UNIVERSIDAD DE MURCIA

Departamento
de Información
y Documentación

Proyecto Esposende

INFORME TÉCNICO

FRANCISCO JAVIER MARTINEZ MENDEZ

ORCID.ORG/0000-0003-1098-9361 || https://webs.um.es/javima

TABLA DE CONTENIDOS

El proyecto: Ciencia Abierta y datos de investigación

El *Real Decreto de acreditación estatal para el acceso a los cuerpos docentes universitarios y el régimen de los concursos de acceso*[1] regula el acceso al cuerpo de Catedráticos y Catedráticas de Universidad, acto administrativo que supone alcanzar el escalón más alto de la carrera docente e investigadora que los candidatos y candidatas han iniciado muchos años atrás. Los legisladores, de forma muy inteligente, introducen la presentación de un "proyecto" que la RAE define, en su tercera acepción, como "designio o pensamiento de ejecutar algo". De esta forma, el desarrollo de estos concursos de acceso no ha de consistir exclusivamente en una justificación de méritos adquiridos, que ya han sido evaluados previamente por una comisión de acreditación de ANECA, sino también de una exposición de una serie de propuestas de mejora continua en la búsqueda de la excelencia y calidad educativa exigibles al profesorado de las instituciones públicas de educación superior.

La *Ley Orgánica de Universidades*[2] introdujo en el seno de la comunidad universitaria la metáfora global de la gestión de la calidad como algo esencial de nuestra actividad. Uno de los cambios introducidos en ese momento, hoy plenamente asentados en nuestra cotidianeidad, son las guías docentes que aprobamos todos los años en nuestros departamentos y que no son exclusivamente, programaciones didácticas porque las mismas representan un acuerdo entre el profesorado y el estudiantado para el cumplimiento de los objetivos educativos del plan de estudio como frutos del proceso de enseñanza-aprendizaje. De la misma manera, cuando un profesor o profesora titular de Universidad participa en un concurso de esta naturaleza, el proyecto que ha de presentar para su aprobación debe plasmar, ineludiblemente, qué compromisos se asumirán para la mejora de la excelencia y calidad del departamento facultad y universidad correspondiente.

La *Ley Orgánica del Sistema Universitario*[3], en su artículo primero entiende por universidades "aquellas instituciones, públicas o privadas, que desarrollan las funciones centrales de docencia, investigación y transferencia e intercambio del conocimiento", por tanto, entendemos que todo proyecto que se presente en un concurso de esta naturaleza debe tener estos tres perfiles de aplicación, ya que se aspira a aplicarlo en el seno de una universidad. Este proyecto, además, no puede representar una ruptura con el trabajo previo desarrollado, sino más bien una consolidación y una expansión del mismo fruto de la experiencia adquirido y el estudio desarrollado.

El ámbito de aplicación de este proyecto es la Ciencia Abierta (en general) y la gestión de los conjuntos de datos de investigación (en particular), disciplina y materia científicas encuadradas dentro de nuestra área de conocimiento y en la que ya vienen trabajando activamente distintos equipos docentes y de investigación de otras universidades con tremendo éxito y repercusión.

---

[1] BOE núm. 213, de 6 de septiembre de 2023. https://www.boe.es/eli/es/rd/2023/07/18/678

[2] BOE núm. 307, de 24 de diciembre de 2001. https://www.boe.es/buscar/pdf/2001/BOE-A-2001-24515-consolidado.pdf

[3] BOE núm. 70, de 23 de marzo de 2023. https://www.boe.es/eli/es/lo/2023/03/22/2/con

La Ciencia Abierta representa un nuevo paradigma que incorpora una visión holística del proceso de generación de conocimiento, a partir del diseño inicial de un proyecto y su desarrollo hasta la comunicación, difusión y preservación de sus resultados (Abadal et al, 2023).

El movimiento de la Ciencia Abierta surge de la comunidad científica y se extiende de un país a otro, instando a que se abran las puertas del conocimiento. UNESCO (2021) la define como "un constructo inclusivo que combina diversos movimientos y prácticas con el fin de que los conocimientos científicos multilingües estén abiertamente disponibles y sean accesibles para todos, así como reutilizables por todos, se incrementen las colaboraciones científicas y el intercambio de información en beneficio de la ciencia y la sociedad, y se abran los procesos de creación, evaluación y comunicación de los conocimientos científicos a los agentes sociales más allá de la comunidad científica tradicional". Así entendida, la Ciencia Abierta abarca todas las disciplinas científicas incluyendo las ciencias básicas y aplicadas, las ciencias naturales y sociales y las humanidades. Este nuevo ecosistema de la comunicación de la ciencia se basa en los siguientes pilares clave: conocimiento científico abierto, infraestructuras de la Ciencia Abierta, comunicación científica, participación abierta de los agentes sociales y diálogo abierto con otros sistemas de conocimiento (UNESCO, 2021).

Se requiere un conocimiento más sólido a escala mundial de las oportunidades que ofrece la Ciencia Abierta para hacer frente a retos mayúsculos. El mejor ejemplo que confirma esta afirmación ha sido la lucha de los científicos en la búsqueda de vacunas contra la pandemia COVID-19 (Ferrer Sapena et al. 202; López Carreño y Martínez Méndez, 2020 y Torres Salinas y Robinson García,2020). En el ámbito europeo, la apuesta decidida por el **depósito y publicación en abierto de los resultados de investigación** financiados con cargo a los presupuestos comunitarios, era un mandato ya recogido en el Programa «Horizonte 2020», donde, en la aceptación de la financiación se incluía la obligación de garantizar el acceso abierto a todas las publicaciones científicas derivadas de la investigación (EC, 2014) y que se mantiene en el Programa Marco de Investigación e Innovación «Horizonte Europa», donde los principios y prácticas de la Ciencia Abierta estarán integrados en todo el programa (EC, 2012).

La *Estrategia Nacional de Ciencia Abierta* (ENCA, 2023), presentada el pasado mes de junio en unas jornadas organizadas por CRUE y FECYT en la Universidad de Vigo, la define como el acceso abierto a los resultados de investigación (publicaciones, datos, protocolos, código, metodologías, software, etc.), la utilización de plataformas digitales basadas en código abierto y la apertura de todo el proceso científico (incluyendo prácticas como la revisión por pares en abierto, los recursos educativos en abierto, el fomento de la ciencia ciudadana y el desarrollo de nuevas formas de medir el rendimiento investigador). Esta estrategia define seis grandes áreas y cuatro objetivos estratégicos.

Imagen 1: Áreas principales de la Estrategia Nacional de Ciencia Abierta. Fuente: ENCA (2023).

Los cuatro objetivos principales de esta estrategia son:

1. Garantizar la existencia de infraestructuras digitales interoperables suficientemente robustas y bien articuladas como para absorber el impacto de la implementación de una política nacional de Ciencia Abierta y facilitar su integración en el ecosistema internacional y su integración, cuando proceda, en la EOSC[4].
2. Fomentar la adecuada gestión de los datos de investigación generados por el sistema nacional de I+D+i a través de los principios FAIR [5] para mejorar su localización, accesibilidad, interoperabilidad y reusabilidad.
3. Implementar el acceso abierto y gratuito por defecto a las publicaciones y resultados científicos financiados de forma directa o indirecta con fondos públicos, para toda la ciudadanía.
4. Establecer nuevos mecanismos de evaluación de la investigación y un sistema de incentivos y reconocimientos dirigidos a **impulsar las prácticas de Ciencia Abierta**, **así como capacitar a todo el personal** (investigador, gestor, financiador, evaluador) para alinear su desempeño profesional con los principios de Ciencia Abierta.

[4] EOSC son las siglas de 'European Open Science Cloud', la infraestructura de soporte a la Ciencia Abierta que están poniendo en marcha la Unión Europea. Más información en https://eosc-portal.eu/.
5 FAIR son las siglas de '(Findable, Accesible, Interoperable, Reusable', los principios básicos para hacer que los datos de investigación sean fáciles de encontrar, accesibles, interoperables y reutilizables.

Propuesta de investigación: el prototipo ‘esposende 1.0’

Esta propuesta consiste en el desarrollo de un **prototipo de evaluador semiautomático** del cumplimiento de las buenas prácticas de gestión de datos en la web del W3C en los conjuntos de datos de investigación. La villa de **Esposende**[6] es un municipio del distrito de Braga, sede de la *Universidade do Minho* donde he realizado una estancia de investigación desde enero a junio de 2023. Precisamente en la biblioteca pública de esta localidad costera es donde comenzaron a redactarse las primeras líneas de este proyecto investigador.

Los datos de investigación y la Ciencia Abierta

En el transcurso de muchas investigaciones se generan datos que sirven para la obtención de resultados y la extracción de conclusiones. Dentro del campo de la Ciencia Abierta (UNESCO, 2021) ha cobrado importancia publicar, junto al artículo o informe derivado de la investigación, el conjunto de datos empleados o generados en la misma con vistas a favorecer su reutilización por parte de otros investigadores, fomentando así la transparencia de la labor investigadora y garantizando la integridad de todo el proceso (Borghi & Van Gulick, 2019).

En el ecosistema de la Ciencia Abierta es especialmente importante la gestión de los datos de investigación: “proceso diseñado para gestionar y difundir conjuntos de datos de alta calidad, que cumplan con los requisitos académicos, legales y éticos establecidos” (Alonso-Arévalo, 2019). La propia ENCA le asigna un área de actuación y le dedica el segundo de sus objetivos estratégicos porque buena parte del éxito de los esfuerzos que se están dedicando para el fomento de la Ciencia Abierta depende de una buena gestión de estos datos, dispuestos habitualmente dentro de un conjunto de datos o ‘dataset’. Disponer de un adecuado acceso a los datos de investigación es imprescindible para la reproducibilidad de los resultados científicos, facilita la cooperación interdisciplinar, estimula el crecimiento económico a través de mejores oportunidades para la innovación, permite la reutilización de datos, aumenta la eficiencia de los recursos, mejora la transparencia, la rendición de cuentas y la confianza en los resultados de la investigación científica (OCDE, 2021). Todas estas características convergen en el potencial de estos conjuntos como aceleradores de la investigación tanto en entornos académicos como en colaboración con empresas privadas (Alexandre Beanvent et al., 2019). Esta gestión se lleva a cabo sobre los conjuntos de datos que recogen información estructurada y organizada recopilada o generada en el desarrollo de una investigación. Estos conjuntos contienen información relevante y detallada que se utiliza para respaldar los objetivos, análisis y resultados del estudio científico. Los datos pueden ser primarios y secundarios, cuantitativos o cualitativos, y pueden estar disponibles en diferentes formatos y tamaños. Es totalmente recomendable que los datos cumplan los principios FAIR que establecen pautas para la correcta gestión de datos y que fueron introducidos por Wilkinson et al. (2016).

Si bien la responsabilidad de la elaboración de estos conjuntos de datos recae sobre los investigadores, no es menos cierto que en ella participan, en mayor o menor medida, los profesionales de las bibliotecas y centros de documentación responsables de la gestión de los repositorios donde se almacenan los resultados de la investigación. Por tanto, esta gestión se

[6] Más información sobre esta localidad en https://www.visitesposende.com/es

halla entre las principales líneas de trabajo, no ya futuras sino actuales, de las bibliotecas académicas (Federer & Qin, 2019), si bien todavía no está implantada de manera uniforme (Ayris & Ignat, 2018; Ashiq, 2022). Esta actividad representa "un proceso diseñado para gestionar y difundir conjuntos de datos de alta calidad, que cumplan con los requisitos académicos, legales y éticos establecidos" (Alonso-Arévalo, 2019). Estamos convencidos de que los investigadores van a necesitar apoyo para adaptarse a los nuevos requisitos específicos de esta gestión (Marín-Arraiza et al., 2019). Por tanto, aparte de un reto considerable, las bibliotecas académicas tienen una nueva oportunidad de mejorar sus actuales servicios de apoyo a la investigación y aumentar así su presencia y relevancia en el seno de sus instituciones (Angelozzi, S. M., 2020; Sheikh et al., 2023). Ante lo ingente de esta tarea, en dos comunidades autónoma de España (Cataluña y Madrid) se han constituido consorcios regionales para "acometer de forma colectiva los retos derivados de la Ciencia Abierta y procurar que su adopción se realice con el menor esfuerzo por parte de las universidades" (Alcalá y Anglada, 2019). En la búsqueda de esa simplicidad, las políticas de acceso abierto de i+D+i que se están implantando deben indicar de manera explícita los procedimientos, estándares, formatos, licencias y lenguajes a adoptar para una normalización e identificación de los datos de investigación (EC, 2023). También es importante la revisión de los propios contenedores de los conjuntos de datos con el objeto de asegurar que se encuentren en las debidas condiciones de ofrecer una respuesta útil en su identificación, descripción, catalogación clasificación y métricas de uso.

En las revistas científicas, no termina de alcanzarse la integración de este tipo de contenidos y apenas tienen cabida los conjuntos de datos. Esto ocurre a pesar de las iniciativas de grandes editoriales en diseñar buscadores especializados en repositorios de datos, como 'Data Citation Index' de *Web of Science Group*, 'Mendeley Data' de *Elsevier* y 'Google Data Search'. Este interés evidencia la dimensión que puede llegar a alcanzar el volumen de los conjuntos de datos de investigación generados (preceptiva o voluntariamente) y muestra claramente la necesidad de su localización y curación, como es el caso del proyecto 'Data Curation Network' (Johnston et al., 2017; Johnston et al., 2018).

Los conjuntos de datos de investigación son una fuente de información en continuo crecimiento. Alcanzar un efectivo desarrollo en su gestión es uno de los retos a los que se enfrenta el movimiento hacia la Ciencia Abierta (UNESCO, 2021; Bethencourt-Aguilar, 2022). El momento actual parece el más adecuado para reconsiderar el proceso de creación/mantenimiento, ayudando a establecer líneas y estrategias de gestión y depósito institucional que formen parte del signo distintivo de calidad científica propio de las instituciones de investigación. En España, y en todo el contexto europeo, la mayor parte de la investigación se lleva a cabo en universidades y centros públicos de i+d+i (Fundación CYD, 2023). Corresponde a estas instituciones implementar los repositorios de conjuntos de datos aplicando las mejores prácticas posibles, sin olvidar la tarea de hacer partícipe a sus comunidades investigadoras de los beneficios de disponer en abierto los resultados y los datos de investigación, especialmente por la transparencia (De Giusti, 2020). En el momento presente, este punto en particular cobra especial relevancia debido a la proliferación de malas prácticas entre la comunidad científica que se han descubierto en los últimos meses, especialmente las vinculadas con una

hiperproducción científica incompatible con la calidad y la excelencia que se le supone a la investigación.

Prueba de la importancia que han adquirido los conjuntos de datos de investigación es su presencia como mérito para la obtención de un sexenio de investigación en la reciente convocatoria realizada por la ANECA en diciembre de 2023[7]. Este documento resalta la necesidad de adaptar los criterios tradicionalmente empleados a la situación general de la ciencia en España y, en consecuencia, estimar que se consideren méritos de la actividad investigadora "los conjuntos de datos, las metodologías y el código de las aplicaciones informáticas desarrolladas" dentro del reajuste en la combinación de los métodos cualitativos y los indicadores cuantitativos utilizados para la valoración de las aportaciones que está llevando a cabo la agencia evaluadora en aplicación de su adhesión a la DORA y CoARA.

Los principios FAIR

Como se ha indicado anteriormente, estos principios fueron introducidos por Wilkinson et al. (2016) en un artículo publicado en la revista *Scientific Data* (del grupo *Nature*). Consisten en una serie de pautas que aspiran a establecer un conjunto de buenas prácticas para asegurar que los datos científicos sean "encontrables, accesibles, interoperables y reutilizables" (el acrónimo FAIR está formado por las iniciales en inglés de las palabras '**Findable**, **Accessible**, **Interoperable** and **Reusable'**). Los principios se han desarrollado para abordar los problemas que se presentan al intentar utilizar los datos científicos en el contexto actual de la investigación, ya que muchas veces estos datos están almacenados en diferentes formatos, sistemas y lugares, lo que dificulta enormemente su uso y reutilización. La implementación de estos principios busca eliminar estas barreras y hacer que los datos sean más accesibles y útiles.

F	Es la capacidad de los datos de ser localizados y accedidos con facilidad. Esto se logra a través de la asignación de identificadores únicos y persistentes (como los dois) para identificar de manera inequívoca a los datos. Además, se deben proporcionar metadatos descriptivos claros y precisos que permitan encontrar los datos en un catálogo o motor de búsqueda.
A	Es la capacidad de los datos de ser descargados y utilizados por cualquier persona. Esto se logra mediante la eliminación de barreras de acceso, como restricciones de acceso o de formato. Los datos deben estar disponibles en un formato estándar y legible por máquina, y se deben proporcionar herramientas para la descarga y acceso a los datos.
I	Este principio se refiere a la capacidad de los datos de ser utilizados en conjunto con otros datos. Esto se logra mediante la utilización de estándares comunes y la adopción de un conjunto común de vocabularios y ontologías que permitan una interpretación uniforme de los datos.
R	Es la capacidad de los datos de ser utilizados para diferentes fines. Esto se logra mediante la publicación de datos bajo licencias abiertas y claras que permitan su uso y reutilización. Además, los datos deben ser estructurados de manera tal que permitan su uso en diferentes contextos y aplicaciones.

Tabla I: Resumen de los principios FAIR.

[7] Resolución de 19 de diciembre de 2023, de la Secretaría General de Universidades, por la que se aprueba la convocatoria de evaluación de la actividad investigadora (BOE de 22 de diciembre de 2023): https://www.boe.es/boe/dias/2023/12/22/pdfs/BOE-A-2023-26094.pdf

El cumplimiento de estos principios es beneficioso para toda la comunidad científica porque favorecen la reutilización y la combinación de datos de diferentes fuentes, lo que puede llevar a nuevos descubrimientos e innovaciones. Su implementación está siendo impulsada por varias organizaciones y gobiernos en todo el mundo. La Comisión Europea, por ejemplo, promueve la Estrategia Europea de Datos Abiertos, con el objetivo de establecer un ecosistema de datos abierto y FAIR en Europa, convirtiéndola en "un modelo de sociedad capacitada por los datos" (2020).

Para cumplir con los principios FAIR, se han desarrollado diversas recomendaciones específicas para el desarrollo y adquisición de buenas prácticas de esta gestión, tales como las elaboradas por la RDA[8] en el documento '*FAIR Data Maturity Model. Specification and Guidelines*' (2020) que ofrecen un conjunto de directrices y una lista de comprobación para satisfacer el cumplimento de estos principios. También están las "recetas" que recogen Rocca-Serra el al. en su proyecto 'FAIR Cookbook' (2023): instrucciones detalladas y guías sobre cómo llevar a cabo tareas específicas relacionadas con la gestión de datos de acuerdo con los principios FAIR. También se encuentra relacionada la comunidad '*The Filesharing Team*' (2019) que, si bien no proporciona una lista explícita y prescriptiva de "buenas prácticas", su contenido y estructura promueven y apoyan buenas prácticas al guiar a los usuarios hacia herramientas, estándares y políticas que ayudarán a mejorar la gestión y compartición de datos, ayudando a los investigadores a navegar y adoptar buenas prácticas en la gestión de datos, alineadas con los principios FAIR y las expectativas de la comunidad científica global. Dentro de la iniciativa europea FAIRSFAIR encontramos '*CoreTrustSeal+FAIRenabling Capability Maturity Model*' binomio formado por dos aspectos relacionados con la certificación y evaluación de repositorios de datos y su capacidad para apoyar los principios FAIR.

- '*CoreTrustSeal*' es un certificado otorgado a repositorios de datos que cumplen con ciertos estándares y requisitos establecidos, indicando que un repositorio de datos es confiable y sigue las mejores prácticas en términos de acceso, preservación y seguridad de los datos que almacena.
- '*FAIR-enabling Capability Maturity Model*' es una herramienta o marco de evaluación que se usa para medir el grado en que un repositorio o conjunto de datos es compatible con los principios FAIR (L'Hours et al., 2022).

En España, REBIUN [9] elaboró su propio documento de recomendaciones (2017). Los requerimientos técnicos, tecnológicos, así como la implicación bibliotecaria y de los responsables de los repositorios, son una pieza fundamental para lograr una óptima implantación y, por tanto, para la visibilidad y reutilización de dichos datos en futuras

[8] RDA es el acrónimo de 'Research Data Alliance'. Organización creada en 2013 por la Comisión Europea, la 'National Science Foundation' y el 'National Institute of Standards and Technology' de EE.UU. y el Departamento de Innovación del gobierno de Australia con el objetivo de construir la infraestructura social y técnica que permitiera compartir y reutilizar los datos de forma abierta. Más información en https://www.rd-alliance.org/about-rda

[9] REBIUN es el acrónimo de la Red de Bibliotecas Universitarias y Científicas Españolas (REBIUN), red asociada a la sectorial de investigación de la CRUE (Conferencia de Rectores y Universidades Españolas). Más información en https://www.rebiun.org

investigaciones. A un nivel algo más genérico, no hay que olvidar que se están gestionando datos en la web por lo que también es interesante tener en cuenta las recomendaciones de buenas prácticas para la publicación de datos elaboradas por el W3C[10] (Lòscio et al., 2017).

A un nivel algo más específico, pero relacionado con conjuntos de datos de instituciones documentales, encontramos el trabajo de Koster y Woutersen-Windhouwer (2018) donde los autores proponen un conjunto de directrices y buenas prácticas para facilitar el proceso de reutilización de colecciones de bibliotecas, archivos y museos. Estas directrices se basan en los principios FAIR teniendo en cuenta otras iniciativas con la misma finalidad. Las recomendaciones se centran en tres niveles: objetos, metadatos y registros de metadatos. Estas directrices vienen acompañadas de aclaraciones y ejemplos, así como recomendaciones para la evaluación de las situaciones actuales y la aplicación de los principios.

Garijo y Poveda-Villalón (2020) describen directrices de aplicación y recomendaciones para conseguir que las ontologías sean localizables (mediante registros de metadatos y anotaciones); accesibles (mediante buenas prácticas en el diseño de URI y la negociación de contenidos), interoperables (mostrando cómo servir ontologías en diferentes serializaciones estándar) y reutilizables (describiendo los metadatos y las directrices de diagramas necesarias para su correcta comprensión) en la web, todo ello en formato de datos abiertos enlazados. Muestran también cómo llevar a cabo las recomendaciones con una ontología de ejemplo y punteros a herramientas de la Web Semántica.

Los conjuntos de datos y su importancia en la Ciencia Abierta

Para UNESCO (2020), la idea básica de la Ciencia Abierta es "permitir que la información, los datos y los resultados científicos sean más accesibles (acceso abierto) y se utilicen de manera más fiable (datos abiertos), con la participación activa de todas las partes interesadas (apertura a la sociedad). El movimiento de la Ciencia Abierta ha surgido de la comunidad científica y se ha extendido rápidamente de un país a otro, instando a que se abran las puertas del conocimiento". En este contexto, los conjuntos de datos son colecciones de datos que se utilizan en la investigación científica y se comparten públicamente para su acceso, uso y reutilización (y también, en cierto modo, transparencia).

Estos conjuntos de datos pueden contener información de diferentes fuentes, como experimentos, encuestas, observaciones, entre otros, y pueden ser utilizados para diferentes propósitos, como la validación de resultados, la replicación de experimentos, la exploración de nuevas hipótesis y la generación de nuevas ideas y descubrimientos. Su importancia reside en su capacidad para facilitar la transparencia, la colaboración y el avance del conocimiento científico. Al compartir públicamente los datos, los investigadores pueden validar sus resultados, permitir que otros reproduzcan sus experimentos y ampliar el número de colaboradores. Los conjuntos de datos están destinados a ser utilizados por otros investigadores para la generación de nuevas ideas y descubrimientos. La reutilización de los datos permite a los investigadores

[10] W3C son las siglas de 'World Wide Web Consortium', la organización internacional que genera recomendaciones y estándares que aseguran el crecimiento de la web a largo plazo. Este consorcio fue creado en octubre de 1994 por **Tim Berners-Lee**, el creador de la web original. Más información en https://www.w3.org

explorar nuevas líneas de investigación y responder preguntas que no se habían considerado anteriormente. Sin duda alguna, es su principal propósito.

Otro beneficio es permitir una mayor eficiencia en la investigación científica. Al compartir los datos, se evita la duplicación de esfuerzos y se pueden realizar análisis más profundos y precisos. Esto también puede llevar a una mayor productividad científica, ya que se pueden utilizar los datos existentes para responder preguntas importantes sin necesidad de recopilar nuevos datos. También pueden ser utilizados para abordar desafíos globales en áreas como la salud, el medio ambiente y la energía. Los datos recopilados a partir de diferentes fuentes y lugares pueden ser utilizados para entender mejor los problemas globales y para desarrollar soluciones más efectivas. La reciente pandemia COVID-19 ha proporcionado muchos ejemplos de ello[11] [12].

La creación y el uso de conjuntos de datos en la Ciencia Abierta involucra una serie de consideraciones éticas, legales y técnicas que resultan fundamentales para garantizar la integridad de la investigación, proteger los derechos de los participantes y fomentar la transparencia y el acceso abierto a los datos científicos. Por ejemplo, es importante asegurarse de que se han obtenido los permisos necesarios para la recopilación y el uso de los datos, y que se han aplicado las medidas necesarias para proteger la privacidad de los participantes. También es importante utilizar formatos de datos abiertos y estándares comunes que permitan la interoperabilidad y la reutilización de los datos. En este sentido, es importante que los investigadores y el personal técnico de las bibliotecas estén familiarizados con las mejores prácticas en la gestión y publicación de datos abiertos: en particular, estos principios. Otro factor importante en la creación y uso de conjuntos de datos es la necesidad de asegurar que sean de alta calidad y hayan pasado los controles necesarios que confirmen su precisión y confiabilidad. Esto incluye validar los datos, eliminar datos erróneos o duplicados, normalización de los datos y disponer de una documentación suficiente de los procesos de recopilación y análisis.

Las buenas prácticas de gestión de datos en la web.

El World Wide Consortium (W3C) publicó en el año 2017 el documento '*Data on the Web Best Practices: W3C Recommendation'.* Es una guía detallada para el diseño, publicación y uso de datos enlazados en la web, con el objeto de promover su accesibilidad, interoperabilidad y

[11] Basta recordar el 'dataset' de la *Johns Hopkins University* sobre la COVID-19: una colección de datos públicos que rastreaba la propagación del virus a nivel global, incluyendo información sobre el número de casos confirmados, muertes y recuperaciones, así como la ubicación geográfica de los casos. Este 'dataset' se convirtió en una de las fuentes de datos más confiables y ampliamente utilizadas durante la pandemia y permitió a los investigadores y responsables de políticas comprender mejor la propagación del virus y tomar decisiones informadas sobre la prevención y el control de la enfermedad. Los datos recopilados en el 'dataset' permitieron la creación de herramientas y visualizaciones interactivas para informar al público sobre la pandemia, incluyendo el popular "dashboard" de la universidad que mostraba las estadísticas de COVID-19 en tiempo real.

