The State of Open Data

Limits of Current Open Data Platforms

Katrin Braunschweig, Julian Eberius, Maik Thiele and Wolfgang Lehner
Technische Universität Dresden
Faculty of Computer Science, Database Technology Group
01062 Dresden, Germany
firstname.lastname@tu-dresden.de

ABSTRACT

Following the *Open Data* trend, governments and public agencies have started making their data available to the public using web portals, web services or REST interfaces. Ideally, making this data available on the web would lead to more transparency, participation and innovation throughout society. However, just publishing the data on the web is not enough. To truly advance the open society, the publication platforms need to fulfill certain legal, administrative as well as technical requirements. In this paper we present a survey of existing Open Government Data platforms, focusing on the technical aspects. We studied over fifty Open Data repositories operated by national, regional and communal governments, as well as international organizations. Features such as standardization, discoverability and machine-readability of data were taken into account. Furthermore, a subset of five repositories was examined in more detail, additionally analyzing data and metadata quality. We introduce a number of aspects of *openness* in order to classify the surveyed repositories and asses the state of Open Data on the web. We point out shortcomings of the existing portals regarding reusability and sketch our vision of an improved Open Data repository.
For detailed information about our Open Data survey please also visit the following website:
`http://wwwdb.inf.tu-dresden.de/opendatasurvey/`

Categories and Subject Descriptors

A.1 [**Introductory and Survey**]; J.1 [**Computer Applications**]: Administrative Data Processing; H.2.m [**Database Management**]: Miscellaneous

General Terms

Measurement, Human Factors, Standardization

Keywords

Open Data, Open Government, Survey

1. INTRODUCTION

The concept of "Open Data" describes data that is freely available and can be used as well as republished by everyone without restrictions from copyright or patents. The goal of Open Data initiatives is to open all non-personal and non-commercial data, especially (but not exclusively) all data collected and processed by government organizations. It is very similar in spirit to the Open Source or Open Access movements. In the course of this trend, public agencies have started making governmental data available on web portals, as web services or via REST interfaces. Ideally, making this data available on the web would lead to more transparency, participation and innovation throughout society.
The published data covers a wide range of domains, from environmental data over employment statistics to the budgets of municipalities. Publishers can be individual government agencies or providers of larger repositories that collect public datasets and make them available in a centralized and possibly standardized way.
Though set up with the same goal in mind, these repositories show very different characteristics. They vary, for example, in size, domains, comprehensiveness, or the application of technical standards. But most importantly, we argue that they vary in their usefulness and suitability to their task. To truly advance the open society, these publication platforms need to fulfill certain legal and administrative, as well as technical requirements.
In this paper, we first introduce a number of technical requirements of an ideal Open Data repository in Section 2. We then present our survey of existing Open Data repositories in Section 3. Our goal is to classify them according to our model and quantify the amount and quality of the data available. We discuss the results of our survey in Section 4 and conclude our work in Section 5.

2. (RE)USING OPEN DATA

Open governmental or municipal data can be a valuable resource of information if published in a useful manner. The current state of Open Data requires citizens who wish to access and use this data to first identify relevant datasets manually. This includes finding organizations or agencies that publish open datasets on platforms that provide a central and responsive entry point where users can search for data. If a single dataset can be found, that contains all the relevant data, the user can directly extract the required information. However, it is rather unlikely to find all relevant data in a single file. Given that several relevant data sets have been found, this loosely coupled information has little

