Manejo de Datos Abiertos en la Agricultura y Nutrición

Este curso de aprendizaje digital (e-learning) es el resultado de una colaboración entre socios de GODAN Action, incluyendo a ***Investigaciones Ambientales Wageininen (WUR)***, ***AgroKnow***, ***AidData***, la ***Organización de las Naciones Unidas para la Alimentación y la Agricultura*** (FAO por sus siglas en Ingles), ***El Foro Global sobre Investigaciones de Agricultura*** (GFAR), y el ***Instituto de los Estudios del Desarrollo*** (IDS), ***The Land Portal***, ***el Instituto de Datos Abiertos*** (IDI) y el ***Centro Técnico de Agricultura y cooperación Rural*** (CTA).

GODAN Action es un proyecto de tres años [por] el Departamento del Desarrollo Internacional del Reino Unido para capacitar a los que usan, producen, e intermediarios de datos para conectarse efectivamente con datos abiertos y maximizar la potencial por su impacto en los sectores de agricultura y nutrición. En particular, trabajamos para mejorar la capacitación, promover estándares comunes y mejores prácticas para medir el impacto. [www.godan.info]

Unidad 3: Creando datos abiertos

Lección 3.1: LA GESTION DE CONJUNTOS DE DATOS

Foto por Niel Palmer (CIAT) licenciado bajo CC BY 2.0

Objetivos y metas de aprendizaje

Se puede utilizar a datos abiertos cuando

Esta lección tiene el objetivo de:

- Explicar los principios básicos detrás de la gestión de datos
- Introduce el proceso de preparar un plan de gestión de datos
- Explicar conceptos de almacenamiento de datos, de usar versiones y prácticas de documentación

Después de estudiar esta lección debes saber cómo:

- Identificar los pasos para asegurar que se sigue un proceso de buenas prácticas en la gestión de datos científicos
- Preparar un plan de gestión de datos
- Entender la potencial de usar buenas prácticas de almacenamiento de datos y de usar versiones.

Contenido

Lista de gráficas

Lista de tablas

1. La gestión de datos

Muchos conjuntos de datos son efémeros por su naturaleza: datos del mercado, datos de producción y el consumo, y datos sobre el clima son buenos ejemplos. Estos datos no significan nada a menos que sabemos del periodo de tiempo que representan y si están actualizados con frecuencia. Datos de tierra, por ejemplo, son similares hasta si no están actualizados con tanta frecuencia que los del clima o el mercado. Estos cambian por ubicación y durante el curso de – y dentro de – las temporadas.

Para crear y mantener la confianza en datos abiertos como estos y en otros, es necesario tener principios y prácticas del manejo de datos establecidos. Buenos **principios de la gestión de datos** ayudan asegurar que los datos producidos o usados están registrados, almacenados, disponibles para la utilización y reutilización, mantenidos en el tiempo y/o desechados según requisitos términos legales, éticas, y de acuerdo con patrocinadores y buenas prácticas. Para consumidores de datos abiertos, la confianza depende de varios factores:

- *Conociendo a la fuente.* La confianza en los datos empieza con el conocimiento de su fuente.
- *Confiando en la fuente*. Si conoces que los datos vienen de una fuente confiable, puedes confiar en ellos, y las conclusiones que vienen de ellos.
- *Cuan actualizados están los datos*. Aun cuando los datos provienen de una fuente confiable, no son útiles si no están actualizados.
- *La calidad de los datos*. Datos confiables deben reflejar lo que miden con precisión
- *Sostenibilidad*. Un conjunto de datos confiable tiene que tener una forma de garantía de estar disponible
- *Facilidad de encontrar*. Como documentos, datos solamente son útiles si es fácil encontrarlos (más en este factor en la Unidad 1, Lección 2.1)
- *La documentación y atención al usuario.* Los consumidores deben de tener acceso a apoyo técnico si necesitan.
- *Interacción*. Los consumidores deben poder dar retroalimento si hay problemas con los datos.

Por lo tanto, la **gestión de datos** es un proceso que involucra una gama amplia de actividades desde lo administrativo a aspectos técnicos del manejo de datos[1] en una manera que abarca a los factores previamente mencionados. Una política sólida del manejo de datos definirá metas estratégicas de largo-plazo para el manejo de datos a través de todos los aspectos de un proyecto o empresa.