[12] Otro ejemplo es el 'dataset' **CORD-19:** una colección de datos de una amplia gama de informaciones sobre la enfermedad, incluyendo investigaciones científicas, noticias, informes gubernamentales y otra información relevante. Fue creado por un grupo de colaboradores de la comunidad científica liderados por el *Allen Institute for AI*. La colección de datos está disponible públicamente y se ha utilizado ampliamente para el desarrollo de tratamientos y vacunas para COVID-19.

reutilización (Teixeira dos Santos, 2023), proporcionando orientación a los editores de datos en línea sobre cómo representarlos y compartirlos en un formato estándar y accesible. Las prácticas se han desarrollado para fomentar y permitir la expansión continua de la web como medio para el intercambio de datos. El documento menciona el crecimiento en la publicación de datos abiertos por parte de los gobiernos en todo el mundo, la publicación en línea de los datos de investigación, la recolección y análisis de datos de redes sociales, la presencia de importantes colecciones de patrimonio cultural y, en general, el crecimiento sostenido de los datos abiertos en la nube, destacando la necesidad de una comprensión común entre editores y consumidores de datos, junto con la necesidad de mejorar la consistencia en el manejo de los datos.

Estas **buenas prácticas** (DWBP a partir de ahora) cubren diferentes aspectos relacionados con la publicación y el consumo de datos, como son los formatos, el acceso, los identificadores y la gestión de los metadatos. Con el fin de delimitar el alcance y obtener las características necesarias para implementarlas, se recopilaron casos de uso que representan escenarios de cómo se publican habitualmente estos datos y cómo se utilizan. El conjunto de requisitos derivados de esta recopilación se utilizó para guiar el desarrollo de las DWBP, independientes del dominio y la aplicación. Estas recomendaciones pueden ampliarse o complementarse con otros documentos de similar naturaleza. Si bien las DWBP recomiendan usar datos enlazados, también promueven el empleo de otros formatos abiertos como son CSV o json, maximizando más si cabe el potencial de este contexto para establecer vínculos.

CATEGORÍA	BUENA PRÁCTICA
Metadatos Requisito fundamental. Los datos no podrán ser descubiertos o reutilizados por nadie más que el editor si no se proporcionan metadatos suficientes.	**BP** 1: Proporcionar metadatos **BP** 2: Proporcionar metadatos descriptivos **BP** 3: Proporcionar metadatos estructurales
Licencias Según el tipo de licencia adoptada por el editor, puede haber más o menos restricciones a la hora de compartir y reutilizar los datos.	**BP** 4: Proporcionar información sobre la licencia de los datos
Procedencia El reto de publicar datos en la web es proporcionar un nivel adecuado de detalle sobre su origen.	**BP** 5: Proporcionar información sobre la procedencia de los datos
Calidad Puede tener un gran impacto en la calidad de las aplicaciones que utilizan un conjunto de datos.	**BP** 6: Proporcionar información sobre la calidad de los datos
Versiones Los conjuntos de datos pueden cambiar con el tiempo. Algunos tienen previsto ese cambio y otros se modifican a medida que las mejoras en la recogida de datos hacen que merezca la pena actualizarlos.	**BP** 7: Proporcionar un indicador de versión **BP** 8: Proporcionar el historial de versiones
Identificadores El descubrimiento, uso y citación de datos en la web depende fundamentalmente del uso de URI HTTP (o HTTPS): identificadores únicos globales.	**BP** 9: Utilizar URIs persistentes como identificadores de conjuntos de datos **BP** 10: Utilizar URIs persistentes como identificadores dentro de conjuntos de datos **BP** 11: Asignar URIs a versiones y series de conjuntos de datos
Formatos El mejor y más flexible mecanismo de acceso del	**BP** 12: Utilizar formatos de datos estandarizados legibles por máquina

CATEGORÍA	BUENA PRÁCTICA
mundo carece de sentido si no se sirven los datos en formatos que permitan su uso y reutilización.	**BP** 13: Utilizar representaciones de datos neutras respecto a la localización **BP** 14: Proporcionar datos en múltiples formatos
Vocabularios Se utiliza para clasificar los términos que pueden utilizarse en una aplicación concreta, caracterizar las posibles relaciones y definir las posibles restricciones en su uso.	**BP** 15: Reutilizar vocabularios, preferentemente estandarizados **BP** 16: Elegir el nivel adecuado de formalización
Acceso a los datos Facilitar el acceso a los datos permite tanto a las personas como a las máquinas aprovechar las ventajas de compartir datos utilizando la infraestructura de la red.	**BP** 17: Proporcionar descarga masiva **BP** 18: Proporcionar subconjuntos para conjuntos de datos grandes **BP** 19: Utilizar negociación de contenidos para servir datos disponibles en múltiples formatos **BP** 20: Proporcionar acceso en tiempo real **BP** 21: Proporcionar datos actualizados **BP** 22: Proporcionar una explicación para datos que no están disponibles **BP** 23: Hacer datos disponibles a través de una API **BP** 24: Utilizar estándares web como base de las APIs **BP** 25: Proporcionar documentación completa para su API **BP** 26: Evitar cambios que rompan su API
Preservación Las medidas deben tomar los editores para indicar que los datos se han eliminado o archivado.	**BP** 27: Preservar identificadores **BP** 28: Evaluar la cobertura del conjunto de datos
Retroalimentación ('feedback') Ayuda a los editores en la mejora de la integridad de los datos, además de fomentar la publicación de nuevos datos. Permite a los consumidores de datos tener voz describiendo experiencias de uso.	**BP** 29: Recopilar comentarios de los consumidores de datos **BP** 30: Hacer comentarios disponibles
Enriquecimiento Procesos que pueden utilizarse para mejorar, perfeccionar los datos brutos o previamente procesados. Esta idea y otros conceptos similares contribuyen a hacer de los datos un activo valioso para casi cualquier negocio o empresa moderna.	**BP** 31: Enriquecer datos generando nuevos datos **BP** 32: Proporcionar presentaciones complementarias
Republicación Combinar datos existentes con otros conjuntos de datos, crear aplicaciones web o visualizaciones, o reempaquetar los datos en una nueva forma.	**BP** 33: Proporcionar comentarios al editor original **BP** 34: Seguir los términos de la licencia **BP** 35: Citar la publicación original

Tabla II: Lista de buenas prácticas en la gestión de datos en la web del W3C.
Fuente: https://www.w3.org/TR/dwbp/#bestPractices

RDA FAIR data maturity model

Los principios FAIR datan del año 2016. Como todas las normas genéricas, dan lugar a distintas interpretaciones en su aplicación. Para remediar la proliferación de medidas del cumplimiento de los principios FAIR ('FAIRness' en inglés), la 'Research Data Alliance' creó un grupo de trabajo para desarrollar un modelo de madurez FAIR en la implementación de los conjuntos de datos (2020). Consiste en una serie criterios básicos de evaluación que establece indicadores y niveles de madurez asociados. Se produjo un primer conjunto de directrices y una lista de verificación relacionada con la implementación de los indicadores, alineando así las directrices para evaluar el nivel de cumplimiento FAIR con las necesidades de la comunidad. Los indicadores se derivan de los principios FAIR y pretenden formular aspectos mensurables de cada principio que puedan ser utilizados por los enfoques de evaluación.

Los principios FAIR se toman tal cual; es decir, los indicadores no amplían o modifican los principios, sólo cubren aspectos que se mencionan en los propios principios o en aclaraciones adicionales. El planteamiento fue crear un indicador para cada aspecto distinguible en la descripción del principio; por ejemplo, cuando se habla de un identificador persistente y globalmente único, se definen dos, uno para evaluar la persistencia y otro para evaluar la unicidad global. También se definen indicadores distintos para los metadatos y para los datos, siempre que un principio se refiera a "(meta)datos" y la evaluación del aspecto para los metadatos sea distinta de la evaluación para los datos.

Principio FAIR	Indicador	Propósito	Naturaleza
F1	RDA-F1-01M	Los metadatos se identifican mediante un identificador persistente	Esencial
	RDA-F1-01D	Los datos se identifican mediante un identificador persistente	Esencial
	RDA-F1-02M	Los metadatos se identifican mediante un identificador único global	Esencial
	RDA-F1-02D	Los datos se identifican mediante un identificador único global	Esencial
F2	RDA-F2-01M	Se proporcionan metadatos enriquecidos para permitir la localización	Esencial
F3	RDA-F3-01M	Los metadatos incluyen el identificador de los datos	Esencial
F4	RDA-F4-01M	Los metadatos se presentan de forma que puedan ser recolectados e indexados.	Esencial
A1	RDA-A1-01M	Los metadatos contienen información que permite al usuario acceder a los datos.	Importante
	RDA-A1-02M	Los metadatos pueden ser accedidos manualmente (por ejemplo, con intervención humana).	Esencial
	RDA-A1-02D	Los datos pueden ser accedidos manualmente (por ejemplo, con intervención humana).	Esencial
	RDA-A1-03M	El identificador de los metadatos resuelve un registro de metadatos.	Esencial
	RDA-A1-03D	El identificador de los datos resuelve un objeto digital.	Esencial
	RDA-A1-04M	Se accede a los metadatos a través de un protocolo estandarizado.	Esencial
	RDA-A1-04D	Se accede a los datos a través de un protocolo estandarizado.	Esencial
	RDA-A1-05D	Los datos pueden ser accedidos de forma automática (por ejemplo, por medio de un programa de ordenador).	Importante
A1.1	RDA-A1.1-01M	Los metadatos son accesibles a través de un protocolo de acceso libre.	Esencial
	RDA-A1.1-01D	Los datos son accesibles a través de un protocolo de acceso libre.	Importante
A1.2	RDA-A1.2-01D	Los datos son accesibles por medio de un protocolo de acceso que soporta autenticación y autorización.	Útil
A2	RDA-A2-01M	Se garantiza que los metadatos seguirán disponibles después de que los datos dejen de estarlo.	Esencial

Principio FAIR	Indicador	Propósito	Naturaleza
I1	RDA-I1-01M	Los metadatos usan representación del conocimiento expresada en formatos estandarizados.	Importante
	RDA-I1-01D	Los datos usan representación del conocimiento expresada en formatos estandarizados.	Importante
	RDA-I1-02M	Los metadatos utilizan una representación del conocimiento comprensible para las máquinas	Importante
	RDA-I1-02D	Los datos utilizan una representación del conocimiento comprensible para las máquinas	Importante
I2	RDA-I2-01M	Los metadatos utilizan vocabularios conformes con los principios FAIR	Importante
	RDA-I2-01D	Los datos utilizan vocabularios conformes con los principios FAIR	Útil
I3	RDA-I3-01M	Los metadatos incluyen referencias a otros metadatos	Importante
	RDA-I3-01D	Los datos incluyen referencias a otros metadatos	Útil
	RDA-I3-02M	Los metadatos incluyen referencias a otros datos	Útil

Tabla III: Lista de recomendaciones del FAIR Data Maturity Model de RDA.
Fuente: https://zenodo.org/record/3909563

La evaluación de cada indicador se lleva a cabo estableciendo cinco niveles de cumplimiento:

- 0, no aplicable
- 1, aún no se está considerando
- 2, en estudio o en fase de planificación
- 3, en fase de implementación
- 4, totalmente implementado

Se ofrece la posibilidad de "descartar un indicador", ya que este podría no ser relevante para una comunidad concreta. La razón de ser de este enfoque es dar crédito a la evolución y ayudar a mejorar la gestión de datos. Este enfoque puede ser muy útil para los proveedores y editores de datos que quieran hacer una prueba de autoevaluación y tener una idea más clara de dónde concentrar los esfuerzos para que sus conjuntos de datos satisfagan mejor los principios FAIR.

Relación entre buenas prácticas DWBP y los principios FAIR

Cláudia Sofia Teixeira dos Santos (2023) elaboró en la Universidade do Minho la tesis de máster '*OGD Lens: avaliação automática da qualidade dos dados do European Data Portal*' sobre la evaluación de la calidad de los conjuntos de datos publicados en abierto para proporcionar una guía de mejora de su calidad. Para medir esa calidad desarrolló una metodología basada en los siguientes criterios:

1. Facilidad de uso con la que los usuarios pueden acceder a los datos y utilizarlos para fines de investigación.
2. La disponibilidad de los datos en un formato estándar y abierto y la accesibilidad de los conjuntos de datos a través de un repositorio en línea o una página web.
3. Calidad técnica: la precisión, integridad y consistencia de los datos.
4. Documentación: la información proporcionada junto con los datos, como la descripción de la fuente de los datos, las limitaciones de uso y la frecuencia de actualización.
5. Legalidad: la conformidad de los datos con las leyes de privacidad y derechos de autor.

Los metadatos empleados para describir los conjuntos de datos constituyen una valiosa fuente de información para satisfacer esos niveles de calidad. Por ello, la investigadora acometió un estudio empírico desarrollando una análisis semiautomático de evaluación de la información aportada por esos metadatos en una serie de conjuntos de datos publicados en el Portal Europeo de Datos [13]. Este estudio demostró que la calidad de esos conjuntos varía significativamente. En general, tienen una buena facilidad de uso, pero la calidad técnica, la documentación y la legalidad presentan deficiencias significativas. En cuanto a la facilidad de uso, disponen de una buena documentación sobre cómo acceder a los datos y cómo reutilizarlos. En cambio, en cuanto a la calidad técnica, muchos carecen de información sobre la fuente de los datos, las limitaciones de uso y la frecuencia de actualización. En cuanto a la legalidad, muchos conjuntos de datos no proporcionan información sobre los derechos de autor y la privacidad, lo que puede derivar en un uso inadecuado de los datos. En este estudio se analizó la calidad de **los catálogos**[14] y de **los conjuntos de datos**.

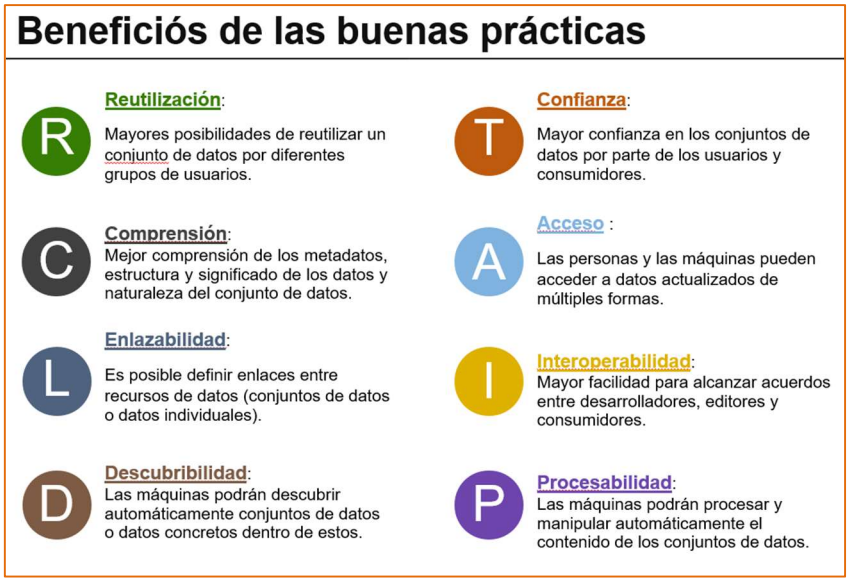


Imagen 2: Beneficios de la aplicación de las buenas prácticas en gestión de datos en la web. Fuente: Pastor Sánchez (2016).

La autora elaboró una tabla que presentaba la serie de beneficios asociada a cada buena práctica (en la tabla siguiente recogemos, a modo de ejemplo, las cinco primeras).

[13] El portal https://data.europa.eu aloja más de millón y medio de conjuntos de datos y 179 catálogos de datos puestos a disposición de la ciudadanía para su posterior reutilización.

[14] Un **catálogo de datos** es un repositorio que contiene información detallada de los **conjuntos de datos** disponibles en una organización. Ofrece metadatos sobre los conjuntos de datos: descripción, origen, estructura, formatos, licencias, fechas de actualización. etc. También puede proporcionar información sobre cómo acceder y utilizarlos.

Buena práctica	Beneficios
BP 1: Proporcionar metadatos	Reusabilidad, comprensibilidad, descubribilidad y procesabilidad
BP 2: Proporcionar metadatos descriptivos	Reusabilidad, comprensibilidad y descubribilidad
BP 3: Proporcionar metadatos estructurales	Reusabilidad, comprensibilidad y procesabilidad
BP 4: Proporcionar información sobre la licencia de los datos	Reusabilidad y confiabilidad
BP 5: Proporcionar información sobre la procedencia de los datos	Reusabilidad y confiabilidad

Tabla IV: Fragmento de la asociación de las buenas prácticas de gestión de datos en la web con sus beneficios (Teixeira dos Santos, 2023, 26-27).

La siguiente tabla muestra las buenas prácticas cuyo cumplimiento proporciona cada uno de estos beneficios siendo la reusabilidad el que agrupa a un número mayor de buenas prácticas seguido de la confianza.

Beneficio	Buenas prácticas
Accesibilidad	BP17, BP18, BP19, BP20, BP21, BP23, BP24, BP32
Comprensión	BP1, BP2, BP3. BP13 , BP15, BP16, BP29, BP31, BP32, BP33
Conectividad	BP9, BP10, BP18, BP24
Confianza	BP4, BP5, BP6, BP7, BP8, BP11, BP15, BP22, BP25, BP26, BP27, BP28, BP29, BP30, BP31, BP32, BP34, BP35
Descubribilidad	BP1, BP2, BP9, BP10, BP11, BP24, BP35
Interoperabilidad	BP9, BP10, BP15, BP16, BP23, BP24, BP26, BP33
Procesabilidad	BP1, BP3, BP12, BP14, BP15, BP18, BP23, BP24, BP31
Reusabilidad	BP1, BP2, BP3, BP4, BP5, BV6, BP7, BP8, BP9, BP10, BP11, BP12, BP13, BP14, BP15, BP16, BP17, BP18, BP19, BP20, BP21, BP22, BP23, BP24, BP25, BP26, BP27, BP28, BP29, BP30, BP31, BP32, BP33, BP34, BP35

Tabla V: Clasificación de las DWBP según beneficio asociado a su uso. Fuente: elaboración propia a partir de Teixeira dos Santos (2023, 26-27).

La investigadora analizó el objeto de cada buena práctica y elaboró una tabla con los metadatos que permitirían asegurar el cumplimiento de cada una de ellas, el valor ideal de esos metadatos (“no nulo” la mayoría de las veces) y los beneficios asociados. En la siguiente tabla se muestra la sección de esa tabla correspondiente a las buenas prácticas 1, 2, 3, 4 y 5.

Buena práctica	Propiedades a cumplir		Valor ideal	Beneficios
	Obligatorias	Recomendado		
BP1: Proporcionar metadatos	dct:title dct:description dct:theme dct:distribution dcat:contactPoint dcat:accessURL dct:license dct:format	dcat:keyword owl:versionInfo foaf:page dct:type dct:temporal dct:spatial dct:source dct:relation dct:provenance dct:modified dct:language	No nulo	Reusabilidad Comprensibilidad Descubribilidad Procesabilidad

Buena práctica	Propiedades a cumplir		Valor ideal	Beneficios
	Obligatorias	Recomendado		
		dct:isVersionOf dct:issued dct:identifier dct:hasVersion dct:creator dct:conformsTo dct:accrucalPeriodicity dct:accessRights dcat:landing_page adms:versionNotes adms:sample adms:identifier		
BP 2: Proporcionar metadatos descriptivos	dct:title dct:description dct:contactPoint dcat:keyword dct:theme dct:distribution dct:publisher dct:format dct:license dct:issued	adms:identifier dct:accessRights dct:accrualPeriodicity dct:conformsTo dct:creator dct:identifier dct:isVersionOf dct:language dct:modified owl:versionInfo adms:status dcat:byteSize dcat:mediaType dct:modified dct:rights	No nulo	Reusabilidad Comprensibilidad Descubribilidad
BP 3: Proporcionar metadatos estructurales	dct:conformsTo		No nulo	Reusabilidad Comprensibilidad Procesabilidad
BP 4: Proporcionar información sobre la licencia de los datos	dct:Accessright dct_license dct:rights		No nulo	Reusabilidad Confiabilidad
BP 5: Proporcionar información sobre la procedencia de los datos	dct_provenance	dct:publiser dct:creator dct_source	No nulo	Reusabilidad Confiabilidad

Tabla VI: Parte de los criterios de validez de la aplicación de las DWBP (Teixeira dos Santos, 2023, 51-53).

La autora diseñó un indicador que mide el porcentaje de "completitud" en la descripción del catálogo o conjunto de datos a partir de los metadatos presentes para cumplir cada buena práctica: el "Ïndice de Calidad de Buenas Prácticas" (ICBP). Si en la **BP5** estuvieran presentes el metadato obligatorio ('provenance') y uno de los optativos ('publisher', por ejemplo), ese factor valdría 1,1, cifra que al dividirse entre 1,3 (el máximo posible a alcanzar en esa buena práctica) proporcionaría un ICBP igual al 84,61%.

La investigación: objetivo y método

Nuestra investigación nos ha llevado a preguntarnos sobre la calidad de la descripción de la información que se publica en los conjuntos de datos de investigación. Para poder evaluarla, tomamos la idea de Teixeira dos Santos de poder llevarla a cabo por medio de un analizador semiautomático que mida la presencia de elementos de metadatos en la descripción de estos objetos y su nivel de cumplimiento de buenas prácticas para hacer posible el cumplimiento de los principios FAIR.

Relacionado con esta temática, a principios del año 2023, realizamos una revisión de la implantación de los repositorios de conjuntos de datos de investigación en las universidades públicas españolas (Martínez Méndez et al., 2023). En este estudio constatamos que las bibliotecas y los servicios de documentación de estas universidades vienen dando pasos importantes en esa gestión, aumentando de forma significativa el volumen de objetos descritos y publicados para su consulta por parte de la comunidad investigadora. Los datos de este estudio se recogieron durante los dos primeros meses del este año y es previsible que, en estos momentos, hayan mejorado considerablemente gracias al esfuerzo que se está desarrollando por parte de profesionales, de sus entidades y de diferentes organizaciones (administraciones, consorcios y redes de bibliotecas universitarias). Los investigadores pueden depositar sus conjuntos de datos en los repositorios institucionales de sus universidades, en los de los consorcios que se han creado a nivel autonómico y también en el repositorio comunitario **Zenodo**. Ese archivo lo pueden llevar a cabo los autores o delegarlo en su biblioteca, pudiendo ser publicado por los autores, pero con revisión posterior de los profesionales de la información.

De la calidad de la descripción llevada a cabo en el registro de esos conjuntos de datos va a depender su posible reutilización y de cumplimiento de los principios FAIR. Si la calidad de esa descripción es baja, el esfuerzo servirá para poco. Evaluar el nivel de calidad aplicado a estos procesos, a partir de la presencia de los metadatos empleados en la descripción, junto con el grado de cumplimiento de buenas prácticas en la gestión de datos en la web proporcionará, sin duda alguna, un conjunto de reflexiones y recomendaciones muy válidas que permitirán corregir (si fuera necesario) alguno fallos identificados y, por consiguiente, mejorar el cumplimiento de los principios FAIR, garantizando una mayor reproducibilidad de los resultados de la investigación.

Como se ha indicado en los párrafos anteriores, el trabajo de Teixeira dos Santos (2023) es un buen punto de partida. Como hipótesis de partida, asumimos que su método puede trasladarse al ámbito de la gestión de los conjuntos de datos de investigación, algo más específico que el estudiado por la investigadora, adaptando su propuesta aplicada en el portal de datos abiertos de la Unión Europea a otro entorno de propósito y características parecidas y más específicas: el portal **Zenodo** [15], el repositorio de acceso abierto a los resultados de la investigación de propósito general desarrollado bajo el programa marco europeo OpenAIRE[16] y operado por el

[15] Más información sobre este repositorio enhttps://zenodo.org/

[16] OpenAIRE es un proyecto europeo de apoyo al movimiento de la Ciencia Abierta para apoyar a la red de expertos en Ciencia Abierta la promoción y formación sobre sus fundamentos. También aporta infraestructura técnica para reunir los resultados de la investigación de los proveedores de datos

CERN[17]. Este repositorio permite a los investigadores depositar artículos de investigación, conjuntos de datos, software de investigación, informes y cualquier otro artefacto digital relacionado con la investigación.

En la fase actual, nuestra investigación tiene **dos propósitos fundamentales**:

1. Averiguar si es posible desarrollar un evaluador semiautomático que analice la metainformación presente en los conjuntos de datos publicados en **Zenodo** y verifique el cumplimiento de las buenas prácticas DWBP. Además del desarrollo del software, que de por sí ya representa un reto, también se pretende ampliar el conjunto de buenas prácticas que se pueden analizar de forma automática (en el trabajo de Teixeira dos Santos (2023), casi un tercio de las DWBP no pudieron formar parte del análisis al no disponer de metadatos que aportaran información sobre su cumplimiento.
2. Si es posible desarrollar e implementar este evaluador, emplearlo para verificar si las buenas prácticas DWBP sirven para llevar a cabo evaluaciones en un contexto algo diferente y si los esquemas de metadatos empleados para la descripción de los conjuntos de datos de investigación resultan suficientes o si necesitan mejorarse.

La elección de **Zenodo** como repositorio para la evaluación se debe, fundamentalmente, a la importancia y trascendencia del mismo. Si somos capaces de desarrollar un método de evaluación semiautomático en el portal de referencia y de destino de la investigación comunitaria, creemos viable aplicarlo posteriormente en entornos más específicos tanto por usar otra plataforma (Dataverse[18] o con DSpace[19], por ejemplo). , como por el tipo de organización (universidad, consorcio u otro tipo de institución científica, por ejemplo). Una posible continuación de esta línea investigadora es verificar otros conjuntos de buenas prácticas, como las elaboradas por la RDA que hemos presentado anteriormente.

El método: pasos a seguir

En el desarrollo de este proyecto investigador se han seguido los siguientes pasos:

1. Selección del repositorio fuente de los conjuntos de dato (**Zenodo**).
2. Identificación de los metadatos que es posible extraer desde la plataforma de forma automática gracias a la API-REST y que constituirán la fuente de información para la evaluación

conectados. OpenAIRE aspira a instituir una infraestructura de comunicación académica abierta y sostenible, responsable de la gestión, el análisis, el manejo, la entrega, el seguimiento y la unión de todos los materiales de investigación. Más información en https://www.openaire.eu/mission-and-vision

[17] El CERN es el Consejo Europeo de Investigación Nuclear, el principal laboratorio de física de partículas del mundo. Es la institución donde Tim Berners-Lee desarrolló el proyecto de la **WWW**.