value at this point and needs to be aligned on a technical and semantical level first. This can be very time consuming and hard work. Since each agency or organization probably has it's own publishing policies, the data can be very heterogeneous. Software tool could support users with the integration of their data. After the integration, users can explore the new data, extract the desired information and, if desired, visualize the results. Again, these steps can be performed manually or with the support of software tools.
The way people access and use Open Data is greatly influenced by the way the data is published. Many government agencies or organizations collect large amounts of data. In it's original, raw form, this data is often not very useful for end users. Therefore, many datasets are cleaned and customized before being published. Looking at current Open Data platforms, two to some extend contrary styles to publish the data can be identified. While some publishers prefer the data to be in a human-readable format, others prefer a machine-readable format. We take a look at both styles to see which format is better suited to maximize the benefits of Open Data repositories:
The first approach aims at providing the public with human-readable information instead of raw data. This usually means that existing raw data is transformed into documents containing, for example, aggregated values, visualization or, more generally, simplifications to make data comprehensible. For some use cases, a good human-readable report is sufficient, if it contains the desired information. But it might also miss some important detail or not contain the required information, since it only contains the information that the report's author intended it to contain. Such a manufactured report does not allow users to gain other insights than the ones provided by the author. It does also not allow reuse of the underlying raw data for other use cases. The process of regaining machine-readable data from human-readable documents, i.e. data extraction, can be very challenging and forms its own branch of research in computer science.
The second approach is publishing the raw data in a machine-readable format to enable software tools to process it. Instead of preprocessing the data, users can directly access the raw data and customize the data for their own needs. However, not all users have the expertise required to use the majority of data processing tools available today. Still, if the raw data is available in a machine-readable format, it enables automatic processing. This is a necessary requirement for the development of new, end user friendly analysis tools that all users can benefit from. And since the data is freely available in this machine-readable format, it can be reused for another use case without extra processing.
We argue in favor of the second approach, since it does not limit the number of possible use cases, but instead supports reusability. Therefore, an ideal Open Data platform should be optimized towards technical users and programmatic reuse.

We identified a number of requirement for such a platform, which are listed in the top of Figure 1, grouped into categories regarding standardization, API, materialization, integration and policies. A certain level of standardization is required to enable automatic processing of the data, while an API is required to provide automatic access. Materialization enables better quality control over datasets and prevents users from experiencing problems such as dead download links. Integration is concerned with the establishment

Open Data Platform

Standardization	API
- standardized technical (file-) formats - common metadata properties - concept hierarchies (common tags, domains etc)	- API existing - access to metadata - access to datasets - fine-grain access(e.g. row-based)

Integration	Policies	Materialisation
- common set of allowed metadata property values - common metamodel / domain ontology	- common (free) licenses - provenance metadata - up-to-dateness policies	- just a collection of links - all datasets mirrored - centralized database

Dataset

Discoverability	Format
- existence of - descriptions - tags	- machine readable - non-proprietary - separation of metadata and data

Validity	Quality	Granularity
- area (e.g. state, county) - time of validity	- clear attribute semantics - no missing values	- raw data instead of highly aggregated data

Figure 1: Possible Features of an Open Data Plaforms and Datasets

of uniformity across multiple datasets, which enables users to combine data from different datasets. Finally, policies are important to ensure that users are allowed to access the data but also to enable provenance.
We further identified requirements for individual datasets, which are also illustrated in Figure 1. Discoverability is an important factor for datasets. Without sufficient metadata, such as descriptions or tags, neither manual nor automatic search can find the dataset and it will not be helpful for any user. Furthermore, a machine-readable file format is important to support automatic tools. Features regarding validity, quality and granularity are required to support a wide range of use cases and enable analysis results of high quality.
The discussed Open Data platform and dataset features form the basis of our survey which is covered by the next two sections.

3. A SURVEY OF OPEN DATA REPOSITORIES

3.1 Methodology

With the derived requirements in mind, we surveyed about fifty Open Data repositories operated by national, regional and communal governments, as well as international organizations. We based our list of repositories on a catalog published by the Open Knowledge Foundation[1]. Then, we expanded and refined this list using the following steps: (1) We diversified the list by adding more repositories from so far unrepresented continents. (2) We added regional and communal repositories in addition to national initiatives. (3) We included both official, i.e., state- or agency sponsored platforms, as well as community-driven efforts. (4) We removed very small or inactive repositories, except for countries for which we could not find other repositories. The full list can be found in the appendix and on the survey web-

[1] http://lod2.okfn.org/eu-data-catalogues/

site. It forms the basis of our global assessment of open data repositories, which is presented in Section 3.2. In addition to that, a subset of five repositories was examined in more detail. The results of this detailed analysis are presented in Section 3.3.

