

**INFORME TAREAS PARA DESARROLLO DE PLATAFORMA INTEGRADA DE DATOS**

Elaborado por:

Karla Escobar Parada

28 de diciembre de 2018

**Índice**

[**1. Introducción 3**](#_Toc533698345)

[**2. Requisitos y estándares mínimos de bases de datos para Open Science 4**](#_Toc533698346)

[**2.1. Tipos de formatos de datasets 5**](#_Toc533698347)

[**2.2. Estructura de Datasets 6**](#_Toc533698348)

[**2.3. Metadatos 6**](#_Toc533698349)

[**2.3.1. Metadatos solicitados por DataVerse para Ciencias Sociales 8**](#_Toc533698350)

[**2.4. Libro de Códigos 9**](#_Toc533698351)

[**2.5. Normas éticas y Licencias 12**](#_Toc533698352)

[**2.5.1. Normas éticas 12**](#_Toc533698353)

[**2.5.2. Licencias 13**](#_Toc533698354)

[**2.6. Catálogo 15**](#_Toc533698355)

[**3. Evaluación Bases de Datos COES 16**](#_Toc533698356)

[**3.1. Evaluación de interfaz plataforma DataVerse 16**](#_Toc533698357)

[**3.2. Evaluación Manual de Usuario y Libro de Códigos 17**](#_Toc533698358)

[**3.3. Evaluación Base de Datos 23**](#_Toc533698359)

[**4. Plataformas digitales disponibles para almacenamiento de bases de datos. 25**](#_Toc533698360)

[**5. Conclusiones 30**](#_Toc533698361)

[**6. Referencias 32**](#_Toc533698362)

# Introducción

El presente documento surge de la necesidad manifestada por el Centro de Estudios de Conflicto y Cohesión Social (COES) sobre conocer los estándares, requisitos o recomendaciones para compartir y disponer al público bases de datos en el marco de la denominada *Open Science*.

La *Open Science* se podría definir como un movimiento científico que tiene por objetivos promover la apertura y la accesibilidad del conocimiento creado a partir de las investigaciones científicas hacia todo público, sin restricciones de propiedad. Dentro de ella se encuentran el *Open Access*, que implica la apertura de las publicaciones académicas y bajas restricciones de *copyright*, el *Open Source*, es decir, la creación de softwares sin restricción de uso o propiedad, y, el *Open Data*, que se relaciona con la distribución y disponibilidad de acceso y uso de los datos obtenidos en el marco de investigaciones científicas. Es justamente en este último donde se inserta del análisis realizado en este documento.

En el primer apartado se presenta una sistematización de los requisitos y estándares mínimos que el *Open Science* y, por consiguiente, *Open Data*, sostiene para los procesos de preservación e intercambio de datos. A través de centros de referencia mundial como el *Digital Curation Centrer,* la *Data Documentation Iniciative Alliance,* la *Statistical Data and Metadata Exchange,* el *Polar Data Catalogue*, etc. junto a los estándares chilenos establecidos por el Gobierno, se entregan requisitos y recomendaciones para la estructura de bases de datos, formatos, metadatos, el libro de códigos, normas éticas y licencias, entre otros.

En el segundo apartado, se realiza un análisis de las bases de datos priorizadas por COES; ELSOC, Observatorio de Conflictos, y ENACOES; conforme a los estándares, requisitos y recomendaciones presentados en el primer apartado.

En el tercer apartado, se entrega una sistematización y análisis de los repositorios de datos abiertos disponibles gratuitamente para la preservación e intercambio de bases de datos en ciencias sociales, añadiendo los softwares más utilizados para el diseño, puesta en marcha y desarrollo de repositorios de datos institucionales.

El documento cierra presentando conclusiones o elementos que sirvan para discusión y toma de decisiones de COES respecto de la aplicación de los estándares, requisitos y recomendaciones a sus bases de datos, y del uso o creación de una plataforma para compartirlas hacia el público académico o interesado en la utilización de los datos para nuevas investigaciones o estudios.

# Requisitos y estándares mínimos de bases de datos para *Open Science*

Este punto implica adentrarse en el mundo de la gestión de información; en específico en la preservación e intercambio de datos. Según Rusbridge, et al (2005) la preservación de datos se define como la administración de datos científicos digitales a lo que se les entrega un valor añadido dándoles contexto y una vinculación a su origen, de manera que se facilite su reutilización, integración y promoción.

En este sentido se han estudiado los estándares establecidos por el *Digital Curation Centrer (DCC)*, centro de referencia mundial que crea estándares para la *data curation[[1]](#footnote-1)* y para la elaboración de planes de gestión de datos.

El *DCC* refiere al *Data Documentation Initiative (DDI) Alliance,* como la institución que entrega estándares para ciencias sociales y humanidades. Esta alianza establece estándares para describir datos producidos mediante encuestas o métodos observacionales en ciencias sociales, de la conducta, económicas y de la salud. Estos estándares son altamente técnicos y específicos para redes computacionales y formatos de software, sin embargo, dentro de esto define lo requerido para un libro de códigos de una base de datos, entre otros.

La *DCC* también propone seguir los estándares de la *Statistical Data and Metadata Exchange* *(SDMX)* en el caso de utilizar bases de datos que contengan datos estadísticos. La *SDMX* es una iniciativa internacional que estandariza y moderniza mecanismos y procesos de intercambio de datos estadísticos y sus metadatos.

Otro aspecto relevante en este punto, son los requisitos para lo que se conoce como *Open Data*, considerando que COES tiene por objetivo último avanzar hacia una política de *Open Science*. De acuerdo con el *Open Definition Advisory Council* para que un dato sea considerado “abierto” debe cumplir con las siguientes condiciones: acceso (disponible integralmente, a un coste razonable y de forma que pueda ser modificable), redistribución, reutilización, ausencia de restricciones tecnológicas, reconocimiento, integridad, sin discriminación de personas o grupos, sin discriminación de ámbitos de trabajo, distribución de la licencia, la licencia no debe ser específica de un paquete y la licencia no debe restringir la distribución de otras obras (Hernández-Pérez, T. y García – Moreno, MA. 2013, p. 260)

De acuerdo con la Norma Técnica Chilena para la publicación de datos abiertos (2013), se establecen ciertas condiciones generales para los conjuntos de datos publicados de modo de facilitar el acceso expedito, abierto y sin restricciones de uso.

1. Los *datasets* deben ser publicados de la manera menos procesada posible. Sin agregaciones estadísticas previas
2. Publicarse en formatos abiertos, estables y de amplio uso (Ver más adelante)
3. Disponibles en varios formatos
4. Cada conjunto de datos debe tener un número de identificación único y permanente
5. Se recomienda que el nombre de los archivos se presente en el estilo Lower Camel Case, es decir, se escriben las palabras juntas y sin espacios, con la primera letra de la primera palabra en minúscula y la primera letra de las siguientes palabras en mayúscula. Ej: encuestaElsoc.csv
6. Entregar datos y metadatos de cada conjunto de datos. Incluyendo el sentido y significación de la codificación utilizada (libro de códigos).
7. La estructura de los metadatos está dada por los estándares DCAT[[2]](#footnote-2).
8. Indicar procedencia de los datos en caso de *datasets* derivados de otros *datasets*, e incorporar los metadatos de los *datasets* primarios.
9. Indicar el tiempo de validez de los *datasets*.
10. Indicar periodicidad de los *datasets*.
11. Para *datasets* de datos georefenciados o de coberturas se recomienda incorporar metadatos según la norma ISO 19115 (utilizada por el Servicio Nacional de Información Territorial)
12. La codificación de caracteres estándar es UTF-8.
13. Cuidar que terceros puedan acceder a información personal privada a través de *datasets* (cruces de variables)
14. No se pueden publicar datos definidos como “datos personales” o “datos sensibles”[[3]](#footnote-3).

## **Tipos de formatos de datasets**

La Norma Chilena (2013) mencionada, establece 4 tipos de formatos de *datasets*, de acuerdo con los datos que contengan. Se pueden apreciar en la tabla 1:

***Tabla Nº1: Formatos de archivos según tipo de dato.***

|  |  |
| --- | --- |
| Tipo de dato | Formato |
| Genérico: no requieren aplicaciones especializadas | .CSV, .XML |
| Geográficos: requieren aplicaciones geográficas, representan superficie terrestre u objetos en el espacio geográfico | .JPEG, .PNG, .GeoTIFF, .WFS, .WCS, .GML, .ArcGIS, .KML |
| Imágenes: fotografías | .JPEG, .PNG, .SVG |
| Estadísticos | .SDMX[[4]](#footnote-4), .SPSS[[5]](#footnote-5), .STATA[[6]](#footnote-6), .CSV |

En base al *Polar Data Catalogue (PDC)* (2014), los archivos de los conjuntos de datos deben ser gestionados fácilmente por diversos software de datos, por tanto, los nombres de los archivos deben contener solo números, letras, guiones y guiones bajos. Espacios y signos especiales no están permitidos. Deben utilizarse minúsculas y no más de 150 caracteres de longitud. Debe incluir el número de versión del *dataset* o la fecha de creación en el título (ejemplo: encuestaElsoc\_v1.csv) El tipo de archivo preferido es .CSV, y cuando requieren ser comprimidos los preferidos son .ZIP, .TAR.

