INCLIVA | VLC
Instituto de Investigación Sanitaria

Fecha: 20/03/2025

Edición: 1.0

GESTIÓN DE DATOS DE INVESTIGACIÓN

METADATOS

GUÍA PARA LA ELABORACIÓN DE UN README

ÍNDICE

1. OBJETO

Esta guía tiene como objeto explicar la utilidad del archivo README, que debe incluirse dentro de la carpeta de los datos de investigaciones científicas. A su vez se van a desarrollar las pautas y buenas prácticas que se deben seguir para crear estos ficheros.

2. ALCANCE

Esta guía se circunscribe en torno al fichero README, pudiendo cubrir otras áreas relacionadas con la ciencia abierta o la presentación de datos, pero sin extenderse.

3. DEFINICIONES

.txt: formato de fichero que guarda la información en texto plano. Muy utilizado debido a su reducida dimensión y las diversas aplicaciones para abrir el formato. Este formato puede ser abierto con el bloc de notas o notepad.

.md: formato de fichero con etiquetas. Es utilizado en el campo de la informática para crear README e informes, permite dar formato al texto sin ser excesivamente complicado. Este formato puede ser abierto con el bloc de notas o Notepad, aunque para ver el texto con formato se debe abrir con programas específicos como Visual Studio Code o Ghostwriter (para Linux).

Ciencia abierta: enfoque de investigación y divulgación científica que promueve la accesibilidad, transparencia y colaboración en todas las etapas del proceso científico. Entre sus principios se encuentran los datos abiertos.

Datos abiertos: datos que se encuentran accesibles al público por defecto. Entre los requisitos se debe asegurar la interoperabilidad, calidad y disponibilidad de estos datos.

Metadatos: datos que acompañan a los datos, sirven para explicar los propios datos y las circunstancias que los rodean.

Conjunto de datasets: uno o más datasets que contienen datos sobre un mismo tema.

Dataset: colección organizada de datos que puede contener diferentes tipos de datos, como números, texto, imágenes o videos, y está estructurado de manera que facilite su acceso y procesamiento.

4. INTRODUCCIÓN

4.1 Los datos abiertos y la utilidad de los metadatos

En primer lugar, los datos sólo tienen valor si se utilizan y cuanta mayor difusión tengan mayor será su impacto. Es por ello que las instituciones de distintos niveles, desde la Comisión Europea hasta la propia fundación INCLIVA, cada vez se preocupan más por la transparencia y accesibilidad de sus datos. En consecuencia, existe extensa legislación y recomendaciones para conseguir que se publiquen los datos que obren en su poder o hayan sido producidos en su entorno. Dentro de este entorno se encuentran proyectos financiados con fondos públicos.

Este concepto no es ajeno al campo de la investigación, puesto que la ciencia abierta consiste en compartir los resultados de las investigaciones publicando *papers* en distintas revistas y medios científicos. La comunidad científica divulga los descubrimientos para demostrar sus avances y, así mismo, se nutre de ellos para sus propias investigaciones. Entonces no nos debe resultar tan extraño que se pida compartir los datos como uno de los productos resultantes de la investigación científica.

Para que estos datos puedan ser fácilmente interpretables se suelen acompañar de metadatos que aportan información sobre los datos. Aunque en los siguientes apartados se entrará en más detalle, los metadatos contienen información general que describen el conjunto del dataset e información sobre la estructura y el contenido de los datos.

Los metadatos son de especialidad utilidad cuando existen una gran cantidad de datos o éstos son creados por procesos automáticos, teniendo un documento anexo que te ayuda a interpretarlos. Los metadatos también ayudan al propio propietario de los datos por lo que es recomendable crear los metadatos al mismo tiempo que se generan los datos.