Una política de la gestión de datos es un conjunto de principios de alto-nivel que establecen un marco de referencia para el manejo de datos. Una política de manejo de datos puede ser utilizado para asuntos estratégicos como acceso a datos, asuntos legales relevantes, asuntos de la mayordomía de datos y responsabilidades custodiales, adquisición de datos, y otros asuntos. Como provee un marco de referencia de alto nivel, la política del manejo de datos debe ser flexible y dinámico. Esto hace que sea capaz de adaptar a retos no esperados, diferentes tipos de proyectos, y potencialmente asociaciones oportunistas mientras que mantiene su enfoque estratégico guía. La política del manejo de datos ayudará informar y desarrollar un **plan de la gestión de datos**, que se describirá con más detalle más adelante en esta lección.

---

[1] A. Gordon (ed.) 2015. *Official (ISC)^2 Guide to the CISSP CBK* (4th edn) CRC Press, Boca Raton, FL, USA

Para cumplir con las metas y estándares del manejo de datos, todos las partes interesadas deben entender los roles y responsabilidades asociados. Los objetivos de delinear estos roles y responsabilidades son para:

- Claramente definir roles asociados con funciones
- Establecer la pertenencia de datos a través de todas las fases de un proyecto
- Instar a la responsabilidad en datos
- Asegurar que medidas de calidad de datos y metadatos adecuadas y acordadas están mantenidos continuamente

La calidad en relación con los datos ha sido definida como la suficiencia para el uso potencial. Muchos de los principios para la calidad de datos aplican cuando se trata de datos de especies y con los aspectos espaciales de ellos. Estos principios están involucrados en todas las etapas del proceso del manejo de datos, empezando con la colección de datos y captura. Una baja de calidad de datos en cualquier de estas etapas reduce la aplicabilidad y las áreas del uso adecuado de ellos.

Todas estas cosas afectan a la calidad final o la suficiencia de los datos para el uso, y aplican para todos los aspectos de ellos. Estándares para la calidad de datos pueden estar disponibles para:

- Exactitud
- Precisión
- Resolución
- Confiabilidad
- Reproducibilidad
- Actualidad
- Relevancia
- Habilidad de auditar
- Lo completo
- La oportunidad

La calidad de datos se evalúa por aplicar procedimientos de verificación y validación como parte del proceso de control de calidad. La verificación y la validación son componentes importantes en cuestión del manejo de datos que ayuda asegurar que los datos son válidos y confiables. Esto se describe en la unidad 2 lección 2.2

Algunos de los principios del manejo de datos descritos en esta unidad también están en común con los principios del manejo de datos para la investigación (MDI). Buenos principios del MDI se pueden aplicar a iniciativas de datos abiertos con éxito, asegurando a:

- Mejor visibilidad y descubrimientos de los conjuntos de datos
- La facilitación de la durabilidad, el intercambio y utilización de conjuntos de datos
- El entendimiento del contexto y limitaciones del contenido de los conjuntos de datos
- La reducción de la pérdida de datos por mantenerlos seguros
- La interoperabilidad de conjuntos de datos y el intercambio de datos
- La conformidad con las expectaciones y políticas de patrocinadores y/o instituciones
- La provisión de oportunidades para colaboración con otros que podrían interactuar con los datos.

1.1. La diversidad de datos

Los datos pueden ser vistos al nivel más básico de donde se derivan la información y conocimiento. Los datos también pueden ser vistos como el nivel de donde las medidas se tomen originalmente, por ejemplo, respuestas individuales a una encuesta o censo; medidas de temperatura horarias, la velocidad y dirección del viento; el número y precio de acciones intercambiado en cada transacción de venta o compra; etc.

No obstante, la palabra datos significa algo diferente para gente diferente en contextos diferentes. Disciplinas diferentes tienen y utilizan un vocabulario diferente para los datos de la investigación sujeto[2].