## **Estructura de Datasets**

El *Polar Data Catalogue (PDC) (2014)* en su documento sobre buenas prácticas de *Open Data* indica que, para la estructura de la base de datos, entendido como el ordenamiento que debe tener la base o planilla de los datos, se recomienda lo siguiente:

* Usar el mismo formato en todo el archivo.
* Dicho formato debe aplicarse a todos los archivos que formen parte del mismo proyecto
* Se deben informar cifras y análisis en documentos anexos
* Los encabezados de las columnas deben describir el contenido de cada columna (nombres de variables)
* Los encabezados de las columnas deben estar construidos de manera de facilitar la importación con varios sistemas. Deben contener solo números, letras y guiones bajos, sin espacios ni otros caracteres. Se prefieren letras minúsculas (ej: “edad\_años”)
* En las filas se incorporan los casos a analizar.

## **Metadatos**

¿Qué son los metadatos? son la descripción de un conjunto de datos, proporcionan el qué, dónde y cuándo de los datos, por quién fue recolectado, y su ubicación actual. Los metadatos deben definirse y presentarse para que un usuario que no esté familiarizado con el proyecto, sitios, métodos u observaciones, pueda hacer uso fácilmente del conjunto de datos. Por tanto, facilitan la comprensión, el uso y la gestión de los datos. Los metadatos también sirven para facilitar la creación de redes, la colaboración y el acceso a los datos.

Los metadatos obligatorios que define la Norma Técnica Chilena (2013) se pueden observar en la tabla 2:

***Tabla Nº2: Metadatos obligatorios según Norma Técnica Chilena (2013)***

|  |  |
| --- | --- |
| Metadatos | Definición |
| Título | Nombre definido para el *dataset* |
| Descripción | Resumen del contenido del *dataset* |
| Institución | Nombre de institución responsable de los datos |
| Licencia | Derechos de autor del *dataset* |
| Categoría | Clasificación conceptual básica del *dataset* |
| Fecha de publicación | Fecha del momento en que se publican los datos |
| Recurso | Contenido y características físicas del *dataset*: Enlace del archivo (URL), Tamaño, Formato, Descripción (características adicionales necesarias para procesamiento, ej. Versiones), Fecha de los metadatos |

Además, define metadatos recomendados: Etiquetas del *dataset* (palabras clave), frecuencia de actualización, cobertura geográfica, cobertura temporal, granularidad (nivel de detalle de los datos), documentos (información adjunta que complementa el *dataset*)

Según las buenas prácticas recopiladas por *Polar Data Catalogue (PDC)* en 2014, los metadatos más comunes debieran ser:

* Título de los datos
* Cómo se deben citar los datos (DOI u otros)
* Sitio de estudio
* Propósito: razón científica por la cual los datos fueron recolectados
* Resumen: ¿Qué datos se recogieron?, Qué instrumentos y fuente (s) fueron utilizados, Cuándo, dónde y con qué frecuencia se recopilaron los datos, Cómo se midieron o produjeron, unidades de medida, precisión y exactitud si se conocen, y la relación con otros datos en el conjunto de datos, si corresponde.
* Creador de datos, Investigador principal y otras partes responsables
* Enlace a datos (URL).
* Programa de investigación que recopila los datos
* Estado de la recopilación de datos
* Mantenimiento y actualización de la frecuencia de los datos.
* Período de tiempo
* Palabras clave
* Restricciones de acceso: “Público” o “Limitado”

Considerando los antecedentes de la Norma Técnica Chilena entregados y que COES ya ha utilizado la plataforma Harvard Dataverse para preservar y compartir algunas de sus bases de datos, y por consecuencia ha cumplido con los requisitos de publicación de dicha plataforma, se propone establecer como base los requisitos mínimos de Dataverse, además de la Norma Técnica Chilena, de modo de mantener el avance en calidad de los metadatos que ya se ha adquirido en COES.

Cabe señalar que la Norma Técnica Chilena (2013) establece su aplicación a la administración pública de modo que los organismos privados no necesariamente deben regirse por dicha norma. Sin embargo, COES obtiene recursos de fondos públicos por tanto sería óptimo que contara con la aplicación de los estándares establecidos por ésta como mínimo.

## **Metadatos solicitados por DataVerse para Ciencias Sociales**

En la tabla 3, se listan una serie de metadatos solicitados por Dataverse para la publicación de los *datasets* correspondientes y exclusivos a Ciencias Sociales. Cabe señalar que no todos los metadatos son obligatorios.

***Tabla Nº3: Metadatos solicitados por DataVerse para Ciencias Sociales***

|  |  |
| --- | --- |
| Nombre | Definición |
| Título | Nombre del *dataset* |
| Unidad de análisis | Unidad de análisis o de observación de la base de datos |
| Universo | Descripción de la población cubierta por la base de datos. |
| Tipo de metodología | Metodología utilizada para el estudio: transversal, longitudinal, etc. |
| Recolector de datos | Institución u otro responsable de la recolección de datos |
| Entrenamiento de recolección de datos | Tipo de entrenamiento o capacitación provista para el proceso de recolección de datos |
| Frecuencia | Frecuencia de recolección de datos: mensual, anual, etc. |
| Procedimiento muestral (diseño) | Tipo de diseño muestral utilizado |
| Target tamaño muestral | Información específica del target para la muestra |
| Tamaño muestral actual | Nº de muestra actual |
| Fórmula tamaño muestral | Fórmula para calcular la muestra |
| Mayores desviaciones del diseño muestral | Discrepancias entre la muestra y la población del estudio (estimaciones de sexo, nivel socioeconómico, etc.) |
| Modo de recolección de datos | Método de recolección de datos |
| Tipo de instrumento de recolección | Características del instrumento utilizado. Estructurado, semi estructurado, etc. |
| Características de la recolección de datos | Descripción de aspectos relevantes sobre la recolección de datos. Ejemplo: Duración, tasa de respuesta o cooperación, etc. |
| Acciones para minimizar pérdidas de datos | Listado de acciones tomadas para minimizar la pérdida de datos. Ej: Supervisiones. |
| Control de recolección de datos | Métodos para controlar la recogida de los datos. |
| Ponderador (muestral) | Cómo se calcularon los pesos muestrales, incluyendo la fórmula si corresponde |
| Operaciones de limpieza de datos | Cómo se limpiaron los datos, por ejemplo: chequeos de consistencia. |
| Notas sobre errores | Cualquier elemento no descrito anteriormente que aclare el procedimiento o la metodología del estudio |
| Tasa de respuesta | El porcentaje de los encuestados que respondió la encuesta o estudio. |
| Medidas de errores muestrales | Medidas de precisión del estudio para estimar la población a través de la muestra indicada. |
| Otras formas de evaluación de datos | Otras cuestiones relacionadas con la evaluación de los datos. Describa problemas como la varianza de la respuesta, la tasa de no respuesta y las pruebas de sesgo, sesgo del entrevistador y de respuesta, niveles de confianza, sesgo de preguntas o similares. |
| Notas generales | Notas generales sobre el *dataset* |
| Tipo de nota |
| Tema de la nota |
| Texto de la nota |

## **Libro de Códigos**

Tal como se ha mencionado anteriormente, el libro de códigos es uno de los documentos anexos que se plantean como estándares para compartir datos. Se observa como un documento central que permite entender la estructura de la base de datos, en cuánto variables, dimensiones y subdimensiones.

