Artículos originales

Calidad en los repositorios digitales. Los principios TRUST para repositorios de datos

Quality in digital repositories. The TRUST principles for data repositories

Marisa Raquel De Giusti[1, 2]

[1] PREBI-SEDICI, Universidad Nacional de La Plata, La Plata, Argentina
[2] Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, La Plata, Argentina

marisa.degiusti@sedici.unlp.edu.ar

Recibido: 23/06/2020 | **Corregido:** 30/12/2020 | **Aceptado**: 05/01/2021

Cita sugerida: M. R. De Giusti, "Calidad en los repositorios digitales. Los principios TRUST para repositorios de datos," *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*, no. 29, pp. 55-59, 2021. doi: 10.24215/18509959.29.e6

Esta obra se distribuye bajo **Licencia Creative Commons CC-BY-NC 4.0**

Resumen

El presente trabajo pretende dar a conocer los principios TRUST que dan directrices para considerar confiables a los repositorios de datos. Estos principios abarcan la Transparencia, la Responsabilidad, el Foco en el Usuario, la Sostenibilidad y la Tecnología, y presentan un marco común para guiar a los gestores de repositorios de datos en la implementación de las mejores prácticas en cuanto a preservación digital. Para lograr el objetivo propuesto se dan definiciones para el área, conceptos vinculados a la calidad, normativa vigente para auditoría y certificación y finalmente, se explican y ejemplifican los modos de cumplir con los mencionados principios que aseguran que los datos almacenados son FAIR es decir que son datos encontrables, accesibles, interoperables y reusables por cualquier comunidad de modo de avanzar hacia una ciencia abierta, participativa y socialmente comprometida.

Palabras clave: Repositorios; Ciencia abierta; Datos; Confiabilidad.

Abstract

This work presents TRUST principles, which provide guidelines for trustworthy data repositories. These principles, namely, Transparency, Responsibility, User Community and Sustainability, and Technology (TRUST), set a shared framework to guide repository managers in implementing the best practices for Digital Preservation. In order to achieve this goal, domain definitions are provided, along with concepts related to quality, current regulations for auditing and certification. Finally, this work also explains –with illustrative examples– how to comply with the aforementioned principles and thus ensure the data stored in the repositories are Findable, Accessible, Interoperable, Reusable (FAIR) by any community to move towards an open, participatory and socially responsible science.

Keywords: Repositories; Open science; Data; Reliability.

Es muy importante que se tenga la oportunidad de conocer y comprender los resultados del trabajo de investigación científica. No es suficiente que el conocimiento adquirido sea registrado, desarrollado y aplicado sólo por algunos especialistas. La limitación del capital de conocimientos a su propio círculo es la muerte del espíritu filosófico de todo un pueblo y conduce al empobrecimiento intelectual.

Albert Einstein (1948)

1. Introducción y antecedentes de Argentina

La República Argentina destaca desde hace más de una década en el avance por la compartición en abierto de su producción académica y científica, lo que hoy en día, en tiempos de pandemia, resulta aún más imprescindible.

En el año 2013, con un esfuerzo de los referentes de los distintos repositorios del país reunidos en el marco del Sistema Nacional de Repositorios Digitales del Ministerio de Ciencia, Tecnología e Innovación Productiva, se logró la sanción de la Ley 26.899 [1] que obliga a las instituciones educativas que reciben financiamiento del Sistema Nacional de Ciencia y Tecnología a crear repositorios propios o compartidos donde alojar su producción en acceso abierto, y, a los investigadores a dejar una copia de su producción (publicaciones y datos que la sustentan) archivada en el repositorio de su institución. En el 2016, se estableció la Resolución E 753 [2] que reglamenta la ley y deja al SNRD como instrumento técnico-operativo para el cumplimiento de las responsabilidades del ministerio emanadas de la mencionada ley.

A fines de 2019, Paola Azrilevich, referente del SNRD, compartió en un congreso en la ciudad de Córdoba (Argentina), el estado de situación de los repositorios en el país, sus avances y las tareas pendientes [3].

Esta posición de privilegio de nuestro país, en cuanto a contar con una ley que, además de lo dicho previamente, obliga al establecimiento de políticas institucionales de acceso abierto, compromete a avanzar en cuestiones vinculadas a la calidad de los repositorios dedicados a albergar publicaciones y datos.

