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]

GODAN ACTION
Global Open Data for Agriculture & Nutrition

UKaid
from the British people

Unidad 4: Compartir datos abiertos

Lección 4.2: Introduccion a la interoperabilidad de los datos

Objetivos y resultados del aprendizaje

El objetivo de esta lección es explicar los fundamentos de la interoperabilidad de datos.

Después de estudiar esta lección, deberías ser capaz de;

- comprender los fundamentos básicos de la interoperabilidad de datos
- identificar los diferentes tipos y niveles de interoperabilidad de datos.

Contenidos

Lista de figuras

1. Marco de referencia para los datos: De lo abierto a lo FAIR

La definición más utilizada de "interoperabilidad" en la web es: "la capacidad de un sistema o producto para trabajar con otros sistemas o productos sin un esfuerzo especial por parte del usuario". Wikipedia la define como "una característica de un producto o sistema, cuyas interfaces se entienden completamente, para trabajar con otros productos o sistemas, actuales o futuros, tanto en su implementación como en su acceso, sin ninguna restricción".

Cuando hablamos de interoperabilidad de datos, considerando que los datos se serializan en conjuntos de datos, las definiciones anteriores pueden aplicarse fácilmente a un dataset como un producto.

La interoperabilidad de datos es la posibilidad de que un conjunto de datos pueda trabajar con otros sistemas o conjuntos de datos sin un gran esfuerzo

La interoperabilidad de los datos es la capacidad de un dataset para trabajar con otros sistemas o conjuntos de datos sin un esfuerzo especial por ambas partes

En las actas de una conferencia organizada por la comunidad CIARD sobre la interoperabilidad de los datos para la agricultura, se definió la interoperabilidad de los datos como "una característica de los conjuntos de datos... por la que los datos pueden ser fácilmente recuperados, procesados, reutilizados y reempaquetados ("operados") por otros sistemas"[1].

Hay algunas definiciones que la concretan como la interoperabilidad "entre dos sistemas", pero la opinión generalizada es que algo es realmente interoperable (o más interoperable) cuantos más sistemas pueden interoperarlo.

Es más, veremos que utilizando determinados formatos de datos y aplicando ciertos estándares de datos, los datos pueden ser "interoperables por diseño" sin conocer, necesariamente, con qué sistema serán interoperables: la interoperabilidad planificada con sistemas específicos significa que los datos estarán "estrechamente acoplados" con esos sistemas, mientras que la interoperabilidad maximizada tiene como objetivo el acoplamiento con el mayor número posible de sistemas.

Sin embargo, nunca habrá algo así como una interoperabilidad universal o perfecta, es decir, una forma de exponer los datos que sea adecuada para todos los casos posibles. La interoperabilidad es siempre relativa a un sistema de normas compartidas y formas comunes de utilizar los datos, que en algunos casos son muy amplios y polivalentes (como Dublin Core o schema.org y un caso de uso genérico de los motores de búsqueda), y en otros casos son muy específicos de las comunidades científicas o de interés (como especificaciones de datos y visualizaciones de secuencias genéticas).

En realidad, las definiciones anteriores señalan a la interoperabilidad como una característica de los datos(conjuntos), lo cual es correcto porque son el objeto de la interoperación, pero el ecosistema de actores y productos que tienen que cooperar para lograr la completa interoperabilidad es más amplio: una definición interesante de la interoperabilidad que destaca la importancia de las "expectativas

[1] Interim Proceedings of International Expert Consultation on "Building the CIARD Framework for Data and Information Sharing", CIARD (2011) http://gfar.net/documents/interim-proceedings-international-expert-consultation-building-ciard-framework-data-and http://www.fao.org/docs/eims/upload/297074/IECProceedings-main-doc.pdf

compartidas" es la del Data Interoperability Standards Consortium (DISC): "La interoperabilidad de datos aborda la capacidad de los sistemas y servicios que crean, intercambian y consumen datos, de tener expectativas claras y compartidas sobre el contenido, el contexto y el significado de esos datos"[2].

2. Niveles de interoperabilidad

La interoperabilidad puede alcanzarse a diferentes niveles. Mientras que Wikipedia distingue entre interoperabilidad sintáctica e interoperabilidad semántica una distinción más granular se hace en la clasificación de los tipos de interoperabilidad de la Healthcare Information and Management Systems Society (HIMSS)[3]:

La interoperabilidad "Fundacional" permite que el intercambio de datos de un sistema informático sea recibido por otro y no requiere que el sistema informático receptor pueda interpretar los datos.

