# Quality assessment & evolution of Open Data portals

Jürgen Umbrich*, Sebastian Neumaier*, Axel Polleres*
Vienna University of Economics and Business, Vienna, Austria
Email: *firstname.lastname@wu.ac.at

***Abstract*—Despite the enthusiasm caused by the availability of a steadily increasing amount of openly available, structured data, first critical voices appear addressing the emerging issue of low quality in the meta data and data source of Open Data portals which is a serious risk that could disrupt the Open Data project. However, there exist no comprehensive reports about the actual quality of Open Data portals. In this work, we present our efforts to monitor and assess the quality of 82 active Open Data portals, powered by organisations across 35 different countries. We discuss our quality metrics and report comprehensive findings by analysing the data and the evolution of the portals since September 2014. Our results include findings about a steady growth of information, a high heterogeneity across the portals for various aspects and also insights on openness, contactability and the availability of meta data.**

## I. 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 services or the chance to outsource innovative use of the published data [1]. However, first critical voices appear addressing – to the public – the emerging issue of low quality for the meta data and data sources in data portals which is a serious risk that could disrupt the Open Data project.[1]

The risk of low (meta-)data quality affects the discovery and consumption of a dataset in a single portal and across portals. On the one hand, missing meta data directly affects the search and discovery services to locate relevant and related datasets for particular consumer needs. On the other hand, incorrect descriptions of the datasets pose several challenges for their processing and integration with other datasets. These risks are common to any search and data integration scenario and there exist, to the best of our knowledge, no comprehensive, quantitative and objective reports about the actual quality of Open Data portals.

We present in this work *Open Data Portal Watch*, a framework that monitors and assess the quality of 82 Open Data portals, which are powered by the CKAN software[2], since September 2014 and report on quality issues by analysing over 160k datasets and 512k resources. We start by highlighting related efforts in assessing the meta data quality in the context of Open Data and present the meta data structure of CKAN portals (§ II). Next, we introduce our intrinsic and contextual quality metrics along 6 dimensions (§ III) and present our publicly available framework (§ IV). We provide a detailed overview about the overall landscape and heterogeneity of the 82 portals, present our findings by analysing the various quality metrics and give insights into evolutionary pattern (§ V). Eventually, we conclude and highlight future directions (§ VI).

## II. Background

Data quality assessment (QA) and improvement methodologies are widely used in various research areas such as in relational databases, data warehouses, information or process management systems, but also to assess the quality of Linked Open Data [2]. Over times, different areas established catalogues of various measures and techniques to assess the quality of data and services and to keep up with the increasing complexity of the tasks [3]. Batini et. al.[4] published in 2009 a detailed and systematic description of methodologies to assess and improve data quality.

### A. Open Data Quality Assessment

Various efforts already exist to study different aspects of Open Data portals which are the main platforms to publish and consume datasets. For instance, the Open Data Barometer project assesses the readiness of countries to exploit their Open Data efforts and measures the achieved impact based on expert judgements.[3] Similarly, the Open Data Census provides a survey to analyse the data of portals in more detail.[4] In addition, the Open Data Monitor project was recently released which also aims to provide a general overview about various data portals without the focus on quality assessment. [5] More related to the actual data quality assessment is the OPQUAST project[6] which provides a checklist for Open Data publishing, including questions related to quality aspects. In relation to data quality assessment in Open Data portals, such as CKAN, recent work discussed the quality of such catalogues and general improvement techniques, however the authors do not inspect the actual data [5]. An earlier survey in 2012 analysed 50 Open Data portals wrt. standardization, discoverability and machine-readability of data [6]. Unfortunately, formula and details are mainly missing in the article and it is unclear how exactly the findings where derived and what data was used.

Most closely related to our effort is the work of Reiche et.al. [7] that also identified the need for an automatic quality assessment and monitoring framework to better understand quality issues in Open Data portals and to study the impact of improvement methods over time. The authors developed a prototype of such a framework which is unfortunately now offline.[7]

To the best of our knowledge, this is the first work which continuously analyses a large set of Open Data portals, reports

[1] http://www.business2community.com/big-data/open-data-risk-poor-data-quality-01010535
[2] http://ckan.org/instances/

[3] http://opendatabarometer.org, last accessed 2015-02-11
[4] http://census.okfn.org
[5] http://www.opendatamonitor.eu/frontend/web/index.php
[6] http://checklists.opquast.com/en/opendata
[7] http://metadata-census.com, last accessed 06.03.2015

```
d: {
  "license_id": "cc-by",
  "author":"National ...",
  ...
  "extras":{
     "schema_language": "ger",
     ...
     "k_m^e": value(k_m^e),
  },
  "resources": [
    {
      "format":"CSV",
      "url": r_1,
      ...
      "k_k^r": value(k_k^r),
    },{ "format":"RDF", ··· }
  ],
  ...
  "k_n^c": value(k_n^c)
}
```

Fig. 1. High level structure of the meta data for a CKAN dataset.

in detail about the quality metrics and provides comprehensive findings.

### B. Open Data Portals (CKAN)

There exists two prominent frameworks for publishing Open Data, i) the commercial Socrata Open Data portal and ii) the open source framework CKAN, developed by the Open Knowledge Foundation and proudly advertising 117 deployed public instances.[8] Both frameworks provide an ecosystem to describe, publish and consume datasets. Since this work focuses on monitoring CKAN portal, we now discuss the necessary meta data schema and provide the formal definition used in the remainder of this work.

