre sí, con otros sistemas y con servicios de valor agregado relacionados con la investigación. Esto significa que deben seguir una serie de criterios y normas que aseguren que los recursos digitales sean descritos, identificados, localizados y transferidos adecuadamente. A continuación, se describen los más importantes:

1. **Protocolos de comunicación:** son reglas y procedimientos que regulan el intercambio de información entre sistemas informáticos. El protocolo de comunicación más utilizado actualmente, en el contexto de las redes de repositorios, es OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*) (Open Archives Initiative, 2017), que permite la recolección e intercambio de metadatos entre repositorios.
2. **Esquemas de metadatos:** son conjuntos de elementos y reglas que definen qué información debe incluirse en los metadatos y cómo debe estructurarse.
3. **Directrices de interoperabilidad:** son recomendaciones y buenas prácticas que orientan a la población gestora y desarrolladora de repositorios de acceso abierto para asegurar la calidad y la consistencia de sus sistemas.

Al momento del presente diagnóstico (julio de 2023), la comunidad global de repositorios se encuentra en transición a la adopción de directrices de metadatos con mayor capacidad de descripción, por ejemplo, OpenAIRE 4.0 (OpenAIRE, 2018), basadas en esquemas con mayor poder expresivo como Datacite (2021). Este proceso de transición depende en gran medida de la actualización y modernización de las plataformas de *software* de repositorios, de la evolución de los procedimientos de depósito y de la transformación de los registros de metadatos existentes a los nuevos estándares. Dados los recursos implicados y las limitaciones de muchas instituciones latinoamericanas, no es posible asegurar con certeza cuánto le tomará a la región llevar a toda la red de repositorios a estos nuevos estándares.

El esquema de metadatos Dublin Core (2023) es el estándar utilizado actualmente en los repositorios de la región. Fue rápidamente adoptado por su simplicidad y flexibilidad, pero tiene serias limitaciones que afectan, especialmente, a los metadatos de repositorios como fuente. La principal limitación es su escasa expresividad semántica: a pesar de que las directrices pueden proveer términos y vocabularios para describir los recursos, la estructura de este esquema no permite expresar matices o relaciones más complejas entre los elementos de metadatos. Esto afecta, especialmente, a los metadatos de autoría (nombres, identificadores y datos de afiliación, puesto que no pueden ser relacionados entre sí debido a la estructura “plana” de Dublin Core).

Una dimensión ausente en los metadatos actuales pero relevante es el tipo de evaluación que un documento ha recibido. Esta información no está adecuadamente representada en los estándares actuales.

A continuación, se presentan los resultados del diagnóstico realizado sobre los metadatos recogidos por LA Referencia mediante el protocolo OAI-PMH, el esquema de metadatos Dublin Core y las directrices de interoperabilidad de LA Referencia basadas en las de OpenAIRE 3.0.

3.1. Nombres de autores y afiliaciones

El nombre de los autores y sus afiliaciones constituyen un elemento esencial para calcular métricas e indicadores bibliométricos básicos. Sin embargo, este suele ser un punto débil de las bases de datos bibliográficas en general (incluso las comerciales como Scopus o Web of Science) y de los metadatos de repositorios, en particular, debido a la carencia de convenciones, estándares y buenas prácticas a la hora de registrar correctamente estas informaciones.

En el caso de los metadatos de repositorios analizados, las mencionadas limitaciones del esquema de metadatos Dublin Core restringen, en su mayoría, la información disponible a los nombres de las personas autoras en el campo `dc.creator` o `dc.contributor.author`, sin posibilidad de relacionar estos nombres con datos de afiliación, identificadores persistentes o información contenida en otros campos. Respecto a la calidad de los nombres como dato, se observan los siguientes problemas:

1. **Variaciones en la nomenclatura:** los nombres de las personas autoras pueden ser escritos de diferentes maneras (inversiones de nombres, iniciales completas, abreviaturas, etc.), lo que dificulta la identificación y la consolidación de la producción científica de una persona autora específica. Esto puede variar aun dentro de un mismo país e incluso dentro de un mismo repositorio.
2. **Caracteres especiales:** la presencia de caracteres especiales, diacríticos o caracteres no latinos en los nombres de autores puede generar problemas de codificación, recuperación y correcta agregación.
3. **Nombre de instituciones:** es bastante frecuente encontrar distintas variantes de nombres de institución consignadas en el campo `autor`, ya sea combinadas con los nombres o separadas.

4. **Identificadores persistentes:** en algunas instituciones se ha adoptado la práctica de incluir identificadores nacionales u ORCID (*Open Research and Contributor ID*) en los campos de autor en un intento por superar las limitaciones de un esquema plano como Dublin Core. Como resultado, aunque infrecuente, pueden encontrarse estas informaciones en ocurrencias aisladas o concatenadas a los nombres de las personas autoras. No existen convenciones o estándares aceptados respecto a esta práctica, lo que resulta en soluciones de compromiso con poca utilidad para los agregadores nacionales, regionales o globales.

Respecto a la información de afiliación, las directrices utilizadas en la actualidad no responden de manera adecuada a este tema. Sí existen algunos esfuerzos locales para consignar las afiliaciones en otros campos, pero, nuevamente, la imposibilidad de vincular los distintos elementos le resta utilidad práctica real.