[18] Dataverse es una aplicación web de código abierto para compartir, preservar, citar, explorar y analizar datos de investigación. Más información en https://dataverse.org/

[19] DSpace es una aplicación de sistema de archivo digital que permite a investigadores y académicos publicar documentos y datos. Fue desarrollado por el MIT y Hewlett Packard. Sirve para el almacenamiento a largo plazo, acceso y preservación de contenidos digitales. Es el preferido por las organizaciones académicas. Más información en https://dspace.lyrasis.org/

3. Ajuste (y ampliación si es posible) de la equivalencia entre los metadatos extraídos del repositorio y los necesarios para la identificación de las buenas prácticas DWBP propuestos por Teixeira dos Santos (2023).
4. Adaptación de la propuesta original de Teixeira dos Santos (2023) de evaluación del cumplimiento de las buenas prácticas DWBP al entorno de nuestra investigación.
5. Diseño del código fuente preciso para la extracción de los metadatos de los conjuntos de datos de investigación.
6. Identificación de las buenas prácticas que sí son aplicables en este experimento: especificación de los requerimientos.
7. Elaboración de los criterios de evaluación y determinación del grado de completitud empleado en la descripción de estos conjuntos de datos.
8. Diseño del código fuente que lleve a cabo la evaluación y la ponderación del nivel de calidad de la descripción.
9. Presentación de resultados, elaboración de recomendaciones y de propuestas de aplicación en otros entornos.

El código fuente ha sido programado con *python*, lenguaje de alto nivel de programación interpretado cuya filosofía hace hincapié en la legibilidad de su código que simplifica el desarrollo de aplicaciones de todo tipo. Existe una amplia cantidad de librerías a libre disposición en distintos repositorios de software de acceso abierto. Los ficheros ejecutables de las rutinas que se han programado en esta investigación han sido depositados en el portal GitHub[20], siguiendo los principios de la Ciencia Abierta.

Desarrollo de la investigación

Zenodo almacena una gran variedad de tipos documentales relacionados con la investigación, muchos de ellos depositados as veces porque es donde deben depositarse para ser accesibles en abierto los resultados de las investigaciones financiadas con fondos comunitarios, tal como recogen las instrucciones de los distintos programas marco. El alcance del proyecto no se va a quedar en este repositorio, se aspira expandir los desarrollos obtenidos a otros sistemas de archivo de conjuntos de datos de investigación. Este propósito está presente en los procesos de toma de decisiones durante la investigación: no proponemos un método específico, los criterios tendrán mayor alcance y serán adaptables a cada contexto específico de aplicación.

[20] GitHub es una plataforma y un servicio basado en la nube para el desarrollo de software y el control de versiones que permite a los desarrolladores almacenar y gestionar su código. Además del sversionado, permite control de acceso, seguimiento de errores, solicitudes de características de software, gestión de tareas, integración continua y wikis para cada proyecto. Es una filial de Microsoft desde 2018. Más información en https://github.com/

Metadatos más frecuentes empleados en los conjuntos de datos

En **Zenodo**, los metadatos siguen un esquema JSON definido[21] y se exportan en los formatos estándares MARCXML[22], Dublin Core y DataCite Metadata Schema[23], según las directrices de OpenAIRE. La lista completa de metadatos usados está disponible en https://developers.zenodo.org/#representation, dividiéndose en cuatro categorías:

1. **Generales**, aplicables a todos los tipos de recursos.
2. **Artículos**: se aplican a los artículos de investigación.
3. **Datos**: aplicables a los conjuntos de datos.
4. **Otros**: se aplican a otros tipos de recursos, como libros, software, arte, etc.

Los metadatos generales informan del título, autores, editores, fecha de publicación, licencia, etc. Los metadatos de artículos incluyen información adicional como el título del artículo, el volumen, el número, las páginas, la revista, etc. Los metadatos de datos incluyen información sobre el conjunto de datos, como el tipo de datos, la descripción, las dimensiones, etc. Los metadatos de otros tipos de recursos incluyen información específica del tipo de recurso.

Datos	título; fecha de creación; DOI; URL del DOI; lista de ficheros; lista de metadatos; fecha de modificación; propietario; identificador del recurso; URL del identificador del recurso; estatus; publicado
Metadatos obligatorios	tipo de fichero; tipo de publicación; tipo de imagen; fecha de publicación; título; creadores; descripción; derechos de acceso; licencia; fecha de embargo; condiciones de acceso
Metadatos optativos de la publicación	DOI; DOI para preservación; palabras clave; notas; identificadores relacionados; colaboradores; referencias; comunidades; subvenciones

Tabla VII: Lista de datos y metadatos de la publicación de un objeto en Zenodo. Fuente: https://developers.zenodo.org/#entities

La siguiente tabla resume la estructura de metadatos de un conjunto de datos de **Zenodo**.

Grupo de metadatos	Claves
Identificadores	created, modified, id, conceptdoi, doi, doi_URL
Generales: **metadata**	title, doi, publication_date, description, access_right. creators, keywords, version, language, resource_type. license, relations
Enlaces: **links**	self, self_html, self_doi, parent, parent_html,parent_doi, self_iiif_manifest, self_iiif_sequence, files, media_files, archive, archive_media, latest, latest_html, draft, versions, access_links, access_users, access_request, access, reserve_doi, communities, communities, suggestions, requests
Revisión	updated, recid, revision
Ficheros	id, filename, filesize, checksum, key, links, self, download
Propiedad	owners
Estado	status, state, submitted

[21] Más información en https://zenodo.org/schemas/records/record-v1.0.0.json
[22] Marco de trabajo que manipula datos MARC en un ambiente XML flexible, extensible y por permitir a los usuarios manipular datos cercanos a sus necesidades. Aúna esquemas, CSS y herramientas software de la Biblioteca del Congreso. Más información en https://www.loc.gov/standards/marcxml//
[23] DataCite recoge las principales propiedades de metadatos seleccionadas para identificar precisa y coherentemente un objeto. Más información en http://schema.datacite.org

Grupo de metadatos	Claves
Estadísticos: **stats**	downloads, unique_downloads, views, unique_views, version_downloads, version_unique_downloads, version_unique_views, version_views

Tabla VIII: Distribución habitual de los metadatos en un 'dataset' de Zenodo. Fuente: elaboración propia.

Zenodo emplea una gran variedad de metadatos que resultan una fuente útil de información para nuestro experimento. Para hacernos una idea más concreta de cuáles son los más frecuentemente empleados en el repositorio hemos implementado el 'script' **lee-frecuencias-metadatos.py**[24] que extrae las claves de los metadatos presentes en la descripción de los 1000 conjuntos de datos'[25] que van a formar parte de la muestra de nuestro experimento (se presenta posteriormente). El resultado de este análisis es el siguiente:

Metadato	%	Metadato	%	Metadato	%
access	100			requests	100
access_links	100	Files	100	reserve_doi	100
access_request	100	Grants	15	resource_type	100
access_right	100	Id	100	revision	100
access_users	100	Imprint	0,9	self	100
alternate_identifiers	0,3	Journal	6,4	self_doi	100
archive	100	Keywords	56,3	self_html	100
archive_media	100	Language	35,6	self_iiif_manifest	100
communities	65,9	Latest	100	self_iiif_sequence	100
communities-suggestions	100	latest_html	100	state	100
conceptdoi	99,9	License	91,2	status	100
conceptrecid	100	media_files	100	subjects	2
contributors	7,6	Meeting	1,7	submitted	100
created	100	Modified	100	title	100
creators	100	Owners	100	unique_views	100
dates	0,5	Parent	100	updated	100
description	100	parent_doi	100	version	45,5
doi	100	publication_date	100	version_downloads	100
doi_url	100	Recid	100	version_unique_downloads	100
downloads	100	References	7,6	version_unique_views	100
draft	100	related_identifiers	22,2	versions	100
embargo_date	0,1	relations	100	views	100

Tabla IX: Metadatos empleados en el repositorio Zenodo con su frecuencia de aparición. Fuente: elaboración propia.

Esta sencilla prueba identifica tres subconjuntos de elementos de metadatos a partir de su frecuencia:

1. Los que aparecen **en todos o en casi todos** (más del 90% de los casos analizados en la muestra), los conjuntos de datos: ('acces', 'access_right', 'access_links'.

[24] El código fuente de este 'script' está recogido en el Anexo II y publicado en Github.

[25] Se analiza el fichero "**1000-dois.txt**" que contiene 100 dois de 'datasets' con cada una de estas 10 palabras: "architecture", "books", "cáncer", "covid-19", "ecosystem", "human", "marine", "migration", "mobility" y "molecules", se han eliminado los duplicados y sustituidos por otros 'datasets' y se ha calculado la frecuencia de aparición de cada metadato.

'access_request', 'access_users', 'archives', 'archive_media', 'communities_suggestioms', 'conceptdoi', 'conceptrecid', 'created', 'creators', 'description', 'doi', 'doi_url', 'files', 'id', 'latest', 'license', modified', 'owners', 'parent', 'parent_doi', 'publication_date', 'recid ', 'relations', 'requests', 'reserve_doi', 'resource_type', 'revisión', 'self', 'self_doi', 'self_html', 'self_iiif_manifest', 'self_iiif_sequence', 'state', 'status', 'submitted', 'title', 'unique_views', 'updated', 'version_downloads', 'version_unique_downloads', 'version_unique_views', 'versions', 'views'.

2. Los que aparecen en un **porcentaje significativo** (un valor comprendido entre el 33% y el 90% de los 'datasets' de la muestra): 'communities', 'keywords', 'language', 'version'
3. De uso ocasional: 'contributors', 'grants', 'imprints', 'journal', 'references', 'meeting', 'subjects' y 'related_identifiers'.

Hay unos pocos metadatos cuya frecuencia es apenas residual y que no incluimos en ningún grupo. Se trata de 'date' y de 'embargo_date'.

Lectura y análisis de los metadatos

El desarrollo del script '**esposende.py**'[26] ha centrado buena parte de esta investigación. Se trata del evaluador semiautomático que lee una lista de 'dois' de entrada (previamente recuperadas por medio del script '**recupera-datasets.py'**[27]) de la que se han eliminado los duplicados. Cada 'doi' equivale a un conjunto de datos publicado en **Zenodo** del cual se extraen los metadatos que exporta la API-REST del repositorio en formato de fichero JSON. Estos metadatos sirven de base para la verificación del cumplimiento de las Data Web Best Practices (2017) del W3C a partir de los criterios de valoración que presentamos en el siguiente apartado.

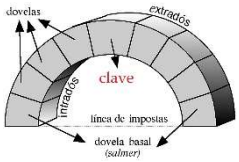
Para poder leer debidamente los metadatos extractados desde el fichero JSON de respuesta, se definen dos grandes listas a nivel interno del 'script': **metadata:** (esta lista agrupa a los elementos de metadatos básicos de Dublin Core) y **record_info:** (en la que se recoge el resto de metadatos). Los resultados de estas verificaciones se almacenan en un archivo CSV denominado "**zenodo.csv**". Sobre el contenido de este fichero de salida se ha llevado a cabo el proceso de análisis estadístico descriptivo.

Propuesta de criterios de valoración

Tras la extracción del conjunto de metadatos con el que vamos a contar para evaluar el cumplimiento de las DWBP, corresponde llevar a cabo la propuesta de criterios de valoración a partir de la elaborada por Teixeira dos Santos (2023).

[26] El código fuente de este 'script' se encuentra publicado en el Anexo IV y en el portal *github*.
[27] El código fuente de este 'script' se encuentra publicado en el Anexo III y en el portal *github*.

La **clave de bóveda** de su propuesta residía en diferenciar entre elementos de metadatos "obligatorios" (sirven para que la buena práctica se cumpla en su totalidad) y elementos "recomendables" [28] (que indican que los autores de los conjuntos de datos aportan información más allá de la básica). Se valora la presencia de ambos, asignando un mayor valor a los obligatorios. En su análisis se toman en consideración tanto metadatos del catálogo de datos como de los conjuntos de datos.



La autora trabajaba en un contexto diferente al de nuestro estudio, el Portal de Datos abiertos de la Unión Europea), donde están disponibles más esquemas de metadatos que en **Zenodo**. Nuestra propuesta, en principio, cuenta con menos elementos de metadatos, aunque, tal como hemos visto anteriormente, el repositorio ofrece abundante información gracias al elevado número de metadatos empleados.

Cada buena práctica DWBP da lugar una regla para verificar su cumplimiento. Una vez establecida cada regla, la misma se introduce en el 'script' analizador semiautomático que va a extraer los metadatos y valorará, a partir de esa información, el nivel de cumplimiento, buena práctica a buena práctica, en cada conjunto de datos de investigación recogidos en la muestra objeto de estudio.

En cada tabla de las que presentamos a continuación recogemos:

1. La buena práctica DWBP.
2. Los beneficios asociados a su uso, relacionados siempre con los principios FAIR, cuyo cumplimiento es el objetivo final de las DWBP.
3. Los elementos de metadatos propuestos para su cumplimiento en la tesis de máster original de Teixeira dos Santos (2023), tanto de los elementos obligatorios como de los elementos opcionales.
4. Nuestra propuesta de elementos de metadatos obligatorios para satisfacer cada DWBP y ña propuesta de elementos recomendables para completar ese nivel de cumplimiento.
5. El criterio de evaluación establecido. Por regla general se toma como base el seguido en la tesis de máster original.
6. La regla escrita debidamente en la sintaxis del analizador semiautomático por medio del lenguaje de programación *python*.

En el trabajo original de Teixeira dos Santos (2023), había algunas buenas prácticas que no pudieron ser objeto de evaluación automática. Nosotros hemos podido incorporarlas a la evaluación de forma automática o de forma semiautomática, asignándoles un valor por defecto en algunos casos. Por esta razón, calificamos a nuestro evaluador como "semi" automático.

De poder analizarse todo el conjunto de las buenas prácticas sin necesidad de definir valores por defecto dependientes de la plataforma **Zenodo**, sería un analizador "automático".

[28] La autora los llama "optativos", pero nosotros hemos cambiado la terminología.

Metadatos

Difícilmente, los datos podrán ser descubiertos y reutilizados si no se proporcionan metadatos suficientes de los mismos aportando información adicional que ayude a los consumidores de datos a comprender mejor el significado de los datos, su estructura y a aclarar otras cuestiones, como los derechos y los términos de la licencia, la organización que generó los datos, la calidad de los datos, los métodos de acceso a los datos y el calendario de actualización de los conjuntos de datos.

Buena práctica	Beneficios	Tese – items	Items metadatos
BP1: Proporcionar metadatos. Esencial para hacer que los datos sean "encontrables".	Reusabilidad Comprensibilidad Descubribilidad Procesabilidad	**Obligatorios** título, descripción, fecha, tema[29], distribución[30], punto de contacto[31], identificador, licencia, formato[32] **Recomendados** dc:keyword dc:version dc:type dc:temporal dc:spatial dc:source[33] dc:relation dc:provenance dc:modified dc:language dc:isVersionOf dc:issued dc:hasVersion dc:creator dc:conformsTo dc:accrucalPeriodicity dc:accessRights	**Obligatorios** título, descripción, fecha, palabras clave, estadísticas, propietarios, doi, licencia. **Recomendables** dc:relations dc:related_identifiers dc:language dc:access_right dc:creators dc:resource_type
BP1: criterio de evaluación	**"1"** si en la descripción del objeto analizado aparecen todos y cada uno de los elementos de metadatos obligatorios + **"0,1"** por cada elemento de metadatos recomendables presente en la descripción del 'dataset'.		

[29] El metadato '**keywords**' describe la materia o tema. Es el que vamos a emplear en este experimento.
[30] El metadato '**dct:distribution**' describe la distribución de un recurso. Puede utilizarse para indicar el formato de archivo en el que está disponible el recurso, el tamaño del recurso, el número de descargas del recurso y alguna otra información. En **Zenodo**, de l formato se encarga la clave "**key**" dentro del metadato '**files**' (se hace uso de ellas en otras buenas prácticas). También vamos a usar el metadato '**stats**' que informa de las descargas.
[31] El metadato más afín aportado por **Zenodo** es '**owners**' que identifica a la persona que ha publicado el 'dataset'.
[32] En **Zenodo**, el elemento de metadatos '**files**' proporciona información sobre la representación física de un recurso digital. Más concretamente su clave '**key**'.
[33] Este elemento de metadatos existe en Dublin Core ('dct:source") pero no se emplea en **Zenodo**. No podemos emplearlo en nuestra revisión.

Buena práctica	Beneficios	Tese – items	Items metadatos

Regla	**# Verificar BP1 - Proporcionar metadatos de forma extensiva**
	
	required_fields = title", "description", "keywords", "publication_date", "doi", "license",]
	if all(field in metadata for field in required_fields) and 'files' in record_data and 'owners' in record_data and 'stats' in record_data:
	bp1_obl = 1
	# Verifica los puntos de referencia BP1_opt
	optional_fields = ["relations", "language", "access_right", "related_identifiers", "creators", "resource_type"]
	bp1_opt = sum(0.1 for field in optional_fields if field in metadata)
	# Calcula BP1 como la suma de BP1_obl y BP1_opt
	bp1 = bp1_obl + bp1_opt		

Valor máximo BP1	**1,6**		

Tabla X: BP1: Proporcionar metadatos.

Buena práctica	Beneficios	Tese – items	Items metadatos
BP2: Proporcionar metadatos descriptivos Esencial para hacer que los datos sean "encontrables".	Reusabilidad Comprensibilidad Descubribilidad	**Obligatorios** título, descripción, punto de contacto, tema, editor[34], formato, licencia, fecha **Recomendados** dct:accessRights dct:conformsTo dct:creator dct:identifier[35] dct:isVersionOf[36] dct:language dct:modified dct:rights	**Obligatorios** título, descripción, palabras clave, propietarios, licencia, fecha. **Recomendables** dc:access_right dc:relations dc:related_identifiers dc:creators dc:doi dc:language dc:updated
BP2: criterio de evaluación	**"1"** si en la descripción del objeto analizado aparecen todos y cada uno de los elementos de metadatos obligatorios + **"0,1"** por cada elemento de metadatos recomendables presente en la descripción del 'dataset'.		

[34] En **Zenodo** se entiende que el '**publisher**' es el propio repositorio porque proporciona el soporte para ello. También se puede considerar que quien publica los datos es el editor, opción que vamos a emplear porque disponemos del metadato '**owner**' para identificar a quien se ha responsabilizado de la publicación del conjunto de datos.

[35] En los conjuntos de datos de investigación, el identificador 'doi' es de **obligada presencia**. En realidad, lo es para todos los tipos de contenidos que se depositan y/o publican en el respositorio.

[36] La cláusula '**isVersionOf**' forma parte del elemento metadatos '**related_identifiers**' en **Zenodo,** es una clave de elemento de metadato.

Buena práctica	Beneficios	Tese – items	Items metadatos

Regla	**# Verificar BP2 - Proporcionar metadatos descriptivos**
	
	required_fields = ["title", "description", "publication_date", "keywords", "license"]
	if all(field in metadata for field in required_fields) and 'owners' in record_data:
	bp2_obl = 1
	# Verifica los puntos de referencia BP2_opt
	optional_fields = ["access_right", "relations", "creators", "language", "related_identifiers", "doi",]
	bp2_opt = sum(0.1 for field in optional_fields if field in metadata)
	if 'updated' in record_data:
	bp2_opt += 0.1
	# Calcula BP2 como la suma de BP2_obl y BP2_opt
	bp2 = bp2_obl + bp2_opt		

Valor máximo BP2	**1,7**		

Tabla XI: BP2: Proporcionar metadatos descriptivos.

Buena práctica	Beneficios	Tese – items	Items metadatos
BP3: Proporcionar metadatos estructurales. Estructura o y esquema de una distribución[37]	Reusabilidad Comprensibilidad Procesabilidad	**Obligatorio** dc:conformsTo[38]	**Obligatorio:** **Recomendables** dc:resource_type dc:related_identifiers
BP3 Criterio de evaluación	"**1**" si en la descripción del objeto analizado aparece el elemento de metadatos obligatorio + "**0,1**" por cada elemento de metadatos recomendable presente en la descripción del 'dataset'.		

Regla	**# Verifica BP3 - Proporcionar metadatos estructurales**
	bp3_obl = 0
	# Verifica el punto de referencia BP3_opt
	if "resource_type" in metadata and "related_identifiers" in metadata:
	bp3_opt = 0.2
	else:
	bp3_opt = 0.1 if "resource_type" in metadata or "related_identifiers" in metadata else 0
	# Calcula BP3 como la suma de BP3_obl y BP3_opt
	bp3 = bp3_obl + bp3_opt		

Valor máximo BP3	**0,2**		

Tabla XII: BP3: Proporcionar metadatos estructurales.

[37] Esta BP verifica si se ofrece información sobre la **estructura interna** de los ficheros de un conjunto de datos (la descripción de columnas de un fichero CSV, por ejemplo). **Zenodo** informa sobre formato, fecha de actualización, extensión y URL si bien no alcanza ese nivel. No se dispone de metadatos para satisfacer el requisito obligatorio, pero podemos valorar los elementos 'resource_type' y related_identifiers' que aportan información indirectamente.

[38] Este elemento de metadatos indica la calidad de un recurso (si cumple con estándares o especificaciones) No hemos encontrado ningún metadato en **Zenodo** que aporte esta información. Es **una posibilidad de mejora** en la descripción, si bien es posible que esta ausencia se deba a que el repositorio cumpla con estándares establecidos (como *OpenAIRE FAIR Data* o *Registry Specification*) y que no se vea necesario aportar más información.

Licencias

Según el tipo de licencia adoptado por el editor habrá más o menos restricciones para compartir y reutilizar los datos. En el contexto de la web, la licencia de un conjunto de datos puede especificarse dentro de los metadatos, o fuera de ellos en un documento independiente al que esté vinculado.

Buena práctica	Beneficios	Tese – items	Items metadatos
BP4: Proporcionar un enlace o copia de la licencia que controla el uso de los datos. El acceso y uso de los datos puede estar restringido por la licencia. Esta información permite hacerlos "accesibles" y "encontrables".	Reusabilidad Confiabilidad	**Obligatorio** dc:rights;	**Obligatorio:** licencia **Recomendables** dc:access_right dc:access_users dc:access_links
BP4 Criterio de evaluación	"1" si en la descripción del objeto analizado aparece el elemento de metadato obligatorio + "0,1" si también aparece el elemento de metadatos Recomendado.		

Regla	**# Verifica BP4 - datos de la licencia**
	if "license" in metadata:
	bp4_obl = 1
	else:
	bp4_obl = 0
	# Verifica el punto de referencia BP4_opt
	bp4_opt = 0
	if "access_right" in metadata:
	bp4_opt += 0.1
	if 'access_users' in record_info.get('links', {}):
	bp4_opt += 0.1
	if 'access_links' in record_info.get('links', {}):
	bp4_opt += 0.1
	# Calcula BP4 como la suma de BP4_obl y BP4_opt
	bp4 = bp4_obl + bp4_opt		

Valor máximo BP4	**1,3**		

Tabla XIII: BP4: Proporcionar licencias.

Procedencia de los datos

El reto que plantea la publicación de datos en la web es proporcionar un nivel adecuado de detalle sobre su origen. El productor de los datos puede no ser necesariamente el editor de los mismos, por lo que es especialmente importante recopilar y transmitir los metadatos correspondientes. Sin procedencia, los consumidores no van a tener una forma inherente de confiar en la integridad y credibilidad de los datos que se comparten. A su vez, los editores de datos deben ser conscientes de las necesidades de las comunidades de consumidores potenciales para saber cuántos detalles de procedencia son apropiados.

Buena práctica	Beneficios	Tese – items	Items metadatos
BP5: Proporcionar información sobre la procedencia de los datos. Es importante para su reutilización y verificación.	Reusabilidad Confiabilidad	**Obligatorio** dct:provenance[39] **Recomendados** dct:publisher dct:creator	**Obligatorio:** contribuidores **Recomendables** dc:archive dc:creators dc:related_identifiers dc:grants
BP5 Criterio de evaluación	"1" si en la descripción del objeto analizado aparece el elemento de metadatos obligatorio + "0,1" por cada uno de los elementos de datos recomendados presentes en la descripción.		

Regla	**# Verifica BP5 - datos sobre la procedencia**
	if "source" in metadata:
	bp5_obl = 1
	#Verifica el punto de referencia BP5_opt
	optional_fields = ["creators", "related_identifiers", "grants"]
	bp5_opt = sum(0.1 for field in optional_fields if field in metadata)
	if 'archive' in record_info.get('links', {}):
	bp5_opt += 0.1
	# Calcula BP5 como la suma de BP5_obl y BP5_opt
	bp5 = bp5_obl + bp5_opt		

Valor máximo BP5	1,4		

Tabla XIV BP5: Proporcionar información sobre la procedencia.

Calidad de los datos

Este aspecto tiene incidencia en las aplicaciones que los utilizan. Incluir información sobre la calidad de los datos en la publicación de datos es primordial. Implica a varias dimensiones de calidad que representan grupos de características que son relevantes para editores y consumidores.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP6: Proporcionar información sobre la calidad de los datos y su adecuación Valora la idoneidad que ayudará a mejorar la confianza y un uso futuro.	Reusabilidad Confiabilidad	**Obligatorio** dqv:hasQualityMeasurement[40]	**Obligatorio:** **Recomendables** dc:updated dc:creators dc:owners
BP6 Criterio de evaluación	"1" si en la descripción del objeto analizado aparece el elemento de metadatos obligatorio; "0" en caso contrario.		

[39] En Zenodo, el metadato más relacionado con la procedencia es el elemento '**contributors**', que proporciona una lista de personas u organizaciones que han contribuido a la creación o desarrollo de un 'dataset'.

[40] El metadato **dqv:hasQualityMeasurement** pertenece a VOCAB-DQV, vocabulario del **W3C** que informa sobre la calidad de un recurso: precisión, exhaustividad; relevancia, actualidad y facilidad de uso.

Buena práctica	Beneficios	Tese – ítems	Tipo
Regla	**# Verifica BP6 - Proporcionar información sobre la calidad y adecuación.** BP6_obl = 0 **if 'creators' in metadata:** bp6_opt += 0.1 **if 'updated' in record_info:** bp6_opt += 0.1 **if 'owners' in record_info:** bp6_opt += 0.1 bp6 = bp6_obl + bp6_opt		
Valor máximo BP6	**0,3**		

Tabla XV: BP6: Calidad en los datos.

