ISBN: 978-84-1091-135-2 (pdf) / 978-84-1091-136-9 (Impreso POD)
DOI: https://doi.org/10.14201/0AQ03865168

LA CURACIÓN DE DATOS COMO METODOLOGÍA DE INVESTIGACIÓN

DATA CURATION AS RESEARCH METHODOLOGY

Valeria Montoya-Roncancio[1] e Inmaculada Bravo García[2]

[1]*Universidad de Salamanca. España*
valeriamontoya@usal.es
https://orcid.org/0000-0001-7357-3326

[2]*Universidad de Salamanca. España*
inma@usal.es
https://orcid.org/0000-0003-3388-0545

Resumen: En el contexto actual de creciente dependencia de los datos abiertos y reutilizables en la evaluación de la ciencia, la curación de datos se presenta como una metodología esencial en el campo de la Documentación. En esta investigación, se propone una metodología que combina técnicas de recuperación, normalización y análisis de datos, centradas en la verificación de registros provenientes de plataformas bibliográficas de datos como Web of Science (WoS) y Scopus utilizando identificadores persistentes como DOI, ORCID, ResearcherID, ScopusID e ISSN, con el fin de garantizar la fiabilidad del estudio. Este enfoque permite identificar inconsistencias, duplicidades o ausencias de información relevante en los perfiles de autor y publicaciones científicas. Los resultados muestran la producción de la Universidad de Salamanca, como estudio de caso, donde se identificaron perfiles de autores con datos que requerían corrección en ambas plataformas y en el portal de producción científica institucional. Estas incidencias evidencian la necesidad de procesos sistemáticos de curación en la gestión documental, destacando el uso de este modelo para mejorar la calidad de los sistemas de información científica. Su aplicación en el ámbito de la Documentación favorece la generación de métricas confiables y la toma de decisiones basada en información verificada y reutilizable.

Palabras clave: curación de datos; evaluación científica; metadatos; identificadores persistentes; Web of Science; Scopus.

Abstract: In the current context of increasing reliance on open and reusable data in the evaluation of science, data curation is presented as an essential methodology in the field

of Documentation. In this research, we propose a methodology that combines data retrieval, normalization and analysis techniques, focused on the verification of records coming from bibliographic data platforms such as Web of Science (WoS) and Scopus using persistent identifiers such as DOI, ORCID, ResearcherID, ScopusID and ISSN, in order to ensure the reliability of the study. This approach allows identifying inconsistencies, duplicities or absence of relevant information in author profiles and scientific publications. The results show the production of the University of Salamanca, as a case study, where author profiles were identified with data that required correction in both platforms and in the institutional scientific production portal. These incidences show the need for systematic curation processes in document management, highlighting the use of this model to improve the quality of scientific information systems. Its application in the field of Documentation favors the generation of reliable metrics and decision making based on verified and reusable information.

Keywords: data curation; scientific evaluation; metadata; persistent identifiers; Web of Science; Scopus.

1. INTRODUCCIÓN

La creciente disponibilidad de información científica y la transformación digital del conocimiento han situado a los datos en el centro de procesos de investigación, gestión, evaluación y reutilización de contenidos. En este contexto, la curación de datos se presenta como una metodología clave que permite garantizar la consistencia, la calidad y la veracidad de la información proveniente de plataformas bibliográficas y bibliométricas que se encuentra depositada en los portales de investigación de las universidades.

Este capítulo ofrece una aproximación a la curación de datos desde el ámbito de las Ciencias de la Documentación, abordando su definición, su aplicación en la localización y tratamiento de datos académicos, así como su utilidad para la evaluación científica. Se examinan las principales herramientas de referencia, como Web of Science y Scopus junto con el papel de los identificadores persistentes en la recuperación masiva de registros bibliográficos. La propuesta metodológica se orienta a optimizar el uso de estos datos contribuyendo a una gestión documental más transparente, eficiente y alineada con los estándares de calidad a los que se enfrenta la comunidad científica.

1.1. Definición y fundamentos de la curación de datos

La curación de datos se entiende como el proceso que permite identificar las fuentes de datos, organizarlas, descargarlas, limpiarlas y prepararlas para la creación de análisis y extracción de información. Estas actividades están destinadas para que los datos de investigación sean comprensibles y reutilizables e incluyen la documentación, estandarización, formateo y asociación de metadatos relevantes para promover la reutilización interdisciplinaria de los datos científicos (Minamiyama, Takeda, Hayashi, Asaoka y Yamaji, 2024).

La curación de datos también se puede definir más específicamente como «las verificaciones y acciones realizadas por los curadores con el fin de asegurar que el conjunto de datos esté estructurado y documentado de la manera más completa posible y de acuerdo con las mejores prácticas» (SciELO, 2023). A su vez, la curación de datos se describe como una actividad de gestión que implica el desarrollo de infraestructuras lógicas y físicas que permite recolectar, indexar, almacenar y facilitar la consulta de los datos para análisis posteriores. Este enfoque evidencia la importancia de las infraestructuras sociotécnicas y de la gestión para acceder a los datos (Chua *et al*., 2022; citado por Parmiggiani *et al*., 2024).

Según Choi y Xin (2021), la curación de datos es el proceso que permite gestionar datos para que estén disponibles para su reutilización y preservación, permitiendo usos que cumplan con los requisitos FAIR (Findable, Accessible, Interoperable, and Reusable). Los principios FAIR establecen que los datos científicos deben ser localizables, accesibles, interoperables y reutilizables, con el fin de maximizar su visibilidad, integrabilidad y valor en la investigación. Estos autores destacan la necesidad de la curación en el ciclo de vida de la investigación, asegurando que los datos sean descubiertos y reutilizados eficientemente.

Yakel (2007) define la curación de datos como la gestión activa y continua de los datos a lo largo de su ciclo de vida, con el fin de conservar su utilidad para la ciencia, la investigación y la educación a lo largo del tiempo, resaltando el alcance de la curación en el acceso y la preservación de datos útiles para la comunidad académica.

