

CONVIÉRTETE EN UN BIBILIOTECARIO DE DATOS



LUIS ENRIQUE LESCANO BORREGO



ÍNDICE

<u>Introducción</u>	4
<u>Ciclo de vida de los datos</u>	7
<u>Creando un plan</u>	11
<u>Organización</u>	17
<u>Trabajo con archivos</u>	27
<u>Metadatos y documentación</u>	33
<u>Almacenamiento de datos</u>	39
<u>El valor de compartir</u>	45
<u>Visualización de datos</u>	52

PRESENTACIÓN

El bibliotecario de datos es un profesional capacitado para organizar, preservar y facilitar el acceso a grandes cantidades de datos. Este rol se ha vuelto especialmente relevante en universidades y centros de investigación, donde la gestión de datos de investigación es una de las áreas más demandadas. En Estados Unidos, los sueldos promedio para bibliotecarios de datos rondan los \$70,000 al año. Este campo ofrece oportunidades de crecimiento y desarrollo profesional, así como una compensación monetaria competitiva.

Esta guía es una recopilación de conocimientos adquiridos a lo largo de años de experiencia en el campo de la bibliotecología de datos. En ella, encontrarás conceptos clave, recursos recomendados y herramientas prácticas para formarte en esta área de manera gratuita.

Como parte de esta guía, recibirás una [plantilla exclusiva](#) en Notion para registrar tu aprendizaje y organizar tus conocimientos de manera eficiente.



INTRODUCCIÓN

¿QUÉ ES UN BIBLIOTECARIO DE DATOS?

Los bibliotecarios de datos son expertos en la gestión, organización y conservación de datos digitales en el ámbito de las bibliotecas. La explosión de datos en la era digital abarca una amplia gama de actividades humanas y campos de conocimiento, que van desde algoritmos hasta interacciones sociales. La bibliotecología de datos se originó en la década de 1960 y se enfoca en adaptar los servicios bibliotecarios al uso de datos digitales, repositorios y gestión de datos, requiriendo habilidades tradicionales de bibliotecarios junto con competencias en datos contemporáneas.

Este campo opera en la intersección de la ciencia de la información, la e-ciencia y la ciencia de datos. La ciencia de la información explora la interacción entre la información y la tecnología, mientras que la e-ciencia y la ciencia de datos aprovechan la computación y las técnicas estadísticas para la investigación intensiva en datos y la extracción de conocimiento. El rol del bibliotecario de datos incluye actividades como la gestión, curación y visualización de datos, así como servicios avanzados de referencia, facilitando la investigación basada en datos y garantizando la preservación y accesibilidad de los datos de investigación.

El diagrama de Venn del bibliotecario de datos también resalta la importancia de la colaboración

interdisciplinaria y el trabajo en equipo en este campo emergente. En la intersección de la ciencia de la información, la e-ciencia y la ciencia de datos, los bibliotecarios de datos deben colaborar estrechamente con investigadores, científicos de datos y profesionales de la información para asegurar la gestión eficaz de los datos, la preservación y la accesibilidad. Esto implica no solo tener habilidades técnicas y conocimientos especializados, sino también la capacidad de comunicarse de manera efectiva con diversos equipos y comprender las necesidades específicas de cada disciplina. Así, el diagrama de Venn subraya el papel crucial de los bibliotecarios de datos como facilitadores de la investigación basada en datos y la preservación del conocimiento científico en la era digital.

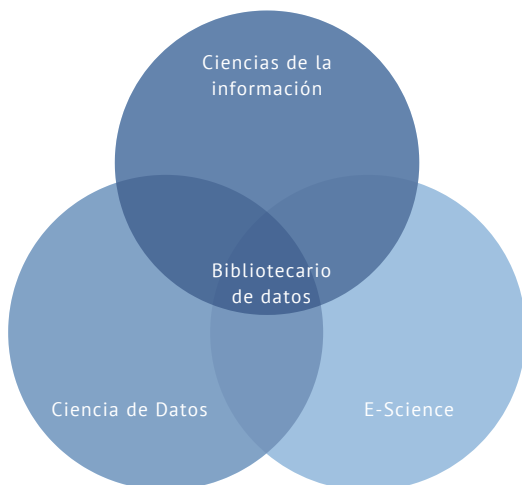


Diagrama de Venn sobre el bibliotecario de datos

Para desempeñar eficazmente su función, los bibliotecarios de datos necesitan habilidades técnicas en la gestión y manipulación de datos, así como un entendimiento profundo de las metodologías de investigación. Además, colaboran estrechamente con los investigadores para planificar la gestión de datos, crear metadatos y cumplir con los requisitos de acceso abierto, contribuyendo así al avance del conocimiento científico mediante la utilización efectiva y la difusión de los datos.

OPORTUNIDADES LABORALES

Las oportunidades profesionales para los bibliotecarios de datos están experimentando un auge notable, especialmente en Estados Unidos, donde la demanda de expertos en gestión y organización de datos digitales está en constante crecimiento. Según los datos recopilados por la plataforma [ZipRecruiter](#), se registran más de 20,000 ofertas de trabajo para roles relacionados con la bibliotecología de datos.

Los salarios para los bibliotecarios de datos son atractivos, con un rango que va desde más de \$50,000 hasta \$100,000 anuales, con una media de \$74,164 por año, según las estadísticas de ZipRecruiter para los Estados Unidos de América. Además, aproximadamente el 15% de estas oportunidades ofrecen la posibilidad de trabajo remoto, lo que amplía significativamente las opciones para profesionales en este campo.

En particular, las universidades y centros de investigación son los principales sectores que requieren servicios especializados de bibliotecarios de datos. Esto se debe al aumento de la investigación intensiva en datos y la necesidad de gestionar grandes volúmenes de información. Por lo tanto, las instituciones académicas están buscando activamente a profesionales con habilidades para la gestión de sus datos de investigación.



NOTA

Los títulos de los puestos de trabajo incluyen roles como **bibliotecarios de servicios de datos**, **bibliotecarios de datos** y **bibliotecarios especializados en la gestión de datos de investigación**.

CICLO DE VIDA DE LOS DATOS

GESTIÓN DE DATOS DE INVESTIGACIÓN

La gestión de datos es fundamental en el contexto de la investigación y el desarrollo, ya que implica una serie de actividades y prácticas destinadas a garantizar la conservación, acceso y uso efectivo de los datos a lo largo del tiempo. Estas actividades pueden abarcar desde la planificación inicial de cómo se van a recopilar y almacenar los datos, hasta la documentación detallada de los mismos, su formato, almacenamiento, anonimización y control de acceso.

¿Por qué es tan importante gestionar los datos? La gestión de datos no solo facilita el trabajo de los investigadores durante el proyecto activo de investigación, sino que también fomenta la colaboración entre ellos y optimiza el uso de los datos disponibles. Es como organizar los recibos durante el año para hacer más sencillo el proceso de declaración de impuestos; tener un plan y seguirlo no solo ahorra tiempo, sino que también evita el estrés y facilita la maximización de los recursos.

Además, la gestión adecuada de los datos garantiza su preservación a largo plazo, permitiendo que futuros investigadores puedan descubrir, interpretar y reutilizar los datos para responder a nuevas preguntas o ampliar estudios longitudinales. En esencia, la gestión de datos asegura que el valor de la información generada en un proyecto de

investigación se mantenga intacto y disponible para futuras investigaciones y aplicaciones.

Hoy en día, cada vez más agencias de financiación, editores de revistas e instituciones de investigación exigen una gestión transparente y responsable de los datos como requisito para el financiamiento y la publicación de proyectos. Esto se debe a la creciente necesidad de aumentar la transparencia en la investigación y garantizar el máximo rendimiento de la inversión realizada, promoviendo la disponibilidad y reutilización de los datos para impulsar futuras innovaciones y descubrimientos en diversos campos del conocimiento.

MODELO DATAONE

DataONE es un proyecto impulsado por la comunidad que proporciona acceso a los datos a través de múltiples repositorios de los miembros asociados, lo que permite mejorar la búsqueda y el descubrimiento de datos de las ciencias naturales y del medio ambiente. DataONE promueve las mejores prácticas en la gestión de datos a través de recursos y materiales educativos receptivos. DataONE comprende una red distribuida de centros de datos, redes científicas u organizaciones. Estas organizaciones pueden exponer sus datos dentro de la red DataONE. Además de los datos científicos, los miembros pueden proporcionar recursos informáticos, o servicios como la replicación de datos, a la comunidad DataONE.

El ciclo de vida de los datos es un proceso que describe las diferentes etapas por las que atraviesan los datos desde su creación hasta su preservación y posible reutilización. Este ciclo proporciona una estructura para gestionar y aprovechar eficazmente la información digital. El modelo de ciclo de vida de los datos propuesto por DataONE se fundamenta en la necesidad de establecer un marco estructurado y sistemático para gestionar eficazmente la información digital en todas sus etapas, desde su creación hasta su posible reutilización.

El uso de datos y las métricas de citación disponibles en la plataforma de búsqueda y descubrimiento de datos, <https://search.dataone.org>, incluyen recuentos de citas, descargas y vistas para cada conjunto de datos en la red de repositorios de Ciencias de la Tierra y del Medio Ambiente. Estas métricas son interactivas y se pueden explorar más a fondo para identificar los artículos que citan los datos o, en el caso de las descargas y las visitas, para profundizar en los detalles de cuándo se accedió a los datos. La implementación de estas métricas de uso y citación de datos es el resultado del proyecto Make Data Count y complementa las métricas recientemente implementadas en Dash.

Este enfoque surge en respuesta a los desafíos que enfrentan los investigadores y científicos en la era de la ciencia de datos, donde la cantidad y complejidad de los datos generados son cada vez mayores.