Muchas de estas limitaciones están contempladas (y en su mayoría resueltas) en esquemas y directrices de metadatos como Datacite y OpenAIRE 4.0, entre otros, que aportan la solución de agrupar los metadatos de la autoría en una estructura jerárquica que permite registrar, detalladamente, los nombres, identificadores varios de persona, afiliación e identificador de institución (Datacite). Este concepto de “entidad”, anteriormente solo implementado en los metadatos de Sistemas de Gestión de la Investigación (CRIS, por sus siglas en inglés), parece marcar el camino hacia una mejora de calidad y utilidad de los metadatos de autoría en el contexto de los sistemas de repositorios.

3.2. Identificadores persistentes, cobertura y niveles de adopción

Los identificadores persistentes se han convertido en elementos clave para los ecosistemas de información científica, ya que ofrecen una referencia perdurable (en ocasiones única) a recursos como investigaciones, datos, publicaciones y objetos digitales en la web. La función principal de los identificadores persistentes es crear enlaces fiables y permanentes entre diferentes elementos dentro del ecosistema global de investigación.

En el contexto general del cálculo de medidas, tanto tradicionales como alternativas, para dar muestra de los procesos de investigación y publicación científica, es necesario identificar de manera persistente (y deseablemente única) los objetos de producción académica en sus distintas formas, y a la población productora y consumidora del sistema científico.

Los sistemas de identificadores persistentes como PURL, HANDLE, DOI o ARK para resultados de investigación y ORCID, junto a identificadores nacionales para personas que contribuyen a resolver problemas como la obsolescencia de los enlaces, el problema de los homónimos y la falta de trazabilidad en la atribución adecuada (Digital Object Identifiers, 2022; Corporation for National Research Initiatives, 2023; ARK Alliance, 2023; ORCID, 2023a).

Se presenta a continuación el resultado del diagnóstico de los principales identificadores persistentes, su nivel de adopción e interoperabilidad en los metadatos de los repositorios analizados.

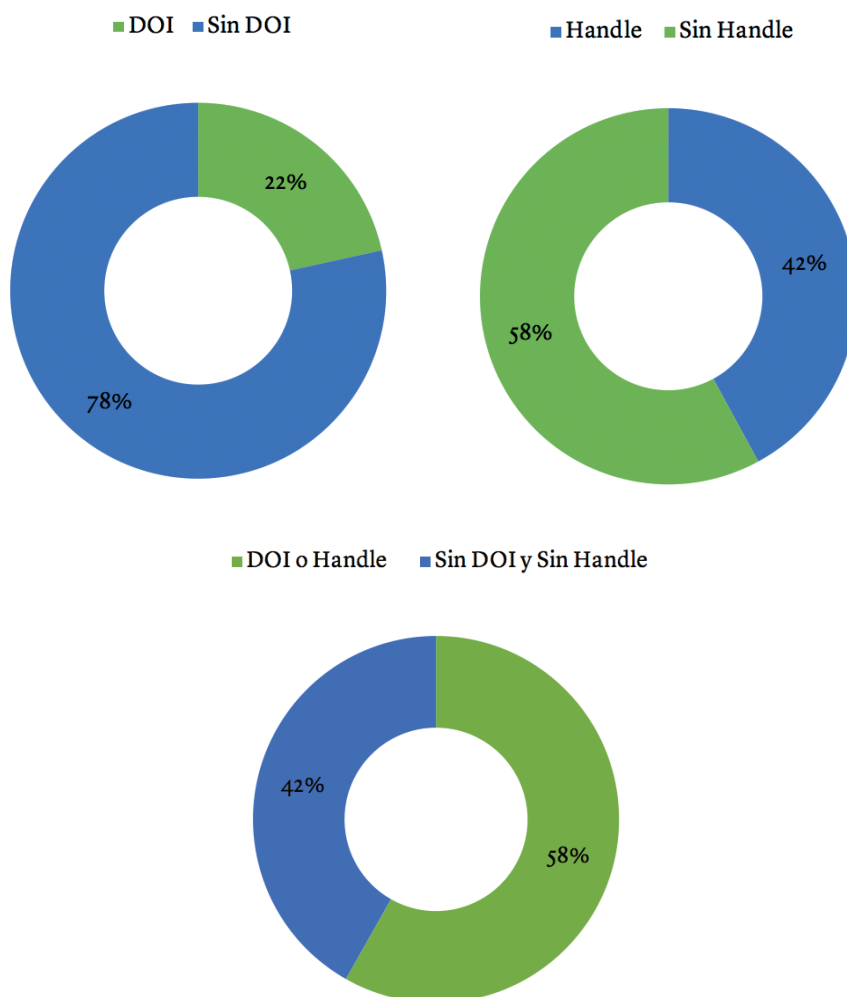
Identificadores de objeto digital

PURL: su uso dentro de la colección es prácticamente inexistente; se entiende que ha sido reemplazado por alternativas más modernas. Como mecanismo de referencia persistente sigue siendo muy utilizado en los vocabularios (COAR) y clasificaciones disciplinares nacionales.