Tabla 1: ***Varios tipos de datos como se encuentra en diferentes contextos y disciplinas***

Tipos de Datos		
General	Ciencias Sociales	Datos Físicos Agrícolas
• Imágenes • Video • Mapas/datos GIS • Medidas numéricas	• Respuestas de encuestas • Transcritos de entrevistas individuales o de grupos de enfoque • Indicadores económicos • Datos demográficos • Encuestas de opinión sobre votación	• Medidas generadas por sensores/ instrumentos de laboratorio • Modelaje de computadores • Simulacros • Observaciones y/o estudios de campo • especímenes

Por lo tanto, los datos pueden ser generados para propósitos diferentes y procesos diferentes en una multitud de formatos digitales. Las siguientes clasificaciones fueron compiladas por la *Research Information Network* (Red de Información de Investigaciones):

- **Observacionales**: datos capturados en tiempo real, que suelen ser únicas e irremplazables, por ejemplo, imágenes del cerebro, datos de encuestas
- **Experimentales**: datos de resultados experimentales, por ejemplo, de equipos de laboratorio, y con frecuencia reproducible, pero a veces caro, por ejemplo, cromatogramas y microensayos
- **De simulacro**: datos generados por modelos de pruebas donde datos del modelo y metadatos podrían ser más importantes que datos resultados, por ejemplo, modelos económicos o climatólogos.
- **Derivados o compilados**: resultando de procesamiento o por combinar datos en bruto, frecuentemente reproducible pero caro, por ejemplo, bases de datos compilados, minería de datos, o datos censales agregados.

[2] http://mantra.edina.ac.uk

- **De referencia o canónicos**: una conglomeración (estático u orgánico) o colección de conjuntos de datos más pequeños (revisados por pares), a lo mejor publicado y curado; por ejemplo, banco de datos para genes, bases de datos cristalográficos.

En la próxima sección, describiremos las consideraciones básicas en cuestión de la gestión de datos que se hace durante un ciclo de vida de datos, por ejemplo, desde la creación y almacenamiento inicial hasta el tiempo que se vuelven obsoletos y eliminados.

1.2. Metadatos y la documentación

Todos los conjuntos de datos deben estar identificados y documentados para facilitar su segunda identificación, gestión adecuada y uso efectivo, y para evitar coleccionar los mismos datos más de una vez. Para proveer una lista precisa de conjuntos de datos mantenidos por una organización, debe de compilar un catálogo de datos. Esta es una colección de metadatos al nivel del descubrimiento para cada conjunto de datos, en un formato adecuado para que los usuarios utilicen como referencia. Estos metadatos deben proveer información sobre el contenido, extensión geográfica, accesibilidad y actualidad de los datos, junto con detalles de contacto para más información sobre los datos.

Todo conjunto de datos, una vez catalogado, debe también estar documentado de un formato adecuado para usuarios como referencia mientras que usan los datos. Estos metadatos detallados deben describir el contenido, las características y el uso del conjunto de datos, utilizando un marco estándar de metadatos detallados.

Metadatos

Los metadatos, o "datos sobre datos" explican a su conjunto de datos y permiten que se documente información importante para:

- Encontrar a los datos más tarde
- Entender de qué son los datos más tarde
- Compartiendo los datos (ambos con contribuidores y usuarios secundarios de los datos en el futuro)

Se debe considerar como una inversión de tiempo que te ahorrará de la molestia de hacerlo más tarde.

Ejemplos:

- *Dublin Core*
- *Darwin Core*
- *FGDC (Federal Geographic Data Committee,* Comité de Datos Geográficos Federales*)*
- *DDI (Data Documentation Initiative,* Iniciativa de Documentación de Datos*)*
- *ABCD (Access to Biological Collections Data,* Accesso a Colecciones de Datos Biológicos)
- *CSDGM (Content Standard for Digital Geospatial Metadata,* Estándar de Contenido para Metadatos Geográficos Digitales*).*

Una distinción citada frecuentemente con la gestión de datos es entre los datos y metadatos (o datos sobre datos). Hay una gran cantidad de distinciones específicas que se refiere eso:

- *Metadatos como esquema*. Cuando coleccionamos datos tabulares, necesitamos a que se refieren las "columna" en los datos. Aun en datos no-tabulares, alguna información sobre el esquema es útil para interpretar a los datos. Muchos formatos de datos incluyen maneras para especificar metadatos de esquema, por ejemplo, XSD para XML, los RDF par RDF, y DDL para bases de datos.
- *Metadatos bibliográficos*. Bibliotecarios y científicos de biblioteca han usado metadatos para describir documentos (libros, artículos, fotos, etc.) por siglos, y han determinado a estructuras para registrar y buscar estos tipos de metadatos. Este tipo de metadatos incluye a la información de procedencia (autoría, datos de publicación), fechas, tamaño de la publicación (por ejemplo, un conteo de páginas), y se puede aplicar a conjuntos de datos también (por ejemplo, DCAT27 y VoID28).
- *Vocabulario Común*. Alinear a conjuntos de datos diferentes es un desafío en cualquier ambiente distribuido. Una herramienta clave en gobernar tales conjuntos de datos es el uso de un vocabulario común. El vocabulario se usa en el contenido de los datos, en vez de describir a los datos en sí. Por ejemplo, el AGROVOC (para Vocabulario en la Agricultura)[3] provee (entre otras cosas) terminología para hablar de varios productos, por ejemplo, leche, sub productos de leche, grasa de leche, etc. *Agroportal*[4] y el registro *VEST*[5] proveen acceso a muchos vocabularios relacionados con la agricultura.

