MÓDULO 3

3.3. DATOS Y FORMATOS ABIERTOS

POR NURIA RICO CASTRO
PROFESORA DEL DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA DE LA UGR

INTRODUCCIÓN

El origen de lo que hoy se conoce como datos abiertos surge en el año 2009, cuando el gobierno británico pone a disposición de la ciudadanía su portal de datos abiertos, presentándolo como una oportunidad de empoderamiento para la población, un ejercicio de transparencia y una fuente de información para las empresas. Desde entonces, el gobierno británico ha seguido proporcionando datos e implementando iniciativas para estimular la reutilización de la información y la creación de oportunidades.

Hoy en día, los datos abiertos constituyen una de las fuentes de conocimiento abierto más potentes y con mayor proyección existentes. Gracias a la tecnología, la humanidad está cambiando a un ritmo sin precedentes. En una sociedad hiperconectada, con una infraestructura tecnológica avanzada, el tráfico de datos promete ser clave para el funcionamiento de nuevas funcionalidades robóticas, de inteligencia artificial y de computación al servicio de la ciudadanía. Los centros de gestión de datos van a tener un papel relevante en un futuro próximo que ya ha comenzado a perfilarse gracias a la revolución tecnológica. Los datos son, como muchos expertos señalan, el petróleo del nuestro siglo.

En este capítulo hablaremos de los datos abiertos, un pilar fundamental en el progreso de la humanidad.

1. ¿QUÉ SON LOS DATOS ABIERTOS?

Los datos abiertos (*open data* en inglés) son conjuntos de datos, oficiales o de otro tipo, que son accesibles, de forma que pueden ser utilizados para cualquier propósito y redistribuidos a terceras personas.

Con el objeto de denominar o no una obra como “abierta”, por ejemplo, en el caso de que esa

obra sea un conjunto de datos y queramos saber si se trata de datos abiertos, tendremos que ceñirnos a unas directrices estándar. En este caso encontramos en el portal de la Open Knowledge Foundation el documento "Open Definitions" donde se establecen las condiciones necesarias para que pueda denominarse "abierta" (open) cualquier obra. En esta definición existen condiciones que deben cumplirse y otras que pueden, o no, darse. De acuerdo con esta definición, para que un conjunto de datos se considere "open data" deben cumplirse las siguientes condiciones:

Condición 1: licencia abierta

Los datos abiertos tienen una licencia abierta. Esto significa que quien tiene la titularidad de los derechos de autoría (ya sea la persona que los creó u otra) concede al público en general permiso jurídico para utilizar los datos. Esta licencia concede libertad para, como mínimo, utilizar los datos, adaptar los mismos (incorporarlos a otros datos, seleccionar una parte de ellos, cambiarlos de formato, etc.) y reproducir los datos originales o los resultantes de su adaptación y compartirlos con otras personas.

Condición 2: disponibilidad

Los datos abiertos están disponibles y su coste, en caso de haberlo, no es superior al de una reproducción. Habitualmente los datos abiertos son directamente descargables desde internet, pero puede ocurrir que algunos solamente se encuentren en un soporte físico como papel, o deban descargarse desde un ordenador que no está en la red, en cuyo caso su coste no debería superar al de realizar una copia del archivo físico, copia en papel o copia en soporte electrónico adecuado. Los datos adicionales, como por ejemplo la referencia a la autoría o a la licencia utilizada, siempre acompañan a los datos, tanto si están disponibles online como si necesitan otro soporte.

Condición 3: legibilidad por máquinas

Para que un conjunto de datos pueda ser leído por una máquina debe tener un formato compatible con el lenguaje de las computadoras, de forma que se pueda acceder a sus elementos y modificar los mismos. En este sentido, unos datos que se encuentren disponibles online pero en un formato, por ejemplo, de imagen no son considerados como datos abiertos, puesto que no se pueden leer de forma fácil ni hacer las modificaciones que se desee sobre ellos, aunque efectivamente sean datos que están disponibles. Sin embargo, en general unos datos que no puedan ser leídos directamente por una máquina o sea necesaria una

transformación de la forma en que se presentan para ello, pueden considerarse como datos abiertos en un sentido amplio, aunque no es deseable que no se cumpla la condición de legibilidad.