3.2 Global View

For the first part of the survey, we studied features of open data platforms that can be measured by just browsing the page or writing a simple crawler for platforms that do not provide statistics on their content. More complex features that require downloading and analyzing the available datasets were studied for a limited number of platforms in the second part of our survey (see Section 3.3).

The goal of the global survey is to assess each repository's suitability for automatic reuse of the data in general. Specifically, we surveyed ten features, which will be described below. For every platform, we also recorded the country of origin, the administrative level it represents (country, region, city, etc.), as well as policies regarding licensing, up-to-dateness and provenance.

We will now describe the individual features. If not stated otherwise, the feature values were measured manually.

Number of published datasets (ND): Almost all platforms have the notion of a dataset: an independent unit of data, describing one special topic or aspect of the world. One of the basic questions for an Open Data repository is: does it provide any interesting datasets? Although the absolute number of datasets can easily be skewed, as we will see later, a repository with more datasets is still more likely to useful data. Many sites provide the number of published datasets, for all others, we wrote simple web crawlers to count them.

Existence of standardized metadata attributes (SM): Many platforms provide the option of adding metadata to datasets. However, for automatic processing of the datasets, they should have a set of standardized attributes, which can be looked up for any dataset.

Standardized file formats (SF): Offering data in standardized file formats makes reuse much easier because it is immediately apparent from the metadata how to process the published files. In contrast, platforms allowing upload of any file format will always require manual processing to enable reuse of the data sets.

Standardized domain categories (SC): Anchoring a dataset in a particular domain is helpful for retrieval and standardizing the available domains can further improve the processing.

Standardized spatial (SS)/ temporal metadata (ST): The majority of open datasets, especially those offered by public agencies, only describe specific geographic entities and/or time intervals. Reuse of these datasets is facilitated when the respective spatial or temporal information is provided as metadata. Again, standardized formats, e.g. a format for dates and date intervals for all datasets on the platform, greatly simplifies the subsequent processing steps.

Existence of an API (EA): This should be an obvious feature of every Open Data platform. Without an API, a platform specific crawler would be required to access the data sets. That does not support automatic processing.

API granularity: access to metadata or data (AG): An API that offers access to standardized package metadata is a good first step. If a platform allows direct access to the datasets through the API, instead of providing file download links, users could automatically access that part of the data that is of interest, without the need to download the entire dataset. This usually requires that the raw data is stored in a database management system.

Curation (as opposed to Wiki-style public editing) (CR): A curated platform is not necessarily better than a public editing platform. We included this feature in the study to differentiate between agency-run platforms, which are usually curated, and platforms driven by an interest group, which usually allow unlimited dataset upload.

Latest date of activity (DA): This feature is used to evaluate whether the trend to publish Open Data is still going strong, or if the initial effort is already slowing down. We've looked at the last uploaded dataset, or if not found, the last blog or news post on the platform.

3.3 Detailed Analysis

The detailed analysis was performed for the *data.gov* (US), *opendata.ke.gov* (Kenya) and *data.gov.uk* (UK) repositories as well as the global repositories *data.worldbank.org* of the Worldbank and *data.un.org* of the United Nations. For the detailed analysis, we surveyed some additional features of the datasets. The goal of these new features is to evaluate how easily the published datasets can be reused. As described in Section 2, required human involvement should be minimal. We added four more featues, which will be described in the following paragraphs.

Downloadable Datasets (DD): The fraction of the available datasets where we succeeded in downloading actual files. One of our first observations was that many of the offered datasets did not include working downloads: many of the provided links were dead, returning HTTP error codes or timing out on request. Other links led not to the data, but to HTML documents containing links to the data, or sometimes just the home page of the organization distributing the data. In both cases, manual browsing is necessary to obtain the actual data. Of course, this prevents automatic retrieval and thus automatic search and analysis on the data. We measured this feature by obtaining all available download links from the respective platform's API and following them. If there was a valid HTTP response, and the content was not in HTML format, we counted the dataset as downloadable.

