THE 5 STARS OF OPEN DATA PORTALS

Pieter Colpaert, Sarah Joye, Peter Mechant, Erik Mannens, Rik Van de Walle
iMinds - UGent – MMLab – MICT
Gaston Crommenlaan 8, 9000 Ghent
{firstname}.{lastname}@ugent.be

A key element in an Open Data Policy is an Open Data Portal. While advocacy tools are in place for assessing Open Datasets (the 5 stars of Linked Open Data by Tim Berners-Lee) and Open Data Communities (the 5 stars of Open Data Engagement by Tim Davies et al.), there is no guidance for determining the functionality and affordances of an Open Data Portal. Therefore, this paper introduces the 5 stars of Open Data Portals which can serve as a guide for Public Administrations just entering the Open Data field. For existing and established Open Data Portals, the 5 stars can help to progressively work towards more involved and advanced goals.

1. – Introduction

In order to become more transparent and to work closer with citizens and companies, Public Administrations worldwide are starting up Open Data Portals stimulated by the idea that open government data reuse can open up economic opportunities, can promote transparency and accountability or can support the reform of public services [1].

As not one dataset is a perfect representation of reality and remains a work in progress for a long time, the Pareto principle comes in: it would take 80% of the time, to get the last 20% of the data right. When Public Administrations open up data, the Open Data Portal can support processing feedback on the data to get the last 20% right by community effort.

Two '5 star' systems are in place for Public Administrations to take up (Linked) Open Data. From a technological perspective Tim Berners-Lee proposed the 5 stars of Linked Open Data [2], from a social perspective, there are the 5 stars of Open Data Engagement [3]. Both these 'star systems' have been adopted widely and are a real success in the Open Data ecology. Currently however, there are no clear pointers or guidelines for Open Data Portals aiming at stimulating data reuse and improving data quality.

In order to fill this gap, this paper provides a 5 star system to guide Public Administrations in setting up an Open Data Portal. This 5 star system for Open Data Portals is complementary to the star system of Open Data Engagement and Linked Open Data and thus helps Public Administrations to maximize their stars in these two other evaluation frameworks.

The paper is structured in four sections. First, we discuss related work that tried to assess Open Data Portals. Next, a 5 star evaluation system for Open Data Portals is proposed. In the third section relevant and real-world examples of Open Data Portals are discussed using the 5 star evaluation systems. The paper ends with a conclusion pinpointing future work and research activities.

2. – Related work

Four important studies have been conducted on the functionality and affordances of Open Data Portals.

In 2010 Socrata conducted an Open Government Data (OGD) study [4] to benchmark the current state of OGD, taking into account the three constituent stakeholders; the government, citizens and developers. The study wanted to gain insight in the different attitudes, motivations, requirements and needs of these three stakeholders in the Open Data field in order to understand and to facilitate the adoption and implementation of Open Data. Importantly, the benchmark study underlined the need for Open Data portal guidelines, showing that more than 56% of all levels of government already have an Open Data mandate but that only half of those cases had specified that mandate in a central data site or data catalogue. The Socrata benchmark study offers better understanding of the current state of Open Data and the requirements, needs and motivations of the different stakeholders.

In December 2007 30 open government advocates developed a set of principles of Open Government Data [5]. Even though their guidelines are not legally binding, they are considered to be widely accepted among the Open Government Data community. In order to be fully 'open', data must be:

- complete (in the broadest sense of the word)

- primary (collected at the source), timely (i.e. open as quickly as possible)

- accessible (to the widest range of users and for the widest range of purposes)

- machine-readable (can be processed in an automated way)

- non-discriminatory (available to anyone, no registration requirement)

- non-proprietary (available in an open format)

- license-free (not subject to Intellectual Property Rights)

- permanent (findable over time)

- as costless as possible (as usage costs are one of the greatest barriers to access).

