Manejo de Datos Abiertos en la Agricultura y Nutrición

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

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

Unidad 4: Compartir datos abiertos

Lección 4.1: Marco de referencia para el intercambio de datos

Foto de Fintrac Inc. Licencia CC BY NC 2.0

Objetivos y resultados de aprendizaje

Esta lección tiene como objetivo;

- Presentar los marcos de referencia recomendados para compartir los datos (FAIR, 5 estrellas de TBL, Linked Open Data, las mejores prácticas del W3C).
- Hacer coincidir los principios e indicaciones de dichos marcos con directrices prácticas sobre cómo aplicarlos

Después de estudiar esta lección, se debe ser capaz de;

- comprender los principios de los marcos de referencia más importantes para compartir los datos abiertos
- comprender las principales implicaciones practicas de estos marcos
- evaluar diferentes formas de publicar datos, incluidas las herramientas relacionadas con estos marcos.
- identificar los problemas de sostenibilidad en una política de datos abiertos, y ser consciente de cómo superar estos problemas.

Contenidos

Lista de figuras

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

El intercambio de datos es algo que ha ocurrido a lo largo de los siglos, especialmente entre las oficinas de estadística y los investigadores. En la agricultura el intercambio de datos es necesario, de la misma manera que lo es para la investigación, la innovación y los negocios (por ejemplo, precios de mercado, infraestructuras, información meteorológica). También son importantes para la política y la legislación (por ejemplo, para el seguimiento de los productos alimentarios y el uso de pesticidas). Así que las partes implicadas son variadas y sus intereses unas veces están alineados y otras no, pero al final todos necesitan datos de otros actores o de otras partes de la cadena de valor de los datos

Un factor clave que ha cambiado el juego del intercambio de datos en las últimas décadas ha sido la llegada de Internet y, más recientemente, las tecnologías en la nube y del big data, que han multiplicado el potencial de la capacidad de procesamiento de datos. Al principio de la web, su inventor Tim Berners-Lee la vio como una "red de datos". En la última década se ha reconocido que los principales consumidores de datos e intermediarios en el proceso de intercambio de datos son las máquinas y, por lo tanto, para ser compartidos, los datos deben ser legibles por ellas.

La facilidad para compartir datos trae consigo preocupaciones legales sobre el uso que se puede hacer de ellos. El hecho de que estén en la web sólo significa que se pueden leer, pero nada se dice sobre el permiso para reutilizarlos, modificarlos o redistribuirlos. Si los permisos no se explicitan, los datos no pueden reutilizarse legalmente.

Por lo tanto, compartir datos requiere acuerdos sobre cómo deben "escribirse" los datos para las máquinas (porque las máquinas tienen que estar programadas para leerlos y tienen que conocer las reglas), pero también sobre cómo se pueden utilizar: si se pueden modificar, si se pueden redistribuir, etc. En resumen, los derechos de uso y cómo formalizarlos (en la unidad 5 se puede encontrar más información sobre el tema de las licencias de los datos). Acordar esto puede ser sólo una cuestión de acuerdos ad hoc o de especificaciones técnicas o, con una visión más ambiciosa y a largo plazo. Se puede llevar a la creación de marcos rectores generales autorizados que lleguen a ser ampliamente respaldados. En esta lección presentaremos los dos principales marcos que se han establecido y aprobado para compartir datos.

1.1. Desarrollo el esquema de 5 estrellas de Tim Berners-Lee para los datos abiertos

El primer marco de referencia para compartir datos fue diseñado por el propio inventor de la web, Sir Tim Berners-Lee. El marco técnico que diseñó para la web de los datos es el "Linked Open Data" (LOD) o simplemente Linked Data1 que se formalizó por completo en 2006. Este consiste en unas directrices para que los datos estén totalmente enlazados y básicamente recomienda:

- utilizar URIs como nombres de cosas
- utilizar URIs HTTP para que la gente pueda buscarlos
- cuando alguien busque una URI, proporcionar información útil; y
- incluir enlaces a otras URI, para que puedan descubrir más cosas.

Debido a la barrera de entrada que supone este tipo de soluciones tecnológicas lo que hizo que en el 2010 publicara el esquema de despliegue de 5 estrellas para los datos abiertos, "con el fin de animar

[1] http://linkeddata.org/ and https://www.w3.org/wiki/LinkedData

a la población -especialmente a los propietarios de datos gubernamentales2-. El esquema de 5 estrellas ilustra el "camino" en la publicación de datos que lleva a los pasos finales de los datos abiertos totalmente enlazados (el marco LOD es sólo un subconjunto de él, las dos últimas estrellas).