Condición 4: formato abierto

El formato en que se distribuyen los datos debe ser accesible. No serán válidos conjuntos de datos que solamente se puedan leer utilizando un software propietario, sino solamente aquellos que sean legibles utilizando al menos un programa de software libre o código abierto, sin que se puedan imponer condiciones económicas o de otro tipo para ser leído.

Cualquier conjunto de datos que cumpla estas cuatro condiciones puede ser considerado como un conjunto de datos abiertos u *open data*, incluso aunque se tengan otras restricciones que no afecten a estas condiciones. Así, los datos pueden estar sujetos a alguna restricción, como por ejemplo que si se utilizan o reproducen deba referirse la autoría (restricción de atribución) o que la obra que se derive del conjunto original de datos debe compartirse de la misma forma en que se encuentran estos (restricción de compartir igual).

Existen, además de las características que definen a los datos como datos abiertos, cualidades que si bien no son obligatorias sí son deseables en los conjuntos de datos abiertos puesto que hacen que estos sean especialmente valiosos; la primera de ellas es que los datos abiertos estén en un lugar de fácil acceso (ordenados, listados, en un portal propio, que aparezcan en los listados existentes, que se encuentren en las búsquedas de forma rápida, que no se necesite ningún software específico para acceder a ellos, etc.)

Otra característica deseable es que se encuentren en el mismo lugar de forma estable. Lo más habitual es que los datos abiertos se encuentren en alguna plataforma que les proporcione una dirección web fija. Esto permite una explotación de los datos de forma automática, su integración con otros conjuntos de datos y la creación de productos basados en la información. Por ejemplo, una aplicación que tome datos a lo largo del tiempo para dar información a usuarios debe tener la certeza de que los datos que alimentan la misma no van a moverse o a desaparecer.

La tercera característica que da valor a un conjunto de datos abiertos es que se proporcione la información en un formato que permita la interoperabilidad, es decir que se puedan

integrar distintos conjuntos de datos sin que existan problemas de incompatibilidades, de forma que respondan a un mismo estándar.

Una última característica que podemos observar es que se trate de datos no triviales, es decir, que aporten información válida para los intereses de la ciudadanía en su conjunto y que vengan con la suficiente descripción (metadatos) de su origen para que no existan dudas sobre su naturaleza.

2. CLASIFICACIÓN DE LOS DATOS ABIERTOS

Existe una clasificación de los datos abiertos, creada en 2010 por Tim Berners-Lee, basada en una puntuación de 1 a 5 estrellas (más información en https://5stardata.info/es/), clasificando los conjuntos de datos abiertos según sea menor o mayor la facilidad con que se pueden reutilizar los mismos. La clasificación de las iniciativas de datos abiertos propuesta incluye 5 niveles.

★ Datos abiertos nivel "una estrella"

Los conjuntos de datos que se encuentran con una licencia libre ya tienen una estrella por el hecho de ser públicos y libres, aunque se proporcionen en imágenes, vídeos, documentos con formatos difícilmente manipulables o legibles de forma automática mediante ordenador para las personas que los utilicen y que no sean expertas. Un conjunto de datos que tenga por ejemplo un formato *jpeg* de imagen o *pdf* puede considerarse un conjunto de nivel una estrella.

★★ Datos abiertos nivel "dos estrellas"

Para que un conjunto de datos consiga este nivel, los datos deben estar estructurados, es decir, tener forma de matriz y ser legible de forma automática y sencilla, aunque sea haciendo uso de software propietario. Un conjunto de datos que esté estructurado pero que se ofrezca en formato *xlsx* de hoja de cálculo o *sav* del paquete SPSS tendrá dos estrellas.

★★★ Datos abiertos nivel "tres estrellas"

Un conjunto de datos se clasifica en este nivel si los datos, además de estar estructurados, se pueden descargar y trabajar con ellos utilizando software libre. Habitualmente, los datos de este nivel se ofrecen en formato *csv* (valores separados por comas), *txt* (texto plano) o *tsv*

(valores separados por tabulaciones).

