ISSN 1695-5498

HTXT

HIPERTEXT.NET · REVISTA ACADÉMICA SOBRE DOCUMENTACIÓN DIGITAL Y COMUNICACIÓN INTERACTIVA

Universitat Pompeu Fabra Barcelona

Número 27 · Noviembre de 2023

Planes de gestión de datos brasileños en DMPTool: caracterización y diversidad de datos científicos

Ketlen Stueber
Instituto Brasileiro de Informações em Ciência e Tecnologia

Laura V. R. Rezende
Universidade Federal do Goiás

Elizabete Cristina de Souza de Aguiar Monteiro
Universidade Estadual Paulista

Fabiano Couto Corrêa da Silva
Universidade Federal do Rio Grande do Sul

Rômulo Arantes Alves
Universidade Federal do Goiás

Alexandre Faria de Oliveira
Instituto Brasileiro de Informações em Ciência e Tecnologia

Overview of Brazilian data management plans in DMPTool: Scientific data characterization and diversity

RESUMEN

El estudio presenta un análisis de los planes de gestión de datos públicos elaborados en la herramienta DMPTool, centrándose en los planes elaborados por investigadores brasileños. El objetivo es verificar el área de conocimiento, instituciones de investigación y agencias de fomento científico, además de los tipos de datos generados. La investigación con enfoque cualitativo y cuantitativo utiliza técnicas de análisis de contenido para recolectar, organizar e interpretar datos. El estudio presenta orientaciones para futuras investigaciones en esta área y destaca la necesidad de incrementar el proceso de formación y apoyar a los investigadores para que puedan comprender la dinámica de planificación y gestión de datos científicos para que, con el tiempo, se convierta en una práctica incorporada a la práctica científica en de una manera tan general. Los principales resultados apuntan que el área de Ciencias de la Salud presenta 166 planes registrados, conformando la mayoría del total. En términos de instituciones, la mayoría de los planes estaban asociados a universidades e institutos de investigación de la región Sudeste de Brasil. El análisis tipológico respecto al Origen de los datos es mayoritario, el 42,2% son derivados, mientras que el análisis tipológico respecto a la Naturaleza arroja que el 67,8% de los datos recogidos y generados son de textos e imágenes. Se concluyó que la herramienta para la elaboración de planes de gestión de datos DMPTool es un recurso que favorece la inclusión del ecosistema científico brasileño en la lista de mejores prácticas en Ciencia Abierta, en particular la gestión de datos científicos.

PALABRAS CLAVE

Ciencia Abierta; Datos científicos; Gestión de la información; Preservación digital.

ABSTRACT

The study presents an analysis of the public data management plans developed in the DMPTool, focusing on the plans developed by Brazilian researchers. The objective is to verify the area of knowledge, research institutions, funding agencies, and the types of data generated. A qualitative-quantitative research approach uses content analysis techniques for data collection, organization, and interpretation. The study presents reflections for future research in this area and highlights the need to increase the training process and support for researchers so that they can understand the dynamics of planning and management of scientific data so that over time it becomes a practice incorporated into scientific work in general. The main results indicate that the area of Health Sciences stands out with 166 registered plans, forming the majority of the total. In terms of institutions, most plans were associated with universities and research institutes in the Southeast region of Brazil. The typological analysis regarding the Origin of the data forms the majority, 42.2% are derivatives, while the typological analysis regarding the Nature results that 67.8% of the data collected and generated are from texts and images. It is concluded that DMPTool is a resource that favors the insertion of the Brazilian scientific ecosystem in the list of best practices in Open Science, especially the management of scientific data.

KEYWORDS

Open Science; Scientific data; Information management; Digital preservation.

Stueber, K., Rezende, L.V. R., Monteiro, E.C.S.A; Da Silva, F.C.C, Alves, R.A. y De Oliveira, A.F. (2023). Planes de gestión de datos brasileños en DMPTool: caracterización y diversidad de datos científicos. *Hipertext.net*, (27), 47-56. https://doi.org/10.31009/hipertext.net.2023.i27.05

1. Introducción

Las iniciativas para abrir la investigación científica, así como la gestión de los datos generados durante las investigaciones, se ampliaron a nivel internacional. Un ejemplo de esta relevancia y alcance se puede ver en el documento de recomendación sobre Ciencia Abierta publicado por la UNESCO en 2021, que considera que prácticas científicas más abiertas, transparentes, colaborativas e inclusivas, combinadas con conocimientos científicos más accesibles y verificables, mejoran la reproducibilidad y el impacto de la ciencia. Por lo tanto, aumenta la confiabilidad de la evidencia necesaria para una toma de decisiones y políticas sólidas, así como una mayor confianza en la ciencia (UNESCO, 2021).

Cada vez más, los gobiernos, las agencias de financiación y las universidades requieren o instan a los investigadores a planificar no sólo sus investigaciones científicas, sino también aspectos relacionados con los datos que se recopilarán o generarán. Para ello, una recomendación relevante es el desarrollo de Planes de Gestión de Datos (PGD) en la fase inicial o de planificación de las investigaciones. Las agencias de financiación científica pertinentes de todo el mundo están empezando a exigir, al solicitar financiación, que los investigadores también preparen y presenten un PGD vinculados a proyectos de investigación.

En general, el PGD es un documento formal, normalmente preparado antes del inicio del proyecto de investigación y a lo largo del desarrollo de los proyectos, que describe cómo se recopilarán, generarán y tratarán los datos. También puede presentar detalles sobre el software u otros recursos a utilizar, dónde se almacenarán los datos, cómo se puede acceder a ellos, qué licencias de uso se practicarán, entre otras informaciones.

