Towards assessing the quality evolution of Open Data portals

Jürgen Umbrich, Sebastian Neumaier, Axel Polleres
Vienna University of Economics and Business, Vienna, Austria

In this work, we present the Open Data Portal Watch project, a public framework to continuously monitor and assess the (meta-)data quality in Open Data portals. We critically discuss the objectiveness of various quality metrics. Further, we report on early findings based on 22 weekly snapshots of 90 CKAN portals and highlight interesting observations and challenges.

1 Motivation

As of today, the Open Data movement enjoys great popularity among governments and public institutions and also increasingly in industry by promising transparency for the citizens, more efficient and effective public sectors or the chance to outsource innovative use of the published data. However, first critical voices appear addressing the emerging issue of low quality in the meta data and data source which is serious risk that could throw off the open data project.[1] However, to the best of our knowledge there exists no comprehensive quantitative and objective reports about the actual quality of Open Data.

Various efforts already exist to study different aspects of Open Data portals which are the main platforms to publish and find datasets. For instance, the Open Data Barometer project assesses the readiness of countries to exploit their Open Data efforts and the achieved impact based on expert judgements.[2] Similarly, the Open Data Census provides a survey for data portal owners to analyse their data in more detail.[3] The PoDQA project (Calero, Caro, & Piattini, 2008) addresses the evaluation of the quality of a Web Portal by defining a data quality model containing 42 characteristics. While some aspects of Open Data quality align with the ones of Web Portals, we identified domain-specific quality dimension in the context of Open Data (e.g., the openness of provided data based on the license or format). More related to a data quality assessment is the OPQUAST project[4] which provides a check-list for Open Data publishing, including some quality aspects.

We can conclude and also identified by Reiche, Höfig, and Schieferdecker (2014), that there is a need for an automatic quality assessment and monitoring framework to better understand quality issues in Open Data portals and study the impact of improvement methods over time. In this work, we present our effort for such a framework, critically discuss our intrinsic and contextual quality metrics and report on early findings which provide insights for future directions.

[1] http://www.business2community.com/big-data/open-data-risk-poor-data-quality-01010535
[2] http://opendatabarometer.org, last accessed 2015-02-11
[3] http://census.okfn.org
[4] http://checklists.opquast.com/en/opendata

	Dimension	Description
Q_r	Retrievability	The extent to which meta data and resources can be retrieved.
Q_u	Usage	The extent to which available meta data keys are used to describe a dataset.
Q_c	Completeness	The extent to which the used meta data keys are non empty.
Q_a	Accuracy	The extent to which certain meta data values accurately describe the resources.
Q_o	Openness	The extent to which licenses and file formats conform to the open definition.
Q_i	Contactability	The extent to which the data publisher provide contact information.

Table 1: Quality metrics together with their informal description.

2 Quality assessment monitoring framework

We develop the Open Data Portal Watch framework, that periodically monitors the content of CKAN portals and compute a set of quality metrics to gain insights about the the evolution of the (meta-)data. While a similar effort is presented by (Reiche et al., 2014), one of our core requirement is to perform all tasks automatically and use objective quality metrics which have a clear interpretation.

2.1 Framework description:

In brief, our framework consists of the following components:

- A **fetching** component to retrieve a snapshot of all dataset descriptions of a portal by accessing the CKAN API and storing the JSON responses in a document store. We also download the actual data sources per snapshot to inspect their content for an in-depth quality analysis.
- A **quality assessment** component to compute several quality metrics per dataset, portal and snapshot with the flexibility to add or modify metrics on the fly.
- A **dashboard** to display vital quality metrics for each portal to users using various views and charts such as a compact summary view of the most important quality measures.

Our dashboard component is publicly available[5] and we publish the collected raw meta data snapshots of all monitored portals to motivate and engage other researchers in analysing it.[6]

2.2 Quality metrics

Our quality metrics and dimensions are partially aligned with Reiche et al. (2014) and Batini, Cappiello, Francalanci, and Maurino (2009) and briefly summarised in Table 1. Due to space limitations, we provide a detailed formal definition of these quality metrics online.[7] In general, we compute for each portal p the metric by averaging the quality metrics for the the single datasets. More formally, the general formula is:

$$Q_\mu(p) = \frac{\sum_{d \in p} \mu(d)}{|p|}$$

with the function μ as a place-holder for the quality metric specific formula.

Next, we critically discuss our used quality metrics, provide an intuition about the actual computation and comment about some of the metrics not yet consider in our framework but used in the literature.

[5] http://data.wu.ac.at/portalwatch/
[6] http://data.wu.ac.at/portalwatch/data
[7] http://data.wu.ac.at/portalwatch/docs/metrics.pdf

Retrievability: The retrievability, or also sometimes referred to as availability, is a clear and objective measure which can be easily computed by performing HTTP lookups on the datasets and resources.

