Dades obertes

PID_00271442

Blas Torregrosa García

Temps mínim de dedicació recomanat: 2 hores

Universitat Oberta de Catalunya

Blas Torregrosa García

Enginyer en Informàtica i màster universitari en Seguretat de les Tecnologies de la Informació i de les Comunicacions (MISTIC) per la Universitat Oberta de Catalunya (UOC). Especialitzat en ciberseguretat. Professor col·laborador del màster de Ciència de Dades de la UOC i professor associat a la Universitat de Valladolid (UVA).

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Ferran Prados Carrasco (2020)

Primera edició: febrer 2020

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

Índex

Introducció

En aquest mòdul definirem el que s'entén per dades obertes (*open data*), encara que la definició no és única i, en principi, és independent de les dades enllaçades i de la web semàntica. També exposarem els beneficis que aporten les dades obertes i la seva publicació.

1. Què són les dades obertes?

La definició de dades obertes o *open data* no és única en l'actualitat. Hi ha diferents plantejaments amb lleugers matisos, encara que en essència no hi ha diferències importants entre elles.

L'organització Open Knowledge Foundation defineix les dades obertes com:

> Les **dades obertes** són dades que es poden usar, reutilitzar i redistribuir lliurement per qualsevol persona, i que estan subjectes, almenys, al requisit d'atribució i de compartir-se de la mateixa manera en què apareixen.

Els factors més importants relatius al concepte de dades obertes es resumeixen en els tres punts següents:

1) Disponibilitat i accés: la informació ha d'estar disponible com un tot i a un cost raonable de reproducció, preferiblement descarregant-la d'internet. A més, la informació ha d'estar disponible en una forma convenient i modificable.

2) Reutilització i redistribució: les dades s'han de proporcionar amb termes que permetin reutilitzar-les i redistribuir-les, i, fins i tot, integrar-les amb altres conjunts de dades.

3) Participació universal: tothom ha de poder utilitzar, reutilitzar i redistribuir la informació. No hi ha d'haver cap discriminació en termes d'esforç, persones o grups. No es permeten restricions «no comercials» que impedirien l'ús comercial de les dades o restriccions d'ús per a certs propòsits (per exemple, solament per a educació).

Això ens porta fins al concepte d'**interoperabilitat**.

> L'IEEE[1] defineix interoperabilitat com la capacitat de dos o més sistemes o components per intercanviar informació i utilitzar la informació intercanviada.

(1) Institut d'Enginyers Elèctrics i Electrònics, en anglès, *Institute of Electrical and Electronics Engineers*.

En aquest cas, és la possibilitat per interoperar o integrar diferents fonts de dades. La interoperabilitat és important perquè permet que diferents components puguin treballar junts i aquesta capacitat d'integrar components és essencial per construir sistemes més complexos i grans.

L'essència de les dades compartides és que una part del material obert pugui, a partir d'aquí, barrejar-se amb un altre material obert. Aquesta interoperabilitat és absolutament fonamental per entendre el principal benefici pràctic de l'obertura de dades: la capacitat de combinar diferents fonts de dades o conjunts de dades i, així, desenvolupar més i millors productes i serveis.

1.1. Beneficis de les dades obertes

En la societat actual hi ha multitud d'individus, organitzacions i, especialment, administracions públiques que generen i gestionen una gran quantitat i varietat de dades per dur a terme les seves tasques quotidianes. En aquest context, les administracions públiques (com ara els ajuntaments, els governs regionals o estatals) exerceixen un paper especialment important per la quantitat de dades que manegen, però també perquè una part considerable d'aquesta informació és oberta i es posa a la disposició de qualsevol individu o institució que desitgi utilitzar-la.

Tecnologies de la informació

Les tecnologies de la informació fan possible el desenvolupament de serveis que permeten respondre a qüestions sobre les dades que generen els organismes públics de forma automàtica. No obstant això, freqüentment aquestes dades no estan disponibles de manera que siguin senzilles d'utilitzar per poder tractar-les i obtenir coneixement.

Hi ha molts camps en què podem veure que les dades obertes han creat valor afegit a la societat. A manera d'exemple, podem enumerar algunes de les més rellevants:

- Transparència i control democràtic.
- Participació ciutadana.
- Creació de nous productes i serveis.
- Innovació.
- Millores en l'eficiència i eficàcia dels serveis oferts.
- Mesurament de l'impacte de les polítiques.

