Datos abiertos

PID_00271451

Blas Torregrosa García

Tiempo mínimo de dedicación recomendado: 2 horas

Universitat Oberta de Catalunya

Blas Torregrosa García

Ingeniero en Informática y máster universitario en Seguridad de las Tecnologías de la Información y de las Comunicaciones (MISTIC) por la Universitat Oberta de Catalunya (UOC). Especializado en ciberseguridad. Profesor colaborador en el máster de Ciencia de Datos de la UOC y profesor asociado en la Universidad de Valladolid (UVA).

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por el profesor: Ferran Prados Carrasco (2020)

Primera edición: febrero 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Realización editorial: FUOC

Índice

Introducción

En este módulo definiremos lo que se entiende por datos abiertos (*open data*), aunque la definición no es única y, en principio, es independiente de los datos enlazados y de la web semántica. Expondremos también los beneficios que aportan los datos abiertos y la publicación de los mismos.

1. ¿Qué son los datos abiertos?

La definición de datos abiertos u *open data* no es única en la actualidad. Existen diferentes planteamientos con ligeros matices, aunque en esencia no hay diferencias importantes entre las ellas.

La organización Open Knowledge Foundation define los datos abiertos como:

> Los **datos abiertos** son datos que pueden usarse, reutilizarse y redistribuirse libremente por cualquier persona, y que se encuentran sujetos, por lo menos, al requisito de atribución y de compartirse de la misma manera en la que aparecen.

Los factores más importantes relativos al concepto de datos abiertos se resumen en los tres puntos siguientes:

1) Disponibilidad y acceso: la información debe estar disponible como un todo y a un coste razonable de reproducción, preferiblemente descargándola de Internet. Además la información debe estar disponible en una forma conveniente y modificable.

2) Reutilización y redistribución: los datos deben proporcionarse bajo términos que permitan reutilizarlos y redistribuirlos, e incluso integrarlos con otros conjuntos de datos.

3) Participación universal: todo el mundo debe poder utilizar, reutilizar y redistribuir la información. No debe haber discriminación alguna en términos de esfuerzo, personas o grupos. No se permiten restriciones «no cormerciales» que impedirían el uso comercial de los datos o restricciones de uso para ciertos propósitos (por ejemplo, solo para educación).

Esto nos lleva hasta el concepto de **interoperabilidad**.

> IEEE[1] define interoperabilidad como la capacidad de dos o más sistemas o componentes para intercambiar información y utilizar la información intercambiada.

(1)Instituto de Ingenieros Eléctricos y Electrónicos, en inglés, *Institute of Electrical and Electronics Engineers*.

En este caso, es la posibilidad para interoperar o integrar diferentes fuentes de datos. La interoperabilidad es importante porque permite que distintos componentes puedan trabajar juntos y esta capacidad de integrar componentes es esencial para construir sistemas más complejos y grandes.

La esencia de los datos compartidos es que una parte del material abierto pueda a partir de ahí mezclarse con otro material abierto. Esta interoperabilidad es absolutamente fundamental para entender el principal beneficio práctico de la apertura de datos: la capacidad de combinar distintas fuentes de datos o conjuntos de datos y así desarrollar más y mejores productos y servicios.

1.1. Beneficios de los datos abiertos

En la sociedad actual hay multitud de individuos, organizaciones y, especialmente, administraciones públicas que generan y gestionan una gran cantidad y variedad de datos para llevar a cabo sus tareas cotidianas. En este contexto las administraciones públicas (como ayuntamientos, gobiernos regionales o estatales) desempeñan un papel especialmente importante por la cantidad de datos que manejan, pero también porque una parte considerable de esta información es abierta y se pone a disposición de cualquier individuo o institución que desee utilizarla.

Tecnologías de la información

Las tecnologías de la información hacen posible el desarrollo de servicios que permiten responder a cuestiones sobre los datos que generan los organismos públicos de forma automática. Sin embargo, frecuentemente estos datos no están disponibles de forma que sean sencillos de utilizar para poder tratarlos y obtener conocimiento.

Existen muchos campos en los que podemos ver que los datos abiertos han creado valor añadido a la sociedad. A modo de ejemplo podemos enumerar algunas de las más relevantes:

- Transparencia y control democrático.
- Participación ciudadana.
- Creación de nuevos productos y servicios.
- Innovación.
- Mejoras en la eficiencia y eficacia de los servicios ofrecidos.
- Medición del impacto de políticas.