A continuación, se detallan los puntos clave de este ciclo:

1. **Planificación:** En esta fase se describe en detalle la información que se recopilará, así como la estrategia para gestionarla y hacerla accesible a lo largo del tiempo.
2. **Recopilación:** Aquí se realizan las observaciones y se transforman en datos digitales, ya sea manualmente o mediante el uso de instrumentos y sensores.
3. **Aseguramiento:** Se llevan a cabo controles y revisiones para garantizar la precisión y fiabilidad de los datos.
4. **Descripción:** Los datos son descritos de manera precisa utilizando estándares de metadatos apropiados, lo que facilita su identificación y comprensión.
5. **Preservación:** Los datos son enviados a un archivo a largo plazo, como un centro de datos, para garantizar su conservación y accesibilidad en el futuro.
6. **Descubrimiento:** Se localizan y obtienen datos potencialmente útiles junto con su información relevante, como metadatos, para facilitar su acceso.

7. **Descubrimiento:** Se localizan y obtienen datos potencialmente útiles junto con su información relevante, como metadatos, para facilitar su acceso.
8. **Integración:** Los datos de diferentes fuentes se combinan para formar un conjunto homogéneo que pueda ser analizado fácilmente.
9. **Análisis:** Finalmente, los datos son analizados para extraer conocimiento y obtener conclusiones significativas.

Modelo de ciclo de vida de DataONE



El objetivo de la gestión de datos con este modelo es producir conjuntos de datos autoexplicativos. Si entregas tus datos a un científico o colega que no ha estado involucrado en tu proyecto, ¿podrá entenderlos? ¿Podrá utilizarlos de manera efectiva y adecuada? Esta guía describe algunas prácticas fundamentales de gestión de datos que te permitirán desarrollar un plan de gestión de datos, así como también cómo crear, organizar, gestionar, describir, preservar y compartir datos de manera efectiva.

CREANDO UN PLAN

PLAN DE GESTIÓN DE DATOS

Un Plan de Gestión de Datos (PGD), también conocido como Data Management Plan (DMP) en inglés, es un documento formal que detalla los datos producidos durante un proyecto de investigación y establece las estrategias para su gestión durante y después de la fase activa del proyecto. Además, describe en detalle todas las actividades de gestión de datos que se llevarán a cabo a lo largo del ciclo de vida de los datos de investigación, desde su recopilación y organización hasta su preservación y compartición.

La demanda de los financiadores de investigación, como la Fundación Nacional de la Ciencia y otros organismos similares, de presentar Planes de Gestión de Datos se ha incrementado. Esto se debe a que estos planes garantizan que los datos producidos sean accesibles y utilizables a largo plazo, lo que apoya la transparencia y la apertura en la investigación financiada con fondos públicos. Además, maximizan el retorno de la inversión al hacer que los datos sean descubribles, accesibles y reutilizables para futuras investigaciones, lo que impulsa la innovación y el progreso científico.



Lecturas

[Gestión de datos de investigación y bibliotecas: preservando los nuevos bienes científicos](#) (2019)

[Cómo crear tu Plan de Gestión de Datos de Investigación](#) (2021)

AGENCIAS DE FINANCIACIÓN

La gestión efectiva de datos es fundamental en el ámbito de la investigación científica y académica, y las agencias de financiación de todo el mundo reconocen su importancia para garantizar la integridad, la transparencia y el valor duradero de los resultados de la investigación. A continuación, exploraremos por qué un plan de gestión de datos resulta fundamental para diversas agencias de financiamiento:

1. Institutos Nacionales de Salud (NIH) de EE.UU.:

- Los NIH tienen una política de gestión y compartición de datos que requiere que los investigadores desarrollen planes de gestión de datos para proyectos financiados.
- Un plan de gestión de datos sólido ayuda a los investigadores a cumplir con los requisitos de los NIH y promueve la transparencia y la reproducibilidad en la investigación biomédica.
- Más información: [Política de Gestión y Compartición de Datos de NIH](#)

2. Horizon 2020 de la Unión Europea:

- El programa Horizon 2020 de la Unión Europea promueve el acceso abierto a los resultados de la investigación y la gestión adecuada de los datos generados.
- Los beneficiarios de financiación deben seguir las directrices de gestión de datos de Horizon 2020, lo que incluye la preparación de planes de gestión de datos detallados.
- Un plan de gestión de datos robusto ayuda a los investigadores a cumplir con los estándares de acceso abierto y a maximizar el impacto de su investigación en la comunidad científica y la sociedad en general.
- Más información: [Directrices de Gestión de Datos de Horizon 2020](#)

3. Agencia de Financiación de Investigación del Reino Unido:

- Las agencias de financiación del Reino Unido también requieren que los investigadores desarrollen planes de gestión de datos para proyectos financiados.
- Estos planes ayudan a garantizar que los datos generados sean gestionados de manera efectiva durante y después del proyecto, maximizando así su valor y su potencial de impacto.
- Los investigadores pueden utilizar recursos y orientación proporcionados por organizaciones como el Digital Curation Centre (DCC) para desarrollar planes de gestión de datos sólidos.
- Más información: [Planes de Gestión de Datos - DCC](#)

ASPECTOS DE UN PGD

Para crear un PGD efectivo, es fundamental tener en cuenta una serie de aspectos importantes. Aquí están los puntos a considerar al crear un plan:

- **¿Qué datos se recopilarán o crearán?**

Describe los datos, incluyendo tipo, formato y volumen, justificando la elección del formato y considerando implicaciones de almacenamiento.

- **¿Cómo se recopilarán o crearán los datos?**

Explique los métodos de recopilación de datos, las normas que dictan estos métodos, y discuta planes para organizar los archivos de datos y aplicar control de versiones y protocolos de garantía de calidad.

- **¿Qué documentación de metadatos acompañará a los datos?**

Determine qué información será necesaria para que los datos sean comprensibles para usuarios secundarios, describiendo el tipo de documentación requerida.

- **¿Cómo gestionará las cuestiones éticas?**

Aborde las consideraciones legales y éticas, especialmente si implica investigación con participantes humanos, describiendo disposiciones para proteger la confidencialidad.

- **¿Cómo gestionará las cuestiones de derechos de autor y propiedad intelectual?**

Identifique al propietario de los datos, las condiciones para su uso, y los permisos necesarios si está reutilizando datos de terceros.

- **¿Cómo gestionará el acceso y la seguridad?**

Detalle las medidas de seguridad para proteger los datos y normas de cumplimiento, especialmente para datos sensibles.

- **¿Qué datos deben conservarse, compartirse y/o preservarse?**

Considere el valor potencial de los datos y los esfuerzos necesarios para su conservación y acceso futuro.

- **¿Cuál es el plan de conservación a largo plazo del conjunto de datos?**

Identifique el repositorio para archivar los datos y planifique la preparación y documentación para su conservación a largo plazo.

- **¿Cómo compartirá los datos?**

Identifique el mecanismo de compartir datos y especifique cómo otros podrán encontrar y acceder a los datos.

- **¿Se requiere alguna restricción para compartir los datos?**

Aborde las preocupaciones sobre privacidad y establezca mecanismos para superar estas restricciones si es necesario.

- **¿Quién será responsable de la gestión de los datos?**
Identifique quién supervisará la implementación del plan de gestión de datos.
- **¿Qué recursos necesitará para aplicar su plan?**
Considere los conocimientos especializados, equipos especiales y costos asociados para la gestión y conservación de datos.

Crear un PGD completo y bien pensado es crucial para garantizar la transparencia, apertura y reutilización efectiva de los datos de investigación.

Herramientas de utilidad

DMPOnline

Es una plataforma que ayuda a los equipos de investigación a cumplir con los requisitos de los financiadores para crear un Plan de Gestión de Datos (DMP). Basada en el código abierto de DMPRoadmap, ofrece plantillas adaptadas a diferentes financiadores e instituciones. Con una interfaz amigable y guías disponibles, DMPOnline facilita la gestión eficaz de los datos según las expectativas de los financiadores y las instituciones.

DMPTool

Es una herramienta gratuita y de código abierto que ayuda a los investigadores a crear planes de gestión de datos (DMP). Se inició en 2011 como un esfuerzo colaborativo entre ocho instituciones en respuesta a las demandas de agencias de financiamiento. Proporciona un asistente para crear DMP que cumplen con los requisitos de los financiadores, enlaces a sitios web de financiadores y recursos de mejores prácticas. Continúa siendo respaldada por una Junta Editorial que garantiza la actualización de requisitos y mejores prácticas.



DMPTool

Ejemplos de PGD

- [NIH sample data management and sharing plans](#)
- [UK examples from the Digital Curation Centre](#)

RECURSOS

1. [Guía Breve - Crear un Plan de Gestión de Datos Efectivo](#) (Alianza de Investigación Digital de Canadá)
2. [Guía Breve - Plan de Gestión de Datos](#) (Alianza de Investigación Digital de Canadá)
3. [Serie de Tutoriales en Video del Asistente de Plan de Gestión de Datos](#) (Alianza de Investigación Digital de Canadá)
4. [Manual - Planes de Gestión de Datos](#) (Alianza de Investigación Digital de Canadá)
5. [Guía para Preparar un Plan de Gestión de Datos](#) (SSHRC - Consejo de Investigación en Ciencias Sociales y Humanidades)
6. [Planes de Gestión de Datos para Propuestas y Premios de la Oficina de Humanidades Digitales de NEH](#) (NEH - Fondo Nacional para las Humanidades)
7. [Escribir un Plan de Gestión y Compartición de Datos](#) (NIH - Institutos Nacionales de Salud)
8. [Preparación de tu Plan de Gestión de Datos](#) (NSF - Fundación Nacional de Ciencia)
9. [Costos de Gestión de Datos](#) (Universidad de Utrecht)
10. [Herramienta y Lista de Verificación para Costos de Gestión de Datos](#) (Servicio de Datos del Reino Unido)
11. [Cómo Desarrollar un Plan de Gestión y Compartición de Datos](#) (Centro de Curación Digital)

ORGANIZACIÓN

UNA BUENA GESTIÓN DE DATOS

La gestión efectiva de los datos de investigación es una piedra angular en cualquier proyecto académico o científico. En este capítulo, exploraremos en detalle el concepto de organización de datos, también conocido como "housekeeping", así como las prácticas óptimas de gestión de archivos. Al concluir la lectura, comprenderá la importancia fundamental de estos procesos, junto con los beneficios tangibles que ofrecen la versión de archivos, las convenciones de nombres y el renombrado.