En Brasil, la utilización del PGD como requisito obligatorio en las convocatorias para recibir fondos públicos (totales o parciales) para investigación ha mostrado una tendencia creciente, convirtiéndose en una realidad cada vez más en el escenario científico nacional. El caso más expresivo es el de la *Fundação de Amparo à Pesquisa do Estado de São Paulo* (FAPESP), que desde septiembre de 2020 exige un PGD a los investigadores como condición para financiar investigaciones en todas las áreas del conocimiento (Arantes, 2020). Aunque se trata de un paso importante, considerando que esta exigencia proviene de la agencia de desarrollo más destacada a nivel regional, la ausencia de una estructura centralizada orientada a la integración sistemática de los aspectos técnicos y tecnológicos de la gestión y la compartición de los PGD contemplando la realidad brasileña y en consecuencia todo el ámbito nacional.

La siguiente tabla presenta una descripción general de algunas de las herramientas disponibles para contribuir en la creación y gestión de Planes de Gestión de Datos, ayudando a los investigadores a cumplir con los requisitos de las agencias de financiación y promover el intercambio y la reutilización de datos científicos.

Nombre de la herramienta	Descripción
DMP online	Impulsado por el *Digital Curation Centre* (DCC) del Reino Unido, proporciona una plataforma basada en web para crear, compartir y exportar DGP.
Opidor DMP	Herramienta francesa que ayuda en la creación y gestión de PGD, ofreciendo plantillas y directrices adaptadas a los requisitos locales y europeos.
Data Stewardship Wizard	Solución de software para la creación de PGD con funciones de orientación y plantillas personalizables.
DMP NSD	Creada por el Centro Noruego de Datos de Investigación (NSD), es una herramienta que ofrece modelos y directrices específicos para los investigadores de Noruega.
Argos	Impulsado por OpenAIRE, ayuda a los investigadores a crear, compartir y mantener PGD de acuerdo con las directrices y requisitos de las agencias de financiación europeas.
Easy DMP	Solución noruega que simplifica el proceso de creación y gestión de PGD al ofrecer plantillas, orientación y funciones de colaboración.
DMPTool	Herramienta ampliamente utilizada que ofrece pautas, ejemplos y recursos para ayudar a los investigadores a crear PGD de acuerdo con las pautas y requisitos de las agencias de financiación, siguiendo las mejores prácticas internacionales.
RDMO (Research Data Management Organizer)	Es una herramienta de código abierto que permite la creación y gestión de PGD colaborativos, además de proporcionar recursos para apoyar la toma de decisiones sobre la gestión de datos.

Tabla 1. Herramientas para la elaboración de Planes de Gestión de Datos, citadas en las referencias. Fuente: elaboración propia.

Entre las herramientas presentadas, se destaca el DMPTool, de idioma inglés, desarrollado por la Universidad de California, Estados Unidos, ofrece recursos de traducción al portugués y español, además de seguir los principales estándares y recomendaciones para la elaboración de PGD existentes encaminados a la integración sistematizada de información, permitiendo el diseño de PGD activados por máquina (maDMP). Esta integración permite, por ejemplo, la integración con identificadores únicos de publicaciones científicas,

conjuntos de datos, ORCID, entre otra información presente en varios sistemas.

DMPTool es la herramienta de elaboración de PGD más utilizada por los investigadores brasileños. Este hecho motivó el presente estudio, que buscó analizar cómo investigadores brasileños desarrollaron PGD utilizando el DMPTool. Así, surgen las siguientes preguntas: ¿*Los PGD de investigadores brasileños presentes en DMPTool están vinculados a investigaciones en qué áreas del conocimiento? ¿Cuáles son las instituciones de investigación registradas en la herramienta con información personalizada? ¿Cuáles son las características de los datos que generarán las encuestas descritas en los PGD?*

Frente a estas preguntas, este estudio tiene como objetivo identificar: las áreas de conocimiento de donde se originan los proyectos de investigación de los PGD brasileños, sus respectivas instituciones y los tipos de datos a generar/crear durante el proceso investigativo. La idea es proponer una reflexión sobre las posibles complejidades y diversidades identificadas en los temas y tipos de datos generados, con el objetivo de resaltar la importancia de planificar la investigación científica, especialmente la información registrada en los PGD.

2. Datos científicos y su caracterización

Los datos científicos son fundamentales para el desarrollo de la investigación, ya que son la base sobre la que se realizan los descubrimientos y conclusiones. A través de los datos se produce evidencia científica que permite a los investigadores probar sus hipótesis, verificar sus teorías y evaluar sus descubrimientos. Los datos deben recolectarse de manera sistemática, precisa y objetiva, de modo que se consideren confiables y, preferiblemente, deben estar disponibles para que puedan ser replicados y reutilizados por otros investigadores. Esto es fundamental para la construcción del conocimiento científico, brindando colaboración, solución de problemas complejos, transparencia y confiabilidad en la validación de las conclusiones.

Borgman (2015) afirma que el concepto de dato trasciende un significado de esencia pura o natural, ya que la existencia de datos exige contextos de origen e interpretación por parte del observador. De esta manera, los datos científicos pueden entenderse como un sustrato de la ciencia, sirviendo de base para el avance del conocimiento en diversas áreas, caracterizándose por aspectos terminológicos, ejemplos, definiciones operativas, categorías, tipologías o por niveles de procesamiento.

Los aspectos terminológicos se refieren al análisis y definición de términos y conceptos utilizados en el área de datos científicos, con el objetivo de aclarar su significado y evitar malentendidos. *Las definiciones operativas* son definiciones que especifican los procedimientos y métodos utilizados para recopilar, procesar y analizar datos científicos. Describen, por ejemplo, qué instrumentos y técnicas se utilizan, cómo se almacenan y organizan los datos y cómo se llevan a cabo los análisis estadísticos.