Estas definiciones reflejan la evolución del concepto de curación de datos en la literatura científica, mostrando el papel crucial que este proceso posee en la gestión, la preservación y la reutilización de datos en diversos contextos de investigación. La curación de datos implica diversas acciones para asegurar que los datos sean accesibles y útiles para futuras investigaciones.

El objetivo de este estudio es describir una estrategia de automatización para la curación de datos en el campo de la Documentación. Un campo científico en el que es necesario no solo almacenar o recuperar información; sino que estudia cómo se produce, organiza, conserva y difunde el conocimiento, y requiere métodos claros y sistemáticos para que sus resultados sean válidos y replicables. Actualmente, en esta rama del conocimiento los datos se tratan cada vez más desde una perspectiva técnica, mediante herramientas que permiten identificar metadatos incompletos, mejorar la interoperabilidad entre plataformas y verificar la persistencia de identificadores. La curación de datos se consolida así como una metodología clave para garantizar que los registros documentales sean fiables, estandarizados y útiles para la evaluación, la investigación y la toma de decisiones informadas (Peng y Wyborn, 2022).

El modelo de curación de datos que se presenta en esta investigación se basa en la extracción y la estructuración de datos abiertos de carácter de autoría y bibliográfico. La propuesta parte de la localización de datos de interés para la investigación relativos a publicaciones y autores mediante el tratamiento de datos de identificadores persistentes procedentes de plataformas bibliográficas y bibliométricas como Web of Science y Scopus.

1.2. Localización de datos en plataformas bibliográficas y bibliométricas

Web of Science (WoS), operada por Clarivate, es una plataforma de información científica que proporciona acceso a publicaciones académicas. Su principal función es facilitar la búsqueda, el análisis y la evaluación de literatura científica y técnica de alta calidad, cubriendo una amplia gama de disciplinas. Esta base de datos bibliométrica es clave para calcular indicadores como el Índice H y el Factor de Impacto (FI). Se caracteriza por su rigurosidad en la selectividad de revistas científicas y en sus índices especializados como el Science Citation Index o el Social Sciences Citation Index.

Scopus, gestionado por Elsevier, es una base de datos multidisciplinar que ofrece una amplia literatura académica, científica y técnica. Contiene información bibliográfica y de citación de artículos publicados en revistas, libros y actas de congresos. Es especialmente valorada por su capacidad de generar métricas de impacto y colaboración científica. Scopus permite identificar tendencias de investigación, redes de coautoría y autores destacados. Ambas plataformas, WoS y Scopus, son esenciales para el análisis y la medición de la actividad investigadora y constituyen las principales fuentes de referencia para la extracción de indicadores bibliométricos empleados en la evaluación de la ciencia. Además, permiten calcular métricas como el Índice H, el número de citas, el impacto y la visibilidad de las publicaciones.

1.3. Recuperación masiva de registros bibliográficos

Las interfaces de programación de aplicaciones (API, por sus siglas en inglés) constituyen un conjunto estandarizado de métodos que permiten acceder a los datos ofrecidos por plataformas tales como Scopus y Web of Science, de forma controlada, eficiente y con campos y formatos predefinidos por cada proveedor (Torres-Salinas *et al*., 2022; Vélez-Estévez *et al*., 2023). Esta vía facilita la recuperación masiva y precisa de registros bibliográficos, así como su integración en sistemas propios de análisis documental. Por su parte, el *web scraping* es una técnica informática que permite extraer información directamente de páginas web en las que no se ofrece acceso a través de una API oficial, mediante el análisis del código HTML y la simulación de la navegación humana.

La API de Web of Science, desarrollada por Clarivate, permite acceder a los datos indexados en la Web of Science Core Collection. Esta interfaz está diseñada para usuarios institucionales que cuenten con una suscripción activa y una clave de autenticación. A través de esta API, es posible recuperar metadatos bibliográficos detallados, información sobre autores, afiliaciones, recuentos de citas, referencias bibliográficas y otros elementos clave para el análisis bibliométrico. Por su parte, la API de Scopus, gestionada por Elsevier, proporciona acceso estructurado a la base de datos bibliográfica Scopus mediante una clave de API asociada a una suscripción institucional. Esta interfaz ofrece funcionalidades similares a las de Web of Science, como la búsqueda y la recuperación de publicaciones, citas, referencias y metadatos de autores y afiliaciones.

Para el caso de DOI, la IDF (International DOI Foundation) gestiona un sistema de resolución basado en el Handle System y proporciona una API en https://doi.org/ que actúa como un sistema de resolución de identificadores DOI, redirigiendo cada DOI a la ubicación actual del recurso digital al que hace referencia. Si bien la IDF se estableció en 1997, la infraestructura para solicitar y asignar DOI no estuvo operativa hasta 2000. CrossRef fue el primer registro de DOI especializado en publicaciones académicas, facilitando la adopción masiva del sistema por revistas científicas y editoriales.

Este estudio describe una estrategia para la recuperación de datos a través de API, llevando a cabo la extracción, la transformación y la carga de los datos que, convenientemente estructurados y relacionados, ofrecerán información para la realización de otros estudios e investigaciones. La curación de datos permitirá detectar duplicidades, errores e inconsistencias garantizando que los datos consultados sean precisos, útiles y sirvan como base para la toma de decisiones, implementación de políticas y gestión documental. En la metodología que se presenta, se evidenciaron las herramientas utilizadas para la recuperación de información, el tratamiento de los datos y la exportación de los mismos, lo que facilitará la obtención de indicadores relevantes, así como la identificación de incidencias, al consultar algunos identificadores persistentes.