Contenido de Archivo

Para que otros utilicen los datos, tienen que entender el contenido del conjunto de datos, incluso los nombres de parámetros, unidades de medir, formatos, y definiciones en valores en código. Al parte superior del archivo, incluye varios renglones de encabezado de información con descriptores que se relacionen a los datos con el conjunto de datos (por ejemplo, el nombre del archivo, el título del conjunto de datos, autor, la fecha de hoy, la última fecha en que el archivo fue modificado, y los nombres de archivos relacionados). Otros renglones de encabezado deben describir el contenido de cada columna, con un renglón para nombres de parámetros y otro para las unidades del parámetro. Para los conjuntos de datos que son granes y complejos y que pueden requerir mucha información descriptiva sobre el contenido del conjunto de datos, esa información se puede incluir en un documento separado vinculado en vez de estar en renglones de encabezado dentro del archivo mismo.

Parámetros: Los parámetros reportados en los conjuntos de datos necesitan tener nombres que describen su contenido, y sus unidades necesitan estar definidos para que otros entiendan a que se reporta. Usa nombres aceptados comúnmente. Un nombre bueno es corto, único (por lo menos dentro del conjunto de datos), y descriptivo del contenido del parámetro. Los nombres de la columna se deben construir para ser importado fácilmente por una variedad de sistemas de datos. Usa mayúsculas de manera consistente y solamente usa letras, números, y guiones bajo – no un espacio o punto – en el nombre del parámetro. Escoja un formato consistente para cada parámetro y usa ese formato en todo el conjunto de datos. Donde sea posible, trata de usar un formato estándar, como los que usan para fechas, tiempos, y coordenados espaciales.

---

[3] FAO. 2016. *AGROVOC Multilingual agricultural thesaurus (Tesauro Multilingual de la Agricultura)*. Disponible en: http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus

[4] IBC Agroportal. Available at: http://agroportal.lirmm.fr/

[5] VEST Directory, Agricultural Information Management Standards (AIMS). http://aims.fao.org/vest-registry

Todas las células dentro de la columna deben contener solamente un tipo de información (por ejemplo, texto, números, etc.). Tipos de datos comunes incluyen texto, numérico, fecha/tiempo, booleano (Si/No Cierto/Falso). Y comentarios, que se usan para almacenar cantidades grandes de texto.

- *Campos de Datos Codificados*. Campos de datos codificados, en comparación a campos de texto libre, suelen tener listas estándares de valores predefinidos de los cuales el proveedor de datos puede escoger. Recopiladores de datos pueden establecer sus propios campos codificados con valores definidos para ser utilizados consistentemente a través de varios archivos de datos. Los campos de datos codificados son más eficientes para el almacenamiento y recuperación de datos que campos de texto libre.
- *Valores Faltantes*. Hay varias opciones para lidiar con un valor faltante. Uno es dejarlo en blanco, pero esto puede crear un problema porque algunos programas no diferencian entre un valor en blanco o un cero, o un usuario puede dudar si el proveedor por accidente faltó un dato. Otra opción es poner un periodo donde un numero iría. Esto le hace claro que un valor debe estar allí, aunque no dice nada porque falta ese dato. Otra opción más será usar un código diferente para indicar diferentes razones del por qué hay datos faltantes.