★	Publicar la información en la web (en cualquier formato) con una licencia abierta
★★	Publicar los datos estructurados (por ejemplo, Excel en lugar de imagen escaneada)
★★★	Formato no propietario (e.g. CSV en lugar de Excel)
★★★★	Utiliza URIs para identificar objetos, para que la gente pueda apuntar a tus objetos.
★★★★★	Vincule sus datos con los de otras personas para proporcionarles un contexto.

Figura 1. Tim Berners-Lee's 5 estrellas del esquema open data

Las 5 estrellas de TBL siguen siendo un marco de referencia para todos los que trabajan con datos abiertos. Normalmente se han interpretado como "acumulativas", en el sentido de que "cada estrella adicional presupone que los datos cumplen los criterios de la(s) etapa(s) anterior(es)", lo que siempre ha dado un peso enorme a la primera etapa, una licencia abierta, como una especie de requisito en ausencia del cual incluso la aplicación de las otras cuatro estrellas no te llevaría a datos realmente abiertos3. Todas las demás estrellas están relacionadas con la interoperabilidad de los datos (lo veremos mejor en la unidad 2 y en las lecciones 4.2 y 4.3, junto con el marco de Linked Open Data estrechamente asociado), mientras que la primera estrella tiene que ver con la apertura para la reutilización. Las 5 estrellas de TBL son, en realidad, un marco para la apertura y, por ello, también se denominan "las 5 estrellas de la apertura".

[Para ser más precisos, para distinguir entre Linked Data y Linked Open Data, TBL dijo: "Por supuesto, Linked Data no tiene por qué ser abierto en principio: hay muchos usos importantes de los datos enlazados a nivel interno para datos personales y de grupos. Se pueden tener datos enlazados de 5 estrellas sin que sean abiertos. Sin embargo, si se dice que es Linked Open Data, para obtener alguna estrella tienen que ser abiertos4]

El concepto de "abierto" ha sido, durante décadas, la piedra angular de todas las iniciativas de intercambio de conocimientos y datos. La definición oficial de "datos abiertos" es que se trata de datos que pueden ser utilizados, reutilizados (modificados) y redistribuidos (compartidos) libremente por cualquiera5. Es un concepto general que se aplica a cualquier tipo de recurso (documento, imagen, dataset), tiene unas pocas reglas y es bastante fácil de aplicar.

Sin embargo, en los últimos años se ha considerado que algunos de los condicionantes de este marco pueden desalentar el intercambio de datos, especialmente en el contexto de la investigación intensiva de datos donde los datos transmitidos en los distintos pasos de la cadena de valor de los datos son muy estrictos:

- La definición de "licencia abierta" es muy estricta: en el curso básico sobre datos abiertos, el Instituto de Datos Abiertos (ODI) afirma que "esta licencia debe permitir a las personas utilizar los datos de la forma que deseen, incluyendo su transformación, combinación y compartición

[2] Tim Berners-Lee. Linked Data. https://www.w3.org/DesignIssues/LinkedData.html
[3] https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data
[4] Tim Berners-Lee. Linked Data. https://www.w3.org/DesignIssues/LinkedData.html
[5] http://opendatahandbook.org/guide/en/what-is-open-data/

con otros, incluso con fines comerciales6". La exigencia de una licencia abierta podría impedir que se compartan datos que tienen algunas restricciones de acceso, pero que se pueden reutilizarse en comunidades mediante acuerdos sencillos y, por tanto, pueden seguir teniendo un gran impacto.

- Las estrellas cuarta y quinta, especialmente en las versiones que describen las estrellas con mucho detalle, se consideran en algunas ocasiones demasiado ligadas al enfoque técnico del Resource Description Framework (RDF) en lugar de ser principios genéricos que puedan aplicarse con cualquier tecnología (véanse las siguientes secciones y la lección 4.2)

Esto no significa que este marco haya sido superado: sigue siendo el marco de referencia para la alta interoperabilidad de los datos y para una red de datos abierta de abajo hacia arriba y con poca conexión.

1.2. Los principios FAIR

Recientemente, dado que se reconoce unánimemente que la reutilización de los datos es el gran motor de la innovación, y que la forma de compartir los datos es clave para su reutilización, ha surgido un nuevo interés en torno a la definición de un marco más formal y coordinado que pueda atender más a la investigación con uso intensivo de datos y a la puesta en común de los mismos en toda la cadena de valor.