1.2. Publicació de dades obertes

Quan una organització o institució desitja publicar dades en obert s'ha de plantejar cinc passos principals que la conduiran a una correcta publicació de les dades en obert.

1) Identificació dels conjunts de dades

El primer pas en el procés de publicació de dades en obert és seleccionar el conjunt o els conjunts de dades que es proposi obrir. Aquest procés és iteratiu i es poden incloure nous conjunts de dades en el futur. En general, no hi ha

requisits per crear una llista completa de conjunts de dades que siguin candidates per a la seva publicació. Hi ha dos punts principals que s'haurien de tenir en compte:

a) En primer lloc, hem d'assegurar-nos que és viable publicar totes (o una part de) les dades.

b) En segon lloc, cal assegurar-se que no hi hagi dades personals o privades de persones individuals en el conjunt de dades que es desitgi publicar. Generalment, es publiquen conjunts de dades que no contenen dades de caràcter personal. En cas contrari, cal aplicar certs processos d'anonimització i protecció de la privadesa que garanteixin que les dades personals estaran correctament protegides en el conjunt de dades obertes.

Anonimització

D'*anonimitzar*: «Expressar una dada relativa a entitats o persones, eliminant la referència a la seva identitat» (RAE).

2) Selecció del format de les dades

En segon lloc, és important triar un format adequat per a la publicació de les dades obertes. El format triat dependrà de diversos factors, encara que el primer és l'estructura i el model de les dades.

A continuació, es mostren alguns dels tipus d'arxius més utilitzats en la publicació de dades:

- **Arxius PDF** (format de document portàtil).[2] És un format no estructurat d'emmagatzematge per a documents digitals multiplataforma que poden incorporar text, imatges vectorials i mapes de bits.

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

- **Arxius XLS o XSLX**. És un format estructurat propietari de Microsoft Office per al full de càlcul Excel utilitzat en tasques financeres i comptables.

- **Arxius de valors separats per comes** (CSV).[3] És un tipus de document estructurat en format obert que permet representar dades en forma de taula en què les columnes se separen per comes i les files per salts de línia. Hi ha variants del mateix format en què les columnes se separen utilitzant altres caràcters, per exemple, tabuladors (TSV).[4]

(3)Acrònim de l'anglès, *Comma-Separated Values*.

(4)Acrònim de l'anglès, *Tab-Separated Values*.

- **Arxius XML**. Un arxiu XML[5] és un tipus de document semiestructurat compost per dades bàsiques, però la definició de les quals no està determinada per endavant i disposa d'etiquetes per descriure la seva pròpia definició.

(5)Acrònim de l'anglès, *eXtensible Markup Language*.

- **Arxius JSON**.[6] És un estàndard obert basat en text, dissenyat per a l'intercanvi de dades llegible per humans i que permet representar estructures d'objectes i llistes.

(6)Acrònim de l'anglès, *JavaScript Object Notation*.

- **Arxius RDF**.[7] És una especificació que proposa un model de dades per descriure vocabularis i enllaçar dades de diferents àmbits. Les dades es relacionen mitjançant tripletes. És el llenguatge utilitzat per enllaçar dades (*Linked Open Data*).

(7)Acrònim de l'anglès, *Resource Description Framework*.

El tipus d'arxiu dependrà, en gran manera, del tipus de dades que necessitem publicar. Per exemple, si desitgem publicar dades en format de taula, llavors l'aconsellable és emprar XLS o CSV. Per contra, si desitgem publicar dades amb una certa estructura flexible, l'opció és XML o JSON. Òbviament, es poden utilitzar diferents formats per publicar les mateixes dades, sent sempre aconsellable utilitzar formats de fitxers oberts.

3) Escollir una llicència oberta

El pas següent és triar una llicència oberta per a la publicació d'aquestes dades. És important seleccionar i utilitzar una llicència que estableixi de forma clara els usos possibles de les dades publicades.

En aquest context, hi ha dos atributs de les llicències obertes que són d'especial importància per seleccionar aquella llicència que millor s'adapti a les necessitats de publicació:

a) Atribució (BY, *Attribution*): indica que el conjunt de dades solament es pot reutilitzar si es reconeix l'autoria original en la nova publicació.

b) Compartir igual (SA, *Share-Alike*): indica que el conjunt de dades solament es pot reproduir o reutilitzar com a base per a la creació d'un nou conjunt de dades si també es fa amb una llicència oberta.