La *Data Documentation Initiative (DDI) Alliance,* establece como estándares para el libro de códigos los que se presentan en la tabla 4:

***Tabla Nº4: Estándares para libro de códigos según DDI Alliance.***

|  |  |  |
| --- | --- | --- |
| Dimensiones | Subdimensiones | Especificación/Explicación |
| **Producción de Metadatos** | Productores de Metadatos | Quiénes son los producen los metadatos de la Base de datos |
| Fecha de la producción | Cuando se produjeron estos metadatos |
| Versión (metadatos) | Qué versión tienen |
| Nº de Identificación (metadatos) | Nº de identificación de los metadatos |
| **General** | Nombre | Nombre de la BD |
| Nº de Identificación | Nº único de identificación |
| Versión y fecha | Versión y fecha de la BD |
| Serie | Incluye la fecha de aplicación, y el nº de veces aplicado. |
| Resumen | Un resumen sobre la BD y el estudio. |
| Tipo de data | Encuesta u otro |
| Unidad de Análisis | Cuál es la unidad de análisis |
| **Alcance y Cobertura** | Palabras clave | Cuáles son las palabras claves de la BD |
| Tópicos | Los tópicos que aborda |
| Período de tiempo | El período de tiempo que aborda (ej: año 2018) |
| País | El/los país(es) dónde se aplicó |
| Cobertura geográfica | Cuál fue el área de cobertura |
| Universo | Cuál es el universo de la muestra |
| **Productores y Patrocinadores** | Investigadores primarios | Nombre y contacto |
| Otros | Nombre y contacto |
| **Muestra** | Procedimiento muestral | Cómo se estimó la muestra |
| Tasa de Respuesta | La tasa de respuesta del cuestionario |
| Ponderación | Cómo se ponderaron los casos. Si corresponde |
| **Recolección de Datos** | Fecha | Período de recolección de los datos |
| Técnica de Recolección | Qué técnicas se utilizaron |
| Notas | Otra información relevante |
| Tipo de cuestionario | Estructurado – Semiestructurado - Abierto. |
| **Procesamiento de datos** | Estimación de errores muestrales | Cómo se estimaron los errores muestrales |
| **Accesibilidad** | Autoridad de Acceso | Propietario de BD |
| Contacto |  |
| Distribuidor | Quién tiene derecho de distribución de la BD |
| Depositario | Dónde están depositados los datos |
| Condiciones de acceso | Principalmente el tipo de licencia |
| Requerimientos de Citación | Cómo citar |
| Copyright | Si existe algún copyright asociado |
| **Descripción del archivo** | Nº de archivos | Archivos que contiene la BD |
| Nombre | Nombre de la BD |
| Casos | Nº de casos que contiene |
| Variables | Nº de variables que contiene |

Para cada variable o grupo de variables, la *DDI Alliance,* define los siguientes requerimientos, a saber, en la tabla 5:

***Tabla Nº5: Estándares para variables o grupo de variables en libro de códigos, según DDI Alliance.***

|  |  |  |
| --- | --- | --- |
| Dimensión | Subdimensión | Explicación |
| **Grupo de variables** | Nombre Grupo | Ej: Variables de identificación |
| Nombre variable | Nombre de la variable en la BD (ej: edad\_años) |
| Etiqueta | La explicación de la variable |
| Tipo | Discreta, continua, etc. |
| Formato | Numérico, cadena, etc. |
| Casos Válidos | Nº de casos válidos para la variable |
| Casos inválidos | Nº de casos inválidos para la variable |
| Pregunta | Cuál es la pregunta. Si corresponde |

Luego, en la tabla 6, se puede observar para cada variable, que la *DDI Alliance* establece los siguientes requerimientos:

***Tabla Nº6: Estándares para variables en el libro de códigos, según DDI Alliance.***

|  |  |  |
| --- | --- | --- |
| Dimensión | Subdimensión | Explicación |
| Descripción de la variable | Nombre de variable | Nombre en Base de Datos. |
| Información | Tipo, formato, Rango, valores perdidos. |
| Estadísticas | Casos válidos, casos inválidos |
| Definición conceptual | Qué se entiende por |
| Universo | La población de la muestra que contiene esta variable |
| Notas | Cualquier otra información que se considere importante |
| Categorías | Para cada categoría entrega el valor, etiqueta, Nº casos, ponderación, porcentaje ponderado. |

Como puede observarse, más que un libro de códigos, lo que recomienda la *DDI Alliance*, es presentar un manual de usuario con la mayor cantidad de información necesaria para poder utilizar el *dataset* fácilmente. Si se considera necesario se recomienda incorporar cualquier documento anexo, tales como: cuestionario, estudios previos, análisis previos realizados, pre-análisis de la BD, etc. Como parte anexa de este manual o libro de códigos. Sin embargo, se debe considerar que la información que se entrega es para uso y acceso del *dataset* correspondiente, no pretender justificar el desarrollo del estudio ni otros objetivos que no sean la mera utilización de los datos en el contexto correspondiente.

## **Normas éticas y Licencias**

## **Normas éticas**

Luego de un análisis de la información disponible sobre normas éticas en el *Open Science,* se ha logrado identificar que no existen estándares de uso a nivel global, dado que los requisitos principalmente están regidos por la legislación existente en cada país.

En el caso de Chile, existen dos leyes que se relacionan con el acceso a datos personales o datos sensibles que podrían entregar un fundamento legal para definir estándares éticos, estas son:

* Ley 19.628, que otorga protección sobre la vida privada, considerando datos de carácter personas aquellos relativos a cualquier información concerniente a personas naturales, identificadas o identificables. Esta ley indica que el tratamiento de los datos puede ser efectuado por cualquier persona siempre que se haga de manera concordante con la legislación y que se respeten los derechos fundamentales de los titulares de los datos.
* Ley 21.096, publicada recientemente (junio de 2018) que otorga el derecho de protección de los datos personales.

A pesar de la no existencia de estándares éticos, el *Open Data Institute* plantea recomendaciones para evaluar los aspectos éticos del uso de datos identificando el potencial impacto adverso en las personas, y para garantizar que la recopilación, uso y distribución de los datos estará acorde a las legislaciones y aminorar los problemas relativos a la privacidad.

Para ello, el *Open Data Institute* creó un esquema que permite definir las propias pautas éticas a través de tópicos y preguntas a responder en cada situación de acceso, uso o distribución de datos. Este esquema se denomina *The Data Ethics Canvas[[7]](#footnote-7)*. A través de 15 tópicos, permite preguntarse acerca de los dilemas éticos que rodean la investigación.

Los puntos son:

1. Fuentes de datos: dónde se recolectarán o cómo se accederá a ellos, etc.
2. Limitaciones de las fuentes de datos: ¿hay limitaciones que influencien los resultados? Considerar sensibilidades respecto de las categorizaciones de los datos, etc.
3. Compartiendo estos datos con otras organizaciones: ¿Cómo? ¿Bajo qué condiciones? (acá remite a licencias y copyright que se abordará a continuación), etc.
4. Legislación y políticas relevantes: ¿Qué legislación es relevante para el uso de estos datos?, etc.
5. Derechos sobre las fuentes de datos: ¿Tiene permiso para usar los datos?, etc.
6. Marco ético existente: países, sectores o comunidades que tienen marcos o códigos éticos, etc.
7. Razones para usar estos datos: ¿Cuáles son los objetivos para el uso de estos datos?, etc.
8. Comunicando tu propósito: Los titulares de los datos a utilizar, ¿están en conocimiento del propósito? ¿Entienden el propósito?, etc.
9. Efectos positivos en las personas: ¿quiénes podrían tener efectos positivos producto del uso de estos datos? ¿se medirá el impacto positivo?, etc.
10. Efectos negativos en las personas: ¿quiénes podrían tener efectos negativos? ¿Cuáles? ¿Causan daño?, etc.
11. Minimizando efectos negativos: ¿qué pasos se pueden usar para minimizar efectos negativos?
12. Compromiso con la gente: ¿cómo las personas se pueden comprometer con el estudio? ¿pueden solicitar cambios?, etc.
13. Comunicando riesgos y otros: ¿Se consideran las ideas o pensamientos de las personas afectadas por el proyecto?, etc.
14. Revisiones e iteraciones: ¿Cómo se monitorearán y discutirán los problemas relacionados con la ética de los datos?, etc.
15. Tus acciones: ¿Qué acciones se tomarán previo al desarrollo del estudio? ¿Cuáles serán prioritarias?, etc.

## **Licencias**

Los tipos de licencias disponibles para proporcionar y compartir datos abiertos son:

1. Creative Commons (CC)[[8]](#footnote-8): esta organización ofrece licencias de acuerdo con cuatro condiciones:
   1. Reconocimiento (Attribution): El material creado puede ser copiado, utilizado, o distribuido por terceras personas siempre que se cite la fuente.
   2. No Comercial (Non Commercial): el material creado puede ser copiado, utilizado o distribuido por terceras personas siempre que su uso no sea comercial.
   3. Sin Derivar (No Derivate Works): El material creado por las terceras personas puede ser copiado, utilizado o distribuido, pero no se puede utilizar para realizar un trabajo derivado del original.
   4. Licenciar Igual (Share alike): El material creado por terceras personas puede ser modificado y distribuido, pero bajo la misma licencia que el material original.

En función de la combinación de estas cuatro condiciones CC entrega seis opciones de licencias:

1. Reconocimiento (CC-BY)
2. Reconocimiento – Compartir Igual (CC BY-SA)
3. Reconocimiento – Sin Derivar (CC BY-ND)
4. Reconocimiento – No Comercial (CC BY-NC)
5. Reconocimiento – No Comercial – Compartir igual (CC BY-NC-SA)
6. Reconocimiento – No Comercial – Sin Derivar (CC BY-NC-ND)

Además, entrega la opción de una licencia para dominio público:

* Dedicación a dominio público (CC0): implica que el titular de derechos de obra o de bases de datos renuncia a los intereses de su obra en el mundo.