Este nivel se refiere principalmente a los protocolos de transmisión, que no analizaremos, ya que no es de interés para el gestor de datos, pero también abarca algunos protocolos de nivel superior, que en su mayoría funcionan sobre el protocolo HTTP común (por ejemplo, APIs REST especiales como OAI-PMH, SPARQL, Linked Data API o negociación de contenidos basada en solicitudes de tipo de contenido HTTP). Estas pueden ser de interés para los gestores de datos porque antes de ser leídos y comprendidos, los datos tienen que ser transmitidos, y, además de las descargas FTP, hay diferentes y más convenientes modos de hacerlo. Veremos este nivel de interoperabilidad en la lección 4.3.

La interoperabilidad "Estructural" es un nivel intermedio que define la estructura o formato del intercambio de datos (es decir, las normas de formato de los mensajes). La interoperabilidad estructural define la sintaxis del intercambio de datos. Esto garantiza que los intercambios de datos entre los sistemas de tecnología de la información puedan ser interpretados a nivel de los campos de datos.

Este es el nivel en el que los formatos de archivo y los formatos de datos desempeñan el papel más importante y es el nivel en el que los (meta)datos se convierten en "legibles por la máquina" y pueden ser analizados por las máquinas: cuanto más fácil sea para la máquina analizar el formato/sintaxis de los (meta)datos (XML, Json, CSV...), los datos son más interoperables estructuralmente. En la lección 4.3 analizaremos más a fondo este nivel de interoperabilidad.

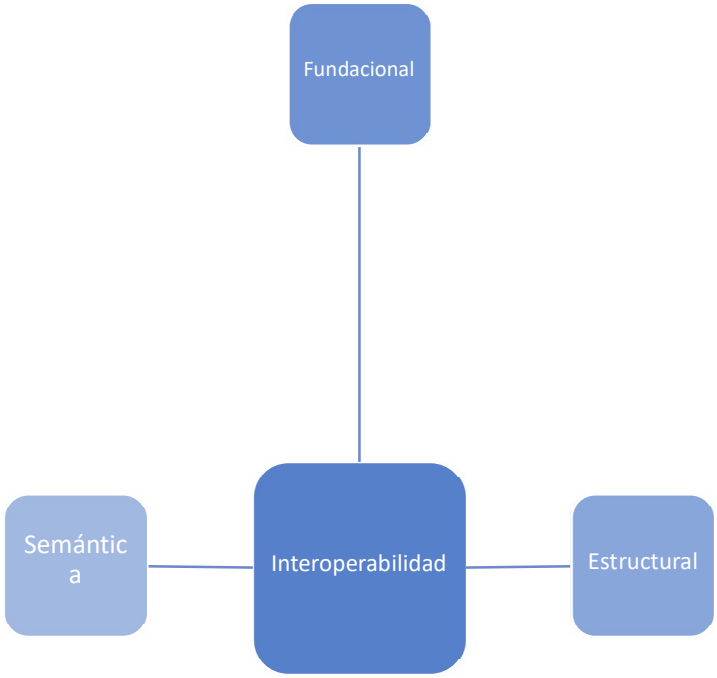
La interoperabilidad "Semántica" proporciona la interoperabilidad al más alto nivel, que es la capacidad de dos o más sistemas o elementos para intercambiar información y utilizar la información intercambiada. La interoperabilidad semántica aprovecha tanto la estructuración del intercambio de datos como la codificación de los datos, incluido el vocabulario, para que los sistemas informáticos receptores de la información puedan interpretar los datos.

Mientras que con la interoperabilidad estructural las máquinas entienden cuáles son los diferentes elementos (y su relación estructural recíproca), con la interoperabilidad semántica también entienden el significado de esos elementos y pueden procesarlos con herramientas que tengan en cuenta la semántica y realizar un razonamiento avanzado. Los formatos de datos no son suficientes para ello: las tecnologías semánticas (véase la lección 4.4) permiten incorporar elementos semánticos legibles por la máquina, de acuerdo, a vocabularios de (meta)datos serializados en la mayoría de los formatos de

[2] http://datainteroperability.org/

[3] http://www.himss.org/library/interoperability-standards/what-is-interoperability

datos existentes, aunque los formatos más adecuados hasta ahora son los distintos tipos de RDF (RDF/XML, Turtle, N-Triples) y JSON-LD. Veremos este nivel de interoperabilidad en la lección 4.4.



En la lección 4.3 veremos cómo se puede implementar la interoperabilidad estructural y arquitectónica. En la lección 4.4 veremos cómo puede implementarse la interoperabilidad semántica, con ejemplos más específicos en la lección 4.4.1.