4.2 Metadatos como README

Existen distintas formas de generar los metadatos, dependiendo del tipo de datos o tecnología empleada, cada uno con sus ventajas y desventajas. Desde la organización se ha decidido apostar por un fichero README para actuar como metadatos. Entre las ventajas de este sistema se encuentra:

- sencillez para elaborar los metadatos.
- facilidad para leer el contenido de los metadatos.
- universalidad del formato empleado, pudiendo abrirse con aplicaciones sin importar el sistema operativo.
- inclusión del archivo junto a los datos, formando un paquete en el que datos y metadatos se pueden distribuir juntos.

Por ello, a la hora de distribuir o subir los datasets a un repositorio se debe incluir el archivo README pudiendo estar en el mismo paquete o como carpeta comprimida.

5. DESARROLLO

5.1 Estructura del README

Pasemos a detallar que elementos debe contener el fichero README, junto con los ejemplos de cómo incluirlos. Esta estructura recoge la información mínima que debe estar presente en los metadatos, pero puede modificarse según el caso concreto. Entre apartados se recomienda dejar espacio. El idioma del README podrá ser castellano o inglés, dependiendo del público objetivo, aunque es preferible el inglés.

5.1.1 Cabecera

Contendrá el título del conjunto del dataset y su descripción. En caso de que la subida se componga de más de un dataset deberá hacer referencia a todos ellos.

DatasetTitle: [Título Dataset, Ej. Calles de Valencia].
Description: [Breve resumen del contenido, Ej: dataset with the adress and postal code of the streets of Valencia]

5.1.2 Organización e investigadores

Al ser la investigación apoyada por INCLIVA se incluirá nombre, dirección e información de contacto como aparece en el ejemplo. A continuación, se detallarán los miembros del equipo, incluyendo nombre, ID de investigador, teléfono móvil (es opcional) y correo electrónico institucional. Por último, se debería incluir una persona de contacto (puede ser integrante del grupo de investigación) para recibir consultas de los datos.

Organization: FUNDACION INCLIVA, Avd. Menendez y Pelayo, 4 acc., 961973517, contacto@incliva.es
Principal Investigator: John Smith, [id(ORCID: , ISNI: ; GND:)],[móvil(opcional)], jsmith@incliva.es
Associate 1: John Smith, [id(ORCID: , ISNI: ; GND:)], [móvil(opcional)], jsmith@incliva.es
Associate 2: John Smith, [id(ORCID: , ISNI: ; GND:)], [móvil(opcional)], jsmith@incliva.es
(opcional)Contact person: John Smith, [id(ORCID: , ISNI: ; GND:)], [móvil(opcional)], jsmith@incliva.es

5.1.3 Información general

Se deberá incluir las palabras clave separadas por una coma, dependiendo del tipo de investigación se incluirán temáticas distintas. Después se incluirá el artículo o revista donde se publica el artículo (si aún no está publicado o no va a publicarse se puede dejar en blanco). A continuación, se incluirá el idioma en el que estén los datos, si tiene varios se deberán incluir. Para marcar el espacio temporal se indicará cuando fue creado el dataset y que cobertura temporal abarcan los datos. Las fechas deberán estar en formato ISO: año, mes y día separados por guiones. Además de la temporal también se añadirá la cobertura geográfica que debería estar representado como coordenadas, pero puede representarse con texto. Por último, se enumerarán los fondos púbicos recibidos para la investigación, incluyendo premios u otros conceptos si se ha recibido una cantidad económica.

Keywords: [Etiquetas]
Publication: [Estudio/Publicación de donde vienen los datos]
Language: [Idioma del dataset, si tiene varios ponerlos todos. Ej: ESP, ENG]
Created Date: YYYY-mm-dd
Date coverage: YYYY-mm-dd [Covertura temporal,(si es un rango incluir inicio y fin)]
Spatial coverage: [Lugar de extracción de datos, Ej. Valencia, Spain]
Grants: [Subvenciones/Financiacion del estudio]

5.1.4 Datasets incluidos

Un conjunto de datasets puede incluir uno o varios datasets, si se incluyeran varios se tendría que repetir los mismos campos para cada uno de ellos. Por cada dataset se incluirá su nombre, formato, descripción del dataset y los atributos del dataset. Entre los formatos más habituales se encuentran: csv, json, sql, yaml. También se deberá explicar la manera en la que se han obtenido estos datos o una breve explicación del ensayo a partir del que se han obtenido los datos.