Los pequeños portales de datos abiertos, de pequeñas o medianas entidades, corporaciones o agrupaciones, suelen proporcionar datos acerca de sus procesos en estos formatos, haciendo uso de páginas web y bases de datos de tamaño mediano.

★★★★ **Datos abiertos nivel "cuatro estrellas"**

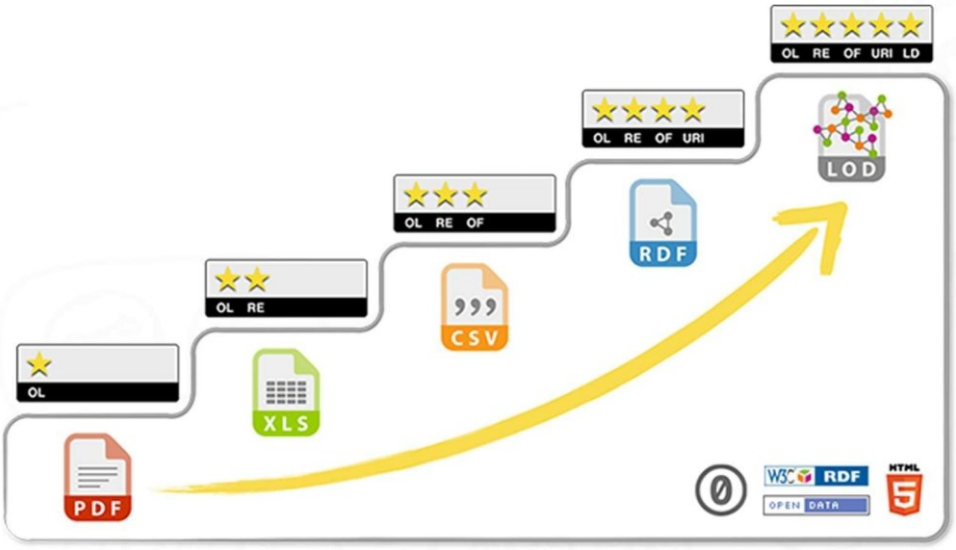
Para que un conjunto se considere de nivel cuatro estrellas es necesario que, además de lo anterior, los datos dispongan de URIs que identifiquen los recursos. Un URI es un Identificador Uniforme de Recursos, (Uniform Resource Identifier) lo que comúnmente se confunde con los URLs (Localizadores Uniformes de Recursos), pero que se diferencian porque el identificador es estable en el tiempo, frente al localizador, que puede no serlo, y por la posibilidad del URI frente al URL de incluir una subdirección dentro de la dirección. Los conjuntos de datos de nivel cuatro estrellas utilizan el estándar RDF (Resource Description Framework o Marco de Descripción de Recursos), que son especificaciones de la World Wide Web Consortium (W3C) para establecer un método común en la descripción de la información que se ofrece en los conjuntos de datos.

Cuando se llega a este nivel, los conjuntos de datos necesitan una estructura que sostenga y dé cuerpo al volumen de información que se maneja. Esta información es la que se encuentra en portales de datos abiertos que aglutinan diferentes iniciativas de liberación de datos y le dan soporte. En España, por ejemplo, se encuentran alojados en el portal de datos abiertos, conjuntos de diferentes niveles y diferentes temáticas. El catálogo, como puede verse en la imagen, es extenso y variado.

★★★★★ **Datos abiertos nivel "cinco estrellas"**

Para que un conjunto de datos alcance el nivel más alto debe cumplir todos los requisitos impuestos hasta el nivel anterior y además los datos deben estar enlazados con otros similares de otras organizaciones de forma que los datos tienen contexto y se facilitan las búsquedas, las comparaciones y la reutilización de la información.

Este tipo de conjunto de datos es bastante más laborioso y costoso de conseguir, pero existe un gran número de iniciativas que consiguen interconectar los conjuntos de datos gracias a la web semántica. Un ejemplo es el portal Big Data Europe o los que se encuentran en el portal de portales de datos abiertos Data Portals.