En 2014, la necesidad de definir mejor las "reglas" para una puesta en común más eficaz de los datos llevó a un grupo de representantes de diferentes partes -el mundo académico, la industria, las editoriales y los organismos de financiación- a reunirse en Leiden (Países Bajos) y debatir un "conjunto mínimo de principios rectores y prácticas que fueran acordadas por la comunidad7"

Lo que surgió de estos debates fue un conjunto de principios denominados principios FAIR: según estos principios los datos deben ser Localizables, Accesibles, Interoperables y Reutilizables (FAIR).

Al leer los cuatro principios, este marco parece estar muy en consonancia con el concepto de abierto y el esquema de 5 estrellas del TBL (y de hecho no hay discontinuidad entre ambos, y el marco FAIR no contradice ni sustituye al marco de 5 estrellas del TBL); sin embargo, al leer los detalles sobre cada uno de los principios, hay algunas diferencias clave y un nivel más alto de generalización, que hacen que el principio FAIR sea más un verdadero "marco" formal:

A. La atención se centra en unos derechos de acceso claros más que en la apertura, en consonancia con la necesidad de una mayor flexibilidad (pero también de una mayor precisión) en la puesta en común de los datos. El fin es facilitar también la puesta en común de los datos que pueden tener algunas restricciones de acceso, pero que pueden ser reutilizados en condiciones específicas: "Apreciamos las excepciones al acceso abierto total de los datos (por ejemplo, por razones de privacidad de los pacientes o de propiedad intelectual). Por lo tanto, consideramos que la concesión de licencias adecuadas para los datos (o incluso para algunos de los complementos de los datos) es clave para la publicación de datos FAIR".
B. Se presta especial atención a la procedencia, atribución y preservación, en consonancia con el hecho de que los principios FAIR han sido acordados por una comunidad que desea trabajar conjuntamente y compartir datos. Para ello necesita un entorno de confianza con algunas normas básicas. Bajo este aspecto, los principios FAIR necesitan reglas más rigurosas en lo que relacionado con la confidencialidad de los datos que la diseñada por TBL que es más de abajo

[6] http://training.theodi.org/InADay/#/id/co-01; ver también http://opendefinition.org/od/2.1/en/
[7] Force11. Guiding Principles For Findable, Accessible, Interoperable And Re-usable Data Publishing Version B1.0. https://www.force11.org/fairprinciples

hacia arriba.

C. Los principios no quieren llegar al nivel de las especificaciones técnicas; son una "guía general para el carácter FAIR de los datos", no una "especificación". Al recopilar los principios rectores FAIR para este documento, se han evitado conscientemente las opciones de implementación técnica. Los [Principios Rectores FAIR] mínimos pretenden guiar a los implementadores de entornos de datos FAIR para que comprueben si sus elecciones de implementación particulares hacen que los datos resultantes sean realmente FAIR".
D. En contraste con las 5 estrellas de TBL los principios FAIR no son “acumulativos”: “Estas facetas FAIR están obviamente relacionadas, pero técnicamente son algo independientes entre sí, y pueden ser implementadas en cualquier combinación, de forma incremental, a medida que los proveedores de datos […] evolucionan hacia grados crecientes de carácter FAIR”.

A continuación, se presentan algunos de los detalles clave de cada principio que ejemplifican las características descritas anteriormente:

1. Algunos detalles del principio encontrable
   a) "Los conjuntos de datos deben ser persistentes, con énfasis en sus metadatos".
   b) "Los identificadores de cualquier concepto utilizado en los conjuntos de datos deben ser, por tanto, únicos y persistentes".

2. Algunos detalles del principio de accesibilidad:
   a) "previa autorización adecuada" - esto significa que incluso los datos que requieren un permiso especial se consideran FAIR si aplican los otros principios, por lo que los datos no tienen que ser necesariamente utilizables "por cualquiera".
   b) "a través de un protocolo bien definido": esto significa que la interoperabilidad no está vinculada a un protocolo (por ejemplo, HTTP y, más concretamente, una API REST o un punto final SPARQL, que subyacen en el marco de Linked Data).

3. Algunos detalles del principio de interoperabilidad:
   a) "Los (metadatos) son procesables por la máquina".
   b) "Los formatos de (los metadatos) utilizan vocabularios y/o ontologías compartidas".
   c) "Los (metadatos) dentro del dataset deben ser sintácticamente analizables y semánticamente accesibles para la máquina". Estas dos indicaciones se aplican perfectamente con RDF, pero no se menciona ninguna tecnología específica. En general, todo el principio de interoperabilidad está muy en consonancia con las estrellas de TBL y todos los principios se basan en el supuesto clave de que los datos tienen que ser "adecuados tanto para las máquinas como para las personas" y que "el hecho de que los metadatos sean legibles por las máquinas es una condición sine qua non para la imparcialidad".