1. Open Data Commons[[9]](#footnote-9): Sitio de la *Open Knowledge* *Foundation* que entrega las herramientas legales para el uso de *Open Data,* específicamente para bases de datos*.*

Entrega tres tipos de licencias:

* 1. Public Domain Dedication and License (PDDL): Licencia para bases de datos de dominio público, es decir, renuncia a todos los derechos.
  2. Attribution License: Permite compartir, copiar, distribuir, producir trabajos, modificar, transformar y construir nuevas bases de datos, siempre que se reconozca su origen. En caso de producir trabajos y construir nuevas bases de datos se debe dejar claro la licencia de la base y mantener intacta cualquier nota de la base original.
  3. Open Database License (ODbl): Permite compartir, copiar, distribuir, producir trabajos, modificar, transformar y construir nuevas bases de datos, siempre que se reconozca su origen, se compartan los trabajos o nuevas bases de datos bajo la misma licencia y se mantengan los datos abiertos.

1. Sistema GNU[[10]](#footnote-10): Es una comunidad que impulsa el software libre, que comenzó con el desarrollo de un sistema operativo libre GNU. Entrega cuatro tipos de licencias, principalmente para softwares.
   1. Licencia Pública General (GPL GNU): Es una licencia para software de dominio público. Se mantienen los derechos del propietario, pero solo para garantizar que cada usuario tenga libertad para usar el software para cualquier propósito, cambiarlo, compartirlo, y compartir los cambios realizados.
   2. Licencia Pública General Reducida (LGPL GNU): Licencia para software que permite generar softwares propietarios.
   3. Licencia Pública General Affero (AGPL): Se basa en la GPL, pero contiene una cláusula adicional que permite a los usuarios interactuar con el programa licenciado a través de una red para recibir el código fuente de ese programa. Se recomienda utilizar para cualquier programa que vaya a ejecutarse a través de una red.
   4. Licencia de Documentación Libre (FDL GNU): pensada para manuales, libros de texto u otros documentos, para garantizar que todo el mundo tenga la libertad de copiar y redistribuir la obra, con o sin modificaciones, de modo comercial o no comercial.

La Norma Técnica Chilena (2013) establece tres tipos de licencia aceptadas:

* Creative Commons 3.0 (actualmente CC va en la versión 4.0)
* GNU GLP
* Open Database License (ODbl)

## **Catálogo**

En el caso del catálogo de datos, la Norma Técnica Chilena (2013) define ciertas condiciones para que las instituciones construyan su catálogo:

1. Listado de los *datasets* existentes en el repositorio, los cuáles deben ser accesibles desde el catálogo central.
2. Debe permitir navegación, búsqueda y consulta simple de los *datasets* contenidos en el repositorio
3. El catálogo debe entregar los metadatos establecidos como mínimos en la norma técnica chilena (2013)
4. Debe permitir retroalimentación de los usuarios, como consultas y/o sugerencias
5. El responsable del catálogo debe velar por su funcionamiento, mantener su disponibilidad y analizar estadísticas de uso.

Las condiciones del catálogo remiten inmediatamente a la plataforma web o software que será utilizado para la preservación e intercambio de datos, opciones que se desarrollan en el punto tres.

# Evaluación Bases de Datos COES

En este apartado se presenta la evaluación de las Bases de Datos COES de acuerdo con el cumplimiento de los estándares y/o requisitos que plantea la *Open Science* y que se han descrito en el apartado anterior.

El análisis se ha divido en tres puntos: 1) evaluación de la interfaz de la plataforma Harvard DataVerse, que incluye identificar la facilidad de acceso y disponibilidad de cada base de datos a través de esta plataforma. 2) evaluación del Manual de usuario y Libro de códigos, que implica el análisis de los metadatos y las dimensiones del libro de códigos. 3) evaluación de la base de datos, en términos de estructura y formato.

## **Evaluación de interfaz plataforma DataVerse**

En este punto se analizan las bases de datos COES de acuerdo con los estándares requeridos por la Norma Técnica Chilena (NTCh, 2013) respecto de lo que debe incluir un catálogo de datos. Se han identificado con Sí (presencia) o No (ausencia) cada una de las categorías de análisis para cada base de datos, y en algunos casos se ha añadido más información relevante sobre la evaluación.

***Tabla Nº7: Evaluación plataforma DataVerse por cada Base de datos COES***

|  |  |  |  |
| --- | --- | --- | --- |
| Categorías por evaluar | *Datasets* | | |
| ***ELSOC*** | ***ENACOES*** | ***Obs. Conflictos*** |
| Nº identificación único | SÍ, tiene DOI | SÍ, tiene DOI | No está disponible para descarga |
| Nombre archivos (minúscula, sin espacios) | NO, se usan mayúsculas para algunos archivos | NO, todos en mayúsculas. | No está disponible para descarga |
| Enlace a web COES y viceversa | NO | NO | No está disponible para descarga |
| Metadatos disponibles[[11]](#footnote-11) | SÍ | SÍ | No está disponible para descarga |
| Posibilidad retroalimentación usuarios | NO | NO | No está disponible para descarga |
| Facilidad de descarga | SÍ | SÍ | No está disponible para descarga |
| Documentos adjuntos | Manual de usuario y libro de códigos (están en el mismo archivo) | Manual de usuario (incluye como anexo libro de códigos y otros) | No está disponible para descarga |
| Otros | Sólo versión ELSOC 2016 (transversal) | La referencia/cita menciona ELSOC. El nº de versión que registra la plataforma no corresponde al nº de versión de la base de datos. | No está disponible para descarga |

## **Evaluación Manual de Usuario y Libro de Códigos**

En este punto se analizan las bases de datos COES de acuerdo con los estándares requeridos por la Norma Técnica Chilena (NTCh, 2013), *DDI Alliance y Harvard DataVerse* respecto de lo que debe incluir un manual de usuario y un libro de códigos, los cuáles se analizan por separado. Se han identificado con Sí (presencia) o No (ausencia) cada una de las categorías de análisis para cada base de datos, y en algunos casos se ha añadido más información relevante sobre la evaluación.

***Tabla Nº8: Evaluación del Manual de usuario de cada Base de datos COES***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Categorías por evaluar  Manual Usuario | | *Datasets* | | |
| **Metadatos NTCh[[12]](#footnote-12)** | | ***ELSOC*** | ***ENACOES*** | ***Obs. Conflictos*** |
| Título | | SÍ | SÍ | SÍ |
| Descripción | | SÍ | SÍ | SÍ |
| Institución | | SÍ | SÍ | SÍ |
| Licencia | | NO, no indica la licencia, pero indica las limitaciones del uso de la base de datos | NO, no indica la licencia, pero indica las limitaciones del uso de la base de datos | NO, no indica la licencia, pero indica las limitaciones del uso de la base de datos |
| Clasificación conceptual Dataset | | NO, pero se puede desprender de la presentación. | NO, pero se puede desprender de la presentación. | NO, pero se puede desprender de la presentación. |
| Fecha de publicación | | SÍ | NO, se puede obtener en Harvard DataVerse | NO |
| Enlace URL | | NO, se menciona descarga en sitio web oficial, pero no indica cuál es. | SÍ, menciona coes.cl; sin embargo, ese sitio no dirige directamente a la base de datos. | SÍ, menciona enlace; sin embargo, ese sitio no dirige directamente a la base de datos. |
| Tamaño archivo (MB/GB) | | NO | NO | NO |
| Formato | | SÍ | NO | NO |
| Versiones | | SÍ | SÍ | NO |
| Fecha metadatos | | NO, se podría considerar la fecha de publicación del manual | NO, se podría considerar la fecha de publicación del manual | NO |
| Palabras clave | | SÍ | NO | NO |
| Frecuencia actualización dataset | | SÍ, sin embargo, a la fecha debiera estar disponible la Ola 2. | SÍ, indica que es solo 2014. | NO, pero a la lectura da la impresión que es anual |
| Cobertura geográfica | | SÍ | NO, pero se desprende de la lectura de población objetivo | SÍ |
| Cobertura temporal | | SÍ | SÍ | SÍ |
| Documentos adjuntos | | NO, se mencionan documentos para no se disponibilizan, se deben solicitar. | SÍ, se anexan al Manual el libro de códigos, sintaxis, y otros. | SÍ, se anexan al manual, libro de códigos, listado de variables en Excel y archivo README |
| **Metadatos DataVerse[[13]](#footnote-13)** | | | | |
|  | *Unidad de análisis* | SÍ | SÍ | SÍ |
| *Universo* | SÍ, pero como población objetivo, no como universo. | SÍ, pero como población objetivo, no como universo. | SÍ, se desprende |
| *Metodología* | SÍ | SÍ | SÍ |
| *Responsable recolección de datos* | SÍ | SÍ | SÍ, se desprende |
| *Tipo entrenamiento recolección de datos* | SÍ | SÍ | NO |
| *Frecuencia de recolección de datos* | SÍ | NO | NO |
| *Diseño muestral* | SÍ | SÍ | NO |
| *Target Tamaño muestral[[14]](#footnote-14)* | NO | NO | NO |
| *Tamaño muestral actual* | SÍ | SÍ | NO |
| *Fórmula tamaño muestral* | SÍ | NO | NO |
| *Mayores desviaciones del diseño muestral* | SÍ | SÍ | NO |
| *Modo de recolección de datos* | SÍ | SÍ | SÍ |
| *Tipo de instrumento de recolección* | SÍ | SÍ | NO |
| *Características de la recolección de datos* | SÍ | SÍ | SÍ |
| *Acciones para minimizar pérdidas de datos* | SÍ | SÍ | NO |
| *Control de recolección de datos* | SÍ | SÍ | SÍ |
| *Ponderación (muestral)* | Sí | NO | NO |
| *Operaciones de limpieza de datos* | NO | NO | NO |
| *Notas sobre errores* | SÍ | SÍ | NO |
| *Tasa de respuesta* | SÍ | SÍ | NO |
| *Medidas de errores muestrales* | SÍ | SÍ | NO |
| *Otras formas de evaluación de datos* | SÍ | NO | NO |