1.4. Identificadores persistentes como elementos clave

En el ecosistema de la información científica, los identificadores persistentes desempeñan un papel clave para garantizar la trazabilidad, la desambiguación y la vinculación correcta de entidades como autores, instituciones y publicaciones. Los identificadores persistentes se emplean como el medio para localizar información de investigación de modo eficiente y precisa de autores con firmas parecidas.

Entre los identificadores más relevantes se encuentra el DOI (Digital Object Identifier). Un DOI (Digital Object Identifier) es un identificador único y persistente asignado a un objeto digital, como un artículo académico, para garantizar su localización y acceso a largo plazo. Su sintaxis siempre consta de un prefijo numérico «10» seguido de un código numérico asignado al registrante (editor o publicador) y un sufijo único para cada objeto. La forma típica de un DOI es: 10.xxxx/xxxxxxxx. La asignación de DOI a publicaciones comenzó en el año 2000, cuando CrossRef, una organización fundada por editores académicos, empezó a registrar identificadores DOI para artículos científicos y otros documentos.

Si bien la International DOI Foundation (IDF) se estableció en 1997, la infraestructura para solicitar y asignar DOIs no estuvo operativa hasta 2000. CrossRef fue el primer registrador de DOIs especializado en publicaciones académicas, facilitando la adopción masiva del sistema por revistas científicas y editoriales. Desde entonces, el uso del DOI se ha expandido a otros ámbitos, incluyendo *datasets*, reportes, tesis y hasta materiales audiovisuales.

En el ámbito de la identificación de autores, destaca el ORCID (Open Researcher and Contributor ID); es un identificador digital persistente lanzado en 2012 por ORCID. Es gestionado por el propio investigador y permite unificar su producción científica en distintas

plataformas, promoviendo la interoperabilidad y la correcta atribución de la actividad académica. ORCID es un identificador abierto que permite consolidar la producción científica de un investigador más allá de las variaciones nominales.

ResearcherID es un sistema de identificación de autores científicos. El sistema fue introducido en enero de 2008 por Thomson Reuters Corporation. Este identificador único tiene como objetivo solucionar el problema de la identificación de autores y la correcta atribución de obras dentro de su base de datos. En 2019 se integra con Publons y en 2022 Publons desaparece y se integra en Web of Science (Fundación Española para la Ciencia y la Tecnología, 2022).

Asimismo, ScopusID es un identificador único generado automáticamente por la base de datos bibliográfica Scopus, gestionada por Elsevier, que agrupa todas las publicaciones de un autor bajo un mismo perfil. Este sistema facilita la correcta identificación y la desambiguación de autores, especialmente en casos de homónimos o variaciones en la forma de escribir el nombre.

Ambos identificadores, ResearcherID y ScopusID, permiten consolidar métricas bibliométricas, como el número de citas y el Índice H, asociados a la producción científica de cada investigador. Los autores pueden solicitar la corrección o la fusión de sus perfiles para asegurar la precisión de los datos. Son herramientas fundamentales para la evaluación del rendimiento académico y el análisis de redes de colaboración.

A nivel institucional, el ROR (Research Organization Registry) actúa como identificador persistente para instituciones académicas, contribuyendo a mejorar la interoperabilidad de los datos de filiación.

Para identificar las revistas científicas se utiliza el ISSN. El ISSN (International Standard Serial Number) es un identificador numérico normalizado de ocho dígitos que se asigna a las publicaciones seriadas, como revistas científicas, boletines y anuarios, con el fin de facilitar su identificación unívoca a nivel internacional. Establecido por la norma ISO 3297, el ISSN no guarda relación con el contenido ni con la editorial, sino que distingue cada título de publicación seriada de forma independiente. El octavo dígito es un código de control que se calcula en función de un algoritmo Módulo 11, sobre la base de los 7 dígitos anteriores.

La gestión global del sistema ISSN está a cargo del Centro Internacional del ISSN, cuya plataforma principal es el sitio web ISSN.org. Desde allí, se coordina una red de centros nacionales e institucionales que se encargan de asignar y mantener los ISSN en sus respectivos países. Además, el sitio proporciona acceso a la base de datos ISSN Portal, un registro internacional donde es posible consultar información normalizada sobre publicaciones seriadas registradas, incluyendo detalles como título, editor, país, frecuencia y soporte. Este sistema cumple un papel esencial en la trazabilidad, la evaluación y la preservación del registro bibliográfico de la literatura seriada en el ámbito académico y editorial. El Centro Internacional del ISSN ha asignado más de 2.5 millones de ISSN. La base de datos del ISSN se actualiza constantemente y aumenta aproximadamente entre 50.000 y 70.000 ISSN por año.

Cuando una publicación está disponible en más de un soporte, por ejemplo, en formato impreso y electrónico, se le asignan dos ISSN diferentes: el ISSN para la versión impresa y el e-ISSN (o ISSN electrónico) para la versión digital. Esta distinción es fundamental en los procesos de catalogación, citación, indexación en bases de datos y evaluación de publicaciones,

ya que permite diferenciar claramente entre ediciones que pueden tener distintas fechas de publicación, paginación u otras características editoriales.

El número ISSN de enlace, o ISSN-L, es un número ISSN específico que reúne las ediciones en diferentes soportes (impreso, electrónico) de una misma publicación seriada (cada edición tiene su propio número ISSN). La tabla de números ISSN-L es un archivo con el cual los administradores de bases de datos o de catálogos pueden procesar con mayor eficacia los números ISSN presentes en sus fuentes, debido a que el número ISSN-L se utiliza como clave de enlace entre las diferentes ediciones identificadas por sus números ISSN específicos (ISSN International Standard Serial Number, s. f.).

Tal como señala Paloma Marín-Arraiza (2022), estos identificadores persistentes no solo favorecen la desambiguación en contextos bibliométricos complejos, sino que «forman parte de una infraestructura crítica para una ciencia abierta, conectada e interoperable», permitiendo el desarrollo de servicios avanzados en los sistemas de información científica. En este contexto, el análisis de la producción científica requiere no solo de identificadores normalizados, sino de herramientas que reflejen su impacto y visibilidad.