Esquemáticamente, como se puede ver en la imagen, podríamos decir que si los datos tienen licencia abierta (OL por *Open License*) están en el primer nivel; si además permiten una lectura de filas y columnas fácil (RE por *Readable*) están en el segundo nivel; si se ofrecen en un formato de software libre (OF por *Open Format*) alcanzan el tercer nivel; si además están dotados de identificadores uniformes de recursos (URI por *Uniform Resource Identifier*) alcanzan un nivel más y, en caso de estar enlazados, (LD por *Linked Data*) tendrían el nivel máximo de cinco estrellas.

3. ¿PARA QUÉ SE UTILIZAN LOS DATOS ABIERTOS?

Los datos abiertos constituyen un arma muy poderosa para fomentar la transparencia, la rendición de cuentas y la participación ciudadana. Las instituciones y entidades que ofrecen datos a la ciudadanía ofrecen mucho más que simples resúmenes de información, están proporcionando una fuente necesaria para crear conocimiento, detectar posibles anomalías y ejercer con plenitud de información nuestros derechos. Aunque los campos en los que los datos abiertos son útiles son muy numerosos, se destacan los datos abiertos en el gobierno y en ciencia.

3.1 Datos abiertos gubernamentales

Hay un gran abanico de argumentos para apoyar la apertura de datos oficiales desde los gobiernos. Por una parte, se facilita la transparencia, la responsabilidad y la participación

pública. Esto quiere decir que un gobierno que pone a disposición de la ciudadanía los datos referentes a su territorio (referentes a la población, geográficos, de infraestructuras, datos sanitarios, de educación, sobre inversiones, gastos, recaudaciones, etc.) hace un ejercicio de transparencia que solamente es factible cuando los datos que se tienen son coherentes y defendibles ante análisis y preguntas derivadas de los mismos. Por lo tanto, constituyen un ejercicio de rendición de cuentas inigualable frente a la ciudadanía, que dispone de toda la información para vigilar el buen gobierno y que además puede utilizarlos para reclamar una mejor gestión.

Otro argumento a favor de la apertura de la información oficial es que los datos pueden favorecer la innovación tecnológica y el crecimiento económico, al permitir que terceras personas desarrollen nuevos tipos de aplicaciones y servicios digitales. Los datos abiertos pueden ayudar a identificar desafíos sociales y económicos y pueden así surgir servicios que utilicen los datos oficiales proporcionados por los gobiernos para mejorar la vida de la ciudadanía, proporcionando, por ejemplo, aplicaciones que informen sobre servicios concretos o que faciliten la lectura y comprensión de informaciones válidas para una población. Un pequeño ejemplo de este tipo de valor es cuando una empresa utiliza los datos gubernamentales para recomendar rutas que sean accesibles para personas invidentes o con movilidad reducida.

Hoy en día, la apertura de la información también persigue la creación de un "gobierno abierto", lo cual consiste en informar y empoderar a la ciudadanía para que sea partícipe de forma efectiva en las decisiones gubernamentales analizando de forma crítica la información que tiene a su disposición. De esta forma, la apertura de los datos de un gobierno puede contribuir a la mejora de la educación, de las políticas públicas y la construcción de herramientas para resolver problemas.

En octubre de 2015, la Alianza para el Gobierno Abierto (OGP por sus siglas en inglés) estableció la Carta Internacional de los Datos Abiertos, un conjunto de principios y buenas prácticas para la publicación de datos gubernamentales abiertos.

El Parlamento Europeo aprobó en 2019 una Directiva relativa a los datos abiertos y la reutilización de la información del sector público que persigue la armonización de las normas y prácticas de los Estados miembros en relación con la explotación de la información del sector

público, reconociendo que la evolución hacia una sociedad basada en datos, que utiliza datos de distintos ámbitos y actividades, afecta a la vida de toda la ciudadanía de la Unión, entre otras cosas al permitirles contar con nuevos medios para acceder y adquirir el conocimiento.

Esta directiva, a la que España se acoge en 2021, insta a los Estados miembros a ir más allá de la armonización y estandarización de los datos, promoviendo iniciativas de difusión y aprendizaje de la ciudadanía para una explotación efectiva de los datos puestos a su disposición. Además, promueve la apertura y reutilización de los datos de investigación, indicando que «procede imponer a los Estados miembros la obligación de adoptar políticas de acceso abierto con respecto a los datos de la investigación financiada públicamente y garantizar que dichas políticas son ejecutadas por todas las organizaciones que realizan actividades de investigación y las organizaciones que financian la investigación».