The central entities in any CKAN portal are datasets which contain general meta data to describe important contextual information about the dataset itself and the actual data resources, such as the publisher, used license, the data format or its encoding. Figure 1 shows an excerpt of a meta data description for a CKAN dataset ($d$) in the JSON format. We distinguish in our work between three categories of meta data keys in a CKAN portal:

**core keys:** a set of predefined keys which are generic and restrictive and by default available in any CKAN portal, such as the `license_id` key in Figure 1.

**extra keys:** a set of arbitrary additional meta data keys to describe a datasets defined by the portal provider. These keys are listed under the `extras` key (cf. `schema_language` in Figure 1)

**resource keys:** a mix between some default keys and additional keys defined by the portal provider to describe the particular resources (e.g., a datafile or also an API). Each resource is described under the `resources` key.

*1) Formal definition of a CKAN portal:* Formally, let $p$ be a CKAN portal consisting of $|p| = m$ datasets ($\mathcal{D}(p) = \{d_1, d_2, \cdots, d_m\}$) which describe $n$ resources ($\mathcal{R}(p) = \{r_1, r_2, \cdots, r_n\}$). Let the function $\mathrm{res}(d) \in \mathcal{R}$ denote all resources described by a dataset $d$ and $\mathrm{res}(p) \in \mathcal{R}$ denote all resources described in a portal $p$, respectively. Next, let $\mathrm{keys}(p) \subseteq \mathcal{K} = \mathcal{K}^{\mathcal{C}} \cup \mathcal{K}^{\mathcal{E}} \cup \mathcal{K}^{\mathcal{R}}$ return the set of used meta data keys for a portal $p$, with $\mathcal{K}^{\mathcal{C}} = \left\{k_1^c, k_2^c, \cdots, k_n^c\right\}$ be the set of core meta data keys, $\mathcal{K}^{\mathcal{E}}$ the set of extra meta data keys and $\mathcal{K}^{\mathcal{R}}$ the set of meta data keys for resources (cf. Figure 1). Eventually, let $\mathrm{keys}(.) \subseteq \mathcal{K}$ be the set of keys used in a portal $p$, dataset $d$ or resource $r$ and $\mathrm{keys}(.|\mathcal{K}^*) \subseteq \mathcal{K}^*$ be the keys belonging to e specific key set ($\mathcal{K}^*$) used in a dataset $d$.

[8]http://ckan.org/instances/

Table I presents the necessary overview of our notation used in the remainder of this work.

TABLE I. META DATA KEY SETS

| | |
|---|---|
| $\mathcal{K} = \mathcal{K}^{\mathcal{C}} \cup \mathcal{K}^{\mathcal{E}} \cup \mathcal{K}^{\mathcal{R}}$ | Set of all available meta keys |
| $\mathcal{K}^{\mathcal{C}}$ | Set of all available **core** meta keys |
| $\mathcal{K}^{\mathcal{E}}$ | Set of all available **extra** meta keys |
| $\mathcal{K}^{\mathcal{R}}$ | Set of all available **resource** meta keys |
| $\mathrm{keys}(.) \subseteq \mathcal{K}$ | All unique keys used in a portal $p$, dataset $d$ or resource $r$ |
| $\mathrm{keys}(.\|\mathcal{K}^*) \subseteq \mathcal{K}^*$ | All unique keys belonging to a certain set $\mathcal{K}^*$ used in a portal $p$, dataset $d$ or resource $r$ |

## III. QUALITY DIMENSIONS & METRICS

Next, we discuss in detail our six quality dimensions and metrics which are *retrievability*, *usage*, *completeness*, *accuracy*, *openness* and *contacability* (cf.Table II for a short description). Our metrics are partially aligned with existing ones [7] and extended by the openness and contactability dimension. We selected the metrics that can be assessed in an scalable, automated and objective way and next, provide the necessary details and formula in the remainder of this section.

TABLE II. BRIEF DESCRIPTION OF OUR QUALITY METRICS.

| | 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. |

### A. Retrievability

Our first quality metric is the *retrievability* of dataset information for a portal and of the actual resources. The metric measures if a legal or software agent can retrieve the content of a portal and its resources without any errors or access restrictions.

*Definition 1 (Retrievability):* The degree to which the description of a dataset and the content of a resource can be retrieved based on an `HTTP GET` operation. Let the function $\mathrm{status}(d)$ ($\mathrm{status}(r)$) return the HTTP response status code of a `HTTP GET` request for a particular dataset $d$ or resource $r$ both identified by a URL. Further, let $\mathrm{ret}(d) = 1$ if $\mathrm{status}(d)$ equals 200, otherwise 0, analogously for $\mathrm{ret}(r)$. The aggregated average retrievability for the datasets (resources) of a portal is defined as:

$$Q_r^{\mathcal{D}}(p) = \frac{\sum_{d \in p} \mathrm{ret}(d)}{|p|} \qquad Q_r^{\mathcal{R}}(p) = \frac{\sum_{r \in res(p)} \mathrm{ret}(r)}{|res(p)|}$$

## B. Usage

Our second quality dimension is the availability or *usage* of meta data keys across the datasets of a portal. We use this measure since we observed that not all portals make all meta data keys available to the data publishers or because keys can be left out if publishers use the CKAN API. While this usage metric is a rather weak quality measure, it can be used either as a weight for other quality formula or as a filter, e.g., one can compute a certain metric by considering only the keys which are used in all datasets. We define the usage quality metric as follows:

*Definition 2 (Usage):* The usage defines the degree to which the available meta data key are used in the datasets for a given portal. The general key usage of a dataset is:

$$\text{usage}(d) = \frac{|\,\text{keys}(d|\mathcal{K}^{C} \cup \mathcal{K}^{\mathcal{E}})| + \sum_{r\in d} |\,\text{keys}(r)|}{|\,\text{keys}(p|\mathcal{K}^{C} \cup \mathcal{K}^{\mathcal{E}})| + (|\,\text{keys}(p|\mathcal{K}^{\mathcal{R}})| * |res(d)|)}$$

and for a portal as the average of the key usage per dataset:

$$\mathrm{Q_u}(p) = \frac{\sum_{d\in p} \text{usage}(d)}{|p|} \text{ with}$$

The usage metric can be easily modified to compute the metric for only a particular set of keys.