1.3. Seguridad y almacenamiento

El compartir de datos efectivo depende den una red fuerte de confianza entre proveedores y consumidores. Una infraestructura para el compartir de datos no será utilizada si los partes que proveen y utilizan los datos no confían en la infraestructura o el uno en el otro. Si datos sensitivos serán compartidos, hay que tener provisiones en la plataforma para asegurar la seguridad de estos datos. Si los datos están cerrados o compartidos con individuos específicos u organizaciones, tendrán que ser alojados en una manera controlada. Dependiendo en cuan sensitivos sean los datos, esto incluye alguna forma de seguridad, por ejemplo, contra el hackeo. En los casos más extremos, los requisitos de seguridad para datos compartidos en la agricultura pueden ser tan severos que por datos compartidos en las fuerzas militares. Estos principios no son únicos para los datos en agricultura, y han sido estudiados profundamente.

Los conceptos básicos detrás de estos principios son los servicios que deben ser difíciles de transigir, que una transigencia debe ser fácil de detectar, y que el impacto de la transigencia podría ser contenido. Para datos abiertos, esto no representa una preocupación grande, pero para generar la confianza entre proveedores de datos, alguna medida de apoyo para la seguridad debe existir.

Algunas consideraciones importantes para el almacenamiento físico y el archivamiento de datos electrónicos/digitales incluyen:

- *El hardware y software de servidores*. ¿Qué clase de base de datos sería necesario para los datos? ¿Tendrá que crear una infraestructura de sistema física, o una ya existe? ¿Será necesario un producto principal de bases de datos? (vea la lección 3.3 en Crear Repositorios de Datos Abiertos para plataformas de software.) ¿Quién estaría encargado de la gestión de este sistema?
- *La infraestructura de la red*. Para datos abiertos, un base de datos necesita estar conectado al internet para ser accesible. ¿Cuán ancha de banda necesita para servir a los usuarios? ¿Cuáles horas del día necesita ser accesible?

- *Tamaño y formato de conjuntos de datos.* Se debe estimar el tamaño de un conjunto de datos para que se tome en cuenta el espacio de almacenamiento. Se debe identificar a los tipos y formatos para que no surjan sorpresas en cuanto a capacidades de la base de datos y compatibilidad.
- *Mantenimiento y actualización de los bases de datos.* Se debe tener protocolos cuidadosamente definidos para la actualización de una base de datos o conjunto de datos. Si un base de datos está corriendo, esto incluirá cosas como adiciones, modificaciones, eliminaciones, y también la frecuencia de actualizaciones. Hacer varias versiones será imprescindible mientras que trabaja en un ambiente multiusuario.
- *Un respaldo (backup) de datos y requisitos de recuperación.* Los requisitos para el respaldo o recuperación de una base de datos en caso de un error, fallo de software/medios, o desastre, deben estar claramente definidos y acordados para asegurar la longevidad de un conjunto de datos. Los mecanismos, horarios, frecuencia, y tipos de respaldos, y planes adecuados de recuperaciones deben estar preparados y especificados. Esto puede incluir tipos de medios para almacenamiento para respaldos de datos en sitio y si sea necesario un respaldo fuera de sitio.

1.4. Acceso a datos y diseminación

La producción de datos es una cosa, pero la diseminación de ellos es otra. Los datos abiertos son útiles cuando se pueden entregar a las manos (o máquina) correctas y dentro de un contexto donde podrá ser más valioso. En algunos casos, esto puede ser un laboratorio investigando la eficacia de un tratamiento (por ejemplo, la efectividad de un herbicida para lidiar con una plaga). A veces los datos se tienen que entregar al campo para que se puedan ayudar a un agricultor pequeño para tomar decisiones informados en cuales variedades de cultivos para plantar o cuales tratamientos para aplicar. Tienen que estar una variedad de medios de entrega de datos, afinados para cada caso de entrega. Este “afino” de los medios de entrega de datos puede convertirse en una oportunidad de negocio para intermediarios de datos en el caso donde los datos están completamente “abiertos.” Un intermediario puede proveer servicios para personalizar la entrega de datos para el rango vasto de clientes que podrían existir para los datos. Los datos abiertos crean la posibilidad de un mercado donde fuentes alternativas de datos relevantes están disponibles.

Para ser hechos más disponibles, los datos tienen que estar almacenados de una manera que son accesibles. Aun en la era moderna del lanzamiento por la nube, los datos y aplicaciones están almacenados en algún *hardware* en algún lado, aun si esta virtualizado. Una estrategia para compartir datos de escala global tiene que especificar donde será almacenado y cuales acuerdos al nivel de servicio (ANS, o *service level agreements*) serán mantenidos (tiempo corriendo, procesamiento, y controles de acceso, etc).