El objetivo de estas notas es compartir algunas premisas básicas para avanzar hacia la calidad de los repositorios y con ello la calidad de la práctica de compartir conocimiento en abierto; así, se describen brevemente consideraciones sobre calidad y estándares, para hacer hincapié finalmente en los principios TRUST para los repositorios de datos.

2. Definiciones: repositorios, calidad y auditoría

Los repositorios digitales son estructuras web capaces de almacenar, preservar y difundir contenidos, en principio académicos y científicos diversos, los cuales han sido archivados por los propios autores, por tareas de la administración que gestiona los contenidos, o por operaciones informáticas. Un repositorio debe brindar una interfaz para usuarios internos y externos. Si se trata de usuarios internos está claro que deben adecuarse a las numerosas tareas de gestión que realizan (catalogación, adecuación y agregado de metadatos, borrado, corrección de metadatos). En el caso de los usuarios externos, estos deben poder acceder, subir contenidos, buscar, navegar el repositorio, según el grado de permisos.

La definición de la Norma ISO 9000 [4] no deja dudas sobre el término "calidad" al definirla como "grado en el que un conjunto de características inherentes a un objeto (producto, servicio, proceso, persona, organización, sistema o recurso) cumple con los requisitos".

El término "requisitos" es un problema porque se entiende como la "expectativa" del usuario y el repositorio tiene muy distintos usuarios: desde aquella persona que ingresa a navegar y recorrer el repositorio por curiosidad, a los autores y a los administradores y a cada grupo hay que brindarle los servicios necesarios; el repositorio además tiene usuarios que no son personas sino otros sistemas informáticos con los cuales interopera y de los cuales recibe requerimientos.

Vale aclarar que en lo referido a "demostrar la calidad", se trate del ámbito que se trate, hay una amplia gama de alcances en esa demostración. Existen chequeos de autoevaluación con la intervención de pares que, incluso, puede llegar a alcanzar el grado de una auditoría interna; tales auditorías pueden ser realizadas también por externos, las llamadas de forma genérica de terceras partes e incluso alcanzar la tan ansiada certificación.

2.1. Estándares y normas para auditoría y certificación

Una redacción detallada de este apartado nos llevaría a extendernos más allá de las posibilidades e intención de este trabajo, sin embargo se muestran aquí algunos ejemplos de normas y estándares adecuados para hacer un seguimiento y mejorar la calidad de los repositorios. Estándares como el de niveles NDSA [5] o la Certificación DINI [6] para Alemania se tornan muy interesantes y de fácil comprensión a la hora de autoevaluar un repositorio; por supuesto que cualquier repositorio desearía una certificación en ISO 16363 [7], pero esta norma tiene alrededor de cien criterios y exige un desarrollo completo y parejo en cuanto a gestión, objeto digital e infraestructura; semejante complejidad dificulta alcanzar la tan ansiada marca de calidad. La mención casi exclusiva de los estándares mencionados responde ante todo a la experiencia de revisión de la distinta normativa bastante más extensa vinculada a auditoría y certificación.

DINI es un certificado que resulta muy interesante ya que es simple y a la vez concreto. Aborda 8 criterios, correctamente explicados que deben atender los servicios de publicación digital (entre ellos pero no excluyentemente los repositorios) aborda la visibilidad del servicio, las políticas, el asesoramiento a autores y

editoriales, los aspectos legales, la seguridad de la información, la indexación e interfaces, las estadísticas de acceso y la disponibilidad a largo plazo. Como puede observarse los ocho criterios balancean aspectos vinculados a la gestión, la administración y las políticas pero también hay criterios relacionados con lo técnico, todos ellos orientan adecuadamente al gestor del repositorio y el personal para mejorar el servicio. Para detalles mayores es posible consultar la última versión en inglés del mencionado certificado [6].

NDSA por su parte resulta muy interesante porque habilita una autoevaluación progresiva del repositorio. En tanto trabaja sobre cinco grandes áreas: almacenamiento y localización geográfica de los datos, no alteración e integridad de archivos, seguridad de la información, metadatos y formatos de archivos, todos ellos aspectos técnicos informáticos y bibliotecarios, permite evaluar, según el nivel de cumplimiento, si el repositorio es capaz de, en un primer nivel, de proteger los datos, en un segundo nivel de conocerlos, en un tercer nivel de controlarlos y en un cuarto nivel de repararlos. Los aspectos para el cumplimiento se expresan de manera muy comprensible y hasta es posible sacar un "valor" en todos y cada uno de los aspectos que permite perfilar el estado del repositorio. Una versión en español ha sido publicada recientemente y se encuentra disponible en [13].