## C. Completeness

The completeness of the meta data description is a widely used and important measure to provide an indication of how much meta information is available for a given dataset.

*Definition 3 (Completeness):* The completeness of a portal is the degree to which the available meta data keys to describe a dataset have non empty values.

Slightly reformulating the metric in [7], we define the completeness function for a key $k$ and a dataset $d$ as $\text{compl}(k, d)$, returning 1 if $k \in \mathcal{K}^{C} \cup \mathcal{K}^{\mathcal{E}}$ and if the value of key $k$ in $d \neq Null$, If the key is a resource key ($k \in \mathcal{K}^{\mathcal{R}}$), the function returns the average completeness of $k$ over all resources in $d$ ($\sum_{r\in res(d)} \text{compl}(k, r)/|res(d)|$), otherwise return 0.

$$\text{compl}(d) = \frac{\sum_{k\in key(d)} \text{compl}(k, d)}{|key(d)|}$$

$$\mathrm{Q_c}(p) = \frac{\sum_{d\in p} \text{compl}(d)}{|p|} \text{ with}$$

## D. Accuracy

A common definition of the accuracy metric is the degree of which the available meta data values accurately describe the actual data . One can distinguish between a syntactic and semantic accuracy [4]. Syntactic accuracy is defined as the closeness of a value to the corresponding definition domain, e.g., does the value of an author email correspond to the email format, or do date values conform to a particular date format. On the other hand, semantic accuracy compares the value with its real-world value, e.g, comparing the content size value with the real content size [7]. In general, we define the accuracy of a meta data key for a portal as follows:

*Definition 4 (Accuracy):* The accuracy is the degree of closeness between meta data values and their actual values. In general, let $\text{accr}(k, r)$ be the distance function for a certain key and a resource. Further, let $\text{accr}(k, d) = \sum_{r\in \text{res}(d)} \text{accr}(k, r)/|\,\text{res}(d)|$ be the average accuracy for key $k$ in a dataset $d$ over all dataset resources. As such, let the overall accuracy for portal $p$ and a key $k$ be

$$\mathrm{Q_a}(k, p) = \frac{\sum_{d\in p} \text{accr}(k, d)}{|p|}$$

As of now, we only compute the semantic accuracy for different keys which describe a resource, which is ideally measured by inspecting the content of all resource files. However, retrieving the actual data can be very resource consuming in terms of bandwidth and disk space. As such, we decided to perform for all resources a `HTTP HEAD` lookup, store the response header and automatically compute the accuracy values using these header information for the following keys:

**format** (*file format accuracy*): To check the accuracy of the specified format value for a given resource, we firstly normalise the specified meta data value (e.g., mapping ".csv" to "csv") and compare it to the file extension of the resource, if available. In addition, we also take into account the format specification in the content-type header field. Figure 2 lists the pseudo code of our algorithm to compute the format accuracy if a file-extension and/or header format information is available.

```
def format_accuracy(meta_data, resource):
  score = 0
  count = 0
  if meta_data.format is not None:
    // check file extensions
    if resource.extension is not None:
      count += 1
      if get_format(resource.extension) == meta_data.format:
        score += 1
    // check mime type
    if resource.header.mime_type is not None:
      count += 1
      for ext in guess_extensions(resource.header.mime_type):
        if get_format(ext) in meta_data.format:
          score += 1
          break
  return score/count
```

Fig. 2. Code fragment for calculating the format accuracy value.

**mime-type** (*mime-type accuracy*): We compare the specified resource meta data mime-type value with the value of the content-type header field.

**size** (*content size accuracy*): The accuracy of the specified content size can be computed based on the header information or the actual resource if downloaded.

## E. Openness

Our next metric focuses on measuring how "open" the dataset of a portal is.

*Definition 5 (Openness):* The openness of a portal is the degree to which datasets provide a confirmed open license and if the resources are available in an open data format. Let $\text{open}(d)$ be a user defined function that determines the openness of a dataset based on the license (subscript $l$) or based on the available formats for the resources of a dataset (subscript $f$). The average openness of a portal is computed as follows:

$$\mathrm{Q}_{\mathrm{o}}^{\mathrm{l}}(p) = \frac{\sum_{d \in p} \mathrm{open}_l(d)}{|p|} \qquad \mathrm{Q}_{\mathrm{o}}^{\mathrm{f}}(p) = \frac{\sum_{d \in p} \mathrm{open}_f(d)}{|p|}$$

Similar to the accuracy measure, one can define a semantic distance of how "open" a license or format is, e.g., usage of the data is allowed but not the redistribution of modified values or the format is not fully open but there exists open source tools. However, the problem with this approach is that it is very complex to define such a distance for all licenses and formats.

We confirm the license openness per dataset by evaluating the specified license against the list provided by the Open Definition[9]. This list contains details about 108 different licenses including their typical id, url, title and an assessment if they are considered as "open" or not. The license information of a dataset in CKAN can be described with three different CKAN keys, namely *license_id*, *license_title* and *license_url*. Our algorithm tries to match a dataset license to one of the defined ones in the list by performing the following steps. Firstly, we try to perform the match using the `license_id` value, if available. If this check fails we use next the `license_title`, which is match either against the id or title in the opendefintion license list. [10] If this check also fails, we use as a fall back solution the `license_url` value for the match. Once a match was successful we decide on the openness based on the assessment of the open definition. Note, that as such, our metric reports only a value about the confirmed licenses and it might be that the non-confirmed licenses are also adhering to the open definition.