Machine-readable Datasets (RD): The fraction of the downloadable datasets where we succeeded in opening the downloaded files using standard, of-the-shelf parsers for the respective file formats. In our research on Open Data, we quickly noticed that a great share of the datasets are in obscure or proprietary formats, or not in the format given in the metadata. Obviously, this prevents automatic searching and processing of the data. We measured this feature by looking at the file type specified in the metadata and applying a standard parser for the respective format. When no file type was given, we guessed using the file name extension. If the file could not be parsed using this method, we counted it as not machine-readable.

Existence and number of tags (NT): Almost all surveyed platforms use tags to make datasets discoverable and searchable. However, tags are only a useful instrument of navigation if they are applied to all datasets, in an adequate number and without repeating generic tags for many datasets. We counted the average number of individual tags per dataset. We did not distinguish between single word and

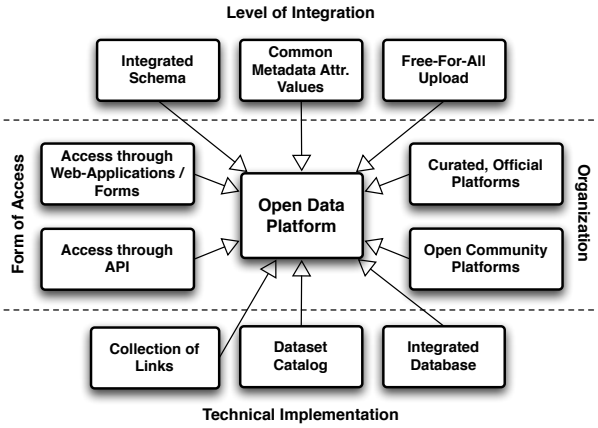


Figure 2: Classification of open data platforms

multiple word tags.

Existence and length of description (LD): Almost all surveyed platforms offer some textual information describing the dataset. Some include a detailed description as a file in the downloadable content of the dataset. These non-standardized descriptions are not included in this feature, as they can not be processed automatically. We argue that a longer description might give more detailed information. Hence, we counted the average number of characters in the description of a dataset. The complete results of the survey can be found at the survey website. The next section will interpret and discuss these results.

4. DISCUSSION

This section will introduce a classification of Open Data platforms derived from our survey. We will discuss the properties of these classes and support them using our data. Finally, we will discuss general phenomena we came across.

4.1 Classes of Open Data Platforms

In the course of our study, we identified several dimensions in which Open Data platforms can be classified. Figure 2 illustrates the four dimensions and possible values for each one. Open Data platforms can be split according to the technical implementation of the platform, the allowed forms of access, the level of integration between the datasets, and the organization form.

Technical: Link Collection Concerning the technical implementation, the most common type is *collection of links*, which accounts for 55% of the surveyed platforms. These platforms host only metadata, and store URLs as the only way of accessing data. This class is not only the largest, but also generally the least useful one. As our detailed analysis of data.gov and data.gov.uk in Section 3.3 showed, large percentages of these links do not resolve to a working file download. Apart from the dead-link issue, our survey also shows that these platforms have a lower level of standardization and integration between datasets. For example, only 20% of the link collections feature standardized file formats (SF), while the rate over all surveyed platforms is 42%. Generally speaking, leaving the actual data with a multitude of different providers leads to a lower data reusability. We argue that this is due to the distribution of the data over many different "spheres of influence", each with its own standards, formats, and levels of dedication to Open Data.