1.2. Publicación de datos abiertos

Cuando una organización o institución desea publicar datos en abierto debe plantearse cinco pasos principales que le conducirán a una correcta publicación de los datos en abierto.

1) Identificación de los conjuntos de datos

El primer paso en el proceso de publicación de datos en abierto es seleccionar el conjunto o los conjuntos de datos que se proponga abrir. Este proceso es iterativo y se pueden incluir nuevos conjuntos de datos en el futuro. En general

no hay requisitos para crear una lista completa de conjuntos de datos que sean candidatos para su publicación. Existen dos puntos principales que se deberían tener en cuenta:

a) En primer lugar debemos asegurarnos de que es viable publicar todos (o parte de) los datos.

b) En segundo lugar hay que asegurarse de que no haya datos personales o privados de personas individuales en el conjunto de datos que se desee publicar. Generalmente se publican conjuntos de datos que no contienen datos de carácter personal. En caso contrario hay que aplicar ciertos procesos de anonimización y protección de la privacidad que garanticen que los datos personales estarán correctamente protegidos en el conjunto de datos abiertos.

Anonimización

De *anonimizar*: «Expresar un dato relativo a entidades o personas, eliminando la referencia a su identidad» (RAE).

2) Selección del formato de datos

En segundo lugar es importante elegir un formato adecuado para la publicación de los datos abiertos. El formato elegido dependerá de varios factores, aunque el primero es la estructura y el modelo de los datos.

A continuación se muestran algunos de los tipos de archivos más utilizados en la publicación de datos:

- **Archivos PDF** (formato de documento portátil).[2] Es un formato no estructurado de almacenamiento para documentos digitales multiplataforma que pueden incorporar texto, imágenes vectoriales y mapas de bits.

(2)En inglés, *Portable Document Format*.

- **Archivos XLS o XSLX**. Es un formato estructurado propietario de Microsoft Office para la hoja de cálculo Excel utilizado en tareas financieras y contables.

- **Archivos de valores separados por comas** (CSV).[3] Es un tipo de documento estructurado en formato abierto que permite representar datos en forma de tabla en la que las columnas se separan por comas y las filas por saltos de línea. Existen variantes del mismo formato en las que las columnas se separan utilizando otros caracteres, por ejemplo, tabulados (TSV).[4]

(3)Acrónimo del inglés, *Comma-Separated Values*.

(4)Acrónimo del inglés, *Tab-Separated Values*.

- **Archivos XML**. Un archivo XML[5] es un tipo de documento semiestructurado compuesto por datos básicos, pero cuya definición no está determinada de antemano y dispone de etiquetas para describir su propia definición.

(5)Acrónimo del inglés, *eXtensible Markup Language*.

- **Archivos JSON**.[6] Es un estándar abierto basado en texto, diseñado para el intercambio de datos legible por humanos y que permite representar estructuras de objetos y listas.

(6)Acrónimo del inglés, *JavaScript Object Notation*.

- **Archivos RDF**.[7] Es una especificación que propone un modelo de datos para describir vocabularios y enlazar datos de distintos ámbitos. Los datos se relacionan mediante tripletas. Es el lenguaje utilizado para enlazar datos (*Linked Open Data*).

(7)Acrónimo del inglés, *Resource Description Framework*.

El tipo de archivo dependerá, en gran medida, del tipo de datos que necesitemos publicar. Por ejemplo, si deseamos publicar datos en formato de tabla, entonces lo aconsejable es emplear XLS o CSV. Por el contrario, si deseamos publicar datos con una cierta estructura flexible, la opción es XML o JSON. Obviamente se pueden utilizar distintos formatos para publicar los mismos datos, siendo siempre aconsejable utilizar formatos de ficheros abiertos.

3) Escoger una licencia abierta

El siguiente paso es elegir una licencia abierta para la publicación de dichos datos. Es importante seleccionar y utilizar una licencia que establezca de forma clara los usos posibles de los datos publicados.

En este contexto hay dos atributos de las licencias abiertas que son de especial importancia para seleccionar aquella licencia que mejor se adapte a las necesidades de publicación:

a) Atribución (BY, *Attribution*): indica que el conjunto de datos solo puede reutilizarse si se reconoce la autoría original en la nueva publicación.

b) Compartir igual (SA, *Share-Alike*): indica que el conjunto de datos solo puede reproducirse o reutilizarse como base para la creación de un nuevo conjunto de datos si también se hace bajo una licencia abierta.