Handle: la presencia de esta alternativa en los registros de la colección es cercano al 42%, más extenso que DOI. Sin embargo, Handle necesita un servicio por parte del repositorio que realice la resolución final del enlace y un pago anual mínimo para garantizar el servicio de resolución central. En muchos casos, pese a tener Handle asignado, los repositorios no resuelven correctamente, ya sea por cesación del pago o por problemas técnicos de la plataforma. Esto compromete la fiabilidad del Handle como identificador persistente (Figura 2).

DOI: la cobertura de este identificador es relativamente baja, alcanza menos del 22% de los registros de metadatos. Esto se debe en gran medida al costo que la atribución de los DOI representa para las instituciones, ya que se requiere el pago de membresía individual o consorciado a las agencias autorizadas (Crossref, Datacite). Probablemente no todos los DOI atribuidos estén presentes en los metadatos, pero esa verificación está fuera de los alcances de este diagnóstico (Figura 2).

FIGURA 2. Distribución del uso de identificadores persistentes en los metadatos analizados



Fuente: Elaboración propia.

ARK: se observa muy baja presencia de este identificador en los metadatos de los repositorios analizados. Este identificador es una alternativa a DOI que no requiere el pago de membresías para operar un servicio de atribución y resolución, pero en su lugar requiere que la institución provea la infraestructura de soporte, por lo que existen en la región algunas iniciativas como ARK-CAICYT (Authier *et*

al., 2021) o dARK (IBICT/LA Referencia) (Segundo *et al.*, 2022), que buscan aportar soluciones de identificadores persistentes más sostenibles para los ecosistemas nacionales de Ciencia Abierta.

Identificadores de personas autoras

ORCID (Open Researcher and Contributor ID) es un identificador único y persistente para investigadores y académicos/as ampliamente utilizado a escala global. No tiene costo de registro para las personas, pero ofrece membresías a las instituciones que desean una integración avanzada de servicios. De acuerdo con las estadísticas publicadas por ORCID (2023b) existen más de 750.000 registros ORCID correspondientes a los países analizados. Sin embargo, debido a las restricciones del esquema Dublin Core y a la falta de estándares claros de cómo exponer este identificador de forma consistente, los niveles de adopción y de presencia de ORCID en repositorios no se ven reflejados en los metadatos recolectados, ya que solo están presentes en menos del 2,5% de los registros. Es esperable que este problema sea resuelto con la actualización de las plataformas de *software* y la adopción de directrices de interoperabilidad más recientes.

Identificadores nacionales: en el plano de los sistemas nacionales existen distintos identificadores de personas registrados en los metadatos: documentos nacionales de identidad, identificadores de sistemas nacionales de currículum (Lattes en Brasil, por ejemplo) y otros identificadores institucionales de carácter local. Sin embargo, su nivel de registro en los metadatos no es exhaustivo ni consistente en todas las instituciones, aun dentro de un mismo país. Por otra parte, aunque es deseable contar con identificadores nacionales relacionados con los identificadores internacionales, su utilidad para la construcción de medidas en el contexto internacional es baja.

Otros identificadores relevantes

ISSN (International Standard Serial Number): este identificador de publicaciones seriadas se utiliza, principalmente, en las revistas de artículos depositados en los repositorios, y presenta una cobertura de alrededor del 52% de registros de tipo artículo recolectado. Este identificador puede resultar de gran utilidad a la hora de recuperar y valorar distintos parámetros de la revista asociada, lo cual puede ser utilizado para segmentar las colecciones y ofrecer distintas dimensiones de análisis de calidad.

3.3. Título, resumen, descriptores y clasificaciones disciplinares

Los elementos textuales como el título, el resumen (dc.description) y los descriptores (dc.subject) están considerados dentro de los campos básicos del esquema Dublin Core y cuentan con una cobertura superior al 85% dentro de los registros analizados.

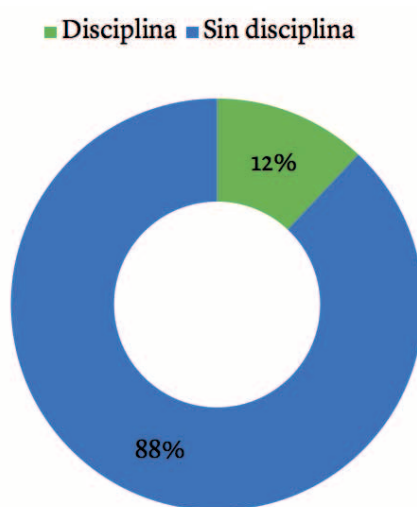
Título: es considerado un elemento obligatorio por las directrices de LA Referencia, de manera que todos los registros analizados presentan al menos un título.

Resumen: aproximadamente el 86% de los registros poseen algún tipo de texto en el campo de resumen; esto es importante como elemento de descripción, pero también permite proyectar el uso de distintas técnicas de análisis posteriores que permitan extraer o enriquecer otros metadatos.

Descriptores: alrededor del 84% de los casos analizados contienen descriptores. El problema es que, en su mayoría, no se sigue ningún estándar en la región, al menos en forma generalizada, por lo que en este campo conviven: textos desnormalizados, clasificaciones disciplinares variadas, vocabularios y términos normalizados según distintas normas y otras informaciones que impiden pensar en una metodología común para los cálculos de medidas representativas sobre esta dimensión.