In 2012, the Pricing of Public Sector Information Study (POPSIS) [6] was conducted "to assess the extent to which Open Data Portals have an impact on, or are influenced by, the PSI pricing policy" [6, p. 6]. This study developed a set of key indicators that measure the level of openness and the impact of these portals, which are:

- number of datasets available

- implementation of the 5 stars of Linked Open Data

- clear re-use conditions

- position of the governing body

- accessibility

- take-up by citizens and application developers

- availability of the take-up data on the website

- range of government agencies contributing dataset listings

- availability of 'suggest a dataset' form

- presence of guidelines or events

- indicators for community engagement.

Finally, a 2013 project by the Open Knowledge Foundation (OKFN) called 'The Open Data Census' [7] aggregates a list of all data portals with their geographical coverage. At the time of writing (June 2013), there are 281 Open Data Catalogues and 77,400 openly licensed datasets available.

An Open Data Portal is the main externalization of an Open Data Policy. The Portal brings together data re-users, data enrichers and data maintainers. While the four studies mentioned above are great advocacy tools for Open Data Policies, they are not directly referring to the features of an Open Data Portal.

3. – The 5 stars of Open Data Portals

An Open Data Portal is a collection of systems set up to make Open Data used and useful[1]. It ensures that potential reusers can discover data, but that is only the start of it. To make data used, data needs to be well described and tools have to be available for *reusers*. The data needs to be in good quality for others to transform it into knowledge and make it *useful*. The same reasons for opening data in the first place are also valid for Open Data Portals: ensure that the tools are cocreated with the reuse community, remove the hurdles to start reusing data and try to collaborate with projects outside of one's organisation to leverage the data quality.
Our proposed 5 star system for Open Data Portals uses stars to represent the main function or affordance that the data portal is built or used for. The stars are ordered by the investment of time needed to be able to fully implement the star. Below, we explain the requirements for each star and list pointers on how to implement these. The stars categorization starts with portals linking to various datasets and continues towards a meta-data portal for both the datasets and the re-use of the datasets. The fourth star category of Open Data Portals takes care of the data publication itself. Finally, a data hub is set up where data becomes a common resource.

★ A dataset registry
A dataset registry is a moderated list of links towards datasets or links to other portals. It provides a clear answer to the question: "Who has which (open) dataset and where can I find it?".
A dataset registry can contain direct links to datasets which are or are not yet openly licensed (or the license might be unknown) and/or links to other pages/portals offering datasets. A point of contact should be announced where third parties can request datasets (an e-mail address is sufficient). The registry has a three folded goal: to be the reference for Open Datasets concerning the organization, to involve on-going and future projects and to raise data awareness within the organization.

---

1   "Opening up knowledge and seeing it used and useful" is the goal of the Open Knowledge Foundation, see http://okfn.org/about/our-vision/.

The Socrata benchmark study shows that in 2010, approximately 56% of all levels of government had an Open Data mandate, but only in 50% of those cases, the mandate had been specified in an Open Data initiative or an Open Data Portal [4, p. 9].

★★ A meta-data provider

This category extends on the 'one star' Open Data Portal category by storing more data about the datasets (meta-data). The meta-data contains information about which license is used, the geographic context, how many times per week they are updated, and so on. For more specific pointers about what to include, the W3C has a vocabulary to describe datasets: DCAT [8].

Meta-data is published as Open Data as well (they become Open Data if they are licensed under an Open License). Consequently, other Open Data Portals will be able to integrate these registers in theirs (federation).

★★★ A co-creation platform

The 'third star' Open Data Portal category goes beyond the aforementioned functionalities of one or two star Open Data Portals by supporting interaction and conversations about the data and potential applications with the data.

Conversations about the datasets themselves should be supported unobtrusively in order to safeguard serendipitous reuse; the data should be directly accessible and not restricted by obligatory procedures (such as filling out an email address before downloading a dataset). This is also reflected in the ten principles of Open Data [5], e.g. in their requirements that 'access must be non-discriminatory' or that people need to be able to subscribe to updates on a certain dataset using e-mail, social media or newsfeeds (in atom or RSS). Conversations should also support communication about idea-generation and code-reuse (announcing hack days, data dives, co-creation events, etc.) and bring together citizens and developers.