Finalmente la Norma ISO 16363 es, sin lugar a dudas el estándar que cubre todos los aspectos: organizativos, de gestión del objeto digital y de seguridad. Ninguna otra norma o principio alcanza tal nivel de detalle, aunque todas, sin lugar a dudas tocan diferentes aspectos que la norma considera, por citar un aspecto la Infraestructura Organizativa por ejemplo, requiere datos y documentos sobre toda la estructura organizativa y su sostenibilidad a largo plazo, con información económica adecuada; en los aspectos técnicos se requiere que exista planificación documentada sobre preservación digital. Estos dos puntos mencionados sólo pretenden mostrar que, la certificación en una Norma ISO como esta requiere un nivel de organización y disposición de personal que normalmente no alcanzan sino los repositorios de instituciones con una economía privilegiada.

3. La ciencia abierta y sus necesidades

La ciencia abierta supone la compartición de todo el proceso de avance de una investigación, propone discusión e interacción de principio a fin y esto extiende los requisitos necesarios para dar cuenta de este nuevo desafío que incluye, entre otros, al acceso abierto.

Compartir los conocimientos generados en una investigación implica, entre muchas otras cuestiones, la necesidad de guardar también los datos que acompañan el proceso y sirven de base a las publicaciones, lo que plantea nuevos retos. En Europa, con el objetivo de desarrollar la filosofía de la Ciencia Abierta, la Comisión Europea requiere, desde enero de 2017, que todos los proyectos financiados con el Programa Horizonte 2020 [8], salvo excepciones justificadas, garanticen el acceso abierto a los datos de investigación. En nuestro país, muy ligado a las prácticas de la ciencia abierta y en estricto acuerdo con la reglamentación de la Ley 26.899 se solicita el plan de gestión de datos y el depósito en abierto de los datos crudos, éstos últimos con un plazo máximo desde su producción de cinco años; es claro que la ley contempla excepciones.

Por lo precedente, en el ámbito europeo, se solicita que los datos sigan los Principios FAIR (acrónimo en inglés de Findable, Accessible, Interoperable and Reusable) publicados en 2016 [9].

Los Principios FAIR dan cuenta de un conjunto de cualidades para conseguir que los datos sean calificados de este modo; en español se habla de que deben ser encontrables, accesibles, interoperables y reutilizables; brevemente cada una de las mencionadas cualidades significa:

Encontrables: que cuenten con un identificador único y persistente DOI o handle, que esos datos estén adecuadamente descritos por metadatos enriquecidos, incluyendo ese identificador asignado y que sean adecuadamente indexados en un recurso de búsqueda, por ejemplo un repositorio.

Accesibles: que sea posible acceder a los datos utilizando protocolos estandarizados de comunicación que sean abiertos y gratuitos. Cuando los datos no puedan ser abiertos por razones de privacidad, seguridad nacional o intereses comerciales, el protocolo debe permitir procedimientos para la autentificación y la autorización.

Interoperables: el formato de los metadatos debe ser estándar y aceptado por la comunidad, del mismo modo los lenguajes y vocabularios deben ser controlados y acordados por la comunidad y contener enlaces a información relacionada mediante identificadores.

Reutilizables: asignación de metadatos con atributos que proporcionen información contextual y de procedencia. Deben utilizar una licencia abierta y legible por las máquinas y estándares que use la comunidad del dominio concreto, para permitir su reutilización.

Publicar los datos de investigación en acceso abierto y siguiendo los Principios FAIR permite cumplir con el siguiente lema "tan abiertos como sea posible y tan cerrados como sea necesario", lo que asegura la protección de datos personales, sensibles e incluso confidenciales. El avance en el cumplimiento de estos principios depende grandemente de su conocimiento por parte de los investigadores y también del apoyo institucional al reconocimiento del trabajo que demanda la tarea y a su acompañamiento con los incentivos adecuados. Un informe de JISC [10], que relata la investigación de un grupo seleccionado de expertos para determinar el uso de los Principios FAIR en el Reino Unido, pone de manifiesto el uso limitado de los mismos a aspectos más conceptuales, aún en el caso de comunidades en las que están los principios bien establecidos,

destacando una variabilidad en el uso vinculada a las distintas disciplinas, instituciones e incluso investigadores.

La importancia reconocida de los Principios FAIR para los datos lleva a la inevitable cuestión de cuándo dar ese carácter de confiables a los conjuntos de datos.