Clasificaciones disciplinares: realizando un análisis de identificación de patrones, se han logrado capturar clasificaciones disciplinares dentro de los metadatos de descriptor (dc.subject). Se trata de algunas clasificaciones basadas en vocabularios PURL (Argentina y Perú), prefijos específicos como en el caso de CNPq (Brasil), REMERI (México) u otras clasificaciones en áreas de conocimiento presentes en algunas instituciones de varios países. Sin embargo, en ninguno de los casos las clasificaciones cubren porcentajes altos de las colecciones nacionales, lo que resulta insuficiente para poder proyectar cortes disciplinares coherentes o representativos a escala nacional. Con respecto a toda la colección, solamente el 12% presenta algún tipo de clasificación normalizada identificable, lo que obliga a pensar en aplicar alguna metodología de clasificación posterior que permita realizar estudios por cortes disciplinares, aunque sea en grandes áreas del conocimiento. En ese contexto, la alta cobertura de título, resumen y descriptores puede resultar de utilidad para la aplicación de distintas técnicas de clasificación automática.

FIGURA 3. Presencia vocabularios controlados en el campo disciplina



Fuente: Elaboración propia.

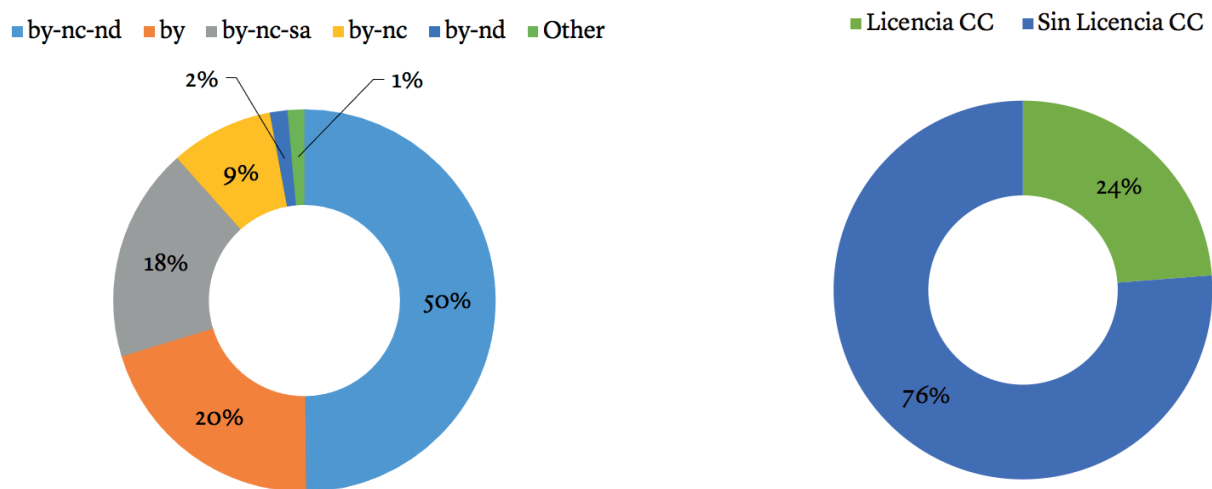
3.4. Metadatos sobre licencias

En el ámbito del acceso abierto, los tipos específicos de licencia con los que los objetos digitales son compartidos representan una dimensión de análisis relevante en el contexto de la medición del uso, reúso e impacto de los resultados de investigación financiados con fondos públicos.

Los metadatos recolectados muestran un universo bastante heterogéneo de licencias; muchas basadas en políticas institucionales particulares, y otras en formas más estandarizadas como Creative Commons.

Las licencias Creative Commons (2023) son un conjunto de herramientas legales que permiten a la población creadora de contenido compartir su trabajo, mientras mantienen ciertos derechos y control sobre cómo se utiliza su producción. Estas licencias ofrecen una alternativa flexible a los tradicionales derechos de autor “todos los derechos reservados”, al permitir a las personas autoras elegir cómo quieren que se utilice su obra. La cobertura de este tipo de licencias en los metadatos analizados alcanza alrededor del 23% de los registros, aunque presentan un alto nivel de desnormalización por las diferentes formas en que un mismo tipo de licencia puede ser representado. Luego de un proceso de normalización, la composición de los distintos tipos de licencias puede observarse en la Figura 4.

FIGURA 4. Distribución de los tipos de licencia Creative Commons en los documentos de LA Referencia



Fuente: Elaboración propia. CC BY (Atribución), CC BY-SA (Atribución-CompartirIgual), CC BY-ND (Atribución-NoDerivadas), CC BY-NC (Atribución-NoComercial), CC BY-NC-SA (Atribución-NoComercial-CompartirIgual), CC BY-NC-ND (Atribución-NoComercial-NoDerivadas), CCo (Dominio Público).

3.5. Metadatos sobre proyectos y financiadores

Los metadatos referentes al proyecto que dio origen a un resultado de investigación y a la fuente de financiamiento asociada son altamente relevantes en el contexto del monitoreo y evaluación de políticas institucionales, nacionales e incluso regionales.