1.5. España y sus principales herramientas de referencias

En el caso de España, Web of Science y Scopus son las principales herramientas de referencia empleadas por las agencias evaluadoras como la FECYT (Fundación Española para la Ciencia y la Tecnología) y ANECA (Agencia Nacional de Evaluación de calidad y Acreditación) para valorar la calidad y el impacto para la producción científica de investigadores, revistas y universidades.

La FECYT, en el ámbito de la evaluación científica, no solo gestiona el acceso a las bases de datos como WoS y Scopus, sino que coordina los procesos de acreditación de revistas científicas españolas garantizando su calidad y visibilidad internacional. Este reconocimiento se otorga a través del Sello de Calidad FECYT, donde el impacto y la indexación en WoS y Scopus contribuyen positivamente en la evaluación de las revistas (FECYT, 2025). La ANECA es la entidad responsable de la acreditación académica, la medición de la calidad del profesorado universitario, la mejora institucional, la evaluación de sexenios y de los títulos oficiales en España. Para esta evaluación utiliza también criterios bibliométricos e indicadores de calidad basados en la calidad, la relevancia y el impacto de las contribuciones científicas en bases de datos como WoS y Scopus (ANECA, 2024).

Este estudio parte de un planteamiento sobre la falabilidad de la producción científica difundida en plataformas bibliográficas como son Web of Science y Scopus. En este sentido, Harari (2024) señala que «rechazar la fantasía de la infalibilidad, permite construir una red de información que considera que el error es inevitable». Esta perspectiva refuerza la necesidad de contar con una metodología rigurosa que permita el tratamiento de datos fiables, estructurados y persistentes, estableciendo un puente entre los requisitos institucionales de calidad, la gestión técnica de la información y la toma de decisiones fundamentada y transparente.

En este contexto, la curación de datos se convierte en una herramienta metodológica fundamental para acceder a los datos de una forma más eficaz y facilitar su reutilización. Esta metodología permite optimizar el uso de los datos extraídos en plataformas como Web of Science y Scopus, especialmente en el análisis científico y en los procesos de evaluación académica. Asimismo, al visibilizar y depurar estos datos, se pueden identificar incidencias que influyen directamente en la medición de la producción científica de los investigadores y, por tanto, en su impacto real dentro del ecosistema científico.

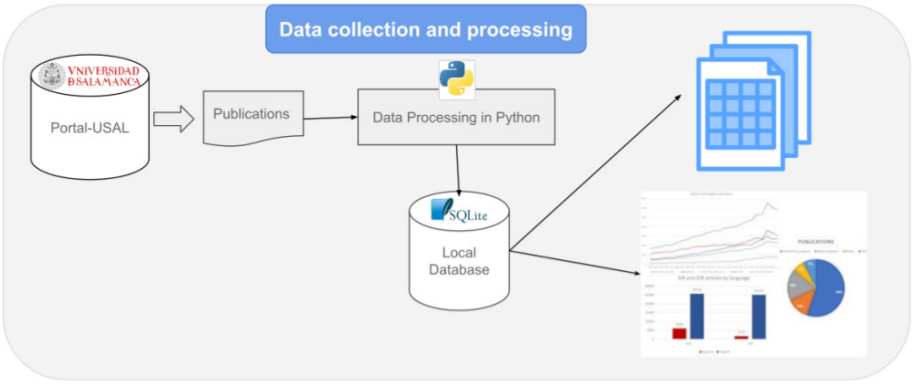
2. METODOLOGÍA

La metodología parte de una aproximación teórica y de un estudio de caso, planteado como estrategia de un caso real de curación de datos de investigación. La primera metodología empleada fue la investigación bibliográfica, mediante la que se realizó el contexto teórico de la curación de datos de investigación y los aspectos relevantes tratados en este estudio: extracción, transformación y estructuración de datos, métodos de gestión de datos y plataformas bibliográficas de referencia. El estado de la cuestión priorizará la descripción de las técnicas de gestión de datos más útiles, como son el uso de API, el *web scraping* o extracción de datos automatizada desde la web y los identificadores persistentes.

La segunda metodología es el estudio de caso, donde se presenta la estrategia práctica de curación de datos de aplicación en la investigación. Se introduce el modelo ETL como proceso de integración y carga de datos para el cálculo y el análisis de la información consultada. El procedimiento ETL consiste en Extract (Extracción), Transform (Transformación) y Load (Cargar). Por lo que se extraen los datos de sistemas heredados, se depuran para mejorar su calidad y establecer una coherencia y, por último, se introducen los datos en una base de datos de destino para su análisis (García García-Doncel, 2023).

En la arquitectura de procesamiento de datos bibliográficos, el modelo ETL constituye una metodología fundamental para sistematizar la curación de grandes volúmenes de información procedente de fuentes heterogéneas. Esta estrategia se estructura en tres fases diferenciadas pero interdependientes. En primer lugar, la fase de extracción (Extract) implica la recuperación de registros bibliográficos desde diferentes orígenes. Posteriormente, en la etapa de transformación (Transform), los datos extraídos son sometidos a procesos de depuración y estandarización que incluyen la eliminación de duplicados, la corrección de errores tipográficos y la normalización de los campos recuperados. Esta fase es crucial para garantizar la calidad, la coherencia y la comparabilidad de los datos. Finalmente, la etapa de carga (Load) consiste en la incorporación de los datos transformados a un sistema de almacenamiento o base de datos centralizada, optimizada para su posterior análisis bibliométrico, visualización o explotación. Este proceso se realizó tomando como estudio de caso el Portal de investigación de la Universidad de Salamanca y se puede observar en la Figura 1.

Figura 1. Modelo ETL



Fuente: Elaboración propia.