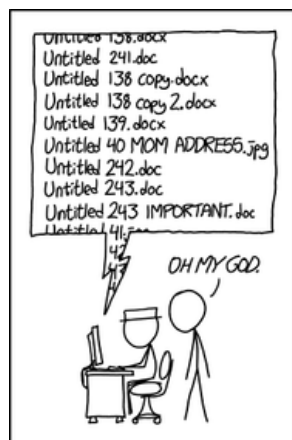
Al comienzo de un proyecto de investigación, es tentador creer que siempre recordará cómo nombrar sus archivos y dónde se encuentran almacenados. Sin embargo, una vez que la investigación está en marcha, la acumulación rápida de una gran cantidad de datos hace que mantener el control sea una tarea ardua. Múltiples archivos en diversos formatos y numerosas versiones se suman a la complejidad del panorama. Además de los archivos de datos, es probable que se generen otros elementos como sitios web, blogs, artículos, citas, metodologías, notas y hojas de cálculo, todos ellos relacionados con la investigación. Por consiguiente, encontrar un archivo de datos específico puede convertirse en una tarea frustrante y una pérdida significativa de tiempo si no se ha nombrado con precisión o consistencia.

Las buenas prácticas de gestión de archivos son esenciales para facilitar la identificación, localización y utilización efectiva de los datos. Las convenciones de nombres de archivos adquieren especial relevancia cuando se busca compartir los datos con colaboradores en un entorno grupal. Garantizar que todos los involucrados apliquen las mismas convenciones de manera consistente es crucial para evitar confusiones y maximizar la eficiencia.

Entonces, ¿cuáles son las mejores prácticas para organizar los datos? En primer lugar, los archivos y carpetas de datos de investigación deben etiquetarse y organizarse de manera sistemática para garantizar su identificabilidad y accesibilidad tanto en el presente como en el futuro. Los beneficios de esta etiquetación consistente son numerosos: los archivos son fácilmente distinguibles entre sí, más fáciles de localizar y explorar, y pueden ser recuperados por cualquier usuario autorizado. Además, la posibilidad de ordenarlos en una secuencia lógica y la protección contra sobrescrituras o eliminaciones accidentales contribuyen a mantener la integridad de los datos.

NOMENCLATURA DE ARCHIVOS

Una convención de nombres de archivos bien definida es esencial para una gestión de datos efectiva, proporcionando un marco estructurado para etiquetar los archivos de manera que refleje con precisión su contenido y relaciones con otros archivos. Al adherirse a una convención de nombres consistente, no solo se mejora la organización personal del investigador, sino que también se facilita la colaboración fluida dentro de equipos de trabajo y repositorios compartidos. Tanto si se



PROTIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

Imagen de XKCD

trabaja de manera individual como en equipo, los beneficios de seguir un esquema de nombres estandarizados son numerosos: desde una mayor capacidad de búsqueda hasta una reducción en la posibilidad de pérdida de datos y un incremento en la eficiencia en la recuperación de archivos. Por lo tanto, establecer una convención clara desde el inicio del proyecto se vuelve fundamental, evitando así la acumulación de un archivo de contenido desorganizado que podría obstaculizar el avance del trabajo y comprometer la integridad de los datos.

EJEMPLO DE NOMBRES DE ARCHIVOS

- ✓ 20240201_ProyectoA_Ex1Test1_Morales_v1.xlsx
- ✓ 20240201_ProyectoD_Notas_González_v2.docx
- ✓ Experimento_Instrumento_Fecha_ImagenID.ti

- ✗ Test data 2016.xlsx
- ✗ Notas de reunión Febrero 17.doc
- ✗ Investigación Final FINAL última versión.docx

CONSEJOS PARA NOMBRES DE ARCHIVOS

Los siguientes consejos fueron adaptados de Briney, K. A. (2020, junio 2). Hoja de trabajo de convención de nombres de archivos. Instituto de Tecnología de California. Este recurso recientemente ha sido galardonado con el Premio de Logros Significativos DataWorks! 2023.

Piensa en tus archivos:

- ¿Qué archivos relacionados estás utilizando?
- Identifica qué grupo de archivos abarcará tu convención de nombres.
- Puedes usar convenciones diferentes para diferentes conjuntos de archivos.
- Verifica si existen convenciones de nombres de archivos establecidas en tu disciplina o grupo.

Identifica metadatos (por ejemplo, fecha, muestra, experimento):

- ¿Qué información se necesita para localizar fácilmente un archivo específico?
- La computadora ordena los archivos por nombre, carácter por carácter. Por lo tanto, coloca la información más importante primero.
- El nombre del archivo debe ser descriptivo y proporcionar suficiente información contextual.

Considera incluir una combinación de la siguiente información:

- Condiciones del experimento
- Tipo de datos
- Nombre/Iniciales del investigador
- Nombre/Lugar del laboratorio
- Nombre o acrónimo del proyecto o experimento
- Fecha o rango de fechas del experimento
- Número del experimento o ID de la muestra
- Si utilizas un sistema de numeración secuencial, utiliza ceros principales para mayor claridad y asegúrate de que los archivos se ordenen de manera secuencial.



Lecturas

[Nombres de archivos y jerarquía de carpetas](#) (MIT)

[Convenciones para nombrar archivos: reglas sencillas que ahorran tiempo y esfuerzo](#) (2017)

Abrevia o codifica metadatos:

- No olvides documentar cualquier código.
- Decide qué información abreviada mantener.
- Estandariza las categorías y/o reemplázalas con códigos de 2 o 3 letras.
- Asegúrate de documentar estos códigos.

Usa versiones:

- ¿Estás manteniendo diferentes versiones del mismo archivo?
- Usa la versión para indicar la versión más actual de un archivo.
- Haz un seguimiento de las versiones de un archivo agregando información de versión al final del nombre del archivo.

Piensa en cómo buscarás tus archivos:

- ¿Qué información es más relevante?
- Decide qué metadatos deberían aparecer al principio.
- Deliberadamente separa los elementos de metadatos.

Escribe tus convenciones de nombres:

- Incluye un archivo README de nivel superior sobre cómo navegar la estructura.
- Las convenciones de nombres deben documentarse para que otros en tu laboratorio o departamento puedan seguir este estándar.
- Si encuentras que estás codificando una gran cantidad de metadatos en los nombres de los archivos, considera almacenar estos metadatos en una hoja de cálculo maestra con tus datos.

**NOTA**

Para la creación de un archivo README se utiliza generalmente el lenguaje de Markdown por lo que es necesario que lo conozcas a profundidad

Markdown



Markdown es una sintaxis de marcado ligero ampliamente utilizada en la redacción y formateo de texto, ofreciendo una forma simple y legible de dar estilo a documentos, notas, informes, páginas web y más. Fue concebido por John Gruber y Aaron Swartz con el propósito de simplificar el proceso de escritura y publicación en línea, permitiendo que los escritores se concentren en el contenido sin preocuparse por la complejidad del formato.

Una de las características más básicas de Markdown es la capacidad de crear títulos con diferentes niveles de importancia. Por ejemplo, para crear un título principal, se utiliza el símbolo de almohadilla (#) seguido de un espacio y el texto del título. A continuación te muestro un ejemplo de la sintaxis:

```
# Título principal
## Título segundo nivel
### Título tercer nivel

*Cursiva*
**Negrita**

- Elemento 1
- Elemento 2
- Elemento 3

[Texto del enlace](URL_del_enlace)
![Texto alternativo](URL_de_la_imagen)
```

Si estás interesado en aprender Markdown, hay varios recursos en línea que pueden ayudarte a dominar esta sintaxis de marcado ligero. Uno de los recursos más completos y accesibles es el sitio web [Markdown.es](https://www.markdown.es), que ofrece una guía detallada sobre la sintaxis y el uso de Markdown.



Checklist para establecer nombres de archivos

- ☐ Verifica las convenciones de nombres de archivo establecidas en tu disciplina o grupo. Estas convenciones deben ser documentadas para que otros en tu laboratorio o departamento puedan seguirlas.
- ☐ Los nombres de archivo deben ser descriptivos y proporcionar la información contextual justa.
- ☐ Intenta no hacer que los nombres de archivo sean demasiado largos. Los sistemas operativos tienen límites diferentes en el número de caracteres. Como regla general, intenta apuntar a un límite de 40-50 caracteres.
- ☐ Coloca la información más importante primero. La computadora organiza los archivos por nombre, carácter por carácter. Si anticipas que querrás encontrar un archivo por fecha, entonces coloca la fecha primero.
- ☐ Utiliza el estándar ISO 8601 para la fecha (YYYYMMDD). Para agregar una marca de tiempo a tu nombre de archivo, utiliza el formato YYYYMMDDThhmm. Usa el tiempo militar de 24 horas para evitar cualquier confusión sobre a.m./p.m.
- ☐ Cuando uses un sistema de numeración secuencial, utiliza ceros principales para asegurarte de que los archivos se ordenen de manera secuencial. Ejemplo: 001, 002, ...010, 011 ... 100, 101
- ☐ Utiliza la versión para indicar la versión más actual de un archivo. Ejemplo: filename_v2.xxx
- ☐ Evita caracteres especiales, como: ~ ! @ # \$ % ^ & * () ` ' ; : < > ? . , [] { } ' " |
- ☐ No uses espacios, ya que algunos programas no reconocerán nombres de archivo con espacios. Otras opciones incluyen guiones bajos, guiones, sin separación.

VERSIONADO

El control de versiones permite mantener un registro de los cambios realizados en los archivos a lo largo del tiempo. Imagina que estás trabajando en un proyecto y haces una modificación en un archivo, pero luego te das cuenta de que necesitas volver a una versión anterior. Con el control de versiones, puedes retroceder fácilmente a esa versión anterior y recuperar los datos perdidos.

Además, el control de versiones facilita la colaboración entre diferentes personas en un proyecto. Permite a múltiples usuarios trabajar en los mismos archivos simultáneamente sin temor a sobrescribir o perder cambios. Cada modificación se registra claramente, lo que facilita la identificación de quién hizo qué y cuándo.

En esta sección, exploraremos cómo aplicar el versionado de datos de investigación utilizando Git y GitHub, incluyendo una explicación detallada de su funcionamiento y cómo aprender a utilizarlos.

¿Qué es Git?