De acuerdo con las directrices de metadatos actualmente utilizadas, basadas en OpenAIRE 3.0, existe un estándar claro para registrar estas informaciones (OpenAIRE, 2010). Sin embargo, en relación con los datos analizados, menos del 2% de los registros contienen datos de proyecto que cumplen esta recomendación. Esta falencia debe ser atendida, ya sea por medio de la implementación de la recomendación a nivel de los repositorios o por el enriquecimiento de los metadatos a nivel de agregador, mediante el procesamiento de otros metadatos o del objeto digital.

3.6. Metadatos de citación y referencias a otros resultados asociados

Los metadatos de citación y referencias a otros resultados de investigación asociados representan una fuente de recursos de análisis tanto desde el punto de vista de la bibliometría más tradicional, como para el cálculo de nuevas métricas de Ciencia Abierta (reúso de datos de investigación, por ejemplo).

En el contexto de los repositorios institucionales no existe, dentro de los metadatos actualmente disponibles, registro detallado de la información de citación utilizable. Sí se encuentran algunas referencias a otros resultados asociados dentro del campo `dc.relation`, aunque nuevamente la cobertura es baja y depende en gran medida de la existencia de identificadores persistentes. La actualización de la red a esquemas de metadatos con mayor capacidad de expresión posibilitaría el registro; sin embargo, no es esperable que esos metadatos sean consignados durante el proceso de depósito debido al esfuerzo implicado.

Algunas alternativas para posibilitar el enriquecimiento de los metadatos de repositorios con datos de citación son la interoperabilidad con los sistemas de revistas (Redalyc, Latindex, SciELO), la implementación

de mecanismos de interoperabilidad con otros sistemas (Datos de Investigación, CRIS o CV) como BRCRIS (Instituto Brasileiro de Informação em Ciência e Tecnologia, 2023) o COAR Notify (Confederation of Open Access Repositories, 2021), y la aplicación de técnicas de procesamiento más avanzadas sobre los objetos digitales con la finalidad de extraer información e integrarla en los metadatos.

4. Recomendaciones para la consolidación de la red de repositorios como fuente para monitoreo y evaluación de la investigación

En esta sección se proponen distintas acciones (algunas se encuentran en ejecución) que tienden a resolver gran parte de los problemas presentados en el diagnóstico. Las estrategias propuestas, fundamentalmente, apuntan a incrementar la cobertura, representatividad estadística y calidad de los metadatos presentes en la red de repositorios.

1. Integrar la producción científica disponible en otros sistemas. Un recurso técnico imprescindible para la población de los repositorios es la interoperabilidad con otros sistemas pertenecientes a los ecosistemas institucionales, nacionales y regionales de investigación, con el fin de recuperar la producción de la institución que por distintas razones no está presente en los repositorios. En ese sentido se recomiendan las fuentes potenciales:

- **Sistemas CRIS y de Currículo Nacionales:** los Sistemas de Gestión de la Investigación y de CV representan una fuente de importancia para los sistemas de monitoreo y evaluación (CLACSO, 2021), por lo que existen distintos proyectos nacionales y regionales para la integración de datos CRIS (PeruCRIS, BRCRIS, CONECTO) y LA Referencia ha integrado esta posibilidad en su plataforma de agregación. En ese sentido, es necesario desarrollar mecanismos para que todas estas integraciones alimenten, al menos al nivel de metadatos, a los repositorios. Una posible vía de interoperabilidad, actualmente en desarrollo, es la propuesta de COAR Notify, que permite intercambiar notificaciones con enlaces a los recursos entre los repositorios y distintos sistemas (Confederation Open Access Repositories, 2021).
- **Portales de revistas:** una parte importante de la producción científica regional corresponde a revistas gestionadas y alojadas en portales institucionales. Pero si bien algunos repositorios integran las revistas propias de su institución, incluso lo hacen en forma completa, no discriminan qué parte de los artículos corresponden a personas autoras de la institución. Al mismo tiempo no se integran metadatos de artículos de población investigadora propia publicada en revistas de otras instituciones. Esto representa un problema a la hora de poblar el repositorio y segmentar la producción real de la institución con fines estadísticos y de evaluación. De tal manera que es necesario adoptar una metodología consistente que permita resolver esta situación a medio plazo.
- **Sistemas regionales de revistas:** gran parte de las revistas de la región se encuentran indexadas/agregadas en sistemas regionales como Redalyc y SciELO. Asimismo, Latindex ofrece índices y criterios de calidad editorial que resultan muy valiosos en el contexto de la evaluación y que debieran estar integrados a los metadatos. En ese sentido, los sistemas regionales son una fuente confiable y normalizada que puede colaborar con poblar los repositorios de una forma consistente.
- **Respecto al problema expuesto en el punto anterior,** Redalyc permite la recuperación de los artículos a partir de la institución de las personas autoras. La Referencia, como red regional, está colaborando precisamente con las tres iniciativas de forma activa, en los mecanismos de interoperabilidad entre los repositorios y los sistemas de revistas, por lo que es imprescindible profundizar en esta línea de trabajo y hacer llegar estos servicios a los repositorios.