La curación de datos aplicada a este portal de investigación se llevó a cabo a través de las siguientes seis fases: recopilación de DOI, extracción de datos mediante API, obtención del listado de revistas científicas, análisis de identificadores de revistas (ISSN y EISSN), análisis de identificadores de autores y procesamiento y almacenamiento de datos.

2.1. Primera fase: Recopilación de datos del portal de producción científica de la universidad

En esta primera fase se realizó la compilación de los investigadores de la universidad y de las publicaciones con DOI correspondientes a cada autor. Esta información fue consolidada desde bases de datos internas o información proporcionada por los participantes.

Se decidió utilizar únicamente las publicaciones con DOI informado, para partir de un dato unívoco y así poder recuperar y comparar la misma publicación en diferentes fuentes, evitando generar errores derivados de la posible ambigüedad que se produciría de otra manera; de igual modo, los autores seleccionados para el análisis fueron los autores con al menos una publicación con DOI.

En esta fase, además de extraer las publicaciones con DOI, se desarrolló una depuración de datos, excluyendo las publicaciones afectadas para evitar propagar errores. Así mismo, se generaron informes con los tres errores analizados para que el personal técnico pudiera proceder a su resolución.

Se verificaron los siguientes posibles errores:

- DOI duplicados: es decir el mismo DOI hace referencia a dos o más publicaciones diferentes en el portal de producción científica. Lo cual no es coherente por la propia definición y funcionalidad del DOI que se ha descrito.

- DOI mal formados: no cumplen con el formato estandarizado de un DOI, no se ajusta a esta sintaxis: 10.xxxx/xxxxxxxx.
- DOI no activos: identificadores que, pese a tener una sintaxis válida, no conducen a ninguna publicación accesible a través del sistema de resolución DOI. Adicionalmente, se realiza un análisis de todos los DOI utilizando la API de DOI.ORG.

2.2. Segunda fase: Extracción de datos mediante API

En esta segunda fase se extrae la información necesaria de las API de WoS y Scopus y se almacena en una base de datos local para su posterior tratamiento.

Las consultas a estas API pueden realizarse utilizando diversos parámetros. En este estudio, se ejecutaron búsquedas por DOI y por autor, utilizando el ReseracherID y el ScopusID respectivamente. Los metadatos que se recuperaron fueron, principalmente, el DOI de la publicación, los nombres de los autores, sus identificadores en la plataforma y sus ORCID.

En Web of Science, se emplearon solicitudes HTTP con *requests* para extraer información equivalente a través de su API REST. En Scopus, se utilizó la Biblioteca Pybliometrics para interactuar con la API REST de Scopus y recuperar información en formato JSON, incluyendo identificadores de autores (ScopusID) y metadatos de publicaciones (Rose y Kitchin, 2024) (Figura 2).

Figura 2. Extracción de datos

Fuente: Elaboración propia.

2.3. Tercera fase: Obtención del listado de revistas científicas

Para poder hacer una comparativa de revistas indexadas en cada plataforma se requería de un listado completo de las revistas con sus respectivos ISSN para realizar la comparativa de manera precisa. En Scopus, se descargó el listado completo de revistas indexadas proporcionado en su sitio web; en Web of Science, se emplearon técnicas de *web scraping* con selenium y beautifulsoup4 para extraer el listado completo desde su sitio web. Se trataron los datos extraídos, unificándolos y cargándose en la base de datos local SQLite para su posterior tratamiento.

Ambos procesos se ejecutaron en octubre de 2024.

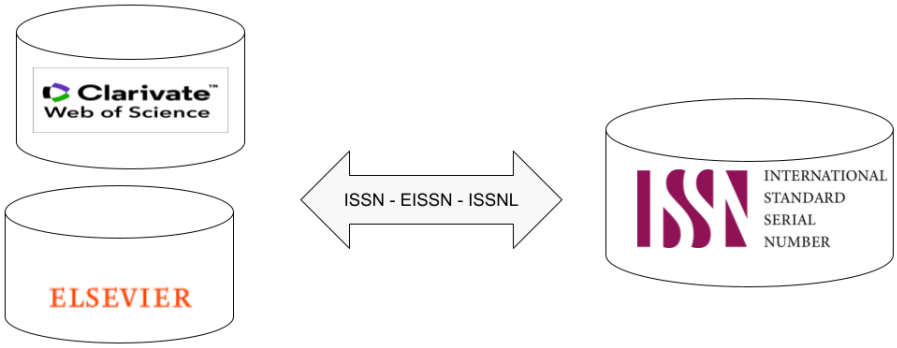
2.4. Cuarta fase: Análisis de identificadores de revistas (ISSN y EISSN)

En esta fase se realizó una depuración de los datos sobre revistas obtenidos en la fase anterior. Primero, se comprobó la existencia de al menos un código ISSN en cada revista. Seguidamente, se comprobó la validez de los códigos ISSN, verificando que el octavo dígito, correspondiente al dígito de control, cumpliera correctamente su función según el algoritmo de verificación Módulo 11, aplicado sobre los siete dígitos numéricos previos. Para este proceso se realizó un script en Python para su procesamiento automático.

En el proceso de comparación de los listados de revistas de ambas plataformas se emplearon como claves principales los identificadores ISSN y EISSN. No obstante, se detectaron inconsistencias en la asignación de estos códigos: en algunos casos, un mismo identificador aparecía como ISSN en una base de datos y como EISSN en la otra. Para subsanar estas discrepancias y asegurar una comparación homogénea, se recurrió al uso del ISSN-L (Linking ISSN), que permite agrupar todas las versiones de una publicación seriada bajo un único identificador, lo que facilitó la normalización de los registros y una comparación fiable entre las dos fuentes. Se obtuvo la tabla completa de estos identificadores solicitándolo a través del portal oficial ISSN.org.