El *National Science Board* (2005), afirma que los "datos" pueden generarse a través de observaciones, cálculos o experimentos con el fin de proporcionar cualquier información que pueda almacenarse en formato digital (textos, números, imágenes, audio, videos o películas). software, ecuaciones, algoritmos, modelos, simulaciones, entre otros).

Los enfoques categóricos consisten en clasificar los datos científicos en categorías o grupos, en función de sus características y propiedades, considerando las infraestructuras, organizaciones e individuos involucrados en los procesos de generación, almacenamiento, uso y reutilización de datos. Esto le permite identificar patrones y relaciones entre diferentes tipos de datos.

Las tipologías son *clasificaciones* que agrupan datos científicos en función de características como su origen, naturaleza de los datos (cualitativos o cuantitativos, además de caracterizarse también como texto, imágenes, audio, entre otros), la fuente de los datos (primaria o secundaria), el tipo de análisis (descriptivo, inferencial, etc.) y otros criterios relevantes.

Los *niveles de procesamiento,* por otro lado, hacen referencia a las diferentes etapas por las que pasan los datos científicos a lo largo del proceso de investigación, desde la recolección hasta el análisis y la interpretación. Esto incluye limpiar los datos, transformar datos sin procesar en formatos estandarizados, integrar diferentes fuentes de datos y aplicar análisis estadísticos o técnicas de aprendizaje automático.

Para esta investigación, los datos científicos extraídos del corpus de análisis fueron clasificados con base en los tipos de *Origen* y *Naturaleza* (National Science Board, 2005). Se entiende por naturaleza de los datos los formatos en que se generan: textos; números; imágenes; audios y; vídeos. La tipología de datos en función de su *origen* se clasifica en orden: observacionales; experimental; computacional; derivados y; intermediarios. Estas tipologías abordan diferentes dimensiones de los datos científicos y ayudan a proporcionar una comprensión más detallada de las formas en que los datos pueden recopilarse, analizarse y utilizarse en la investigación, clasificándose de la siguiente manera:

a) Datos observacionales: se recopilan directamente a través de la observación del objeto o evento, son registros históricos que deben archivarse indefinidamente. Por ejemplo: datos meteorológicos, datos de sensores ambientales y datos de imágenes de satélite.

b) datos experimentales: recopilados a través de experimentos controlados, en los que el recolector manipula una o

más variables para estudiar el efecto sobre otras variables. Incluyen datos de ensayos clínicos, datos de laboratorio y datos de campo.

c) datos computacionales: son generados por computadoras, incluidas simulaciones, modelos matemáticos y algoritmos. Pueden generarse de forma independiente o derivarse de otros tipos de datos.

d) datos derivados: se generan a partir de otro tipo de datos, mediante procesos de transformación, análisis o combinación. Se pueden utilizar para apoyar la toma de decisiones, identificar tendencias y describir relaciones entre variables.

e) datos intermedios: se recogen durante investigaciones preliminares y se utilizan para verificar las posibles variaciones de un experimento o para analizar datos recogidos en diversas circunstancias y obtener de este conjunto sólo los resultados que consideran más interesantes.

3. Plan de gestión de datos (PGD) y la herramienta DMPTool

Como se presentó en la parte introductoria de este estudio, la planificación de la gestión de datos científicos puede implicar la creación de un documento que delinee la investigación científica, el Plan de Gestión de Datos (PGD). Se destaca aquí su importancia en el sentido de considerar los usos y potenciales (re)usos de los datos generados, además de posibilitar también informar elementos relacionados con la preservación a largo plazo. Las herramientas para la preparación del PGD brindan orientación sobre aspectos y recomendaciones relacionadas con la gestión de datos científicos que cumplen con los requisitos previos de los agentes financiadores. Los PGD pueden seguir modelos estandarizados con estructuras predefinidas, con preguntas abiertas o cerradas.

DMPTool es una herramienta online gratuita y de código abierto, disponible para todos los usuarios, independientemente de su afiliación institucional (universidades y centros de investigación), orientada a la creación de PGD. Se utiliza comúnmente en Estados Unidos y países de América Latina. Proporciona una interfaz de búsqueda para recuperar planes que son públicos, con filtros: organismo financiador, institución, idioma y materia.

La herramienta proporciona instrucciones que guían a los usuarios para crear un PGD. Además, proporciona *plantillas* de diferentes instituciones y directrices que cumplen con los requisitos de los financiadores. Las preguntas a responder son agrupadas según el modelo *de Digital Curation Center*, abordando las siguientes cuestiones involucradas en la gestión de datos (DMPTool, 2023):

a. **Recopilación de datos** (cuáles y cómo se recopilarán o crearán los datos),

b. **Documentación y Metadatos** (qué documentos y metadatos acompañan a los datos);

c. **Ética y cumplimiento legal** (cómo se gestionan las cuestiones éticas, cómo se gestionarán los derechos de autor y de propiedad intelectual)

d. **Almacenamiento y copia de seguridad** (cómo se almacenarán y respaldarán los datos durante la encuesta, así como solicitar información sobre gestión de acceso y seguridad)

e. **Selección y Preservación** (qué datos tienen valor a largo plazo y deben aceptarse, compartirse y/o preservarse y si existe algún plan de preservación a largo plazo para el conjunto de datos);

f. **Intercambio de datos** (cómo se compartirán los datos y si se aplican restricciones necesarias) y;

g. **Responsabilidades y recursos** (quién será responsable de la gestión de datos y qué recursos se necesitan para llevar a cabo la recopilación/generación de datos)

4. Procedimientos metodológicos

Se trata de una investigación mixta, que contempla enfoques cualitativos y cuantitativos (Creswell & Clark, 2017). Desde el punto de vista de los procedimientos y en línea con los objetivos propuestos, la investigación se basa en el Análisis de Contenido (Bardin, 2016) para los procesos de recogida, interpretación y análisis del conjunto de datos estudiados. Para Bardin (2016), el análisis de contenido (CA) se centra en la identificación y sistematización de la información y se organiza en tres etapas principales: ***preanálisis***, ***exploración material*** y ***tratamiento de resultados e interpretaciones***. La etapa ***de preanálisis*** incluye la búsqueda y organización del material a analizar. En el contexto de este estudio, hay investigaciones realizadas en la herramienta DMPTool, con foco en planes disponibles públicamente, realizadas por investigadores de instituciones brasileñas. La colección tiene un corte temporal que va desde la creación de la herramienta en 2011 hasta el 27 de julio de 2022 (fecha de inicio de esta investigación) abarcando 404 PGD brasileños.