3.2 Datos abiertos en ciencia

En el ámbito científico se genera una gran cantidad de datos, ya sea por observación, medición o mediante la generación pseudoaleatoria de conjuntos de datos con ciertas características.

Uno de los proyectos pioneros en la apertura de datos es el Proyecto Genoma Humano, basado en que toda la información sobre la secuencia genómica humana debe estar disponible libremente y en el dominio público para alentar la investigación y el desarrollo y maximizar su beneficio para la sociedad.

En el año 2004, los ministerios de Ciencia de todas las naciones de la Organización para la Cooperación y el Desarrollo Económicos (OCDE), que incluye a la mayoría de los países desarrollados del mundo, firmaron una declaración donde establecen que todos los datos de investigación financiados con fondos públicos deberían ponerse a disposición del público ya que el acceso abierto y el uso intensivo de los datos de investigación mejoran la calidad y la productividad de los sistemas científicos en todo el mundo. Esta declaración y el posterior documento de Principios y Directrices de la OCDE para el Acceso a los Datos de Investigación de Financiación Pública apuntan la dirección que mejores resultados augura para el progreso científico, basados en los pilares de apertura, transparencia, interoperabilidad, profesionalidad, calidad, eficiencia, sostenibilidad, seguridad, flexibilidad, conformidad legal y protección de la propiedad intelectual

Podemos considerar que, según diferentes definiciones, es común considerar que los datos de investigación se definen como registros de hechos (resultados numéricos, textos, imágenes y sonidos) utilizados como fuentes primarias para la investigación científica y comúnmente aceptados en la comunidad científica por permitir la validación de resultados en la investigación científica. Un set de datos de investigación constituye una representación sistemática parcial del objeto que está siendo investigado. La información científica interpela a las conclusiones obtenidas del análisis de datos y a los resultados de una investigación. Por tanto, los datos producidos en la investigación forman un grupo de materiales extremadamente heterogéneo y complejo, creado para distintos propósitos y mediante procesos también diferentes. Los datos son el "alma" de la investigación, rara vez son objetos sencillos que pueden ser fácilmente compartidos, sino que encarnan las perspectivas epistemológicas quien los crea.

Los datos abiertos en investigación científica son una pieza clave del movimiento que se conoce como *Open Science* y que pretende hacer llegar los resultados de todas las investigaciones a todos los rincones del globo. Un sistema ideal donde la investigación científica fuera accesible y no estuviera sujeta a derechos de autoría aseguraría, si fuera acompañado de la libertad de acceder a los datos de investigación y trabajar con ellos, que los experimentos fueran fácilmente replicados, discutidos, enriquecidos y mejorados.

Uno de los portales más visitados sobre datos científicos es el de Scholarly Publishing and Academic Resources Coalition (SPARC), aunque existe un creciente y abundante número de portales que facilitan la búsqueda y alojamiento de datos científicos a nivel mundial.

4. ¿QUIÉNES SON LOS AGENTES PRINCIPALES EN LOS DATOS ABIERTOS?

Se pueden destacar dos tipos de agentes fundamentales: quienes proveen los datos y quienes los utilizan. A estos dos grupos se deben añadir, por una parte, los legisladores que establecen los marcos legales para que se produzca una liberación efectiva, útil y veraz de los datos y, por otra parte, debemos añadir a quienes establecen los mecanismos técnicos, físicos y computacionales que hacen posible compartir grandes conjuntos de datos interconectados, las empresas que proveen a los gobiernos de plataformas y protocolos que permiten una custodia segura de la información, y en general todo un ecosistema creado para que compartir información de forma fácil y útil sea posible.