Hi ha multitud de llicències aplicables a dades o conjunts de dades. És possible trobar llistats més detallats a la web de recomanacions de l'Open Definition i en la guia d'Open Data Commons.

4) Assegurar l'accessibilitat

Per assegurar que les dades obertes són realment «obertes», han de ser-ho des d'un punt de vista legal i, a més, ho han de ser des d'un punt de vista tècnic. És a dir, cal facilitar que siguin fàcilment accessibles i, preferiblement, llegibles per una màquina.

Hi ha moltes alternatives per fer que les dades estiguin disponibles per a altres organitzacions de forma ràpida i eficient. La forma més natural és la publicació en els seus propis llocs web. No obstant això, quan la grandària de les dades és extremadament gran, la distribució per mitjà d'altres formats pot presentar alguns avantatges.

A continuació, veurem les formes d'accessibilitat a dades obertes més habituals:

- **Per mitjà del lloc web de la institució o organització**. Sol ser l'opció més elemental i senzilla d'implementar en molts contextos. Generalment els costos de l'emmagatzematge de les dades i del tràfic generat per les descàrregues dels usuaris són molt baixos, per la qual cosa aquesta és una opció molt interessant per a la publicació de dades d'una grandària raonable.

- **Per mitjà del lloc web de tercers**. Hi ha repositoris de dades generalistes i també repositoris especialitzats en diferents camps. Els llocs web de tercers poden ser molt útils, atès que generalment faciliten l'accés a una comunitat de persones interessades i posen en comú diversos conjunts de dades similars o complementaris. A més, aquest tipus de plataformes proporciona una infraestructura adequada que pot suportar un volum de descàrregues important i ofereix anàlisis i informació d'utilització.

- **Per mitjà de les xarxes P2P**. Les xarxes punt-a-punt o entre iguals (P2P) són una alternativa eficient per a la distribució de volums molt grans de dades, ja que reparteixen els arxius entre la comunitat que accedeix a aquests arxius.

- **Per mitjà d'una API**.[8] Les dades poden ser publicades mitjançant una API com les accessibles de Google, Twitter o Facebook, que ofereixen accés a les dades per mitjà d'aquest tipus d'interfícies. Les API permeten que els programadors seleccionin a quines dades s'accedeix i solen estar connectades a bases de dades actualitzades en temps real, la qual cosa implica que l'accés per mitjà d'una API proporciona dades actualitzades. L'ús de l'API evita haver de generar i actualitzar grans arxius contínuament. Encara que també s'ha de tenir en compte que cal desenvolupar el codi de les API.

- **Per mitjà d'un punt d'accés SPARQL**. Amb RDF es poden representar les dades i la relacions entre elles. SPARQL és un llenguatge de consulta proposat per W3C (World Wide Web Consortium) que permet consultar dades en format RDF. Els punts d'accés SPARQL[9] són serveis web que permeten consultar un determinat conjunt de dades obertes en format RDF.

(8) Interfície de programació d'aplicacions o en anglès *Application Programming Interface.*

(9) *SPARQL End Points* en anglès.

5) Facilitar el descobriment de les dades

És important aconseguir que les dades obertes puguin ser trobades per la comunitat d'usuaris potencials. Actualment hi ha una sèrie d'eines o llocs web dissenyats expressament per donar visibilitat a les dades obertes.

La mateixa Open Knowledge Foundation ens ofereix dues eines que ens permeten donar visibilitat a les dades obertes. D'una banda, **CKAN** és una eina per a la gestió i publicació de col·leccions de dades. Aquesta eina ha estat uti-

litzada per diferents governs nacionals i locals, institucions de recerca i altres organitzacions que recullen una gran quantitat de dades. Els usuaris, siguin ciutadans, desenvolupadors, periodistes o investigadors, entre d'altres, poden buscar dades, registrar conjunts de dades publicades, crear i administrar grups de conjunts de dades, i obtenir actualitzacions de bases de dades i dels grups que resultin d'interès.