Git es un sistema de control de versiones distribuido que registra cambios en archivos a lo largo del tiempo en un repositorio de código. Git se utiliza para rastrear cambios en conjuntos de datos, scripts de análisis, documentos y cualquier otro archivo relacionado con el proyecto. Esto permite a los investigadores mantener un registro histórico de los cambios realizados en los datos, así como revertir a versiones anteriores si es necesario.

¿Cómo funciona Git?

Git registra cambios en los archivos utilizando un modelo de snapshots, donde cada cambio se guarda como una instantánea del estado actual de los archivos en un momento determinado. Esto facilita la colaboración entre investigadores, ya que pueden trabajar en diferentes aspectos

del proyecto simultáneamente y fusionar sus cambios de manera segura. Además, Git proporciona herramientas para gestionar ramas de desarrollo, etiquetar versiones importantes y realizar seguimiento de problemas y tareas relacionadas con los datos de investigación.

¿Qué es GitHub ?

GitHub es una plataforma de alojamiento de código basada en la web que utiliza Git para el control de versiones. Permite a los investigadores colaborar en proyectos, realizar un seguimiento de problemas, gestionar solicitudes de extracción y más. En el contexto de la investigación científica, GitHub proporciona un espacio centralizado para compartir datos, colaborar en análisis y facilitar la revisión por pares y la reproducción de resultados.

Recursos para aprender a utilizar Git y GitHub

Aprender a utilizar Git y GitHub en la gestión de datos de investigación puede ser un proceso gratificante pero desafiante. Algunos recursos que te recomiendo son:

- Utilizar Git:
 - [Comenzar con Git](#)
 - [Utilizar GitHub](#)
- Recursos adicionales de lectura en línea para Git:
 - [Sitio oficial del proyecto de Git](#)
 - [Guía de Git para profesionales](#)
 - [Lista de comandos de Git](#)
- Utilizar GitHub:
 - [Flujo de GitHub](#)
 - [Conexión a GitHub](#)
 - [Comunicarse en GitHub](#)
- Cursos en línea:
 - [GitHub Skills](#)
 - [La bifurcación de Git](#)
 - [Curso interactivo de Git en línea de Pluralsight](#)



Ruta de aprendizaje para Git y Github



TRABAJO CON ARCHIVOS

En este módulo, exploraremos los fundamentos esenciales de los formatos de archivo, la compresión de datos y las transformaciones. Al finalizar este módulo, comprenderás la importancia de estos conceptos y podrás tomar decisiones informadas sobre qué enfoque utilizar en cada situación. Además, podrás aplicar este conocimiento para mejorar tus prácticas de gestión de datos de investigación.

FORMATOS

Un formato de archivo es una estructura de codificación que organiza la información dentro de un archivo informático. Para acceder al contenido de un archivo, un programa o aplicación debe reconocer su formato específico. Por ejemplo, un navegador web reconoce el lenguaje de marcado de hipertexto (HTML) y puede abrir archivos con esta extensión, mostrándolos como páginas web. Si un archivo utiliza un formato diferente, es posible que se necesite un complemento especializado para abrirlo, o el navegador podría sugerir su descarga para abrirlo con el software adecuado.

Las extensiones de archivo, que generalmente consisten en tres o cuatro letras después de un punto en el nombre del archivo, indican su formato. Por ejemplo, "jpg" denota archivos JPEG, mientras que "docx" representa documentos de Word.

Los formatos de archivo propietarios suelen requerir software específico para abrirlos, lo que puede limitar su accesibilidad. En contraste, los formatos abiertos están disponibles para el público en general, lo que facilita su apertura y visualización mediante una variedad de aplicaciones.

Texto vs. Binario

Los archivos pueden clasificarse en dos categorías principales: texto y binario. Los archivos de texto son legibles por máquina a través de estándares de codificación de caracteres como ASCII y Unicode, lo que los hace accesibles a una amplia gama de aplicaciones y sistemas operativos. Por otro lado, los archivos binarios contienen datos ejecutables por software específico y pueden ser propietarios. La elección entre formato de texto y binario depende de los requisitos de la aplicación y de la legibilidad humana versus la eficiencia de almacenamiento y procesamiento.

Compresión de datos

La compresión de datos, también conocida como reducción de la tasa de bits, es un proceso que codifica la información en un formato más eficiente en términos de espacio de almacenamiento. Un formato de compresión común es ZIP, utilizado en diversas plataformas como Windows, Mac, Linux y Unix. ZIP ofrece compresión sin pérdidas, lo que significa que el archivo comprimido es idéntico al original después de descomprimirlo. Sin embargo, algunos formatos de compresión pueden provocar pérdida de calidad en archivos multimedia, lo que debe tenerse en cuenta al seleccionar el método de compresión adecuado.

Transformaciones de formato

En ocasiones, es necesario convertir o migrar archivos de un formato a otro, ya sea para adaptarse a nuevos sistemas, compartir datos con colaboradores que utilizan software diferente, o garantizar la accesibilidad a largo plazo. Las herramientas de conversión y migración facilitan este proceso,

aunque es importante tener en cuenta posibles pérdidas de formato o funcionalidad durante la transición. Además, se deben considerar las implicaciones de la compresión y la transformación en la integridad y la accesibilidad de los datos a lo largo del tiempo.

FORMATOS ESTABLES PARA ALMACENAR DATOS

Los formatos de archivo desempeñan un papel crucial en la comprensión y la integridad de los datos almacenados. Al seleccionar un formato adecuado, es fundamental considerar los siguientes aspectos:

- ¿Cumple el formato de archivo con uno o más estándares reconocidos?
- ¿Se trata de un formato abierto, accesible para todos, o es propietario y limitado?
- ¿Es necesario un software específico para acceder y manipular los datos? En caso afirmativo, se deben detallar el software, su versión y el sistema operativo compatible en los metadatos.
- ¿El archivo de datos se compone de múltiples archivos? En tal caso, se debe especificar esta estructura en los metadatos.

Tipos de archivo recomendados

- **No propietario:** Formatos abiertos y bien documentados.
- **Comúnmente utilizado:** Representaciones estándar como ASCII o Unicode.
- **No encriptado:** Para facilitar la accesibilidad y el intercambio de datos.
- **No comprimido:** Para preservar la integridad de los datos sin pérdida de calidad.
- **Formato ASCII:** Para garantizar la legibilidad futura, especialmente en datos tabulares.

FORMATOS RECOMENDADOS

1. **PDF/A:** Este formato se destaca por su capacidad para preservar la apariencia visual de un documento con el tiempo, independientemente de las herramientas y sistemas utilizados para crear, almacenar o renderizar los archivos. Al prohibir características inadecuadas para el archivado a largo plazo, como el uso de definiciones de fuentes externas y el cifrado, garantiza la integridad y accesibilidad de los documentos en el futuro.
2. **TIFF:** El formato TIFF es ampliamente reconocido en la industria de la impresión y la imagen debido a su capacidad para almacenar imágenes comprimidas sin pérdida. Esta capacidad de compresión sin pérdida asegura que no se produzca degradación de la calidad de la imagen con el tiempo, lo que es esencial para la preservación a largo plazo de documentos e imágenes.
3. **QuickTime:** Es un formato de contenedor ampliamente utilizado en la industria del video digital. Permite almacenar video, audio y texto en pistas separadas, lo que facilita la edición individual. El formato descomprimido 4:2:2 asegura la preservación de toda la información de la imagen sin pérdidas. QuickTime es compatible con una amplia gama de sistemas operativos y software, lo que garantiza su accesibilidad a largo plazo.
4. **BWAV:** El Broadcast Wave Format (BWAV) es un formato similar a WAV pero incluye metadatos adicionales para preservación. Admite LPCM y tiene un límite de tamaño de archivo de 2 GB. Puede enlazar múltiples archivos para reproducciones más largas. Las extensiones de archivo para BWAV son .bwf y .wav.



Lecturas



[Recomendación de formatos para la preservación digital de archivos](#) (2021)

[Sostenibilidad de los Formatos Digitales](#) (2017)

Seleccionar formatos de archivo estables y bien documentados es esencial para preservar la integridad y la accesibilidad de los datos a lo largo del tiempo. Proporcionar información detallada sobre los formatos elegidos facilita su comprensión y promueve su reutilización en futuras investigaciones.

TRANSFORMACIÓN EXITOSA DE DATOS

La transformación de datos es un proceso que puede realizarse durante o después de un proyecto de investigación. Existen diversas razones para transformar datos, desde la simplificación y estandarización hasta la visualización más efectiva de la información. A diferencia de la migración de archivos entre diferentes formatos, las transformaciones de datos implican modificar los propios datos, lo que puede tener un impacto significativo en su utilidad y comprensión.

Razones para transformar datos

1. Estandarización y simplificación: Por ejemplo, la codificación de respuestas de encuestas como números en lugar de cadenas de caracteres puede facilitar la introducción de datos y evitar errores tipográficos.
2. Transformación de datos cualitativos a cuantitativos: Las transcripciones de entrevistas pueden convertirse en datos cuantitativos mediante técnicas de codificación textual y categorización.
3. Visualización eficaz: Convertir datos de proporciones a porcentajes, por ejemplo, facilita su representación visual en gráficos de barras o circulares.

Herramientas para la transformación de datos

- Software de Análisis de Datos: Paquetes como [R](#), [Python](#) con bibliotecas como pandas, y software de análisis estadístico como SPSS y SAS ofrecen amplias capacidades para transformar datos de diversas maneras.
- Herramientas de visualización de datos: Aplicaciones como [Tableau](#), [Power BI](#) y matplotlib en Python permiten transformar datos para su visualización efectiva.

- Herramientas de Anonimización y Agregación: Para proteger la confidencialidad de los datos, herramientas especializadas como ARX y DataMasking pueden utilizarse para anonimizar y agregar datos sensibles.



Anonimización

La anonimización de datos es un proceso crucial en la gestión de datos que implica modificar o eliminar información identificable de un conjunto de datos, con el fin de proteger la privacidad y confidencialidad de las personas involucradas. Es especialmente importante en entornos donde se manejan datos sensibles, como registros médicos, información financiera o datos de encuestas.