Versionado de los datos

Los conjuntos de datos publicados en la web pueden cambiar con el paso del tiempo. Algunos se actualizan de forma programada. Pueden crearse nuevas versiones de un conjunto de datos, no hay consenso sobre cuándo es un conjunto de datos diferente en lugar de una nueva versión.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP7: Proporcionar un indicador de versión. Asigne e indique un #versión o una fecha para cada dataset. La versión de los datos es importante para la transparencia y la reproducibilidad. Permite rastrear el historial de cambios y saber si se está utilizando la versión correcta.	Reusabilidad Confiabilidad	**Obligatorio** owl:versionInfo[41] **Recomendado** dct:modified[42]	**Obligatorio:** versión **Recomendables** dc:revision dc:updated
BP7 Criterio de evaluación	**"1"** si en la descripción del objeto analizado aparece el elemento de metadatos obligatorio más **"0.1"** por cada clave de metadato que informe de la versión o de las fechas de actualización		

Regla	**# Verifica BP7 - Proporcionar un indicador de versión.**
	if "version" in metadata:
	bp7_obl = 1
	# Verifica bp7_opt - si tenemos info de versión y fecha de actualización
	optional_fields = ["updated", "modified"]
	bp7_opt = sum(0.1 for field in optional_fields if field in record_data)
	bp7 = bp7_obl + bp7_opt		

Valor máximo BP7	**1,2**		

Tabla XVI: BP7: Proporcionar un indicador de versión.

[41] El elemento '**owl:versionInfo**' se refiere a la versión de una ontología. Más información en http://ontolog.cim3.net/file/resource/OOR/OMV/OMV-Reportv2.4.1.pdf

[42] En **Zenodo** exisetn los metadatos 'updated' y 'modified'. El primero informa de cuándo se actualiza el conjunto de datos, el segundo de la fecha de modificación del contenido.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP8: Proporcione un historial de versiones completo que explique los cambios realizados. Proporcionar un historial de versiones completo que explique los cambios realizados en cada versión.	Reusabilidad Confiabilidad	**Obligatorio** dct:isVersionOf dct:hasVersion **Recomendado** dct:modified dct:source	**Obligatorio** **metadato** relaciones **clave** version **Recomendables** dc:revision dc:updated
BP8 Criterio de evaluación	"1" si en la descripción del objeto analizado aparece el elemento de metadatos obligatorio más "0.1" si en el campo "relations" se informa del historial de versión.		

Regla	**# Verifica BP8 - Proporcione un historial de versiones.**
	if "relations" in metadata and "version" in metadata:
	bp8_obl = 1
	#Verifica bp8_opt - el número de versiones del dataset
	bp8_opt = 0
	if "revision" in record_data:
	bp8_opt += 0.1
	if "updated" in record_data:
	bp8_opt += 0.1
	# Calcula BP8 como la suma de BP8_obl y BP8_opt
	bp8 = bp8_obl + bp8_opt		

Valor máximo BP8	**1,2**		

Tabla XVII: BP8: Proporcionar historial de versiones.

Identificadores de datos

Los identificadores adoptan muchas formas y se utilizan ampliamente en todos los sistemas de información. El descubrimiento, uso y citación de datos en la web dependen fundamentalmente del uso de URI de http (o https): identificadores únicos a nivel mundial.

Buena práctica	Beneficios	Tese - ítems	Tipo
BP9: Utilizar URI persistentes como identificadores de los conjuntos de datos. Identifique cada 'dataset' mediante un URI persistente elegida de forma cuidadosa.	Reusabilidad Conectividad Descubribilidad Interoperabilidad	**Obligatorio** dct:identifier	**Obligatorio** doi
BP9 Criterio de evaluación	"1" si en la descripción del objeto analizado aparece el elemento obligatorio.		

Regla	**# Verifica BP9 - Utilizar URI persistentes como identificadores**
	if "doi" in metadata:
	bp9_obl = 1
	# No hay elementos optativos en BP9
	bp9_opt = 0
	# Calcula BP9 como la suma de BP9_obl y BP9_opt
	bp9 = bp9_obl + bp9_opt		

Valor máximo BP9	**1**		

Tabla XVIII: BP9 URIs persistentes en los conjuntos de datos.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP10: Utilizar URI persistentes como identificadores dentro de los conjuntos de datos. Reutilice los URI de otros como identificadores en los conjuntos de datos siempre que sea posible.	Reusabilidad Conectividad Descubribilidad Interoperabilidad	La autora no considera esta buena práctica para una evaluación automática. **Zenodo** asigna un único identificador a cada fichero del 'dataset'. Es la clave "id".	**Obligatorio:** **metadato** ficheros **clave** id **Recomendable**: dc:related_identifiers
BP10 Criterio de evaluación	**"1"** si en la descripción del objeto analizado aparece el elemento obligatorio.		

Regla	**# Verifica la presencia de la clave "id" en "files"**
	if "files" in record_data:
	for file_info in record_data["files"]:
	if "id" in file_info:
	bp10_obl = 1
	# Verifica elementos recomendables para BP10
	if "related_identifiers" in metadata:
	bp10_opt = 0,1
	# Calcula BP10 como la suma de BP10_obl y BP10_opt
	bp10 = bp10_obl + bp10_opt		

Valor máximo BP10	**1,1**		

Tabla XIX: BP10: Emplear URIs persistentes dentro de los conjuntos de datos.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP11: Asignar URI a versiones de conjuntos de datos y series. Asignar URI a versiones individuales y a la serie global.	Reusabilidad Descubribilidad Confiabilidad	**Obligatorio** dcat:accessURL	**Obligatorio:** doi (tema)[43] **Recomendables** dc:related_identifiers
BP11 Criterio de evaluación	**"1"** si en la descripción del objeto analizado aparece el elemento obligatorio.		

Regla	**# Verifica BP11 - Asignar URI a versiones de conjuntos de datos y series.**
	if 'conceptdoi' in record_info:
	bp11_obl = 1
	# Verifica bp11_opt - si está el elemento "related_identifiers"
	if 'related_identifiers' in metadata:
	bp11_opt = 0.1
	# Calcula BP11 como la suma de BP11_obl y BP11_opt
	bp11 = bp11_obl + bp11_opt		

Valor máximo BP11	**1,1**		

Tabla XX: BP11: Asignar URIs a versiones de conjuntos de datos y series.

[43] El elemento de metadato 'conceptdoi' es el identificador único que se asigna a un concepto científico.

Formatos de los datos

El formato en el que se ponen a disposición los conjuntos de datos es un aspecto clave para hacer que sean utilizables. El mejor mecanismo de acceso no tiene sentido a menos que proporcione datos en formatos que permitan su uso y reutilización.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP12: Utilizar formatos de datos normalizados legibles por máquina. Poner los datos a disposición en un formato de datos normalizado y legible por máquina que se adapte bien a su uso previsto o potencial.	Reusabilidad Procesabilidad	**Obligatorios** dct:conformsTo dct:format	**Obligatorio:** **metadato** ficheros **clave** key
BP12 Criterio de evaluación	"**1**", por defecto.		

Regla	**#VerificaBP12-Formatos datos legibles x máquina.**
	if "files" in record_data:
	for file_info in record_data["files"]:
	if "key" in file_info:
	bp12_obl = 1
	#No existen elementos recomendables.
	bp12_opt = 0
	# BP12 es la suma de BP12_obl y bp12_opt
	bp12 = bp12_obl + bp12_opt		

Valor máximo BP12	**1**		

Tabla XXI: BP12: Utilizar formatos legibles por máquina.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP13: Utilizar representaciones de datos neutras desde el punto de vista local. Usar estructuras de datos y valores neutros en la configuración regional o aportar metadatos sobre esa configuración.	Reusabilidad Confiabilidad	**Obligatorio** La autora no considera esta buena práctica para la evaluación automática[44].	**Obligatorio:** idioma
BP13 Criterio de evaluación	"**1**" si en la descripción del objeto analizado aparece el elemento obligatorio.		

Regla	**# Verifica BP13 - idioma del dataset.**
	if "language" in metadata:
	bp13_obl = 1
	# No hay elementos de verificación para BP13_opt
	bp13 = bp13_obl		

Valor máximo BP13	**1**		

Tabla XXII: BP13: Utilizar representaciones de datos neutras.

[44] Las DWBP incluyen la verificación de la zona horaria, información que no proporciona **Zenodo**.

Buena práctica	Bene	Tese – ítems	Tipo
BP14: Proporcionar datos en múltiples formatos. Facilitar los datos en varios formatos cuando más de uno se adapte a su uso previsto o potencial.	Reusabilidad Procesabilidad	**Obligatorio** dct:format dcat:mediaType En **Zenodo**, el elemento de metadatos '**files**' aúna la información de estos dos elementos obligatorios.	**Obligatorio:** **metadato** ficheros **clave** key
BP14 Criterio de evaluación	**"1"** si en la descripción del objeto analizado aparece el elemento obligatorio.		

Regla	**# Verifica punto de referencia BP14 - proporcionar la información en múltiples formatos**
	
	if "files" in record_data:
	for file_info in record_data["files"]:
	if "key" in file_info:
	bp14_obl = 1
	# No hay elementos de verificación para BP14_opt
	bp14_opt = 0
	# Calcula BP14 como la suma de BP14_obl y BP14_opt
	bp14 = bp14_obl		

Valor máximo BP14	**1**		

Tabla XXIII: BP14: Proporcionar datos en múltiples formatos.

Vocabularios de los datos

Los vocabularios definen los conceptos y relaciones utilizados para describir y representar un área de interés. Sirven para clasificar los términos a usar en una aplicación concreta, caracterizar las posibles relaciones y definir las posibles restricciones en su uso. Se han acuñado varios términos equivalentes: ontología, vocabulario controlado, tesauro, taxonomía, lista de códigos o red semántica.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP15: Reutilizar vocabularios, preferiblemente normalizados. Emplear términos de vocabularios compartidos. Si es posible se normalizan, para codificar datos y metadatos.	Reusabilidad Comprensibilidad Confiabilidad Interoperabilidad Procesabilidad	**Obligatorios** dct:language dct:theme dct:spatial dct:mediaType dct:license	**Obligatorios** idioma palabras clave tipo license **Recomendables** dc:access_right dc:access_conditions
BP15 Criterio de evaluación	"**1**" si en la descripción del objeto analizado aparecen todos y cada uno de los metadatos obligatorios + "**0,1**" si aparece el elemento metadatos recomendado en la descripción del dataset.		

Regla	**# Verifica BP15 reutilizar vocabularies**
	required_fields = ["language", "keywords", "license", "resource_type"]
	if all(field in metadata for field in required_fields):
	bp15_obl = 1
	# Verifica los puntos de referencia BP15_opt
	if "access_right" in metadata:
	bp15_opt = 0.1
	if "access" in record_info:
	bp15_opt += 0.1# Calcula BP15
	bp15 = bp15_obl + bp15_opt

Valor máximo BP15	**1,2**

Tabla XXIV: BP15: Reutilizar vocabularios normalizados.

Buena práctica	**Beneficios**	**Tese – ítems**	**Tipo**
BP16: Elegir el nivel de formalización adecuado. Opte por un nivel de semántica formal que se adapte tanto a los datos como a las aplicaciones más probables.	Reusabilidad Comprensibilidad Interoperabilidad	**Obligatorios** dct:conformsTo dct:format	**Obligatorio** **metadato** ficheros **clave** key. **Recomendables** dc=license dc=standard dc=vocabularies[45]
BP16 Criterio de evaluación	"**1**" si en la descripción del objeto analizado aparecen los tres elementos de metadatos obligatorios.		

Regla	**# Verifica BP16 - Elegir el nivel de formalización adecuado.**
	if "files" in record_data:
	for file_info in record_data["files"]:
	if "key" in file_info:
	bp16_obl = 1
	# Verifica puntos de referencia para bp16_opt
	bp16_opt = 0
	if "license" in metadata:
	bp16_opt += 0.1
	# Calcula BP16
	bp16 = bp16_obl + bp16_opt		

Valor máximo BP16	**1,1**		

Tabla XXV: BP16: Elegir el nivel de formalización adecuado.

Acceso a los datos

Facilitar el acceso a los datos en la web permite a personas y máquinas aprovechar las ventajas de compartir datos utilizando infraestructura web. Por defecto se accede mediante el Protocolo de Transferencia de Hipertexto (HTTP) a nivel de transacción atómica. Puede ser a través de la simple descarga masiva de un archivo o, cuando los datos están distribuidos en varios archivos o requieren métodos de recuperación más sofisticados.

[45] La API-REST de **Zenodo** no informa de la presencia y contenido de los elementos de metadatos 'standard' y 'vocabularies'. Esto podría ser objeto de una **posible mejora**.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP17: Descarga masiva del 'dataset'. Permite a los consumidores recuperar el conjunto de datos completo con una sola solicitud.	Reusabilidad Accesibilidad	**Obligatorio** dcat:downloadURL	**Obligatorio:** **metadato** enlaces **clave** archive
BP17 Criterio de evaluación	"1" si en la descripción del objeto analizado aparece el elemento de metadatos obligatorio; "0" en caso contrario.		
Regla	**# Verifica BP17 - descarga masiva del dataset** **if "links" in record_data and "archive" in record_data["links"]:** bp17_obl = 1 **# No hay puntos de referencia para bp17_opt** bp17_opt = 0 bp17 = bp17_obl + bp17_opt		
Valor máximo BP17	1		

Tabla XXVI: BP17: Descarga masiva del 'dataset'.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP18: Proporcione subconjuntos para grandes conjuntos de datos. Si el 'dataset' es grande, los subconjuntos permitirán que usuarios y aplicaciones trabajen fácilmente.	Reusabilidad Conectividad Accesibilidad Procesabilidad	La autora no considera esta buena práctica para la evaluación automática.	**Obligatorio:** **Recomendable**: **metadato** dc: **related_identifiers** **clave** isPartOf[46] isContinuedBy
BP18 Criterio de evaluación	"1" si en la descripción del objeto analizado aparece el elemento de metadatos obligatorio; "0" en caso contrario.		

Regla	**# Verifica BP18- - proporciones subconjuntos para grandes conjuntos**
	# No existen metadatos para bp18_obl
	bp18_obl = 0
	# Verifica bp18_opt
	bp18_opt = 0
	if "related_identifiers" in metadata:
	for file_info in record_data["files"]:
	if "isPartOF" in file_info:
	bp18_obl = 0.1
	if "PartOF" in file_info:
	bp18_opt += 0.1
	if "isContinuedBy" in file_info:
	bp18_obl += 0.1
	if "continues" in file_info:
	bp18_opt += 0.1
	bp18 = bp18_obl + bp18_opt		

Valor máximo BP18	0,2		

Tabla XXVII: BP18: Proporcione subconjuntos de grandes conjuntos de datos.

[46]Podemos emplear también la clave 'PartOf'. Asimismo, para la clave 'isContinuedBy' también se puede emplear 'continues'. Los resultados del estudio indican que, en la práctica, **su uso es nulo**.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP19: Utilizar la negociación de contenidos para servir datos disponibles en varios formatos. Usar negociación de contenidos y las extensiones de archivo para servir datos disponibles en varios formatos.	Reusabilidad Accesibilidad	La autora no considera esta buena práctica para la evaluación automática.	**Obligatorio:** **Recomendables** dc:files (su clave "keys") dc:access_right dc:access_links El metadato '**files**', por medio de la clave 'files' aporta información sobre el formato de datos de cada archivo individual, información que resulta útil para verificar el cumplimiento.
BP19 Criterio de evaluación	**"1"** si se verifica que el servidor cumple la buena práctica más **"0.1"** si se cumple el requisito recomendable.		

Regla	**# Verifica BP19 - Utilizar la negociación de contenidos para servir datos**
	bp19_obl = 0
	# Verifica bp19_opt – usar la negociación de contenidos
	bp19_opt = 0
	if "files" in record_data:
	for file_info in record_data["files"]:
	if "key" in file_info:
	bp19_opt = 0.1
	if "links" in record_data and "access_links" in record_data["links"]:
	bp19_opt += 0.1
	if "access_right" in metadata:
	bp19_opt += 0.1
	bp19 = bp19_obl + bp19_opt		

Valor máximo BP19	**0,3**		

Tabla XXVIII: BP19: Proporcione subconjuntos de grandes conjuntos de datos.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP20: Proporcionar acceso en tiempo real[47]. Cuando los datos se produzcan en tiempo real, se deben poner a disposición en la web en tiempo real (o "casi real").	Reusabilidad Accesibilidad	**Obligatorio** dct:accrualPeriodicity[48]	**Obligatorio** **Recomendable:** dc:updated Este metadato aporta información sobre la última fecha de modificación de un conjunto de datos. Esto es útil para verificar el cumplimiento.

[47] Existen conjuntos de datos con información sismológica, climática, financiera, etc. creados en tiempo real, aspecto quizá no demasiado trascendente en una posterior reutilización en futuras investigaciones.

[48] En **Zenodo** no se emplea este metadato. Se ha consultado a los responsables del repositorio y la causa se debe a la propia naturaleza del repositorio, concebido para publicar y preservar resultados de investigaciones. Lo más afín es la fecha de actualización que aporta el metadato 'updated'.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP20 Criterio de evaluación	**"1"** si se verifica que el servidor cumple la buena práctica más **"0.1"** si se cumple el requisito recomendable.		

Regla	**# Verifica BP20 - Proporcionar acceso en tiempo real.**
	# No hay metadatos para verificar bp20_obl
	bp20_obl = 0
	#Verifica bp20_opt
	if "updated" in "record_info":
	bp20_opt = 0.1
	if "modified" in record_info:
	bp20_opt += 0.1
	# Calcula BP20 como la suma de BP20_obl y BP20_opt
	bp20 = bp20_obl + bp20_opt		

Valor máximo BP20	**0,2**		

Tabla XXIX: BP20: Proporcione acceso a los datos en tiempo real.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP21: Proporcionar datos actualizados. Facilitar datos actualizados explicitando la frecuencia de actualización.	Reusabilidad Acessibilidad	**Obligatorios** dct:accrualPeriodicity dct:modified	**Obligatorio** fecha de actualización **Recomendables** dc:created dc:modified
BP21 Criterio de evaluación	**"1"** si se verifica que el servidor cumple la buena práctica más **"0.1"** si se cumple el requisito recomendable.		

Regla	**# Verifica BP21 - Informar de la fecha de actualización de la información .**
	if "updated" in record_info:
	bp21_obl = 1
	#Verifica bp21_opt
	bp21_opt = 0
	if "created" in record_info:
	bp21_opt += 0.1
	if "modified" in record_info:
	bp21_opt += 0.1
	# Calcula BP21 como la suma de BP21_obl y BP21_opt
	bp21 = bp21_obl + bp21_opt		

Valor máximo BP21	**1,2**		

Tabla XXX: BP21: Proporcione datos actualizados.

Buena práctica	Beneficios	Tese - ítems	Tipo
BP22: Explicación acerca de la no presencia ni disponibilidad de algunos datos.[49] Si los datos no están disponibles, explique cómo y quién puede acceder a ellos.	Reusabilidad Confiabilidad	**Obligatorio** adms:status	**Obligatorio:** estado **Recomendable:** dc:state
BP22 Criterio de evaluación	**"1"** si en la descripción del objeto analizado aparece el elemento de metadatos obligatorio; **"0"** en caso contrario.		

Regla	**# Verifica BP22 - Explicación de los datos no disponibles.**
	if "status" in record_info:
	bp22_obl = 1
	#Verifica BP22_opt -
	if "state" in record_info:
	bp22_opt = 0.1
	else:
	bp22_opt = 0
	# Calcula BP22 como la suma de BP22_obl y BP22_opt
	bp22 = bp22_obl + bp22_opt		

Valor máximo BP22	**1,1**		

Tabla XXXI: BP22: Explicación acerca de la no presencia de algunos datos.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP23: Ofrecer datos a través de una API. Ofrezca una API para servir datos si dispone de los recursos para hacerlo.	Reusabilidad Confianza	La autora no considera esta buena práctica para la evaluación automática.	**Obligatorio** derechos de acceso En **Zenodo** se verifica si el metadato '**access_right**' tiene como valor "open"[50]. Si es así, la respuesta es **afirmativa**.
BP23 Criterio de evaluación	**"1"** si en la descripción del objeto analizado aparece el elemento de metadatos obligatorio con el contenido indicado; **"0"** en caso contrario.		
Regla	**# Verifica BP23 – Comprueba presencia de 'access_right' y si su valor es "open".** **if "access_right" in metadata and metadata["access_right"] == "open":** bp23_obl = 1 **# No hay elementos para verificar bp23_opt** bp23_opt = 0 bp23 = bp23_obl + bp23_opt		
Valor máximo BP23	**1**		

Tabla XXXII: BP23: Ofrecer datos a través de una API.

[49] Para indicar la no disponibilidad de datos se puede publicar un documento HTML con una explicación. También se pueden utilizar códigos de estado HTTP apropiados con mensajes personalizados legibles por humanos y máquinas, como el metadato '**status**' que describe el estado de un recurso. Su valor puede ser una palabra como "completado", "incompleto", "pendiente", etc., o una expresión como "en revisión".

[50] Este metadato puede tener los siguientes valores: 'open' (acceso abierto), 'embargged' (acceso embargado), 'restricted' (acceso restringido) y 'closed' (acceso cerrado).

Buena práctica	Beneficios	Tese – ítems	Tipo
BP24: Utilizar los estándares web como base de las API. Cuando se diseñen APIs, se debe usar un estilo arquitectónico basado en tecnologías web.	Reusabilidad Conectividad Descubribilidad Accesibilidad Interoperabilidad Procesabilidad	La autora no considera esta buena práctica para la evaluación automática.	**Obligatorio:** solicitudes de acceso **Zenodo** cuenta con el metadato "access_request" que permite a los desarrolladores gestionar las solicitudes de acceso a los registros del repositorio. Esta gestión se realiza mediante una solicitud HTTP POST, que es un **estándar web**.
BP24 Criterio de evaluación	"**1**" si en la descripción del objeto analizado aparece el elemento de metadatos obligatorio con el contenido indicado; "**0**" en caso contrario.		

Regla	**# Verifica BP24 – Si la API sigue estándares web**
	if 'access_request' in record_info.get('links', {}):
	bp24_obl = 1
	# No hay elemento de metadatos para verificar bp24_opt
	bp24_opt = 0
	# Calcula BP24 como la suma de BP24_ob l y BP24_opt
	bp24 = bp24_obl + bp24_opt		

Valor máximo BP24	**1**		

Tabla XXXIII: BP24: Utilizar estándares web como base de la API.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP25: Proporcione documentación completa sobre su API. Proporcione información completa y actualizada sobre la API.	Reusabilidad Confianza	La autora no considera esta buena práctica para la evaluación automática.	Se puede evaluar en función del repositorio y fijar su valor antes del análisis. En el caso de **Zenodo**, **la respuesta es afirmativa**[51]**.** Su valor es "1".
BP25 Criterio de evaluación	Esta buena práctica exige consultar cada repositorio y verificar si existe o no esa documentación completa y de calidad. Si es así, se asigna a la variable correspondiente un valor de "**1**" en el propio script.		
Regla	**# Verifica BP25 - Si hay acceso a la documentación de la API pública** **#Se cumple en Zenodo**. bp25 = 1		
Valor máximo BP25	**1**		

Tabla XXXIV: BP25: Proporcione documentación completa sobre su API.

[51] La API-REST de **Zenodo** está disponible en la URL https://developers.zenodo.org/

Buena práctica	Beneficios	Tese - ítems	Tipo
BP26: Evite cambios perjudiciales en su API. Evitar cambios de código que provoquen "rupturas" en los programas de clientes que usen la API. Todo cambio preciso debe comunicarse pensando en la actualización de las rutinas vinculadas.		La autora no considera esta buena práctica para la evaluación automática.	Se debe evaluar en función del repositorio y fijar su valor antes del análisis. En el caso de **Zenodo**, la **respuesta es afirmativa.** De la consulta de la documentación se desprende que informa debidamente de ese tipo de cambios[52].
BP26 Criterio de evaluación	Para verificar esta BP hay que consultar el repositorio y verificar si en la documentación se indica cómo se informa a los usuarios. En caso afirmativo, se asigna directamente a la variable correspondiente un valor de "**1**", en caso contrario "**0**".		
Reglas	**# Verifica BP26 - Evitar cambios bruscos en la API-REST publica.** **# Se cumple en Zenodo.** bp26 = 1		
Valor máximo BP26	**1**		

Tabla XXXV: BP26: Evite cambios perjudiciales en la API.

Preservación de los datos

Es importante que los editores indiquen cuándo los datos han sido eliminados o archivados. Eliminarlos directamente, sin más explicación, es una mala práctica porque provoca confusión y errores. Los editores deben usar redirecciones, archivos de lápida o metadatos para indicar qué ha sucedido. Esto ayuda a los usuarios a encontrar los datos que necesitan, incluso si ya no están.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP27: Conservar los identificadores. Al retirar datos de la web, conserve el identificador y facilite información sobre el recurso archivado.		La autora no considera esta buena práctica para la evaluación automática. **Política de repositorios**: En principio, se conserva este identificador.	Se debe evaluar en función del repositorio y fijar su valor antes del análisis. Cuando un objeto se retira de **Zenodo**, se elimina de la base de datos. Esto incluye el identificador 'doi' del objeto que ya **no se asociará con el objeto retirado** ni con ningún otro.
BP27 Criterio de evaluación	Se asigna un valor "**1**" si la política del repositorio indica que un DOI no se reutiliza. Se asigna "**0**" en caso contrario o si no conocemos ese aspecto en concreto.		
Regla	**# Verifica BP27 - Conservar los identificadores.** **# En Zenodo se cumple.** bp27 = 1		
Valor máximo BP27	**1**		

Tabla XXXVI: BP26: Conservar los identificadores.

[52] Como ejemplo reciente, a finales del pasado mes de octubre de 2023 se llevó a cabo una actualización del sistema y se informó de ello de forma suficiente por medio de mensajes emergentes en la plataforma.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP28: Evaluar la cobertura de un conjunto de datos. Se debe hacer antes de su conservación.	Reusabilidad Confianza	La autora no considera esta buena práctica para la evaluación automática.	**Obligatorio** cobertura Este requisito obligatorio no se va cumplir en ningún 'dataset' porque ese metadato no forma parte de la respuesta de la API de Zenodo. **Debería ser una mejora**.
BP28 Criterio de evaluación	"1" si en la descripción del objeto analizado aparece el elemento de metadatos obligatorio con el contenido indicado; "0" en caso contrario.		
Regla	**# Verifica BP28 - Evaluar la cobertura del repositorio.** **if "coverage" in metadata:** bp28_obl = 1 **#Verifica BP28_opt - no hay elementos recomendados** bp28_opt = 0 bp28 = bp28_obl + bp28_opt		
Valor máximo BP28	1		

Tabla XXXVI: BP28: Evaluar la cobertura de un conjunto de datos.