La segunda etapa del análisis de contenido se refiere a la ***exploración del material***. Para Bardin (2016), esta etapa consiste en administrar las técnicas sobre el ***corpus*** y elaborar las categorías. Para este estudio, luego del perfeccionamiento realizado entre los 404 planes brasileños recolectados con el fin de suprimir posibles repeticiones, se mantuvieron 393

planes. Los datos recopilados se ingresaron en tablas para su análisis. Estos datos son formados por los campos de conocimiento atribuido mediante el análisis del área de formación de los investigadores que crearon los planos y sus respectivos programas de posgrado, la universidad que pertenecen y diante de la secuencia de respuestas insertadas en las preguntas planteadas en el plan de gestión de datos:

a. Recopilación de datos

b. Documentación y Metadatos

c. Ética y cumplimiento legal

d. Almacenamiento y copia de seguridad

e. Selección y Preservación

f. Intercambio de datos

g. Responsabilidades y recursos

Los ejes organizados habilitan la creación de diferentes publicaciones. El enfoque deste artículo son los ítems "a" y "b" (recopilación de datos, documentación y metadatos).

La tercera etapa del análisis se refiere al ***tratamiento de los resultados y las interpretaciones*** que se llevó a cabo en dos fases: En la primera fase se obtuvieron los resultados directamente en la página oficial de DMPTool. En este sentido, se presentan estadísticas generales de los ejes analisados, seguidas de un panorama de la investigación brasileña que se muestra en la página de Planes Públicos de la herramienta [1]. Se buscó verificar el organismo financiador, la institución del investigador, el idioma y el tema de la investigación. Mientras que el segundo movimiento analiza los planes públicos brasileños centrándose en la clasificación tipológica (origen y naturaleza) de los datos científicos generados y producidos.

Los planes fueron recolectados mediante un formulario de llenado creado por ***Google Forms***[2], en el que se enumeraban las preguntas relacionadas a los PGD de la muestra, más la opción de categorización de áreas de conocimiento con base en la entidad brasileña del Conselho ***Nacional de Desenvolvimento Ciêntífico e Técnológico*** (CNPq, 2022).

Específicamente en este trabajo se analizan los tipos de datos generados y compartidos considerando lo declarado por los investigadores en los PGD. La clasificación de datos se basa en dos categorías preestablecidas según el fundamento teórico del ***National Science Board*** (2005) para la tipología de datos de ***Origen*** (observacional; experimental; derivado; intermedio y; computacional) y ***Naturaleza*** (texto; números; imagen; audio y; video).

La elección de las categorías de clasificación de datos se debe a que son ampliamente reconocidas en la literatura científica como una forma coherente de describir y clasificar los diferentes tipos de datos generados y compartidos en la investigación científica. La categorización según su origen permite distinguir diferentes tipos de datos según cómo fueron obtenidos, mientras que la categorización según Naturaleza permite distinguir los diferentes formatos en los que se crean/generan y almacenan los datos.

5. Resultados

Desde su creación en 2011 hasta el 27 de julio de 2022, DMPTool se contabilizaron 73.543 planes, 75.524 usuarios registrados y 330 instituciones participantes (DMPTOOL, 2023). En la página de búsqueda es posible recuperar los PGD públicos por campos: Agencia financiadora; Institución; Idioma; Sujeto. Esta página presentó 722 planes públicos vinculados a 197 instituciones, de las cuales 153 eran de múltiples países y 44 brasileñas. Del universo de 722 planes indica que 468 (en promedio, 64%) son de investigadores brasileños, disponibles en acceso abierto. Con el objetivo de presentar un panorama de los planes públicos brasileños, se identificaron tres categorías para el análisis: sujeto, financiador e institución.

5.1. Planes públicos brasileños categorizados por sujeto, agencia financiadora e institución

Para el análisis de esta primera fase, cabe resaltar la diferencia de resultados entre las categorías (sujeto; agencia financiadora ; institución), en la herramienta se utiliza el término *financista y tiene el mismo significado que agencia financiadora*. Considerando también la presentación de resultados extraídos directamente de las estadísticas de la herramienta, cada categoría presenta resultados diferentes al compararse entre sí. Por ejemplo, el número de planes públicos brasileños es de 468 considerando la categoría *Asunto/tema*, mientras que el número total de planes públicos brasileños por Organismo *de Financiamiento* correspondió a 302 y por *Institución* surgieron 389 resultados. Con el análisis de los datos, se verificó que tal diferencia se debe a que hay PGD en los que no se completaron los campos que indican la entidad financiadora (Financiador) y la institución del investigador (Institución). También se identificó que en el ámbito de la agencia financiadora, el término *patrocinador de la investigación en sí* o declaraciones de que los estudios no cuentan con financiamiento. Todas estas cuestiones implican la variabilidad de las estadísticas planteadas.