4. Entre las indicaciones prácticas que detallan el principio de Reusabilidad tenemos:
   a) "Los Conjuntos de Datos Publicados deben hacer referencia a sus fuentes y procedencia con unos metadatos lo suficientemente ricos como para permitir una cita adecuada" - esto está en línea con el objetivo de construir una infraestructura de repositorios de datos de confianza donde la autoría y la atribución son particularmente importantes.

Los principios FAIR están siendo adoptados rápidamente por distintos colectivos, especialmente por los financiadores de la investigación. Recientemente han sido adoptados por las directrices de la

Comisión Europea sobre la gestión de datos FAIR en el programa Horizonte 2020 y su continuación, Horizonte Europa8.

1.3. Otros marcos generales

Hay otros marcos que hay que tener en cuenta a la hora de publicar datos abiertos, uno más genérico que los dos anteriores (Acceso Abierto), otro más estricto en cuanto a licencias (Contenido Abierto) y otro que es técnicamente una "mejor práctica" muy detallada con claras opciones de implementación (W3C Data on the Web Best Practices).

Los movimientos de Acceso Abierto y Ciencia Abierta abogan, respectivamente, por la publicación de los resultados de la investigación (desde artículos de revistas hasta tesis y conjuntos de datos) de forma accesible (ya sea mediante el autoarchivo o a través de revistas de acceso abierto) y por facilitar "el acceso a la investigación, los datos y la divulgación científica a todos los niveles de una sociedad inquieta, aficionada o profesional". Seguir las indicaciones sobre cómo lograr el acceso abierto siguiendo la vía verde9 (autoarchivo) o la vía dorada10 (publicando en revistas de acceso abierto, algunas de las cuales publican ahora también conjuntos de datos) es ya un paso esencial para compartir los datos.

En los nuevos proyectos financiados por la UE, el acceso abierto a los datos de investigación es un requisito opcional pero será obligatorio en el próximo programa marco. Ejemplos de lugares en los que podrían publicarse los conjuntos de datos son los catálogos públicos como OpenAIRE11 o Dryad12.

Las Mejores Prácticas de Datos en la Web del W3C13 ofrecen directrices muy detalladas relacionadas con la publicación y el uso de datos en la Web, diseñadas para ayudar a apoyar "un ecosistema autosuficiente". En comparación con los principios FAIR, estas mejores prácticas profundizan mucho más en la aplicación técnica real y recomiendan soluciones específicas.

El enfoque de compartir datos es muy similar al de los principios FAIR, destacando la necesidad de atender también a la publicación de datos con acceso controlado, la necesidad de fiabilidad y persistencia de los datos y la necesidad de acordar un conjunto de normas comunes.

Las mejores prácticas están vinculadas a una serie de "beneficios": cada beneficio representa una mejora en la forma en la que los conjuntos de datos se encuentran disponibles en la web: Comprensión (humana), Procesabilidad, Descubrimiento, Reutilización, Confianza, Vinculación, Acceso, Interoperabilidad.

8 https://ec.europa.eu/info/sites/default/files/research_and_innovation/strategy_on_research_and_innovation/presentations/horizon_europe_es_invertir_para_dar_forma_a_nuestro_futuro.pdf
9 https://en.wikipedia.org/wiki/Open_access#Self-archiving:_green_open_access
10 https://en.wikipedia.org/wiki/Open_access#Journals:_gold_open_access
11 https://www.openaire.eu
12 http://datadryad.org/
13 https://www.w3.org/TR/dwbp/

Figura 2. Las ventajas del W3C para aplicar las mejores prácticas de datos en la web.

Como se puede ver, los principios FAIR están todos cubiertos, pero el marco del W3C cubre también un paso atrás (comprensión humana) y un paso más allá (procesabilidad de los datos). En la siguiente sección mencionaremos algunas de las soluciones técnicas recomendadas por las mejores prácticas del W3C.

Una definición aún más abierta de lo que significa "abierto" es la que ofrecen las 5R del contenido abierto14: El contenido abierto es "cualquier obra susceptible de ser protegida por derechos de autor (tradicionalmente excluyendo el software, que se describe con otros términos como "código abierto") que tiene una licencia que proporciona a los usuarios un permiso libre y perpetuo para realizar las 5R: Conservar, Reutilizar, Revisar, Remezclar, Redistribuir".