Proveedores de datos

Con respecto al primer colectivo, las diferentes normativas locales, autonómicas, estatales, etc., de los países desarrollados, marcan que las instituciones públicas deben rendir cuentas poniendo a disposición de la ciudadanía sus documentos para su reutilización y remarcan que las instituciones deben ser proactivas en la rendición de cuentas. Esto quiere decir que las instituciones deben hacer públicos los datos, en la forma que mejor permita la reutilización de la información, antes de que los datos en cuestión sean solicitados por parte de la ciudadanía. Uno de los motivos principales es que las instituciones públicas deben conocerse a sí mismas en profundidad, para poder ser capaces de tomar las mejores decisiones en base a la evidencia.

Sin embargo, no solamente los gobiernos son proveedores de datos. Existen también numerosas instituciones públicas (universidades, centros sociales, hospitales, colegios, etc) y todas ellas están sometidas a las mismas normas de transparencia y buen gobierno. En España, además de las diferentes legislaciones autonómicas, está vigente la Ley de transparencia, acceso a la información pública y buen gobierno donde se establece de forma clara el derecho de acceso a la información pública: todas las personas tienen derecho a acceder a la información pública.

En cuanto a las instituciones privadas, existen algunas que proveen a la ciudadanía de sus datos, sobre todo aquellas que están inmersas en la filosofía del software libre y el conocimiento abierto. Para una empresa privada, realizar este ejercicio no es obligatorio por ley, pero cada vez más entidades del ámbito privado están adoptando la apertura de datos y están observando el valor añadido que aporta notando el impulso que supone para la adopción de estrategias de inteligencia de negocio (Business Intelligence: aplicaciones, infraestructura y herramientas, y mejores prácticas que permiten el acceso y el análisis de la información para mejorar y optimizar las decisiones y el rendimiento).

5. EL LÍMITE

Una vez observada la gran utilidad que tienen los conjuntos de datos abiertos, los esfuerzos de los diferentes gobiernos por armonizar sus datos y compartirlos, es necesario establecer el límite de la apertura de información, ya que no todos los datos son susceptibles de ser

abiertos. Como se indica en la directiva de la Unión Europea: «deben tenerse debidamente en cuenta las inquietudes relacionadas con la privacidad, la protección de datos personales, la confidencialidad, la seguridad nacional, los intereses comerciales legítimos, como los secretos comerciales, y los derechos de propiedad intelectual de terceros, conforme al principio "tan abiertos como sea posible, tan cerrados como sea necesario"».

En la Ley de transparencia, acceso a la información pública y buen gobierno también se indica que «el derecho de acceso podrá ser limitado cuando acceder a la información suponga un perjuicio para: la seguridad nacional, la defensa, las relaciones exteriores, la seguridad pública», y un extenso etcétera que recorre las posibles incompatibilidades entre la publicación de información y la seguridad y bienestar general de la ciudadanía. También alude a la Protección de Datos Personales y garantía de los derechos digitales, donde se establece claramente un límite que no puede ser sobrepasado ni siquiera amparado por el derecho de información.

6. RECURSOS

- Open Knowledge Foundation - The Open Data Handbook (website https://opendatahandbook.org/)
- Open Data for Development (website: https://www.od4d.net/)
- Open Government Data (website: https://opengovernmentdata.org/)
- Data Portals (website: http://dataportals.org/
- The Open Definition (website: https://opendefinition.org/)
- Clasificación de datos abiertos con el sistema 5 estrellas (website: https://5stardata.info/es/)
- Open AIRE (website: https://www.openaire.eu/)
- Datos abiertos del gobierno de España (website: https://datos.gob.es/
- NORMA TÉCNICA DE INTEROPERABILIDAD DE REUTILIZACIÓN DE RECURSOS DE INFORMACIÓN disponible en https://www.boe.es/diario_boe/txt.php?id=BOE-A-2013-2380
- Proyecto Linking Open Data https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
- DIRECTIVA (UE) 2019/1024 DEL PARLAMENTO EUROPEO Y DEL CONSEJO de 20 de junio de 2019 relativa a los datos abiertos y la reutilización de la información del sector público (versión refundida) https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=uriserv:OJ.L_.2019.172.01.0056.01.SPA
- Definición de Business Intelligence (BI) de Gartner: (http://www.gartner.com/it-glossary/business-intelligence-bi)
- Protección de Datos Personales y garantía de los derechos digitales https://www.boe.es/eli/es/lo/2018/12/05/3/con