Las variables o atributos son algo más complejo. Por lo general, para los atributos se incluirá su nombre, descripción y el tipo de datos. Los distintos tipos de datos son:

Tipo	Descripción	Ejemplo
Texto	El tipo más extenso, usualmente mientras haya una letra será texto.	Hola, año 1992, x = a+b, 54€
Entero	Número sin decimal, no puede incluir letras o caracteres especiales.	1, 56, 336273, 41247354
Decimal	Número con decimales, no puede incluir letras o caracteres especiales.	1.0, 1.74395, 0.32, 2679.478
Boolean	Solo dos estados, usualmente Verdadero y Falso ó 0 y 1.	V, T, F, 0, 1
Fecha	Representa una fecha, puede tener múltiples formatos, no puede incluir letras o caracteres especiales fuera de los aceptados.	13/12/1994, 12/13/1994, 1994-12-13

Files:
(por cada archivo)
Name: [Nombre completo del archivo, Ej: Calles.json]
SourceType: [el tipo de recogida de los datos, Elegir entre: Observed, Experiment, Computational]
DataSource: [Explicar como se han obtenido, Ej: Data collected from clinical trials. Test subjects have given consent to use their data.]
FileFormat: [Formato del archivo, Ej: JSON]
Description: [Breve resumen del contenido, Ej: dataset with the adress and postal code of the streets of Valencia]
Variables: [variable,descripcion,tipo;variable,descripcion,tipo, Ej: direccion,nombreConLaCalleDeValenciatext;CP,CodigoPostalDondeSeEncuentraUbicadaLaCalle,number]

En muchos casos existirá información adicional que debamos explicar sobre esa variable. Para datos numéricos quizá sea preciso especificar la unidad de medida (se añade un campo más) o es necesario conocer información que rodea a esa variable como si la medida se ha tomado antes o después de la ingesta de medicamentos (se redacta en la descripción). Así pues, la explicación de las variables se puede hacer todo lo simple o complejo que queramos, pero cuanta más información aportemos mejor se podrá entender esa variable.

5.1.5 Licencia y tratamiento de los datos

En este último apartado se incluye el tipo de licencia con la que se cede los datos, el enlace a la licencia y los procesos que se han llevado para captar y tratar estos datos. La licencia que se recomienda es la CC 4.0 en la que se debe citar al autor e indicar los cambios efectuados, en caso de querer otra especifica se puede indicar en este apartado. A continuación, se incluye el link a la política de la organización que regula cómo se captan los datos y cómo se procesan, pudiendo añadir información adicional sobre el tratamiento de datos. Esta información adicional será sobre tratamientos masivos o estándares que se han utilizado en todos los datasets, pudiendo ser la política de tratamiento de datos nulos, cómo se nombran los identificadores o que los datos numéricos se redondean con tres decimales.

License: Creative Commons 4.0 International (cambiar si se desea, este es abierto con obligacion de citar y mencionar cambios)
LicenseLink: https://creativecommons.org/licenses/by/4.0/legalcode
How To Cite: [Poner cita del dataset]
DataCollection: link
DataTransformation: link
OtherTreatment: [si se ha hecho tratamiento especial]

5.2 FORMATO DEL README

Una vez vista la importancia y la estructura que debe tener el archivo README, tan sólo queda por concretar el formato. Para seguir los principios expuestos en el punto 4.2, el fichero puede ser .txt o .md (Markdown).

El formato txt es más simple puesto que sólo guarda su contenido en texto plano. La mayoría de los usuarios están habituados a emplear este fichero y no se necesita de ningún conocimiento para escribir el texto. Sin embargo, no se le puede dar forma al texto.

El formato Markdown permite dar forma al texto, no es complicado de utilizar y con la suficiente experiencia puede generar textos al mismo nivel que un procesador de textos. Requiere de conocimientos básicos sobre el funcionamiento de las etiquetas y una aplicación que interprete el formato si se quiere ver la estructura del texto.