The format openness metric has to consider that a dataset can have various resources with different formats. We label a dataset as open as soon as one resource of the dataset has an open format. Regarding the openness of a format we currently use the following set of file formats:

$$\{csv, html, latex, dvi, postscript, json, rdf, xml,$$
$$txt, ical, rss, geojson, ods, ttf, otf, svg, gif, png\}$$

Please note that we excluded formats such as *zip* or *xls* (Microsoft Excel) since there exists not yet a clear agreement if they should be considered as open or closed. Again, we only can measure the confirmed open formats and might miss other formats that are considered as open but not included in our list. However, we can easily adapt this by including new formats or licenses as required and identified.

### F. Contactability

Another important issue concerning datasets in Open Data portals is the contactability of their creates/maintainers, that is, if information are available to a user to contact the data provider.

*Definition 6 (Contactability):* The degree of which the datasets of a portal provide a value, an email address or HTTP URL to contact the data publisher. To provide less restrictive contactability results, we define the $\mathrm{Q}_{\mathrm{i}}^{\mathrm{v}}$ metric, indicating that the dataset has some kind of contact information by adapting the completeness metric for a particular set of keys. Let $\mathrm{cont}_v(d)$ return 1 if and only if $\mathrm{compl}(k, d) = 1$ for one of the following CKAN meta data fields: `maintainer`, `maintainer_email`, `author` and `author_email`.

$$\mathrm{Q}_{\mathrm{i}}^{\mathrm{v}}(p) = \frac{\sum_{d \in p} \mathrm{cont}_v(d)}{|p|}$$

Further, let $\mathrm{email}(d)$ be a verification function that returns "1" if a dataset $d$ has an email address and "0" otherwise, respectively, let $\mathrm{hasURL}(d)$ be the function to denote if a dataset has a maintainer or author http address.

$$\mathrm{Q}_{\mathrm{i}}^{\mathrm{e}}(p) = \frac{\sum_{d \in p} \mathrm{email}(d)}{|p|} \qquad \mathrm{Q}_{\mathrm{i}}^{\mathrm{u}}(p) = \frac{\sum_{d \in p} \mathrm{hasURL}(d)}{|p|}$$

## IV. Quality Assessment Framework

We developed a monitoring framework, termed "*Open Data Portal Watch*", to continuously assess the quality of CKAN portals similarly to [7]. The components of our framework are depict in Figure 3. The *fetch component* periodically retrieves the dataset information of a given portal and store the meta data in a document store. We currently use the CKAN API Version 1 which is supported by all portals. Newer API versions (e.g. version 3) are not available for all portals and also return a slightly different JSON format which would require some internal transformations. The stored information are analysed by the *quality assessment component* which computes our defined quality metrics for the various dimensions. A publicly available dashboard component[11] displays vital quality metrics for each portal using various views and charts. An example of a "portal-evolution view" for an Austrian portal is depict in Figure 4. 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). This example illustrates the usefulness of the monitoring and quality assessment over time, since we can clearly see how added or removed datasets influence certain quality metrics.

The fetching and analysis code is implemented in Python and all data are stored in a MongoDB instance. The frontend dashboard is based on NodeJS and various JavaScript libraries (e.g. the jQuery library for table rendering and interaction and D3.js for the visualisations). We also make all snapshots of collected raw meta data for all monitored portals publicly available to motivate and engage other researchers in analysing it.[12]

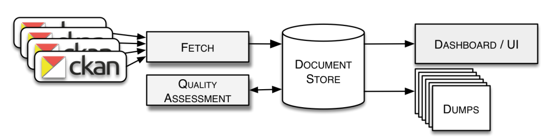



Fig. 3. The *Open Data Portal Watch* components

## V. Findings

In this section, we present our comprehensive findings about the current landscape of 82 active Open Data CKAN portals, including quality assessment results and observed evolutionary patterns. Initially, we had 89 portals in our monitoring list, however, 3 portals send a `403 FORBIDDEN`

[9] http://opendefinition.org/

[10] We perform the additional id match because we observed that in several cases the datasets contain the license id in the license title field.

[11] http://data.wu.ac.at/portalwatch/

[12] http://data.wu.ac.at/portalwatch/data


Fig. 4. Screenshot of a evolution view of a portal.

response code upon an API lookup, 2 portals went off-line in 2015 and 2 portals continuously return connection time-out errors. The full list of all current portals is available at our public framework interface.[13] This section discuss the results for the portal snapshots gathered in the last week of February 2015.

### A. Portals overview

Currently, we actively monitor 82 CKAN portals, consisting of 160K datasets describing 512k resources. We found 180 different unique license IDs, 724 file format descriptions and ~78k tags (see Table III). The portals use a total of ~3.1k different meta data keys, of which 68 keys belong to the core or default keys ($\mathcal{K}^C$), 2906 to the extra keys ($\mathcal{K}^E$) and 261 unique keys are used to describe resources ($\mathcal{K}^R$).

TABLE III. Basic statistics of 82 portals

| $\lvert\mathcal{D}\rvert$ | $\lvert\mathcal{R}\rvert$ | $\lvert\mathcal{K}^C\rvert$ | $\lvert\mathcal{K}^\mathcal{E}\rvert$ | $\lvert\mathcal{K}^\mathcal{R}\rvert$ | Licenses | Formats | Tags |
|---|---|---|---|---|---|---|---|
| 160069 | 512543 | 68 | 2906 | 261 | 180 | 724 | 78179 |

Regarding the portal sizes, Table IV shows the distribution of the portals based on their datasets and resources. Please not that the table cells should be interpreted as intervals. We can clearly see that half of the portals have less than 500 datasets or resources and around 25% of all portals are in the range between $10^3$ – $10^4$ datasets or resources.