Existen multitud de licencias aplicables a datos o conjuntos de datos. Es posible encontrar listados más detallados en la web de recomendaciones de Open Definition y en la guía de Open Data Commons.

4) Asegurar la accesibilidad

Para asegurar que los datos abiertos son realmente «abiertos», deben serlo desde un punto de vista legal y, además, lo tienen que ser desde un punto de vista técnico. Es decir, hay que facilitar que sean fácilmente accesibles y, preferiblemente, legibles por una máquina.

Hay muchas alternativas para hacer que los datos estén disponibles para otras organizaciones de forma rápida y eficiente. La forma más natural es la publicación en sus propios sitios web. Sin embargo, cuando el tamaño de los datos es extremadamente grande, la distribución a través de otros formatos puede presentar algunas ventajas.

A continuación veremos las formas de accesibilidad a datos abiertos más habituales:

- **A través del sitio web de la institución u organización**. Suele ser la opción más elemental y sencilla de implementar en muchos contextos. Generalmente los costes del almacenamiento de los datos y del tráfico generado por las descargas de los usuarios son muy bajos, por lo que esta es una opción muy interesante para la publicación de datos de un tamaño razonable.

- **A través del sitio web de terceros**. Existen repositorios de datos generalistas y también repositorios especializados en distintos campos. Los sitios web de terceros pueden ser muy útiles, dado que generalmente facilitan el acceso a una comunidad de personas interesadas y ponen en común diversos conjuntos de datos similares o complementarios. Además este tipo de plataformas proporciona: una infraestructura adecuada que puede soportar un volumen de descargas importante y ofrece análisis e información de utilización.

- **A través de las redes P2P**. Las redes punto-a-punto o entre iguales (P2P) son una alternativa eficiente para la distribución de volúmenes muy grandes de datos, ya que reparten los archivos entre la comunidad que accede a estos archivos.

- **A través de una API**.[8] Los datos pueden ser publicados mediante una API como las accesibles de Google, Twitter o Facebook, que ofrecen acceso a los datos a través de este tipo de interfaces. Las API permiten que los programadores seleccionen a qué datos se accede y suelen estar conectadas a bases de datos actualizadas en tiempo real. Lo que implica que el acceso a través de una API proporciona datos actualizados. El uso de API evita tener que generar y actualizar grandes archivos continuamente. Aunque también hay que tener en cuenta que es necesario el desarrollo del código de las API.

(8) Interfaz de programación de aplicaciones o en inglés *Application Programming Interface.*

- **A través de un punto de acceso SPARQL**. Con RDF se pueden representar datos y la relaciones entre ellos. SPARQL es un lenguaje de consulta propuesto por W3C (World Wide Web Consortium) que permite consultar datos en formato RDF. Los puntos de acceso SPARQL[9] son servicios web que permiten consultar un determinado conjunto de datos abiertos en formato RDF.

(9) *SPARQL End Points* en inglés.

5) Facilitar el descubrimiento de los datos

Es importante conseguir que los datos abiertos puedan ser encontrados por la comunidad de usuarios potenciales. Actualmente hay una serie de herramientas o sitios web diseñados expresamente para dar visibilidad a los datos abiertos.

La misma Open Knowledge Foundation nos ofrece dos herramientas que nos permiten dar visibilidad a los datos abiertos. Por un lado **CKAN** es una herramienta para la gestión y publicación de colecciones de datos. Esta herramienta ha sido utilizada por distintos gobiernos nacionales y locales, instituciones de investigación y otras organizaciones que recogen una gran cantidad de datos. Los usuarios, sean ciudadanos, desarrolladores, periodistas o investigadores, entre otros, pueden buscar datos, registrar conjuntos de datos publicados, crear y administrar grupos de conjuntos de datos, y obtener actualizaciones de bases de datos y de los grupos que resulten de interés.

En este contexto CKAN y SOCRATA son las principales soluciones adoptadas para catalogación de datos abiertos. CKAN es un software libre que permite crear portales de datos y proporciona también publicación, almacenamiento y gestión de conjuntos de datos. CKAN tiene una API con funcionalidad para previsualización de datos, creación de grafos y mapas, y búsquedas en datos georreferenciados. SOCRATA es una solución propietaria basada en la nube que permite crear visualizaciones de datos más complejas.