Debe considerar fuertemente lo siguiente cuando decidiendo como diseminar a los datos:

- Acceso a los datos debe estar provisto según la política de datos de la organización y las leyes nacionales para acceso a información.
- Acceso a los datos debe estar permitido sin violar los derechos del autor o de propiedad intelectual de los datos o cualquieras obligaciones estatutarias/departamentales.

2. Planes de la gestión de datos

Los fundadores son cada vez más demandantes en el desarrollo de planes de gestión de datos como condición de recibir fondos. [6] Los requisitos normalmente son para permitir el compartir de resultados de proyectos o investigaciones, incluso a datos (y publicaciones). Por ejemplo, la UE ha publicado nuevas guías para la gestión de datos en proyectos de investigación de Horizonte 2020 a partir de Diciembre del año 2013:
ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

- El proceso de planificación para la gestión de datos empieza con una lista de chequeo para planificación de datos. Una lista de chequeo ayuda más tarde en el desarrollo de un plan de gestión de datos. Esa lista de chequeo puede incluir considerar algunas o todas de las siguientes preguntas:
- ¿Cuáles datos coleccionarás o crearas, cómo los vas a crear, y para qué propósito?
- ¿Cómo lidiaras con asuntos de ética? ¿Cómo podrás manejar asuntos de derechos de autor o de propiedad intelectual?
- ¿Cuáles formatos usaras? ¿Son no propietarios, transparentes y sostenibles? ¿Cuáles convenciones de directorio y nombramiento de archivos usarás? ¿Hay estándares formales que adoptarás? ¿Qué tipo de documentación y metadatos acompañarán a los datos?
- ¿Cómo vas a almacenar y respaldar a los datos? ¿Cómo manejarás el acceso y seguridad? ¿Quién será responsable para la gestión de los datos?
- ¿Hay protocolos existentes donde puedes basar su plan? Por ejemplo, ¿Hay políticas institucionales de protección de datos o de seguridad para seguir, guías de la gestión de datos de departamento o de grupo, definidos por su institución o fundador que se tiene que considerar?
- ¿Cuál es su plan de preservación de largo plazo para el conjunto de datos? Por ejemplo, ¿Cuáles datos deben estar retenidos, compartidos, y/o preservados? ¿Cómo compartirás a los datos, y hay algunas restricciones en el compartimiento de datos requisitos?
- ¿Cuáles recursos vas a requerir para realizar tu plan? Por ejemplo, ¿Hay herramientas o software necesario para crear, procesar, o visualizar a los datos?

Se puede agregar o restar de esta lista sugerida dependiendo en la naturaleza de su proyecto. Una lista de chequeo más detallada está disponible en el Centro de Curación Digital (*Digital Curation Centre*) en www.dcc.ac.uk/sites/default/files/documents/resource/DMP_Checklist_2013.pdf.

Nota importante: Planes de gestión de datos deben estar mantenido continuamente y actualizados a través de la vida de un proyecto o investigación

6

Tabla 2: ***Componentes de un plan de gestión de datos***

Información administrativa	• Nombre e identificación del proyecto • Descripción del proyecto • Cuerpo(s) de fundación • Contacto para datos del proyecto • Políticas relacionadas • Fecha de primera versión • Fecha de la última actualización
Colección de datos	• Descripción de datos, incluso al tipo, formato, y volumen anticipado • Conjuntos de datos existentes para ser reutilizados • Métodos para la colección o creación de datos • Estructuras, sistemas de nombramiento y creación de versiones para carpetas y archivos • Procesos para asegurar la calidad
Documentación y metadatos	• Una lista de información que esperas que necesitarías para ser leídos e interpretados en el futuro. • Como planificas coleccionar o crear esta documentación o metadatos • Los estándares de metadatos que utilizaras **Algunos ejemplos de documentación de datos:** • Cuadernos de laboratorio y protocolos experimentales • Encuestas, catálogos de códigos, diccionarios de datos • La sintaxis de software y archivos de resultados • Información sobre configuraciones de equipos y calibración de instrumentos • Esquemas de bases de datos • Reportes de metodología • Información de procedencia de fuentes y datos derivados Agregar descripciones detalladas para colecciones o archivos, por ejemplo, lo que está dentro de un archivo, de donde salió, como podría recuperarla si necesario, cualquier problema existente, etc.
Ética y consentimiento legal	**Ética** • Detalles de consentimiento necesarios para la preservación y el compartir de datos • Pasos para tomar, si es necesario, para proteger la identidad de cualquier participante • Pasos para tomar, si es necesario, para asegurar que datos sensitivos están almacenados de manera segura **Derechos de autor y de propiedad intelectual** • Nombre(s) de dueño(s) de los datos