La elección de la herramienta adecuada depende del propósito para el que se desee anonimizar los datos. Por ejemplo, para una anonimización única, como la de datos de encuestas, puede ser suficiente una herramienta de anonimización estática de código abierto.

A continuación, se presentan algunas herramientas de anonimización de datos de código abierto:

1. [ARX Data Anonymization Tool](#).
2. [Amnesia](#).
3. [μ-ARGUS](#).
4. [sdcMicro](#).
5. [Anonimatron](#).

Para casos que requieren una alta precisión y cumplimiento con regulaciones como el GDPR, herramientas de anonimización profesional como Aircloak Insights ofrecen anonimización interactiva y conforme a GDPR que permite un alto nivel de utilidad de datos para análisis precisos.

METADATOS Y DOCUMENTACIÓN

METADATOS

Los metadatos están presentes en todas partes, capturando información que permite descubrir, comprender y utilizar eficazmente los conjuntos de datos. Su valor se extiende a través de diferentes dominios y partes interesadas:

- Los metadatos permiten a los desarrolladores evitar la duplicación de datos, compartir información confiable y promover sus contribuciones a un campo de estudio. Al reutilizar los metadatos, los desarrolladores pueden ahorrar tiempo y recursos a largo plazo.
- Los metadatos proporcionan a los usuarios la capacidad de buscar, recuperar y evaluar información sobre conjuntos de datos. Ayudan a los usuarios a encontrar datos relevantes, determinar su aplicabilidad a sus necesidades y comprender su contenido y estructura.
- Los metadatos ayudan a las organizaciones a proteger su inversión en datos al documentar los pasos de procesamiento de datos, las medidas de control de calidad y las restricciones de uso. También permiten a las organizaciones rastrear el uso de datos, facilitar la publicación y promocionar sus esfuerzos de investigación.

ESTÁNDARES DE METADATOS

La elección del estándar de metadatos adecuado es crucial para garantizar la coherencia, la interoperabilidad y la usabilidad de los metadatos en diferentes conjuntos de datos y disciplinas. Existen varios estándares de metadatos, cada uno con su propio enfoque y aplicabilidad:

- [Conjunto de elementos Dublin Core](#): Se enfoca en recursos web y publicaciones.
- [Data Cite](#): Esquema de metadatos enfocado en la descripción de datos
- [Estándar de contenido FGDC para metadatos geoespaciales Digitales \(CSDGM\)](#): Se centra en datos geoespaciales y ofrece perfiles especializados para datos biológicos e información costera.
- [Metadatos de Información Geográfica ISO 19115/19139](#): Diseñado específicamente para datos y servicios geoespaciales.
- [Lenguaje de metadatos ecológicos \(EML\)](#): Adaptado para datos ecológicos.
- [Núcleo de Darwin \(DwC\)](#): Diseñado para especímenes de museo.
- [Lenguaje de Mercado Geográfico \(GML\)](#): Principalmente utilizado para características geográficas.

Al crear metadatos, es esencial seguir las mejores prácticas para garantizar la calidad, precisión y completitud del registro de metadatos. Esto incluye organizar la información de manera efectiva, escribir metadatos claros y descriptivos, y utilizar terminología estándar y vocabularios controlados. Además, los metadatos deben ser revisados por colegas para identificar y corregir cualquier error o inconsistencia antes de su publicación.

CITACIÓN DE DATOS

Los metadatos desempeñan un papel fundamental en el proceso de citación de datos. Estos proporcionan información contextual y descriptiva sobre los conjuntos de datos, lo que facilita su identificación, comprensión y localización. En el contexto de la citación de datos, los metadatos son esenciales para asegurar que la referencia sea precisa y completa.

Al proporcionar una cita para sus datos, usted hace que sea más fácil para otros usuarios identificarlos y reconocerlos. La citación de datos ayuda a promover la reproducibilidad de los resultados de la investigación, nos permite hacer un seguimiento del uso y el impacto de los datos y proporciona una estructura mediante la cual podemos reconocer y recompensar a los creadores de datos.

Como regla general, debe asegurarse de incluir suficiente información para que otro usuario pueda localizar el conjunto de datos o la parte de la base de datos a la que se refiere. DataCite, un organismo internacional de normalización fundado en 2009, recomienda los siguientes cinco elementos mínimos de citación: un creador, un año de publicación, un título, un editor y un identificador. Aquí, el identificador se refiere a cualquier ID persistente que identifique de forma única el objeto, incluido un DOI o incluso una URL, aunque esta última puede ser menos persistente. Dos propiedades adicionales que pueden añadirse son Versión y Tipo de recurso.

Las buenas prácticas en la documentación y citación de datos pueden contribuir a alcanzar un mejor nivel de reproducibilidad de la investigación.



Lecturas



[Declaración Conjunta de Principios de Citación de Datos](#) (2014)

[Análisis de los principios para la descripción de datos de investigación mediante Datacite](#) (2020)

DOCUMENTACIÓN DE DATOS

La documentación sistemática te ayuda a ti mismo y a otros a comprender y utilizar tus datos de manera efectiva. Mientras creas y analizas datos, es fundamental documentarlos para asegurarnos de que su significado y contexto estén claros. Esta documentación es esencial para otros investigadores que deseen revisar tus resultados, validar tus hallazgos o utilizar tus datos en nuevos estudios.

Existen diversos tipos de documentación de datos, como cuadernos de laboratorio, protocolos experimentales, cuestionarios, libros de códigos, diccionarios de datos, sintaxis de software y archivos de salida. La información sobre la configuración y calibración del equipo, el esquema de la base de datos, los informes metodológicos y la información de procedencia también son importantes para una documentación completa.

La documentación de datos se puede dividir en tres niveles: proyecto, base de datos y variable. En el nivel del proyecto, se documentan los objetivos de estudio, las preguntas de investigación, las hipótesis y las metodologías utilizadas. En el nivel de base de datos, se documentan todos los archivos relacionados entre sí. Finalmente, en el nivel de variable, se define el significado de cada variable incluida en el conjunto de datos.

Cuaderno de laboratorio

Los cuadernos de laboratorio son registros detallados de actividades científicas en el laboratorio. Desde experimentos hasta resultados, todo se documenta para facilitar la reproducibilidad y la verificación de resultados. Para conocer más:

1. [Guía de cuadernos de laboratorio electrónico](#)
2. [Acerca de los cuadernos de laboratorio electrónico](#)

Jupyter Notebooks

Jupyter Notebook es una herramienta interactiva basada en web que permite crear y compartir documentos con código en vivo, ecuaciones, visualizaciones y texto explicativo. Es ampliamente utilizado en ciencia de datos, matemáticas, investigación científica y educación debido a su capacidad para combinar código ejecutable, resultados visuales y narrativa.



Características de Jupyter Notebook

- Soporte para múltiples lenguajes de programación.
- Interactividad que permite ejecutar fragmentos de código directamente en las celdas.
- Visualizaciones integradas utilizando bibliotecas como Matplotlib y Plotly.
- Posibilidad de compartir y colaborar en documentos.
- Amplio ecosistema de extensiones y bibliotecas.

Conceptos clave de Jupyter Notebook

- Notebook: Documento interactivo que contiene celdas de código, texto enriquecido y elementos multimedia.
- Celda: Unidad básica dentro de un notebook que puede contener código, texto o elementos multimedia.
- Kernel: Entorno de ejecución para un lenguaje de programación específico asociado a cada notebook.
- Markdown: Sintaxis de formato de texto enriquecido para dar formato al texto en las celdas de texto.

Para instalar Jupyter Notebook visite la [guía oficial](#).

Para aprender más sobre Jupyter Notebook y Python, te recomiendo explorar el curso introductorio disponible en este [enlace](#).



Diccionario de datos



Un diccionario de datos es una herramienta utilizada en la gestión de datos que proporciona una descripción detallada de los elementos de datos en un conjunto de datos o base de datos. Consiste en un registro estructurado que documenta la semántica, la estructura y el significado de los datos almacenados.

Importancia:

1. Ayudan a los usuarios a comprender la estructura y el contenido de los datos de manera clara y consistente.
2. Facilitan la comunicación entre diferentes partes interesadas, como científicos de datos, analistas y desarrolladores.
3. Proporcionan una fuente centralizada de documentación que ayuda en la gestión y mantenimiento de los datos a lo largo del tiempo.
4. Permiten la reutilización de datos al proporcionar información detallada sobre su contexto y uso.

Principios:

1. Completo: Debe proporcionar una descripción completa de todos los elementos de datos relevantes.
2. Preciso: La información en el diccionario de datos debe ser precisa y actualizada.
3. Estandarizado: Se deben seguir estándares de nomenclatura y formato para garantizar la coherencia.
4. Accesible: Debe ser fácilmente accesible para todos los usuarios que trabajan con los datos.

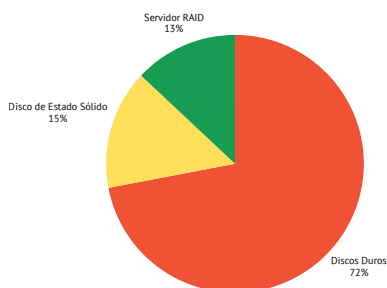
Plantilla de diccionario de datos en Excel: [Descargar aquí](#)

Ejemplo de diccionario de datos: [Descargar aquí](#)

ALMACENAMIENTO DE DATOS

ALMACENAMIENTO

Según una encuesta publicada en 2014 por Kroll Ontrack, proveedor de productos y servicios de recuperación de datos y e-discovery, las caídas de los discos duros se han más que triplicado en los últimos cuatro años, convirtiéndose en la causa más común de pérdida de datos. Sus datos son la savia de su investigación. Si pierde sus datos, la recuperación puede ser lenta, costosa o, peor aún, puede que ni siquiera sea posible. Por eso es tan importante almacenar y hacer copias de seguridad de sus datos de forma segura.



Planificación del Almacenamiento

Es útil pensar en la cantidad de almacenamiento que necesitará y planificar eso desde el principio. Piense también en quién necesitará acceder a sus datos y lo que eso significa en términos de dónde deben almacenarse. Si tiene en cuenta estas cuestiones desde el principio, podrá planificarlas adecuadamente y también incluir los costes de gestión de datos en sus solicitudes de subvención.