De esta manera, se analizó el ISSN-L de todas las revistas incluidas en Web of Science y Scopus, bajo el supuesto de que dicho identificador debería ser único para cada título, al vincular todas sus variantes de ISSN, tanto impresas como electrónicas. No obstante, se identificaron casos en los que una misma revista presentaba múltiples ISSN-L, lo que contradecía su función como identificador unificador (Figura 3). Ante esta inconsistencia, se adoptó un criterio operativo más flexible: se consideró que dos registros correspondían a la misma revista si coincidían sus ISSN o EISSN o si el ISSN de una base coincidía con el EISSN de la otra.

Figura 3. Identificadores de revistas



Fuente: Elaboración propia.

2.5. Quinta fase: Análisis de identificadores de autores

Se examinaron los identificadores persistentes de autores, como ResearcherID (de Web of Science) y ScopusID de Scopus, realizando una comparación sistemática con los identificadores presentes en el portal de producción científica en cada publicación. El análisis detectó casos de perfiles duplicados y asignación incorrecta de autorías de publicaciones.

Otras incidencias menores que se detectaron fueron las siguientes: error de orden, es decir, el autor no se encuentra en el mismo orden de firma de la publicación en origen, esto es, en el portal institucional y en la plataforma analizada; error de orden excedido, lo cual ocurre cuando en origen el orden de firma supera al número de autores recuperados para una publicación a través de la API.

2.6. Sexta fase: Procesamiento y almacenamiento de datos

Los datos extraídos fueron procesados y almacenados en una base de datos local SQLite, garantizando su portabilidad y facilidad para futuros análisis; también se crearon informes en formato de hojas de cálculo con enlaces a las fuentes originales de cada dato para facilitar la verificación y la corrección manual de las incoherencias detectadas. Estos informes se generaron de manera automatizada utilizando la librería 'xlsxwriter' de Python.

3. RESULTADOS

Una vez aplicada la metodología, se hallaron resultados de investigación de dos tipos: teóricos y de aplicación práctica. Los resultados teóricos consistieron en la descripción de las técnicas de mayor utilidad para la curación de datos de investigación, especialmente en lo relativo al uso de las API de las bases de datos, así como a los métodos de extracción de datos desde la web o *web scraping*, los identificadores persistentes y principales herramientas de referencia (Khder, 2021).

Los resultados prácticos se orientan hacia la aplicación de una estrategia útil para la curación de datos de investigación; su procesamiento, estructuración y reutilización para la obtención de indicadores, y como metodología para la demostración de hipótesis. Se facilitan informes en formato excel a los técnicos responsables de la curación de datos.

3.1. Resultados obtenidos en la recopilación de datos del portal de producción científica de la universidad

En la primera fase, al extraer los datos del portal institucional de las publicaciones con DOI y analizarlos, se obtuvo que el 76,9 % están presentes en WoS y el 87,9 % están presentes en Scopus. Sobre los errores en origen detectados, el 0,7 % son DOI duplicados, es decir, que

el mismo DOI hace referencia a dos publicaciones diferentes dentro del portal; el 0,6 % son DOI no activos, es decir, que no remiten a una publicación disponible, y únicamente un 0,01 % con errores de formato (Tabla 2).

Tabla 1. Publicaciones

DOI	En WoS	En Scopus	Duplicados	No activos	Error formato
50841	38849 (76,9 %)	44402 (87,9 %)	361 (0,7 %)	314 (0,6 %)	6 (0,01 %)

Fuente: Elaboración propia.

Con los datos de los errores detectados se generan informes en formato excel, incluyendo enlaces al portal de producción científica para facilitar su corrección: DOI duplicados en el portal de producción científica, DOI con errores de formato y DOI que no existe el documento al que deberían hacer referencia.

3.2. Resultados obtenidos en el análisis de identificadores de revistas (ISSN y EISSN)

En la tercera fase, el análisis de identificadores de revistas científicas muestra los siguientes datos:

El número de revistas indexadas en WoS es de 21.273.

ISSN o EISSN no valido: 5.

Revistas sin ISSN ni EISSN: 0.

El número de revistas indexadas en Scopus es de 46.535.

ISSN o EISSN no valido: 33.

Revistas sin ISSN ni EISSN: 253.

El número de revistas indexadas por ambas plataformas es de 19.705.

En este estudio, al analizar los datos, se observó que revistas en Scopus y Web of Science tienen ambos identificadores, ISSN y E-ISSN, que deberían converger en el mismo ISSN-L, sin embargo, se encontraron 581 revistas con más de un ISSN-L. Se consultó a ISSN.org cómo deberíamos proceder para intentar resolver esta aparente inconsistencia en los datos. ISSN.org solicitó que se le enviara el listado completo y se les envió por correo electrónico el informe en formato excel con enlaces a la revista en WoS o Scopus y a su portal en febrero de 2025 (Tabla 2).

Tabla 2. Ejemplo de revistas con más de un ISSN-L

NOMBRE	ISSN	ISSN-L	E-ISSN	ISSN-L
ACTA CHIRURGICA BELGICA	0001-5458	0001-5458	2577-0160	1784-3421
Ornithology	0004-8038	0004-8038	2732-4613	2732-4613
BLUMEA	0006-5196	0006-5196	0373-4293	0373-4293
Ornithological Applications	0010-5422	0010-5422	2732-4621	2732-4621
Foreign Trade Review	0015-7325	0015-7325	0971-7625	0971-7625
Geofisica Internacional	0016-7169	0016-7169	2954-436X	2954-436X
Acta Cardiologica	0001-5385	0001-5385	0373-7934	0373-7934
German Journal of Agricultural Economics	0002-1121	0002-1121	2191-4028	2191-4028
Annales de Medecine Veterinaire	0003-4118	0003-4118	1781-3875	1781-3875

Fuente: Elaboración propia.

3.3. Resultados obtenidos en el análisis de identificadores de autores

El análisis de perfiles de autor en WoS evidenció que en el portal de origen estaba informado el ResearcherID del 48.3 % de los autores; se halló que el 37.1 % de los autores, aunque no disponen de este identificador en origen, sí tienen un perfil de autor en WoS.