La categoría Asunto presenta el universo de 468 PGD, con indicación de 38 temas que representan diferentes áreas del conocimiento. Estas fueron registradas/declaradas en preguntas abiertas respondidas por los propios investigadores. Como ejemplo se puede destacar la categoría indicada como Ciencias Biológicas en la herramienta DMPTool. Esta categoría presentó la siguiente variación: Ciencias de la Tierra

y ambientales afines; Ciencias biologicas; Ciencias Naturales; Otras ciencias naturales.

Así, al enmarcar los temas encontrados con la Tabla de Áreas de Conocimiento del CNPq (2022), se verifica la presencia de planes en diez grandes áreas. La mayoría de los planes proceden de Ciencias Biológicas, seguidas de Ciencias de la Salud y Ciencias Humanas, que en conjunto representan 253 planes, más de la mitad del total. Otros 215 planes se distribuyen en las cinco áreas restantes: Ciencias Exactas y de la Tierra; Ciencias Sociales Aplicadas; Ingeniería; Ciencias Agrícolas; Lingüística, Letras y Artes.

categoría ***Financiador/Agencia de desarrollo*** suman 302 planes de 10 instituciones brasileñas. Se destaca el papel protagónico de la ***Fundação de Amparo à Pesquisas do Estado de São Paulo*** (FAPESP), que solicita la realización de planes de gestión de datos como prerrequisitos para la asistencia a la investigación. Las agencias federales Capes y CNPq, por no haber iniciado la exigencia para la elaboración del PGD, presentan una cantidad menor que las vinculadas a la FAPESP. Al observar el cumplimiento de los planes, se advierte, en algunos casos, se mencionan a las instituciones en general, sin distinguir si son agencias de promoción de la investigación o si son instituciones de enseñanza, como las Universidades. Esta característica evoca dos hipótesis. La primera es que tal actitud puede surgir de la necesidad de responder a los campos a completar en los planes para no dejarlos en blanco y/o considerar a la institución a la que se está vinculado como la fuente que proporciona los fondos. La segunda hipótesis sería la relación de los investigadores y sus proyectos de investigación con los Programas de Posgrado (PPG) que también pueden financiar algunas etapas de la investigación durante los procesos de investigación, recolección y análisis de datos. Este tipo de incentivo convierte a los Programas de Posgrado en potenciales financiadores de investigaciones, considerando la cultura científica brasileña.

La categoría ***Institución*** se refiere a 44 instituciones brasileñas. El registro de instituciones en diferentes campos del formulario del PGD (institución educativa y/o financiadora), el uso de nomenclaturas variadas (algunas utilizaron sólo siglas de las instituciones, otras informaron el nombre completo) y traducciones de los nombres oficiales generaron el registro de 389 planes de gestión de datos. Sin embargo, al analizar y depurar la información obtenida, se determinaron 29 instituciones brasileñas representadas por 386 planes, y no 44 según los resultados obtenidos inicialmente en el campo de búsqueda del DMPTool .

La tabla 2 ofrece una lista de los planes de gestión de datos completados por investigadores de las siete instituciones más representativas en la elaboración de planes de gestión a disposición pública.

Institución	Planes registrados
Universidade de São Paulo (USP)	136
Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP)	101
Universidade Estadual de Campinas (UNICAMP)	57
Universidade Federal de São Carlos (UFSCAR)	26
Universidade Federal De São Paulo (UFSP)	26
Universidade Federal do ABC (UFABC)	06
Instituto Tecnológico de Aeronáutica (ITA)	04
Total	**356**

Tabla 2. Principales instituciones brasileñas registradas en DmpTool. Fuente: elaboración propia.

En esta tabla se muestra que las cinco primeras instituciones listadas provienen del Estado de São Paulo y corresponden a la generación del 89% de los PGDs. Otras 18 instituciones tienen sólo un PGD registrado. Reiteramos una vez más la importancia del trabajo de las entidades que financian la investigación y el énfasis de la Fapesp en su espíritu pionero al priorizar acciones de Ciencia Abierta que involucran la planificación de la gestión de datos científicos. Entre las instituciones enumeradas, más del 90% están ubicadas en la región Sudeste de Brasil.

5.2. Planes de gestion públicos brasileños, procesos de minería de datos y su tipología según áreas de conocimiento

La recopilación de todos los planes de gestión disponibles públicamente generó una hoja de cálculo con 404 respuestas que fueron cuidadosamente refinadas para que no hubiera errores en el análisis de planes repetidos a través de creaciones duplicadas, durante su registro en Google Forms, o incluso al crear planes duplicados en el ***mismo proyecto*** por el investigador que hubiera utilizado el DMPTool. Tras este perfeccionamiento, se validaron 393 planes que, posteriormente, se contabilizaron y organizaron dentro de sus respectivas áreas de conocimiento (ver tabla 3).

Para evitar duplicaciones de planes, fue necesario atribuir a la investigación de perfil multidisciplinario el reconocimiento de apenas una gran área. Se estableció como criterio la verificación de los currículos del investigador en la ***Plataforma Lattes***[3], el área del Programa de Posgrado y el objeto declarado de la investigación, por ejemplo, investigación en Biomedicina, declarada interdisciplinaria en Ciencias Biológicas y Ciencias

Área de conocimiento	Número de planes
Ciencias de la salud	166
Ciencias Humanas	51
Ciencias Sociales Aplicadas	37
Ciencias Exactas y de la Tierra	35
Ciencias Agrícolas	32
Ingeniería	28
Ciencias biologicas	27
Lingüística, Letras y Artes	17

Tabla 3. Distribución de planes de gestión de datos en diferentes áreas de conocimiento. Fuente: elaboración propia.

de la Salud, cuando se desarrolle para estudios dirigidos sobre la curación de alguna enfermedad se consideraron estudios pertenecientes al área de Ciencias de la Salud.