TABLE IV. Number of Open Data Portals with a given size.

| | $<10^2$ | $<5\times10^2$ | $<10^3$ | $<10^4$ | $<5\times10^4$ | $<10^5$ | $>10^5$ |
|---|---|---|---|---|---|---|---|
| $\lvert\mathcal{R}\rvert$ | 10 | 28 | 11 | 22 | 9 | 1 | 1 |
| $\lvert\mathcal{D}\rvert$ | 29 | 30 | 3 | 16 | 3 | 1 | 0 |

We located 512k values for the `url` resource meta key, of which 328k are unique and syntactically valid URIs. Overall, we managed to successfully download 248k unique resource URLs, resulting in a total content size of around 1.09TB.

*a) Portal overlap:* The difference between total resource URL values and unique once indicates that resources are multiple times described, either in the same datasets, portals or across portals. A closer look reveals that 93k unique resource values (28%) appear more then once, out of which the majority of 79k resources are described in datasets in different portals and the remaining 14k resources are described in the same portal several times. Surprisingly, looking into the overlapping portals, we discovered that the majority of ~71k resources appear in the Pan European data portal[14]. The main aim of this portal is to harvest other European portals to provide a single-point of access. The Pan European data portal itself contains 140k resources in total, with 108k unique resource URIs. Overall, these findings show a very small overlap across the portals with only 71k out of 312k URIs are described in more than one portal.

*Heterogeneity:* The low overlap of resources across portals already indicate that currently the portals can be seen as data silos. This suspicion becomes even stronger if we look at the overall used extra meta data keys, tags and format values.

A first interesting observation is that out of the 2906 used extra meta data keys a total of 2383 keys appear in only one portal, indicating that the extra keys are highly portal specific and not much re-use or alignment happens between portals. We found only 523 keys in more than one portal of which 174 are used in more than two portals. Only 2 keys are shared in more than 14 portals. Similar observations can be found for the tags used to describe a dataset. Out of the 78179 used tags, 50448 appear in exactly one portal, 27731 in more than one portal and 8894 in more than two portals. In addition and surprisingly, we discovered 724 different values to describe the format of a resource. The main reason for this is that there exists no standards for describing the resource formats. For instance, we observed several values for the comma-separate file format such as, *csv*, *comma-separate-values*, *character-separate-values* or *csv-format*, just to name a few.

### B. Retrievability ($Q_r$)

Table V shows the results of our dataset and resource retrievability analysis. We grouped the response codes by their first digit; *others* indicate socket or connection timeouts. As expected, nearly all datasets could be retrieved without any errors. The 641 datasets that could not be retrieved responded with a 403 Forbidden HTTP status code, indicating that an access token is required to retrieve the information. A slightly different picture can be observed if we try to retrieve the content of the actual resources. As mentioned above, out of the 512k described resource, only ~328k are unique valid URLs. We performed lookups on 313k URIs, resulting in the response code distribution in Table V. An slightly alarming observation is that 30k described resources point to a non-existing data source and returned a response code of 4xx and 21k resources caused some socket or timeout exception upon the lookup (indicated with others).

TABLE V. Distribution of response codes.

| | № | 2xx | 4xx | 5xx | others | 3xx |
|---|---|---|---|---|---|---|
| $\mathcal{R}$ | 313733 | 248855 | 30788 | 3638 | 21484 | 8950 |
| $\mathcal{D}$ | 160710 | 160069 | 641 | 0 | 0 | - |

Overall, the retrievability of datasets is very good and wrt. the resource retrievability, we successfully downloaded the content of 79% of the tested resources.

[13] http://data.wu.ac.at/portalwatch/overview/

[14] https://publicdata.eu/

### C. Meta Data usage ($Q_u$) and completeness ($Q_c$)

Next, we analyse the usage and completeness of meta data keys across all 82 portals. Figure 5 plots the average usage ($Q_u$) against the average completeness ($Q_c$) for each portal for the three different key subsets with the $Q_c$ distribution on the horizontal and $Q_u$ distribution on the vertical axes of the plot. The distributions also contain the total values over all keys (black bars) for a general overview.

Looking at the histogram on the vertical axis, we observe that 60% of the portals have an average $Q_u$ value per dataset and all keys of more than 0.9. Drilling deeper, we see that nearly all portals have a $Q_u$ value of over 0.9 for the core meta data keys (red bar) and around 80% of the portals have a $Q_u$ value of over 0.9 for the resource meta data keys (green bar). In contrast, the usage value for extra meta data keys is widely spread across the portals with around 50% of the portals having a value of less than 0.3 (see the lower part of the axis and the blue bars).

One explanation for the low usage values of the extra keys might be that for these particular portals the datasets are mainly uploaded by software agents using the CKAN API which does not require that all keys are used and thus are left out. In contrast, a high usage value for portals might be because the datasets are mainly created by using the UI for humans. This UI has a predefined mask using the full set of keys. In addition, the better usage value for the core and resource keys might be because those keys become more standard and documented (e.g. on CKAN documents) and as such, are known to the data publishers and portal specific extra keys might be not well documented or advertised.

