Assessment and Visualization of Metadata Quality for Open Government Data

Konrad Johannes Reiche*, Edzard Höfig**, Ina Schieferdecker***

**Fraunhofer FOKUS/FU Berlin, konrad.reiche@fokus.fraunhofer.de*
***FU Berlin, edzard.hoefig@fu-berlin.de*
****Fraunhofer FOKUS/FU Berlin, ina.schieferdecker@fokus.fraunhofer.de*

Abstract: *An increasing number of datasets from government, public organizations and institutions are published as open data. Metadata that describes them, are cataloged at central places to enable a better access to these datasets. Quantifying the metadata quality can help to measure the efficiency of a catalog and discover low-quality metadata records which prevent the user from finding what she is looking for. We researched and implemented a range of metrics from the field of metadata quality assessment as part of an open data platform. This paper describes the platform that automatically assesses the quality of different open government data portals using the CKAN catalog software. The results are aggregated and visualized through a web application in order to establish a continuous and sustainable monitoring service.*

Keywords*: Open Data, Metadata, Data Quality, Government Data, CKAN*

Acknowledgement: *This work has been partially supported by EC FP7 TAGCLOUD project (Grant agreement no: 600924).*

Introduction

With rise of the open data movement, government and public agencies start to open up their data for public use. The technical tools for implementing this infrastructure are, often distributed, repositories for the datasets and typically centralized catalogs for metadata. Metadata are used to describe the datasets and provide information and search capabilities. Central to the operation and success of the metadata catalogs and their interoperability is the quality of the metadata they provide. In this context, we understand Metadata quality as “fitness for a purpose”.

We have been designing and implementing the German Open Data Portal1 (GovData) which harvests metadata from different German portals on municipal, state, and federal level and from different domains such as statistics, geo information or environmental information. One of the issues that we encountered was the diverse metadata quality of the different portals, which does

[1] https://www.govdata.de

not only complicate the harvesting but also limits the services of GovData in terms of adequateness, completeness, accuracy and correctness of the metadata provided. We started our research in order to get an overview to survey and improve metadata quality.

Metadata

Metadata is used to catalog and index the datasets. Data about data has become the most used, yet underspecified, definition for the term metadata as it allows different interpretations by various professional communities. Not too long ago, metadata was only a concern of information professionals engaging in cataloging, classification and indexing. Often cited examples are libraries and librarians using catalog cards to assess the content and location of a book. Today, there are much more creators and consumers of digital content which also needs to be cataloged. Arguably, the term metadata is used a lot less, but the digital content is described, indexed, and cataloged by metadata. Metadata consist of a set of information pieces about information objects it describes. Thus, the term metadata can be refined in its definition as follows:

> ***Definition 1****: Metadata. The sum of statements that is associated with any (set of) information objects at any level of aggregation.*

Please note that such an information object can consist of a single information resource (an image), multiple information resources (a data series) or be even a whole information system like a database. The structure of metadata can be highly diverse. The intended use, context but also technical circumstances determine, how much metadata is structured, how well this structure is defined and how strict the structure is enforced.

Catalogs

Catalogs, sometimes also called repositories, are a commonly used technical tool for implementing a metadata infrastructure. Catalogs facilitate the collection, publication, presentation and search of metadata. Metadata describe information resources and provide information like authors, maintainers, formats, descriptive free text, etc. The referenced resources typically do not reside in the same repository. Metadata, in turn, is organized in a centralized and possibly standardized way using catalogs.

Metadata Quality

Quality is both objective and subjective. It depends on the context what quality means, how quality can be determined and what the implications are. Government data is primarily opened to enable transparency, innovation and new businesses building on the open government datasets. By that, it is not only crucial that the datasets themselves are of high-quality, but likewise the metadata need to be of high-quality.

Today, the number of available datasets on an open government platform is also a political issue. The platforms advertise their effectiveness by displaying the total number of datasets available. While this is a great quantity factor, it is not a quality factor. Making the data accessible, does not imply that the users will find the resources they are looking for. Content publisher have to ensure that the resources are credible and discoverable. The credibility is bound to the quality of the content. The discoverability is bound to the quality of the metadata.

