Good practices for identifying high value datasets and engaging with re-users: the case of public tendering data

Brecht Wyns, Leda Bargiotti, Nicolas Loozen, Nikolaos Loutas, Makx Dekkers, Michiel De Keyser, Stijn Goedertier
PwC EU Services, Belgium
firstname.lastname@be.pwc.com

This paper lays out the base for a plenary presentation about good practices for identifying high value datasets and engaging with re-users.

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 criteria to identify high-value datasets from the perspective of the data owner, but also from the point of view of the re-user. We used the high value dataset criteria to identify and prioritise datasets owned by European Institutions to be listed on the European Union Open Data Portal (EU ODP). One such dataset is the public tendering data of the Publications Office. Therefore, we also interacted with re-users of data from Tenders Electronic Daily (TED) to identify their needs and collect suggestions. Their input was used to formulate recommendations and the functional requirements for the future TED service. Brought together, this experience will allow EU institutions to determine which new datasets should be published with priority or which high-value datasets already listed on the EU ODP should be improved with priority, while taking into account the point of view of re-users.

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

Introduction

Open Government is supported by the availability of 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 used, re-used and redistributed by anyone and for free. Open data is key to achieve a number of goals both from a data publisher and a data re-user perspective, which could include transparency, research or the creation of business opportunities.

Furthermore, 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.

Theoretically, the more data is made available, the better. But given the limited resources that publishers have at their disposal, one of the objectives of this activity was to **prioritise datasets for publication**.

Our proposed approach for prioritising datasets includes gathering input from re-users. Based on the work done with the Publications Office to **engage with re-users** of public procurement data from TED, this paper proposes **good practices** for prioritising datasets for publication and engaging with re-users.

Good practice: identifying high-value data sets

One of the methods for prioritizing dataset publication is to identify high-value data sets. Indeed, the publication of high-value datasets as a priority for opening up data by governments has 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.

For publishers of datasets to be able to properly prioritise based on the value of datasets, it is important to establish a common understanding on what can be considered as high-value datasets. In the context of Action 1.1[1] of the ISA Programme and in close collaboration with the Publications Office of the EU, we elaborated on a working definition for high-value datasets. In order to establish this definition, we looked at datasets from the point of view of data owners, but also, and more importantly, from the perspective of re-users.

For identifying 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, a survey on social media was 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.

Based on this analysis, we identified the characteristics of "high-value datasets", looking at them from three different perspectives:

- Reusability
- Value for data owners
- Value for re-users

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 re-use potential for this dataset is rather low. In order to increase its potential for re-use, the data should be made 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.

The 5-star schema of Tim Berners-Lee can be used as a benchmark to assess the reusability of datasets. This schema constitutes a set of best practices to maximise reusability when publishing 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

[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.

data ranked as 3-star is still rather simple to do since normally there is no need to use new tools. For instance, the data can be made available as CSV or Open Document Spreadsheet (ODS) simply by selecting this particular format when saving the document.

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

- 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.

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

Value for data owners

From the viewpoint of the data owner, there can be different reasons to make a dataset available. Particularly in the case of public administrations, 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, government expenditures, or staff cost of public administrations all contribute to increased transparency for public administrations.
- The 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 helps with **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, which reduces costs and increases interoperability. Data from 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.

Value for re-users

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 activity. The use and re-use potential of a dataset is defined by

- The size and the dynamics of the **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 high 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.

- The **number of systems or services** that 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.

Datasets contributing to transparency deserve a special mention. As these datasets have a strong social impact, re-users' interest is high. An example of open data serving transparency is the case of apps such as zewerkenvoorjou.be and TheyWorkForYou.com, which are based on activity data of parliament members.

However, the data owner cannot foresee all the possible creative use cases for its datasets prior to their publication. For this reason, it is important to engage directly with re-users.

Good practice: engaging with re-users

Following the high value dataset criteria, public tendering data is a good example of dataset to prioritise for publication. According to EU directive 2004/18, public procurement notices for amounts above certain thresholds must be published online. Recently, the Publications Office of the EU also made its datasets available in a re-usable open format. Consequently, more effort should also be spent on engaging with the re-users.

In the context of ISA action 1.1 and together with the Publications Office, we collected needs and suggestions from Tenders Electronic Daily (TED) re-users as to how they use TED notices. The main objective of this work was to collect needs and suggestions from TED re-users as to how they use TED notices, what information they want to have and what formats and access technologies they want TED to support. This feedback contributed to formulating 14 recommendations and the functional requirements for the future TED service.

Together with the Publications Office, we identified 21 re-users of TED data from a list of organisations that have requested access to the TED data, and contacts from the open spending network. These re-users included re-publishers of TED data, data journalists, researchers, academia, and public administrations both at national and European level. In order to collect information from re-users, we carried out the following activities:

1. Interviews were held with re-users identified by the Publications Office and PwC, via conference call or in a face to face meeting;
2. A questionnaire was sent to re-users of tender re-publication and support services in several Member States (Finland, Ireland, Spain, and UK).

In many of the interactions that we had with TED re-users, they expressed interest and willingness to share their suggestions with the Publications Office. This involves the types of uses that they see for the data but also the technical infrastructure that they would like to see. Engaging these re-users would give better insights in what they need and want. In addition, it would open up a source of ideas and suggestions, and possibly even access to additional tools developed by re-users.

Therefore, we formulated the following recommendations with regard to re-users engagement:

- Establish a communication channel:
  The approach could be simple at first, for example with a mailing list or a community on Joinup or on the Open Data portal that could be used to make announcements to re-users about the redesign project and other issues of interest. The announcements, and any other

relevant communications, should also be made available to topic specific mailing lists and fora. The re-users could use those channels for providing feedback and bringing up issues that they want to discuss with the Publications Office.

- Use collaborative tools:
  At a later stage, a community site could be established, either as a Wiki-based collaboration platform or as a repository for joint development of applications, tools and services based on the data and metadata. Joinup offers each project a free Subversion (SVN) repository, alternatively, Github can be another solution for joint development. This community would also encourage the collaboration between re-users and the cross-fertilisation of ideas and business opportunities.

Conclusion

The criteria for identification of high-value datasets were used in practice to identify and prioritise 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 and prioritise a total of **261** new high-value datasets.

In addition, engaging with re-users allowed us to gather needs and suggestions to improve the TED service and the high-value TED dataset from the perspective of re-users.

Prioritizing datasets based on their value is a continuous process in which re-users should be involved. The working definition for high-value datasets 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

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. Retrieved from http://ec.europa.eu/isa/documents/isa_annex_ii_eif_en.pdf

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. Retrieved from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32003L0098:En:HTML

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

(2012). Open Data White Paper - Unleashing the Potential. Norwich: The Stationery Office. Retrieved from https://www.gov.uk/government/publications/open-data-white-paper-unleashing-the-potential

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

PwC. (2014). Do current licensing practices hinder the commercial reuse of open data? Retrieved from https://www.w3.org/2013/share-psi/wiki/images/5/5f/Bargiotti_Wyns.pdf.

PwC. (2014). Value-based prioritisation of Open Government Data investments. Retrieved from http://www.w3.org/2013/share-psi/workshop/samos/report.

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