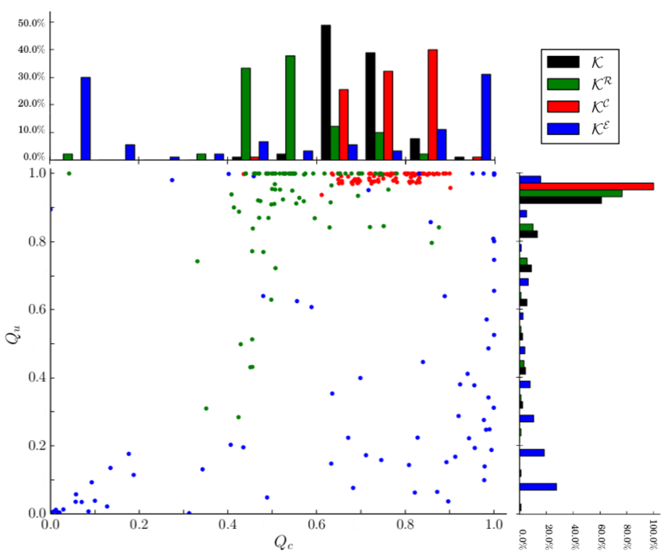



Fig. 5. Usage and completeness distribution.

The horizontal histogram shows the respective completeness distribution for the different key sets. Overall, we can observe that the majority of the portals have an average meta data key completeness value in the range of 0.6 - 0.8 and only a few portals have value of over 0.9. Inspecting the key subsets, we can see that the overall values are highly influenced by the set of extra meta data keys which have a completeness value of less than 0.1 for around 33% of the portals. In contrast, over 40% of the portals have a $Q_c$ value of over 0.8 for the core keys. Looking at the key set used to describe resources, we also observe that the majority of the portals provide a specific $Q_c$ value between 0.4 and 0.6.

The scatter plot helps to highlight groups of portals with different average usage and completeness values. For instance, we can see in the bottom left part a group, covering around 35% of the portals, for which the extra keys show very low usage and completeness values. In such a case, a portal owner could rethink the necessity of the extra keys.

### D. Openness ($Q_o$)

It is crucial for the Open Data movement that published datasets and formats are adhering to the open definition and that everybody is allowed to use, re-use and modify the information which should be provided in an open format. Table VI shows the top-10 used licenses per dataset and top-10 used formats per total and unique resources together with their number of portals they appear in. Bold highlighted values indicate that the license or format is considered by our metric as open. Please note, that we count the number of datasets for the licenses and the number of resources for the formats. The

TABLE VI. TOP-10 FORMATS AND LICENSES.

| license_id | № | $\|p\|$ | format | № | $\|p\|$ | unique |
|---|---|---|---|---|---|---|
| *empty* | 32428 | 32 | **csv** | 122828 | 74 | 88458 |
| **cc-by** | 27207 | 53 | *empty* | 113094 | 55 | 52632 |
| **uk-ogl** | 22999 | 11 | xls | 48404 | 63 | 37619 |
| **cc0** | 7254 | 1 | pdf | 37855 | 66 | 32145 |
| dl-de-by-1.0 | 6961 | 2 | **html** | 30588 | 51 | 16194 |
| other-nc | 5279 | 14 | wms | 17827 | 22 | 4959 |
| http://...° | 5096 | 1 | **xml** | 17721 | 53 | 13488 |
| dl-de-by-2.0 | 4907 | 1 | **json** | 11931 | 51 | 8251 |
| **cc-zero** | 4816 | 25 | kartenviewer | 11307 | 3 | 2986 |
| *notspecified* | 4685 | 52 | zip | 10469 | 54 | 7294 |
| others | 34859 | | others | 90519 | | 72060 |

° http://open-data.europa.eu/kos/licence/EuropeanCommission

first surprising observation is that ∼23% of the datasets and ∼22% of all the resources have no license or format specified (see the italic values in Table VI). Secondly, we see that the confirmed open licenses in the top-10 cover only ∼39% of all datasets. Similarly, the top-10 used open formats cover only 35% of all resources. Again, note that for now we do not consider *zip* and *xls* files as open. Another interesting observation is that only 1396 unique resources, which appear in more than one portal, have more than one different format description in two or more datasets. This indicates that the format description across datasets for the same resource seems to be very stable. In addition, Figure 6 shows the distribution of the average $Q_o$ values per portal. From the plot we can see that around 40% of the portals have an average license openness value of over 0.9 and around 30% of the portals have an format openness value of over 0.9. There is also a group of around 20% of the portals for which we could only confirm an average license openness per dataset of less than 0.1. The average values for the remaining portal spread more or less equally from 0.1 to 0.9.

Overall, we could not confirm for the majority of the portals that the datasets provide an open license and their resources are available in open formats. In future work, we will investigate methods to address the unconfirmed licenses and formats.

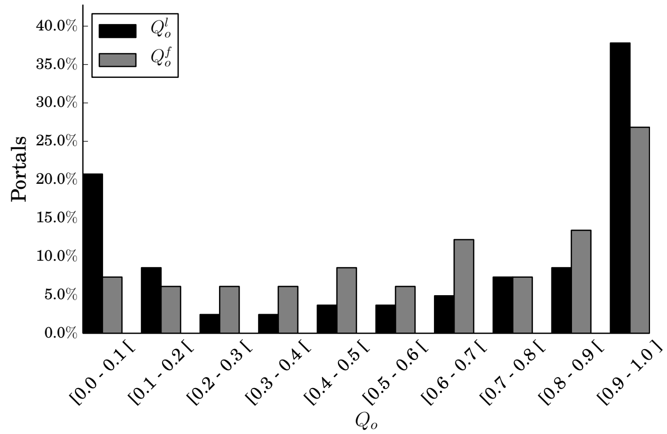



Fig. 6. Distribution of $\mathrm{Q_o}$ metrics.