Hence, the fitness of metadata, i.e. the metadata quality, can be defined by the effectiveness in supporting the functional requirements of the users it is designed for. With this in mind, the following definition for metadata quality is proposed:

> ***Definition 2**: Metadata Quality. Metadata quality is the fitness of the metadata to make use of the data, i.e. of the information resources, it is describing. Metadata's fitness determines the level of enabling to find, identify, select, and eventually obtain the information resources. Metadata quality is inversely proportional to the metadata user's uncertainty about the described information resources.*

Quality Metrics

Because of its subjective dimensions, quality is not easy to measure. Often, only objective quality attributes are measured. Furthermore, there are complex attributes which have no single measure. For example, in the case of metadata records there are attributes like accuracy, accessibility, conformance to expectations, completeness, comprehensibility or timeliness. For each of these attributes another measure is more appropriate. Thus, the measures are by no means equivalent, but rather measure different aspects of an attribute.

Xavier Ochoa and Erik Duval (2006) have aggregated a rich set of metadata quality metrics. These metrics were developed for repositories managing metadata records of learning objects, but we find that they are defined in such a general manner, that they are suitable for application to open government metadata. A selection of their metrics together with refinements and additional metadata quality metrics developed in our research are discussed in the following text.

Completeness

A metadata record is considered complete, if the record contains all the information required to have an ideal representation of the described resource. While the attribute of completeness again can be very vague, one way of constructing a metric for this is to simply count the total number of fields and all fields, which have been set to a value which is not *null*. The completeness metric q_c is then defined as the ratio of number of fields and number of completed fields:

$$q_c(record) = \frac{\sum_{i=1}^{n}[record[field_i] \ \neq null]}{n}$$

Weighted Completeness

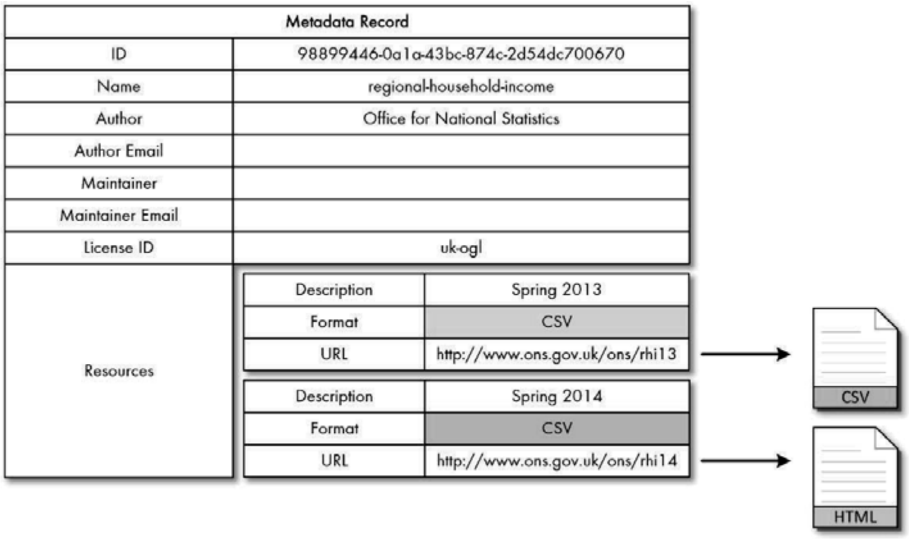
While the completeness metric is straightforward it comes with the drawback of treating every field with the same importance. The relevance of a certain metadata field depends strongly on the context. The problem is addressed by specifying a weight to each field. The weight w_i is a numerical value which expresses the relative importance for the fields to each other. This would allow to assign a weight of 1 for semi-important or regular fields, a weight of 3 for important fields, but also a weight of 0 for fields which should be excluded completely from the measurement. The weighted completeness q_w is then defined as follows:

$$q_w(record) = \frac{\sum_{i=0}^{n} w_i \cdot [\, record[field_i] \;\neq null \,]}{w_i}$$

Accuracy

The accuracy metric measures how accurate the metadata record represents the associated resources. There are field types where this can be expressed with a Boolean value. Either the given information is correct or not. This example is illustrated in Figure 1, where the resource format type is checked against the actual format returned by the host.

Xavier Ochoa (2008) proposes that the correctness can be understood as the semantic distance between the information given through the metadata record and the information given through the resource. The semantic distance d_i is the difference between the information a user can extract from the record and the information the same user could extract from the referenced resource itself.