A partir de esta recopilación y de las respuestas obtenidas, buscamos comprender la relación entre los investigadores y los tipos de datos que obtuvieron y generaron para el desarrollo de sus estudios, según sus áreas de conocimiento. La cuantificación de la tipología de datos se realizó tanto de forma general (considerando valores unitarios de clasificación) como de forma combinada, considerando los conjuntos de datos presentes en cada encuesta.

5.2.1. Clasificación de los tipos de datos de los PGD según su origen y naturaleza

La Tabla 4 presenta la cantidad de recurrencia de los tipos de datos desde el punto de vista de su Origen (observacional; experimental; derivados; intermedio; computacional) y Naturaleza (texto; números; imagen; audio y; video), con base en la información contenida en los PGD disponibles en acceso abierto (planes públicos) en DMPTool.

Los resultados demuestran la cantidad unitaria, es decir, las investigaciones analizadas que tienen uno o más tipos de datos implícitos en cada estudio y los resultados presentados consideran el total general de los datos de forma unitaria. Por ejemplo, hay estudios que se clasificaron en base a un solo tipo de datos (según su origen y naturaleza), mientras que otros de la misma área de conocimiento se clasificaron con dos o más tipos de datos en cuanto a su naturaleza y origen.

A la vista de los resultados generales respecto al origen de los datos, se desprende que los derivados y observacionales son los más recurrentes con 42,2% y 35,6% respectivamente. Sin embargo, los datos experimentales (20,1%) también desempeñan un papel destacado en la investigación brasileña. Se identificaron en menor medida datos intermedios (1,4%) y computacionales (0,5%).

Datos	**Áreas de conocimiento**								**Ocurrencias**
tipología: origen	CS	CH	CET	CB	CSA	CA	E	LLA	Parcial total
de observación	107	6	22	22	3	25	18	--	203 [35,6%]
Experimental	64	--	19	10	--	8	14	--	115 [20,1%]
Derivados	94	50	12	11	37	10	10	17	241 [42,2%]
Intermedios	4	1	1	--	2	--	--	--	08 [1,4%]
computacional	--	--	2	--	--	--	1	--	03 [0,5%]
Total general									570 [100%]
tipología: naturaleza									Parcial total
Texto	75	47	14	11	28	9	9	15	208 [30.15]
Números	124	15	26	23	17	29	25	2	261 [37,7%]
Imagen	64	13	18	10	7	5	13	3	133 [19,2%]
Audio	11	14	2	1	9	2	--	8	47 [6,8%]
Video	11	11	5	1	7	--	1	6	42 [6,0%]
Total general									**691 [100%]**

Tabla 4. Cantidad de datos de Origen y Naturaleza producidos por encuestas brasileñas declaradas en los Planes de Gestión de Datos de DMPTool (hasta el 27 de julio de 2022). Leyenda: CS - Ciencias de la Salud; CH- Ciencias Humanas; CET- Ciencias Exactas y de la Tierra; CB - Ciencias Biológicas; CSA - Ciencias Sociales Aplicadas; CA - Ciencias Agrícolas; E – Ingeniería y; LLA Lingüística, Letras y Artes. Fuente: elaboración propia.

En cuanto a la naturaleza de los datos, es decir, su tipo de fuente documental, estos se distribuyen en : datos numéricos (37,7%); textos (30,15%); imágenes (19,2%); audio (6,8%) y vídeo (6,0%).

Al cruzar los resultados de los tipos de datos con las diferentes áreas del conocimiento, se observa que la investigación realizada en Ciencias de la Salud genera mayoritariamente datos observacionales y derivados, respectivamente, seguidos de datos experimentales e intermedios. La naturaleza de los datos en esta área se basa principalmente en datos numéricos, textuales e imágenes, respectivamente, seguidos en menor cantidad por datos de audio y video.

En Ciencias Humanas destacan los datos derivados, provenientes de datos con fuentes documentales mayoritariamente textuales y seguidos proporcionalmente de datos numéricos, de imagen, audio y vídeo. Mientras que las Ciencias Exactas y de la Tierra distribuyen sus estudios proporcionalmente entre datos observacionales y experimentales, seguidos de datos derivados. Los datos numéricos están presentes en mayor cantidad en los estudios de este campo (Ciencias Exactas y de la Tierra) que también se apropian de datos de imágenes y textos en una proporción considerable, seguidos de datos de vídeo y audio, en menor frecuencia.

Las Ciencias Biológicas desarrollan gran parte de sus estudios a partir de datos observacionales y a través de datos experimentales y derivados. El enfoque cuantitativo es predominante en investigaciones que buscan resultados obtenidos a partir de datos numéricos. Proporcionalmente, los datos obtenidos y generados en formato texto e imagen aparecen en secuencia. Las Ciencias Sociales Aplicadas recurren a investigaciones que producen mayoritariamente datos derivados, distribuidos en gran medida por datos textuales y numéricos, seguidos de datos de audio, imágenes y vídeo en cantidades proporcionales.

Las Ciencias Agrícolas se basan principalmente en datos observacionales, seguidos de datos experimentales y derivados. En gran medida, este campo utiliza datos numéricos y, en menor medida, datos textuales, de imagen y de audio. La ingeniería se centra en la investigación basada en datos observacionales y, posteriormente, en base a datos experimentales y derivados. Se resaltan los datos numéricos, seguidos de los datos de imagen y texto; los datos de la fuente videográfica tuvieron solo una ocurrencia. Lingüística, Letras y Artes se basan en datos derivados en su totalidad, distribuidos en datos textuales en su mayor parte, seguidos de audio y vídeo y en menor medida de imágenes y datos numéricos, respectivamente.

