Case Study:

European Data Portal

Type: Website

Organisation(s): European Commission, Capgemini, , Fraunhofer Fokus, the Open Data Institute

Tags: open data, process, standards, website

The European Data Portal (EDP) was created to harvest the metadata of public sector information from open data portals in European countries – including European Union member states and the EFTA countries. The website provides a single access point for all data, using APIs to link to other existing open data portals.

The portal is very large, providing acces to a total of 1,108,438 datasets from across the continent. This comes with both advantages and challenges, but there is a significant variation in levels of engagement between member states.

The team behind the EDP also acts as a focal point for activity in the field of open data across Europe, producing reports, analysis, and rankings which help encourage member states and individual organisations to release more open data and to use it more effectively.

Background

The European Data Portal has its origins in the EU's strategy towards public sector transparency. In 2003, the EU issued the Public Sector Information Directive (2003/98/EC), which secured equal treatment of potential re-users where public sector bodies release information. In 2013, the Directive was amended (2013/37/EU) to make all information accessible under Member State legislation re-usable. The amendment also stated that administrative charges should not exceed the actual cost of information release.

The European Data Portal was launched in 2015, three years after the separate EU Open Data portal was founded. Whereas the EUDP is a catalogue of datasets from European Union institutions, the EDP is larger in scope. It involves the harvesting of data held by the EU Member States and several other affiliated countries (including the EFTA, EEA, and some Eastern Partnership countries) within a single portal. It also harvests data directly from the EU Open Data Portal.

GREATER MANCHESTER
DOING THINGS DIFFERENTLY

The European Commission tasked the consultancy [Capgemini](#) with building the platform at a total cost of €6.5 million. Several organisations active in the open data field also contributed to the project, including the [Open Data Institute](#) and [Fraunhofer FOKUS.](#)

Both the EDP and EUDP are maintained by the Luxembourg-based [Publications Office of the European Union](#) (POEU). POEU is also responsible for other open data sites such as [EURlex](#), a library of European Union legislation; [CORDIS](#) which aggregates the results of EU-supported research across the continent; [EU Whoiswho](#), a directory of EU institutions and agencies; and [TED](#) (tenders electronic daily), which aggregates all open public procurement tenders across the European Union.

The EU also launched a [portal dedicated to the COVID-19 pandemic](#), which shares data and research on the disease and its spread to support the global effort against it.

Open data opportunities

At the time of launch, the European Commission (EC) identified many [opportunities around the release of open data](#), including increased efficiency of public services, growth of the knowledge economy, and reduced costs of goods and services. The EC also estimated that the total market size for open data in the then EU28+ would be around €280 billion, supporting up to 100,000 jobs, most prominently in government administration but also in other sectors. These include transport, logistics, professional services, and real estate. Other predicted benefits include better-targeted services and reduced road fatalities.

Important considerations

Metadata quality

When data owners register to have their datasets appear on the portal, they must go through a checklist to ensure the dataset meets the standards of the portal and can be easily categorised. These include asking that data publishers provide metadata and use standard date and time products. Publishers are also asked to provide information on:

- How often data should be harvested
- Metadata standards used
- API used and whether authentification required

A [Metadata Quality Assurance Framework](#) has been established as a set of guidelines to ensure high and consistent metadata standards across datasets. A series of measures are applied as shown below, to determine the overall quality rating for each dataset:

- **Interoperability** - such as machine readability, format availability, and proprietary status;
- **Findability** - such as availability of temporal and spatial information, keywords, and category;
- **Contextuality** - such as availability of date issued and modified, as well as file size and rights associated with dataset;
- **Accessibility** - such as availability of download URL, and most frequent download statistics; and
- **Reusability** - such as availability of a contact point and licence.

Very few datasets achieve a score of 'Good', showing that there is a great deal of work to be done to roll out consistent standards across multiple states.

Content and quality

It is easy to gain a quick overview of the types datasets contained in the EDP using statistics section that shows the makeup up the data portal by category and country of the publisher. This allows for easy benchmarking and comparisons.

Some participant countries are much better represented in the data portal than others. The portal is linked to 341,388 datasets hosted by the Czech Republic alone, and 210,854 by Germany. At the bottom of the scale, Hungary only hosts 57 datasets and Liechtenstein 79. Participating in the EDP requires data owners to actively engage and to register with the portal for the metadata to be harvested.

The most populated dataset categories meanwhile are 'Justice, legal system and public safety' with 121,560 datasets. The second-largest category is 'Regions and cities' with 83,707, which includes many datasets containing neighbourhood and regional boundaries as well as planning-related data. Many datasets belong to two or more categories

Some of the datasets are very niche and only of local significance, such as one dataset showing the locations of public water fountains in Trier, Germany. It is sometimes possible to access other related datasets by clicking on 'similar datasets', but in many cases this function is not available.

Data access shortcuts

There are alternative ways by which users can access datasets without manually filtering and selecting them. One data access option is to make SPARQL queries of linked datasets through the Virtuoso SPARQL query engine. Users can also connect to either an RSS feed or an ATOM feed, both of which allow users to automatically keep track of updates in a machine-readable format. All of these methods are more efficient than filtering through categories to find one of over a million datasets.

Community engagement

The EDP team is very active on [Twitter](#) and shares the latest articles and insights from the portal several times a day. There are also over a thousand [news stories](#) that have been published on the portal which discuss new datasets, reports, and research projects which are happening elsewhere in the world of open data. There is also an [in-built contact form](#) allowing users to make suggestions or follow up on any concerns.