The retrievability function of a dataset $\text{ret}(d)$ ($\text{ret}(r)$ for a resource respectively) returns 1 if and only if the status code of the `HTTP GET` request returns `200 OK`.

Usage & completeness: Low values generally indicate a poor description of the datasets. However, there are several reasons why meta data keys are not used or no values are provided: i) some *keys are optional* and are not known to a publisher, ii) other keys are not relevant to the publisher or iii) some keys are kind of redundant (e.g., author and maintainer email). Because of this, a clear interpretation might be hard and as such, we consider these metrics rather as indicator and a good input for other quality metrics (e.g., as weights or filters for certain keys).

The usage of a dataset d in a portal p is basically $\text{usage}(d) = |\text{keys}(d)|/|\text{keys}(p)|$ (where $\text{keys}(d)$ is the set of available meta data keys in the dataset, and $\text{keys}(p)$ the set of all unique meta data keys in a portal). In contrast, the completeness of a dataset is the ratio of non-empty meta data keys to the available keys in the dataset.

Openness: This metric determines if the license and used data formats of a dataset classifies as open which requires a complete and up-to-date list of all available licenses and formats. We confirm the openness of a license by evaluating if the specified license is conform with a list provided by the Open Definition.[8] However, since we cannot guarantee the integrity and completeness of this list of 108 different licenses, our metric only reports on confirmed open licenses. Note, we can miss open licenses if they are not listed or not yet fully evaluated in the aforementioned list.

Looking into more detail, the license of a dataset can be specified by three different keys, namely *license_id*, *license_title*, *license_url*. Our metric is computed by matching the *license_id* or the *license_title* against the id or title of the open definition list. If this check fails, we try to match the *license_url* with the provided license URLs in the list. A license specification for a dataset is then considered as open (i.e. evaluates to 1) if and only if the matched license is approved by open-definitionorg.

Regarding the format openness of a dataset, we selected a set of file formats (containing common open data formats as *csv*, *json* or *rdf*) and considered a dataset as open if there is a corresponding resource with a format description matching our list. Similarly to the licenses, this metric only reports on the confirmed open formats and might miss other formats considered as open. It is worth mentioning that we excluded formats such as *xls* (Microsoft Excel) or *zip* since there exists no clear agreement if they should be considered as open or closed.

Contactability: The contactability metric computes the availability of contact information for the author or maintainer of a dataset, by checking if for valid email addresses or URLs in certain meta data keys. However, to correctly compute the metric, one would need to check in addition of such email addresses exist or URLs are actually pointing to contact details.

As of now, we provide three different contactability indicators. The first general metric checks if any of the meta data keys for contact information[9] is non-emtpy. Secondly and more specific, we check if any of the values is a syntactically correct email address and thirdly, if the values are syntactically valid URLs.

Accuracy: Typically one uses a distance function between the described and real values of a dataset to compute the accuracy of the meta data. Yet, it is challenging to provide meaningful distance functions for all meta data keys, e.g., computing the distance between the description of a resource and the content. In addition, we need to monitor the resources over time to verify the correctness of certain meta data keys (e.g. update frequency or change rate). As such, we decided to consider only keys for our accuracy metric that can be clearly measured, such as the format,

[8] http://opendefinition.org/

[9] The related CKAN keys are *author*,*author_email*,*maintainer* and *maintainer_email*

size or content encoding.

As for concrete metrics, we use currently only the header information of a resource to check the accuracy of the format, mime-type and content size. However, we will improve this metric in the future by actively monitoring and using the content of a resource for the accuracy metric. This also allows for other checks such as verifying the specified language or the change frequency.

Metrics not yet considered.
Timeliness: The timeliness is in general a measure of how sufficiently up-to-date a data set is for a certain task (e.g. live timetables or current election results). However, it is hard to automatically understand the time dimension from the content of a dataset, e.g., to distinguish between historical data vs. real-time data.

Information richness/uniqueness: Another frequently used metric is the information richness of the meta data description typically measured by how much unique information is provided compared to all other datasets. Nevertheless, portal owners want in certain cases a low uniqueness value for certain meta data keys. For instance, all datasets should be published under the same license. Also, a common pattern is to publish the same type of data but for different timestamps (grouped by year or month), in which case, the meta information differ only by the time value and the uniqueness value would be low again. These observations need to be carefully considered for the overall value.

3 Findings

At the time of writing we collected up to 22 weekly snapshots for 90 CKAN portals.

Open Data is growing. Overall, we observed that the number of datasets and resources is continuously growing by comparing the first and latest snapshot of each portal: We observed a average increase of 43% for the number of datasets for 70 portals, while only 5 portals showed an average decrease of 27%. Similarly, the number of resources increases by an average of 58% for 72 portals , while for only 7 portals the resource decreased by around 27%.

