Search, Filter, Fork, and Link Open Data

The ADEQUATe platform: data- and community-driven quality improvements

Sebastian Neumaier
Vienna University of Economics and Business
sebastian.neumaier@wu.ac.at

Lőrinc Thurnay
Danube University Krems
loerinc.thurnay@donau-uni.ac.at

Thomas J. Lampoltshammer
Danube University Krems
thomas.lampoltshammer@donau-uni.ac.at

Tomáš Knap
Semantic Web Company
tomas.knap@semantic-web.com

ABSTRACT

The present work describes the ADEQUATe platform: a framework to monitor the quality of (Governmental) Open Data catalogs, to re-publish improved and linked versions of the datasets and their respective metadata descriptions, and to include the community in the quality improvement process. The information acquired by the linking and (meta)data improvement steps is then integrated in a semantic search engine. In the paper, we first describe the requirements of the platform, which are based on focus group interviews and a web-based survey. Second, we use these requirements to formulate the goals and show the architecture of the overall platform, and third, we showcase the potential and relevance of the platform to resolve the requirements by describing exemplary user journeys exploring the system. The platform is available at: https://www.adequate.at/

ACM Reference Format:
Sebastian Neumaier, Lőrinc Thurnay, Thomas J. Lampoltshammer, and Tomáš Knap. 2018. Search, Filter, Fork, and Link Open Data: The ADEQUATe platform: data- and community-driven quality improvements. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3184558.3191602

1 INTRODUCTION

Open Data has increasingly become a valuable asset, both as social capital of our society as well as an economic factor for businesses and the industry domain [1]. In particular, Open Government Data proliferates in the past years, as the number of publicly available datasets is steadily growing.

Nevertheless, low data quality is the generally recognized factor impeding the broader adoption of Open Data. Data publishers often lack expertise and resources to ensure that the data is published in an optimal way, fully standards compliant and with complete metadata. This potentially incomplete and heterogeneous metadata, and the lack of interoperability between data sources, impedes a more sophisticated search functionality over the datasets, exploiting the semantics of the datasets. In fact, the current search over Open Government Data catalogs (such as the project's use-case portal `data.gv.at`) is limited to specific facets, i.e. metadata fields only, ignoring embedded semantics in the actual dataset.

Also, the currently available metadata descriptions at these catalogs do not contain links to any external knowledge bases, existing ontologies, or other datasets and data catalogs. This lack of external references implies the risk of having data silos instead of connected and interlinked data portals.

These issues have been recognized by the Austrian Open Governmental Data initiative when it supported the research project ADEQUATe: Analytics & Data Enrichment to improve the QUAliTy of Open Data.[1] The project, now at its final phase, aims at providing a programmatic framework that would help (1) data publishers to improve the quality of the data and metadata in an automated, intuitive and efficient way, and (2) data consumers to better search, judge about the usefulness of the data, and reuse the data.

In this particular paper, we focus on the following concrete contributions: We identify the major challenges/needs of Open Data consumers by conducting focus group interviews and surveys (cf. Section 2), we address these issues by building the ADEQUATe platform (outlined in Section 3), and show its basic usage scenarios by picturing different user journeys in Section 4.

2 MOTIVATION & REQUIREMENTS

As an initial step of the project – in order to define clear goals and outcomes – we conducted focus group interviews, together with an online-based survey regarding current issues of data quality on open data portals in Austria [2].

The focus groups were conducted on four separate events in the timeframe of December 2015 - February 2016 (i.e., Open Data meet-up in Vienna, GovCamp in Vienna, OGD platform meeting in Graz, and at the IRIS conference in Salzburg), addressing different stakeholder groups in terms of background and expertise (i.e., lay-persons, open data interested individuals, platform providers, as well as scientists). In total, 106 persons were attending the focus group meetings. A pre-defined set of steering questions was used to initiate the discussion process within the groups yet being open to adopt upcoming additional aspects as well. The discussions were recorded and transcribed, followed by a qualitative analysis including the coding of the responses and subsequently the derivation of categories and subcategories. The statements (184 items in total) of the participants were then aggregated by these categories, resulting

[1] http://adequate.at

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191602

in four main aspects, namely 'community', 'quality', 'search', and 'versioning'. The two most prominent aspects were 'quality' with 84 items as well as 'search' with 53 items.

Based on these findings we derived a set of requirements for the ADEQUATe platform [3]. Below we summarize the most important points and group them by four facets: *quality report*, *quality improvements*, *linkage & search*, and *collaboration*.[2]

Quality report. Users want to:

- know to which extent the data is complete content-wise (e.g., are really all entries of the year 2018 available) as described by the meta data.
- work with structurally-consistent data in terms of used formats (e.g., is a comma used as seperator in a given CSV file).
- know to which extent the data sets have complete and correct descriptions.
- know to which extent the data is up-to-date (e.g., for critical/real-time applications).
- have clarity about legal aspects regarding available data sets (e.g., license information).