Para obtener recomendaciones más detalladas sobre cómo implementar la interoperabilidad de datos, el documento del W3C "Data on the Web Best Practices"[4] es probablemente el mejor documento de referencia. Se basa en gran medida en el enfoque de "Linked Data" (véase la lección 4.3), pero muchas de las recomendaciones ayudan a implementar la interoperabilidad a diferentes niveles, aunque no se pretenda alcanzar el 100% de los datos enlazados.

3. Interoperabilidad de datos y metadatos

Los datos sin metadatos no pueden ser entendidos por las máquinas. Los datos normalmente vienen con metadatos.

La definición de datos en Wikipedia es 'los Datos son valores de variables cualitativas o cuantitativas, pertenecientes a un conjunto de elementos'. Como tal, al final los datos son siempre parte de una colección (un conjunto de elementos) y el elemento individual (fila registro) del conjunto. Los datos son los valores de algunas variables, por lo que siempre estarán encapsulados en alguna forma de par clave-valor donde la clave (la variable) es al final un elemento de metadatos que da sentido a los datos.

[4] W3c. Data on the Web Best Practices. https://www.w3.org/TR/dwbp/

Este par clave-valor es lo que en términos de FAIR[5] se denomina "una asociación única entre dos conceptos", que es "uno de los más pequeños posibles "Elementos de Datos".

Ambas partes del par clave-valor pueden presentar problemas de interoperabilidad, por lo que tratamos los problemas de interoperabilidad de los metadatos (por ejemplo, de acuerdo a esquemas para los elementos de metadatos, o nombres de variables) y cuestiones de interoperabilidad de los datos/valores (por ejemplo, de acuerdo a listas controladas/rangos de los que debe tomarse el valor, o cuestiones de sintaxis). Sin embargo, dado que también es habitual considerar valores controlados y la especificación de la sintaxis o la unidad de medida como "metadato" (sobre todo porque lo ideal es que se definan mediante elementos de metadatos separados y no dentro del propio valor), también podemos decir que la interoperabilidad tiene que ver sobre todo con los metadatos. La interoperabilidad de los datos se consigue mediante la interoperabilidad de los metadatos, por ejemplo, queremos que los metadatos "temperatura del aire" sean interoperables para luego obtener el valor real (el dato) que necesitamos - el número que expresa el valor nunca será encontrable o comprensible, por si mismo, sin los metadatos asociados.

Seguiremos la misma convención adoptada en los principios rectores de FAIR mencionados anteriormente, utilizando el término (meta)datos cuando algo se aplica indistintamente a los datos y a los metadatos.

4. Interoperabilidad de los datos e interoperabilidad de los datasets

Hay metadatos que acompañan a los datos individuales recogidos y hay metadatos generales sobre toda la colección a la que pertenecen los datos.

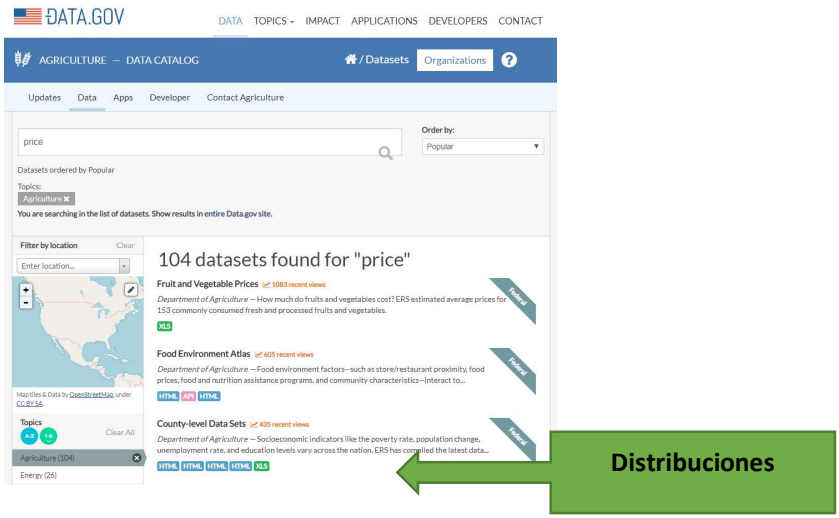
Un dataset puede describirse y hacerse interoperable como un producto en si mismo, una "obra" (con sus propios metadatos) y como los datos contenidos en un datasets (con sus propios metadatos, las variables a las cuáles corresponden los valores).