Metadata Record		
ID	98899446-0a1a-43bc-874c-2d54dc700670	
Name	regional-household-income	
Author	Office for National Statistics	
Author Email		
Maintainer		
Maintainer Email		
License ID	uk-ogl	
Resources	Description	Spring 2013
	Format	CSV
	URL	http://www.ons.gov.uk/ons/rhi13
	Description	Spring 2014
	Format	CSV
	URL	http://www.ons.gov.uk/ons/rhi14

Figure 1: Example of an accuracy metric implementation validating the file format of the resources

A shorter distance implies a higher accuracy of the metadata record. With this approach the metric q_a could be expressed with the following calculation:

$$q_a = 1 - \frac{d_i(record[field_i])}{n}$$

The difficulty resides in d_i, which is the distance measurement of the field value $record[field_i]$. Different fields require different, tailored distance measurements. For numbers and dates the offset can be computed, for categorical values a predefined distance table can be used, e.g. declared language and actual language. The language distance between Spanish and Italian is shorter than between Spanish and Japanese.

Richness of Information

The vocabulary terms and the description used in a metadata record should be meaningful to the user. For that the metadata need to contain enough information for describing uniquely the referred resource. This can be done by measuring the amount of unique information present in the metadata. The approach originates from the field of information theory. In this work the metric will be called richness of information, as it describes the procedure better. In general, the richness of information metric q_i is defined as follows:

$$q_i(record) = \frac{I(record[field_i])}{n}$$

Where the function I returns a quantification of the information content. For numerical and vocabulary values this can be defined as 1 minus the entropy which can be expressed with the following function:

$$I(field) = -\log P(field)$$

Whereas $P(field)$ is the probability for value to occur in a set of metadata records. For free text the term frequency-inverse document frequency (tf-idf) is proposed. A numerical statistic which reflects how important single words are relative to a collection of documents. Here the term frequency tf, the document frequency df, the total number of documents m and the total number of words n is used.

$$I(text) = \frac{\sum_{i=1}^{n} tf(word_i) \cdot \log\left(\frac{m}{df(word_i)}\right)}{n}$$

Readability

The readability metric measures the degree to which a metadata record is cognitive accessible. The readability describes how easy a user can comprehend what the resource is about after reading the metadata record. To implement this metric several readability indexes could be used. One of these is the Flesch-Kincaid Reading Ease which measures the comprehension difficulty when reading an academic text. This reading ease score for English texts can be computed by applying the following function q_r:

$$q_r(record) = 206.836 - 1.015\left(\frac{words}{sentences}\right) - 84.6\left(\frac{syllables}{words}\right)$$

For this calculation the total number of words, sentences and syllables is required. Although the metric aims to describe results for scores on a scale between 0.0 and 100.0, negative values and values above 100.0 are possible, as well.

Availability

Metadata records contain URLs which point to the actual resources. The availability metric assesses the number of reachable resources. A resource is available, if the resource can be retrieved from the given URL. Thus, the following function definition is used for the metric q_{av}:

$$q_{av}(record) = \frac{\sum_{i=1}^{n} [resource\ i\ availabile]}{n}$$

Misspelling

Readers which are proficient in a language might halt for a moment on words written incorrectly. The number of spelling mistakes might not be a very important measure, as opposed to the availability of resources, nevertheless it influences the information quality. For the misspelling metric q_m the number of spelling mistakes are counted:

$$q_m(record) = 1 - \frac{m}{n}$$

Where m is the number of spelling mistakes and n is the total number of words.

Platform: Metadata Census

We implemented the presented quality metrics and applied them to a set of metadata. In order to make them reusable by others, they are implemented as part of a platform: Metadata Census. A web application acts as a “quality-dashboard” to survey the quality of selected CKAN-based catalogues in a continuous way.

There is a range of functional requirements which have been identified for the Metadata Census:

- A continuous, CKAN-based metadata harvester
- A schemaless data store
- The presented quality metrics
- A Scheduler for triggering the harvesting runs
- A module for metric reports
- Some visualization to allow users to grasp the analysis results
- A leaderboard, to enable comparison of metadata quality between repositories

The harvester component is required to gather the metadata locally, but also to access it afterwards, even if the repository is not online at the point in time. Due to the number of metadata records, it would not be feasible to perform all the operations in memory. Different repositories might use a slightly different metadata schema. A schemaless data store can then organize and manage the metadata in a natural way.