Por otro lado **DataHub.io** es una plataforma de código abierto para la gestión de datos de la Open Knowlegde Foundation e impulsada por CKAN. DataHub facilita que instituciones y organizaciones puedan publicar el material, pero también es posible agregar conjuntos de datos publicados en diferentes sitios web.

1.3. Buenas prácticas

Dentro de la web el W3C es el organismo encargado de velar por el desarrollo de estándares abiertos, libres e interoperables para la web. Esta organización ha elaborado una guía de publicación con pautas sobre cómo han de publicar datos los gobiernos. Igualmente existen otras iniciativas impulsoras de manuales de buenas prácticas o de concienciación acerca de los datos abiertos como las proporcionadas por ejemplo por la Sunlight Foundation o por la Open Knowledge Foundation.

Entre ellas destaca el *Decálogo Open Data*, que es un resumen de buenas prácticas a la hora de afrontar políticas de datos abiertos:

1) Armonización entre administraciones. Todos los puntos del decálogo se basan en la premisa de que debe existir una armonización entre todas las administraciones públicas. Todas las iniciativas de datos abiertos deben compartir los mismos principios y definiciones. Este punto es esencial para la interoperabilidad y aprovechamiento eficiente.

2) Publicar datos en formatos abiertos y estándares. Cualquier iniciativa de datos abiertos debería publicar sus conjuntos de datos en formatos abiertos (no-propietarios) y adecuados para permitir la reutilización de los mismos.

3) Usar esquemas y vocabularios consensuados. Además de los formatos abiertos, la estructura de los datos debería seguir convenios o esquemas definidos, si existieran. Los vocabularios o esquemas de representación de la información específicos deberían difundirse públicamente para poder interpretar correctamente la información.

4) Inventario en un catálogo de datos estructurado. Cualquier iniciativa de datos abiertos debe tener un punto de consulta en el que se incluya un inventario con información descriptiva y técnica sobre los conjuntos de datos que se exponen. Los metadatos que informan sobre cada conjunto de datos deberían seguir una estructura común y estándar.

5) Datos accesibles desde direcciones web persistentes y amigables. Tanto las fichas de los conjuntos de datos como la distribución de la propia información (archivos, API de consulta, etc.) deberían estar accesibles desde URL persistentes (PURL). Además deben seguir una estructura homogénea y bien definida, con información legible.

6) Exponer un conjunto mínimo de datos relativos al nivel de competencias del organismo y su estrategia de exposición de datos. Cada administración que impulse una iniciativa de datos debería crear una hoja de ruta en la que manifieste la estrategia de exposición de los conjuntos de datos y sus prioridades. Inicialmente debería publicar los conjuntos de mayor interés según las competencias del propio organismo.

7) Compromiso de servicio, actualización y calidad del dato, manteniendo un canal eficiente de comunicación para la reutilización. La administración debe mantener un mínimo de calidad y servicio en su iniciativa de datos abiertos, cumpliendo lo expuesto en la estrategia de publicación y comprometiéndose con su colectivo reutilizador.

8) Monitorizar y evaluar el uso y servicio mediante métricas. La administración debe crear métricas y evaluar sus indicadores de uso y servicio de la iniciativa de datos abiertos.

9) Datos bajo condiciones de uso no restrictivas y comunes. Las condiciones de uso deberían ser lo menos restrictivas posible y permitir la reutilización libre, incluso con fines comerciales.

10) Evangelizar y educar en el uso de datos. Es necesario educar en el uso de los datos, tanto a los colectivos de reutilización (sector TIC, periodismo, investigación, etc.) como a la sociedad en general y así fomentar el conocimiento y la inquietud por procesar información de una forma autónoma.

2. Ejemplos de publicación de datos abiertos

2.1. Administraciones locales

La mayoría de las administraciones o ayuntamientos de las grandes ciudades disponen de sitios web en los que se ofrecen datos abiertos.