En aquest context, CKAN i SOCRATA són les principals solucions adoptades per a la catalogació de les dades obertes. CKAN és un programari lliure que permet crear portals de dades i també proporciona publicació, emmagatzematge i gestió de conjunts de dades. CKAN té una API amb funcionalitat per a la previsualització de les dades, creació de grafs i mapes, i cerques en dades georeferenciades. SOCRATA és una solució propietària basada en el núvol que permet crear visualitzacions de dades més complexes.

D'altra banda, **Datahub.io** és una plataforma de codi obert per a la gestió de dades de l'Open Knowlegde Foundation i impulsada per CKAN. DataHub facilita que les institucions i les organitzacions puguin publicar el material, però també és possible agregar conjunts de dades publicades a diferents llocs web.

1.3. Bones pràctiques

Dins de la web, el W3C és l'organisme encarregat de vetllar pel desenvolupament d'estàndards oberts, lliures i interoperables per a la web. Aquesta organització ha elaborat una guia de publicació amb pautes sobre com han de publicar les dades els governs. Igualment, hi ha altres iniciatives impulsores de manuals de bones pràctiques o de conscienciació sobre les dades obertes com les proporcionades, per exemple, per la Sunlight Foundation o per l'Open Knowledge Foundation.

Entre elles destaca el *Decàleg Open Data*, que és un resum de bones pràctiques a l'hora d'afrontar polítiques de dades obertes:

1) Harmonització entre les administracions. Tots els punts del decàleg es basen en la premissa que hi ha d'haver una harmonització entre totes les administracions públiques. Totes les iniciatives de dades obertes han de compartir els mateixos principis i definicions. Aquest punt és essencial per a la interoperabilitat i aprofitament eficient.

2) Publicar dades en formats oberts i estàndards. Qualsevol iniciativa de dades obertes hauria de publicar els seus conjunts de dades en formats oberts (no-propietaris) i adequats per permetre la seva reutilització.

3) Usar esquemes i vocabularis consensuats. A més dels formats oberts, l'estructura de les dades hauria de seguir convenis o esquemes definits, si existissin. Els vocabularis o esquemes específics de representació de la informació s'haurien de difondre públicament per poder interpretar correctament la informació.

4) Inventari en un catàleg estructurat de dades. Qualsevol iniciativa de dades obertes ha de tenir un punt de consulta en què s'inclogui un inventari amb informació descriptiva i tècnica sobre els conjunts de dades que s'exposen. Les metadades que informen sobre cada conjunt de dades haurien de seguir una estructura comuna i estàndard.

5) Dades accessibles des d'adreces web persistents i amigables. Tant les fitxes dels conjunts de dades com la distribució de la pròpia informació (arxius, API de consulta, etc.) haurien d'estar accessibles des d'URL persistents (PURL). A més, han de seguir una estructura homogènia i ben definida, amb informació llegible.

6) Exposar un conjunt mínim de dades relatives al nivell de competències de l'organisme i la seva estratègia d'exposició de dades. Cada administració que impulsi una iniciativa de dades hauria de crear un full de ruta en què manifesti l'estratègia d'exposició dels conjunts de dades i les seves prioritats. Inicialment, hauria de publicar els conjunts de major interès segons les competències del propi organisme.

7) Compromís de servei, actualització i qualitat de les dades, mantenint un canal eficient de comunicació per a la reutilització. L'administració ha de mantenir un mínim de qualitat i servei en la seva iniciativa de dades obertes, complint l'exposat en l'estratègia de publicació i comprometent-se amb el seu col·lectiu reutilitzador.

8) Monitoritzar i avaluar l'ús i servei mitjançant mètriques. L'administració ha de crear mètriques i avaluar els seus indicadors d'ús i servei de la iniciativa de dades obertes.

9) Dades amb condicions d'ús no restrictives i comunes. Les condicions d'ús haurien de ser com menys restrictives millor i permetre la reutilització lliure, fins i tot amb finalitats comercials.

10) Evangelitzar i educar en l'ús de dades. Cal educar en l'ús de les dades, tant als col·lectius de reutilització (sector TIC, periodisme, recerca, etc.) com a la societat en general i, així, fomentar el coneixement i la inquietud per processar informació d'una forma autònoma.

2. Exemples de publicació de dades obertes

2.1. Administracions locals

La majoria de les administracions o ajuntaments de les grans ciutats disposen de llocs web en què ofereixen dades obertes.