The quality metrics form the core functionality of the implementation and it should be possible to easily add new quality metrics. The scheduler is required to continuously monitor the quality. For the metric reports we needed to decide how a single metric score is computed. The problem of making a comprehensible assessment is not necessarily solved by a large number of scores. The results also need to be broken down into smaller information pieces to make the outcome better understandable. Visualization can also help to reduce the information noise for a more natural interpretation. Open data is inherently political. In fact, open data has a competitive appeal. A leaderboard could be instrumentalized to compare the metric scores of different repositories with each other and encourage this competition.

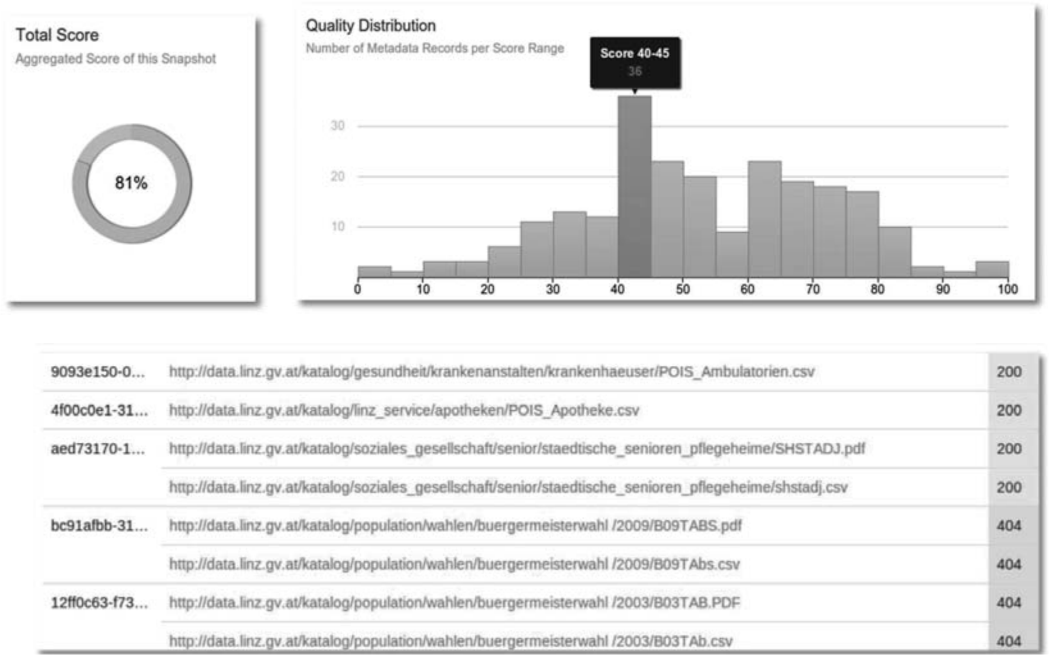
Visualization

An appropriate visualization is crucial to enable the communication of quality assessment in a way that goes beyond a sheer quality metric score. The sustainability is created when the data providers are enabled to investigate the source for the lack of quality through visualization. An effective approach for this is to generalize where possible and specialize otherwise.

This is shown in Figure 2. Every detail page for a quality metric has a score meter and a histogram. The score meter does not induce additional information but it helps to grasp the overall state visually. The histogram shows the metadata quality distribution grouped by the different score ranges. This clearly communicates how many metadata are affected by low-quality and in which seriousness. Below are the more advanced, respectively more specialized visualizations. Visualization is not necessarily a graph or a diagram, thus it can also be a plain table with highlighted fields. For instance, for the availability metric it is relevant which metadata records are affected by dead links. Further, it should be easy to examine the dead links. This requires a dynamic interface, for example input fields in order to filter the result list.