La interoperabilidad puede lograrse a nivel de los metadatos del dataset y a nivel de los datos: unos buenos metadatos sobre el dataset pueden hacer un dataset descubrible; unos buenos metadatos sobre la estructura de los datos pueden hacer los datos analizables (además, el uso de semántica, como nombres de variables estandarizados vinculados a vocabularios acordados o valores controlados en los registros de datos reales, hará que los datos sean comprensibles y totalmente interoperables; véase, para más información al respecto, la lección 4.4).

Los principios FAIR son muy explícitos en cuanto a las diferentes capas de metadatos/datos: según su definición, un Objeto de Datos es "un elemento de Datos con elementos de Datos + Metadatos" y utilizan el término (Meta)datos cuando "el principio es válido tanto para los Metadatos así como para la actual colección de Elementos de Datos en el Objeto de Datos".

Además, el mismo dataset puede distribuirse en diferentes formas (un archivo Excel, un CSV, un formato específico de la aplicación) y para que las máquinas sepan qué distribución deben obtener, también necesitamos metadatos sobre las distribuciones (por ejemplo, formato, licencia, nuevos metadatos estructurales).

[5] Hemos hablado de los principios FAIR en la lección 4.1. Ver https://www.force11.org/fairprinciples y https://www.go-fair.org/fair-principles/ para concretar el proceso de puesta en marcha



Es importante ser consciente de que existen diferentes niveles de metadatos: metadatos que describen el dataset (metadatos de descripción/metadatos de descubrimiento), metadatos que describen las distribuciones y metadatos que describen los datos (metadatos estructurales que describen los nombres, los tipos y los rangos de los valores del dataset, por ejemplo, los nombres y los tipos de las columnas en un dataset tabulado) que pueden estar a nivel del dataset o a nivel del elemento de datos.

Todos los principales modelos de conjuntos de datos y metadatos prevén metadatos al menos a dos niveles: dataset y distribución; la mayoría de ellos también prevén metadatos a nivel de registro o incluso (en el caso de FAIR) de elementos de datos más pequeños, como un par clave-valor.

Layer	Description	Example
FDP (Data repository)	Information about the FDP as a data repository	PID, title, description, license, owner, API version. etc.
Catalog	Information about the catalog of datasets offered	PID, title, description, publisher, etc.
Dataset	Information about each of the offered datasets	Publisher, issue date, theme, etc.
Distribution	Information about how the dataset is distributed	AccessURL, downloadURL, format, mediaType, etc.
Data record	Information about the actual data, types, identifiers, etc.	Data items types, identifiers, domain, range, etc.

Figura 1: ***Capas de metadatos (Luiz Olavo Bonino et al., 2016, 'Fair Data Technology Update')***[6]

[6] https://www.slideshare.net/lolavo/dtl-partners-event-fair-data-tech-overview-day-1

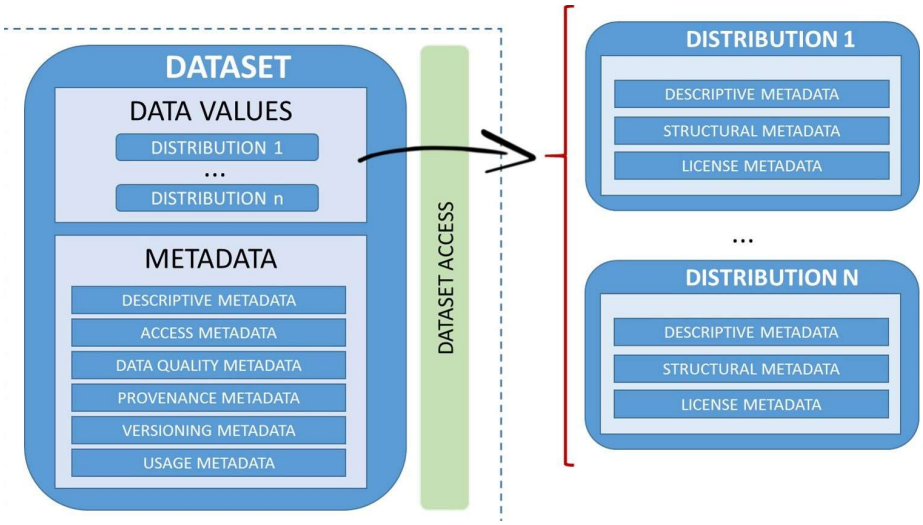


Figura 2: ***Modelo de buenas prácticas del W3C sobre datos en la web***[7]