Opciones de Almacenamiento

- **Unidades de red:** Son gestionadas de forma centralizada por el personal de su escuela, departamento o universidad. Es muy recomendable que almacene sus datos en unidades de red.
- *Ordenadores personales y portátiles, dispositivos externos:* Pueden ser una opción cómoda por su bajo coste y portabilidad, pero nunca deben utilizarse para la copia maestra de sus datos.

Almacenamiento de datos confidenciales

El almacenamiento de datos considerados confidenciales o sensibles requiere una consideración adicional. Legalmente, los datos que contienen información personal deben tratarse con más cuidado que los datos que no lo hacen. Es importante abordar este tema durante los procedimientos de consentimiento para informar a las personas sobre cómo y por qué se almacenarán los datos. Además, los datos personales deben estar protegidos por medidas de seguridad adecuadas, como la encriptación de archivos y el acceso restringido.

Condiciones físicas para el almacenamiento

Las áreas y salas designadas para el almacenamiento de datos deben ser adecuadas para este propósito. Las condiciones bajo las cuales se almacenan los datos afectan significativamente su longevidad. Es importante garantizar un ambiente controlado y seguro para evitar daños físicos o degradación de los medios de almacenamiento.



TENER EN CUENTA

Los datos deben ser almacenados con la documentación generada previamente y teniendo en cuenta las cuestiones de formatos explicadas anteriormente.

SEGURIDAD EN EL ALMACENAMIENTO DE DATOS

La seguridad de los datos es una preocupación central en cualquier entorno donde se manejen información sensible o de investigación. La seguridad de los datos implica asegurar que la información de investigación se mantenga a salvo de la corrupción y que el acceso esté adecuadamente controlado. Esto incluye evitar daños o modificaciones accidentales o malintencionadas, así como proteger contra robos, violaciones de la confidencialidad y divulgaciones prematuras de los datos.

PRINCIPIOS CLAVE DE SEGURIDAD DE DATOS

1. **Control de acceso:** Es fundamental considerar quién necesitará acceder a los datos y establecer permisos o restricciones necesarios para garantizar que solo las personas autorizadas puedan acceder a la información.
2. **Copias de seguridad:** Una política clara y directrices sobre quién puede hacer copias de seguridad de los datos y dónde pueden almacenarse, especialmente en dispositivos móviles como computadoras portátiles, teléfonos inteligentes o memorias USB, son esenciales para proteger datos sensibles.
3. **Seguridad física:** Aunque un ordenador esté desconectado de la red, sigue siendo vulnerable al robo o la manipulación de datos. Por lo tanto, es fundamental guardar datos altamente sensibles en lugares seguros, como habitaciones cerradas con llave o cajas fuertes, y asegurarse de que estén encriptados.
4. **Autenticación:** Es importante seguir prácticas seguras en cuanto a nombres de usuario y contraseñas, evitando utilizar información personal obvia y asegurándose de no revelar contraseñas a personas no autorizadas.



Para más información, consulta el “[Toolkit de Preservación Digital del Gobierno de Canadá](#)”

RESPALDO DE DATOS

En el ámbito de la investigación, donde los datos son el fundamento de descubrimientos e ideas, la gestión de este valioso activo es primordial. Una sólida política de respaldo de datos se erige como centinela, protegiendo contra los peligros de pérdida o corrupción, asegurando la continuidad de los esfuerzos investigativos.

A continuación, se exponen las razones fundamentales que explican la importancia de una política de respaldo de datos bien documentada según [DataONE](#):

- **Claridad y transparencia:** Una política meticulosamente elaborada aporta claridad al ámbito nebuloso de la gestión de datos. Delinea los protocolos, procedimientos y responsabilidades, no dejando lugar a la ambigüedad.
- **Orientación estructurada:** Al dictar el dónde, quién, cuándo y cómo del respaldo y restauración de datos, la política provee un mapa estructurado, asegurando la adherencia metódica a las mejores prácticas.
- **Parámetros críticos:** Faculta a los interesados para estipular parámetros cruciales como ubicaciones de respaldo, privilegios de acceso, frecuencia de respaldos, tipos de respaldo y hardware y software recomendados, fomentando un entorno propicio para la administración de datos.
- **Integración y alineación:** Al identificar políticas y procedimientos existentes, la política de respaldo se integra sin problemas con los marcos organizativos, asegurando la alineación con los objetivos generales.
- **Ejecución sistemática:** Con un horario bien definido y personal designado, la política garantiza la ejecución sistemática de las operaciones de respaldo, mitigando el riesgo de omisión o negligencia.
- **Redundancia y fiabilidad:** Enfatizando la redundancia, la política exige la existencia de múltiples copias de respaldo en diversas ubicaciones, protegiendo contra desastres locales o fallas de hardware. Además, aboga por

el uso de diversos tipos de medios, reconociendo los problemas inherentes de fiabilidad asociados con cada uno.

- **Preservación de la integridad de datos:** Al asegurar que la estructura de los datos respaldados refleje su forma original y al delinear protocolos de archivo, la política resguarda la integridad y coherencia de los datos archivados.
- **Accesibilidad y continuidad:** La centralización de la información en un repositorio accesible facilita el acceso sin problemas para los interesados, asegurando la continuidad incluso frente a cambios o transiciones de personal.
- **Empoderamiento a través de la capacitación:** La provisión de capacitación para el personal responsable de las operaciones de respaldo y los usuarios finales fomenta el empoderamiento, fortaleciendo la resiliencia del ecosistema de datos.
- **Automatización y eficiencia:** Al abogar por la automatización parcial o total, la política subraya los beneficios de eficiencia, liberando recursos para empresas más estratégicas.

CIFRADO DE DATOS

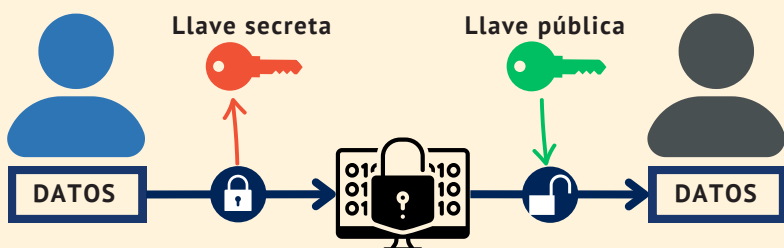
El cifrado es esencial durante el almacenamiento de datos porque garantiza que la información sensible permanezca protegida contra accesos no autorizados. Mediante este proceso, los datos se convierten en un formato ilegible e inaccesible para cualquier persona que no posea la clave de cifrado o la contraseña adecuada. Esta medida de seguridad se vuelve crucial en situaciones de pérdida o robo de dispositivos de almacenamiento, ya que asegura que los datos no puedan ser comprometidos incluso si el dispositivo cae en manos equivocadas.

El cifrado, que se basa en la creación de contraseñas sólidas, asegura que los datos permanezcan inaccesibles para cualquier entidad no autorizada. Sin embargo, es importante

destacar que la seguridad del cifrado depende en gran medida de la integridad de la clave de cifrado. Por lo tanto, es fundamental establecer un procedimiento de copia de seguridad confiable para respaldar tanto los datos cifrados como las contraseñas asociadas. La pérdida de la clave de cifrado podría resultar en la pérdida permanente de los datos, subrayando la importancia de mantener medidas de seguridad adecuadas en todo momento. Más información en este [enlace](#).

Cifrado de datos

Funcionamiento de un sistema de cifrado



Herramientas utilizadas para cifrar datos

1. **BitLocker:** Esta herramienta integrada en sistemas operativos Windows ofrece una sólida protección mediante el cifrado de unidades completas. [Activar en Windows](#)
2. **LibreCrypt:** es una aplicación gratuita y de código abierto que facilita el cifrado de datos mediante la creación de "contenedores". Estos contenedores son fáciles de montar y utilizar, proporcionando una manera segura de proteger datos confidenciales. [Repositorio en GitHub](#)
3. **Unidades flash USB con software de cifrado:** Se recomienda el uso de estas unidades para proteger los datos sensibles en caso de pérdida del dispositivo. Se menciona que, en comparación con las unidades flash USB normales, estas unidades cifradas tienen un costo ligeramente mayor, pero son una buena inversión para proteger los datos.

EL VALOR DE COMPARTIR

El compartir datos es esencial en cada etapa del ciclo de vida de los datos. Implica describir el contenido, las características y los procesos de los datos, depositarlos en ubicaciones accesibles, preservarlos en formatos adecuados para almacenamiento a largo plazo y publicar información sobre los datos para facilitar su descubrimiento.

El compartir datos ofrece numerosos beneficios para diferentes partes interesadas:

- **Para el científico:** Reconocimiento como fuente autoritativa, mejora en la calidad de los datos, aumento de oportunidades para el intercambio de datos y conexiones y colaboraciones potenciales en redes científicas.
- **Para el patrocinador:** Maximizar el retorno de las inversiones en investigación al mejorar el valor de los fondos gastados.
- **Para la comunidad de investigación:** Mejor capacidad para construir sobre trabajos existentes, realizar metaanálisis, aumentar la transparencia, reproducibilidad y comparabilidad de los resultados, ampliar la evaluación de metodologías, recomendaciones y mejoras, y proporcionar una mejor educación para nuevos investigadores sobre hallazgos actuales y significativos.
- **Para el público:** Mejor acceso a la información que conduce a una mejor comprensión y -

contribuciones hacia la toma de decisiones pública y personal efectiva.

ABORDANDO PROBLEMÁTICAS SOBRE COMPARTIR DATOS

Preocupaciones	Soluciones
Uso inapropiado	Proporcionar resúmenes detallados, propósito, limitaciones de uso e información complementaria para mitigar malentendidos.
Seguridad y confidencialidad	Proporcionar metadatos sin datos reales y utilizar restricciones de uso para controlar el acceso.
Falta de reconocimiento	Especificar la cita de datos requerida en las restricciones de uso.
Pérdida de fondos de investigación	Crear una segunda versión pública con descripciones de procesamiento de datos generalizadas

Métodos para crear datos compartibles

Para asegurar que los datos sean compartibles, es crucial:

- Crear metadatos robustos y descubribles.
- Incluir identificadores únicos e información de citación.
- Hacer que los colaboradores revisen los metadatos para garantizar su precisión.
- Publicar metadatos a través de un portal o centro de información.