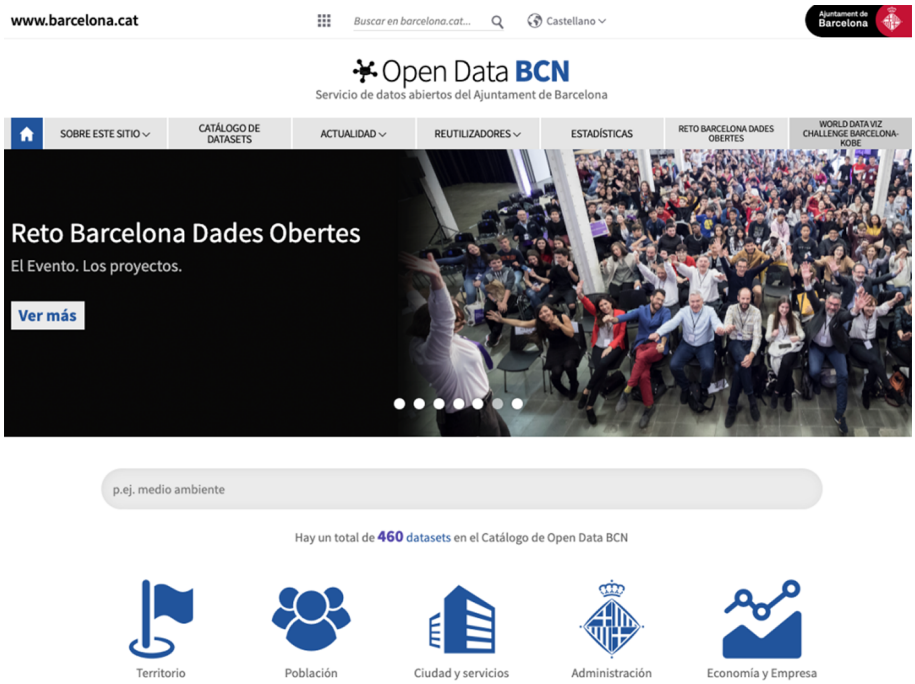
1) Ayuntamiento de Barcelona

El Ayuntamiento de Barcelona dispone del portal Open Data BCN de datos abiertos relacionados con la ciudad. Este catálogo tiene más de 460 conjuntos de datos abiertos, clasificados en distintas categorías, como administración, ciudad y servicios, economía y empresa, población y territorio. Todos los conjuntos de datos que se ofrecen en el servicio Open Data BCN indican qué licencia y condiciones de uso tienen.

Enlace de interés

Open Data BCN: opendata-ajuntament.barcelona.cat/es

Figura 1. Open Data BCN



Fuente: opendata-ajuntament.barcelona.cat/es

2) Ayuntamiento de Madrid

El Ayuntamiento de Madrid también dispone de un portal de datos abiertos que contiene más de 400 conjuntos de datos de muy diversas temáticas, como ciencia, comercio, tráfico, educación, empleo o energía.

Enlace de interés

Portal de datos abiertos del Ayuntamiento de Madrid: datos.madrid.es

Figura 2. Portal de datos abiertos del Ayuntamiento de Madrid

Fuente: http://datos.madrid.es/

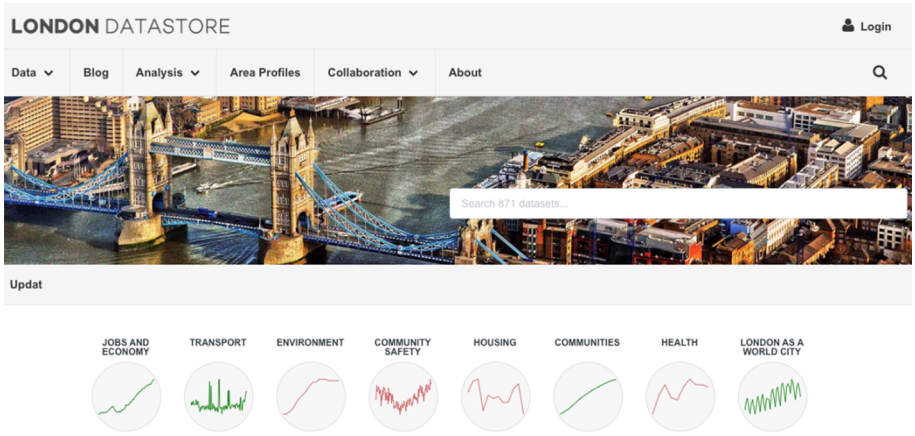
3) Ayuntamiento de Londres

En el portal de datos abiertos del Ayuntamiento de Londres (London Datastore) se encuentran más de 600 conjuntos de datos abiertos relacionados con multitud de categorías, como arte, cultura, crimen y seguridad, educación, medio ambiente, transparencia y transporte. Al igual que en los casos anteriores, los conjuntos de datos pueden descargarse en distintos formatos de datos.