6. Consideraciones Finales

Como resultados obtenidos destaca el área de Ciencias de la Salud con 166 planes registrados, conformando la mayoría del total. En términos de instituciones, la mayoría de los planes estaban asociados a universidades e institutos de investigación de la región Sudeste de Brasil. Con relación al análisis tipológico sobre el Origen de los datos, que forman la mayoría, el 42,2% son derivados, mientras que el de Naturaleza arroja que el 67,8% de los datos recogidos y generados son de textos e imágenes.

Los resultados obtenidos con relación a los datos generados revelan su diversidad de tipologías en su origen y naturaleza. Además, la discrepancia entre el número de PGD por áreas de conocimiento, con énfasis en Ciencias de la Salud, demuestra la relevancia de las discusiones sobre el tema entre todas las disciplinas y la conciencia de que un PGD no está diseñado sólo para atender una convocatoria pública de financiación, y puede ayudar en la gestión de datos durante todo el proceso de investigación.

Es importante señalar que la implementación de lo presentado en el PGD es un proceso complejo y continuo, que requiere colaboración y adaptación constante a los cambios en el ecosistema científico. El intercambio abierto de datos científicos es fundamental para el progreso de la ciencia y se vuelve esencial que las instituciones involucradas en el ecosistema científico tengan la capacidad de gestionar y analizar datos científicos para que puedan extraer conocimientos y tomar decisiones informadas, especialmente en relación con la financiación.

La herramienta DMPTool, entre varias disponibles, ayuda a los investigadores en la creación de PGD, proporcionando una estructura de llenado intuitiva, sin embargo, el estudio demostró la inminente necesidad de detallar claramente, en un lenguaje accesible, pautas respecto de la información que debe incluirse en el PGD. Dicha instrucción puede maximizar el valor de los datos para la comunidad científica, permitiendo también nuevos descubrimientos y validaciones de investigaciones. Además, la implementación de los PGD permite a las instituciones de enseñanza e investigación gestionar de manera más efectiva los datos científicos, minimizando los riesgos asociados a su gestión, como la duplicación de datos y la falta de confianza en los mismos.

Es importante establecer estrategias y procedimientos claros para garantizar la calidad de los datos, así como mecanismos de seguimiento y evaluación. Finalmente, se debe enfatizar la importancia de compartir y abrir el acceso a los datos científi-

ficos para la comunidad, ya que impulsa la transparencia y la innovación, aumentando la eficiencia y la calidad.

En este contexto, la investigación también permitió analizar la distribución de los temas cubiertos por los Planes de Gestión de Datos, ampliando la comprensión de las áreas predominantes y sus respectivas demandas. Este conocimiento ayuda a identificar posibles motivos de participación en diferentes áreas del conocimiento en mayor o menor medida, ya sea por algún aspecto cultural específico, requerimientos de agencias de financiación de actividades de investigación, entre otros factores.

En futuros estudios, será relevante investigar estas causas, con el objetivo de comprender mejor la dinámica involucrada y mejorar aún más la implementación y el uso de los PGD. Un análisis de este tipo puede proporcionar ideas para orientar estrategias y políticas destinadas a promover una cultura más integral y eficaz de gestión e intercambio de datos en todas las áreas de la investigación científica. De esta forma, será posible maximizar los beneficios derivados de los PGD e impulsar la innovación, la transparencia y la calidad de la ciencia a escala global.

Como iniciativa posterior a este estudio, un hecho evidente fue el problema de que los formularios para la creación de DGP traían una importante cantidad de campos a cumplimentar libremente. Las implementaciones actuales de DMPTool cumplen con los requisitos de una herramienta operada por máquina (maDMP), lo que significa que cada vez se implementan más campos con información parametrizada e interoperable entre diferentes sistemas del ecosistema científico. Así, se inició la personalización de una herramienta brasileña para la creación de PGD derivada de DMPTool, adaptándola a la realidad brasileña (IBICT, 2023).

Notas al final

1. https://dmptool.org/public_plans
2. https://www.google.com/intl/pt-BR/forms/about/
3. Sistema virtual de currículos de investigadores brasileños.

Referencias

Arantes, J. T. (3 diciembre 2020). Webnary enseña cómo hacer un plan de gestión de datos. *Agência Fapesp*. https://agencia.fapesp.br/webinario-ensina-como-fazer-plano-de-gestao-de-dados/34749

ARGOS. (23 de enero. 2023). *Sitio web oficial*. Argos. https://argos.openaire.eu/splash/index.html

Bardin L. (2016). *Análisis de Contenido*. Ediciones 70.

Borgman, C. L. (2015). *Big Data, Little Data, No Data: scholarship in the networked World*. MIT Press.

Creswell, J. W. y Plano-Clark, V. L. (2017). *Designing and Conducting Mixed methods Research*. Sage

Conselho Nacional de Desenvolvimento Científico E Tecnológico - CNPq. (2022). *Tabla de Áreas de Conocimiento*. Brasilia, Brasil: CNPq. http://lattes.cnpq.br/documents/11871/24930/TabeladeAreasdoConhecimento.pdf/d192ff6b-3e0a-4074-a74d-c280521bd5f7

DSW (s/f). *Data Stewardship Wizard. [Sitio web oficial].* https://ds-wizard.org/

DMPonline. (2023). *Plan to make data work for you. [Sitio web oficial].* https://dmponline.dcc.ac.uk/

DMPOpidor. (2023). *Bienvenue ! [Sitio web oficial].* https://dmp.opidor.fr/

DMPTool. (2023). *About. [Sitio web oficial].* https://dmptool.org/about_us

Easy DMP. (2023). *Data Management Plan Generator [Sitio web oficial]*. https://easydmp.sigma2.no/

Instituto Brasileiro De Informação em Ciência e Tecnologia - IBICT. (2023). *Plano de Gestão de Dados - PGD BR*. IBICT. http://www.pgd.ibict.br

National Science Board. (2005). *Long-Lived Digital Data Collections: enabling research and education in the 21st Century*. National Science Foundation.