1.4. Un marco para los datos abiertos para la agricultura y la nutrición

Además de estos marcos generales para compartir datos, ha habido iniciativas dedicadas a la defensa de los datos abiertos en sectores específicos. En el caso de la alimentación y la agricultura, tras un acuerdo en la Conferencia Internacional del G8 sobre Datos Abiertos para la Agricultura en 2012, se lanzó la iniciativa Datos Abiertos Globales para la Agricultura y la Nutrición15 (GODAN) en la Conferencia de la Asociación para el Gobierno Abierto celebrada en octubre de 2013. La iniciativa se centra en conseguir apoyos de alto nivel entre gobiernos, responsables políticos, organizaciones internacionales y empresas.

GODAN cuenta con una Declaración de intenciones16 a la que se han adherido más de 500 socios hasta el momento, y que está orientada a los datos totalmente abiertos:

> La iniciativa de Datos Abiertos Globales para la Agricultura y la Nutrición (GODAN) pretende apoyar los esfuerzos globales para que los datos agrícolas y nutricionales estén disponibles, sean accesibles y se puedan utilizar sin restricciones en todo el mundo".

Sin embargo, reconoce las preocupaciones legítimas sobre la apertura total:

> 'la iniciativa aboga por "políticas de datos abiertos y de acceso abierto por defecto, tanto en el sector público como en el privado, al tiempo que respeta y trabaja para equilibrar la apertura con las preocupaciones legítimas en relación con la privacidad, la seguridad, los derechos de la comunidad y los intereses comerciales'.

1.5. Herramientas de evaluación de datos

Se han desarrollado algunas herramientas para evaluar la apertura y/o equidad de los datos: dos ejemplos son los certificados ODI17 desarrollados por el Instituto de Datos Abiertos (una herramienta para evaluar y reconocer la publicación sostenible de estándares de datos abiertos de calidad, basada en marcos como opendefinition.org, 5-star Open Data, principios Sunlight18, DCAT) y el Sello de Aprobación de Datos19 de los Servicios Holandeses de Archivo de Datos y Redes (muy

[14] http://opencontent.org/definition/
[15] http://godan.info
[16] http://www.godan.info/pages/statement-purpose
[17] https://certificates.theodi.org/en/
[18] https://sunlightfoundation.com/policy/documents/ten-open-data-principles/
[19] https://www.datasealofapproval.org

en línea con los principios FAIR, pero más relacionado con la calidad de los repositorios digitales, no con conjuntos de datos individuales).

2. Recomendaciones prácticas para aplicar los principios de intercambio de datos

Algunos de los principios descritos anteriormente tienen implicaciones en términos de política de datos, pero la mayoría de ellos tienen fuertes implicaciones en la implementación técnica. Especialmente en el caso del principio de interoperabilidad, muchas de las implicaciones son muy técnicas y se tratarán con más detalle en las lecciones 4.2-4.4. Los detalles sobre la implementación de un buen repositorio de datos se ven en la lección 3.3 y puedes compararlos con los requisitos aquí enumerados. Aquí sólo ofrecemos algunas recomendaciones generales que pueden ayudar al gestor de datos a seleccionar las herramientas adecuadas o a orientar a los desarrolladores en la elección de soluciones tecnológicas para aplicar los principios más importantes de la compartición de datos. En el texto mencionaremos como ejemplos de herramientas de repositorio de datos CKAN20 y Dataverse21 (más en la lección 3.3).

Aunque los principios FAIR no entran en opciones de implementación técnica explícitas, el esquema de 5 estrellas de TBL ofrece algunos ejemplos e indica tecnologías específicas. Pero, sobre todo, el documento que puede ayudar a identificar soluciones tecnológicas para la publicación de datos es el W3C Data on the Web Best Practices (DWBP) mencionado anteriormente, al que nos referiremos a menudo. Es un documento muy técnico, pero contiene consejos para prácticamente todo lo que se necesita para publicar datos en la web, y se ajusta tanto a las 5 estrellas de TBL como a los principios FAIR.

2.1. Licencias

Como se ha visto, las prescripciones para la concesión de licencias son diferentes en el esquema de TBL (licencia abierta) y en el marco FAIR (cualquier licencia que aclare los derechos de uso).

Se trata, sobre todo, de una elección sobre la política de datos, ya que hay que tener en cuenta varios factores: si los datos han sido producidos disfrutando de fondos públicos y existe un mandato para publicar los resultados en acceso abierto o en un portal de datos; si el dataset incluye datos sensibles; si los datos tienen valor comercial (o si podrían entrar en un espacio precompetitivo); y, por supuesto, la política de la organización.