Enlace de interés

London Datastore: data.london.gov.uk

Figura 3. London Datastore



Fuente: https://data.london.gov.uk/

2.2. Administraciones regionales

A continuación veremos algunos ejemplos de administraciones públicas de ámbito regional.

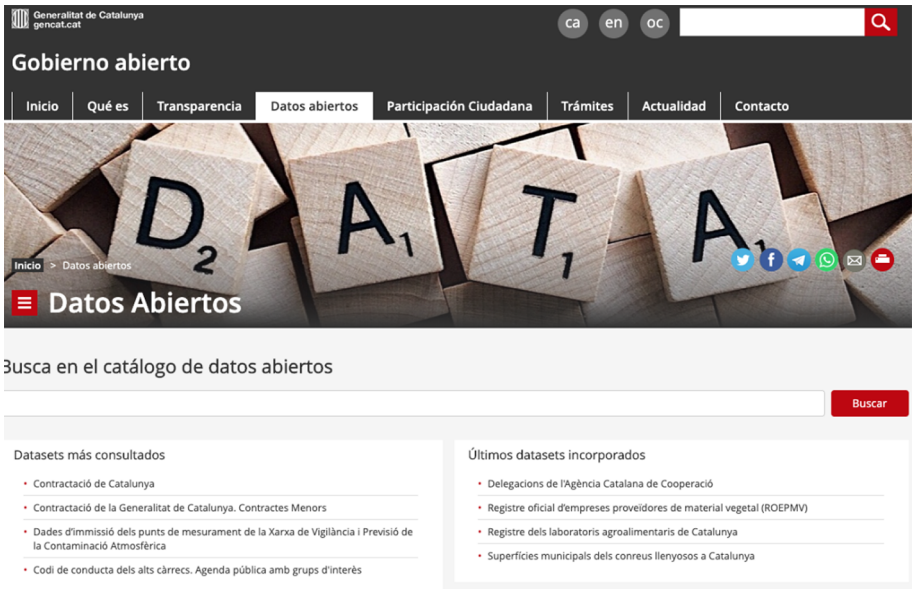
1) Generalitat de Catalunya

El portal de datos abiertos de la Generalitat de Catalunya ofrece conjuntos de datos abiertos del ámbito autonómico catalán. El propio portal ofrece una definición de datos abiertos y un catálogo de más de 570 conjuntos de datos abiertos que contempla multitud de temáticas, como demografía, territorio, urbanismo, agricultura o movilidad entre muchas otras.

Enlace de interés

Portal de datos abiertos de la Generalitat de Catalunya: governobert.gencat.cat/es/dades_obertes

Figura 4. Portal de datos abiertos de la Generalitat de Catalunya



Fuente: http://governobert.gencat.cat/es/dades_obertes/

2) Gobierno vasco

El portal Open Data Euskadi dispone de cerca de 5.000 conjuntos de datos abiertos con información relacionada con la Comunidad Autónoma del País Vasco. El sitio web ofrece una sección en la que se presentan distintas ideas y ejemplos de uso de los datos abiertos publicados en el mismo portal.

Enlace de interés

Open Data Euskadi: opendata.euskadi.eus/inicio

Figura 5. Open Data Euskadi

Fuente: http://opendata.euskadi.eus/inicio/

3) Junta de Andalucía

Como los anteriores, el portal de datos abiertos de la Junta de Andalucía contiene más de 500 conjuntos de datos sobre los temas de agricultura, educación, deporte, empleo, cartografía o salud.

Enlace de interés

Portal de datos abiertos de la Junta de Andalucía: www.juntadeandalucia.es/datosabiertos/portal.html

Figura 6. Portal de datos abiertos de la Junta de Andalucía

Fuente: https://www.juntadeandalucia.es/datosabiertos/portal.html

2.3. Administraciones estatales

1) Gobierno de España

La iniciativa de datos abiertos del Gobierno de España a través del portal datos.gob.es trata de facilitar la puesta a disposición de toda la información para aprovechar la reutilización de la información de España. Cuenta con más de 24.000 conjuntos de datos en varias categorías como sector público, economía o demografía. Accesible mediante archivos en diferentes formatos, una API y un punto de acceso SPARQL.