Impact and Studies

One of the most useful features of the EDP is its analytics pages which give an overview of the state of open data in Europe. These include studies, use cases, and [country insights](#) for each participating country to help users understand the potential of open data.

For instance, for the UK the examples given include [Local Government Inform](#), which is hosted by the Local Government Association and allows users to compare their area with others. For Germany, one example is [Politicindex](#), which is an open-source platform for political information to keep voters informed. All case studies provided are regarding open data held at a national level and do not relate to the EDPdirectly.

This section also includes reports on the state of open data in Europe, including annual [data maturity reports](#) that rank countries by data maturity scores and explores where more needs to be done. Ireland and France are at the top of this ranking in the 2019 report, whilst the United Kingdom is just below the EU28+ average. These rankings are formed by aggregating scores on data policies, open data portals, data quality, and real-world impact.

High-value datasets report

Another useful report that the EDP has produced is one about [high-value datasets (HVDs)](#). Published in January 2020, this report sets out six key recommendations for open data publishers. These were written with international data programmes in mind, incorporating lessons learned from the EDP, but many are also applicable to local data projects. These are, in summary:

1. Create incentives, like additional resources, for data providers to foster their active engagement in identifiying potential high-value datasets.

2. Set clear expectations around roles, responsibilities, and resources relevant to data providers.

3. Standardise HVD assessment and specifications across borders.

4. Provide expert guidance that supports a consistent process and is aware of differences in language, culture, politics, and perceptions of impact.

GREATER
MANCHESTER
DOING THINGS DIFFERENTLY

5. Work in iterative rounds to allow incremental progress and different stakeholders to reach alignment and mutual consent.

6. The views of those with sector / subject-specific user experience must be consulted to capture the potential of HVDs.

The economic impact of open data

The EDP team also produced another [report and literature review on the economic impact of data](#) which helps to capture some of the wider impacts of open data. Key conclusions include:

- The economic impact of open data has turned out to be incremental, not immediate.
- The benefits of open data are often intangible, including improved productivity, wellbeing, safety, and sustainability.
- Open data is, like transport infrastructure, an economic enabler and many of the benefits are felt downstream from those who actively interact with open data.
- This makes it necessary to capture market size by also measuring the products and services that are improved or enabled through open data. The consultancy McKinsey referred to ['causal value chains'.](#)
- The key sectors identified as benefitting most from open data are public administration; professional, scientific and technical; information & communication; and transport & storage.

Blockers and challenges

The scale of the portal, and the fact that there is no centrally-controlled publication protocol, is both a strength and a weakness of the European Data Portal. This has the advantage of allowing publishers across the continent to easily share their datasets. However, it also means that there are so many datasets on the portal it becomes difficult to search through them effectively and gain an overview of what is available.

Despite quality assurance provisions, including the [Metadata Quality Assurance Framework](#), there is still a variation in data maturity between organisations that publish on the EDP. Licensing is inconsistent, with an estimated [90 different types of data licenses](#) across national, regional, and municipal governments in the European Union.

There have been further [efforts to improve interoperability](#) between datasets through the [European Interoperability Framework](#), but this has not yet produced substantial gains. As part of these efforts, the [DCAT Application Profile](#) provides a set of metadata and interoperability standards to help mitigate this problem, but it has not

been fully implemented in every participating country. The EDP also hosts a [Licensing Assistant](#) to help publishers understand how to navigate licensing options.

However, the EDP data harvesting process is automatic and so is not able to discriminate between datasets based on metadata, contents, or file type. This reflects the priority of making the EDP open to data from all affiliated countries.

The fact that metadata is often only available only in the original language also makes it particularly difficult to compare equivalent datasets across national boundaries. There is an ongoing translation programme of this metadata using the machine-learning translation program [MT@EC](#), but this is still [prone to errors.](#)

What can Greater Manchester take from this?

- As EDP research has shown, it is difficult to measure the value of open data since the benefits are often downstream, spread widely, and intangible. Nevertheless, the value of open data is iterative, and comes from its use rather than its storage.
- When publishing high-value datasets, it is sensible to take a gradual, iterative approach to allow alignment between partners and to ensure lessons can be learnt along the way rather than publishing everything at once.
- Just as EU 28+ states display varying levels of engagement with the EDP, any Greater Manchester open data programme may experience similar variations. It is important to run consultation exercises to identify reasons for these differences and then offer tailored support to try to boost engagement.
- When a data portal reaches a certain size, it is useful to publish a dashboard with statistics that trends in numbers of databases and their makeup by category and publishing organisation. This helps data users to make sense of the data portal as a whole and identify what is available.
- A local data team needs to strike a balance as early as possible between size and quality. Permitting different licenses and standards whilst aiming at interoperability may be politically necessary at a European level, but a 'halfway house' approach to applying standards is generally not ideal. This issue may be less likely within Greater Manchester, given the consistent Open Government Licence standard for licensing data in the UK.
- A comprehensive checklist like the EU Metadata Quality Assurance Framework is an important starting point to ensure consistency and searchability of datasets. At a more local scale, a more pragmatic and nuanced approach may also benefit.
- Publishing detailed articles or blog posts with the latest news on an open data programme may seem costly and unimportant but they help to publicise new data releases and provide a chronology of a data portal's development. This helps users gain an easy overview of what is on offer. It is also important to

consider the creation of a communications strategy, providing both longer term information and instant updates (i.e. using both blogs and Twitter).

Find out more:

[European Data Portal website](#)

[European Commission - Open Data](#)

[DCAT Application Profile for data portals in Europe](#)

[European Union Open Data Portal](#)

GREATER MANCHESTER
DOING THINGS DIFFERENTLY