Value-based prioritisation of Open Government Data investments

Leda Bargiotti, Michiel De Keyzer, Stijn Goedertier, Nikolaos Loutas
PwC EU Services, Belgium
firstname.lastname@be.pwc.com

Abstract

Governments increasingly prioritise their investments in Open Government Data on the basis of the value that can be unlocked by opening up government datasets. For example, the G8 Member States, including the EU, committed to the opening up and publishing *high-value* datasets with priority. This was formalised in the "G8 Open Data Charter" and the individual action plans of the G8 Member States and the EU. In the context of Action 1.1 of the Interoperability Solutions for European Public Administrations (ISA) Programme of the European Commission, we elaborated on a working definition for high-value datasets through different dimensions, both from the perspective of the data publisher and data re-user. We used this working definition for identifying and prioritising datasets owned by European Institutions to be listed on the European Union Open Data Portal (EU ODP). This will allow EU institutions to determine which new datasets to be published with priority or for which high-value datasets already listed on the EU ODP, the reusability should be improved with priority.

Keywords: Open Data, PSI, European Union, high-value dataset

Introduction

Open Government is supported by the availability of high-value Open Government Data. Open Government Data refers to data available under an open licence produced or commissioned by governments or government controlled entities which can be freely used, re-used and redistributed by anyone. Open data is instrumental to reach a variety of objectives both from a data publisher and a data re-user perspective. Theoretically the more data are made available, the better. But given the limited resources that publishers have at their disposal, one of the objectives of this activity was to identify high-value datasets to be published with higher priority.

The publication of high-value datasets as a priority for opening up data by governments has, for example, been set as a priority by the G8 and all Member States and the EU committed to the opening up and publishing of such datasets. In the "G8 Open Data Charter" and the individual action plans of the different G8 Member States, the opening of datasets is seen as a key enabler for transparency.

For publishers of datasets to be able to properly prioritise based on the value of datasets, it is important to have a common understanding on what can be considered as high-value datasets. In the context of Action 1.1[1] of the ISA Programme, we elaborated on a working definition for high-value datasets. For this we looked at different perspectives. Of course the view of the data publisher was taken into account. But also, since the value of a dataset lies importantly in its reuse, the perspective of the re-user has been taken into account.

[1] This paper reports on work that was funded in the context of Action 1.1 of the Interoperability Solutions for European Public Administrations (ISA) Programme of the European Commission. Specific acknowledgement is due to: Valentina Frato, Norbert Hohn, Athanasios Karalopoulos, Vassilios Peristeras and Agnieszka Zajac.

In order to support the Commission's Open Data policy, metadata describing datasets published by European Institutions is collected and made available through the European Union Open Data Portal (EU ODP). As a result, users can access datasets produced by the European Institutions and bodies through a single point of access. As part of our work, we provided support by identifying new datasets by applying the definition and building a list of high-value datasets that should be listed on the EU ODP as a priority.

Defining high-value datasets

For defining high-value datasets we first had a look at existing related work and studies. These were identified through our work in the field of (Linked) Open Government Data and Semantic Interoperability and in our collaboration with the European Commission and EU Member States. Additionally, we investigated which apps are currently being developed and what the underlying data (would) be. Also, a survey on social media has been launched, asking the re-users and app community which datasets they would like see opened up. Finally, we also looked at other indicators for instance from the Open Data Request Map, an interactive dashboard showing requests for data since the launch of the data request mechanism on data.gov.uk.

Title	Description
ES - Characterization Study of the Infomediary Sector	The Characterization Study of the Infomediary Sector identified that the following types of datasets are reused the most by business in Spain: • Geographic and cartographic information • Business and financial information • Social-demographic and statistical information • Legal information
DK – Good Basic Data for Everyone	The publication of high-value government data in Denmark is approached through the concept of "Basic Data" and comprises, amongst others, core information about individuals, businesses, real properties, buildings, addresses... These kind of data make a lot of sense of being publically available for re-use because it increases governmental efficiency and it enables the private sector to develop new products and services.
G8 Open Data Charter	The G8 Open Data Charter defines high-value data as data that improve democracy and encourage the innovative re-use of the particular data. The charter mentions the following high-value data domains: Companies (e.g. company registers), Crime and Justice (e.g. crime statistics), Earth observation (e.g. meteorological observations), Education (e.g. list of schools and performance indicators), Energy and Environment (e.g. pollution levels), Finance and contracts (e.g. calls for tenders and budget/spending data), Geospatial (e.g. national maps), Global Development (e.g. aid and food security data), Government Accountability and Democracy (e.g. government contact points), Health (e.g. prescription data), Science and Research (e.g. genome data and experiment results), Statistics (e.g. national statistics and census), Social mobility and welfare (e.g. housing and health insurance), and Transport and Infrastructure (e.g. public transport timetables). National government and the EC should facilitate the opening-up and publication of high-value datasets in machine-readable formats. The G8 Open Data Charter has been transposed in the Open Data Action Plans of the France, Italy, UK and the European Commission.
Open data: Unlocking innovation and performance with liquid information	This paper has the goal to quantify the potential value of using open data in seven "domains" of the global economy: education, transportation, consumer products, electricity, oil and gas, health care, and consumer finance. The paper aims on identifying the "levers" through which open data can create economic value and the barriers for adoption and "enablers" for capturing value by making data more open.