Si está más cómodo con txt puede hacer su README en este formato, pero si tiene experiencia previa con Markdown o quiere intentar esta otra vía quedará más presentable.

A continuación, se incluye como anexo un tutorial con los aspectos básicos del formato Markdown.

6. ANEXO – TUTORIAL FORMATO MARKDOWN

Anexo 1 - CONCEPTOS BÁSICOS DE MARKDOWN

Markdown es un lenguaje sencillo de utilizar que sigue los mismos principios que HTML. En base a símbolos y espacios es posible darle formato al texto, pudiendo crear títulos, apartados, índices y otros elementos. Vamos a ver los elementos básicos.

Títulos

Para añadir un título, utiliza el símbolo "#" al inicio del texto y deja un espacio entre el símbolo y el texto. Cuantos más # añades menor será el nivel del título.

Título 1

Título 2

Título 3

Salto de línea

Para añadir un salto de línea, además de darle al enter añadir dos espacios.

Arriba__

Abajo

Dentro del texto se puede usar la etiqueta <br> para hacer un salto.

Arriba
 Abajo -> Arriba

Abajo

Cursiva y Negrita

Para resaltar un texto coloca * ó _ entre las palabras para ponerlas en cursiva y ** ó _ _ para ponerlas en negrita.

cursiva 1 _cursiva 2_

negrita 1 __negrita 2__

Listas

Para crear una lista utiliza - , * ó + para cada elemento de la lista, puedes generar subniveles al identar el texto. También puedes usar números para realizar listas ordenadas.

Nota: recuerda dejar un espacio entre el símbolo y el texto

Lista

- Elemento 1
  - Elemento 1.1
    - Elemento 1.1
  - Elemento 1.2
- Elemento 2

Lista ordenada

1. Elemento 1
   1. Elemento 1.1
      1. Elemento 1.1.1
   2. Elemento 1.2
2.Elemento 2

Links

En Markdown también se puede añadir enlaces, para ello entre [] se pone el texto para enlazar y a continuación el link entre ().

[Google](https://www.google.com)

Avanzado: Con el mismo método se puede hacer un índice o hacer referencia a otro apartado del documento, tan solo se sustituye el link por el título al que se hace referencia.

[Título 1](# Título 1)

imagen

Muy similar al anterior, se añade una ! al comienzo. Después se añade un nombre para la imagen [] y el enlace a la imagen.

![Imagen 1](https://example.com/image.jpg)

Línea horizontal

Para separar el texto en partes se pueden añadir tres - _ * para crear una línea horizontal.

___

***

Tablas

Por último, para crear una tabla se marcan las celdas con | y entre el encabezado y los datos se añade una fila con ---.

| Columna1 | Columna 2 |

| --- | --- |

| celda1 | celda2 |

| celda3 | celda4 |

Con esto concluye la explicación de los elementos más básicos. Puede resultar complejo al inicio, pero siguiendo estas pautas se puede hacer la mayoría de formatos necesarios. Puedes observar un ejemplo en https://github.com/pytorch/pytorch/blob/main/README.md o a continuación.

Anexo 2 - EJEMPLO DE README UTILIZANDO MARKDOWN

 # README

DatasetTitle

[Título Dataset, Ej. Calles de Valencia]

Description

[Breve resumen del contenido, Ej: dataset with the address and postal code of the streets of Valencia]