## E. Contactability ($\mathrm{Q_i}$)

Next, we report on our findings regarding the contactability information provided by the datasets, plotted in Figure 7. Firstly, considering the availability of any information for contacts, we can see that around 50% of the portals have an average $\mathrm{Q_i^v}$ value of over 0.9 and 15% of the portals a respective value of less than 0.1.

Regarding the contactability by email, we discovered that a subset consisting of 40% of the portals have an average $\mathrm{Q_i^e}$ value of over 0.8 and 28% of the portals do not really contain any email information (average $\mathrm{Q_i^e}$ value of $< 0.1$). The remaining 30% of the portals are more or less equally spread over the range of 0.1 – 0.8.

Regarding the appearance of URLs for either the author or maintainer contact values, we observed an average URL contactability over all portals of 0.0024, meaning that there are basically no URLs provided for contacting the publisher or maintainer of a dataset.

Overall, we can conclude that the majority of the portals have a low contactability value which bears the risk that data consumers stop using dataset if they cannot contact the maintainer or author (e.g., regarding the re-use if the license is not clearly specified or in case of any data related issue).

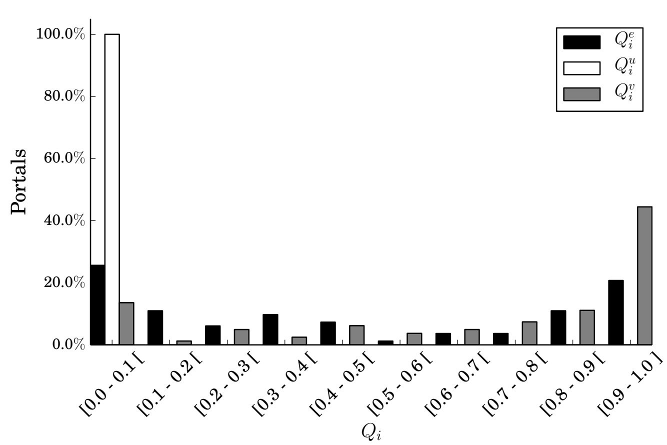



Fig. 7. Distribution of $\mathrm{Q_i}$ metrics.

## F. Accuracy ($\mathrm{Q_a}$)

Our current accuracy analysis is based on header information from the available resource URLs. We performed in total 274k HTTP HEAD lookups, of which 226k successfully returned HTTP response headers and 223k contained the *content-type* field and 183k a *content-length* field. Considering datasets for which we have meta data values available and resources with a HTTP GET response header, we compute the format accuracy $\mathrm{Q_a}(format, .)$ for 72k datasets, the mime-type accuracy for 69k datasets and the size accuracy for only 14k datasets.

Figure 8 shows the $\mathrm{Q_a}$ distribution of the average accuracy per portal. We can see that there exists a subset of 35% of the portals for which the meta data description about the content size is highly accurate with the header content size information of the resource, if available (see $\mathrm{Q_a}(size)$). Similarly, we observe for the mime-type that ~28% of the portals provide in average a mostly correct mime-type information for the describe resources in the datasets. Regarding the file format, we observe that the provided formats information in general do not match with the derived file format from either the file extension or the header. One reason might be that there are over 700 different variations of format descriptions in the datasets which can cause many incorrect format matches.

Overall, we derive two main findings. Firstly, the results reflect only a subset of the datasets and resources since we rely only on header information and not on the actual file content. However, the results show that those header information are not very precise and that there exists a mismatch between the provided meta data and the header information of the resources. Secondly, to provide highly accurate measures we need to download and monitor the actual content of the resources and also need to improve the handling of the various format descriptions .

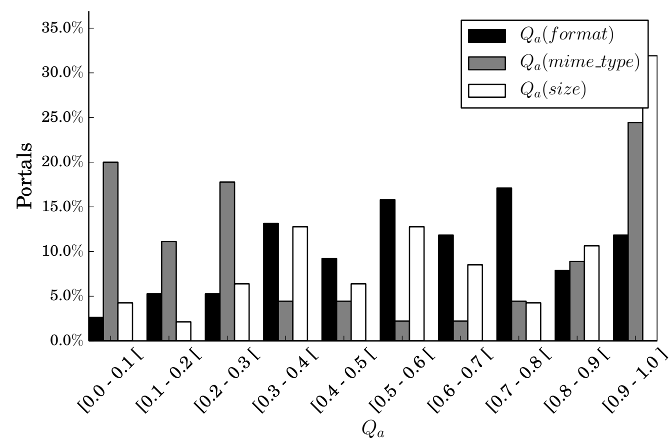



Fig. 8. Accuracy distribution of the keys *mime_type*, *format* and *size*.

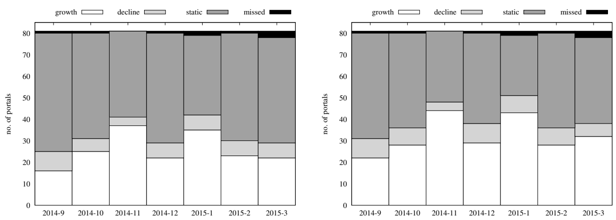
## G. Evolution

Eventually, we present the most interesting evolutionary patterns for our 82 portals.