Next to storing the meta-data about the datasets (as in the two star category), data about data re-use (apps/concepts/ideas) must be stored as well. Similar to the DCAT vocabulary, a vocabulary is being defined in the context of the Apps for Europe project [9]. Federation becomes possible when these data sets are also openly licensed. When data is federated towards a global context, code duplication will be overcome and more business potential may be unleashed.

★★★★ A data publishing platform

The authentic sources of the datasets publish meta-data using the Data Portal. The fourth star category of Open Data Portals implies that these datasets themselves are published using the portal. When doing so, the URIs can be maintained by the Data Portal's URI strategy, resulting in URIs and URLs which can remain the same for ever [10].

Shifting data publishing to the Data Portal also entails that the Portal becomes a tool to generate Linked Open Data. Data quality can be safeguarded as the authentic source will receive automatic feedback on newly added data. Once the data is accepted, datasets can be easily accessed through open formats (e.g. json, xml, csv) and subsequently, generic tools can be built to work with this data. After all, data journalists and other stake holders need simple tools to study the data, application developers and programmers need support for reusing the data directly in their apps, and other stakeholders such as citizens, who happen to notice errors in datasets, need to be able to correct them. Some of the Sunlight Foundation principles reflect this fourth star category of Open Data Portals, stating that data must be accessible, machine processable, permanent and that data formats must be non-proprietary. Also some key indicators, developed in the POPSIS report [6], such as accessibility, clear re-use conditions, availability of take-up data on the website, and availability of a 'suggest a dataset' form, relate to the fourth star category of Open Data Portals.

★★★★★ A common data hub

A common data hub reflects the fifth star category of Open Data Portals enabling data to become a common resource: anyone can implement his/her own Data Life Cycles (see Figure 1) to maintain his/her (enriched) versions of the data. In this way, the Data Portal becomes a hub between organizations and the outside world aggregating data from inside the organization, as well as third party data that may be useful for the organization.

While different versions are maintained, they should stay interoperable: corrections in one version should be able to merge with another version, both on meta-data as on the actual data level. In this way, not only the app concepts, data tools and visualizations are co-created, also the data itself. As such, the Open Data Portal becomes self-correcting with data that keeps improving by input that may come from everywhere. Data is also linked: through meta-data about the datasets, through meta-data about the app concepts and through the data itself.
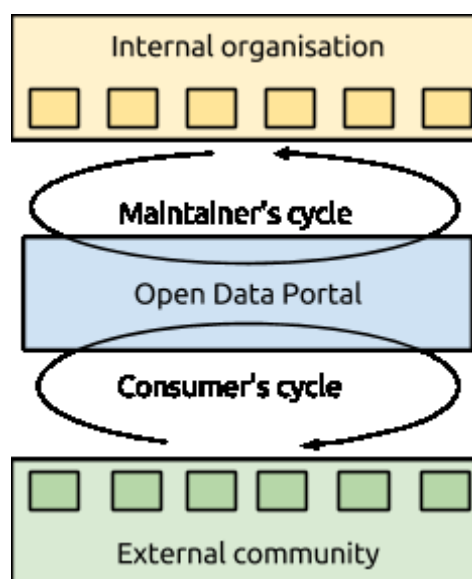

Figure 1: Data Life Cycles

4. – Examples of the 5 star system for Open Data Portals

This section, illustrates the proposed 5 star system for Open Data Portals with a concise set of examples.

★ A dataset registry

The official Belgian Open Data Portal [11] is a good example of a registry on its way to become a meta-data repository. The Belgian ICT organisation FEDICT hosts a website which links to other websites and collect datasets using a Drupal CMS storing meta-data fields such as coverage, granularity, publisher, category, publication format and so on.

★★ A meta-data provider

The official Dutch Open Data Portal [11] exemplifies a Meta-data platform with data structured towards the Overheid.nl Web Meta-data standard (OWMS), openly licensed and available for querying through an API.

★★★ A co-creation platform

People visiting a co-creation platform are called to action: they can participate as a non-technical person in generating ideas and raising issues, developers can add their apps or programs to the portal, request datasets, discuss datasets in the forum and can subscribe to an RSS feed of the meta-data of the apps. Given these affordances, the official Open Data Portal [13] for the UK serves as a perfect an example of the third star category.