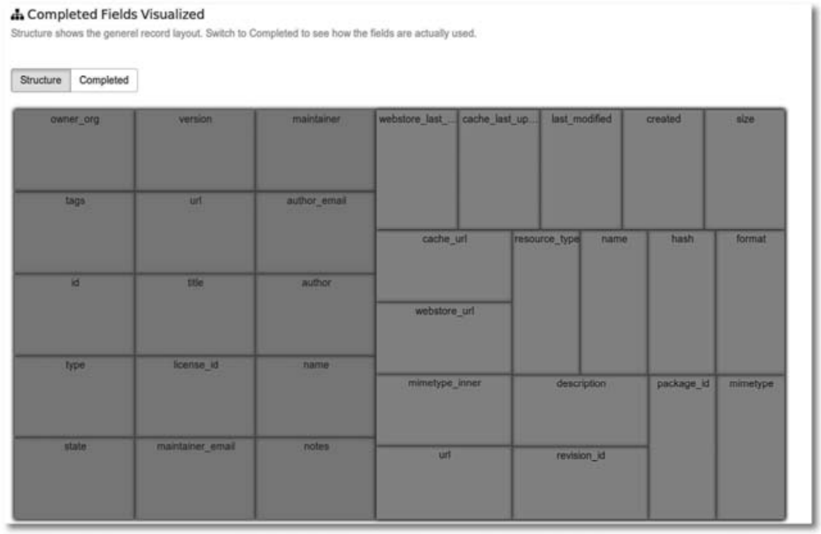
Visualization can also be used to describe the same information in different ways. Treemaps are used to illustrate the results of the completeness metric (Figure 3). This way the nested nature of metadata records is exploited. Again, dynamic interfaces are used to enhance the visualization. The Treemap display two results. On the one hand, how is the metadata record structured in general, like what fields are there and how are they nested, and on other hand how often these fields are actually used. Switching between these two results in an animated transition helps the investigator to see what fields stay and what fields are marginalized because they are not used at all.

For the more general pages like overview of metadata quality of a repository over time the obvious choices are made and the aggregated score is shown on a line chart.


Total Score
Aggregated Score of this Snapshot
81%
Quality Distribution
Number of Metadata Records per Score Range
Score 40-45
36
30
20
10
0
10
20
30
40
50
60
70
80
90
100

9093e150-0...	http://data.linz.gv.at/katalog/gesundheit/krankenanstalten/krankenhaeuser/POIS_Ambulatorien.csv	200
4f00c0e1-31...	http://data.linz.gv.at/katalog/linz_service/apotheken/POIS_Apotheke.csv	200
aed73170-1...	http://data.linz.gv.at/katalog/soziales_gesellschaft/senior/staedtische_senioren_pflegeheime/SHSTADJ.pdf	200
	http://data.linz.gv.at/katalog/soziales_gesellschaft/senior/staedtische_senioren_pflegeheime/shstadj.csv	200
bc91afbb-31...	http://data.linz.gv.at/katalog/population/wahlen/buergermeisterwahl /2009/B09TABS.pdf	404
	http://data.linz.gv.at/katalog/population/wahlen/buergermeisterwahl /2009/B09TAbs.csv	404
12ff0c63-f73...	http://data.linz.gv.at/katalog/population/wahlen/buergermeisterwahl /2003/B03TAB.PDF	404
	http://data.linz.gv.at/katalog/population/wahlen/buergermeisterwahl /2003/B03TAb.csv	404

Figure 2: View components to communicate the results, e.g. metric score meter, quality score distribution (chart), link availability (table)


Completed Fields Visualized

Structure shows the generel record layout. Switch to Completed to see how the fields are actually used.

Figure 3: Treemap2 illustrating the metadata completeness

[2] Treemapping is a method for displaying hierarchical data by using nested rectangle.

Case Study: Open Government Data

The developed approach for an automated quality assessment of metadata and its prototypical implementation by Metadata Census have been tested for a set of open government data portals. For this task catalogues from around the world have been selected. The results are shown in Table 1 in form of the leaderboard. The repositories are sorted by their aggregated score.

The aggregated scores give an overview about the score distribution. More details are reviewed through the metric reports (Figure 2). The completeness metric shows that there are no metadata records which fill out every available field. The completeness metric report also helps to identify fields that are seldom used, for instance *Maintainer Email*. This, then again can be used to plan quality improvements. For example, if the field *Author* has only been completed in 80% of the records, the focus should be to improve the remaining 20%. The weighted completeness metric has a better score than the completeness metric. Now, due to the field weighting there are metadata records which satisfy the completeness for every field.

The accuracy metric has the worst overall results for most of the repositories. Often the MIME type is simply not correct. This can also be an indicator that the actual resource is not available directly through the given URL, but through an additional link.