*1) Growth-rate of portals:* Overall, we observe a steady growth in the number of datasets and resources in the portals since September 2014. Table VII shows the growth or decline

TABLE VII. Dataset/resource evolution of portals

| | DATASETS | | | RESOURCES | | |
|---|---|---|---|---|---|---|
| | № | *avg.* | | № | *avg.* | |
| GROWTH | 66 | 237 | (+54.11%) | 70 | 1018 | (+89.22%) |
| DECLINE | 4 | -46 | (-4.79%) | 5 | -937 | (-14.98%) |




(a) Dataset evolution (b) Resource evolution

Fig. 9. Evolution of the size of portals over time.

of datasets (resources) if we compare the first and last snapshots of each portal. We can observe that 66 portals increased their number of datasets by an average of 237 or ~54%, while only 4 portals slightly removed datasets (by around ~5% or 46 datasets). We observed similar patterns for the resource growth. The majority of 70 portals added an average of 1018 resources to their portal which corresponds to an average resource growth of 89%, while only 5 portals removed in average 937 resources. Figure 9 shows the number of portals which are static or dynamic (split into decline and growth), aggregated by month. We can see that each month 18 to 40 portals show an increase in datasets or resources, while a small amount of portals show a decline. The plot also contains the number of portals which we could not monitor in one month due to consistent server errors causing either empty results or connection timeouts. We will cater for this bottleneck by periodically rechecking the availability of the portal in the future.

*2) Quality evolution:* Figure 10 shows the evolution of the average quality metric values across all 82 portals starting from September 2014. We excluded the accuracy metric and resource retrievability since we do not have that data for the beginning of our experiment. If we want to conclude a trend from this plot, one can say that in general the portals use more open licenses (yellow line) and also show and increase of contactability information in form of emails. The average usage, completeness and format openness per portal seems to remain stable over time.

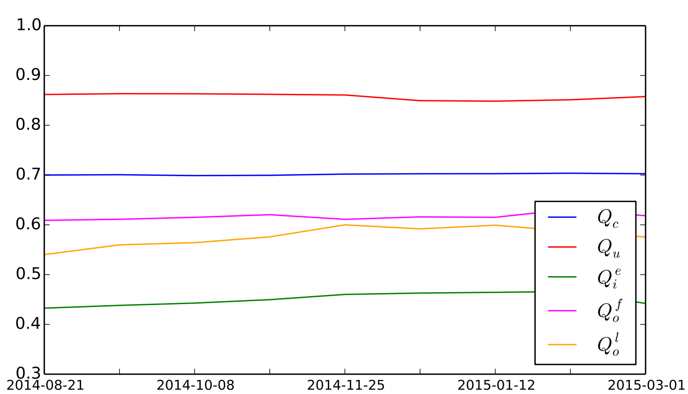



Fig. 10. Evolution of quality metrics.

# VI. Conclusion

As of today, the Open Data movement enjoys great popularity among governments and public institutions and also increasingly in industry, but first critical voices appear addressing the emerging issue of low quality for the meta data in data portals. While first projects emerge to quantify and qualify the quality of Open Data, there exists no comprehensive quality assessment and evolution tracking. We contribute to this efforts by developing a Open Data portal monitoring and quality assessment framework and currently monitor 82 CKAN portals consisting of 160k datasets and 512k resources. Our core findings can be summaries as follows:

- ⋄ We found 328k unique and valid resources URLs of which ~79% can be downloaded, resulting in a total on disk file size of 1.09TB. The most common file formats is currently *csv* (24% of the resources), followed by other structured formats such as xls, JSON or xml.
- ⋄ Strong heterogeneity across portals wrt. format descriptions, extra meta data keys and tags causing serious challenges for a global integration of the portals
- ⋄ We observe a steady growth of datasets and resources for the majority of the portals since Sept. 2014: we measured an average increase of datasets by +54% for 66 portals and an average increase of resources per portal by +89% for 70 portals.
- ⋄ ~50% of the datasets provide a confirmed open license and 20% do not list any license.
- ⋄ Similarly, ~50% of the resources provide a confirmed open format and 21% do not provide any format information.
- ⋄ The majority of the datasets do not provide contact information in form of email addresses or URLs.

Regarding the next improvements for our framework and quality metrics, we prioritise the integration of more portals (also including non CKAN portals), the scalable monitoring of the actual resource content for a better accuracy metric (also expanding on change frequency) and the improvement of the openness metric regarding the various licenses and formats. Eventually, we will research solutions to deal with the high heterogeneity for meta data keys and values in and across the portals with the aim of providing meta data mappings.

# References

[1] M. Janssen, Y. Charalabidis, and A. Zuiderwijk, “Benefits, adoption barriers and myths of open data and open government,” *IS Management*, vol. 29, no. 4, pp. 258–268, 2012.

[2] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, “Quality assessment methodologies for linked open data (under review),” *Semantic Web Journal*, 2014, this article is still under review.

[3] H. Zhu, S. E. Madnick, Y. W. Lee, and R. Y. Wang, “Data and information quality research: Its evolution and future,” in *Computing Handbook, 3rd ed. (2)*, 2014, pp. 16: 1–20.

[4] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, “Methodologies for data quality assessment and improvement,” *ACM Comput. Surv.*, vol. 41, no. 3, pp. 16:1–16:52, Jul. 2009.

[5] J. Kucera, D. Chlapek, and M. Necaský, “Open government data catalogs: Current approaches and quality perspective,” in *EGOVIS/EDEM*, 2013, pp. 152–166.

[6] K. Braunschweig, J. Eberius, M. Thiele, and W. Lehner, “The state of open data - limits of current open data platforms,” in *Proceedings of the 21st World Wide Web Conference 2012, Web Science Track at WWW'12, Lyon, France, April 16-20, 2012*. ACM, 2012.

[7] K. J. Reiche, E. Höfig, and I. Schieferdecker, “Assessment and visualization of metadata quality for open government data,” in *International Conference for E-Democracy and Open Government*, May 2014.