Lecturas



[Intercambio de Datos por Científicos: Prácticas y Percepciones.](#) (Tenopir 2011)

[Intercambio, gestión, uso y reutilización de datos.](#) (Tenopir 2020)

Los buenos metadatos son la solución para abordar las preocupaciones sobre compartir datos de manera efectiva.

PROTECCIÓN DE DATOS AL COMPARTILOS

La gestión adecuada de datos de investigación implica proteger la privacidad y confidencialidad de la información sensible. Aquí se presentan algunos conceptos clave para entender este proceso:

1. **Información de Identificación Personal (PII):** Cualquier información que pueda identificar directa o indirectamente a un individuo.
2. **Información Sanitaria Protegida (PHI):** Datos de salud identificables individualmente, conservados por entidades cubiertas.
3. **Información Sensible:** Datos cuya divulgación podría causar daño físico, emocional, social, psicológico o reputacional.

La protección de la confidencialidad requiere una consideración cuidadosa de los identificadores directos e indirectos, así como el consentimiento informado de los participantes.

Leyes y Normativas: En Estados Unidos, leyes como la HIPAA, FERPA y la Regla Común rigen la gestión de datos confidenciales y la protección de la privacidad de los sujetos.

Propiedad y acceso a datos: La propiedad de los datos y los derechos de propiedad intelectual pueden afectar la compartición de datos. Es esencial resolver cualquier problema relacionado con la propiedad y los derechos de propiedad intelectual de manera temprana.

Restricciones de acceso a datos: Embargos, restricciones tecnológicas y acuerdos de uso de datos son formas comunes de limitar el acceso a los datos. Estas restricciones deben estar claramente descritas en las condiciones de uso, y se sugiere aplicar una licencia estándar, como las proporcionadas por Creative Commons.



Checklist para la protección de datos personales del Data Curation Center

1. La protección de datos está contemplada en cualquier orientación o material de formación sobre gestión de riesgos digitales del que usted sea responsable, y:

- ☐ Incluye enlaces a otras orientaciones sobre ética de la investigación.
- ☐ Se actualizará para reflejar los cambios, por ejemplo, el Reglamento Europeo de Protección de Datos.

2. Al planificar la investigación, se abordan los siguientes puntos, por ejemplo en un Plan de Gestión de Datos:

- ☐ Todos los datos se almacenan y etiquetan con el nivel adecuado de seguridad/confidencialidad, y esto se documenta antes de depositar los datos en un repositorio.
- ☐ Protección de los datos personales recogidos o creados. Cumplimiento de las directrices de protección de datos de la institución y de cualquier política informática o de seguridad de la información.
- ☐ La solicitud de consentimiento informado debe incluir el consentimiento para preservar y compartir los datos, a menos que un comité de ética lo considere inapropiado.
- ☐ Los datos sensibles (si los hay) se almacenan y transfieren de forma segura.

☐ 3. Existe un acuerdo de depósito claramente redactado y controles adecuados de los datos depositados en cualquier repositorio, de modo que sólo se comparten abiertamente si se han anonimizado por completo, y se protegen adecuadamente en caso contrario.

☐ 4. Cualquier repositorio de datos de cuyo mantenimiento sea responsable funciona de forma coherente con la política de seguridad de la información de la institución y las directrices pertinentes del servicio de TI.



Checklist para la protección de datos personales del Data Curation Center

- ☐ 5. Cualquier dato personal que ya se conserve se evalúa periódicamente para determinar si sigue siendo necesario para los fines para los que se conservó, o si podría anonimizarse (para que no entre en el ámbito de la DPA), o eliminarse de forma segura.
- ☐ 6. Los interesados (o participantes) pueden ejercer de forma significativa su derecho a oponerse al tratamiento de los datos, alegando que les causaría un daño o una angustia considerables.
- ☐ 7. La institución planifica los cambios técnicos y organizativos adicionales que puedan ser necesarios para cumplir la Directiva general europea de protección de datos.

PLATAFORMAS PARA COMPARTIR DATOS

Existen varias plataformas dedicadas al compartir datos, algunas de las cuales incluyen:

- **[Figshare](#)**: Una plataforma en línea donde los investigadores pueden hacer públicos sus datos de investigación en formato de archivo.
- **[Dryad](#)**: Un repositorio de datos de investigación que ofrece a los investigadores una forma de publicar, citar y hacer accesibles sus datos de investigación.
- **[Mendeley Data](#)**: Un repositorio de datos de investigación que permite a los investigadores compartir, descubrir y reutilizar datos de investigación.
- **[Open Science Framework](#)**: Una plataforma de colaboración de código abierto que permite a los investigadores compartir proyectos, datos y resultados de investigación de manera colaborativa y transparente.

Estas plataformas proporcionan una infraestructura sólida para compartir datos de manera efectiva y colaborativa. Además, que permiten su integración con otras herramientas.

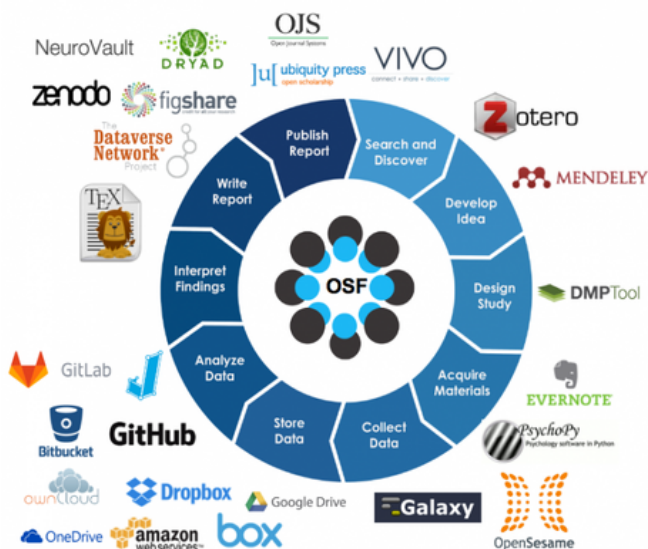


Imagen obtenida del [Blog de Open Science Framework](#)

COMPARTIR DATOS EN REPOSITORIOS CONFIABLES

Según un informe del RLG-OCLC, un repositorio digital de confianza es aquel cuya misión es proporcionar un acceso fiable a largo plazo a los recursos digitales gestionados a su comunidad designada, ahora y en el futuro. Los repositorios demuestran su fiabilidad mediante el cumplimiento de las normas y las mejores prácticas, políticas y procedimientos transparentes y bien documentados, y auditorías y certificaciones.

Las políticas y procedimientos comunican los servicios y capacidades de un repositorio, información esencial para ayudar a los investigadores a seleccionar un repositorio apropiado para sus datos. Además, la auditoría y la certificación ofrecen garantías a los investigadores de que sus datos recibirán el cuidado adecuado.

Tres ejemplos de auditorías y/o certificaciones son:

1. [DRAMBORA](#) (autoauditoría).
2. [Sello de Aprobación de Datos](#) (autoevaluación y revisión por pares).
3. [Auditoría y Certificación ISO 16363 de Repositorios Digitales de Confianza](#) (auditoría externa).

Otras certificaciones incluyen la Certificación Mundial de Sistemas de Datos del Consejo Internacional para la Ciencia y el Sello Nestor para Archivos Digitales de Confianza.

En la base de muchas normas archivísticas se encuentra el Modelo de Referencia ISO 14721 para un [Sistema Abierto de Información Archivística](#), u OAIS. El modelo OAIS presenta un marco de alto nivel para comprender los conceptos archivísticos, define elementos y procesos dentro de los repositorios digitales y establece un conjunto de responsabilidades para la conservación a largo plazo de la información digital.

Los repositorios dignos de confianza no sólo tienen flujos de trabajo y políticas acordes con las normas, como OAIS, sino que también demuestran estabilidad organizativa, física y económica. Un repositorio digno de confianza puede capear el temporal de las transiciones de personal, los contratiempos de infraestructura y los entornos con pocos recursos sin que se interrumpan las actividades de preservación.

Archivar los datos adecuadamente no es una empresa trivial. Los repositorios dignos de confianza dedican mucho trabajo, financiación, investigación y evaluación para salvaguardar los datos depositados en ellos.

No existe un repositorio de talla única. Los investigadores deben evaluar los servicios y puntos fuertes de un repositorio para determinar cuál es el mejor hogar para sus datos.

VISUALIZACIÓN DE DATOS

La visualización de datos se ha convertido en una herramienta indispensable en la gestión de datos de investigación. En este capítulo, exploraremos la importancia de la visualización de datos en la investigación, así como los principales tipos de gráficos y herramientas utilizadas en este proceso.

IMPORTANCIA DE LA VISUALIZACIÓN DE DATOS

La gestión efectiva de datos es esencial para generar conocimiento significativo y tomar decisiones informadas en la investigación. La visualización de datos desempeña un papel crucial al convertir datos crudos en representaciones visuales comprensibles y significativas. Algunas de las razones por las que la visualización de datos es importante en la investigación incluyen:

- **Comprensión de datos:** La visualización de datos permite a los investigadores comprender mejor la estructura y las características de los conjuntos de datos con los que trabajan. Al representar los datos de manera visual, los investigadores pueden identificar patrones, tendencias y relaciones que pueden no ser evidentes en los datos en bruto.
- **Comunicación de resultados:** Las representaciones visuales claras y concisas facilitan la comunicación de resultados y hallazgos a una variedad de audiencias. Desde colegas científicos hasta el público en general, la visualización de datos hace que la

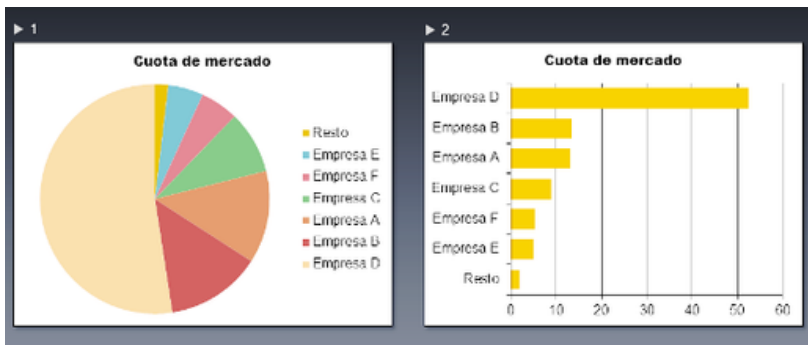
información sea más accesible y comprensible.