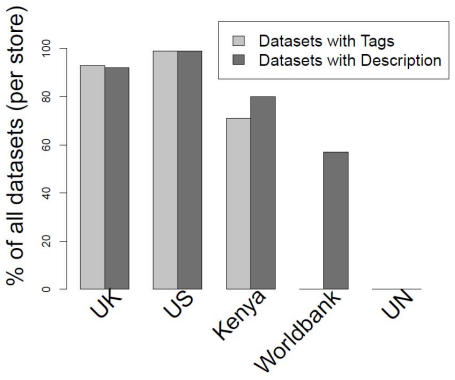
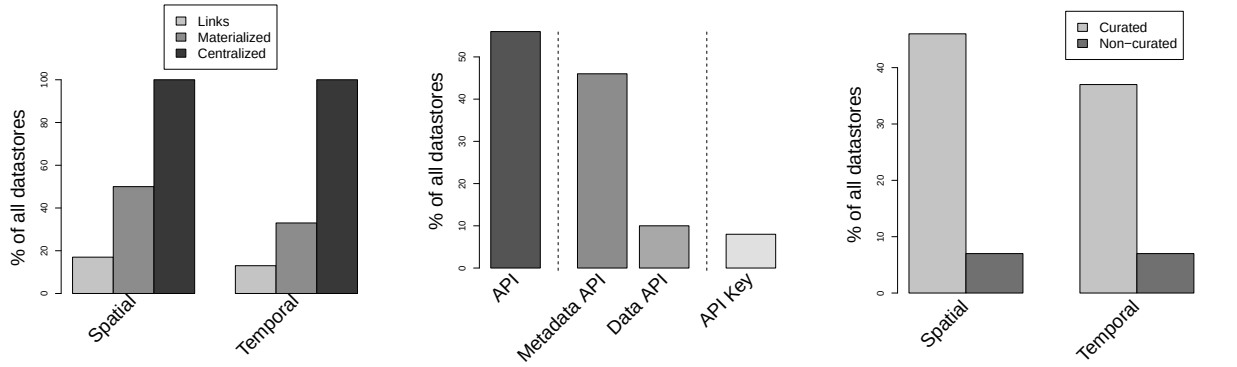


Figure 4: Metadata Availability

Technical: Download Catalog The second largest class regarding technical implementations is *download catalogs* consisting of 29% of the surveyed platforms (plus 6% hybrids between link collection and download catalog). In contrast to link collections, these platforms also host the files, instead of just the links. Interestingly, platforms implementing this simple concept have a much higher reusability, e.g., 80% of the download catalogs have standardized file formats (SF). The other features related to standardization (SC, SM, ST) show similar results. A possible explanation is that these platforms might operate with greater personal or financial resources, allowing more integration and higher technical sophistication.

Technical: Integrated Databases With 8% the integrated databases class is the smallest one within our classification. These platforms do not operate on the level of individual datasets or files, but offer integrated datasets using a database management system. This difference in storage manifests in the actual web platform through the ability to filter and query datasets. Most of these platforms use relational data, but Linked Data (e.g. SPARQL endpoints) is used as well. In contrast to catalogs and especially link collections, these platforms achieve 100% downloadable datasets (DD) as well as 100% machine readable datasets (RD). For the share of these platforms that offer data-level APIs, a database backend of course offers the highest quality, enabling fine-grained, filtered or even aggregated data access via API call. Even though these repositories offer the highest reusability of data, they also have the lowest diversity of available data sets. This is due to the fixed relational schemata they employ: *data.worldbank.org* and *data.un.org*, the two main representatives of this category, both offer only statistical values that are always relative to a country and a year.

Access: Webforms and -applications All surveyed platforms offer a web interface for finding and downloading datasets. While an API seems like a mandatory part of an Open Data web site, 43% of them do not feature an API, but instead allow only manual download through listings of datasets. An important aspect of usability for such a platform is the discoverability of datasets through metadata. We measured the existence and quantity of tags and description (NT and LD) to estimate it. Figure 4 shows for three examples (data.gov, data.gov.uk and opendata.go.ke) from the detailed analysis that these features are widely used in dataset catalogs and link collections. The platforms from the global analysis show similar distributions. In



(a) Spatial and Temporal Concepts in Metadata

(b) API capabilities grouped by implementation type

(c) Spatial and Temporal Concepts in Metadata, grouped by curation

Figure 3: Capabilities of Open Data Repositories

contrast, platforms that focus on providing relational data using an integrated database mostly lack tags or descriptions, instead offering web-based table explorers. However, the database-centric platforms always allow querying data in more advanced ways than text-based information retrieval techniques can provide. Figure 3a shows that temporal or spatial metadata is offered by all of them, while only very view catalog- or collection-style repositories can support this feature.