Retroalimentación ('feedback')

Publicar en la web permite compartir datos a gran escala con una amplia variedad de audiencias con diferentes niveles de experiencia. Los publicadores quieren asegurar que los datos satisfacen las necesidades de los consumidores y, para ello, la retroalimentación del usuario es crucial porque beneficia tanto a los publicadores como a los consumidores, ayudando a los primeros a mejorar la integridad de sus datos publicados y fomentando la publicación de nuevos datos.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP29: Recoger las opiniones de los consumidores de datos. Ofrecer un medio fácil para aportar su opinión.	Reusabilidad Compreensibilidad Confiabilidade	**Obligatorio** dcat:contactpoint	**Obligatorio** propietarios **Recomendable** dc:creators
BP29 Criterio de evaluación	**"1"** si se verifica que el servidor cumple la buena práctica más "**0.1**" por cada requisito recomendable que se cumpla.		
Regla	**# Verifica BP29 – Recoger opiniones de los consumidores de datos.** **if "owners' in metadata:** bp29_obl = 1 else: bp29_obl = 0 **# Verifica el punto de referencia BP29_opt** If "creators" in metadata: bp29_opt = 0.1 bp29 = bp29_obl + bp29_opt		
Valor máximo BP29	1,1		

Tabla XXXVII: BP29: Recoger las opiniones de los consumidores de datos.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP30: Hacer públicas las opiniones. Publicar las opiniones de los consumidores sobre los conjuntos de datos y las distribuciones.	Reusabilidad Comprensibilidad Confiabilidad	La autora no considera esta buena práctica para la evaluación automática[53].	No se dispone de información suficiente para considerar una evaluación automática de esta buena práctica. Es precisa la evaluación manual previa al análisis. En **Zenodo** no se cumple.
BP30 Criterio de evaluación	Esta buena práctica no se puede evaluar de forma automática. Se asigna un valor tras la revisión previa de "**1**" si se satisface, de "**0**" en caso contrario.		
Regla	**# Verifica BP30 – Hacer públicas las opiniones de los consumidores de datos.** **# En Zenodo no se cumple.** BP30 = 0		
Valor máximo BP30	**0**		

Tabla XXXVIII: BP30: Publicar las opiniones de los consumidores de datos.

Enriquecimiento de datos

Es el conjunto de procesos utilizados para mejorar, refinar o mejorar de alguna otra manera los datos crudos o previamente procesados. Esta idea, junto con otros conceptos similares, contribuyen a convertir los datos en un activo valioso. En la investigación científica, se debe tener cuidado de evitar un enriquecimiento que distorsione resultados o resultados estadísticos.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP31: Enriquecer los datos generando otros nuevos. Generación de nuevos datos cuando se considere que ello aumenta su valor.	Reusabilidad Comprensibilidad Confiabilidad Procesabilidad	La autora no considera esta buena práctica para la evaluación automática.	**Obligatorio:** **Recomendables** **metadato** dc:related_identifiers **claves** isDerivedFrom isSourceOf.
BP31 Criterio de evaluación	"**1**" si se verifica que el servidor cumple la buena práctica; "**0**" en el caso contrario.		

Regla	**# Verifica BP31 - Enriquecer los datos generando otros nuevos.**
	# No hay metadatos para verificar bp31
	bp31 = 0
	#Verifica elementos recomendables
	optional_relations = {"isDerivedFrom", "isSourceOf"}
	bp31_opt = [0.1 for key in optional_relations if key in metadata]
	# Calcula BP31 como la suma de BP31_obl y BP31_opt
	bp31 = bp31_obl + bp31_opt		

Valor máximo BP31	**0,2**		

Tabla XXIX: BP31: Enriquecer los datos generando otros nuevos.

[53] Esta buena práctica tiene más que ver con la política de fidelización de cada repositorio.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP32: Proporcionar presentaciones complementarias. Enriquecer los datos aportando presentaciones complementarias e inmediatamente informativas: tablas, resúmenes, visualizaciones o aplicaciones web.	Reusabilidad Comprensibilidad Confiabilidad Accesibilidad	**Obligatorio**: dct:Type	**Obligatorio:** **Recomendables** **metadato** dc:related_idendigiers **claves:** isSupplementTo isSuplementedBy isDescribedBy isPartOf isCompiledBy isReferencedBy
BP32 Criterio de evaluación.	**"1"** si se verifica que el servidor cumple la buena práctica; "**0**" en el caso contrario.		
Regla	**# Verifica BP32 – Ofrecer presentaciones complementarias.** bp32_obl = 0 **# Verifica elementos recomendables** optional_relations = {"isSupplementTo", "isSuplementedBy", "describes", "isDescribedBy"} bp32_opt = [0.1 for key in optional_relations if key in metadata] bp32 = bp32_obl + bp32_opt		
Valor máximo BP32	**0,4**		

Tabla XL: BP32: Proporcionar presentaciones complementarias.

Republicación

Reutilizar datos es otra forma de publicar datos. Puede tomar la forma de combinar datos existentes con otros conjuntos de datos, crear aplicaciones web, visualizaciones o reempaquetar los datos en una traducción. Los "re-editores" tienen responsabilidades únicas para esa forma de publicar.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP33: Informar al editor original. Posibilidad de Informar al editor original cuando reutilice sus datos. Si encuentra algún error o tiene alguna sugerencia o cumplido, hágaselo saber.	Reusabilidad Comprensibilidad Interoperabilidad	**Obligatorio**: dcat:contactPoint	**Obligatorio**: propietarios **Recomendable**: dc:creators
BP33 Criterio de evaluación	**"1"** si se verifica que el servidor cumple la buena práctica más "**0.1**" por cada requisito recomendable que se cumpla.		

Regla	**# Verifica BP33 - Poder informar al editor original.**
	if "owners' in metadata:
	bp33_obl = 1
	# Verifica el punto de referencia BP33_opt
	If "creators" in metadata:
	bp33_opt = 0.1
	# Calcula B33 como la suma de BP33_obl y BP33_opt
	bp33 = bp33_obl + bp33_opt		

Valor máximo BP33	**1,1**		

Tabla XLI: BP33: Informar al editor original.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP34: Cumplir las condiciones de licencia Busque y siga los requisitos de licencia del editor original del 'dataset`.	Reusabilidad Confianza	La autora no considera esta buena práctica para la evaluación automática. .	**Obligatorio:** **metadato** dc:'links' **clave** access **Recomendable:** dc:license
BP34 Criterio de evaluación	"**1**" si se verifica que el servidor cumple la buena práctica más "**0**" en el caso contrario.		
Regla	**# Verifica BP34 - Cumplir las condiciones de licencia del editor original.** **if "access" in record_info.get('links', {}):** bp23_obl = 1 **# Verifica el punto de referencia para bp34_opt** if "license" in metadata: bp34_opt = 0.1 bp34 = bp34_obl + bp34_opt		
Valor máximo BP34	**1,1**		

Tabla XLII: BP34: Cumplir las condiciones de la licencia.

Buena práctica	Beneficios	Tese – ítems	Tipo
BP35: Citar la publicación original Reconocer la fuente de sus datos en los metadatos.	Reusabilidad Descubribilidad Confiabilidad	**Obligatorio:** dcat:source	**Obligatorio:** **Recomendables** **metadato** dc:related_identifiers **claves** references cites isBasedOn isDerivedFrom isSupplementTo, isPartOf
BP35 Criterio de evaluación	"**1**" si en la descripción del objeto está el metadato obligatorio + "**0,1**" por cada uno de los elementos recomendables presentes en la descripción.		
Reglas	**# Verifica punto de referencia BP35 - Citar la publicación original** bp35_obl = 0 **# Verificar BP35_opt - si en 'related_idenfiers hay referencias.** if "related_identifiers" in metadata: cited_keys = {"isBasedOn", "isPartOf", "references", "cites", "isDerivedFrom", "isSupplementTo"} related_identifiers = metadata["related_identifiers"] bp35_opt += sum(0.1 for item in metadata["related_identifiers"] if item.get("relation") in cited_keys) bp35 = bp35_obl + bp35_opt		
Valor máximo BP35	**0,6**		

Tabla XLIII: BP35: Citar la publicación original.

Resumen de criterios de valoración.

La siguiente tabla se muestra la totalidad de los criterios de evaluación.

Propiedades	Buena práctica	Incluida	IBPQ	Observaciones
Metadatos	BP1: proporcionar metadatos.	Sí	**1,8**	
	BP2: Proporcionar metadatos descriptivos	Sí	**1,7**	
	BP3: Proporcionar metadatos estructurales.	Sí	**0,2**	
Licencias	BP4: Proporcionar un enlace o copia de la licencia de uso de los datos.	Sí	**1,3**	
Procedencia	BP5: Proporcionar información sobre la procedencia de los datos.	Sí	**1,4**	
Calidad	BP6: Ofrecer información sobre la calidad de los datos y su adecuación	Sí	**0.3**	
Versionado	BP7: Aportar un indicador de versión	Sí	**1**	
	BP8: Proporcione un historial de versiones completo	Sí	**1,2**	
Identificadores	BP9: Utilizar URI persistentes como identificadores	Sí	**1**	
	BP10: Utilizar URI persistentes como identificadores dentro de los conjuntos de datos.	Sí	**1,1**	
	BP11: Asignar URI a versiones de conjuntos de datos y series.	Sí	**1,1**	
Formatos	BP12: Utilizar formatos de datos normalizados legibles por máquina	Sí	**1**	
	BP13: Utilizar representaciones de datos neutras desde el punto de vista local.	Sí	**1**	
	BP14: Proporcionar datos en múltiples formatos.	Sí	**1**	
Vocabularios	BP15: Reutilizar vocabularios normalizados.	Sí	**1,2**	
	BP16: Elegir el nivel de formalización adecuado.	Sí	**1.3**	
Acceso	BP17: Descarga masiva del 'dataset'.	Sí	**1**	

Propiedades	Buena práctica	Incluida	IBPQ	Observaciones
	BP18: Proporcione subconjuntos para grandes conjuntos de datos.	Sí	0.2	
	BP19: Usar negociación de contenidos para servir datos en varios formatos.	Sí	0,3	
	BP20: Proporcionar acceso en tiempo real.	Sí	0,1	
	BP21: Proporcionar datos actualizados.	Sí	1,2	
	BP22: Explicación acerca de la no presencia ni disponibilidad de datos.	Sí	1,1	
	BP23: Ofrecer datos a través de una API.	Sí	1	
	BP24: Utilizar los estándares web como base de las API.	Sí	1	.
	BP25: Proporcione documentación completa sobre su API.	Sí	1	**Zenodo** ofrece esa documentación sobre su API. El valor va a ser siempre "1".
	BP26: Evite cambios perjudiciales en su API.	Sí	1	**Zenodo** sigue esta recomendación. El valor va a ser siempre "1".
Preservación	BP27: Conservar los identificadores.	Sí	1	**Zenodo** sigue esta recomendación. El valor va a ser siempre "1".
	BP28: Evaluar la cobertura de un conjunto de datos.	Sí	1	
Retroalimentación	BP29: Recoger las opiniones de los consumidores de datos.	Sí	1,2	
	BP30: Hacer públicas las opiniones.	Sí	0	**Zenodo** no lo ofrece. El valor va a ser siempre nulo.
Enriquecimiento	BP31: Enriquecer los datos generando otros nuevos.	Sí	0,2	
	BP32: Proporcionar presentaciones complementarias.	Sí	0,4	
Republicación	BP33: Informar al editor original.	Sí	1,1	
	BP34: Cumplir las condiciones de licencia	Sí	1,1	
	BP35: Citar la publicación original	Sí	0,6	

Tabla XLIV: Resumen de las buenas prácticas aplicadas en nuestro experimento y valoración Fuente: elaboración propia.

Metainvestigación: avances en el evaluador

Hemos conseguido que todas las DWBP formen parte de nuestro análisis semiautomático. La verificación automática del cumplimiento de 31 buenas prácticas (un 88,5% del total) es posible a partir de la información que aportan los metadatos presentes en la descripción de los conjuntos de datos. Las restantes 4 buenas prácticas, las evaluadas de forma semiautomática, necesitan de una revisión previa manual y de la inserción del resultado en el código del analizador porque no disponemos de metadatos que permitan valorar su cumplimiento. Si tomamos como referencia el estudio de Teixeira Dos Santos (2023) llevado a cabo sobre catálogos y conjuntos de datos del Portal Europeo de Datos, en el mismo se excluyeron 11 buenas prácticas. Llegar a este nivel de análisis representa un avance.

Tras el análisis de las frecuencias de aparición de los metadatos, previamente a la revisión de las buenas prácticas, se intuye que algunas de ellas se van a cumplir siempre, por ejemplo, las BP3, BP9 y BP29 que exigen la presencia de los metadatos 'resource_type', 'doi' y 'owners' respectivamente que están presentes en el 100% de los objetos a analizar. En el otro extremo, hay buenas prácticas cuyo nivel de cumplimiento del requisito obligatorio va a ser nulo porque ninguno de los elementos de metadatos que permitirían informar positivamente forma parte de la respuesta que proporciona el fichero JSON generado por la API de Zenodo, como es el caso del metadato 'coverage' exigido por la BP28, por ejemplo.

En aquellas buenas prácticas donde es necesario llevar a cabo una revisión previa manual e introducir el valor de la variable asociada a cada una de ellas en el código del analizador semiautomática. Se trata de las buenas prácticas BP25, BP26. BP27 y BP30.

Buena práctica	Valor	Justificación
BP25: Proporcione documentación completa sobre su API.	1	**Zenodo** ofrece esa documentación sobre su API en https://developers.zenodo.org/ y en GitHub.
BP26: Evite cambios perjudiciales en su API.	1	El repositorio sigue esta recomendación.
BP27: Conservar los identificadores.	1	El repositorio sigue esta recomendación.
BP30: Hacer públicas las opiniones.	0	El repositorio no ofrece esta prestación.

Tabla XLV: Buenas prácticas que precisan de evaluación semiautomática. Fuente: elaboración propia.

En estos casos, el resultado del cumplimiento de cada buena práctica está aquí determinado por la plataforma **Zenodo**. Esto sitúa la dependencia del cumplimiento con respecto a la misma en un porcentaje algo menor del 15%, porcentaje que estimamos como aceptable más si cabe si consideramos que estamos aportando información, algo que no ocurría en el trabajo original de Teixeira do Santos (2023), donde un poco más del 30% de las buenas prácticas no llegaban a evaluarse. El hecho de que algunas buenas prácticas las dicte el diseño de la plataforma y no el nivel y la calidad de la descripción de los conjuntos de datos es lógico, no debemos olvidar que estas buenas prácticas abarcan tanto a los metadatos como al diseño de los sistemas de depósito y/o publicación que las aplican en su desarrollo e implementación.

En el caso de que se pudiera establecer algún tipo de correlación o agrupamiento por distancias entre los resultados, aquellas que vienen determinadas por la plataforma no tienen valor de discriminación alguno porque su valor va a ser siempre el mismo. Todas las buenas prácticas asociadas a la aportación extensiva de metadatos, licencias, procedencia de los datos, calidad, versionado, uso de identificadores, formatos, vocabularios, enriquecimiento y republicación (esta última creemos que con escaso aporte de información), van a poder evaluarse de modo automático. En cuanto a las de acceso a los datos (el grupo más grande de buenas prácticas), tres de ellas (un tercio) son evaluadas de forma semiautomática. En el caso de las buenas prácticas de retroalimentación y enriquecimiento de los datos, una se evalúa de forma automática y la otra de forma semiautomática.

Índice de Calidad de la Buena Práctica -IBPQ

Teixeira do Santos (2023) estableció una medida del nivel de cumplimiento de cada práctica a la que llama "Índice de Calidad de la Buena Práctica" o IQBP. Para calcular este valor, se divide la medida de cumplimiento de una buena práctica que ofrece el evaluador al analizar un conjunto de datos y se divide entre el valor máximo que esta buena práctica podría alcanzar. En el caso de la buena práctica BP1, el valor máximo sería 1,7 puntos a partir del punto obtenido por el cumplimiento de los requisitos obligatorios más 0,7 puntos correspondientes al cumplimiento de hasta los 7 requisitos recomendados. La propia autora reconoce que va a ser complicado llegar al 100% en los valores de este índice en las buenas prácticas que posean muchos elementos recomendables para la descripción (como la anterior BP1), si bien lograrlo conllevará un mayor cumplimiento de los principios FAIR y una mejor calidad del repositorio. Al igual que ella, nosotros vamos a asumir esta circunstancia en el cálculo de este índice.

Nivel cualitativo de cumplimiento de las buenas prácticas

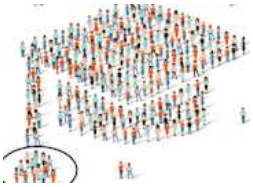
El anterior índice ofrece una medida cuantitativa del nivel de completitud de la descripción de cada conjunto de datos. Queremos aportar una información cualitativa suplementaria a partir del total de buenas prácticas en las que un conjunto de datos satisface el nivel el requisito obligatorio (o requisitos), el total donde solo satisface requisitos recomendables y el número de casos donde no satisface ni obligatorios ni recomendables. En la siguiente tabla recogemos cuatro casos de ejemplo, una muestra pequeña pero que ya adelanta algunos patrones en el cumplimiento de las DWBP (en realidad en el "no cumplimiento"). Si bien esta información parte de datos cuantitativos, va a terminar aportando información cualitativa.

doi	Obligatorios	Recomendables	No se cumplen
	#bp >1	0 < #bp <	#bp = 0
4972098	17	9	9
5443258	22	4	9
5816881	17	4	9
6914806	24	2	9

Tabla XLVI: Nivel cualitativo de cumplimiento de las buenas prácticas (fragmento del listado total). Fuente: elaboración propia.

El experimento

En Zenodo hay 3.124.411 objetos digitales publicados, de los que 211.821 son conjuntos de datos de investigación ('datasets'): un 6,77 del total. Este porcentaje se puede considerar alto si se compara con la presencia de los conjuntos de datos en los repositorios institucionales de las universidades públicas españolas que usan DSpace (Martínez Méndez et al., 2023) o del escasísimo 0,54% de conjuntos de datos presentes en el portal Recolecta (15909 'datasets' dentro de una colección de 2.908.102)[54].



Para determinar el tamaño de la muestra a revisar para asegurar que sea significativa, vamos a usar la fórmula de cálculo para una población finita. La fórmula (en términos porcentuales) para calcular el tamaño de la muestra en una población finita es:

$$\text{Tamaño de la muestra} = \frac{Z^2 \cdot p \cdot (1-p)}{E^2 \cdot (N-1) + Z^2 \cdot p \cdot (1-p)}$$

Donde:

- ***Z*** es el valor crítico que corresponde al nivel de confianza deseado (por ejemplo, 1.96 para un nivel de confianza del 95%).
- **p** es la proporción estimada de la población que tiene la característica objeto de estudio (en tu caso, el cumplimiento de las DWBP en la descripción de los conjuntos de datos publicados en Zenodo).
- **E** es el margen de error deseado (generalmente expresado como un porcentaje).
- **N** es el tamaño de la población total (211.821 conjuntos de datos).

Para determinar el tamaño de la muestra objeto de estudio, estimamos como hipótesis que el porcentaje de la población que satisface el cumplimiento de las buenas prácticas es del 66,66% (dos de cada tres 'datasets'), establecemos el nivel de confianza en el 95% y el margen de error en un 1%. De esta forma, sustituyendo esos valores en la ecuación tendríamos:

$$\text{Tamaño de la muestra} \approx \frac{0.270432 \cdot 0.22}{0.0001 \cdot 211{,}820 + 0.270432 \cdot 0.22}$$

Resolviendo estos cálculos, se obtiene una muestra de 593 objetos como representativa. Nosotros realizaremos el experimento con una muestra de **1000 conjuntos de datos de investigación**, la unión de 10 subconjuntos de 100 registros de Zenodo del tipo de contenido. Se han recuperado por separado los subconjuntos buscando por las siguientes palabras: 'architecture', 'books', 'cancer', 'covid-19', 'ecosystem', 'marine', 'migration', 'molecules', 'mobility', 'human'. Se ha verificado que no coincide ningún resultado de búsqueda y se ha

[54] Los datos de **Zenodo** y de **Recolecta** han sido extraídos de cada portal durante la primera semana de octubre de 2023.

creado el fichero "**1000-dois.txt**", el de entrada en el evaluador semiautomático. Los resultados de aplicar las rutinas definidas en el evaluador para verificar el cumplimiento de las DWBP se almacenan en el fichero "**zenodo.csv**".

Índice de calidad de las buenas prácticas

En primera instancia, este análisis lo realizamos pormenorizando por cada grupo de buenas prácticas afines establecidos por el W3C. En la parte final de este apartado, presentamos los datos de forma general y un resumen de resultados.

Metadatos

Los resultados de estas buenas prácticas, muy importantes porque están destinadas a aportar metainformación consistente y suficiente a los conjuntos de datos, son los siguientes:

	media	mediana	V(BP)_max	ICBP
BP1	0,9688	1,4	1,8	0,53
BP2	1,0688	1,5	1,7	0,62
BP3	0,0444	0	0,2	0,22

Tabla XLVII: Resultados ICBP en las buenas prácticas vinculadas a los metadatos.

La **BP1** verifica si se aportan metadatos descriptivos suficientes y la **BP2** centra su atención en la presencia suficiente de metadatos estructurales. En ambos casos, los requisitos obligatorios son exigentes, pero es lo mínimo para satisfacer lo indicado en las DWBP.

BP1: metadatos requeridos de carácter general	"title", "description", "keywords", "publication_date", "doi", "license"
BP2:metadatos requeridos estructurales	"title", "description", "publication_date", "keywords", "license"

Tabla XLVIII: Metadatos obligatorios de las buenas prácticas BP1 y BP2.

La mediana es el valor que se encuentra justo en el medio de un conjunto de números cuando se ordenan en secuencia. Es decir, la mitad de los números son menores que la mediana y la otra mitad son mayores. En el contexto de nuestro experimento, en la variable BP1, una mediana de 1.4 significa que la mitad de los conjuntos de datos analizados tienen un ICBP menor o igual a 1.4, y la otra mitad tiene un valor que es mayor o igual a 1.5. Si, además, la mediana es mayor que la media (1,4 > 0.96) y el valor de ICBP es 1.8, esto sugiere que la distribución de los datos es asimétrica o **sesgada hacia la izquierda**: hay una mayor concentración de valores más bajos que la mediana. Algo similar ocurre en la BP2. Estas dos variables se han visto especialmente perjudicadas por el escaso porcentaje de aparición del metadato 'keywords' en la muestra de nuestro experimento (solo un 56,3%) que propicia que casi la mitad de los conjuntos de datos analizados tengan valores comprendidos entre 0 y 1 (de ahí el sesgo a la izquierda, más si cabe cuando los máximos son 1,8 y 1,7). En el caso de la BP3, la información que aportan los metadatos recomendables es muy escasa, están presentes en solo una quinta parte de los conjuntos de datos.

Licencias

Los resultados son los siguientes:

	media	mediana	V(BP)_max	ICBP
BP4	1,212	1,3	1,3	0,93

Tabla XLIX: Resultados ICBP en las buenas prácticas vinculadas a las licencias.

La **BP4** verifica la presencia del metadato "license" como requisito obligatorio y de los metadatos "access_right", "access_users' y 'access_links' como metadatos recomendables. Ahora también la mediana es mayor que la media que se ve afectada por el 8,8% de conjuntos de datos donde no aparece el metadato obligatorio, pero la distancia es mucho menor y esta buena práctica tiene un índice de cumplimiento del ICBP del 93% ya que los metadatos recomendables están siempre presentes.

Procedencia

Los resultados son los siguientes:

	media	mediana	V(BP)_max	ICBP
BP5	0,1976	0,1	1,4	0,14

Tabla L: Resultados ICBP en las buenas prácticas vinculadas a las licencias.

En la presentación de los criterios de valoración de esta buena práctica se comentaba que Zenodo no aporta metadatos suficientes para garantizar el cumplimiento del requisito obligatorio. Sólo contamos con el metadato 'contributors' que apenas está presente en un 7,6% de los conjuntos de datos. La presencia más significativa de los metadatos recomendables permite obtener una media de casi 0,2, muy lejos del valor máximo que podría alcanzarse: 1,4. Los estadísticos aportan más información sobre esta variable: se trata. distribución sesgada asimétrica hacia los valores más bajos porque la mayoría de los datos están agrupados en el extremo inferior del rango. También puede haber algunos valores atípicos o un pequeño número de valores relativamente altos que estén alejando la media (0.2) de la mediana (0.1).

Calidad

Los resultados son los siguientes:

	media	mediana	V(BP)_max	ICBP
BP6	0,3	0,3	0.3	1

Tabla LI: Resultados del ICBP en las buenas prácticas vinculadas a la calidad de los datos.

Se trata de una variable que, si bien no satisface un requisito obligatorio porque Zenodo no aporta metadatos para ello, sí cumple íntegramente las condiciones expresadas en las DWBP. En el caso de Zenodo es posible que se aporte esta información en el metadato 'description' pero no es seguro del todo y tampoco es posible analizarlo automáticamente.

Versionado

Los resultados son los siguientes:

	media	mediana	V(BP)_max	ICBP
BP7	1,1455	1,1	1,2	0,95
BP8	1,2	1,2	1,2	1

Tabla LII: Resultados ICBP en las buenas prácticas vinculadas al versionado de los datos.

En la variable **BP7** (historial completo de versiones), la media y la mediana están muy próximas entre sí, lo que sugiere que la distribución es relativamente simétrica con un sesgo levemente positivo con una ligera inclinación hacia la derecha. Esto significa que hay una cola ligeramente más larga en el extremo derecho de la distribución. Como el valor máximo posible es 1,2 (muy cercano a la media y mediana), se puede inferir que la mayoría de los datos están agrupados cerca del mismo sin una variabilidad extrema. Y lo que es más importante, estos valores muestran que **esta variable se verifica** en un porcentaje muy alto de los casos. En cuanto a la variable **BP8** se satisfacen todos los requisitos, el obligatorio y los dos remendables, en todos los casos. Es una buena práctica que **siempre se cumple.**

Identificadores

Los resultados son los siguientes:

	media	mediana	V(BP)_max	ICBP
BP9	1	1	1	1
BP10	0,2162	0	1,1	0,19
BP11	1,02	1	1,1	0,92

Tabla LIII: Resultados ICBP en las buenas prácticas vinculadas al versionado de los datos.

La **BP9** exige la presencia del metadato obligatorio ‘’doi’, lo que **se cumple siempre.** La **BP10** intenta verificar la presencia de URLs persistentes dentro de los conjuntos de datos. Su distribución muestra una asimetría significativa (1,54) lo que apunta a una distribución con una cola larga hacia la derecha, es decir: existe una concentración de valores bajos y un número menor de valores altos, pero lo suficientemente influyentes como para afectar la forma general de la distribución. En líneas generales, el cumplimiento de esta variable es reducido. En cuanto a la **BP11**, que verifica si se asignan URIs a versiones de conjuntos de datos y series, los resultados son altamente positivos porque la mayoría de los valores están concentrados cerca del extremo superior del rango. Esto crea una asimetría positiva y la distribución general está centrada cerca del valor máximo.

Formatos

Los resultados son los siguientes:

	media	mediana	V(BP)_max	ICBP
BP12	0,914	1	1	0,91
BP13	0,36	0	1	0,36
BP14	0,914	1	1	0,91

Tabla LIV: Resultados ICBP en las buenas prácticas vinculadas al formato de los datos.