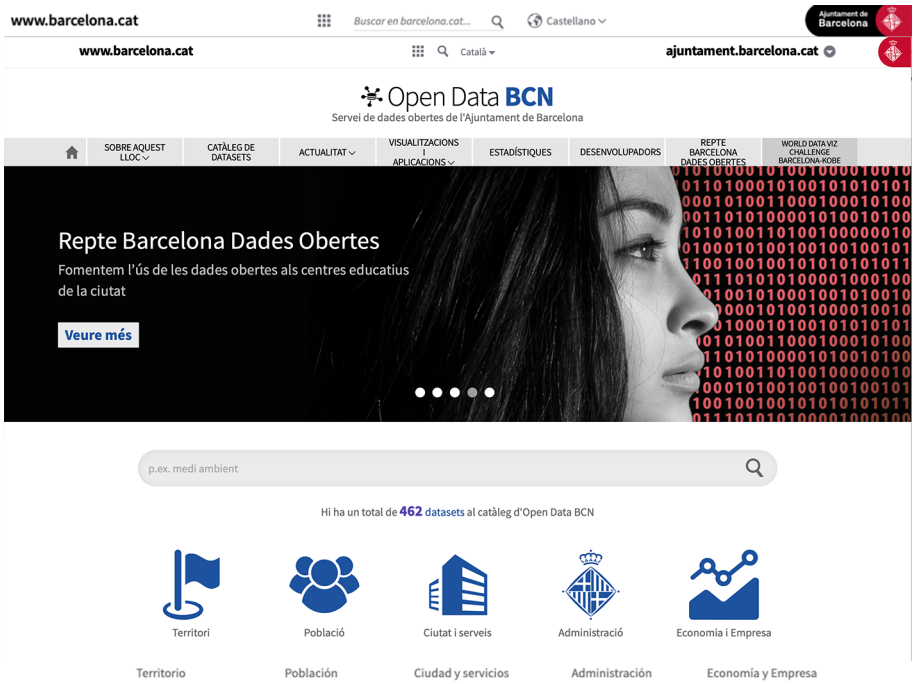
1) Ajuntament de Barcelona

L'Ajuntament de Barcelona disposa del portal Open Data BCN de dades obertes relacionades amb la ciutat. Aquest catàleg té més de 460 conjunts de dades obertes, classificades en diferents categories, com ara administració, ciutat i serveis, economia i empresa, població i territori. Tots els conjunts de dades que s'ofereixen en el servei Open Data BCN indiquen quina llicència i condicions d'ús tenen.

Enllaç d'interès

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

Figura 1. Open Data BCN



Font: opendata-ajuntament.barcelona.cat/ca

2) Ajuntament de Madrid

L'Ajuntament de Madrid també disposa d'un portal de dades obertes que conté més de 400 conjunts de dades de molt diverses temàtiques, com ara ciència, comerç, tràfic, educació, ocupació o energia.

Enllaç d'interès

Portal de dades obertes de l'Ajuntament de Madrid: datos.madrid.es

Figura 2. Portal de dades obertes de l'Ajuntament de Madrid

Font: http://datos.madrid.es/

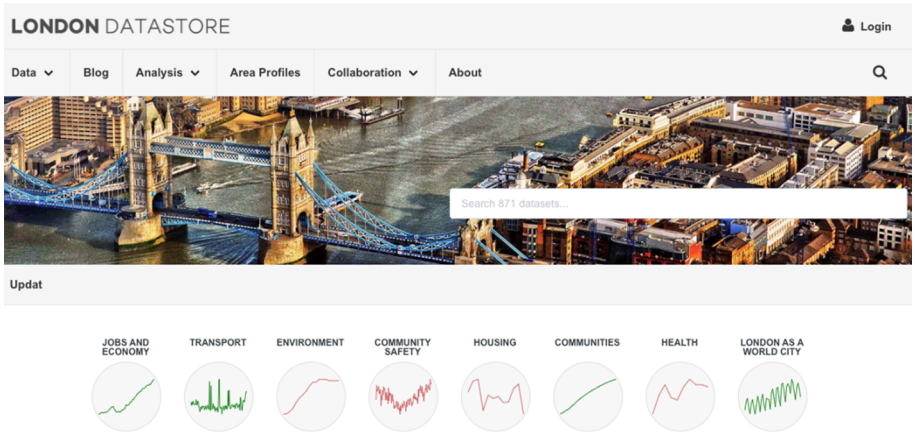
3) Ajuntament de Londres

Al portal de dades obertes de l'Ajuntament de Londres (London Datastore) trobem més de 600 conjunts de dades obertes relacionades amb multitud de categories, com ara art, cultura, crim i seguretat, educació, medi ambient, transparència i transport. Igual que en els casos anteriors, els conjunts de dades es poden descarregar en diferents formats de dades.