Enlace de interés

Datos abiertos del Gobierno de España: datos.gob.es

Figura 7. Datos abiertos del Gobierno de España

Fuente: https://datos.gob.es

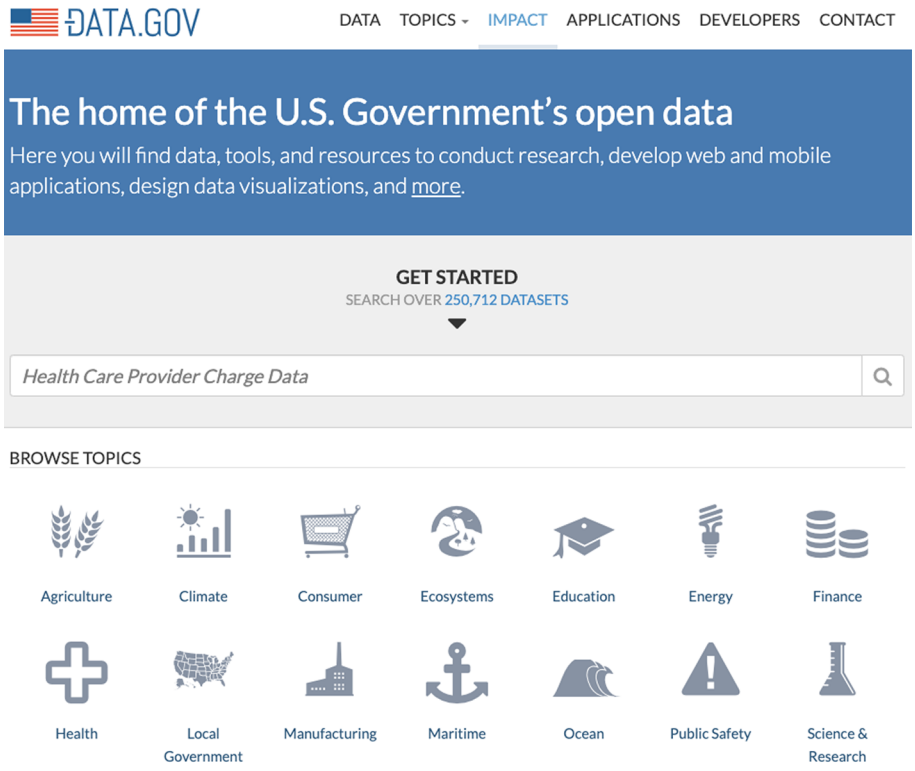
2) Estados Unidos de América

El portal de datos abiertos de Estados Unidos de América contiene actualmente más de 250.000 conjuntos de datos. Este portal engloba conjuntos de datos generados por las organizaciones públicas del país. Los datos pueden buscarse en función de su origen, de su categoría temática o de su contenido.

Enlace de interés

Portal de datos abiertos de Estados Unidos de América: www.data.gov

Figura 8. Portal de datos abiertos de Estados Unidos de América



Fuente: https://www.data.gov/

2.4. Unión Europea

La Unión Europea apuesta desde hace tiempo por una política de apertura de datos. En este sentido ha potenciado políticas que promuevan la publicación de datos abiertos en Europa. Actualmente el portal de datos abiertos de la Unión Europea ofrece más de 500.000 conjuntos de datos en abierto. Estos conjuntos de datos se extraen automáticamente de 73 portales web de datos abiertos pertenecientes al sector público (a nivel nacional y regional).

Enlace de interés

Portal europeo de datos: www.europeandataportal.eu/es/homepage

Figura 8. Portal europeo de datos

Fuente: https://www.europeandataportal.eu/es/homepage

Los datos pueden buscarse en función de su origen (país), de su idioma, de su categoría temática o de su contenido (mediante una búsqueda por palabras clave).

El portal pretende no solo ser un punto de acceso a datos en abierto, sino también fomentar una cultura más propensa al uso de datos abiertos. Para ello promueve la accesibilidad a los datos en abierto que ofrece, analiza el valor que aportan sus datos y proporciona información mediante cursos de aprendizaje en línea (*e-learning*), sobre qué son los datos abiertos y cómo usarlos.