La **BP12** verifica si se hace uso de formatos de datos normalizados legibles por máquina exige la presencia del metadato obligatorio muestra también resultados satisfactorios. Como la media está cerca del valor máximo, hay una cantidad importante de valores cerca del máximo, aunque algunos que no satisfacen el requisito obligatorio inciden en la media final (la asimetría de -2,96 y la curtosis de 6,75 así lo sugieren). El análisis de la **BP13,** la buena práctica que verifica el uso de representaciones neutras (en **Zenodo** solo disponemos del metadato 'language'), muestra una media baja (0.36) pero significativamente más alta que la mediana (0), influida por aquellos casos donde se satisface el requisito obligatorio que elevan la media que, por otro lado, se queda considerablemente por debajo del valor máximo y así el índice de cumplimiento tiene un valor reducido. La asimetría de 0.60 confirma que hay algunos valores más altos que la mayoría e influyen en la media, si bien la distribución no es extremadamente sesgada, dato que también confirma la curtosis: una distribución más aplanada en comparación con una distribución normal con una mezcla de valores bajos y moderadamente altos, pero sin valores extremos. La **BP14** verifica si se proporciona información en múltiples formatos. Para poder verificar su cumplimiento hemos de verificar si la clave 'key' está dentro del metadato 'files' y cuántas veces. Los resultados y la explicación son idénticos que los de **BP12**, se satisface esta buena práctica en un muy considerable porcentaje de los casos estudiados.

Vocabularios

Los resultados son los siguientes:

	media	mediana	V(BP)_max	ICBP
BP15	0,712	1,2	1,2	0,59
BP16	1,005	1,1	1,1	0,91

Tabla LV: Resultados del ICBP en las buenas prácticas vinculadas al uso de vocabularios.

La **BP15** muestra unos valores medios de cumplimento (el índice ICBP es 0,59). Se encuentra muy afectada porque, para cumplir con el requisito obligatorio es precisa la presencia del metadato 'keywords' cuya frecuencia de aparición apenas roza el 60% de los casos analizados. En cambio, la **BP16** obtiene unos valores alto de cumplimiento gracias a la alta presencia de la clave 'key' dentro del metadato 'files', con un alto índice ICBP (0,91). no permite satisfacer el requisito obligatorio en muchos de los conjuntos de datos analizados.

Acceso

Los resultados de este grupo de buenas prácticas, el más amplio, son los siguientes:

	media	mediana	V(BP)_max	ICBP
BP17	1	1	1	1
BP18	0	0	0.1	0
BP19	0,291	0,3	0,3	0,97
BP20	0,1	0,1	0,1	1
BP21	1,2	1,2	1,2	1
BP22	1,1	1,1	1,1	1
BP23	0,914	1	1	0,91
BP24	1	1	1	1
BP25	1	1	1	1
BP26	1	1	1	1

Tabla LVI: Resultados ICBP en las buenas prácticas vinculadas al acceso a los datos.

En general, las DWBP relacionadas con el acceso a los conjuntos de datos poseen un elevado nivel de cumplimiento de los requisitos con una única excepción. La **BP17** verifica si es posible la descarga del conjunto de datos íntegro (no solo fichero a fichero), lo es gracias a la clave 'archive' del metadato 'links' presente en todos los conjuntos analizados. La **BP18** verifica si se proporcionan subconjuntos para grandes conjuntos de datos, **Zenodo** no dispone de esta prestación y tampoco aporta información para satisfacer los requisitos recomendables. En el caso de la **BP19**, que intenta verificar si se hace uso de la negociación de contenidos para servir datos disponibles en varios formatos, tampoco disponemos de metadatos para satisfacer el requisito obligatorio, en cambio sí disponemos de hasta tres recomendaciones de metadatos para aportar información y el nivel de cumplimiento es alto, de ahí que, si bien esta variable adopta valores pequeños (máximo 0,3) el valor del índice ICBP es alto (0,97). Algo parecido ocurre con la **BP20** que no puede verificar de forma obligatoria el requisito de disponer la publicación de datos en tiempo con metadato alguno pero que sí aporta información actualizada de los metadatos recomendables 'updated' y 'modified' en todos los casos analizados, así que el ICBP adopta el valor de 1. Las buenas prácticas **BP21** y **BP22** satisfacen todos los requisitos obligatorios y recomendables para informar sobre fecha de actualización de la información y de la frecuencia de actualización la primera de ellas y la explicación de los datos no disponibles en la segunda. En ambos casos, el ICBQ adopta también el valor de 1. La **BP23** verifica si el conjunto de datos está accesible ('open`) por medio de la API-Rest del repositorio, algo que se satisface en el 91,4% de los casos (ICBP = 0,91). Como esta API-Rest se verifica en todos los casos que sigue estándares web, la **BP24** tiene el mismo valor de media, mediana y valor máximo, con un ICBP de 1. La misma distribución le corresponde las buenas prácticas **BP25** y **BP26,** pero, en este caso, los valores

se asignan de forma semiautomática al no disponer de metadatos que aporten esta información. Se le asigna el valor de 1 porque se ha verificado que hay acceso a la documentación dela API-Rest y que se evitan cambios bruscos en la misma para que las rutinas[55] que se hayan desarrollado haciendo uso de ella no dejen de funcionar súbitamente.

Preservación

Los resultados de este grupo de buenas prácticas, el más amplio, son los siguientes:

	media	mediana	V(BP)_max	ICBP
BP27	1	1	1	1
BP28	0	0	1	0

Tabla LVII: Resultados ICBP en las buenas prácticas vinculadas con la preservación de los datos.

La **BP27** verifica que se conservan los identificadores y que no se reutilizan, aunque se retire de la publicación un conjunto de datos. Su valor es siempre positivo y el índice ICBP es 1. El caso de la **BP28** es uno de los particulares de nuestro análisis porque, a pesar de saber de antemano que **Zenodo** no hace uso del metadato 'coverage', mantenemos ese requisito como obligatorio porque no aceptamos que no emplee para informar la fuente de los datos. Estamos convencidos de que su uso sería una mejora importante. Al no disponer tampoco de metadatos recomendables, esta DWBP vale 0 en todos los casos.

Retroalimentación

Los resultados de este grupo de buenas prácticas, dedicadas a medir el contacto con los usuarios de los conjuntos de datos, son los siguientes:

	media	mediana	V(BP)_max	ICBP
BP29	1,1	1,1	1,1	1
BP30	0	0	0	**indeterminado**

Tabla LVIII: Resultados ICBP en las buenas prácticas de retroalimentación.

Es posible contactar con los editores de los datos gracias a los metadatos 'owners' y 'creators'. **Zenodo** no ofrece la posibilidad de publicar las opiniones de los usuarios, tal como exige la **BP30**. Esta práctica se calcula de forma semiautomática con un valor nulo.

Enriquecimiento

Los resultados de las buenas prácticas que verifican el enriquecimiento de los datos, son:

	media	mediana	V(BP)_max	ICBP
BP31	0,0141	0	0,2	0,07
BP32	0,015	0	0,4	0,03

Tabla LVIX: Resultados del ICBP en las buenas prácticas de retroalimentación.

[55] Todos los 'scripts' desarrollados en esta investigación hacen uso de esa API-REST.

En ambos casos no se dispone de metadatos para satisfacer el requisito obligatorio. Se recurre a un conjunto de claves del tipo 'isDerivedFrom', 'isDescribedby', 'isSupplementTo', etc., apenas presentes en los conjuntos de datos. Por ello, los resultados son muy bajos. Sugerir a los investigadores aumentar el uso de estas claves de metadatos mejoraría el nivel de cumplimiento de estas buenas prácticas y mejoraría la calidad de los conjuntos de datos.

Republicación

Los resultados de este grupo de buenas prácticas son los siguientes:

	Media	mediana	V(BP)_max	ICBP
BP33	1,1	1,1	1,1	1
BP34	1,0032	1,1	1,1	0,91
BP35	0,018	0	0,6	0,03

Tabla LX: Resultados ICBP en las buenas prácticas de republicación.

En las dos primeras buenas prácticas revisadas se detectan valores altos de cumplimiento. Respecto a la **BP33**, en todos los conjuntos de datos están presentes los metadatos 'owners' y 'creators' y, tal como ocurría en la **BP29**, se alcanza el nivel máximo del índice ICBP. En cuanto a la **BP34**, su nivel de cumplimiento es casi total (índice ICBP de 0,91). Por el contrario, el nivel de cumplimiento de la **BP35** es muy escaso al no disponer Zenodo de un metadato que permita satisfacer el requisito obligatorio y porque la presencia de las claves de metadatos recomendables ('isBasedOn', 'isPartOf', 'references', 'cites', 'isDerivedFrom' y 'isSupplementTo'} dentro del metadato 'related_identifiers' es muy residual.

El siguiente diagrama de barras agrupadas (media y valor máximo posible de cada buena práctica) y línea (el valor del índice ICBP) resume gráficamente los datos obtenidos.

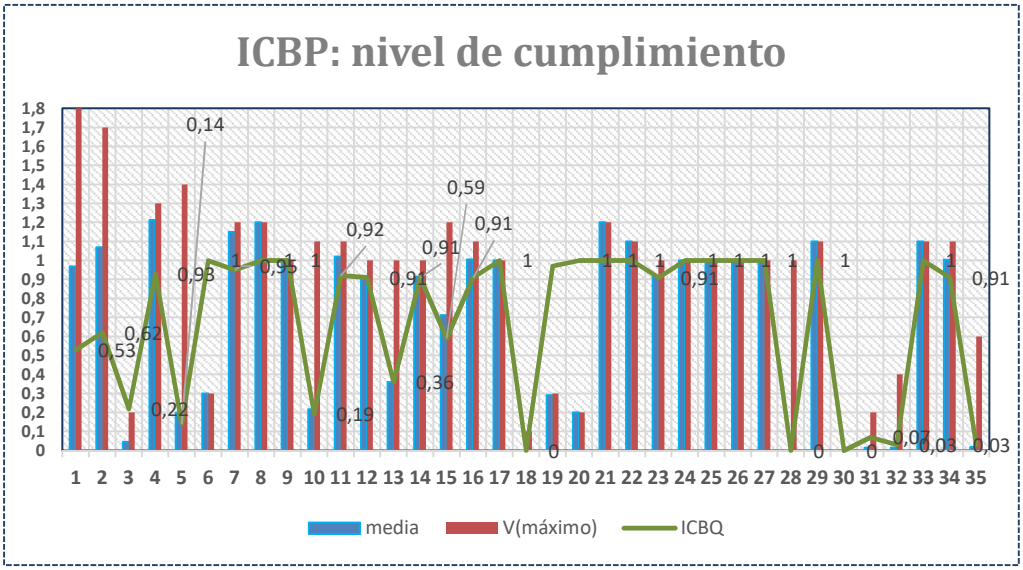


Imagen 3: Diagrama del cumplimiento de las DWBP a partir del ICBP. Fuente: elaboración propia.

Dendograma.

La siguiente figura muestra los agrupamientos más directos de las 35 buenas prácticas DWBP según el método del **vecino más cercano** y con los datos de nuestro experimento.

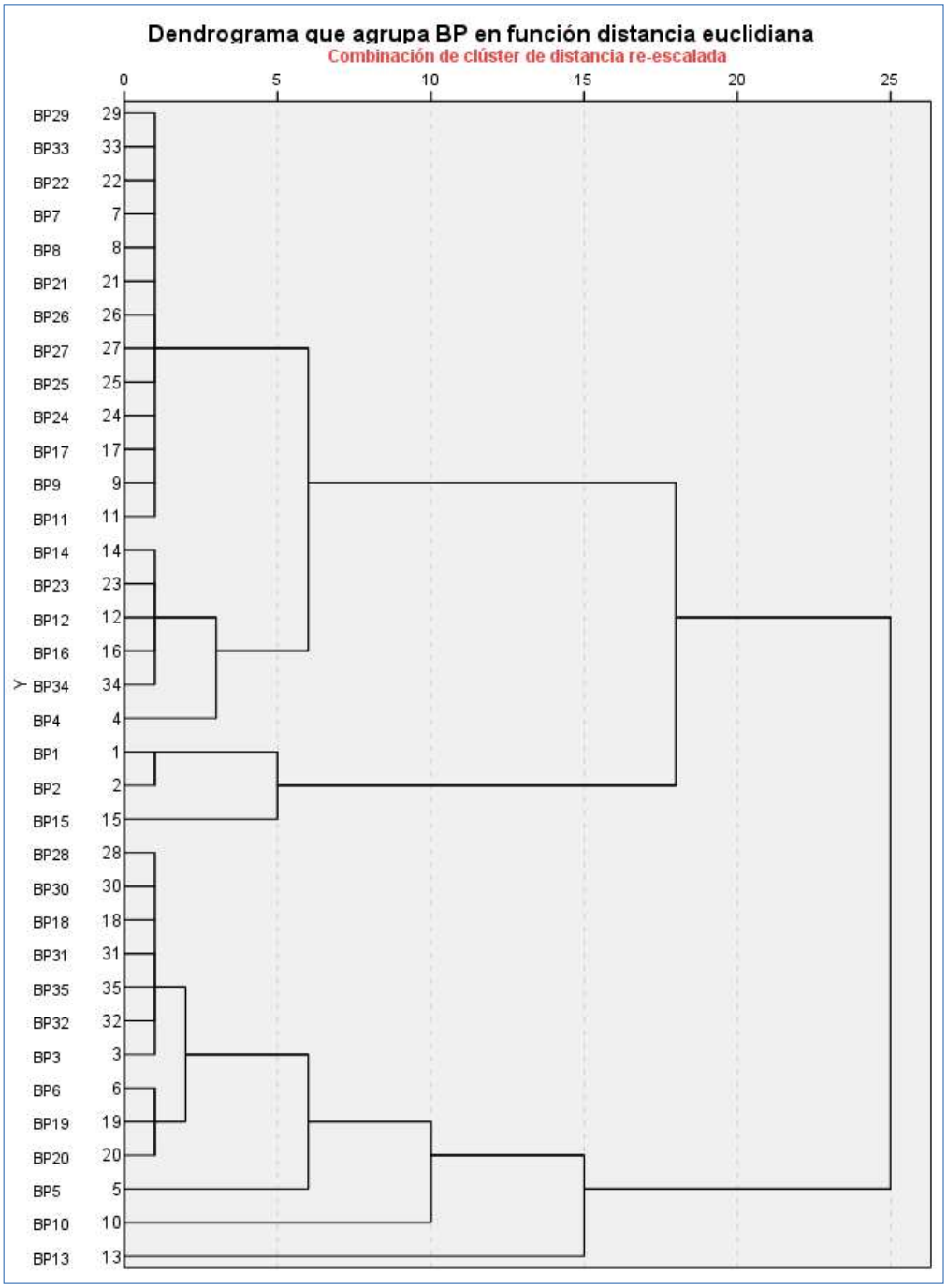


Imagen 4: Dendograma elaborado con los valores medios de cumplimiento de las DWBP en el experimento. Fuente: elaboración propia.

Se han identificado dos grandes grupos de buenas prácticas: en el primero están aquellas cuya media de cumplimiento es cercano o mayor que 1 y, en el segundo, están aquellas DWBP cuya media de cumplimiento no alcanza la unidad, bien porque en los criterios de valoración esto no es posible o bien porque no se satisface, por regla general, el requisito obligatorio.

Iteración	BP agrupadas
1	29, 33, 22, 7, 8, 21, 26, 27, 25, 24, 17, 9, 11
2	29, 33, 22, 7, 8, 21, 26, 27, 25, 24, 17, 9, 11, 14, 23, 12, 16, 34, 4
3	29, 33, 22, 7, 8, 21, 26, 27, 25, 24, 17, 9, 11, 14, 23, 12, 16, 34, 4, 1, 2 , 15
Final	29, 33, 22, 7, 8, 21, 26, 27, 25, 24, 17, 9, 11, 14, 23, 12, 16, 34, 4, 1, 2 , 15, 28, 30, 18, 31. 35, 32, 3, 6, 19, 20, 5, 10, 13
5	28, 30, 18, 31. 35, 32, 3, 6, 19, 20, 5, 10, 13
4	28, 30, 18, 31. 35, 32, 3, 6, 19, 20, 5, 10
3	28, 30, 18, 31. 35, 32, 3, 6, 19, 20, 5
2	28, 30, 18, 31. 35, 32, 3, 6, 19, 20
1	28, 30, 18, 31. 35, 32, 3

Tabla LXI: evolución del agrupamiento por el método del vecino más cercano.
Fuente: elaboración propia.

	Cumplimiento ICBP		Dendograma	
	Bueno	Deficiente	Grupo 1	Grupo 2
Metadatos	1, 2	3	1, 2	3
Licencias	4		4	
Procedencia		5		5
Calidad	6			6
Versionado	7, 8		7, 8	
Identificadores	9, 11	10	9, 11	10
Formatos	12, 14	13	12, 14	13
Vocabularios	15. 16		15, 16	
Acceso	17, 19, 20, 21, 22, 23, 24, 25, 26	18	17, 21. 22, 23. 2, 25, 26, 27	18, 19, 20
Preservación	27	28	27	28
Enriquecimiento		31, 32		31, 32
Retroalimentación	29	30	29	30
Republicación	33, 34, 35		33, 34	35

Tabla LXII: resumen de resultados comparado: ICBP vs posición en dendograma.
Fuente: elaboración propia.

Se resaltan los grupos de buenas prácticas donde se detectan divergencias entre la distribución de "mayores / menores" valores en el índice ICBP y la pertenencia al grupo del dendograma en función del "mayor / menor" valores de la media de cada buena práctica. Son los grupos de calidad, acceso y republicación. En el primero, la **BP6** sí tiene un valor alto de ICBP, pero, en realidad, aporta poca información (solo informan metadatos recomendables, no tiene metadatos para satisfacer el requisito obligatorio); en cuanto al acceso, el que más buenas prácticas tiene asociado, ocurre algo similar con las buenas prácticas **BP19** y **BP20**; por último, en el grupo de la posibilidad de republicación de la información, ocurre lo mismo con la última buena práctica, la **BP35**. Solo los grupos de buenas prácticas sobre licencias, versionado y vocabularios muestran resultados totalmente positivos en cada una de las distribuciones y únicamente el grupo relacionado con el enriquecimiento de los datos muestra valores negativos en ambos casos.

Obtención de beneficios en función del cumplimiento delas buenas prácticas.

Beneficio	Buenas prácticas
Accesibilidad	BP17, **BP18**, BP19, BP20, BP21, BP23, BP24, BP32
Comprensión	BP1, BP2, **BP3**. **BP13** , BP15, BP16, BP29, BP31, BP32, BP33
Conectividad	BP9, **BP10**, **BP18**, BP24
Confianza	BP4, **BP5**, BP6, BP7, BP8, BP11, BP15, BP22, BP25, BP26, BP27, **BP28**, BP29, **BP30**, **BP31**, **BP32**, BP34, BP35
Descubribilidad	BP1, BP2, BP9, **BP10**, BP11, BP24, BP35
Interoperabilidad	BP9, **BP10**, BP15, BP16, BP23, BP24, BP26, BP33
Procesabilidad	BP1, **BP3**, BP12, BP14, BP15, **BP18**, BP23, BP24, **BP31**
Reusabilidad	BP1, BP2, **BP3**, BP4, **BP5**, BP6, BP7, BP8, BP9, **BP10**, BP11, BP12, **BP13**, BP14, BP15, BP16, BP17, **BP18**, BP19, BP20, BP21, BP22, BP23, BP24, BP25, BP26, BP27, **BP28**, BP29, **BP30**, **BP31**, **BP32**, BP33, BP34, BP35

Tabla LXIII Clasificación de las DWBP según beneficio asociado a su uso en función del índice ICBP.
Fuente: elaboración propia a partir de Teixeira dos Santos (2023, 26-27).

Beneficio	Buenas prácticas
Accesibilidad	BP17, **BP18**, **BP19**, **BP20**, BP21, BP23, BP24, BP32
Comprensión	BP1, BP2, **BP3**. **BP13** , BP15, BP16, BP29, **BP31**, **BP32**, BP33
Conectividad	BP9, **BP10**, **BP18**, BP24
Confianza	BP4, **BP5**, **BP6**, BP7, BP8, BP11, BP15, BP22, BP25, BP26, BP27, **BP28**, BP29, **BP30**, **BP31**, **BP32**, BP34, **BP35**
Descubribilidad	BP1, BP2, BP9, **BP10**, BP11, BP24, **BP35**
Interoperabilidad	BP9, **BP10**, BP15, BP16, BP23, BP24, BP26, BP33
Procesabilidad	BP1, **BP3**, BP12, BP14, BP15, **BP18**, BP23, BP24, **BP31**
Reusabilidad	BP1, BP2, **BP3**, BP4, **BP5**, **BP6**, BP7, BP8, BP9, **BP10**, BP11, BP12, **BP13**, BP14, BP15, BP16, BP17, **BP18**, **BP19**, **BP20**, BP21, BP22, BP23, BP24, BP25, BP26, BP27, **BP28**, BP29, **BP30**, **BP31**, **BP32**, BP33, BP34, **BP35**

Tabla LXIV: Clasificación de las DWBP según beneficio asociado a su uso.
Fuente: elaboración propia a partir de Teixeira dos Santos (2023, 26-27).

En las tablas **LXIII** y **LXIV** intentamos reflejar qué clase de beneficios de la aplicación de las DWBP se encuentran afectados por valores negativos del índice ICBP y de los valores medios de cumplimiento de cada buena práctica. En ambos casos es la reusabilidad el más afectado, algo que lógico porque la mayoría de las prácticas se vinculan a este beneficio y, por lógica, la mayoría de las que no se satisfagan estarán, obviamente, asociadas al mismo. En el lado positivo, la descubribilidad y la interoperabilidad son los beneficios donde más buenas prácticas se satisfacen. La accesibilidad, estrechamente relacionada con las anteriores, está al mismo nivel de poca afectación si se analiza solo el ICBP.

Nivel cualitativo de cumplimiento: resultados.

De la revisión del nivel de cumplimiento de las 35 buenas prácticas en los 1000 conjuntos de datos analizados, se han obtenido los siguientes datos:

	Media	Mediana	Mínimo	Máximo
V(BP) >= 1	20,65	22	13	25
0 < V(BP) < 1	6,06	7	3	13
V(BP) = 0	8,30	9	3	14

Tabla LXV: Resultados cualitativos del cumplimiento de las buenas prácticas por variable.

En cuanto al número de variables que satisfacen el requisito obligatorio y, además, cumplen alguno de los recomendables, la media es igual a 20,65 y la mediana tiene un valor igual a 22. En un principio pueden parecer porcentajes moderadamente positivos. Si tomamos la mediana de referencia, la misma posee valores superiores o iguales a 1 en 22 de las 35 buenas prácticas (un 62,8% de los conjuntos de datos). En realidad, esto equivale a un porcentaje más significativo si se tiene en cuenta que hasta en 8 buenas prácticas no se dispone de elemento de metadatos alguno para satisfacer el requisito obligatorio y, además, en una de ellas (la **BP28** que precisa del metadato 'coverage') sabemos de antemano que el resultado va a ser nulo.

Tomando esto en consideración, el valor de mediana debería ponderarse en relación a un máximo posible de 24 buenas prácticas. Bajo esta perspectiva, el porcentaje de nivel de completitud en el cumplimiento de las buenas prácticas sería del 91,6%, muy positivo, por tanto. En cuanto a las variables cuyos promedios se sitúan entre 0 y 1, la mediana tiene un valor de 7, lo que significa que casi la totalidad de variables que solo pueden aportar información por medio de metadatos recomendables así lo hacen (7 de 8 equivale a un 87,5%). Finalmente, si bien partimos de la base de la imposibilidad de cumplir nunca la BP28 anteriormente citada y sabemos que la **BP30** tiene el valor nulo (es de las variables "semiautomáticas" cuyo valor conocemos de antemano), consideramos que una mediana de 9 buenas prácticas no satisfechas sea 9 es un valor alto, sobre el que se debe reflexionar.

Conclusiones de la investigación, futuros estudios y transferencia de resultados

De todo este proyecto investigador se puede concluir, a modo general, lo siguiente:

1. La gestión de los conjuntos de datos de investigación en los repositorios es gestionar datos en la web y, por tanto, se puede y se debe seguir los estándares de buenas prácticas establecidos para ello. Eso es algo absolutamente irrenunciable si se quieren cumplir los principios FAIR, aspecto que representa el pilar de la Ciencia Abierta.
2. Si esa gestión se lleva a cabo, de forma total o parcial, en las bibliotecas y servicios de documentación de las instituciones académicas, unidades dotadas de profesionales acostumbrados a aplicar normas y estándares internacionales en la descripción de documentos, la misma se puede convertir en una oportunidad en doble sentido: propiciar la mejora de la calidad continua de la descripción de esos conjuntos de datos (y potenciar así su futura reutilización) y, en segundo lugar, puede representar una reivindicación del papel a desarrollar por estos profesionales en el seno de sus instituciones, adoptando un rol emergente lejos de estereotipos y obsolescencias.
3. La aplicación de buenas prácticas en general, y de las DWBP del W3C en particular, puede evaluarse de forma semiautomática tal como ha quedado demostrado en nuestro estudio y posterior experimentación.
4. El alcance de esta evaluación puede abarcar total o parcialmente al conjunto de las DWBP gracias a la abundante presencia de metadatos en la descripción de los conjuntos de datos de investigación.

Si a continuación nos centramos de forma más específica en el experimento desarrollado, se obtiene el siguiente conjunto de conclusiones adicionales:

1. **Zenodo** es un repositorio de datos de investigación con un alto nivel de desarrollo y firmemente asentado sobre la base del cumplimiento de los principios FAIR, lo que le convierte en un sitio muy adecuado para publicar en abierto los resultados de la investigación.
2. A pesar de esta idoneidad, se han detectado algunas deficiencias en cuanto a ausencias de algunos metadatos que podrían ser objeto de posibles acciones de mejora. No obstante, el porcentaje de buenas prácticas de gestión de datos en la web verificadas de forma positiva supera ampliamente al de aquellas no tan bien valoradas.
3. La no utilización del metadato 'source' provoca que alguna buena práctica no pueda haber sido analizada debidamente, obviando la presencia del mismo entre los requisitos de evaluación. En el caso de los conjuntos de datos de investigación, especialmente aquellos que son fruto de encuestas o revisiones de bibliografía, esta ausencia provoca que se deja de informar de un aspecto importante: la naturaleza y objetivo de la investigación en general, y de los conjuntos de datos en particular.