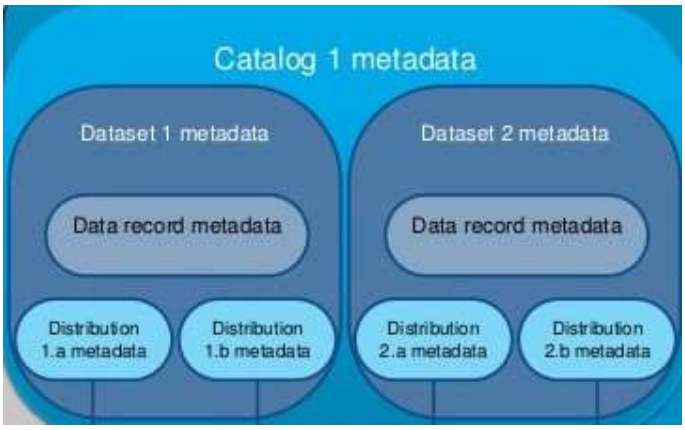


Figura 3: ***Modelo FAIRPort (de Luiz Olavo Bonino et al., 2016, 'Fair Data Technology Update')***[8]

Tradicionalmente (véanse los formatos de datos de larga duración como NetCDF[9] o HDF5[10], o considerar el modelo de la "ISO 19115: Información Geográfica – Metadatos estándar" que abarca docenas de elementos de metadatos) los metadatos a nivel del dataset incluyen información espacial y temporal sobre la recopilación de datos, información sobre la autoría, algunas condiciones ambientales que pueden aplicarse a toda la base de datos, y los nombres de las variables utilizadas en la secuencia

[7] https://www.w3.org/TR/dwbp/

[8] https://www.slideshare.net/lolavo/dtl-partners-event-fair-data-tech-overview-day-1

[9] http://www.unidata.ucar.edu/software/netcdf/docs/

[10] https://support.hdfgroup.org/HDF5/doc/H5.format.html

de datos (que definen la estructura de los registros de datos). Los nombres de las variables en los registros de datos posteriores serían los metadatos de los datos individuales.

Por ejemplo, los nombres de las columnas de un dataset tabulado son metadatos sobre los datos (los metadatos sobre el dataset podrían estar en un archivo de documentación separado). En las estructuras típicas de los datasets de observación, tales como NetCDF o HDF5, los metadatos sobre el dataset están al principio e incluyen los "metadatos de descubrimiento " y "metadatos de uso", que tienen que apoyar el uso de los datos y normalmente contienen todas las dimensiones y los nombres de las variables utilizadas en el dataset; a continuación, una secuencia de registros en forma de tabla o matriz con los datos actuales.

Esto también es importante a la hora de elegir los datos más adecuados, estándares de datos o "vocabularios" (véase la lección 4.4) que se van a utilizar para los datos propios: hay vocabularios para describir conjuntos de datos y hay vocabularios para codificar los registros reales de datos (y hay vocabularios que hacen ambas cosas pero utilizando diferentes clases con diferentes propiedades). La elección del vocabulario/diccionario adecuado, tanto para los metadatos del dataset como para los datos, es esencial.

Y esto es importante a la hora de elegir una herramienta para la gestión de conjuntos de datos: hasta ahora la mayoría de herramientas no admiten un modelo de metadatos que contemple diferentes niveles de metadatos. Es de esperar que las herramientas de repositorio de datos y los vocabularios de conjuntos de datos cubran esto pronto. Por el momento, parece que el único vocabulario de conjuntos de datos que lo cubre es el vocabulario de cubos de datos del W3C.

En particular, en lo que respecta a los metadatos de los conjuntos de datos, la falta de metadatos de "estructura de datos" (por ejemplo, los nombres de las variables y los tipos, unidades de medida y códigos de los valores) puede dificultar seriamente la interoperabilidad real de los datos: las máquinas sólo pueden entender el significado de los datos si los metadatos sobre el dataset incluyen metadatos estructurales que describen las dimensiones y variables utilizadas en los datos, es decir, los verdaderos "metadatos" sobre los datos, que dan sentido a los datos.

En realidad, con una infraestructura distribuida de triples como en la web semántica (véase la lección 4.3), que representan los metadatos en todos los niveles posibles, el concepto de dataset pasa a ser menos una entidad física o un archivo que una entidad conceptual, una colección de registros (que residen en cualquier lugar, vinculados a la colección de pertenencia) con la misma estructura. Es decir, algunos triples sobre un dataset estarán en un sitio, otros triples sobre el mismo dataset en otro lugar, y los triples para los registros contenidos potencialmente en cualquier lugar (pero todos vinculados a la URI del dataset).

En las siguientes lecciones se explicará con más detalle cómo implementar la interoperabilidad estructural e interoperabilidad estructural y semántica.

Lecturas adicionales

Tom Heath and Christian Bizer (2011) Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. http://linkeddatabook.com/editions/1.0/