Incluso se pueden publicar diferentes partes de los conjuntos de datos bajo diferentes licencias o aplicar licencias a elementos de datos específicos (siguiendo el modelo FAIR de objetos de datos "modulares" y "recurrentes").

Ambos enfoques de la concesión de licencias pueden, en cualquier caso, aplicarse siguiendo la Buena Práctica 4 de DWBP: Proporcionar información sobre la licencia de los datos:

> *"En el contexto de los datos en la web, la licencia de un dataset puede especificarse dentro de los datos, o fuera de ellos, en un documento separado vinculado. [...] Las máquinas deberían poder detectar automáticamente la licencia de los datos de una publicación".*

[20] https://ckan.org/
[21] https://dataverse.org/

DWBP también ofrece ejemplos de implementación: vocabularios que pueden utilizarse para expresar los metadatos de las licencias (Dublin Core Terms, schema.org) e incluso lenguajes de derechos legibles por máquina, como el Creative Commons Rights Expression Language, el Open Data Rights Language y el Open Data Rights Statement Vocabulary.

Lo ideal sería que, para ser interpretados de forma unívoca, los metadatos de las licencias apuntaran a un URI o al menos a una URL de una licencia publicada.

Para más información sobre las licencias de datos, véase la unidad 5 de este curso.

2.2. Infraestructura de los repositorios: protocolos, persistencia e identificadores

Agrupamos aquí una serie de requisitos de la arquitectura que conciernen a un repositorio de datos más que al dataset en sí: estos requisitos tienen fuertes implicaciones en la elección de soluciones técnicas para la implementación del catálogo. Los detalles técnicos sobre cómo implementarlo requieren una lección específica, a menos que desarrolle usted mismo la plataforma del repositorio, puede utilizar estos requisitos como criterios de evaluación para la plataforma del repositorio que elija. Puede leer más sobre cómo configurar un repositorio de datos en la lección 3.3 y sobre los identificadores persistentes en la lección 2.5.

La 4ª estrella de TBL recomienda simplemente: "usa URIs para nombrar cosas, para que la gente pueda apuntar a tus cosas".

El marco FAIR tiene unos requisitos mucho más exigentes para el repositorio:

- A los (meta)datos se les asigna un identificador globalmente único y eternamente persistente.
- Los (meta)datos sean recuperables por su identificador mediante un protocolo de comunicación estandarizado.
- El protocolo es abierto, gratuito y de aplicación universal
- El protocolo permite un procedimiento de autenticación y autorización, cuando este resulte necesario
- Los metadatos son accesibles, incluso cuando los datos ya no se encuentren disponibles.

En cuanto a los protocolos, el DWBP tiene la práctica recomendada 21: Utilizar interfaces web estandarizadas: "se recomienda utilizar URI, verificación HTTP, códigos de respuesta HTTP, tipos MIME, enlaces HTTP tipificados y negociación de contenidos al diseñar las API". La mejor práctica recomienda explícitamente las APIs RESTful.

En cuanto a los identificadores persistentes, las prácticas más comunes son utilizar:

- Identificadores únicos de recursos (URI) que se resuelven a URLs.
- Identificadores de objetos digitales (DOI).

La Buena Práctica 9 de DWBP: "Utilizar URIs persistentes como identificadores de conjuntos de datos" se basa en gran medida en el marco de Linked Data y proporciona enlaces a muchos documentos técnicos sobre cómo construir URIs y cómo garantizar la persistencia de estas URIs. Sin embargo, también considera los DOI y sugiere una forma de fusionar ambos enfoques añadiendo el DOI a un patrón URI: "Los identificadores de objetos digitales (DOI) ofrecen una alternativa similar. Estos identificadores se definen independientemente de cualquier tecnología web, pero pueden añadirse a un 'stub URI'. Los DOI son una parte importante de la infraestructura digital para los datos de investigación y las bibliotecas".

Aunque implementar todo lo anterior es técnicamente un reto, la mayoría de las herramientas de repositorio de datos existentes crean URIs (por ejemplo, CKAN) o DOIs (por ejemplo, Dataverse) para los conjuntos de datos cargados y utilizan protocolos abiertos (además de http, utilizan APIs REST como SPARQL u OAI-PMH). (Si se quiere ser totalmente compatible con Linked-Data, es importante que se elija una herramienta que cree URIs desreferenciables, implemente la negociación de contenidos22 y utilice el protocolo SPARQL).