The readability metric does not reveal a lot of information. Some repositories do better, some do worse, but when investigating the results something becomes evident: many descriptions are too short. An improvement would be to compute the Flesch reading ease only on texts with a certain length.

The availability metric is one of the most useful metrics. A repository with too many dead links can quickly render the whole repository useless. The metric has the clear drawback of only delivering the state from the moment the URLs have been checked. Often resources are only temporarily not available, which raises the need for measuring such quality factors over time and for averaging the results.

The misspelling metric detects some typical typos. Not every detection is always an actual typo. The misspelling dictionaries also need to be updated continuously and additional language support is required to cover the full range of all languages in use.

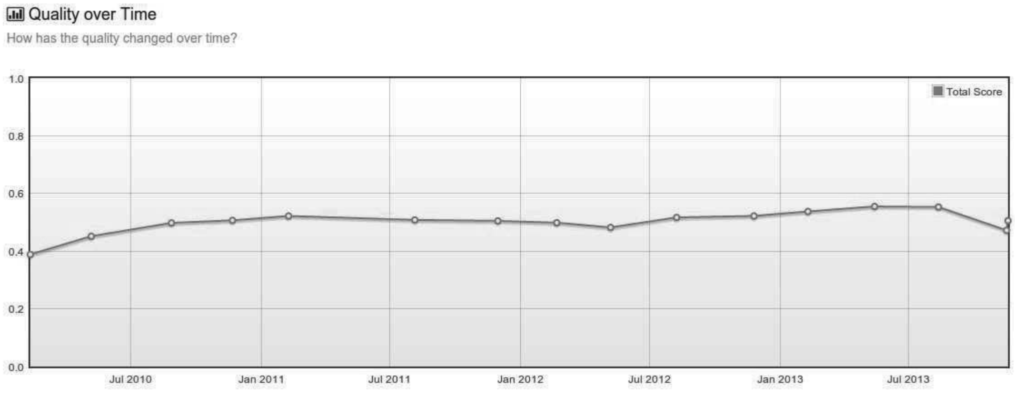


Figure 4: Analyzing the aggregated quality over time shifts the importance towards quality improvement

While single metric results can give interesting insight it is of even more interest to investigate the quality change over time. Such a monitoring can be seen in Figure 4. This way the focus shifts to metadata quality improvement. After all, this has to be the concern when managing a metadata catalog. Small changes in the overall quality go back to different reasons. For example, the quality increased slightly after a large set of metadata have been removed. Thus, further parameters like the number of metadata records should be included in the result, as well.

Table 1: Ranked repositories based on their average score computed through different quality metrics

Rank	Repository	Score	Completeness	Weighted Completeness	Accuracy	Richness of Information	Readability	Availability	Misspelling
1	data.gc.ca	74	79	81	20	86	71	79	97
2	data.sa.gov.au	71	77	82	0	63	72	86	98
3	GovData.de	67	55	87	56	44	79	81	99
4	data.qld.gov.au	66	73	78	0	67	59	60	99
4	PublicData.eu	66	64	67	32	84	42	70	98
4	data.gov.uk	66	62	67	28	85	44	74	97
4	africaopendata.org	66	70	68	53	20	55	87	100
5	datos.codeandomexico.org	65	65	75	0	55	37	100	100
6	catalogodatos.gub.uy	63	70	78	52	64	65	74	100
6	data.openpolice.ru	63	58	81	64	0	100	100	100
7	dados.gov.br	61	53	72	39	87	44	57	100
8	opendata.admin.ch	59	58	68	100	12	35	100	100
9	data.gv.at	57	51	65	0	21	59	68	100
10	data.gov.sk	49	48	58	7	51	37	92	100

Another important feature of Metadata Census is the ability to weight the importance of the quality attributes according to the current purpose of portal evaluation. This flexibility in assessing the metadata quality allows to develop a better understanding of the weaknesses and strengths of a metadata portal and to derive options for improvements.

Conclusion

The experimental results of evaluating the selected portals demonstrate the applicability of the developed platform Metadata Census. The purpose of this research was to assess the metadata quality of open government data portals; the Metadata Census is a first prototype for an automated and flexible evaluation mechanism to carry out such a task.

The quantification of metadata quality attributes was addressed by quality metric functions. Effectively, metrics are used to measure these quality attributes. Although quantifications were performed, it became quickly evident, that they do not cover every possible quality attribute. The presented quantification of metadata quality cannot satisfy a metadata quality assessment to its full end.