Un dato relevante en este estudio es que del 30.8 % de los autores se recuperó más de un identificador ReseracherID en WoS.

El 69 % de los autores de la institución tiene algún dato que corregir, esto es, que no tenía informado el identificador en el portal de origen y sí en WoS, o se ha encontrado más de un perfil en WoS para dicho investigador, o se ha detectado alguna de las incidencias menores descritas en la metodología (Tabla 3).

Tabla 3. Perfiles de autor en WoS

Autores	Con RID en origen	Sin RID en origen y sí en WOS	Con más de un RID en WoS	Con algún dato que corregir
4628	2234 (48,3 %)	1719 (37,1 %)	1433 (30,8 %)	3193 (69,0 %)

Fuente: Elaboración propia.

El análisis de perfiles de autor en Scopus evidenció que en el portal de origen estaba informado el ScopusID del 76.8 % de los autores y se halló que el 10,7 % de los autores, aunque no dispone de este identificador en origen, sí tienen un perfil de autor en Scopus.

Un dato relevante en este estudio es que del 6.8 % de los autores se recuperó más de un identificador ScopusID en Scopus.

El 36,1 % de los autores de la institución tiene algún dato que corregir, esto es, que bien no tenía informado el identificador en el portal de origen y sí en Scopus, o se ha encontrado más de un perfil en Scopus para dicho investigador, o se han detectado alguna de las incidencias menores descritas en la metodología (Tabla 4).

Tabla 4. Perfiles de autor en Scopus

Autores	Con ScopusID en origen	SIN Scopus_ID en origen y sí en Scopus	Con más de un Scopus_ID en Scopus	Con algún dato que corregir
4628	3552 (76,8 %)	495 (10,7 %)	315 (6,8 %)	1669 (36,1 %)

Fuente: Elaboración propia.

3.4. Resultados obtenidos en el procesamiento y almacenamiento de datos

Finalmente, en la sexta fase se genera un informe, por cada plataforma analizada, con la información de los perfiles de autor con algún dato que corregir. En estos informes se agrupan en una fila las publicaciones de cada autor, con el mismo resultado de comparación entre los metadatos de origen, es decir, los publicados en el portal de producción científica y los hallados a través de la API en la plataforma analizada. Para facilitar la toma de decisión en cuanto a la corrección de la incidencia se incluyen enlaces a los perfiles de autor y a los documentos en el portal de origen, en ORCID, en WoS y en Scopus.

Figura 4. Informes

Informe de Scopus

ID	Autor	Scopus_ ID	S_Scopus_ID	S_autor	Clasificación origen/API-BD	Publicaciones
a1	autor1	scopusid1			OtrosErrores	1
						5
					ia_ID	3
					ia_ID	12
					e en Scopus	9

Informe de WoS

ID	Autor	RID	W_RID	W_autor	Clasificación origen/API-BD	Publicaciones
a1	autor1	rid1	rid1	autor1	Correcto	3
a2	autor2	rid2	rid3	autor2	Incidencia_ID	6
a3	autor3				No existe en WOS	9
a1	autor1	rid1			OtrosErrores	2
a4	autor4		rid4	autor4	Incidencia_ID	1

Fuente: Elaboración propia.

4. DISCUSIÓN Y CONCLUSIONES

Los resultados de este estudio ponen en evidencia la importancia de la curación de datos como metodología eficaz para identificar inconsistencias y mejorar la calidad de la información científica. Las bases de datos Web of Science y Scopus, que tradicionalmente han sido consideradas referentes en la evaluación de la producción científica, presentan errores en sus registros. Sin embargo, son utilizadas para realizar análisis bibliométricos y rankings institucionales sin cuestionar la calidad y la consistencia de los datos que proporcionan.

En la revisión de las revistas indexadas por ambas bases de datos se observó que 19.705 están presentes en ambas plataformas, es decir, que el 90 % de las revistas indexadas en WoS también lo están en Scopus. Además, Scopus indexa más del doble de revistas científicas (46.535 en Scopus y 21.273 en Wos). Por este motivo, el porcentaje de publicaciones de la universidad en Scopus es mayor (87.9 % frente al 76.9 % en WoS). No obstante, el hecho de que WoS mantenga una proporción tan alta sugiere que, a pesar de su menor cobertura, sigue siendo la principal referencia para muchos investigadores. Esta preferencia puede explicarse por el valor que las publicaciones en WoS tienen en los procesos de evaluación científica gestionados por agencias como ANECA y FECYT, donde se otorgan puntuaciones decisivas para la obtención de acreditaciones, sexenios, financiación de proyectos, reconocimiento académico y evaluación de publicaciones.

Uno de los hallazgos más significativos fue que el 69 % de los autores de la institución analizada presentaban algún dato que corregir. Estas incidencias incluían que no tenían informado el identificador en el portal de origen y sí en WoS, o se había encontrado más de un perfil en WoS para dicho investigador o algunas incidencias menores.

Otro dato que llamó la atención fue la alta tasa de duplicidad de perfiles de autor en WoS: un 30,8 % de los investigadores tenían más de un perfil, frente al 6,8 % en Scopus. Esta diferencia puede deberse tanto a la mayor eficiencia de los algoritmos de detección y generación de autoría en Scopus como al hecho de que la universidad ha invertido más tiempo en depurar los datos provenientes de esta plataforma.

Por otro lado, se evidenció que el identificador de Scopus estaba ampliamente informado en el portal de producción científica de la universidad (77 % de los autores), mientras que solo el 48 % de los investigadores tienen registrado su ResearcherID de WoS. Esta diferencia se explica por el hecho de que Scopus fue la fuente inicial de datos en la construcción del portal institucional, lo que generó una integración más robusta desde el inicio.