4. Algo parecido ocurre con el metadato 'coverage' que informa de la cobertura de la investigación en general y de cada conjunto de datos en particular. Lo cierto es que su empleo parece una oportunidad de mejora muy oportuna y necesaria.
5. En nuestro experimento ha sorprendido el nivel intermedio de uso del metadato 'keywords' (que en este repositorio desempeña una doble función: informar de la materia del conjunto de datos y de las palabras clave que lo describen). Esto ha lastrado los resultados de las buenas prácticas que analizan el aporte de metadatos generales y descriptivos. Creemos que, de confirmarse esta situación en otros experimentos, el uso de este metadato debería ser más frecuente y mejoraría los resultados en alguna de las buenas prácticas más importantes.
6. Se podría mejorar la metadescripción de los conjuntos de datos de investigación incorporando un metadato (o clave de metadato dentro de otro más amplio) similar a 'conformsTo' o 'scheme' que sí están presentes en algunos otros esquemas de metadatos empleados en el Portal de Datos Europeos Abiertos. Esto aportaría información sobre la estructura de los ficheros que forman los conjuntos de datos y mejoraría el cumplimiento de algunas buenas prácticas (las relacionadas con el enriquecimiento y la republicación de los datos), junto con un uso mayor de las claves de metadatos del tipo 'isBAsedOf' o 'isPartOf' que muestran las relaciones entre distintos conjuntos de datos o resultados de otras de investigaciones
7. **Zenodo** no está concebido como un repositorio de conjuntos de datos accesibles en tiempo real. Por esa razón, nunca va a satisfacer el requisito obligatorio indicado en la **BP30,** se está exigiendo algo no previsto en la metáfora global de todo el sistema. Algo parecido ocurre con la buena práctica que considera el uso de subconjuntos de datos dentro de los conjuntos de datos y exige verificar si la unión de todos estos subconjuntos no pierde información alguna. Estas buenas prácticas están situadas en un plano distinto del habitual en repositorios de datos de investigación y su presencia lastra el resultado de algunas buenas prácticas. Se debería reflexionar sobre la trascendencia de su verificación.
8. A pesar de los casos particulares, las buenas prácticas de identificación de los conjuntos de datos y de acceso a los mismo alcanzan unos niveles de cumplimiento muy elevados en su inmensa mayoría, esto es positivo para el cumplimiento de los principios FAIR. La descubribilidad, la interoperabilidad y la accesibilidad están de los conjuntos de datos están aseguradas en virtud del nivel de cumplimiento de las buenas prácticas vinculadas.
9. Finalmente, no podemos dejar de resaltar el hecho de que el porcentaje de nivel de completitud en el cumplimiento de las buenas prácticas sería del 91,6% si lo calculamos en referencia con el total de aquellas que puedan satisfacer el requisito recomendable. Esto es un hecho muy positivo.

A partir de los resultados del estudio y de las conclusiones extraídas del mismo, surgen distintas **vías futuras de investigación**:

1. Creemos necesario contrastar el experimento llevado a cabo con muestras de conjuntos de datos de investigación (de similar y de mayor volumen), con el objeto de verificar algunos de los resultados obtenidos, en especial el del relativamente escaso uso del metadato 'keywords'.
2. Las DWBP son un conjunto muy amplio de buenas prácticas, algunas de las cuales desbordan el alcance y propósito de un repositorio de datos de investigación. En definitiva, poseen un nivel de exigencia elevado. Podría ser interesante incluir en el prototipo 'esposende 1.0' el análisis del cumplimiento de otros conjuntos de buenas prácticas como las elaboradas por la 'Research Data Alliance' (o similares) para complementar la valoración que hemos llevado a cabo con la aplicación de las DWBP. Sería otra acción de mejora interesante.
3. El prototipo ha demostrado ser capaz de evaluar el cumplimiento de buenas prácticas en un repositorio robusto y avanzado como es **Zenodo**, pero no debemos olvidar la presencia de una cantidad considerable de conjuntos de datos de investigación publicados en otros repositorios basados en las plataformas 'dataverse' y 'DSpace' (de uso muy extendido en universidades y centros de investigación). Otra línea de trabajo necesaria es de adaptar el evaluador a estos entornos.
4. El prototipo de analizador, en su actual estado de desarrollo, puede servir para el desarrollo de una línea de investigación paralela: verificar si las consultas y descargas de los conjuntos de datos de investigación están dando lugar a referencias a los mismos en nuevos artículos y proyectos de investigación, certificando el principal de los objetivos esenciales de los principios FAIR.
5. El analizador 'esposende 1.0' ha demostrado ser una herramienta eficaz como analizador del cumplimiento de las DWBP. Una posible acción de mejora podría ser incorporar un módulo "asesor" que informase a los autores de los conjuntos de datos de las mejoras que deben llevar a cabo en sus descripciones para mejorar el cumplimiento de los principios FAIR.

La **transferencia de los resultados** de esta investigación se manifiesta de distintas maneras.

1. El software del prototipo desarrollado (y otros 'scripts' vinculados a la investigación), junto con toda la documentación del proyecto, se va a publicar en repositorios de acceso abierto, GitHub y Zenodo entre ellos. También se publicarán los artículos científicos y otra producción derivada del mismo.
2. El prototipo de analizador de los metadatos de miles de conjuntos de datos de investigación puede convertirse en un evaluador de la calidad de la metadescripción de un único conjunto de datos de investigación (a modo de test de accesibilidad web como OAW o Wave, por ejemplo) que podríamos denominar '**miesposende**' y que sería un portal online donde los usuarios remitirían los identificadores permanentes

de sus conjuntos de datos de investigación con el objeto de testear si los mismos satisfacen las buenas prácticas y proceder, de ser necesario, a introducir las mejoras oportunas.

3. Finalmente, se pueden organizar una serie de 'webinars' informativos de la necesidad de cumplir las buenas prácticas de gestión de datos en la web y de la oportunidad de hacer uso de estos prototipos con el objeto de mejorar el cumplimiento de los principios FAIR de todos estos resultados de investigación. Esta formación y divulgación podría canalizarse a través de la Cátedra UNESCO en Gestión de Información en las Organizaciones y enmarcarse dentro de las actividades de formación indicadas por la Estrategia Nacional de Ciencia Abierta.

Imagen 5: Atardecer en la ría de Esposende

Referencias

Abadal, E. et al. (2023). *Ciencia Abierta en España 2023: informe de situación y análisis de la percepción*. http://hdl.handle.net/2445/200020

Alcalá, M., & Anglada, L. (2019). *FAIR x FAIR: Requisitos factibles, alcanzables e implementables para un repositorio de datos de investigación FAIR*. https://www.recercat.cat/bitstream/handle/2072/356460/InformeFxF_maquetada_ESP.pdf

Alexandre-Benavent, R., Ferrer Sapena, A. y Peset, F. (2019). Compartir los recursos útiles para la investigación: datos abiertos (open data). *Educación Médica*, Aug. PMCID: PMC7148709. https://europepmc.org/article/pmc/pmc7148709 (02-10-2023)

Alonso-Arévalo, J. (2019). La gestión de datos de investigación en el horizonte de las bibliotecas universitarias y de investigación. *Cuadernos de Documentación Multimedia*, 30, 75-88. https://doi.org/10.5209/CDMU.62806

Angelozzi, S. M. (2020). La gestión de datos de investigación en abierto: introducción al rol emergente para las bibliotecas universitarias y científicas argentinas. *Palabra clave*, 9(2), e091. https://doi.org/10.24215/18539912e091

Ashiq, M., Usmani, M. H., & Naeem, M. (2022). A systematic literature review on research data management practices and services. Global Knowledge, Memory and Communication, 71(8/9), 649-671. https://doi.org/10.1108/GKMC-07-2020-0103

Ayris, P., & Ignat, T. (2018). Defining the role of libraries in the Open Science landscape: a reflection on cur-rent European practice. Open Information Science, 2(1), 1-22. https://doi.org/10.1515/opis-2018-0001

Bethencourt-Aguilar, A., Castellanos-Nieves, D., Sosa-Alonso, J. J., & Area-Moreira, M. (2022). Implicaciones técnicas y prácticas de las Redes Adversarias Generativas a la Ciencia Abierta en Educación. *RiiTE ,* 138-156. https://doi.org/10.6018/riite.545881

Borghi, J.A., & Van Gulick, A.E. (2021). Promoting Open Science through research data management. *arXiv preprint* https://arxiv.org/abs/2110.00888

Comisión Europea. (2014). *Guía del participante Horizonte 2020*. https://www.horizonteeuropa.es/sites/default/files/inline-files/guia-del-participante-h2020.pdf (12-11-2023)

Comisión Europea. (2020). *Estrategia europea de datos: hacer de la UE un modelo de sociedad capacitada por los datos*. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_es (08-11-2023)

Comisión Europea. (2022). *Guía del participante Horizonte Europa*. https://www.horizonteeuropa.es/sites/default/files/noticias/Gu%C3%ADa%20del%20participante%20-%20Horizonte%20Europa%20web_0.pdf (12-11-2023)

Comisión Europea. (2023). *Políticas de acceso abierto en América Latina, el Caribe y la Unión Europea: avances para un diálogo político*. https://data.europa.eu/doi/10.2777/162

De Giusti, M.R. (2021) Calidad en los repositorios digitales: los principios TRUST para repositorios de datos. *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*, 29, 55-59. https://doi.org/10.24215/18509959.29.e6

ESPAÑA. (2022). Ley 17/2022, de 5 de septiembre, por la que se modifica la Ley 14/2011, de 1 de junio, de la Ciencia, la Tecnología y la Innovación. https://www.boe.es/eli/es/l/2022/09/05/17/con

ESPAÑA. (2023). Estrategia Nacional de Ciencia Abierta (ENCA). https://www.ciencia.gob.es/InfoGeneralPortal/documento/c30b29d7-abac-4b31-9156-809927b5ee49 (12-11-2023)

FAIR Data Maturity Model Working Group. (2020). *FAIR Data Maturity Model. Specification and Guidelines (1.0)*. https://doi.org/10.15497/rda00050

Federer, L. M., & Qin, J. (2019). Beyond the data management plan: Expanding roles for librarians in data science and open science. *Proceedings of the AIST*, 56(1), 529-531. https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.82

Ferrer-Sapena, A.; Calabuig, J.M.; Peset, F. y Sánchez-del-Toro, I. (2020). Trabajar con datos abiertos en tiempos de pandemia: uso de covidDATA-19. *El Profesional de la información*, v.29, n. 4, e290421. https://doi.org/10.3145/epi.2020.jul.21

Fundación CYD. (2023). La investigación en España: aumenta el gasto y la producción científica, pero con recursos inferiores a países de su entorno. *Informe CYD*. https://www.fundacioncyd.org/la-investigacion-en-espana-aumenta-el-gasto-y-la-produccion-cientifica-pero-con-recursos-inferiores-a-paises-de-su-entorno/ (08-11-2023)

Garijo, D. & Poveda-Villalón, M. (2020). *Best practices for implementing FAIRr vocabularies and ontologies on the web*. https://arxiv.org/pdf/2003.13084.pdf

Johnston, L.; Carlson, J.; Hswe, P.; Hudson-Vitale, C.; Imker, H.; Kozlowski, W.; Olendorf, R. and Stewart, C. (2017). Data Curation Network: How Do We Com-pare? A Snapshot of Six Academic Library Institutions, Data Repository and Curation Services. *Journal of eScience Librarianship*, 6 (1). e1102. https://doi.org/10.7191/jeslib.2017.1102

Johnston, L.; Carlson, J.; Hswe, P.; Hudson-Vitale, C.; Imker, H.; Kozlowski, W.; Olendorf, R.; Stewart, C.; Blake, M.; Herndon, J.; McGeary, T. and Hull, E. (2018). Data Curation Network: A Cross-Institutional Staffing Model for Curating Research Data. *International Journal of Digital Curation*, vol.13, (1). http://www.ijdc.net/article/view/616 (08-11-2023)

Koster, L. & Woutersen-Windhouwer, S. (2018). FAIR Principles for Library, Archive and Museum Collections: A proposal for standards for reusable collections. *Code4Lib Journal, (40)*. https://journal.code4lib.org/articles/13427 (16-11-2023)

López Carreño, R. y Martínez Méndez, F.J. (2020). Sistemas de recuperación de información implementa-dos a partir de CORD-19: herramientas clave en la gestión de la información sobre COVID-19. Revista Española de Documentación Científica, 43 (4). https://doi.org/10.3989/redc.2020.4.1794

L'Hours, H., Verburg, M. L., de Vries, J., Cepinskas, L., von Stein, I., Huber, R. & Mathers, B. J. (2022). *D4. 6 Report on a maturity model towards FAIR data in FAIR repositories*. https://zenodo.org/record/6699520

Lóscio, B. F., Burle, C., & Calegari, N. (Eds.). (2017). *Data on the Web Best Practices*. https://www.w3.org/TR/dwbp/ (08-11-2023)

Marín-Arraiza, P., Puerta-Díaz, M. y Gregorio-Vidotti, S. (2019). Gestión de datos de investigación y bibliotecas: preservando los nuevos bienes científicos. *Hipertext.net*, (19), 13-31. https://doi.org/10.31009/hipertext.net.2019.i19.02

Martínez Méndez. F.J.; López Carreño, R., Vázquez Delgado, A. y Baptista, A. (2023) Implementación de los repositorios de datos de investigación en las universidades públicas españolas: estado de la cuestión. *Ibersid: revista de sistemas de información y documentación*, Vol.29, 2. https://www.ibersid.eu/ojs/index.php/scire/294/.

Méndez Rodríguez, E. M. (2007). Dublin Core, metadatos y vocabularios. 2007. *ThinkEPI,* 2007, 61-64. https://e-archivo.uc3m.es/bitstream/handle/10016/877/EMendez_DCVocs.pdf

OCDE/LEGAL/0347. (2021). *Recommendation of the Council concerning Access to Research Data from Public Funding*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0347 (07-11-2023)

Pastor-Sánchez, J. A. (2016). Buenas prácticas para la publicación de datos en la Web: La superación del paradigma Linked Open Data. *VIII Encuentros de Centros de Documentación de Arte Contemporáneo*. https://digitum.um.es/digitum/handle/10201/51181

Ramírez Castañeda, L.A. y Sepúlveda López, J.J. Brecha digital e inclusión digital: fenómenos socio – tecnológicos. *Revista de la Escuela de Ingeniería de Antioquía*, vol.15, n.30. https://doi.org/10.24050/reia.v15i30.1152

REBIUN (2017) Plantilla de metadatos para la descripción de datos de investigación. En: Estudio de los repositorios de datos, 2017. http://hdl.handle.net/20.500.11967/137

REBIUN (2018) Gestión de datos de investigación en las universidades españolas y en el CSIC: memoria de buenas prácticas de los servicios ofrecidos. http://hdl.handle.net/20.500.11967/244

Redkina, N. S. (2019). Current Trends in Research Data Management. Scientific & Technical Information Pro-cessing, 46(2), 53-58. https://doi.org/10.3103/S0147688219020035

Rocca-Serra, P. et al. The FAIR Cookbook - the essential resource for and by FAIR doers. Scientific Data 10, 292 (2023). https://doi.org/10.1038/s41597-023-02166-3

Sansone, S. A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A. L. & FAIRsharing Community. (2019). FAIRsharing as a community approach to standards, repositories and policies. *Nature biotechnology*, 37(4), 358-367. https://doi.org/10.1038/s41587-019-0080-8

Sheikh, A., Malik, A., & Adnan, R. (2023). Evolution of research data management in academic libraries: A review of the literature. Information Development, 0(0). https://doi.org/10.1177/02666669231157405

Teixeira dos Santos, C. S. (2023) *OGD Lens: avaliação automática da qualidade dos dados do European Data Portal*. [TFM] Universidade do Minho.

Torres-Salinas, D., Robinson-Garcia, N., & Castillo-Valdivieso, P. A. (2020). Open Access and Altmetrics in the pandemic age: Forescast analysis on COVID-19 literature. BioRxiv, 2020-04. https://doi.org/10.1101/2020.04.23.057307

UNESCO. (2020). Un llamamiento conjunto en pro de la Ciencia Abierta. https://es.unesco.org/sites/default/files/joint_appeal_for_open_sciences_v5_es.pdf

UNESCO (2021) Recomendación de la UNESCO sobre la Ciencia Abierta. https://unesdoc.unesco.org/ark:/48223/pf0000379949_spa (02-03-2023)

Weibel, S., & Lagoze, C. (1997). An element set for metadata description of Internet resources. *D-Lib Magazine*, 3(4). https://doi.org/10.1045/july97-weibel

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A. & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Anexo I: Plataformas y metadatos para la descripción de 'datasets'

El objetivo de este proyecto investigador es conocer el nivel de cumplimiento de DWBP en los repositorios de **las universidades españolas públicas**[56], necesitamos conocer con detalle cómo están estos implementados[57]. Este análisis se va a llevar a cabo por medio de un procedimiento semiautomático de análisis de los metadatos que podemos encontrar en los 'datasets'. Depende de la plataforma que se emplee, los esquemas de metadatos, y la aplicación de los mismos, puede variar.

Investigación preliminar.


Implementación de los repositorios de datos de investigación en las universidades públicas españolas: estado de la cuestión

Antes de este proyecto, realizamos una revisión de la implantación de los repositorios de conjuntos de datos en estas instituciones. Los resultados se recogen en el artículo: **Martínez Méndez et al**. (2023) Implementación de los repositorios de datos de investigación en las universidades públicas espa ñolas: estado de la cuestión. *Scire: representación y organización del conocimiento*, Vol. 29, 2 (https://www.ibersid.eu/ojs/index.php/scire/article/view/4914).

Plataformas de los repositorios

A nivel del software empleado para implementar los repositorios institucionales, las universidades se dividen en dos grandes grupos: aquellas las que forman parte de los consorcios autonómicos (e-Ciencia-Datos en Madrid y CORA.RDR en Cataluña), y aquellas que han desarrollado su propio repositorio para los conjuntos de datos de forma individual, normalmente como una colección de su repositorio institucional.

- Según los datos de febrero de 2023, en España hay 86 universidades: 50 públicas (47 presenciales, 1 no presencial y 2 especiales) y 36 privadas (31 presenciales y 5 no presenciales). Este número puede variar porque hay varios proyectos de creación de universidades privadas pendientes de aprobación[58]
- La UNED es la universidad pública a distancia de mayor implantación a nivel nacional y la Oberta de Catalunya es una universidad pública, también a distancia, perteneciente al sistema universitario catalán, pero con gestión privada (se va a

[56] Está previsto ampliar el alcance del estudio a los institutos y centros de investigación, también de naturaleza pública.

[57] Para el desarrollo de este experimento se van a analizar 'datasets' sobre COVID-19 elaborados por investigadores de España. Se trata de una muestra amplia y, además, cuenta con el factor de la actualidad, se han desarrollado en los últimos tres años cuando ya parece estar asumida la importancia de los conjuntos de datos.

[58] En realidad, la aprobación de estas universidades no influye en nuestro estudio por dos razones: son privadas y no serían recogidas en el mismo y, en segundo lugar, las universidades privadas realizan poca investigación (salvo alguna excepción) y es imposible que tengan 'datasets' que publicar.

considerar "pública" a efectos de nuestro estudio). Las universidades "especiales" son la Menéndez Pelayo y la Internacional de Andalucía, centros más dedicados a la organización de cursos de verano y conferencias que a la investigación (por tanto, no van a ser objeto de estudio). El total de universidades que serán revisadas, en un principio, lo forman las 47 universidades presenciales públicas más las 2 a distancia citadas anteriormente. En total serán 49 las instituciones cuyos repositorios de conjuntos de datos serán objeto de análisis.

- Las universidades de la comunidad de Madrid (menos la Complutense) y la UNED forman el Consorcio Madroño[59] para poner en marcha de un repositorio colectivo: e-Ciencia-Datos[60]. En Cataluña existe el repositorio CORA.RDR[61] puesto en marcha por el consorcio CSUC[62] y en el que participan las universidades catalanas más la Universidad de las Isla Baleares. Este segundo consorcio queda formado por ocho universidades. Con datos de febrero de 2023, estos dos repositorios almacenaban 1211 'datasets' (744 el madrileño y 467 el catalán).
- El resto de las universidades españolas, de forma individual y sin ningún otro tipo de asociación y/o coordinación (excepto los documentos de coordinación que emanan de la red de bibliotecas universitarias REBIUN), han implementado su propio repositorio de conjuntos de datos, formando parte generalmente del repositorio institucional. En el recuento de febrero de 2023, todos estos repositorios totalizaban 744 'datasets'.

En cuanto al **software** utilizado, aspecto Importante y decisivo en la implementación de los esquemas de metadatos:

- Los dos consorcios autonómicos están implementados por medio de la plataforma Dataverse.
- Por regla general, las universidades que han desarrollado individualmente el repositorio de conjuntos de datos de investigación, emplean el software DSpace. Una de ellas, las Palmas de Gran Canaria, emplea DSpace-CRIS, una distribución específica para la gestión de datos de investigación (aunque solo se han depositado 2 conjuntos de datos).
- La Universidad de A Coruña no ha desarrollado repositorio propio, sino que "redirige" a los investigadores a una comunidad propia creada en Zenodo (el repositorio de acceso abierto de propósito general desarrollado por el CERN dentro del programa comunitario OpenAIRE).
- La Universidad de Zaragoza emplea la plataforma Invenio RDM para implementar su repositorio. Se trata del desarrollo matriz de Zenodo.

[59] Más información de este consorcio en http://www.consorciomadrono.es/
[60] La URL para acceder a este repositorio es https://edatos.consorciomadrono.es/
[61] La URL para acceder a la consulta de este repositorio es https://dataverse.csuc.cat/
[62] Siglas de "*Consorci de Serveis Universitaris de Catalunya*". Más información en https://www.csuc.cat/

- La Universidad Complutense de Madrid implementa el software 'eprints' alojando en su repositorio un número pequeño de 'datasets' para ser la universidad pública presencial más grande de España.
- La Universidad de la Laguna, emplea tecnología de pago (Digital Commons de *Elsevier*), por lo que ha quedado excluida del estudio.

En el siguiente gráfico se puede ver la distribución de 'datasets' por software empleado:

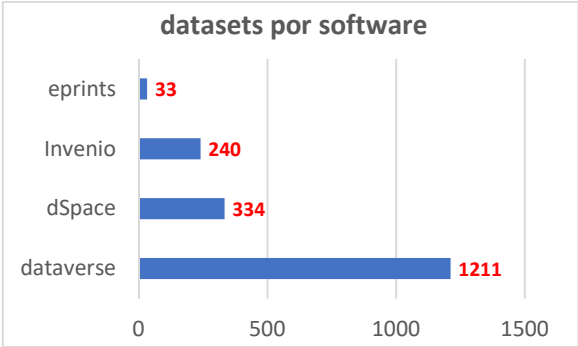


Imagen 6: #datasets por plataforma software en España

Como se ha mencionado antes, hay también 'datasets' elaborados por investigadores de universidades españolas que han sido registrados directamente en Zenodo, bien por indicación de sus propias bibliotecas (la Universidad de A Coruña, por ejemplo) o bien para cumplir lo estipulado en los programas de financiación de la investigación comunitarios.

En las universidades que emplean DSpace y dataverse (por pertenecer a uno de los consorcios), el esquema de metadatos más utilizado es Dublin Core que viene a ser, de facto, el estándar que implementan ambas plataformas (aunque no es el único porque permiten implementar otros esquemas de metadatos de menor uso, como es el caso de Open Graph[63]). El repositorio Zenodo si bien implementa un esquema de metadatos propio (del mismo nombre) basado en el esquema Datacite al que aporta metadatos específicos, también hace uso frecuente del esquema de metadatos Dublin Core.

Para poder revisar de forma automática un 'dataset' independientemente del esquema de metadatos elegido para la descripción hay que llegar a un "punto común". Se puede pensar en establecer una equivalencia entre estos dos esquemas: Dublin Core y Zenodo. La casi totalidad de las bibliotecas universitarias españolas describen sus conjuntos de datos en Dublin Core porque es el esquema más empleado en las plataformas DSpace y dataverse que agrupan a la casi totalidad de estas instituciones (bien de forma individual, bien en consorcio).

[63] Son una serie de etiquetas que se agregan a las páginas web para controlar cómo se muestra la información en las redes sociales. Estas etiquetas definen elementos como el título, la descripción, la imagen y el tipo de contenido de una página web, lo que permite a los motores de búsqueda y a las redes sociales presentar la información de una manera más útil y atractiva para los usuarios. Resultan de gran importancia para el SEO y el marketing en redes sociales, ya que pueden influir en la cantidad de tráfico y la visibilidad de una página web. Más información en https://ogp.me/

Zenodo es usado por la comunidad de investigadores de la Universidad de A Coruña que "redirigen" el depósito al repositorio de la Unión Europea y también hay pequeñas comunidades de bibliotecas universitarias españolas en este portal. Además, no debemos de olvidar que el propósito de Zenodo viene claramente determinado en los marcos de los programas de financiación comunitarios: **depositar en el mismo los resultados de la investigación**, de los que forman parte, de forma inexcusable, en aplicación de los principios FAIR, **los conjuntos de datos**.