***Tabla Nº9: Evaluación del Libro de Códigos de cada Base de datos COES***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Categorías por evaluar | | *Datasets* |  |  |
| **Libro de Códigos[[15]](#footnote-15)** | | ***ELSOC*** | ***ENACOES*** | ***Obs. Conflictos*** |
| Producción de Metadatos | *Productores de Metadatos* | NO, pero se subentiende en el Manual que es COES | NO, pero se subentiende en el Manual que es COES | NO |
| *Fecha de la producción* | NO | NO | NO |
| *Versión (metadatos)* | NO | NO | NO |
| *Nº de Identificación (metadatos)* | NO | NO | NO |
| General | *Nombre* | SÍ, en el Manual | SÍ, en el Manual | SÍ, en el Manual |
| *Nº de Identificación* | SÍ, en Harvard DataVerse | SÍ, en Harvard DataVerse | NO |
| *Versión y fecha* | SÍ, en el Manual | Sólo versión en el Manual | NO |
| *Serie* | NO, aunque en el manual se mencionan las Olas del estudio. | NO | NO |
| *Resumen* | NO | NO | NO |
| *Tipo de data* | SÍ, en el Manual, aunque no se incluye cuestionario | SÍ, en el Manual. Se incluyen cuestionarios. | SÍ, en el Manual |
| *Unidad de Análisis* | SÍ, en el Manual | SÍ, en el Manual | SÍ, en el Manual |
| Alcance y Cobertura | *Palabras clave* | SÍ, se desprende del Manual y está en Harvard DataVerse | NO | NO |
| *Tópicos* | NO | NO | NO |
| *Período de tiempo* | SÍ, en el Manual | SÍ, en el Manual | SÍ, en el Manual |
| *País* | SÍ, en el Manual | SÍ, en el Manual | SÍ, en el Manual |
| *Cobertura geográfica* | SÍ, en el Manual | Se desprende del Manual | SÍ, en el Manual |
| *Universo* | SÍ, en el Manual | SÍ, en el Manual | NO |
| Productores y Patrocinadores | *Investigadores primarios* | SÍ, en el Manual | SÍ, en el Manual | SÍ, en el Manual |
| *Otros* | SÍ, en el Manual | SÍ, en el Manual | NO |
| Muestra | *Procedimiento muestral* | SÍ, en el Manual | SÍ, en el Manual | NO |
| *Tasa de Respuesta* | SÍ, en el Manual | SÍ, en el Manual | NO |
| *Ponderación* | SÍ, en el Manual | NO | NO |
| Recolección de Datos | *Fecha* | SÍ, en el Manual | SÍ, en el Manual | NO |
| *Técnica de Recolección* | SÍ, en el Manual | SÍ, en el Manual | SÍ, en el Manual |
| *Notas* | SÍ, en el Manual | SÍ, en el Manual | NO |
| *Tipo de cuestionario* | SÍ, en el Manual | SÍ, en el Manual | NO |
| Procesamiento de datos | *Estimación de errores muestrales* | SÍ, en el Manual | SÍ, en el Manual | NO |
| Accesibilidad | *Autoridad de Acceso* | SÍ, en el Manual | SÍ, en el Manual | SÍ, en el Manual |
| *Contacto* | SÍ, pero menciona solo un mail, se podría identificar el responsable, en el Manual | SÍ, pero menciona solo un mail, se podría identificar el responsable, en el Manual | SÍ, pero menciona solo un mail, se podría identificar el responsable, en el Manual |
| *Distribuidor* | SÍ, en el Manual | SÍ, en el Manual | SÍ, en el Manual |
| *Depositario* | SÍ, en el Manual | SÍ, en el Manual | NO |
| *Condiciones de acceso* | SÍ, en el Manual | SÍ, en el Manual | SÍ, en el Manual |
| *Requerimientos de Citación* | SÍ, en el Manual | SÍ, en el Manual | SÍ, en el Manual |
| *Copyright* | NO | NO | NO |
| Descripción del archivo | *Nº de archivos* | NO | NO | SÍ, en archivo README |
| *Nombre* | NO | NO | SÍ, en archivo README |
| *Nº Casos* | SÍ | SÍ, en el Manual | NO |
| *Nº Variables* | SÍ | NO | NO |
| Grupo de variables | *Nombre Grupo* | SÍ | NO | NO |
| *Nombre variable* | SÍ | SÍ | SÍ |
| *Etiqueta* | SÍ | SÍ | SÍ |
| *Tipo* | En general NO, pero hay variables que sí presentan el tipo de variable que es. | NO | SÍ, a veces se reporta tipo |
| *Formato* | NO | NO | SÍ, a veces se reporta formato |
| *Casos Válidos* | NO, se subentiende que el nº de casos son casos válidos. | NO | NO |
| *Casos inválidos* | NO | NO | NO |
| *Pregunta* | SÍ | SÍ (reemplaza etiqueta) | NO |
| Descripción de la variable | *Nombre de variable* | SÍ | SÍ | SÍ |
| *Información* | Tipo y formato NO. Rango y valores perdidos SÍ | NO | A veces tipo o formato. Nunca valores perdidos ni rango |
| *Estadísticas* | NO | NO | NO |
| *Definición conceptual* | NO, solo se entrega etiqueta | NO. Para CIT en algunos casos existe breve definición conceptual | A veces se entregan breves definiciones conceptuales |
| *Universo* | NO, aunque hay variables que en “valores perdidos” parece dar cuenta de ello. | NO. Para CIT hay alusión al universo. | NO |
| *Notas* | NO | NO | NO |
| *Categorías* | SÍ (Entrega frecuencias y %) | SÍ, solo valores de categorías. No incluye variables CIT | A veces se entregan las categorías de la variable. Para las variables P13, P15, P18, P19, P20 y P29 hay más información. |

En resumen, a nivel general para las tres bases de datos analizadas, se propone mejorar el Manual de Usuario y Libro de códigos de acuerdo con los estándares *Open Science*, pero además se sugiere ordenar la información presentada.

En el proceso de análisis se observó que si bien, gran parte de la información requerida se encuentra contenida en los documentos, está redactada en formato de informe de investigación social, indicando no solo los metadatos requeridos sino también decisiones y el porqué de esas decisiones, incluyendo referencias académicas y explicaciones para quienes no son versados en el tema. Lo anterior es óptimo en contextos académicos e investigativos, pero no necesariamente para los estándares *Open Science* en donde se espera que la información sea de fácil acceso y comprensión. Se propone utilizar como modelo el sugerido por la *DDI Alliance*, tomando como ejemplo el siguiente documento: <https://www.ddialliance.org/sites/default/files/National%20Household%20Survey,%202011%20%5bCanada%5d%20Public%20Use%20Microdata%20File%20(PUMF)-%20Individuals%20File.pdf>.

Como se podrá observar ahí, el formato es simple y solo incluye la información necesaria como metadato. Cualquier otra información se obtiene en documentos adicionales o a través del contacto con el investigador.

## **Evaluación Base de Datos**

En este punto, se evalúan las bases de datos COES de acuerdo con los estándares presentados en el apartado 1. Se consideró la estructura del *dataset*, los tipos de formato, y la privacidad o identificabilidad de los datos. Se han identificado con Sí (presencia) o No (ausencia) cada una de las categorías de análisis para cada base de datos, y en algunos casos se ha añadido más información relevante sobre la evaluación.