No agreement for additional meta data keys. We observed that the 90 portals use in total 2810 different extra meta data keys[10], with the majority of 98% of the keys appearing in only one portal. We also manually identified one set of semantically similar keys describing geo information[11] which consists of 44 keys and span across 42 portals. One implication of this observation is that combining all portals will be a challenge task.

Combining metrics for a better understanding. In Figure 2 we plot the average completeness vs. usage metrics for three different subsets of the CKAN meta data keys. The plot helps us to better understand how keys are used in the portal. For instance, the set of keys in the top left are very likely provided by default but are not really used (e.g., not relevant, or redundant) and the set of keys in the bottom left are rarely used in the portal. In fact, we see in this example that the keys in the bottom left are all part of the extra keys which are added by the portal provider. However, these keys are either not known to the publishers or not really relevant due to their low completeness and usage values. The portal provider could now either remove those keys or provide them as default and with a better documentation to increase the overall quality of the meta data in his portal. While the keys on the right part are frequently used and are either useful default keys (top right) or useful optional keys (bottom right).

Evolution tracking provides new insights. Eventually, we want to briefly show that tracking the portals and their quality metrics can provide important insights. In Figure 1 we show a screen-

[10] Each CKAN installation consists of a set of standard default meta data keys which can be optionally extended by user-defined extra keys

[11] keys such as or similar to ”geolocation”, ”location” or ”spatial”

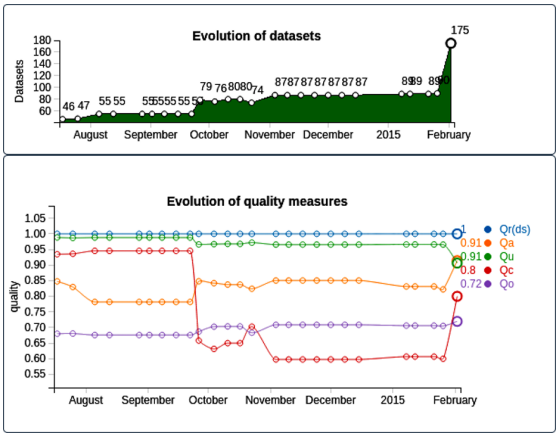

Figure 1: Evolution of a portal.

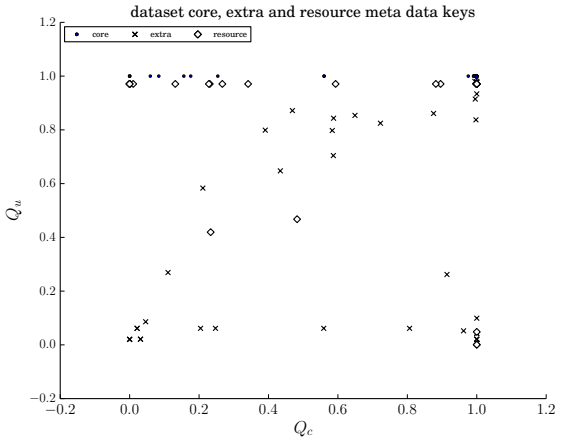

Figure 2: completeness vs usage plot

shot of a ”portal-evolution view“ in our dashboard. The top part shows the evolution of the dataset in the portal and the bottom part the values for our quality metrics (each dot is one snapshot). We can clearly see how added or removed datasets can influence certain quality metrics.

4 Conclusion

We developed and presented our effort towards monitoring and assessing the evolution of the quality of Open Data Portals, discussed a set of objective metrics that can be automatically computed and reported on early findings. In future work we plan on the technical site to provide more information on our dashboard and integrate more portals. Regarding the research direction, we will refine and extend our quality metrics based on the discussed observations. In addition, we will explore the possibility for quality metrics and anaylsis for specific domains, considering domain-specific standards or vocabularies. Further and inspired by (Kučera, Chlapek, & Nečaskỳ, 2013), we plan to evaluate and estimate the impact of automatic improvement methods for the overall data quality. Also of interest is to tackle the identified challenge of the diversity of extra meta data keys by creating and publishing mappings between semantically related keys.

Acknowledgements The authors would like to thank the reviewers for the valuable feedback and improvement suggestions. This work was partially funded by the “Jubiläumsfond der Stadt Wien” for the project OpenData@WU.

References

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Comput. Surv.*, *41*(3), 16:1—-16:52.

Calero, C., Caro, A., & Piattini, M. (2008). An applicable data quality model for web portal data consumers. *World Wide Web*, *11*(4), 465–484.

Kučera, J., Chlapek, D., & Nečaskỳ, M. (2013). Open government data catalogs: Current approaches and quality perspective. In *Technology-enabled innovation for democracy, government and governance* (pp. 152–166). Springer.

Reiche, K. J., Höfig, E., & Schieferdecker, I. (2014). Assessment and Visualization of Metadata Quality for Open Government Data. In *International conference for e-democracy and open government.*