Organization

****Name:**** FUNDACIÓN PARA LA INVESTIGACIÓN DEL HOSPITAL CLÍNICO DE LA COMUNIDAD VALENCIANA (INCLIVA)

****Address:**** Avd. Menendez y Pelayo, 4 acc.

****Phone:**** 961973517

****Email:**** incliva@incliva.es

Researchers

Principal Investigator

****Name:**** John Smith

****ORCID, ISNI, GND:**** 11111

****(optional)Phone:**** [móvil(opcional)]

****Email:**** jsmith@incliva.es

Associate 1

****Name:**** John Smith

****ORCID, ISNI, GND:**** 11111

****(optional)Phone:**** [móvil(opcional)]

****Email:**** jsmith@incliva.es

Associate 2

****Name:**** John Smith

****ORCID, ISNI, GND:**** 11111

****(optional)Phone:**** [móvil(opcional)]

****Email:**** jsmith@incliva.es

(opcional**)Contact Person**

****Name:**** John Smith

****(optional)Phone:**** [móvil(opcional)]

Email: jsmith@incliva.es

General Information

Keywords

[Etiquetas]

Language

[Idioma del dataset, si tiene varios ponerlos todos. Ej: ESP, ENG]

Created Date

YYYY-mm-dd

Date coverage

YYYY-mm-dd [Covertura temporal,(si es un rango incluir inicio y fin)]

Spatial Coverage

[Lugar de extracción de datos, Ej. Valencia, Spain]

Publication

[Estudio/Publicación de donde vienen los datos]

Grants

[Subvenciones/Financiacion del estudio]

Files

File 1

Name: [Nombre completo del archivo, Ej: Calles.json]

FileFormat: [Formato del archivo, Ej: JSON]

SourceType: [el tipo de recogida de los datos, Elegir entre: Observed, Experiment, Computational]

DataSource: [Explicar como se han obtenido, Ej: Data collected from clinical trials. Test subjects have given consent to use their data.]

Description: [Breve resumen del contenido, Ej: dataset with the address and postal code of the streets of Valencia]

Variables: [variable,tipo;variable,tipo, Ej: direccion,text;CP,number]

File 2

Name: [Nombre completo del archivo, Ej: Calles.json]

FileFormat: [Formato del archivo, Ej: JSON]

Description: [Breve resumen del contenido, Ej: dataset with the address and postal code of the streets of Valencia]

Variables: [variable,tipo;variable,tipo, Ej: direccion,text;CP,number]

<!-- O alternativamente puedes emplear: -->

Variables

****Name:**** [Nombre de la variable]

****Variable type:**** [tipo de la variable, Ej: text, numeriacal, decimal]

****Description:**** [Breve resumen de la variable, se puede explicar su contenido]

****Unit measure:**** [Unidad de medida, Ej: meters, kilograms]

Data Information

License

****Type:**** Creative Commons 4.0 International (cambiar si se desea, este es abierto con obligacion de citar y mencionar cambios)

****Link:****
https://creativecommons.org/licenses/by/4.0/legalcode

How to cite

[Como citar correctamente el dataset]

Data Treatment

[Si se ha hecho tratamiento especial,

Ej: The following policy has been followed for this dataset (individual treatment may apply for each variable, see specific chapter above):

- Data anonimization with Openaire's application AMNESIA

- id has three incremental numbers, Ex: 001, 002, 003

- numerical values are rounded to four decimals, Ex: 25.0049

- symbol * is included when value is not checked afterwards

- Patients are from Valencia, Spain; ages between 30 to 55

- Patients are chosen based on random selection procedure consisting on …]

Anexo 3 - RESULTADO DEL README CON MARKDOWN

README

DatasetTitle

[Título Dataset, Ej. Calles de Valencia]

Description

[Breve resumen del contenido, Ej: dataset with the address and postal code of the streets of Valencia]

Organization

Name: FUNDACION INCLIVA
Address: Avd. Menendez y Pelayo, 4 acc.
Phone: 961973517
Email: contacto@incliva.es