	• Licencia(s) para la reutilización que se aplicará (por ejemplo, una de las licencias disponibles de *Creative Commons* u *Open Data Commons*) • Restricciones del uso de terceros • Posibles atrasos esperados para el compartir de datos, por ejemplo, una aplicación de patente pendiente o embargo relacionado a la publicación en un periódico
Almacenamiento y respaldo de datos	• Donde (físicamente) se almacenarán los datos • Provisión para respaldo de datos • Persona o equipo responsable por el respaldo • Protocolo para recuperación de datos **Seguridad** • Riesgos, y como se manejan • Acuerdos de acceso • Acuerdos, si necesarios, para la transferencia segura de datos coleccionados en el campo
Selección y preservación	• Detalles de cuales datos se deben retener, compartir, y/o preservar, con referencia particular a requisitos contractuales, legales, o de regulaciones • Uso probable para investigaciones con los datos • Tiempo para que los datos sean (o deben ser) retenidos más allá de la vida del proyecto • El repositorio o archivo donde se mantienen los datos, y cualquieras cargos asociados • Tiempo y esfuerzo necesario para preparar a los datos para preservar o compartir
Responsabilidades del compartir de datos y recursos	• Persona nombrada responsable para la implementación del Plan de Gestión de Datos • Persona nombrada responsable para cada actividad relacionada con la gestión de datos • El hardware y software necesario (cualquier que sea adicional a la provisión institucional existente) • Experiencia especialista o entrenamiento adicional necesario • Cargos para aplicar por los repositorios de datos

3. Organización de datos

Los archivos de datos y carpetas necesitan estar etiquetados y organizados de una manera sistemática para que sean ambos accesible y pueden ser identificados para usuarios actuales y futuros. Los beneficios de etiquetar de manera consistentes son:

- Los archivos de datos se pueden ser distinguidos dentro de una carpeta.
- El nombramiento de archivos con datos previene que haya confusión cuando más de una persona está trabajando con archivos compartidos.

- Los archivos de datos son más fáciles de encontrar y buscar.
- Los archivos de datos no solamente podrían ser recuperados por el creador de ellos, sino por otros usuarios también.
- Se puede ordenar a los archivos de manera lógica.
- los archivos de datos no son borrados o reemplazados por accidente.
- Se puede identificar las diferentes versiones de archivos de datos.
- Si se traslada los datos a otra plataforma de almacenamiento, los nombres preservarán el contexto útil de ellos.

Hay tres criterios principales para considerar en cuestión de nombrar y etiquetar a los archivos de datos, como:

- *Organización*: importante para el acceso y recuperación futuro, y necesita tomar en cuenta las limitaciones de etiquetar en el sistema donde se encuentra el archivo
- *Contexto*: esto puede incluir información descriptiva o específico al contenido, independientemente de donde se almacenan los datos.
- *Consistencia*: escoja una conversión y asegura que se siga todas las reglas de manera sistemática para siempre incluir la misma información (como fecha y hora) en el mismo orden (por ejemplo, AAAAMMDD).

Elementos comunes de una estrategia de etiquetar

- Numero de versión
- Fecha de creación
- Nombre del creador
- Descripción del contenido
- Nombre de equipo/departamento/unidad asociado con los datos
- Fecha de publicación
- Número de proyecto

Es importante también considerar lo siguiente cuando etiquetas a un archivo:

El uso de nombres genéricos puede crear un conflicto cuando se traslada el archivo de un lugar a otro. Asegúrese que los nombres de archivo son independientes del lugar.

Los nombres de archivos deben ser relevantes independiente de la persona quien creyó al archivo originalmente

Cuan fácil será expandir el protocolo de nombrar a los archivos, por ejemplo, si el número de proyecto se limita a dos dígitos, solamente puedes tener noventa y nueve proyectos.

Control de versiones

Es importante identificar y distinguir entre versiones de conjuntos de datos consistentemente. Esto asegura que haya rastros claros para examinar al desarrollo de un conjunto de datos y poder identificar versiones anteriores si se necesita. Por lo tanto, necesitarás establecer un método que tiene sentido para ti que indicará la versión de su conjunto de datos.