Norwegian Centre for Research Data. (2023). *NSD DMP – Enabling long-term preservation and sharing of Research Data. https://www.rd-alliance.org/sites/default/files/NSD-2.pdf*

RDMO. (s/f). *Welcome to RDMO. [Sitio web oficial]*. https://rdmorganiser.github.io/en/

UNESCO. (2021). *Recomendaciones de la UNESCO sobre ciencia abierta.* https://unesdoc.unesco.org/ark:/48223/pf0000379949_eng

CV

Ketlen Stueber

- ketistueber@hotmail.com
- https://orcid.org/0000-0002-2171-0365

- Doctora en Educación en Ciencias: Química de la Vida y de la Salud (PPGQVS), Universidad Federal de Rio Grande do Sul (UFRGS). Maestría en Comunicación e Información, Universidad Federal de Rio Grande do Sul (UFRGS). Especialista en Biblioteca Escolar Cultura Escrita y Sociedad en Red, Universidad Autónoma de Barcelona en colaboración con la Universidad de Barcelona y la Organización de Estados Iberoamericanos (OEI). Licenciada en Biblioteconomía - Calificación en Gestión de la Información, Universidad Estadual de Santa Catarina (UDESC). Profesora suplente de la carrera de Licenciatura en Biblioteconomía de la Universidad Federal de Rio Grande do Sul (UFRGS) en 2016 y 2017. En julio de 2022, se convirtió en becaria investigadora del Instituto Brasileño de Información en Ciencia y Tecnología (IBICT). Áreas de interés: Acciones culturales y políticas de fomento de la lectura; Imaginarios y representaciones

sociales; Epistemología y sociología del conocimiento; Comunicación Científica; Ciencia Abierta; apertura de datos.

Laura V. R. Rezende
- lauravil.rr@gmail.com
- https://orcid.org/0000-0002-8891-3263
- Colaboradora de Investigación en el Instituto de Ciencias Sociales Cuantitativas - IQSS (Universidad de Harvard) con el Equipo de Sanación Digital del Proyecto Dataverse; Investigación Postdoctoral en Ciencia Abierta: Diagnóstico Brasileño considerando el Escenario Europeo (Universidad de Barcelona); Doctorado y Maestría en Ciencias de la Información (Universidad de Brasilia -UnB); Especialista en Inteligencia Organizacional y Competitiva (Universidad de Brasilia - UnB); Especialista en Redes de Computadores (Universidad Católica de Goiás -PUC); Informático (Universidad Católica de Goiás - PUC). Profesor Asociado - Universidad Federal de Goiás - UFG) de la Facultad de Información y Comunicación.

Elizabete Cristina de Souza de Aguiar Monteiro
- elizabete.monteiro@unesp.br
- https://orcid.org/0000-0002-3797-8139
- Doctora y Magíster en Ciencias de la Información, Universidad Estadual Paulista (Unesp) en Marília. Licenciado en Biblioteconomía, Universidade Estadual Paulista (2006). Miembro del Grupo de Investigación en Tecnologías de Acceso a Datos (GPTAD) (UNESP), Laboratorio de Ecosistemas Informacionales y Estudios y Prácticas de Preservación Digital (IBICT). Bibliotecario de la Unesp de Marília desde 2009, actuando en la Sección de Técnicas de Adquisición y Procesamiento de Información y en la Sección de Referencia Técnica, Atención al Usuario y Documentación. Investigadora Becada en IBICT - Proyecto InovaInfo.

Fabiano Couto Corrêa da Silva
- fabianocc@gmail.com
- https://orcid.org/0000-0001-5014-8853
- Profesor Adjunto del Departamento de Ciencias de la Información/FABICO de la Universidad Federal de Rio Grande do Sul (UFRGS), actuando en la graduación en Biblioteconomía y en el Programa de Posgrado en Ciencias de la Información (PPGCIN), ambos de la misma institución. Licenciado en Biblioteconomía por la Universidad Federal de Rio Grande do Sul (2002), Máster en Ciencias de la Información por la Universidad Federal de Santa Catarina (2008) y Doctorado en Información y Documentación en Sociedad del Conocimiento de la Universitat de Barcelona (2017).

Rômulo Arantes Alves
- romuloarantes20@gmail.com
- https://orcid.org/0009-0000-6571-6315
- Estudiante de graduación en Bibliotecología de la Universidad Federal de Goiás (UFG), Brasil. Becario de iniciación científica con énfasis en Ciencia Abierta y apertura de datos científicos.

Alexandre Faria de Oliveira
- alexandreoliveira@ibict.br
- https://orcid.org/0000-0003-0470-4972
- Estudiante de Maestría en Gestión Estratégica de Organizaciones, Centro Universitário IESB. Postgrado en Sistemas Orientados a Objetos (JAVA), Universidad Católica, Brasilia (2008). Licenciatura en Informática, FACTU. Becario CNPQ DTI-7E del 2006-2008. Proyecto de Investigación y Desarrollo de Tecnologías de la Información para la Construcción de la Sociedad de la Información y el Conocimiento. Investigador - Consultor UNESCO en el Instituto Brasileño de Información en Ciencia y Tecnología (IBICT).entre 2010-2012: Investigación y Análisis de Software de Preservación Digital. Actualmente es investigador en el Instituto Brasileño de Información en Ciencia y Tecnología, IBICT. Coordinador de soluciones tecnológicas en el proyecto de investigación Preservación Digital - "Rede Cariniana". Tiene experiencia en el área de Ciencias de la Información, con énfasis en preservación digital, en particular en las siguientes áreas: datos de investigación, sistemas electrónicos de gestión de información, publicaciones científicas electrónicas, preservación digital, repositorios Institucionales y sistemas de Preservación Distribuida.