Enllaç d'interès

London Datastore: data.london.gov.uk

Figura 3. London Datastore



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

2.2. Administracions regionals

A continuació, veurem alguns exemples d'administracions públiques d'àmbit regional.

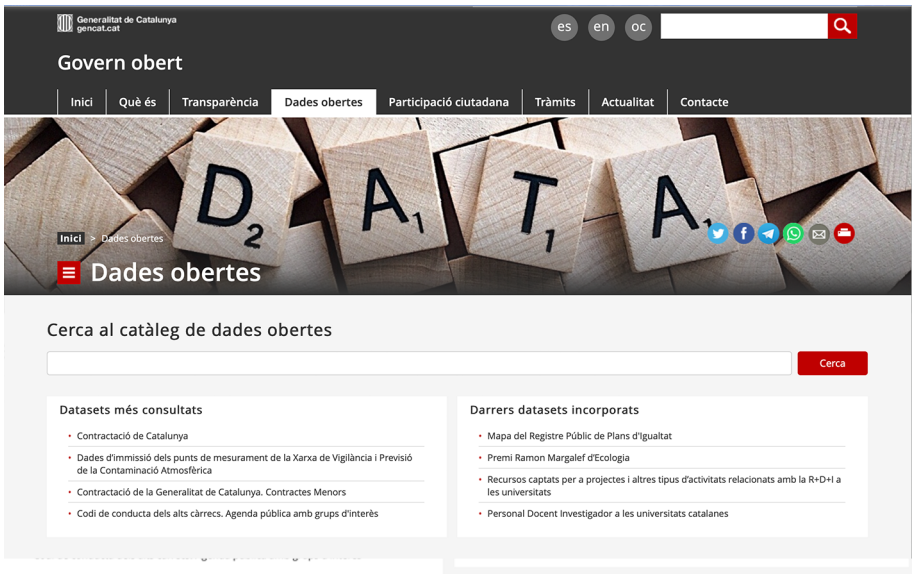
1) Generalitat de Catalunya

El portal de dades obertes de la Generalitat de Catalunya ofereix conjunts de dades obertes de l'àmbit autonòmic català. Aquest portal ofereix una definició de dades obertes i un catàleg de més de 570 conjunts de dades obertes que té en compte multitud de temàtiques, com ara demografia, territori, urbanisme, agricultura o mobilitat entre moltes altres.

Enllaç d'interès

Portal de dades obertes de la Generalitat de Catalunya: governobert.gencat.cat/es/dades_obertes

Figura 4. Portal de dades obertes de la Generalitat de Catalunya



Font: http://governobert.gencat.cat/ca/dades_obertes/

2) Govern basc

El portal Open Data Euskadi disposa de prop de 5.000 conjunts de dades obertes amb informació relacionada amb la Comunitat Autònoma del País Basc. El lloc web ofereix una secció en què es presenten diferents idees i exemples d'ús de les dades obertes publicades al mateix portal.

Enllaç d'interès

Open Data Euskadi: opendata.euskadi.eus/inicio

Figura 5. Open Data Euskadi

Font: http://opendata.euskadi.eus/inici/

3) Junta d'Andalusia

Com els anteriors, el portal de dades obertes de la Junta d'Andalusia conté més de 500 conjunts de dades sobre temes d'agricultura, educació, esport, ocupació, cartografia o salut.

Enllaç d'interès

Portal de dades obertes de la Junta d'Andalusia: www.juntadeandalucia.es/datosabiertos/portal.html

Figura 6. Portal de dades obertes de la Junta d'Andalusia

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

2.3. Administracions estatals

1) Govern d'Espanya

La iniciativa de dades obertes del Govern d'Espanya per mitjà del portal datos.gob.es tracta de facilitar la posada a disposició de tota la informació per aprofitar la reutilització de la informació d'Espanya. Compta amb més de 24.000 conjunts de dades en diverses categories com ara el sector públic, economia o demografia. Accessible mitjançant arxius en diferents formats, una API i un punt d'accés SPARQL.