	While it is recognised that opening up data contributes to societal goals such as improving the transparency and accountability of Institutions, the researchers focuses on economic value that can be created by open data and how this economic value can be reached by making data more "liquid" (open, widely available, and in shareable formats).
EU – Results of the online survey on recommended standard licensing, datasets and charging for the re-use of public sector information	In the context of the implementation of the revised Directive on the re-use of public sector information (2013/37/EU), the European Commission ran an online consultation to seek the views of stakeholders on recommended standard licensing, datasets and charging for the re-use of public sector information, which should be addressed in the future Commission guidelines. The main results with regards to the identification of high-value datasets are the following: • There is an even split of opinions between the stakeholders favouring "high-value for commercial re-use" and the ones preferring "high-value for non-commercial re-use" as determinant for labelling a dataset as "core". • In the survey a number of dataset categories were suggested to the stakeholders, asking them to indicate which ones could be considered as "core" datasets. The results showed there is a general consensus as to the inclusion of virtually all of the suggested datasets in the category of 'core data sets' so as to ensure their immediate availability. Some categories stand out as particularly relevant including geospatial data and data about transport, statistics, earth observation, environment and public finances. • For "core" datasets a majority of the stakeholders indicated that machine-readability and guaranteed quality should be the most important characteristics.
ISA Access to base registers	This report includes a set of recommendations for providing access to base registries (core information like addresses, persons, organisations...) for all layers of interoperability of the base registry interoperability model (legal, organisational, semantic and technical interoperability). A base registry is defined in "D1.2. Base registry definition" as "a trusted authentic source of information under the control of an appointed public administration or organisation appointed by government".

Based on the analysis of the above, we elaborated on a working definition of "high-value dataset", taking into account two different points of view: the one of the data **publishers** and the one of the data **re-users**. Of course both two perspectives may in certain circumstance be strongly interlinked and overlap. For instance, the fact that a publisher may consider a dataset as of high value because it contributes to transparency, does not exclude the fact that also from a re-users' perspective it could be of high value for the same reason.

This said the distinction between these two viewpoints has been developed as an instrument for identifying high-valued datasets rather than providing a definition fitting for all purposes. In fact a given dataset may serve multiple purposes depending for instance on the stakeholders involved, the specific point in time in which the dataset is analysed etc.

High value from a data publisher's perspective

From the viewpoint of the data publisher, there can be different reasons to make a dataset available. Particularly in the case of public administrations, the following dimensions contribute to the determination of the value of a given dataset. From the publisher's perspective, a dataset may be considered of high-value when one or more of the following criteria are met:

- It contributes to **transparency:**
  These datasets are published because they increase the transparency and openness of the government towards its citizens. For instance the publication of parliaments' data, such as election results, or the way governmental budgets are spent, or staff cost of public

administrations all contribute to the transparency of the way public administrations are working.

- Its publication is subject to a **legal obligation:**
  In some cases the publication of data is enforced by law. The PSI Directive for instance, regulates the publication of policy-related documents by (semi)public organisations.
- It directly or indirectly relates to their **public task:**
  A public administration may publish a dataset because it directly relates to its public task. For instance DG CLIMA may publish statistics on CO_2-emission as part of its task for raising awareness about climate change.
- It realises a **cost reduction:**
  The availability and re-use of a dataset, e.g. contact information, code lists, reference data and controlled vocabularies, eliminates the need for duplication of data and effort, reduces costs and increases interoperability. Collections of data housed in the base registers and geospatial data are prime examples of dataset which opening up will lead to direct cost reductions in data management, production and exchange.
- The type and size of its **target audience:**
  A dataset may be useful for/relevant to a large audience (size-based value), for instance traffic data. On the other hand a dataset may bring large value to a specific target audience (target/subject-based value), for instance a dataset containing data of particles colliding at high speed in a particle accelerator.