Access: API The other 57% of surveyed platforms feature an API, although their capabilities vary, as Figure 3b shows. Generally, APIs can be classified in those that only allow to retrieve metadata such as dataset titles, categories, and download links, and those that allow to access the data directly. On first glance, it surprises that there are so many platforms that have an API, but no direct data access. This is mainly due to the prevalence of link collections over download catalogs and databases. It is also due to the fact that link collections themselves have relatively more APIs than other platforms and can of course not provide data directly through their API. As the Figure 3b shows, API keys are very rare in public data platforms.

Organization: Non-curated Platforms These platforms are most often not run by institutions or agencies, but by a community that collects sources of open data or raw datasets. In contrast to curated repositories (see next paragraph), they offer facilities for uploading datasets that can be used by everyone. The lack of a central integration authority usually leads to less structured repositories, as Figure 3c shows for the example of spatial and temporal metadata. However, in countries that do not have an open data legislation, these are often the only platforms that exist.

Organization: Curated Platforms This class of platforms does not allow uploads by users. They are moderated and usually run by some public institution. The moderation allows them to have a higher degree of integration and standardization. Figure 3c shows the difference in metadata (ST,SS) for curated and non-curated platforms. Another difference is average size: while the average number of datasets for curated repositories is about 1600, it is only about 450 for non-curated repositories.

Integration: Free-for-all Upload This category encompasses all repositories that have no regulations concerning datasets that are published. Repositories in this category do not have standardization features such as SF, SC, or SS/ST. Of the surveyed platforms, 43% do not even have one of them.

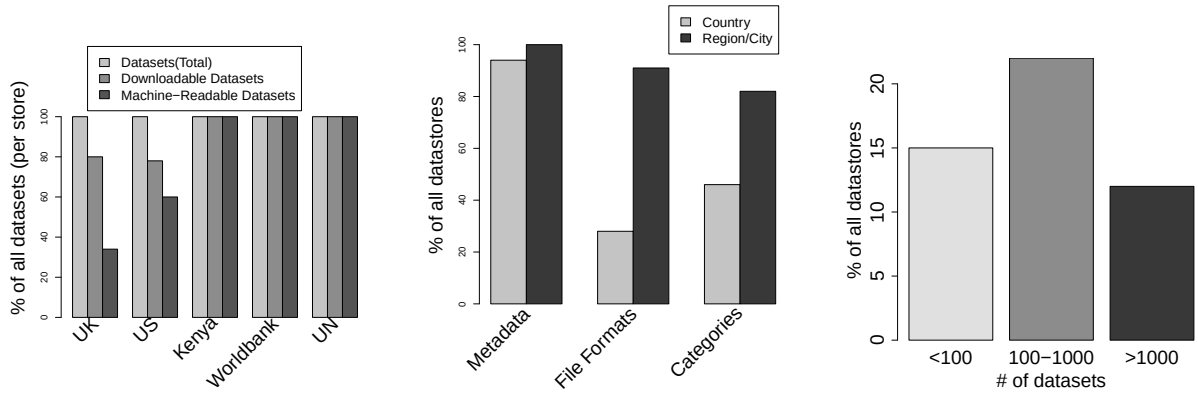
Integration: Common Metadata Attribute Values The next category contains repositories which apply restrictions to the datasets or metadata that can be attached to them. Examples would be restricting the possible file formats, or requesting standardized date formats. This corresponds to the standardization features of the survey (SF, SC, SS or ST). Only 10% of surveyed repositories have all four features, while another 20% have three of them.

Integration: Integrated Schema This class of platforms with integrated schemata offers the highest degree of integration. In our context, integrated schema means repositories that use classic data integration techniques to create explicit schemata (or ontologies) for datasets. Currently, we are not aware of repositories which implement explicitly integrated schemata. Platforms like data.worldbank.org and data.un.org use the same combination of attributes for a large number of datasets, usually mapping from a country and a year to an attribute value. Still, the schema is not being made explicit, and the integration of more complex relations remains an open problem, so that this class of platforms is more of a theoretic construct.

4.2 General Trends

This section will present some general trends we discovered through our survey, which crosscut the classification presented in the last section.