Enllaç d'interès

Dades obertes del Govern d'Espanya: datos.gob.es

Figura 7. Dades obertes del Govern d'Espanya

Font: https://datos.gob.es

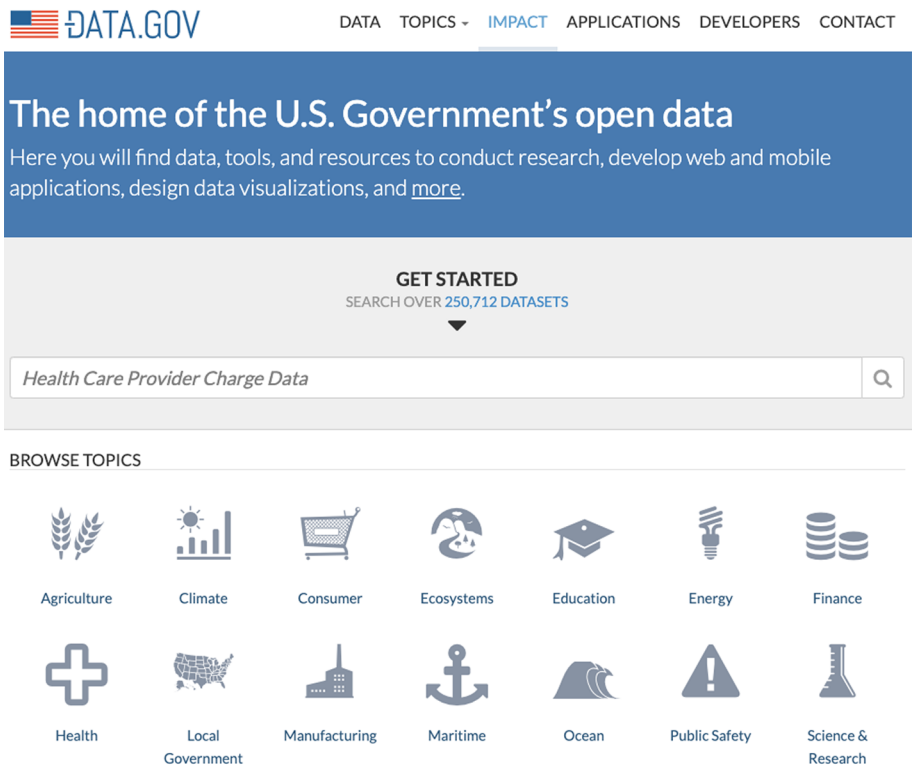
2) Estats Units d'Amèrica

El portal de dades obertes dels Estats Units d'Amèrica conté actualment més de 250.000 conjunts de dades. Aquest portal engloba conjunts de dades generades per les organitzacions públiques del país. Les dades es poden buscar en funció del seu origen, de la seva categoria temàtica o del seu contingut.

Enllaç d'interès

Portal de dades obertes dels Estats Units d'Amèrica: www.data.gov

Figura 8. Portal de dades obertes dels Estats Units d'Amèrica



Font: https://www.data.gov

2.4. Unió Europea

La Unió Europea aposta, des de fa temps, per una política d'obertura de dades. En aquest sentit ha potenciat polítiques que promoguin la publicació de dades obertes a Europa. Actualment el portal de dades obertes de la Unió Europea ofereix més de 500.000 conjunts de dades en obert. Aquests conjunts de dades s'extreuen automàticament de 73 portals web de dades obertes que pertanyen al sector públic (en l'àmbit nacional i regional).

Enllaç d'interès

Portal europeu de dades: www.europeandataportal.eu/es/homepage

Figura 8. Portal europeu de dades

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

Les dades es poden buscar en funció del seu origen (país), del seu idioma, de la seva categoria temàtica o del seu contingut (mitjançant una cerca per paraules clau).

El portal pretén no solament ser un punt d'accés a les dades en obert, sinó també fomentar una cultura més propensa a l'ús de dades obertes. Per a això, promou l'accessibilitat a les dades en obert que ofereix, analitza el valor que aporten les seves dades i proporciona informació mitjançant cursos d'*e-learning*, sobre què són les dades obertes i com usar-les.