High value from a data re-user's perspective

From a data re-user's perspective, the value of a dataset primarily depends on its **use and re-use potential**, which can effectively lead to the generation of (new) **business models.**

The use and re-use potential of a dataset is defined by the size and the dynamics of the target audience of the dataset (see above), as well as by the number of new and existing systems and services that are using (or could use) the particular dataset.

Opening up datasets with a high use and re-use potential is expected to lead to the creation of new products and/or services that have direct or indirect economic or social impact and/or positive economic externalities. The base registers, geospatial data, transport data and statistics constitute prime examples of datasets with a high use and re-use potential.

It is worth making a special reference to the interest of data re-users in datasets which contribute to transparency. Such datasets have a strong social impact. An example of open data used for serving transparency is the case of re-use of parliament data to develop an app about the activity of parliament members, like zewerkenvoorjou.be and TheyWorkForYou.com.

High-value datasets publication requirements

The value of a dataset also depends on its reusability. A dataset can be considered of high value, but its value is greatly diminished if it is not published in a highly reusable, open format. For instance, a dataset containing spending information of a public administration may be considered as high-value data. But if this information is not published with an open licence or in a structured or machine-readable format (as is the case for PDF) the potential re-use of this dataset is rather low. In order to increase its potential for re-use, the data should be made readily available in a non-proprietary, machine-readable format with an open licence and clear re-use conditions. In this way data re-users can easily process this information and develop products and services, such as the app publicspending.net.

For datasets to be re-used extensively, and thus to bring as much value as possible, datasets should be published according to a set of best practices to maximise the reusability. In this context the 5-star schema of Tim Berners-Lee is a good tool for assessing the conformance to

publishing principles of high-value datasets. We advise high-value data to be published at least as 3-star data, which implies making it available on the web under an open license in a non-proprietary structured format. Publishing data ranked as 3-star is still rather simple to do since normally there is no need to learn new tools. For instance, the data can be made available as CSV or Open Document Spreadsheet (ODS) by simply selecting this particular type when saving the document.

For a data re-user already a 3-star dataset gives some important advantages, like:

- Changing and manipulating the data without being confined by the capabilities of any particular software;
- Processing the data;
- Sharing the data; and
- Exporting the data to another format.

Also, to greatly improve the potential for re-use, it is important that metadata related to the re-use condition are duly published. These metadata include for example the licence under which a given dataset is made available and its provenance. However, this information is often missing.

Conclusion

The definition was used in practice for identifying and prioritising datasets owned by European Institutions to be listed on the European Union Open Data Portal (EU ODP). With this work we were able to identify a total of **261** new high-value datasets.

The list containing this high-value datasets according to the elaborated definition can be found through the following link:
https://docs.google.com/spreadsheets/d/1jeb5R2O7YQMxFE7BjJorGXFtUduBhyNh7HsD2eGId5c/edit#gid=1901062878

Additionally, the research showed that identifying apps being developed can help in identifying datasets having a high re-use potential. Since potential re-use of a dataset strongly determines its value, looking at the apps being developed or requested is a good indication of which datasets are of high value (mainly from the data re-user's point of view) and can thus be prioritised for being published.

The work of identifying datasets that may be of high-value for different reasons is a continuous process. The working definition that was elaborated in the context of this work can assist data publishers, in particular government bodies, to prioritise on which data to open up first, taking into account resource restrictions.

References

(2012). *Open Data White Paper - Unleashing the Potential.* Norwich: The Stationery Office.

Official Journal of the European Union. (2013, June 27). Retrieved December 02, 2013, from EUROPA - European Union website, the official EU website: http://eur-lex.europa.eu/JOHtml.do?uri=OJ:L:2013:175:SOM:EN:HTML

Berners-Lee, T. (2006, July 27). *Linked Data*. Retrieved December 02, 2013, from World Wide Web Consortium (W3C): http://www.w3.org/DesignIssues/LinkedData.html

Bizer, Heath, & Berners-Lee. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 1-22.

European Commission. (2010). *European Interoperability Framework for European Public Services.*

ISA. (2012). *How Linked Data is transforming eGovernment.* European Commission.

Official Journal of the European Union. (2003, November 17). Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information.

Open Knowledge Foundation. (2013). *Open Definition*. Retrieved December 02, 2013, from Open Definition: http://opendefinition.org/

The Open Knowledge Foundation (OKFN). (n.d.). *Open Government Data*. Retrieved April 22, 2014, from Welcome to Open Government Data: http://opengovernmentdata.org/

W3C. (2013). *Linked data*. Retrieved December 02, 2013, from World Wide Web Consortium (W3C): http://www.w3.org/standards/semanticweb/data