Repository Sizes Repository sizes vary considerably in our study. The largest repositories are in general those of larger nations and international organizations, reaching up to 4900 for data.gov and 7400 for data.gov.uk, while many countries have only small community driven platforms with sometimes less than 100 datasets. Still, there are exceptions. For example, Estonia provides a statistics database that has over 3000 indicators in a well structured format. The majority of platforms can be found in the 100 to 1000 range (see Figure 5c). **Locality** One interesting discovery we made was that local platforms, those operated by municipalities or states, performed better regarding standardization (SF, SC, SS or ST) than national repositories (see Figure 5b). It



(a) Accessability of Datasets

(b) Standardization, grouped by local/national

(c) Size of Repositories

Figure 5: Capabilities of Open Data Repositories

seems that platforms focusing on a more limited geographic area can more easily introduce constraints to or require information for a dataset. The users of local platforms seem to be more willing to enhance datasets by adding tags or descriptions.

Format Zoo and Dead Links Two major problems are common to almost all platforms: dead links and a plethora of different file formats. The data from the detailed analysis concerning data.gov and data.gov.uk illustrates these problems (Figure 5a). In both cases, only about 79% of the metadata entries on these two platforms could be used to download actual files (as described for feature DD). Some links were dead, while others led to HTML pages or web applications where the data might be found. With regard to our goals of machine readability and reusability, those entries can not be seen as useful open datasets. But even from the downloaded files, only 77% in the case of data.gov and only 42,2% in the case of data.gov.uk could be opened with a set of standard parsers (feature RD).

The difference between these two platforms can be explained by the fact that data.gov uses a small set of standardized file formats, while the resources of data.gov.uk have no limitations regarding the file format. We counted 32 different formats in the data.gov.uk metadata, the overwhelming majority of datasets being marked as "no-format".

5. CONCLUSION

Over the years, public and governmental institutions have generated and collected vast amounts of data which is locked within heterogeneous, proprietary and undocumented data files. Within the Open Government initiatives that emerged recently, municipals and agencies have started making this data available to the public in order to achieve transparency, participation and innovation throughout society. Looking at the web today we can find hundreds of more or less useful open data platforms from which we surveyed the top 50 within this paper. We benchmarked this repositories in terms of reusability which is in our opinion the most important aspect of open data. Furthermore, we analyzed a subset of five repositories in terms of data and metadata quality. Our analysis showed that a majority of the platforms lack of proper standards and APIs, and have a lot of data published that is not machine-readable or in a proprietary format. This prevents reuse by automated tools which means that many open datasets are not open at all. The surveyed repositories also provide varying degrees of metadata which causes problems in integrating open datasets from different platforms. In general, the surveyed repositories show considerable differences in terms of openness, which draws us to the conclusion that the open data community acts very uncoordinated at the moment and needs to be aligned in the future.

Except for Linked Open Data, there has been very little research on Open Data topics. The few papers that exist on this subject unfortunately focus on very specific use cases. Two projects worth mentioning here are MIDAS [1, 3] and GovWild [2]. The MIDAS project integrates public financial data to gain new insights, e.g. to identify critical hub banks to monitor systemic risks. The second project, GovWild, integrates data from several US and EU public sector data sources and links them with datasets from NYT. This data might then help reveal the abuse of authority of politicians or spending affairs. While the techniques developed for these projects in general hold for all kinds of open dataset, the data integration itself requires a lot of manual work. Therefore, we think it is important to develop automatic techniques which can be applied on a much larger scale instead of in these very specific scenarios.

By applying our methodology presented within this paper we would like to repeat this survey annually to see how the open data trend evolves and to infer indications about it's future development.

6. REFERENCES

[1] S. Balakrishnan et al. Midas: integrating public financial data. In *SIGMOD '10*, SIGMOD '10, pages 1187–1190, New York, NY, USA, 2010. ACM.

[2] C. Böhm et al. Linking open government data: what journalists wish they had known. In *I-SEMANTICS '10*, I-SEMANTICS '10, pages 34:1–34:4, New York, NY, USA, 2010. ACM.

[3] A. Sala, C. Lin, and H. Ho. Midas for government: Integration of government spending data on hadoop. *ICDE Workshops*, 0:163–166, 2010.