***Tabla Nº10: Evaluación de archivos de bases de datos COES***

|  |  |  |  |
| --- | --- | --- | --- |
| Categorías por evaluar | *Datasets* | | |
| ***ELSOC (SPSS)*** | ***ENACOES (SPSS)*** | ***Obs. Conflictos (SPSS)*** |
| Datos abiertos (no agregados) | SÍ | SÍ | SÍ |
| Formato abierto (.CSV) | NO .CSV, pero está en R | NO .CSV, pero está en R | NO |
| Formatos disponibles (SPSS, STATA, .CSV) | SPSS – STATA - R | SPSS – STATA - R | SPSS-STATA |
| Variables en columnas | SÍ | SÍ | SÍ |
| Nombres variables (minúsculas, letras, números y guiones bajos) | SÍ | NO, variables en mayúsculas. | SÍ |
| Etiquetas de variables | SÍ, hay claridad de a qué corresponde cada variable. | De 384 a 537 la etiqueta es el nombre de la variable. Las demás podrían mejorar indicando si es pregunta, grado de acuerdo u otro. | Algunas no quedan tan claras para quién no conoce la BD |
| Casos en filas | SÍ | SÍ | SÍ |
| Tipo variable (Unidad de medida) | Revisar variables que toman valores numéricos, pero son nominales, ej: Miembro del Hogar 1: Sexo; Folio identificador del encuestado. | Solo se registran las variables nominales | Solo se registran las variables nominales |
| Datos personales | Peso, estatura, escala de estado de ánimo, ingresos, podrían ser considerados sensibles. | Se registran nombres de pila de conocidos | NO. La unidad de análisis no son personas. |
| Identificabilidad de datos | Variables sociodemográficas como sexo, edad, nivel educacional, ocupación, nivel educacional padre y madre, manzana censal, comuna, etc. Podrían permitir identificar al encuestado/a. al cruzarlo con peso y estatura. | La variable S15\_A y S15\_B cruzadas con variables sociodemográficas como sexo, edad, nivel educacional, ocupación, nivel educacional padre y madre, manzana censal, etc. Podrían permitir identificar al encuestado/a. Sumado a los nombres de pila de conocidos, lo que genera más posibilidades de identificación. | NO. La unidad de análisis no son personas. |

# Plataformas digitales disponibles para almacenamiento de bases de datos.

Considerando lo expuesto en el punto 1, la definición de la plataforma o software para preservar, disponibilizar, y compartir las bases de datos COES es clave para avanzar una política de *Open Science.* La elección, sin embargo, deberá estar condicionada por ciertos aspectos técnicos como la facilidad de recuperación de los datos y el acceso a estos, la capacidad de almacenamiento, la disponibilidad de licencias de uso que se quiere otorgar a los datos, etc.

A continuación, se presentan varias plataformas online o software que tienen por función ser repositorios de datos.

* Dataverse Proyect: <https://dataverse.org/institutions> o <https://dataverse.harvard.edu/>

Es un software/plataforma para gestión de datos. Permite administrar datos, linkear los datos con el sitio web de COES, pertenecer a la comunidad científica de Dataverse. Un dataverse se puede configurar para investigadores individuales, departamentos, revistas y organizaciones. Permite la administración completa del Dataverse, gestionando roles y privilegios de colaboradores, curadores y administradores. Tiene opción de idioma. La opción sin software es Harvard Dataverse. Ambos tienen buscador y buscador avanzado que garantiza el acceso a la base de datos, permiten descargar los datos, genera una DOI, licencia CC0 y cita y permite descargar la cita en los formatos más comunes. Solicita y muestra metadatos, permite la gestión de versiones de los datos (asegura trazabilidad), se basa en la *DDI* para la conservación a largo plazo de los datos.

* Figshare: <https://figshare.com/> o <https://figshare.com/services/institutions>

Repositorio de datos, permite la gestión de datos. A todo el material publicado a través de Figshare se le asigna automáticamente un DOI y una referencia bibliográfica que enlaza directamente a tu contenido en Figshare. Proporciona licencias CC0 para las bases de datos. Genera documentos QR del material y permite integración y difusión en redes sociales y gestores de referencias. Si los datos se guardan en abierto, el espacio disponible es ilimitado; en caso contrario proporciona 1 GB. Se solicitan escasos metadatos y no se muestran al usuario. Existe una versión para instituciones, pero se debe solicitar.

* Open Science Framework (OSF): <https://osf.io/>

Plataforma que permite administrar proyectos de investigación. Incorpora todo el ciclo de la investigación, desde el diseño hasta las bases de datos y los artículos académicos disponibles. Permite enlazar archivos con repositorios de datos abiertos o cerrados como Dropbox, Figshare, DataVerse, GitHub, Google Drive, etc. Permite compartir en redes sociales y con la comunidad de OSF. Es gratis. El tamaño de los archivos no puede exceder los 5GB, sin embargo, de requerir más espacio se puede solicitar. Entrega DOI y dispone de licencias de uso para los datos.

* Dryad: <https://datadryad.org/>

Repositorio de datos, permite la gestión de datos. Proyecto colaborativo con DSpace. Admite datos relacionados con publicaciones, incluso durante la fase de revisión del artículo si la revista está asociada a Dryad. Es de uso gratuito hasta 10GB por fichero y asigna como DOI el del artículo (Nina-Alcocer, V; Blasco-Gil, Y; Peset, F., 2013). Escasos metadatos para *datasets*.

* Zenodo: <https://zenodo.org/>

Repositorio de datos, permite la gestión de datos. Entrega hasta 2GB gratis para almacenar datos. Verifica la relación de los *datasets* con alguna investigación, permite crear comunidades de usuarios. No muestra los metadatos de la base.

* DataHub <https://datahub.io/>

Repositorio y administrador de datos, entre otras cosas permite la previsualización de la base de datos cuando está en formato .csv. Alta calidad de los datos. Aloja datos que también están en otras webs. Permite crear colecciones de datos. No es gratis.

* Open ICPSR <https://www.openicpsr.org/openicpsr/> : Repositorio de datos para ciencias sociales
* Re3data <https://www.re3data.org/>: buscador de repositorios de datos.
* DataSearch Google [https://toolbox.google.com/*datasets*earch](https://toolbox.google.com/datasetsearch) : Buscador de Google de *datasets*.

Además, se incluyen software de gestión y administración de datos que permiten crear el propio repositorio de datos abiertos.

* CKAN <https://ckan.org/>

Software libre para facilitar la creación de catálogos de datos, permite a los usuarios explorar y previsualizar los datos a través de mapas, gráficos o tablas.

Permite publicar y alojar datos en la misma plataforma web de la institución. Cuenta con un buscador de palabras claves con filtros para metadatos. Contiene la posibilidad de incluir metadatos por defecto o de personalizar los metadatos. Permite compartir los conjuntos de datos a través de redes sociales, generar suscripciones a estos y tiene herramientas para recoger el feedback de la comunidad. Los creadores de CKAN también ofrecen servicios de desarrollo y soporte del repositorio (de pago).

Repositorios gestionados por CKAN: <http://datos.gob.cl/> <https://data.gov.au/> <https://www.dataplatform.nl/>

* Junar <http://www.junar.com/>

Plataforma para publicación de datos abiertos en la nube. Permite compartir datos abiertamente o definir como de uso interno. Permite visualizaciones en tablas o mapas. Entrega permisos para diferentes categorías de usuarios en la plataforma. Es de pago.

* DKAN <https://getdkan.org/>

Es un gemelo de CKAN, pero de propiedad del gestor de contenidos Drupal, permite previsualizar, descargar datos estadísticos básicos de los conjuntos de datos.

Repositorios gestionados por DKAN: <https://data.ca.gov/> <https://healthdata.gov/>

A continuación, la tabla 7 resume las ventajas y desventajas de las plataformas y software presentados anteriormente:

***Tabla Nº7: Ventajas y desventajas de plataformas y software***

|  |  |  |  |
| --- | --- | --- | --- |
| TIPO (PLATAFORMA/ SOFTWARE) | NOMBRE | VENTAJAS | DESVENTAJAS |
| Plataforma y Software | DataVerse | Compatible con *Open Science*, Permite administración, entrega diversos tipos de usuarios, permite linkear datos con sitio web, tiene buscador avanzado, permite descargar datos, genera nº de identificación única, licencia CC0 y cita, solicita y muestra a usuarios gran cantidad de metadatos | Tiene restricción de GB para subida de *datasets* |
| Plataforma/ Software | Figshare | Compatible con *Open Science*, Permite gestión completa de los datos, entrega nº de identificación único, licencia CC0 y cita. Permite integración y difusión con redes sociales y gestores de referencias, permite agregar más contenido de estudios previos o análisis. | Tiene restricción de GB para subida de *datasets.* Sólo si se suben en formato abierto es ilimitado. Solicita y muestra muy pocos metadatos a los usuarios. |
| Plataforma | OSF | Compatible con *Open Science*. Permite la gestión de proyectos de investigación. Entrega DOI y dispone de licencias de uso de datos. Permite compartir en redes sociales y enlaza con otros repositorios de datos. | Sólo 5 GB de espacio disponible. Pero permite solicitar más. |
| Plataforma | Dryad | Compatible con *Open Science*. Permite gestión de datos. Asigna DOI basado en el artículo. Entrega licencia CC0. Se vincula a artículo académico y sus archivos. | Sólo admite *datasets* que se relacionen a un artículo a publicar o publicado en alguna revista asociada a Dryad. Solicita escasos metadatos. Sólo 10GB. |
| Plataforma | Zénodo | Compatible con *Open Science*. Permite la gestión de datos. Verifica la relación de los *datasets* con publicaciones científicas o investigaciones. Permite crear comunidades de usuarios. Se puede crear cuenta a partir de GitHub | Sólo entrega 2GB disponibles para repositorio. No muestra al usuario los metadatos previos a la descarga de material. |
| Plataforma | DataHub | Compatible con *Open Science*. Permite gestión de datos. Permite previsualización de los datos cuando están en formato .csv. Permite alojar datos que se encuentran también en otros servidores. | Es de pago. |
| Software | CKAN | Compatible con *Open Science*. Permite administración completa de la gestión de datos. Permite publicar y alojar datos en el sitio web de la institución. Entrega metadatos por defecto o personalizados. Permite linkear con redes sociales y generar suscripciones a los datos. Entrega herramientas para recoger feedback de la comunidad. | Implica desarrollar el repositorio y aprender a utilizarlo. No entrega licencias ni nº de identificación propio (DOI). Está en lenguaje Python lo que lo hace más técnico y complejo de utilizar para no versados en la tématica. |
| Plataforma | Junar | Compatible con *Open Science*. Permite publicar con alojamiento en la nube. Permite compartir datos como abiertos o como de uso interno. Permite visualizaciones de tablas o mapas. | Es de pago. No se sabe si entrega licencias ni nº de identificación. Se desconocen los GB disponibles. |
| Software | DKAN | Mismas ventajas de CKAN. Se encuentra en lenguaje PHP. El software tiene licencia GNU GPL. | Implica desarrollar el repositorio y aprender a utilizarlo. No entrega licencias a bases de datos ni nº de identificación propio (DOI) |

En resumen, a nivel general las ventajas que entregan las plataformas de datos implican la administración, gestión y curación de los datos en varios casos gratuito, además de la definición de un DOI o nº de registro único para el conjunto de datos publicado, una licencia en formato *Open Data*, la conexión con investigaciones u otras publicaciones relacionadas al *dataset*, entre otros. Por otro lado, las desventajas guardan relación con la cuota de espacio que se dispone para el repositorio, en algunos casos aumenta si se paga por el uso de la plataforma, pero en su mayoría termina siendo limitado, y con las restricciones de uso en función de las definiciones de cada plataforma, por ejemplo: las licencias de dominio público.

En este sentido, y en función de las necesidades defina COES para la “Plataforma Integrada de Datos”, sería bueno preguntarse si el uso de las plataformas logra cubrir todas esas necesidades o tal vez, sería propicio utilizar un software como CKAN o DKAN para crear un repositorio de datos propio que no se encontraría regido por las condiciones de uso de una plataforma ni tampoco limitado por una cuota en GB o MB de archivo de datos, y que además dicen entregar la ventaja de ser de fácil uso y administración tanto para personas versadas en el lenguaje de programación como para personas no tan cercanas a éste.

# Conclusiones para la discusión

Lo primero a destacar es que hablar de *Open Data* implica referirse no solo al momento de disponer los datos hacia el público, sino que requiere brindar las condiciones para que ese público pueda acceder y utilizar correctamente los datos proporcionados, eliminando restricciones de uso y barreras que podrían generar discriminación.

En este sentido, cobran relevancia los formatos de los *datasets*, priorizando aquellos que sean libres y no propietarios, los metadatos disponibles que facilitan la comprensión de los *datasets* y que dada su estructura permiten la lectura por personas y redes informáticas/digitales (humanos y no humanos), la codificación y los documentos anexos que entregan las facilidades para la utilización de los datos y permiten descifrarlos. Asimismo, las condiciones de *copyright* como las licencias de uso y la forma en que los datos se ponen a disposición del público, es decir los repositorios, también juegan un papel primordial. Mientras más claros, ordenados y con condiciones de uso y redistribución definidas, y en un ambiente también claro, ordenado y de fácil acceso, sean puestos a disposición los *datasets*, más cercano a los estándares internacionales y, por tanto, más en la línea del *Open Data.*

En Chile existen estándares sobre preservación e intercambio de datos que ha definido el Estado en la lógica de un “gobierno abierto” y que sientan las bases para las instituciones estatales sobre este tema. Estos estándares se rigen principalmente por las recomendaciones de los organismos internacionales pioneros y líderes en el *Open Science*, y por tanto son un buen punto de partida para COES.

En términos del análisis realizado a las bases de datos COES, se puede indicar:

* La plataforma Harvard DataVerse contiene ventajas y desventajas. Por un lado, entrega la posibilidad de compartir las bases de datos, solicitando y mostrando al público los metadatos más comunes y además agregando metadatos que permiten comprender mejor los *datasets*, dispone de un DOI o un número de registro único y permite adjuntar algunos documentos relevantes. Por otro lado, limita las licencias de uso disponibles para los datos, limita el idioma de búsqueda y de acceso a los datos, no permite bajar las bases de datos del sitio en caso de que queden en desuso, no incorpora la posibilidad de generar un link con el sitio web de COES o del estudio, lo que limita la visibilidad de las bases de datos y, tiene una cuota fija de espacio en la web, entre otros.
* Las bases de datos COES, sus manuales y libro de códigos, en general, presentan altos estándares y entregan información suficiente para facilitar la accesibilidad y utilización de la base, siendo la que requiere mayores mejoras la base de datos del observatorio de conflictos.
* Todas las bases de datos requieren ordenar su documentación anexa de acuerdo con la política *Open Science.* Considerando que la facilidad de acceso también considera el tiempo requerido para descifrar y entender la base de datos, el ordenamiento adecuado es un aporte relevante. De esta manera manuales de usuario complejos y extensos en información podrían producir dificultades para la accesibilidad.
* La información de contacto del investigador principal también es una pieza fundamental. Cualquier duda de los metadatos o de los documentos anexos podrían ser respondidos por el investigador, pero además en términos académicos la línea de pensamiento teórica, empírica o ideológica, incluso, del investigador es un dato relevante para la comprensión de la recogida de datos, por ejemplo, lo cual juega un rol preponderante en el posterior re-análisis de datos que podría realizar quien acceda a estos.

De lo anterior, se desprende la necesidad de que COES defina estándares propios y garantice que sus bases de datos se encuentren alineadas con ellos. Por ejemplo, las bases ELSOC y ENACOES tienen la misma unidad de análisis (personas), pero la base del Observatorio de Conflictos tiene una unidad de análisis diferente (medios de prensa) de modo tal que los marcos muestrales son distintos. En base a esto, cabe preguntarse si ¿es necesario que los metadatos sean los mismos para todas las bases de datos o deberán definirse algunos obligatorios y otros según sea la necesidad? ¿Será mejor definir los estándares propios de metadatos o regirse por los metadatos solicitados por una plataforma/repositorio de datos en particular? Cualquiera sea la decisión, los requisitos para considerar o no un metadato deben quedar definidos y claros (fácil comprensión para todo público) a priori.

Respecto de las plataformas para compartir y preservar datos, como ya se mencionó, si bien entregan ciertas oportunidades y facilidades, pareciera ser que lo más conveniente para COES es apuntar al desarrollo de una plataforma propia, no solo porque permitiría sortear las limitaciones de las plataformas actuales (limitación en usos de licencias, en eliminación de bases obsoletas, en cuota de espacio disponible, en idiomas, en conexión con otras investigaciones o estudios sobre el tema, etc.) sino que porque tener un repositorio propio de bases de datos permitiría posicionar al Centro dentro de los pioneros en el país en la apertura de datos, daría más visibilidad de las investigaciones que realiza y potenciaría el impacto de COES en las políticas públicas de la materia.

Este repositorio debería contar con ciertos aspectos técnicos que garantizan la aplicación del *Open Data*, entre ellos se sugieren: panel de búsqueda y búsqueda avanzada, posibilidad de retroalimentación de los usuarios, panel de ayuda, link directo a contacto con investigadores principales según *datasets,* link a documentos anexos, estudios o investigaciones vinculados a los datos, medios estadísticos para evaluar descargar o visualización de los *datasets*, listado de bases de datos con sus versiones disponibles, posibilidad de previsualizar los datos, metadatos disponibles, facilidad de identificar la licencia o los derechos de autor asociados a los *datasets* (incorporando, incluso, una preaceptación de los requisitos generales de la licencia previo a la descarga), DOI o número de registro único, seguridad y encriptación de los datos (especialmente para *datasets* que podrían contener información sensible), opciones para personas con discapacidad (considerando la política de no discriminación del *Open Data*), entre otros.