Anexo II: lee-frecuencias-metadatos.py

Este código fuente corresponde al 'script' que lee un fichero de entrada con los identificadores ('dois') de conjuntos de datos de investigación de Zenodo, lee el contenido de la descripción de cada uno de ellos, extrae las claves de metadatos y calcula la frecuencia de aparición de cada una de ellas.

```
#=============================================================================#
#============================                    ============================#
#============================ LEE FRECUENCIA METADATOS ============================#
#============================                    ============================#
#=============================================================================#

pac requests
import csv

# Nombre del archivo CSV de salida
output_file = "frecuencias.csv"

# Solicita al usuario el nombre del archivo que contiene la lista de DOIs
doi_file = input("Ingrese el nombre del archivo con la lista de DOIs: ")

# Lee la lista de DOIs desde el archivo proporcionado
with open(doi_file, "r") as file:
    doi_list = file.read().splitlines()

# Crea un diccionario para almacenar la frecuencia de aparición de todos los metadatos
metadata_frequency = {}

# Clave de acceso a Zenodo
zenodo_token = "nVyvkXM3ezd5bKA0hesFLc3uMKMgfRenCxA3ZxkfusTAm0ZVzcvr3wZxE2k0"

# Realiza solicitudes a la API de Zenodo para cada DOI en la lista
for doi in doi_list:
    response = requests.get(
        f"https://zenodo.org/api/records/{doi}",
        headers={"Authorization": f"Bearer {zenodo_token}"},
        timeout=500
    )

    if response.status_code == 200:
        record = response.json()

        # Recorre todos los metadatos en el registro y calcula sus frecuencias
        for key, value in record.items():
            # Si es un diccionario, recorre sus elementos internos
            if isinstance(value, dict):
                for sub_key, sub_value in value.items():
                    # Incrementa la frecuencia de este metadato en el diccionario
                    if sub_key in metadata_frequency:
                        metadata_frequency[sub_key] += 1
                    else:
                        metadata_frequency[sub_key] = 1
            else:
                # Incrementa la frecuencia de este metadato en el diccionario
```

```
            if key in metadata_frequency:
               metadata_frequency[key] += 1
            else:
               metadata_frequency[key] = 1

# Ordena el diccionario por clave (nombre de metadato) alfabéticamente
sorted_metadata = sorted(metadata_frequency.items(), key=lambda x: x[0])

# Escribe los resultados en el archivo CSV
with open(output_file, "w", newline="", encoding="utf-8") as csvfile:
   fieldnames = ["Metadato", "Frecuencia"]
   writer = csv.DictWriter(csvfile, fieldnames=fieldnames)
   writer.writeheader()

   for metadata_key, frequency in sorted_metadata:
      writer.writerow({"Metadato": metadata_key, "Frecuencia": frequency})

# Imprime en pantalla la lista de metadatos y sus frecuencias
for metadata_key, frequency in sorted_metadata:
   print(f"{metadata_key}, {frequency}")
```

Anexo III: recupera-datasets.py

```
# ================================================  ==================#
# ==========================                        ====================#
# ========================== LEE FRECUENCIA METADATOS  ==================#
# ==========================                        ====================#
# ================================================  ==================#
import csv
import requests

# Parámetros de búsqueda
# Número de resultados deseados
num_results = 100

# Solicitar al usuario la palabra de búsqueda
query = input("Por favor, ingrese la palabra de búsqueda: ")

# URL de la API de Zenodo
url = "https://zenodo.org/api/records/"

# Parámetros de autenticación (clave de acceso personal)
headers = {"Authorization": "Bearer
nVyvkXM3ezd5bKA0hesFLc3uMKMgfRenCxA3ZxkfusTAm0ZVzcvr3wZxE2k0"}

# Inicializar una lista para almacenar los últimos siete dígitos del DOI
dois = []

# Realizar la búsqueda y recopilar los resultados paginados
page = 1
while len(dois) < num_results:
   response = requests.get(url, params={"q": query, "size": num_results, "type": "dataset", "page": page},
headers=headers)
   data = response.json()
   records = data["hits"]["hits"]
   for record in records:
      if "doi" in record:
         doi = record["doi"][-7:]
         if doi.isdigit():
            dois.append(doi)

   # Verificar si hay más páginas de resultados
   if len(records) < num_results:
      break

   # Incrementar el número de página para obtener más resultados
   page += 1

# Limitar la lista de DOIs a los primeros 100 si se obtuvieron más
dois = dois[:num_results]

# Escribir los últimos siete dígitos del DOI en un archivo TXT
with open("dois.txt", "w", encoding="utf-8") as txtfile:
   for doi in dois:
      txtfile.write(doi + "\n")
print(f"Se han guardado los últimos siete dígitos de los DOIs en dois.txt.")
```

Anexo IV: esposende.py

```
# ========================================================================#
# ========================================================================#
# ============== ESPOSENDE: evaluador semiautomático DWBP v.1 ===============#
# ========================================================================#
# ========================================================================#
# librerías necesarias para el script
import csv
import requests
import sys
import re
import time
import json

# variables a inicializar
processed_dois = 0

# Número máximo de intentos permitidos
max_attempts = 3

# Bucle para solicitar el nombre del archivo
while max_attempts > 0:
   doi_file = input("Ingrese el nombre del archivo con la lista de DOIs de los datasets a analizar: ")

   try:
      with open(doi_file, "r") as f:
         doi_list = [doi.strip() for doi in f.readlines()]
      break  # Sale del bucle si se pudo abrir el archo correctamente
   except FileNotFoundError:
      print(f"El archivo '{doi_file}' no se encontró. Quedan {max_attempts - 1} intentos.")
      max_attempts -= 1

if max_attempts == 0:
   print("Se agotaron los intentos. El programa se detiene.")
   sys.exit()

# Abre el archivo de entrada con los DOIS de los datasets y los lee en una lista, limpiando los espacios en
blanco
with open(doi_file, "r") as f:
   doi_list = [doi.strip() for doi in f.readlines()]
   doi_list = [str(doi) for doi in doi_list]  # Convierte los elementos de la lista a cadenas de texto

# Por verificación de correcta lectura, informa del total de DOIs leídos
print(f"Total de DOIs leídos: {len(doi_list)}")

# Inicializar la lista (diccionario).
record_info = []

# Abre el archivo CSV en modo de escritura
with open("zenodo.csv", mode="w", newline="") as csvfile:
   writer = csv.writer(csvfile)
   writer.writerow(["DOI", "BP1", "BP2", "BP3", "BP4", "BP5", "BP6", "BP7", "BP8", "BP9", "BP10", "BP11",
"BP12", "BP13", "BP14", "BP15", "BP16", "BP17", "BP18", "BP19", "BP20", "BP21", "BP22", "BP23", "BP24",
"BP25", "BP26", "BP27", "BP28", "BP29", "BP30", "BP31", "BP32", "BP33", "BP34", "BP35"])
```

```
    # Inicializa el contador de fila
    row_number = 0

    # Para cada DOI en la lista, gracias a la API de Zenodo, obtenemos los metadatos y se agregan a una lista
    for doi in doi_list:
        response = requests.get(f"https://zenodo.org/api/records/{doi}", headers={"Authorization": "Bearer
nVyvkXM3ezd5bKA0hesFLc3uMKMgfRenCxA3ZxkfusTAm0ZVzcvr3wZxE2k0"}, timeout=500)

        if response.status_code == 200:
            record = response.json()
            metadata = record.get('metadata', {})
            record_data = {
                "DOI": doi,
                # "metadata": metadata,
                "access": record.get("access", ""),
                "access_users": record.get("access_users", ""),
                "access_right": record.get("access_right", ""),
                "archive": record.get("archive", ""),
                "archive_media": record.get("archive_media", ""),
                "communities": record.get("communities", ""),
                "created": record.get("created", ""),
                "creators": record.get("creators", {}),
                "conceptrecid": record.get("conceptrecid", []),
                "conceptdoi": record.get("conceptdoi", []),
                "contributors": record.get("contributors", ""),
                "id": record.get("id", ""),
                "description": record.get("description", ""),
                "doi": record.get("doi", ""),
                "doi_url": record.get("doi_url", ""),
                "draft": record.get("draft", ""),
                "files": record.get("files", []),
                "grants": record.get("grants", ""),
                "imprint": record.get("imprints", ""),
                "journal": record.get("journal", ""),
                "keywords": record.get("imprints", ""),
                "key": record.get ("key", ""),
                "languages": record.get("languages", ""),
                "license": record.get("license", ""),
                "links": record.get("links", []),
                "media_files": record.get("media_files", []),
                "meeting": record.get("meeting", ""),
                "owners": record.get("owners", []),
                "parent": record.get("parent", []),
                "publication_date": record.get("publication_date", []),
                "references": record.get("references", ""),
                "related_identifiers": record.get("related_identifiers", []),
                "relations": record.get("relations", []),
                "resource_type": record.get("resource_type", ""),
                "revision": record.get("revision", ""),
                "recid": record.get("recid", ""),
                "self": record.get("self", ""),
                "self_doi": record.get("self_doi", ""),
                "status": record.get("status", ""),
                "stats": record.get("stats", {}),
                "state": record.get("state", {}),
                "submitted": record.get("submitted", ""),
                "subjects": record.get("subjects", ""),
```

```
            "title": record.get("title", ""),
            "version": record.get("version", ""),
            "modified": record.get("modified", ""),
            "updated": record.get("updated", []),
            "views": record.get("views", ""),
            "downloads": record.get("downloads", ""),

        }
        record_info.append(record_data)

            # Mostrar el contenido de record_info en pantalla
            # if "related_identifiers" in metadata:
            #   print("Otros ids")
            #   print(metadata["related_identifiers"])

# ============================================================================#
# ==== INICIALIZA LAS VARIABLES DE LAS BUENAS PRÁCTICAS DE GESTIÓN DE DATOS  ===#
# ============================================================================#
bp1 = bp1_obl = bp1_opt  = bp2 = bp2_obl = bp2_opt = bp3 = bp3_obl = bp3_opt = bp4 = bp4_obl = bp4_opt
= bp5 = bp5_obl = bp5_opt = bp6 = bp6_obl = bp6_opt = bp7 = bp7_obl = bp7_opt = bp8 = bp8_obl = bp8_opt
= 0
        bp9 = bp9_obl = bp9_opt = bp10 = bp10_obl = bp10_opt = bp11 = bp11_obl = bp11_opt = bp12 =
bp12_obl = bp12_opt = bp13 = bp13_obl = bp13_opt = bp14 = bp14_obl = bp14_opt = bp15 = bp15_obl =
bp15_opt = 0
        bp16 = bp16_obl = bp16_opt = bp17 = bp17_obl = bp17_opt = bp18 = bp18_obl = bp18_opt = bp19 =
bp19_obl = bp19_opt = bp20 = bp20_obl = bp20_opt = bp21 = bp21_obl = bp21_opt = bp22 = bp22_obl =
bp22_opt = bp23 = bp23_obl = bp23_opt = 0
        bp24 = bp24_obl = bp24_opt = bp25 = bp25_obl = bp25_opt = bp26 = bp26_obl = bp26_opt = bp27 =
bp27_obl = bp27_opt = bp28 = bp28_obl = bp28_opt = bp29 = bp29_obl = bp29_opt = 0
        bp30 = bp30_obl = bp30_opt = bp31 = bp31_obl = bp31_opt = bp32 = bp32_obl = bp32_opt = bp33 =
bp33_obl = bp33_opt = bp34 = bp34_obl = bp34_opt = bp35 = bp35_obl = bp35_opt = 0

        # Incrementa el contador de fila para la siguiente iteración
        row_number += 1
        # print(f"Procesada la fila: {row_number}")
        # print (row_number)
        # print (doi)
# ============================================================================#
#  ======================== INICIO DE LAS VERIFICACIONES  ====================#
# ============================================================================#
        # Verificar BP1 - Proporcionar metadatos de forma extensiva
        bp1_obl = 0
        required_fields = ["title", "description", "keywords", "publication_date", "doi", "license"]
        if all(field in metadata for field in required_fields) and 'owners' in record_data and 'files' in record_data
and 'stats' in record_data:
            bp1_obl = 1

        # Verifica el punto de referencia BP1_opt
        optional_fields = ["relations", "related_identifiers", "language", "access_right", "creators",
"resource_type"]
        bp1_opt = sum(0.1 for field in optional_fields if field in metadata)

        # Calcula BP1 como la suma de BP1_obl y BP1_opt
        bp1 = bp1_obl + bp1_opt

        # Verificar BP2 - Proporcionar metadatos descriptivos
```

```
        required_fields = ["title", "description", "publication_date", "keywords", "license"]
        if all(field in metadata for field in required_fields) and 'owners' in record_data and 'owners' in
record_data:
            bp2_obl = 1

        # Verifica el punto de referencia BP2_opt
        optional_fields = ["access_right", "relations", "creators", "language", "related_identifiers", "doi"]
        bp2_opt = sum(0.1 for field in optional_fields if field in metadata)
        if 'updated' in record_data:
            bp2_opt += 0.1

        # Calcula BP2 como la suma de BP2_obl y BP2_opt
            bp2 = bp2_obl + bp2_opt

        # Verifica BP3 - Proporcionar metadatos estructurales - Entendemos que 'self' informa del tipo de
fichero, por la extensión, y que siempre está en "files"
        # No hay metadatos para verificar bp3_obl
            bp3_obl = 0
            bp3_opt = 0

        # Verifica el punto de referencia BP3_opt
        if "resource_type" in metadata:
            bp3_opt = 0.1
        if "related_identifiers" in metadata:
            bp3_opt += 0.1

        # Calcula BP3 como la suma de BP3_obl y BP3_opt
            bp3 = bp3_obl + bp3_opt

        # Verifica BP4 - datos de la licencia
        if "license" in metadata:
            bp4_obl = 1

        # Verifica el punto de referencia BP4_opt
        bp4_opt = 0
        if "access_right" in metadata:
            bp4_opt += 0.1
        if "links" in record_data and "access_users" in record_data["links"]:
            bp4_opt += 0.1
        if "links" in record_data and 'access_links' in record_data["links"]:
            bp4_opt += 0.1

        # Calcula BP4 como la suma de BP4_obl y BP4_opt
            bp4 = bp4_obl + bp4_opt

        # Verifica BP5 - datos sobre la procedencia
        if "contributors" in metadata:
            bp5_obl = 1

        # Verifica el punto de referencia BP5_opt
            optional_fields = ["creators", "related_identifiers", "grants"]
            bp5_opt = sum(0.1 for field in optional_fields if field in metadata)
        if "links" in record_data and "archive" in record_data["links"]:
            bp5_opt += 0.1

        # Calcula BP5 como la suma de BP5_obl y BP5_opt
            bp5 = bp5_obl + bp5_opt
```

```
        # Verifica BP6 - Proporcionar información sobre la calidad de los datos y su adecuación a determinados
fines.
            bp6_obl = 0
            bp6_opt = 0

        # Verifica si 'creators' está en 'metadata' y suma 0.1 si es cierto
        if 'creators' in metadata:
            bp6_opt += 0.1
        # Verifica si 'updated' está en 'record_data' y suma 0.1 si es cierto
        if 'updated' in record_data:
            bp6_opt += 0.1
        # Verifica si 'owners' está en 'record_data' y suma 0.1 si es cierto
        if 'owners' in record_data:
            bp6_opt += 0.1

        # Calcula BP6 como la suma de BP6_obl y BP6_opt
            bp6 = bp6_obl + bp6_opt

        # Verifica BP7 - Proporcionar un indicador de versión.
        if "revision" in record_data:
            bp7_obl = 1

        # Verifica bp7_opt - si tenemos info de versión y si el número de revisiones es mayor que 1
            bp7_opt = 0
        if "version" in metadata:
            bp7_opt += 0.1
        if "updated" in record_data:
            bp7_opt += 0.1

        # Calcula BP7 como la suma de BP7_obl y BP7_opt
            bp7 = bp7_obl + bp7_opt

        # Verifica BP8 - Proporcione un historial de versiones - comprueba si 'relations' está presente en
metadata y dentro del mismo la clave 'version' informa del historial.
        bp8_obl = 0

        if 'relations' in metadata:
        #    print("relations =", metadata['relations'])

            # Si 'version' está dentro de 'relations', imprímelo
            if 'version' in metadata['relations']:
        #       print("version =", metadata['relations']['version'])
                bp8_obl = 1

        #Verifica bp8_opt - el número de versiones del dataset
            bp8_opt = 0
        if "revision" in record_data:
            bp8_opt += 0.1
        if "updated" in record_data:
            bp8_opt += 0.1
        # Calcula BP8 como la suma de BP8_obl y BP8_opt

        bp8 = bp8_obl + bp8_opt
        # print ("bp8, bp8_obl_, bp8_opt ", bp8, bp8_obl, bp8_opt)

        # Verifica BP9 - Utilizar URI persistentes como identificadores de los datasets.
```

```
if "doi" in metadata:
    bp9_obl = 1

# No hay elementos optativos para la BP9
    bp9_opt = 0

# Calcula BP9 como la suma de BP9_obl y BP9_opt
    bp9 = bp9_obl + bp9_opt

# Verifica BP10 - Utilizar URIs persistentes como identificadores dentro de los conjuntos de datos.
bp10_obl = 0
if "files" in record_data:
    for file_info in record_data["files"]:
        if "id" in file_info:
            bp10_obl = 1

# Verifica elementos recomendables para BP10
    bp10_opt = 0
if "related_identifiers" in metadata:
    bp10_opt = 0.1

# Calcula BP10 como la suma de BP10_obl y BP10_opt
    bp10 = bp10_obl + bp10_opt

# Verifica BP11 - Asignar URI a versiones de conjuntos de datos y series. .
if "conceptdoi" in record_data:
    bp11_obl = 1
else:
    bp11_obl = 0

# Verifica bp11_opt - si está el elemento "related_identifiers".
if 'related_identifiers' in metadata:
    bp11_opt = 0.1

# Calcula BP11 como la suma de BP11_obl y BP11_opt
bp11 = bp11_obl + bp11_opt

# Verifica BP12 - Utilizar formatos de datos normalizados legibles por máquina.
if "files" in record_data:
    for file_info in record_data["files"]:
        if "key" in file_info:
            bp12_obl = 1

# No existen elementos recomendables para esta buena práctica.
    bp12_opt = 0

# Calcula BP12 como la suma de BP12_obl  y bp12_opt
    bp12 = bp12_obl + bp12_opt

# Verifica BP13 - idioma del dataset.
if "language" in metadata:
    bp13_obl = 1
else:
    bp13_obl = 0
# No hay elementos de verificación para BP13_opt
    bp13_opt = 0
```

```
        # Calcula BP13 como la suma de BP13_obl y bp13_opt
        bp13 = bp13_obl + bp13_opt

        # Verifica punto de referencia BP14 - proporcionar la información en múltiples formatos
        if "files" in record_data:
           for file_info in record_data["files"]:
              if "key" in file_info:
                 bp14_obl = 1

        # No hay elementos de verificación para BP14_opt
           bp14_opt = 0

        # Calcula BP14 como la suma de BP14_obl y BP14_opt
           bp14 = bp14_obl + bp14_opt

        # Verifica BP15 reutilizar vocabularios
        required_fields = ["keywords", "license", "resource_type"]
        if all(field in metadata for field in required_fields) and "languages" in record_data:
           bp15_obl = 1

        # Verifica los puntos de referencia BP15_opt
           bp15_opt = 0
        if "access_right" in metadata:
           bp15_opt += 0.1
        if "access" in record_data:
           bp15_opt += 0.1

        # Calcula BP15 como la suma de bp15_obl y bp15_opt
           bp15 = bp15_obl + bp15_opt

        # Verifica BP16 - Elegir el nivel de formalización adecuado.

        if "files" in record_data:
         for file_info in record_data["files"]:
            if "key" in file_info:
               bp16_obl = 1

        # Verifica BP16_opt
        # obviamos usar los metadatos 'stadndarized' y 'vocabularies' porque no se usan en Zenodo y
simplificamos el código
        if "license" in metadata:
           bp16_opt = 0.1
        else:
           bp16_opt = 0

        # Calcula bp16
        bp16 = bp16_obl + bp16_opt

        # Verifica BP17 - descarga masiva del dataset
        if "links" in record_data and "archive" in record_data["links"]:
           bp17_obl = 1

        # No hay puntos de referencia opcionales
           bp17_opt = 0

        # Calcula BP17 como la suma de bp17_obl y bp17_opt
           bp17 = bp17_obl + bp17_opt
```

```
        # Verifica BP18 - Proporcione subconjuntos para grandes conjuntos de datos.
        # No hay puntos de referencia para bp18_obl
           bp18_obl = 0

        #Verifica bp18_opt
        bp18_opt = 0
        if "related_identifiers" in metadata:
         for file_info in record_data["files"]:
            if "isPartOF" in file_info:
               bp18_opt = 0.1
            if "PartOF" in file_info:
               bp18_opt += 0.1
            if "isContinuedBy" in file_info:
               bp18_obl += 0.1
            if "continues" in file_info:
               bp18_opt += 0.1

        # Calcula BP18
        bp18 = bp18_obl + bp18_opt

        # Verifica BP19 - Utilizar la negociación de contenidos para servir datos disponibles en varios formatos.
No hay puntos de referencia para bp19_obl
        bp19_obl = 0

        # Verifica bp19_opt - el número de versiones del dataset
        bp19_opt = 0

        if "files" in record_data:
           for file_info in record_data["files"]:
              if "key" in file_info:
                 bp19_opt = 0.1
        if "links" in record_data and "access_links" in record_data["links"]:
           bp19_opt += 0.1
        if "access_right" in metadata:
           bp19_opt += 0.1

        # Calcula BP19 como la suma de BP19_obl y BP19_opt
           bp19 = bp19_obl + bp19_opt

        # Verifica BP20 - Proporcionar acceso en tiempo real. No disponemos de metadatos para verificar
bp20_obl
           bp20_obl = 0

        # Verifica bp20_opt
        if "updated" in record_data:
           bp20_opt = 0.1
        if "modified" in record_data:
           bp20_opt += 0.1

        # Calcula BP20 como la suma de BP20_obl y BP20_opt
           bp20 = bp20_obl + bp20_opt

        # Verifica BP21 - Informar de la fecha de actualización de la información y de la frecuencia de
actualización.
        if "updated" in record_data:
           bp21_obl = 1
```

```
        else:
            bp21_obl = 0

        # Verifica bp21_opt
            bp21_opt = 0
        if "created" in record_data:
            bp21_opt += 0.1
        if "modified" in record_data:
            bp21_opt += 0.1

        # Calcula BP21 como la suma de BP21_obl y BP21_opt
            bp21 = bp21_obl + bp21_opt

        # Verifica BP22 - Explicación de los datos no disponibles.
        bp22_obl = 0
        if "status" in record_data:
            bp22_obl = 1

        # Verifica BP22_opt
        bp22_opt = 0
        if "state" in record_data:
            bp22_opt = 0.1

        # Calcula BP22 como la suma de BP22_obl y BP22_opt
        bp22 = bp22_obl + bp22_opt

        # Verifica BP23 - Si la API-REST es pública comprueba si 'access_right' está presente en metadata y si su
valor es "open".

        if "access_right" in metadata and metadata["access_right"] == "open":
            bp23_obl = 1

        # No hay elementos para verificar bp23_opt

        # Calcula BP23 como la suma de BP23_obl y BP23_opt
        bp23 = bp23_obl


        # Verifica BP24 - Si la API-REST publica sigue estandarew web.
        bp24_obl = 0
        if "links" in record_data and "access_request" in record_data["links"]:
            bp24_obl = 1

        # No hay elemento de metadatos para verificar bp24_opt
            bp24_opt = 0

        # Calcula BP24 como la suma de BP24_ob l y BP24_opt
            bp24 = bp24_obl + bp24_opt

        # Verifica BP25 - Si hay acceso a la documentación de la API pública - esto se cumple en Zenodo.
            bp25 = 1

        # Verifica BP26 - Evitar cambios bruscos en la API-REST publica - esto se cumple en Zenodo.
            bp26 = 1

        # Verifica BP27 - Conservar los identificadores - esto se cunple en Zenodo.
            bp27 = 1
```

```
        # Verifica BP28 - Evaluar la cobertura del repositorio. Se entiende que siempre es posible con
metadatos.
        if "coverage" in metadata:
            bp28_obl = 1
        if "coverage" in record_data:
            bp28_obl = 1

        # Verifica BP28_opt - no hay elementos recomendados
            bp28_opt = 0

        # Calcula BP28 como la suma de BP28_obl y BP28_opt
            bp28 = bp28_obl + bp28_opt

        # Verifica BP29 - Recoger la opinión de los usuarios.
        if "owners" in record_data:
            bp29_obl = 1

        # Verifica el punto de referencia BP29_opt
        if "creators" in metadata:
            bp29_opt = 0.1

        # Calcula BP29 como la suma de BP29_obl y BP29_opt
            bp29 = bp29_obl + bp29_opt

        # Verifica BP30 - Publicar las opiniones de los usuarios - en Zenodo, en principio, no es posible. Lo que
no implica que no debiera hacerse.
            bp30 = 0

        # Verifica BP31 - Enriquecer los datos generando otros nuevos - comprobar si 'related_identifiers' está
presente en metadata y contiene alguna de las claves requeridas.
        # No hay elementos obligatorios para verificar.
            bp31_obl = 0

        # Verifica BP31_opt - no hay elementos recomendados
        if "related_identifiers" in metadata:
            rich_relations = {"isDerivedFrom", "isDescribedby", "isSupplementTo", "isSourceOf", "isRequiredBy",
"isObsoloteby"}
            related_identifiers = metadata["related_identifiers"]
            bp31_opt += sum(0.1 for item in metadata["related_identifiers"] if item.get("relation") in
rich_relations)

        # Calcula BP31 como la suma de BP31_obl y BP31_opt
            bp31 = bp31_obl + bp31_opt

        # Verifica BP32 - Proporcionar presentaciones complementarias - comprobar si 'relations' está
presente y contiene alguna de las claves requeridas.
        # No hay elementos para verificar bp32_obl
            bp32_obl = 0

        # Verifica  elementos recomendables

        if "related_identifiers" in metadata:
            cited_keys = {"isSupplementTo", "isSuplementedBy", "isDescribedBy", "isPartOf", "isCompiledBy",
"isReferencedBy"}
            related_identifiers = metadata["related_identifiers"]
            bp32_opt += sum(0.1 for item in metadata["related_identifiers"] if item.get("relation") in cited_keys)
```

```
        # Calcula BP32 como la suma de BP32_obl y BP32_opt
        bp32 = bp32_obl + bp32_opt

        # Verifica BP33 - Poder informar al editor original.
        if "owners" in record_data:
            bp33_obl = 1

        # Verifica el punto de referencia BP33_opt
        if "creators" in metadata:
            bp33_opt = 0.1

        # Calcula B33 como la suma de BP33_obl y BP33_opt
            bp33 = bp33_obl + bp33_opt

        # Verifica BP34 - Cumplir las condiciones de licencia del editor original.
        if "links" in record_data and "access" in record_data["links"]:
            bp34_obl = 1
        else:
            bp34_obl = 0

        # Verifica el punto de referencia para bp34_opt
        if "license" in metadata:
            bp34_opt = 0.1

        # Calcula BP34 como la suma de BP34_obl y BP34_opt
            bp34 = bp34_obl + bp34_opt

        # Verifica punto de referencia BP35 - Citar la publicación original
        # No hay metadatos para verificar bp35_obl
            bp35_obl = 0

        # Verificar  BP35_opt - comprobar si en 'relations' hay citas a la publicación original
        if "related_identifiers" in metadata:
            cited_keys = {"isBasedOn", "isPartOf", "references", "cites", "isDerivedFrom", "isSupplementTo"}
            related_identifiers = metadata["related_identifiers"]
            bp35_opt += sum(0.1 for item in metadata["related_identifiers"] if item.get("relation") in cited_keys)

        # Calcula BP35 como la suma de BP35_obl y BP35_opt
            bp35 = bp35_obl + bp35_opt

# ========================================================================#
# == ==================== FIN DE LAS VERIFICACIONES  ======================#
# ========================================================================#
# ========================================================================#
# ===  ESCRIBE LOS RESULTADOS DE CADA VARIABLE EN EL FICHERO CSV DE SALIDA  ====#
# ========================================================================#

writer.writerow([doi, float(bp1), float(bp2), float (bp3), float(bp4), float(bp5), float(bp6), float(bp7),
float(bp8), float(bp9), float(bp10), float(bp11), float(bp12), float(bp13), float(bp14), float(bp15), float(bp16),
float(bp17), float(bp18), float(bp19), float (bp20), float(bp21), float(bp22), float(bp23), float(bp24),
float(bp25), float(bp26), float(bp27), float(bp28),float(bp29), float(bp30), float(bp31), float(bp32),
float(bp33), float(bp34), float(bp35)])
```