Researchers

Principal Investigator

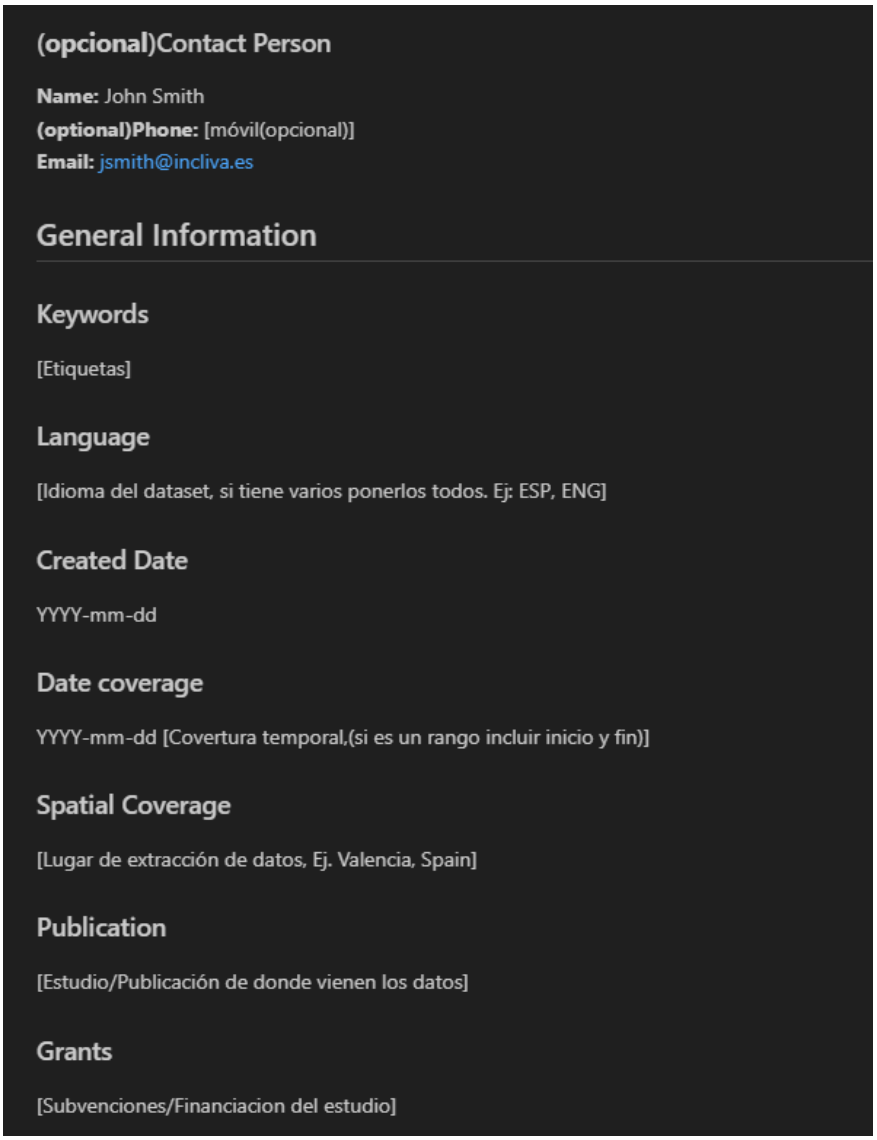
Name: John Smith
ORCID, ISNI, GND: 11111
(optional)Phone: [móvil(opcional)]
Email: jsmith@incliva.es

Associate 1

Name: John Smith
ORCID, ISNI, GND: 11111
(optional)Phone: [móvil(opcional)]
Email: jsmith@incliva.es

Associate 2

Name: John Smith
ORCID, ISNI, GND: 11111
(optional)Phone: [móvil(opcional)]
Email: jsmith@incliva.es



(opcional)Contact Person

Name: John Smith
(optional)Phone: [móvil(opcional)]
Email: jsmith@incliva.es

General Information

Keywords

[Etiquetas]

Language

[Idioma del dataset, si tiene varios ponerlos todos. Ej: ESP, ENG]

Created Date

YYYY-mm-dd

Date coverage

YYYY-mm-dd [Covertura temporal,(si es un rango incluir inicio y fin)]

Spatial Coverage

[Lugar de extracción de datos, Ej. Valencia, Spain]

Publication

[Estudio/Publicación de donde vienen los datos]

Grants

[Subvenciones/Financiacion del estudio]

[...]