Además, una estructura interna que permita la curación de los datos: facilidad de administración a través de carpetas, subida y descarga de *datasets* con facilidad, tiempo de mantención de los *datasets* en el repositorio, espacio disponible y velocidad del servidor para sostener la cantidad de datos y sus actualizaciones constantes, etc. Para ello, se sugiere la utilización del software DKAN, el cual a través de un lenguaje informático simplificado por medio de PHP, facilita el diseño, desarrollo y mantención web del repositorio, lo que implicaría que en primera instancia se requiera la contratación de un profesional versado en el desarrollo de sitios y repositorios digitales, pero para el proceso de mantención se podría capacitar a algunos investigadores o asistentes de investigación, o bien, se podrían bajar costos contratando solo a un profesional vinculado a la gestión de bibliotecas virtuales.

De cualquier forma, el proceso de preservación, mantención y reutilización de los datos, o en otras palabras las curatoría de los datos, debe estar estandarizada, con perfiles de acceso y responsabilidades definidas, solo ello garantiza que la política de “datos abiertos” u *Open Data* se aplique correctamente.

# Referencias

* Centro de Estudios de Conflicto y Cohesión Social (2015). Encuesta sobre el Conflicto y la Cohesión Social en Chile (Encuesta COES 2014 3.0) [Archivo de datos]. Santiago, Chile: Centro de Estudios de Conflicto y Cohesión Social (COES).
* Centro de Estudios de Conflicto y Cohesión Social (2018). Estudio Longitudinal Social de Chile, Primera Ola (ELSOC\_W01\_v2.00) [Archivo de datos]. Santiago, Chile: Centro de Estudios de Conflicto y Cohesión Social (COES).
* Centro de Estudios de Conflicto y Cohesión Social (2018). Manual de Usuario de Estudio Longitudinal Social de Chile, Primera Ola (2016). Corte Transversal. Versión 2.0. Santiago, Chile: Centro de Estudios de Conflicto y Cohesión Social (COES).
* Centro de Estudios de Conflicto y Cohesión Social (2018). Observatorio de Conflictos, Acciones de Protesta 2012-2017. [Archivo de datos]. Santiago, Chile: Centro de Estudios de Conflicto y Cohesión Social (COES).
* Data Documentation Initiative (DDI) Alliance (2018) Explore Documentation. Recuperado de: <http://www.ddialliance.org/explore-documentation>
* Digital Curation Centrer (DCC) (2004-2019) Digital Curation Standars. Recuperado de: <http://www.dcc.ac.uk/resources/standards>
* González, L.; Saorín, T.; Ferrer-Sapena, A.; Aleixandre-Benavent, R.; Peset, F. (2013). “Gestión de datos de investigación: infraestructuras para su difusión”. *El profesional de la información*, septiembre octubre, v. 22, n. 5, pp. 415-423. <http://dx.doi.org/10.3145/epi.2013.sep.06>
* Hernández-Pérez, T.; García-Moreno, M. A. (2013). “Datos abiertos y repositorios de datos: nuevo reto para los bibliotecarios”. *El profesional de la información*, 2013, mayo-junio, v. 22, n. 3, pp. 259-263. <http://dx.doi.org/10.3145/epi.2013.may.10>
* Ley 19.628. Diario Oficial República de Chile, Santiago, Chile. 20 de agosto de 1999
* Ley 21.096. Diario Oficial República de Chile, Santiago, Chile. 16 de junio de 2018
* Ministerio Secretaría General de la Presidencia (2013) Norma Técnica Chilena para la publicación de datos abiertos. Versión 2.1. Unidad de Modernización y Gobierno Digital.
* Nina-Alcocer, V.; Blasco-Gil, Y.; Peset, F. (2013). “*Datasharing*: guía práctica para compartir datos de investigación”. *El profesional de la información*, noviembre-diciembre, v. 22, n. 6, pp. 562-568. http://dx.doi.org/10.3145/epi.2013.nov.09
* Open Data Institute (ODI). The Data Ethics Canvas. Recuperado de: <https://theodi.org/article/data-ethics-canvas/>
* Polar Data Catalogue (PDC) (2014). Best Practices for Sharing and Archiving Datasets. Canadá. Recuperado de: <https://www.polardata.ca/pdcinput/public/PDC_Best_Practices_FULL.pdf>
* Rusbridge, C., et al (2005) “The digital curation centre: a vision for digital curation”. En: *Proceedings from local to global: data interoperability - challenges and technologies*. Forte Village Resort, Sardinia, Italy, pp. 1-11. http://eprints.erpanet.org/82
* Statistical Data and Metadata Exchange (SDMX) (2018) Whats is SDMX. Recuperado de: <https://sdmx.org/?page_id=3425>
* Torres-Salinas, D (2010). “Compartir datos (*data sharing*) en ciencia: contexto de una oportunidad”. *Anuario ThinkEPI*, v. 4, pp. 258-261
* Torres-Salinas, D.; Robinson-García, N.; Cabezas-Clavijo, A. (2012) “Compartir los datos de investigación: introducción al *data sharing*”. *El profesional de la información*. Marzo-abril, v. 21, n. 2, pp. 173-184. <http://dx.doi.org/10.3145/epi.2012.mar.08>

**Páginas web**

* Portal de datos públicos. Recuperado de: <http://datos.gob.cl/>
* Australian Goverment Data. Recuperado de: <https://data.gov.au/>
* DataPlatform. Recuperado de: <https://www.dataplatform.nl/>
* California Open Data Portal. Recuperado de: <https://data.ca.gov/>
* Health data.gov. Recuperado de: <https://healthdata.gov/>
* World Wide Web Consortium (W3C) Recuperado de: <https://www.w3.org/>
* Creative Commons. Recuperado de: <https://creativecommons.org/>
* Open Data Commons. Recuperado de: <https://opendatacommons.org/licenses/index.html>
* Sistema GNU. Recuperado de: <https://www.gnu.org/home.en.html>
* Centro de Estudios de Conflicto y Cohesión Social. Recuperado de: www.coes.cl

**Plataformas y Softwares:**

* Dataverse Proyect: <https://dataverse.org/institutions>; <https://dataverse.harvard.edu/>
* Figshare: <https://figshare.com/>; <https://figshare.com/services/institutions>
* Open Science Framework (OSF): <https://osf.io/>
* Dryad: <https://datadryad.org/>
* Zenodo: <https://zenodo.org/>
* DataHub <https://datahub.io/>
* Open ICPSR <https://www.openicpsr.org/openicpsr/>
* Re3data <https://www.re3data.org/>
* DataSearch Google <https://toolbox.google.com/datasetsearch>
* CKAN <https://ckan.org/>
* Junar <http://www.junar.com/>
* DKAN <https://getdkan.org/>

1. El conjunto de actividades que implica la preservación y reutilización de datos. Incluye aspectos técnicos, de organización, de procedimientos, legales, de propiedad y autoría, restricciones de uso, aspectos financieros, etc. [↑](#footnote-ref-1)
2. Data Catalog Vocabulary (DCAT) del World Wide Web Consortium (W3C) Ver en https://www.w3.org/ [↑](#footnote-ref-2)
3. Datos que se refieren a las características físicas o morales de las personas o a hechos o circunstancias de su vida privada o intimidad. Ej: estados de salud, vida sexual, ideologías u opiniones políticas, etc. [↑](#footnote-ref-3)
4. Basado en el estándar OCDE del Statistical Data and Metadata Exchange. Ver en https://sdmx.org/ [↑](#footnote-ref-4)
5. Dado el uso actual del Software [↑](#footnote-ref-5)
6. Dado el uso actual del Software [↑](#footnote-ref-6)
7. Más información en: https://theodi.org/article/data-ethics-canvas/ [↑](#footnote-ref-7)
8. Más información en: https://creativecommons.org/ [↑](#footnote-ref-8)
9. Más información en: https://opendatacommons.org/licenses/index.html [↑](#footnote-ref-9)
10. Más información en: https://www.gnu.org/home.en.html [↑](#footnote-ref-10)
11. En el apartado 2.2 “Evaluación Manual de Usuario” se analizan los metadatos disponibles. [↑](#footnote-ref-11)
12. Metadatos solicitados por Norma Técnica Chilena (NTCh) (2013), considera los obligatorios y algunos recomendados. [↑](#footnote-ref-12)
13. Se consideraron en el análisis los metadatos de DataVerse que no consideraba la Norma Técnica Chilena (2013), a modo de no repetir información. [↑](#footnote-ref-13)
14. Se espera que se indiquen, al menos, variables sociodemográficas de quiénes serán parte de la muestra. [↑](#footnote-ref-14)
15. De acuerdo con los estándares, un libro de códigos debe entregar un resumen de los metadatos y categorías que permiten el análisis de los datos, por tanto, se repiten algunas categorías de estándares o requisitos evaluados en los puntos, y se indica si están disponibles en otro lugar, en caso de no ser parte del libro de códigos. [↑](#footnote-ref-15)