Las incidencias halladas en esta investigación evidencian el impacto en la visibilidad y la evaluación de la producción científica de los investigadores, así como la falibilidad de los sistemas que sustentan las políticas de evaluación académica. La existencia de múltiples perfiles, la falta de normalización de los metadatos o la ausencia de identificadores puede ocasionar que una parte significativa del trabajo de los investigadores quede invisibilizada, tanto en portales institucionales como en las bases comerciales; comprometiendo la calidad de los análisis y la presencia en el ecosistema científico. En este sentido, cobra valor la frase de Harari (2024) cuando afirma que «rechazar la fantasía de la infalibilidad, permite construir una red de información que considera que el error es inevitable». Esta perspectiva

refuerza la necesidad de contar con una metodología rigurosa que permita el tratamiento de datos fiables, estructurados y persistentes, estableciendo un puente entre los requisitos institucionales de calidad, la gestión técnica de la información y la toma de decisiones fundamentada y transparente.

La curación de datos permite no solo estructurar grandes volúmenes de información, sino también establecer relaciones significativas entre documentos, autores e instituciones. La metodología empleada en este estudio, actualmente replicada en cinco universidades, ha demostrado ser portable y adaptable a distintos contextos institucionales, lo que permite ampliar el análisis hacia nuevas dimensiones, como el estudio de citas, afiliaciones e índices bibliométricos más verificables. Esto resulta fundamental en el campo de la Documentación, donde trazabilidad, interoperabilidad y precisión de los registros son esenciales para asegurar una gestión del conocimiento sostenible, reutilizable y de calidad en la información consultada.

REFERENCIAS BIBLIOGRÁFICAS

Agencia Nacional de Evaluación de la Calidad y Acreditación. (2024). *Resolución de 20 de marzo de 2024, por la que se aprueban los criterios de evaluación y requisitos mínimos de referencia de los méritos y competencias requeridos para obtener la acreditación a Catedrática o Catedrático de Universidad y a Profesora o Profesor Titular de Universidad*. https://participa.aneca.es/processes/evaluacion-meritos-competencias/f/3/ANECA+7ANECA+7partici

Choi, A. J. y Xin, X., (2021). Data Curation in Practice: Extract Tabular Data from PDF Files Using a Data Analytics Tool. *Journal of eScience Librarianship*, *10*(3), 10. https://doi.org/10.7191/jeslib.2021.1209

Fundación Española para la Ciencia y la Tecnología (FECYT). (2022). *Publons se integra en Web of Science*. Recuperado el 3 de mayo de 2025, de https://www.recursoscientificos.fecyt.es/noticias/publons-se-integra-en-web-science

Fundación Española para la Ciencia y la Tecnología. (2022). *FECYT renueva el Sello de la Calidad Editorial y Científica a 514 revistas españolas*. Recuperado el 3 de mayo de 2025 de https://www.fecyt.es/actualidad/fecyt-renueva-el-sello-de-la-calidad-editorial-y-cientifica-514-revistas-espanolas

Fundación Española para la Ciencia y la Tecnología. (2025). *Guía de evaluación: Novena edición de evaluación de la calidad editorial y científica de las revistas académicas españolas*. https://evaluacionarce.fecyt.es/Publico/Bases/__Recursos/2025_9_convocatoria_GuiaEvaluacion.pdf

García García-Doncel, J. (2023, agosto 7). Curado de datos. Data Science. Recuperado el 13 de febrero de 2025 de https://dat-science.com/curado-de-datos/

ISSN International Standard Serial Number. (s. f.). Consultar la tabla de números ISSN-L. Recuperado el 30 de mayo de 2025, de https://www.issn.org/es/servicios-y-prestaciones/servicios-en-linea/consultar-la-tabla-de-numeros-issn-l/

Khder, M. (2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing and Its Applications*, *13*, 145-168. https://doi.org/10.15849/IJASCA.211128.11

Marín-Arraiza, P. (2022). Madurez de sistemas de identificadores persistentes: Oportunidades en el contexto español. *Anuario ThinkEPI*, *16*. https://doi.org/10.3145/thinkepi.2022.e16a06

Minamiyama, Y., Takeda, H., Hayashi, M., Asaoka, M. y Yamaji, K. (2024). A study on formalizing the knowledge of data curation activities across different fields. *PLOS ONE*, *19*(4), e0301772. https://doi.org/10.1371/journal.pone.0301772PLOS

Parmiggiani, E., Amagyei, Nana Kwame y Kollerud, S. K. S. (2024). Data curation as anticipatory generification in data infrastructure. *European Journal of Information Systems*, *33*(5), 748-767. https://doi.org/10.1080/0960085X.2023.2232333

Peng, G., Downs, R. R. y Wyborn, L. (2022). Global Community Guidelines for Documenting, Sharing, and Reusing Quality Information of Individual Digital Datasets. *Data Science Journal*, *21*(1), 8. https://doi.org/10.5334/dsj-2022-008

Rose, M. E. y Kitchin, J. (2024). pybliometrics: Python-based API-Wrapper to access Scopus. pybliometrics. https://pybliometrics.readthedocs.io/en/stable/

SciELO. (2023). *Guía de curación de datos de investigación para equipos editoriales*. Scielo. https://www.scielo.org/es/sobre-el-scielo/scielo-data-es/

Torres-Salinas, D. y Arroyo-Machado, W. (2022). APIs en contextos bibliométricos: Introducción básica y corpus exhaustivo. *Anuario ThinkEPI*, *16*. https://doi.org/10.3145/thinkepi.2022.e16a09

Vélez-Estévez, A., Pérez, I. J., García-Sánchez, P., Moral-Munoz, J. A. y Cobo, M. J. (2023). New trends in bibliometric APIs: A comparative analysis. *Information Processing & Management*, *60*(4), 103385. https://doi.org/10.1016/j.ipm.2023.103385

Yakel, E. (2007). Digital curation. *OCLC Systems and Services: International digital library perspectives*, *23*(4), 335-340.