★★★★ A data publishing platform

The UK's data.gov.uk's [13] fourth star is already on the horizon: the portal has a SPARQL endpoint and all the datasets are described using their own schemas. Also, a URI strategy is given and they have built tools for various user profiles to discover the data.

★★★★★ A common data hub

The 5 star category can be envisioned by comparing this type of Open Data portal with Software Management Systems (e.g. github, bitbucket, Google code). All the required affordances are indeed present: people can create their own version of the software, can implement their own development cycles and can merge their changes with other versions of the software. Although not perfect, a quick implementation can be made by hosting the data on a Software Management System. This has already been done by for example Montpellier [14] and Chicago [15]. Also the Open Data Institute has released a Git Data Viewer [16], Max Ogden started working on Dat [17] and MMLab started working on r&wbase [18].

5 – Conclusion

In this paper, the need was expressed for guidance in building Open Data Portals. A new star system was suggested. The 5 stars of Open Data Portals are:

> ★ A dataset registry: a list of links to Open Datasets
> ★★ A meta-data provider: maintain, structure and open up your meta-data
> ★★★ A co-creation platform: gather tools and stimulate conversations about re-use
> ★★★★ A data publishing platform: provide the data itself in common formats
> ★★★★★ A common data hub: open governance, provenance, trust and versioning

In section 4, examples showed that for each star, tools are already being made and the "5 star Open Data Portal" comes within reach.

6 – Future work

This paper has been written in an interdisciplinary setting by the Department of Communication's Media & ICT and the Department of Engineering's MultiMedia Lab at

Ghent University and the Flemish research institute iMinds (http://www.iminds.be). Both parties will use these 5 stars as a guide line for future work.

MMLab, which is building tools for Linked Open Data, is going to use the stars as a guide for the roadmap of different projects: The DataTank (since 2010) for leveraging Data Portals toward the 4th fourth star, and r&wbase [16] (since 2013) for reaching the fifth star of category of Open Data Portals. The 5 star system will also be used as a roadmap for the Flemish Open Data portal built within a project funded by the EWI (department of Economy, Science and Innovation) and the Flemish Government.

Acknowledgement

References

[1] Tim Davies. Open data, democracy and public sector reform. http://practicalparticipation.co.uk/odi/report/.

[2] Tim Berners-Lee. 5 * Open Data. http://5stardata.info/.

[3] Tim Davies. 5-Stars of Open Data Engagement? http://www.timdavies.org.uk/2012/01/21/5-stars-of-open-data-engagement/.

[4] Socrata. Open Data Benchmark Study Report. http://www.socrata.com/benchmark-study/download-report/ .

[5] Open Government Working Group. https://public.resource.org/8_principles.html.

[6] European Commission, Information Society and Media. Pricing Of Public Sector Information Study . http://ec.europa.eu/information_society/policy/psi/docs/pdfs/report/11_2012/models.pdf.

[7] Open Knowledge Foundation (OKFN). The Open Data Census. http://census.okfn.org/.

[8] W3C. Data Catalog Vocabulary (DCAT). http://www.w3.org/TR/vocab-dcat/#class-catalog.

[9] European Commission. Apps4EU: Apps 4 Europe - Turning Data into Business. http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=325090.

[10] Tim Berners-Lee. Cool URIs don't change. http://www.w3.org/Provider/Style/URI.html.

[11] Belgian Government. http://data.gov.be/nl.

[12] Dutch Government. https://data.overheid.nl.

[13] British Government. http://data.gov.uk/.

[14] City of Montpellier. OpenDataMontpellier. https://github.com/VilleDeMontpellier.

[15] City of Chicago. Chicago. https://github.com/Chicago.

[16] Git Data Viewer. http://www.theodi.org/blog/git-data-publishing

[17] Dat. https://github.com/maxogden/dat/

[18] Miel   Vander   Sande   et   al.   2012.   r&wbase: Git   for   triples.   LDOW2013.
http://events.linkeddata.org/ldow2013/papers/ldow2013-paper-01.pdf.