Los Principios FAIR tienden a centrarse en el estado actual de los datos, pero para que éstos se mantengan auténticos y seguros a lo largo del tiempo se requiere información contextual y repositorios "confiables" (por sus siglas en inglés llamados TDRs) que respalden y mantengan activamente todos los requerimientos de FAIR para sus datos; esto nos obliga a traer a colación la necesidad de conservación y acceso a largo plazo de los datos y con ello la Norma ISO 14721 sobre la que, en definitiva se basa un estándar tan complejo como la Norma de certificación 16363.

En el 13º Plenario de la RDA, en abril de 2019, Dawei Lin [11] expuso la versión ya actualizada de los principios TRUST mostrando de manera clara la complementariedad entre FAIR y TRUST, remarcando que mientras que los Principios FAIR definen las propiedades deseadas para los datos y metadatos, los Principios TRUST describen las características deseadas de los repositorios de datos para que puedan asegurar la gestión y diseminación correcta de todos los conjuntos de datos que albergan, a lo largo del tiempo.

Los Principios TRUST consideran la transparencia, la responsabilidad, el foco en el usuario y la sostenibilidad y la tecnología como los componentes esenciales para definir repositorios de datos confiables (*Trusthworthy Digital Repositories*).

Transparencia se refiere a la capacidad del repositorio para dar acceso público a sus políticas, la comunidad a la que está dedicado, la misión y el alcance de sus tareas, los términos de uso de sus datos y metadatos, los tiempos de conservación de los datos que gestiona, así como su responsabilidad en relación al manejo de confidencialidad.

Responsabilidad se refiere a gestionar los datos de manera adecuada, lo que significa en principio conocer los metadatos de la comunidad, asegurar la autenticidad y permanencia inalterada a largo plazo, prestar los servicios a través de una interfaz, incluida la descarga de datos por humanos o máquinas. Gestionar los derechos y seguir las normas éticas.

Foco en el usuario significa conocer la comunidad a la que presta servicio y monitorear continuamente sus requerimientos brindando métricas adecuadas de sus existencias.

Sostenibilidad significa tener capacidad para brindar servicios a lo largo del tiempo y responder con servicios nuevos o perfeccionados en función de los cambios en los requisitos de la comunidad y para demostrarlo se debe contar con una planificación que incluya a personas, tecnología e incluso los modos de actuar ante riesgos concretos para los datos (desastres, incendios) de modo de asegurar el mantenimiento de los datos. Claramente para todo ello se precisa de un adecuado financiamiento a lo largo del tiempo y hacer una gestión que asegure la preservación y el acceso a futuro.

Tecnología significa contar con la infraestructura y capacidades (humanas y técnicas) para soportar todas las operaciones. Contar con herramientas de software y hardware actualizadas a través de las cuales las personas puedan llevar adelante los procesos necesarios para el mantenimiento correcto de los datos en el tiempo y el repositorio lo demuestra con la implementación de estándares, herramientas y tecnologías relevantes apropiadas para la gestión y curación de los datos, así como su prevención frente a los riesgos. Para un análisis más detallado puede verse el trabajo recientemente publicado de Dawei Lin [12].

Conclusiones

Los repositorios de datos albergan contenidos de muy diversa naturaleza, tamaño y formato de guardado y visualización para ser aptos por parte de aplicaciones que permitan la visualización e incluso el reúso de los contenidos. La complejidad de los acervos lleva a que los repositorios de datos se constituyan en un elemento clave para cumplir con el cometido de la ciencia abierta en cuanto a asegurar la reproducibilidad de la investigación. Para lograr gestionar y mantener adecuadamente sus acervos los repositorios digitales de datos deben realizar operaciones y gestionar sus contenidos de modo seguro. Los principios TRUST tienen el gran valor de traducir al mundo de los repositorios de datos los requerimientos reales para acreditar confianza sobre su capacidad de gestión frente a los diferentes grupos interesados en relación a su quehacer de albergar y dar acceso a largo plazo a datos que aseguren la reproducibilidad de la investigación. Los Principios TRUST guían a los gestores y demás administradores de los repositorios de datos porque permiten hacer una lista, (no necesariamente única ni estática) de los aspectos cumplidos y los aspectos a emprender en pos de cumplir las expectativas y, en definitiva, brindar servicios que aseguren la confianza en sus prácticas y en la ética que debe acompañar una tarea de resguardo de producciones que significan dedicación y tiempo por parte de los autores.