Sin embargo, el dominio URI es responsabilidad del propietario del repositorio y la aplicación del principio de persistencia requerirá algún compromiso político para mantener el dominio URI (o los DOI). Para las URI, DWBP sugiere una solución alternativa: "Cuando un editor de datos no puede o no quiere gestionar su espacio URI directamente para la persistencia, un enfoque alternativo es utilizar un servicio de redirección como purl.org. Esto proporciona URIs persistentes que pueden ser redirigidos según sea necesario con lo que la ubicación puede ser efímera".

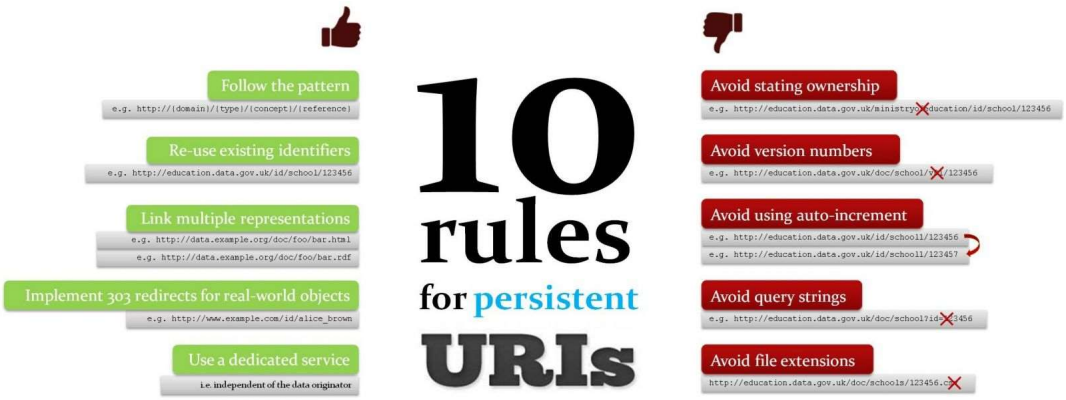

Figura 3: 10 *Reglas para URIs persistentes*23.

2.3. Formatos de los datos

En cuanto a los formatos de los datos, hay recomendaciones sencillas que pueden cumplir los principios de todos los marcos que hemos descrito:

TBL: en el punto 3 recomienda: "ponerlos a disposición en un formato abierto no propietario (por ejemplo, CSV en lugar de Excel)".

FAIR: en el punto 1 de interoperabilidad recomienda: “los (meta)datos utilizan un lenguaje formal, accesible, compartido y ampliamente aplicable para la representación del conocimiento”.

Básicamente, los datos tienen que estar " seriados", expuestos, en un formato legible por las máquinas, que sea fácil de analizar, y preferiblemente sin necesidad de un software propietario. Hay formatos técnicamente legibles por las máquinas (HTML, Excel), pero que no son necesariamente fáciles de analizar, ya sea porque no están rigurosamente estructurados o porque los algoritmos para analizarlos están patentados.

[22] http://linkeddata.org/conneg-303-redirect-code-samples
[23] Fuente: https://joinup.ec.europa.eu/community/semic/document/10-rules-persistent-uris

Recomendamos como mejor práctica la práctica 12 de DWBP del W3C: Utilizar formatos de datos estandarizados legibles por máquina. Hacer que los datos estén disponibles en un formato de datos estandarizado legible por máquina que sea fácilmente analizables, incluyendo pero no limitándose a CSV, XML, Turtle, NetCDF, JSON y RDF".

La mayoría de las herramientas de repositorio de datos existentes ofrecen datos en RDF (por ejemplo, CKAN) o en JSON o XML (por ejemplo, Dataverse). Para más información sobre este tema, véase la lección 4.3 sobre interoperabilidad estructural.

2.4. Metadatos

Las 5 estrellas de TBL no mencionan los metadatos.

Los principios FAIR establecen un papel fundamental de los metadatos, para su localización:

- los datos se describen con metadatos ricos
- los metadatos especifican el identificador de los datos

y para la reutilización:

- los (meta)datos tienen una pluralidad de atributos precisos y relevantes
- los (meta)datos se liberan con una licencia de uso de datos clara y accesible
- los (meta)datos están asociados a su procedencia
- los (meta)datos cumplen con las normas de la comunidad relevantes para el dominio.

Una cosa que hay que tener en cuenta sobre los metadatos, es que, si utilizas una herramienta de repositorio de datos estos se encuentran, normalmente, fuera de tu control. Los repositorios de datos vienen con sus modelos de metadatos y su capa de presentación, y en la mayoría de los casos modificarlos significa hackear el código de programación. Es cierto que estas herramientas tienden a cumplir con los estándares de metadatos existentes (por ejemplo, CKAN expone los metadatos utilizando el vocabulario DCAT, aunque no lo hace en su totalidad), pero esto no siempre es así, y en algunos casos es posible que se quieran añadir estándares específicos de tu área (por ejemplo, para cumplir con que los "(meta)datos cumplan con los estándares de la comunidad relevantes para el área de conocimiento).