The proposed method has another weakness: The use of an algorithmic approach is too limited to discover all subtleties that result in quality flaws. However, keeping the actual objective of improving metadata quality, this is not necessary at all. The importance does not reside in creating very high-quality metadata records, but in improving those who have a very low quality. Metadata Census prototype provides ways to sort these records out. For instance, the quality distribution histogram can list those, which have a very low quality. From there on, a repository can be advanced greatly by improving this group of metadata.

Furthermore, a platform like Metadata Census has two functions. On the one hand as an investigative tool to find metadata of low quality and on the other hand as a competitive one. Open data is instrumental and so can be metadata quality. A leaderboard, such as the one implemented, can be used to engage data provider in improving their metadata. This, of course, requires public provisioning and acceptance of such a tool.

In the future, we will investigate how to improve the definition of metadata quality attributes and of their measurement functions. Besides, the technical implementation of Metadata Census is an early design. There are many ways to improve its functions, as well as the function's behavior including

- Supporting a wider range of repositories
- A metadata revision system
- A live quality feed
- Support for domain-specific languages for metric definition
- Quality measurement as a service

CKAN is just one repository software. Socrata is another widely used open data platform which serializes its metadata to JSON. By further abstracting the metric analysis implementation we could make this option easily available. In addition, with every repository dump added to the database of the Metadata Census, the size increases linearly. This approach introduces a lot of redundant data, which could be eliminated by implementing a metadata revision system.

Furthermore, in order to reveal quality issues in a finer granularity single quality changes could be presented as a live feed. Finally, while new metrics can be easily added, the next step would be the development of a domain-specific language to design quality metrics. Quality is subjective, hence there is a need for more possibilities to create customized metrics and customizations of the Metadata Census.

References

Jens Klessmann, Philipp Denker, Ina Schieferdecker, Sönke E. Schulz. (2012). *Open Government Data Germany: Short Version of the Study on Open Government in Germany*. Published by Federal Ministry of the Interior in Germany. Online available.

D. Lathrop and L. Ruma. (2010*). Open Government: Collaboration, Transparency, and Participation in Practice .* Theory in Practice. O'Reilly Media.

Tony Gill, Anne J. Gilliland, Maureen Whalen, and Mary Woodley. (1998). *Introduction to Metadata*. Getty Research Insitute, Los Angeles.

Abiteboul, P. Buneman, and D. Suciu. Data on the Web: From Relations to Semistructured Data

Ochoa, X. (2008). Learnometrics: Metrics for Learning Objects. PhD Thesis. Katholieke Universiteit Leuven.

W.H. DuBay. (2007*). Unlocking Language: The Classic Readability Studies*. Impact Information.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. (2008*) Introduction to Information Retrieval.* Cambridge University Press, New York, NY, USA

Xavier Ochoa and Erik Duval. (2006). *Quality Metrics for Learning Object Metadata*. In World Conference on Educational Multimedia, Hypermedia and Telecommunications 2006 , pages 1004–1011. AACE.

Besiki Stvilia, Les Gasser, Michael B. Twidale, and Linda C. Smith. (2007). A Framework for Information Quality Assessment. JASIST , 58:1720–1733.

About the Authors

Konrad Reiche

Konrad Reiche is a graduate student at Freie Universität Berlin where he is pursuing a Master's degree in Computer Science. As a working student at Fraunhofer FOKUS he is responsible for the metadata management of GovData.de, the data portal for Germany. For his Master's thesis he researched on metadata quality of open government data.

Edzard Höfig

Edzard Höfig holds a PhD in Engineering from the Technical University of Berlin and is currently working as a Post-Doc at the working group for Model-Driven Engineering and Quality Assurance of Software-Based Systems at Freie Universität Berlin. Edzard's research is focussed on attribution and provenance questions related to open data. He also investigates dynamic data streams in urban environments.

Ina Schieferdecker

Ina Schieferdecker heads the Competence Center on Modeling and Testing of System and Service Solutions at Fraunhofer FOKUS, Berlin, coordinates Open Data and ICT for Smart Cities at Fraunhofer FOKUS and is also a professor of Model-Driven Engineering and Quality Assurance of Software-Based Systems at Freie Universität Berlin. Her research interests include model-driven engineering, software quality assurance, conformance, interoperability, and certification.