- **Detección de errores:** La visualización de datos también ayuda a detectar anomalías o errores en los datos. Al representar los datos de manera visual, los investigadores pueden identificar puntos atípicos o inconsistencias que podrían indicar problemas en la recopilación o el procesamiento de datos.

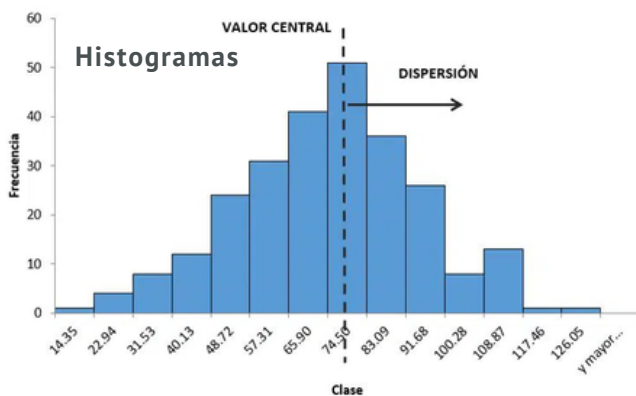
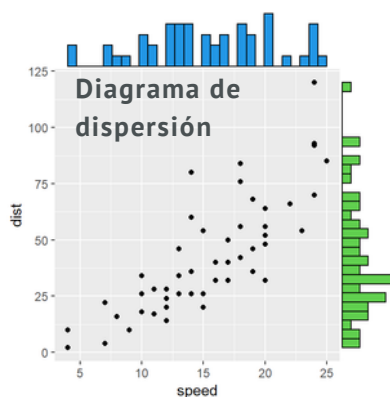
IMPORTANCIA DE LA VISUALIZACIÓN DE DATOS

Existen varios tipos de gráficos que se utilizan comúnmente en la visualización de datos, cada uno con sus propias características y aplicaciones específicas. A continuación, se presentan algunos de los principales tipos de gráficos utilizados en la gestión de datos de investigación:

- **Gráficos de barras:** Los gráficos de barras son una forma efectiva de comparar cantidades entre diferentes categorías o grupos. Consisten en barras rectangulares que representan los valores de cada categoría, y la altura de cada barra es proporcional al valor que representa. Los gráficos de barras son ideales para visualizar datos categóricos y para identificar diferencias y tendencias entre grupos.
- **Gráficos circulares:** También conocidos como gráficos de pastel, los gráficos circulares muestran la proporción de cada categoría en relación con el total. Son útiles para representar la distribución de datos en términos de porcentajes y para visualizar la composición relativa de diferentes categorías.



- **Gráficos de líneas:** Los gráficos de líneas muestran la relación entre dos variables a lo largo del tiempo o en una serie de datos ordenados. Consisten en líneas que conectan puntos de datos, y son ideales para visualizar tendencias y cambios a lo largo de un período específico.
- **Diagramas de dispersión:** Los diagramas de dispersión muestran la relación entre dos variables mediante la disposición de puntos en un plano cartesiano. Son útiles para identificar correlaciones y patrones de dispersión entre variables y para visualizar la distribución de datos en un espacio bidimensional.
- **Histogramas:** Los histogramas representan la distribución de frecuencias de una variable continua mediante barras adyacentes. Son útiles para visualizar la distribución de datos y para identificar posibles sesgos o patrones en los datos.



HERRAMIENTAS DE VISUALIZACIÓN DE DATOS

Para crear visualizaciones de datos efectivas, los investigadores pueden utilizar una variedad de herramientas y software especializado. A continuación, se presentan algunas de las herramientas más populares utilizadas en la visualización de datos:

1. **Python:** es un lenguaje de programación ampliamente utilizado en la ciencia de datos y la visualización de datos. Con bibliotecas como Matplotlib, Seaborn y Plotly, los investigadores pueden crear una amplia gama de visualizaciones de datos interactivas y personalizadas.
2. **R:** es un lenguaje de programación diseñado específicamente para el análisis estadístico y la visualización de datos. Con paquetes como ggplot2 y ggvis, los investigadores pueden crear gráficos de alta calidad y publicables para visualizar sus datos.
3. **LookerStudio:** es una plataforma de análisis y visualización de datos que permite a los usuarios explorar y compartir datos de manera intuitiva. Con una interfaz fácil de usar y capacidades de visualización interactiva, LookerStudio es una herramienta popular para la visualización de datos en entornos empresariales y académicos.
4. **Tableau:** es una herramienta de visualización de datos líder en la industria que permite a los usuarios crear visualizaciones interactivas y dinámicas con facilidad. Con una interfaz intuitiva y una amplia gama de opciones de visualización, Tableau es ampliamente utilizada en diversos campos para explorar datos, analizar tendencias y comunicar hallazgos de manera efectiva.



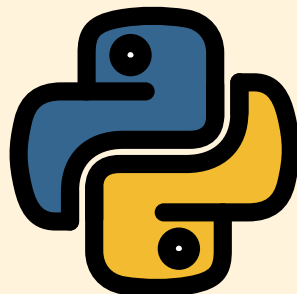
Lecturas

[*Fundamentos de Visualización de Datos*](#) (2021)

[*Interactive web-based data visualization with R, plotly, and shiny*](#) (2019)

Pyhton

Python es un lenguaje de programación de alto nivel, interpretado, de propósito general y de código abierto. Es conocido por su sintaxis clara y legible, lo que lo hace ideal para principiantes y también potente y flexible, lo que lo convierte en una opción popular para una amplia gama de aplicaciones, incluida la visualización de datos.



Bibliotecas para la el análisis y visualización de datos

Python cuenta con una serie de bibliotecas poderosas y fáciles de usar específicamente diseñadas para la visualización de datos. Estas bibliotecas ofrecen una amplia gama de opciones de visualización y permiten a los usuarios crear gráficos personalizados y visualizaciones interactivas con facilidad, algunas de estas bibliotecas son:

- [NumPy](#)
- [Pandas](#)
- [SciPy](#)
- [Matplotlib](#)
- [Seaborn](#)

Recursos para aprender Python

1. [Tutorial oficial de Python](#)
2. [Playlist en Youtube de Python desde 0](#)
3. [Quiero aprender Python](#)
4. [Python Data Science Handbook](#)
5. [Analyzing Data with Python](#)
6. [Zero to Mastery TensorFlow for Deep Learning book](#)
7. [NLP Course](#)

R - Lenguaje

R es un lenguaje de programación y un entorno de software ampliamente utilizado en el análisis de datos y la estadística. Es especialmente popular entre los científicos de datos, analistas estadísticos y profesionales de la investigación debido a su capacidad para manipular, visualizar y modelar datos de manera eficiente.



Bibliotecas para la el análisis y visualización de datos

R cuenta con numerosas bibliotecas especializadas el análisis y visualización, estas proporcionan herramientas y funciones poderosas que facilitan la exploración y comprensión de conjuntos de datos complejos, así como la comunicación efectiva de resultados a través de visualizaciones atractivas y informativas, algunas de estas bibliotecas son:

- dplyr
- ggplot2
- tidyr
- caret
- randomForest
- lubridate
- stringr
- tidymodels
- shiny
- purrr

Recursos para aprender R

1. [R for Data Science](#)
2. [Telling Stories with Data](#)
3. [ggplot2: Elegant graphics for Data Analysis](#)
4. [Mastering Shiny](#)
5. [Text Mining with R](#)
6. [Handbook of Graphs and Networks in People Analytics](#)
7. [R Programming Crash Course](#)
8. [R Programming Tutorial](#)

LookerStudio

LookerStudio es una plataforma de análisis de datos de Google que permite a las organizaciones visualizar, explorar y compartir información crítica de manera efectiva. Se destaca por su enfoque en la creación de paneles de control interactivos y personalizables, así como por su capacidad para integrarse con una variedad de fuentes de datos.



Looker Studio

Características de Looker Studio

- Facilita la creación de paneles de control personalizables para explorar datos de forma dinámica.
- Compatible con diversas fuentes, incluidas bases de datos relacionales y almacenes de datos en la nube.
- Ofrece capacidades avanzadas para definir métricas, crear dimensiones personalizadas y establecer relaciones entre conjuntos de datos.
- Permite realizar análisis en tiempo real para responder preguntas sobre los datos sobre la marcha.
- Facilita la colaboración entre equipos al compartir paneles de control, informes y visualizaciones.
- Proporciona características robustas de seguridad y cumplimiento para proteger datos sensibles y cumplir regulaciones como GDPR y HIPAA.

Recursos para aprender Looker Studio

- [Guía práctica de Google](#)
- [Looker Studio Essentials Beginner Course](#)
- [Google Looker studio | curso completo](#)
- [Taller Gratuito de Looker Studio](#)

DESPEDIDA

¡Enhorabuena por completar esta guía para convertirte en bibliotecario de datos! Esperamos que hayas encontrado útiles los conceptos, recursos y herramientas compartidos aquí.

Queremos expresar nuestro más sincero agradecimiento por tu dedicación y compromiso para aprender sobre este tema tan importante en constante evolución. La gestión de datos es una habilidad invaluable en el mundo actual, donde la información es uno de nuestros activos más preciados.

Como muestra de nuestro agradecimiento, te ofrecemos una plantilla en Notion que te ayudará a organizar y seguir tu progreso en esta área. Puedes acceder a la plantilla [aquí](#).

Si esta guía te ha sido útil, te invitamos cordialmente a dejar una reseña en nuestra página. Tus comentarios nos ayudarán a mejorar y a seguir ofreciendo contenido relevante y de calidad a nuestra comunidad.

Te deseamos mucho éxito en tu trayectoria como bibliotecario de datos y en todos tus futuros proyectos.