Así, por un lado es muy importante que se evalúe el modelo de metadatos de una herramienta de repositorio de datos antes de utilizarla, por otro lado es deseable que la herramienta permita a los gestores de datos ajustarla de manera fácil el modelo de metadatos de su interés.

Incluso si se puede conseguir con la herramienta seleccionada como repositorio de datos, puede ser interesante comprobar la Buena Práctica 1 de DWBP. "Proporcionar metadatos", especialmente en lo que respecta a los metadatos legibles por máquina. En primer lugar, en línea con la mejor práctica de utilizar formatos fácilmente analizables, recomienda serializar los metadatos en formatos como Turtle o JSON (o incrustarlos en la página HTML utilizando RDFa o Json-LD), y luego recomienda el uso de vocabularios existentes: "al definir metadatos legibles por máquina, se recomienda encarecidamente reutilizar los términos estándar existentes y los vocabularios populares. Por ejemplo, los términos del Dublin Core Metadata (DCMI) y del vocabulario del catálogo de datos deberían utilizarse para proporcionar los metadatos descriptivos".

El uso de las normas de metadatos existentes ayuda a aplicar el requisito de FAIR de que los "(meta)datos cumplan las normas comunitarias relevantes para el área" y, con suerte, también los requisitos sobre los metadatos de licencia y procedencia, teniendo en cuenta que los vocabularios adecuados de los conjuntos de datos deben incluir esos metadatos.

Hay más información sobre los metadatos en la lección 4.4 sobre interoperabilidad semántica.

2.4.1. Metadatos y catálogos

Lo ideal sería que los metadatos sobre el dataset estuvieran en el propio conjunto (muchos formatos y vocabularios estructurados permiten modelos jerárquicos o relacionales que incluyen metadatos sobre el dataset y los datos reales) para la autorrecuperación.

Sin embargo, la autorrecuperación supondría una infraestructura existente de conjuntos de datos distribuidos y repositorios de datasets que presentan sus metadatos estandarizados y que se rastrean o federan mediante búsquedas distribuidas, algo que no ocurre. Los principios FAIR prevén que los metadatos de los conjuntos de datos se registren en catálogos de conjuntos de datos en los que puedan encontrarse más fácilmente: "los (meta)datos se registran o indexan en un recurso de búsqueda".

Muchas herramientas de repositorio de datos normalmente también ofrecen buenas funciones de catálogos de datos, proporcionando funcionalidades de búsqueda de metadatos y exponiendo todos los metadatos a través de APIs.

Incluso Tim Berners-Lee añadió una nota al respecto en el diseño de Linked Data: Ahora, en 2010, la gente me ha presionado para que, en el caso de los datos gubernamentales, añada un nuevo requisito, y es que debe haber metadatos sobre los propios datos, y que esos metadatos deben estar disponibles en un catálogo importante. [...] Sí, debe haber metadatos sobre el dataset24".

2.5. Enlaces

La 5ª estrella de TBL recomienda: "vincular los datos con otros datos para proporcionarles un contexto".

Los principios FAIR recomiendan, para la interoperabilidad: "que los (meta)datos incluyan referencias cualificadas a otros (meta)datos".

Este es el núcleo de la arquitectura de Linked Data y el mecanismo básico de la web semántica.

La Buena Práctica 10 de DWBP: Utilizar URIs persistentes como identificadores dentro de los conjuntos de datos recomienda que "los conjuntos de datos deben utilizar y reutilizar los URIs de otras personas como identificadores siempre que sea posible".

Este es otro requisito que debería utilizarse como criterio a la hora de elegir una herramienta de repositorio de datos (y como requisito si se desarrolla una nueva herramienta). Hasta ahora, las herramientas más populares no permiten ni vincular un concepto interno (como una categoría) a uno externo ni utilizar un URI como valor para un elemento de metadatos (excepto como categoría), sin considerarlo un recurso). Por lo tanto, cumplir con este requisito utilizando herramientas como Dataverse o CKAN es por el momento difícil.

---

[24] Tim Berners-Lee. Linked Data. https://www.w3.org/DesignIssues/LinkedData.html

En conclusión, es importante que los gestores de datos analicen las implicaciones prácticas de la aplicación, de los requisitos de los principales marcos de intercambio de datos y que consecuentemente tomen decisiones informadas sobre su repositorio de datos.