- Repositorios temáticos y otros sistemas de información: otra parte importante de la producción científica regional nunca llega a los repositorios institucionales, sino que se mantiene registrada en distintos sistemas (en ocasiones temáticos) nacionales, regionales o de otros continentes, tales como GBIF, SIDALC, BVS, CLACSO, Zenodo, OpenAIRE, Crossref, e incluso en sistemas de editoriales comerciales como Scopus. Por lo tanto, es necesario desarrollar metodologías consistentes que permitan recuperar y “repatriar” esa producción y registrarla en los repositorios institucionales.
2. Fortalecer políticas institucionales y nacionales de depósito de repositorios. Como complemento esencial a los medios técnicos de vinculación e interoperabilidad con las distintas fuentes de producción, es imprescindible fortalecer las políticas y mandatos de depósito de resultados de investigación tanto en el plano institucional como en el de las políticas nacionales de Ciencia Abierta. En ese sentido las recomendaciones de la UNESCO (2021) sobre Ciencia Abierta marcan un camino a los Estados miembros, que deberá ser acompañado por la actualización de las normativas nacionales en materia de Ciencia Abierta y evaluación de la investigación.
 3. Deduplicar las distintas versiones de un mismo producto de investigación. Es esperable que la convergencia de diversas fuentes, muchas veces solapadas, produzca duplicaciones de registros a nivel de los repositorios y agregadores. En ese contexto es fundamental implementar mecanismos de deduplicación confiables y homogéneos a todas las colecciones, para poder contabilizar y atribuir correctamente los distintos resultados de investigación. De tal manera que existen proyectos como BRCRIS (basado en la plataforma abierta de LA Referencia) y otras iniciativas que trabajan en proveer soluciones abiertas a esta problemática.
 4. Monitoreo de los niveles de cobertura y representatividad estadística. Esta acción propone extender la metodología utilizada en este texto con el objetivo de construir una herramienta abierta de monitoreo que permita evaluar, periódicamente, el avance de la red de repositorios con base en el camino trazado en este diagnóstico. En ese sentido, LA Referencia es el posible ámbito político y técnico donde estas capacidades pueden ser desarrolladas y compartidas con la comunidad en general.
 5. Adopción de directrices y esquemas de metadatos más completos. La adopción a nivel general de esquemas y directrices más completas representa, probablemente, el avance más importante y necesario en el camino de la solución de muchos de los problemas detectados. La hoja de ruta y estándares por adoptar han sido consensuados, por lo que existen repositorios que han logrado mejoras en este sentido, pero solo cuando la gran mayoría se actualice será posible visualizar los beneficios y su impacto en los servicios de valor agregado.
 6. Modernización de las plataformas de *software* de repositorios. Actualmente, la principal barrera para la adopción de nuevos estándares y la mejora de los mecanismos de interoperabilidad la representa la obsolescencia de las plataformas de repositorios instaladas en las instituciones. Esta acción debe ser ejecutada por cada institución y he aquí la complejidad de este punto, puesto que existen enormes restricciones de recursos tanto materiales como humanos en muchas instituciones de la región. En este sentido, es necesaria una coordinación con las instancias nacionales y regionales para obtener financiamiento y apoyar una modernización sistemática de la red de repositorios. En este punto, es importante mencionar que existe una coordinación de esfuerzos entre COAR, Lyris (iniciativa responsable de DSpace, la plataforma más utilizada) y LA Referencia con el objetivo de buscar fuentes de financiamiento para avanzar en este tema.
 7. Incremento de la cobertura de identificadores persistentes. Como fue expresado, los identificadores persistentes representan elementos clave en el uso de los metadatos para el monitoreo y la evaluación.

Las acciones por desarrollar en ese sentido dependen, por una parte, de la capacidad de poder registrar y exponer correctamente estas informaciones y, por otra, de contar con los recursos financieros para poder solventar los costos de algunos identificadores. Hay alternativas en desarrollo como ARK-CAICYT (Authier *et al.*, 2021) y dARK (IBICT/LA Referencia) (Segundo *et al.*, 2022), que intentan construir servicios de atribución y resolución de identificadores con perspectiva de bien público, pensadas para ofrecer identificadores a las instituciones que por distintas razones no pueden adquirir los servicios de DOI. En cualquier caso, la recomendación más importante es construir servicios y sistemas que sean agnósticos del tipo de identificador, que no resulten dependientes de ninguno en particular y permitan construir un ecosistema diverso e interoperable.

8. Acuerdos, normalización y curación de dimensiones relevantes para el monitoreo. En este trabajo se han presentado algunas dimensiones importantes que no están normalizadas bajo las directrices ni esquemas de metadatos presentes y, en algunos casos, tampoco de los que se encuentran en proceso de adopción. El caso más relevante es el de las clasificaciones disciplinares, las cuales constituyen una dimensión esencial a la hora de analizar y comprender algunos fenómenos que dependen esencialmente del área donde se produce el conocimiento, por lo que es necesario avanzar con la definición de algunos lineamientos que sean implementados por la red en general, con la colaboración de los ámbitos nacionales y regionales. Un caso aparte son los metadatos de citas, ausente en los metadatos de repositorios, pero valiosos a la hora de recrear medidas tradicionales y de construir medidas alternativas. Sin duda este es un punto en el cual se debe avanzar a medio plazo.
9. Desarrollo y aplicación de métodos automáticos de enriquecimiento de metadatos. En la actualidad, existen diversas técnicas de procesamiento de información. Las más recientes están basadas en los grandes modelos de lenguaje (LLM), que poseen la potencialidad de ser aplicadas al enriquecimiento de los metadatos. En este momento, se cuenta con experiencias tanto regionales (BRCRIS) como internacionales (OpenAIRE Graph) exitosas en el campo. En este sentido, es recomendable la búsqueda de recursos de financiamiento que permitan sistematizar estas técnicas y aplicarlas como solución de algunos de los problemas planteados en las secciones anteriores.