Referencias

[1] República Argentina, Congreso de la Nación, Ley 26.899: Sistema Nacional de Ciencia, Tecnología e Innovación. Repositorios digitales institucionales de acceso abierto. 2013. [Online]. Available: http://servicios.infoleg.gob.ar/infolegInternet/anexos/220000-224999/223459/norma.htm.

[2] República Argentina, Ministerio de Ciencia, Tecnología e Innnovación Productiva, Resolución 753-E/2016. 2016. [Online]. Available:https://www.argentina.gob.ar/normativa/nacional/resoluci%C3%B3n-753-2016-267833.

[3] P. A. Azrilevich and M. R. De Giusti, "El contexto de los repositorios de acceso abierto en la Argentina: Logros y asuntos pendientes", presented at *Asamblea General ISTEC y I Congreso Internacional de Tecnología Aplicada, Innovación y Educación Continua*, Córdoba, 2019. [Online]. Available: http://sedici.unlp.edu.ar/handle/10915/86423 and http://digital.cic.gba.gob.ar/handle/11746/10417

[4] International Organization for Standardization (ISO), UNE-EN ISO 9000:2015, Sistemas de gestión de la calidad. 2015. [Online]. Available: https://www.aenor.com/normas-y-libros/buscador-de-normas/une?c=N0055468.

[5] National Digital Stewardship Alliance (NDSA), Niveles de Preservación Digital (NDSA-LDP Ver. 1), 2020. [Online]. Available: http://www.apredig.org/wp-content/uploads/2019/01/Niveles-de-Preservaci%C3%B3n-Digital-NDSA-LDP-Ver.-1-Traducida..pdf

[6] U. Müller *et al.*, *DINI Certificate for Open Access Repositories and Publication Services 2019*. Humboldt-Universität zu Berlin, 2020. [Online]. Available: https://edoc.hu-berlin.de/handle/18452/22465.

[7] International Organization for Standardization (ISO), UNE-ISO 16363:2017 Sistemas de transferencia de información y datos espaciales. Auditoría y certificación de repositorios digitales de confianza. 2017. [Online]. Available: https://www.aenor.com/normas-y-libros/buscador-de-normas/iso/?c=062542.

[8] Unión Europea, Horizonte2020. 2020. [Online]. Available: https://eshorizonte2020.es/.

[9] M. Wilkinson, M. Dumontier, I. Aalbersberg et al, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci Data 3*, 160018, 2016, doi: https://doi.org/10.1038/sdata.2016.18.

[10] R. Allen and D. Hartland, FAIR in practice—JISC report on the Findable Accessible Interoperable and Reuseable Data Principles, Zenodo, 2018, doi: https://doi.org/10.5281/zenodo.1245568.

[11] D. Lin, The TRUST Principles for Trustworthy Data Repositories – An Update. 2019. [Online]. Available: https://www.rd-alliance.org/trust-principles-trustworthy-data-repositories-%E2%80%93-update.

[12] D. Lin, J. Crabtree, I. Dillo, R. R. Downs, R. Edmunds, D. Giaretta, M. De Giusti, H. L'Hours, W. Hugo, R. Jenkyns, V. Khodiyar, M. E. Martone, M. Mokrane, V. Navale, J. Petters, B. Sierman, D. Sokolova, M. Stockhause, J. Westbrook, "The TRUST Principles for digital repositories," *Scientific Data*, vol. 7, no. 1, p. 144, 2020, doi:https://doi.org/10.1038/s41597-020-0486-7. [Online]. Available: http://sedici.unlp.edu.ar/handle/10915/97465

[13] National Digital Stewardship Alliance (NDSA), «Niveles-de-Preservación-Digital-NDSA-2019-V2.0-Traducción-Español.pdf», 2019, Accedido: dic. 28, 2020. [En línea]. [Online]. Available: https://osf.io/egjk8.

Información de Contacto de la Autora:

Marisa Raquel De Giusti
Calle 24 Nº 709
La Plata
Argentina
marisa.degiusti@sedici.unlp.edu.ar
http://sedici.unlp.edu.ar
ORCID iD: https://orcid.org/0000-0003-2422-6322

Marisa Raquel De Giusti

Ingeniera en Telecomunicaciones y doctora en Ciencias Informáticas (UNLP), investigadora de la Comisión de Investigaciones Científicas (CIC). Directora de PREBI-SEDICI (UNLP) y CESGI (CIC).