- Una forma común de expresar a las versiones de archivos es para usar números ordenales (como 1,2,3, por ejemplo) para cambios mayores de versiones y decimales para cambios menores (como v1, v1.2, v2.6).
- Se debe evitar etiquetas confusas como por ejemplo “revisión”, final, final2, copia definitiva, porque puede ser que estos se acumulan.
- Registrar TODOS los cambios (menores y mayores)
- Desechar o borrar versiones obsoletas (reteniendo a la versión original)
- Usa un sistema de respaldo automático (si disponible) en vez de guardar o archivar múltiples versiones
- Habilita la creación de versiones o rastreado en documentos colaborativos con aplicaciones de almacenamiento como los Wiki, o GoogleDocs
- Considera utilizar un software de control de versiones como *Subversion*, *TortoiseSVN*

Algunos ejemplos estructurados de mantener un sistema de control de versiones [nombre del documento] [número de versión] [estado: borrador/final]:

- Entrevista_Gonzalez_Juilo2010_v1_borrador
- Análisis-lipido-v2_definitivo
- 2001_01_28_ILB_CS3_V6_AB_editado

Resumen

Buenos principios de gestión de datos ayudan a asegurar que los datos producidos o usados están registrados, almacenados, hechos accesibles para utilizar y reutilizar (si adecuado), mantenidos a través del tiempo y /o eliminados, según requisitos legales o éticos del fundador y según la buena práctica.

Por lo tanto, la gestión de datos es un proceso que involucra un rango amplio de actividades desde los aspectos administrativos a técnicos del manejo de datos de una manera que abarca a los factores mencionados arriba. Una buena política de gestión de datos definirá metas estratégicas de largo plazo para la gestión de datos en todos los aspectos de un proyecto o empresa.

Una política de gestión de datos es un conjunto de principios de alto-nivel que establecen un marco guía para la gestión de datos. Una política de gestión de datos puede ser usado para abarcar asuntos estratégicos como el acceso a datos, asuntos legales relevantes, asuntos para la mayordomía en datos y responsabilidades custodiales, adquisición de datos, y otros asuntos. Planes de gestión de datos tienen que ser mantenidos continuamente y actualizados a través de la vida de un proyecto o investigación.

Los componentes de un plan de gestión de datos incluirían:

- Información administrativa
- Métodos de colección de datos y procesos de control de calidad
- Documentación y metadatos
- Conformidad ética y legal
- Almacenamiento y respaldo
- Selección y preservación
- Responsabilidades y recursos del compartir de datos

Lecturas Adicionales

Arms, C. R., Fleischhauer, C. and Murray, K. (2013). Sustainability of digital formats: planning for Library of Congress collections. Library of Congress, Washington DC, USA. Available at: www.digitalpreservation.gov/formats

Beagrie, N. and Houghton, J. (2014). *The value and impact of data sharing and curation - synthesis of three recent UK studies*. Jisc. Available at: repository.jisc.ac.uk/5568/1/iDF308__Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf

Charles Beagrie Ltd (2013). *Keeping research data safe: cost / benefit studies, tools, and methodologies focussing on long-lived data*. Available at: http://www.beagrie.com/krds.php (accessed 4 August 2014)

Digital Curation Centre (DCC). (2010). *Data management plans*. Available at: http://www.dcc.ac.uk/resources/data-management-plans (accessed 4 August 2014)

Drummond, C.G. (2009). Replicability is not reproducibility: nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop*, 26th ICML, Montreal, Canada. Available at: http://cogprints.org/7691

European Commission (EC). (2017). *Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020*. Available at: *https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf* (Version 3.2, 21 March 2017)

GODAN 2016 *A Global Data Ecosystem for Agriculture and Food*. GODAN, Wallingford, UK. Available at: http://www.godan.info/documents/dataecosystem-agriculture-and-food

Jones, S. (2011). *How to Develop a Data Management and Sharing Plan*. DCC How-to Guides, Digital Curation Centre, Edinburgh, UK. Available at: www.dcc.ac.uk/resources/how-guides/develop-data-plan

UK Data Archive (UKDA). *Plan to Share*. Available at: www.dataarchive. Available at: https://www.ukdataservice.ac.uk/manage-data/plan.aspx (accessed 4 August 2014).