5. Consideraciones finales

De acuerdo con los análisis aquí presentados, los repositorios institucionales poseen la potencialidad de convertirse en verdaderas fuentes integradoras y representativas de la producción intelectual de las instituciones académicas y, por tanto, en herramientas valiosas para el monitoreo y evaluación de los procesos de generación y consumo de conocimiento científico. En ese sentido, es necesario actuar de manera constante sobre los mecanismos tanto políticos como técnicos que permitan consolidar esta misión en forma homogénea y consistente en todas las instituciones de la región. Paralelamente, es necesario modificar los incentivos a la comunidad investigadora para lograr el requerido cambio cultural a favor de la Ciencia Abierta.

Para lograr una integración efectiva de los diferentes sistemas de información que pueden incrementar la cobertura y representatividad de la red de repositorios, se requiere una coordinación constante y fluida entre las instancias institucionales involucradas. Las acciones técnicas que se plantean en este trabajo deben ir acompañadas de decisiones políticas firmes que impulsen los cambios necesarios en las normativas y la asignación de recursos que permitan consolidar un verdadero ecosistema regional de Ciencia Abierta con una visión de bien público.

Finalmente, algunos de los elementos y medidas presentados en este diagnóstico pueden ser sistematizados con el objetivo de crear una herramienta de monitoreo periódico del impacto de las acciones propuestas.

Índice de siglas y acrónimos

ARK	Archival Resource Key
BRCRIS	Ecosistema de Informação da Pesquisa Científica Brasileira
BVS	Biblioteca Virtual en Salud
CAICYT	Centro Argentino de Información Científica y Tecnológica
CC BY	Creative Commons Attribution
CC BY-NC	Creative Commons NonCommercial
CC BY-NC-SA	Creative Commons NonCommercial Attribution-ShareAlike
CC BY-ND	Creative Commons Attribution-NoDerivatives
CC BY-SA	Creative Commons Attribution-ShareAlike
CCo	Creative Commons No Rights Reserved
CECIC	Centro de Estudios de Circulación del Conocimiento
CLACSO	Consejo Latinoamericano de Ciencias Sociales
CNPQ	Conselho Nacional de Desenvolvimento Científico e Tecnológico
COAR	Confederación de Repositorios de Acceso Abierto
COARA	Coalición para el Avance de la Evaluación de la Investigación
CONECTO	Sistema de perfiles de la ciencia y tecnología de la República de Panamá
CONICET	Consejo Nacional de Investigaciones Científicas y Técnicas
CRIS	Sistemas de Gestión de la Investigación
dARK	Decentralized Archival Resource Key
DOI	Identificador de objeto digital
FOLEC	Foro Latinoamericano sobre Evaluación Científica
GBIF	Global Biodiversity Information Facility
IBICT	Instituto Brasileiro de Informação em Ciência e Tecnologia
ISSN	International Standard Serial Number
LA Referencia	La Red Federada de Repositorios Institucionales de Publicaciones Científicas
LATINDEX	Sistema Regional de Información en Línea para Revistas Científicas de América Latina, el Caribe, España y Portugal)
LLM	Large Language Model
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OLIVA	Observatorio Latinoamericano de Indicadores de Evaluación
ONCYT	Organismos Nacionales de Ciencia y Tecnología
OpenAIRE	Open Access Infrastructure for Research in Europe
ORCID	Open Research and Contributor ID
PURL	Persistent uniform resource locator
RedALyC	Red de Revistas Científicas de América Latina y el Caribe, España y Portugal
REMERI	Red Mexicana de Repositorios Institucionales
RICYT	Red Iberoamericana de Indicadores de Ciencia y Tecnología
SciELO	Scientific Electronic Library Online
SIDALC	Sistema de Información y Documentación Agropecuaria de América Latina y el Caribe
SJR	SCImago Journal Rank
UNESCO	Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura

Referencias bibliográficas

- ARK ALLIANCE (2023): *ARK Alliance. Home of the Archival Resource Key (ARK)*. Disponible en: <https://arks.org/>.
- AUTHIER, C.; FERREYRA, D. y BIGLIERI, H. (2021): *ARK como identificador persistente gratuito: 2021 un año crucial para los identificadores persistentes en la ciencia*. Disponible en: <https://zenodo.org/record/4665799/files/CAICYT-CONICET%20Argentina.pdf>.
- BRADFORD, S. C. (1934): "Sources of information on specific subjects", *Engineering*, 137. pp. 85-86.
- CLACSO (2021): Herramienta 1: los sistemas CRIS, su potencialidad para visibilizar diversas formas de producción de conocimiento e impulsar nuevas modalidades de evaluación. Hacia la Transformación de los Sistemas de Evaluación en América Latina y el Caribe. Foro Latinoamericano de Evaluación Científica (FOLEC). Disponible en: <https://biblioteca-repositorio.clacso.edu.ar/handle/CLACSO/14763>.
- (2023): *FOLEC-Foro Latinoamericano sobre Evaluación Científica*. Disponible en: <https://www.clacso.org/folec/que-es-el-folec/>.
- COALITION FOR ADVANCING RESEARCH ASSESSMENT (2023): *Coalition for Advancing Research Assessment*. Disponible en: <https://coara.eu/>.
- CONFEDERATION OPEN ACCESS REPOSITORIES (2021): *COAR Notify Initiative*. Disponible en: <https://www.coar-repositories.org/notify/>.
- CONFEDERATION OPEN ACCESS REPOSITORIES (2021): *Repositories*. Disponible en: <https://www.coar-repositories.org/news-updates/european-repositories-infographic/>.
- COMISIÓN EUROPEA (2023): *Políticas de acceso abierto en América Latina, el Caribe y la Unión Europea: avances para un diálogo político*, Dirección General de Investigación e Innovación. Disponible en: <https://data.europa.eu/doi/10.2777/162>.
- CORPORATION FOR NATIONAL RESEARCH INITIATIVESHANDLE.NET REGISTRY (2023): *HDL.NET® Information Services*. Disponible en: <https://www.handle.net/>.
- CREATIVE COMMONS (2023): Sobre las licencias. Lo que hacen nuestras licencias. Disponible en: https://creativecommons.org/licenses/?lang=es_ES.
- DATAcite (2021): Datacite Schema. Disponible en: <https://schema.datacite.org/meta/kernel-4.4/>.
- DIGITAL OBJECT IDENTIFIER (2022): *Doi Foundation*. Disponible en: <https://www.doi.org/>
- DUBLIN CORE (2023): *DCMI Metadata Terms*. Disponible en: <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#section-1>.
- GALLARDO, O. (2002): *La evolución de la producción científica en las revistas indexadas en Iberoamérica*, OEI. Disponible en: <https://observatoriocts.oei.org.ar/wp-content/uploads/2022/09/Papeles23ESP.pdf>.
- INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA (2023): *BRCRIS*. Disponible en: <https://brcris.ibict.br>.
- LA REFERENCIA (2012): *Acuerdo de Cooperación Regional - Acta de Buenos Aires que constituye LA Referencia*. Disponible en: <http://www.lareferencia.info/es/recursos/documentos/acuerdos-politicos/2-acuerdo-de-cooperacion-regional-acta-de-buenos-aires-que-constituye-la-referencia-2012>.
- (2023): *Declaración 10 Años de LA Referencia: Hacia un ecosistema de Ciencia Abierta No Comercial*. Disponible en: <https://www.lareferencia.info/es/component/k2/item/307-declaracion-10-la-referencia>.
- OPEN ARCHIVES INITIATIVE (2017): *Open Archives Initiative*. Disponible en: <https://www.openarchives.org/>
- OPENAIRE (2010): *OpenAIRE Guidelines for Literature Repositories v3*. Disponible en: <https://guidelines.openaire.eu/en/latest/literature/introduction.html>.
- (2018): *Directrices de OpenAIRE para administradores de repositorios de Literatura v4*. Disponible en: <https://guiasopenaire4.readthedocs.io/es/latest/>.
- ORCID (2023): *ORCID statistics*. Disponible en: <https://www.google.com/url?q=https://info.orcid.org/orcid-statistics/&sa=D&source=docs&ust=1692022455816528&usg=AOvVaw34lCeNuyvX7x5dfjiRGB>.
- ORCID (2023a): *About ORCID*. Disponible en: <https://info.orcid.org/what-is-orcid/>
- SCIMAGO (2022): *SJR Scimago Journal & Country Rank*. Disponible en: <https://www.scimagojr.com/>.

- SEGUNDO, W.; MATAS, L.; NÓBREGA, T.; FILHO, F. y MENA-CHALCO, J. (2022): dARK: A decentralized blockchain implementation of ARK Persistent Identifiers (1.1). Disponible en: <https://doi.org/10.5281/zenodo.7686101>.
- SILVEIRA, L. DA; RIBEIRO, N. C.; MELERO, R.; MORA-CAMPOS, A.; PIRAQUIVE-PIRAQUIVE, D. F.; URIBE-TIRADO, A.; SENA, P. M. B.; POLANCO-CORTÉS, J.; SANTILLÁN-ALDANA, J.; DA SILVA, F. C. C.; ARAÚJO, R. F.; ENCISO-BETANCOURT, A. M. y FACHIN, J. (2023): “Taxonomia da Ciência Aberta: revisada e ampliada”, *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, 28, 1-22. Disponible en: <https://doi.org/10.5007/1518-2924.2023.e91712>.
- UNESCO (2021): *Recomendación de la UNESCO sobre la Ciencia Abierta*. Disponible en: https://unesdoc.unesco.org/ark:/48223/pfo000379949_spa.