Quality improvements. User wants to:

- see issues of a particular data set (e.g., reported errors or gaps within a given data set).
- see changes in data sets to identify potential quality improvements (e.g., community members have suggested corrections to identified errors).
- see the development (changes) over time of a particular data set (i.e., the activity rate of the community on a particular issue or data set).

Linkage & search. User wants to:

- see all resources (CSV files) when searching for certain entity (e.g., city of Vienna)
- be able to create and modify filters (e.g., for single keywords, time, format) for search results
- have well-indexed data sets to improve search results.
- have flexible search functionalities beyond pure keyword-based searches. That means e.g. to be able to 1) search data sets based on the types of concepts (types of columns in case of CSV files) they contain; 2) search taking into account hierarchy of types of concepts.

Collaboration. Users want to:

- discuss and collaborate on particular data sets.
- be able to fork particular data sets in order to work independently on certain ideas/issues.

Overall, we can conclude that data quality and search were the most discussed themes within the workshops and the survey. The participants reported that many data sets suffer from issues such as non-standard encoding or a divergence between the stated encoding and the actual encoding of the data set. Besides quality aspects, users see demand for action regarding the process of searching and the provided search functionalities on the given platforms. They claim that it is hard to actually find a certain data set if users do not use the exact search term.

[2]We already filtered out requirements that are beyond our control, e.g., that users want the public data to be qualified as correct.

3 THE ADEQUATE PLATFORM

Figure 1 displays the overall architecture of the ADEQUATe platform. The *Data Monitor* component (cf. the block at the bottom of the figure) is built on-top of the Open Data Portal Watch framework [4]. It harvests the datasets and respective metadata descriptions from the two Austrian Open Data portals `data.gv.at` and `opendataportal.at`. It gets triggered by the *Orchestration* component, which schedules weekly fetching, archiving, annotating, and quality assessment and improvement pipelines for the datasets.

The ADEQUATe platform is centered around the *ADEQUATe knowledge base*, a knowledge vault which contains entities, classes, and relations in the data available at Austrian Open Data portals. ADEQUATe knowledge base is used for (semi)automatically linking terms within the input data to Linked Data entities and also during dataset search to automatically build search facets. The ADEQUATe knowledge base can be maintained via the "PoolParty Thesaurus manager", a component from the PoolParty Semantic Suite[3].

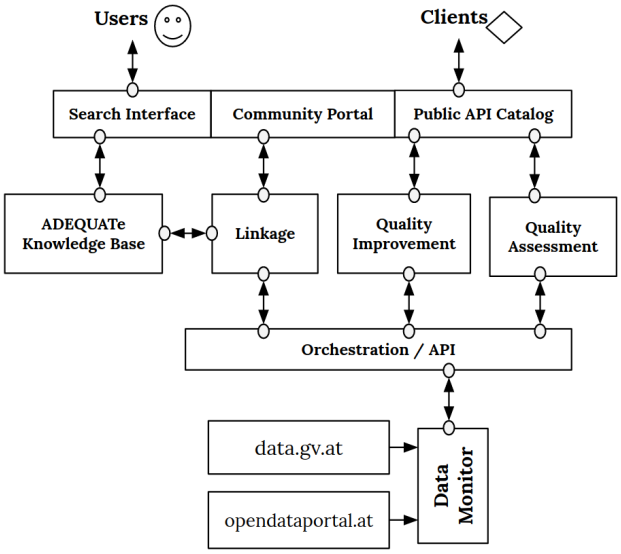

Figure 1: Overall architecture of the ADEQUATe platform.

The platform consists of the following components:

Quality Assessment. The *Quality Assessment* and reporting component computes the Open Data metrics developed in [4]. For instance, it displays quality metrics such as the availability of resources and the Open Data compliance of the dataset's license.

Quality Improvement. The *Quality Improvement* component focuses on the metadata descriptions and on tabular resources (i.e. CSV files): We map and homogenize the descriptions using the Schema.org and DCAT [5] metadata standards, and automatically complete the metadata using information such as the file size and file format, in detail described in [6]. Regarding the CSV files, for each file we normalize the encoding (UTF-8) and delimiter (","), and provide a single header row (in case of missing/multiple header rows). The improved metadata and "cleaned" CSV files get re-published at the community portal build on top of Gitlab[4].

[3]https://www.poolparty.biz/
[4]http://gitlab.com

Linkage & Search. The *Linkage* component is based on Odalic [7], a tool for semantically interpreting input tabular data (CSV files) and publishing them as Linked Data. A user can run Odalic on an input table and get suggestions for table *annotations* – we distinguish different types of annotations: (1) classifications of columns, (2) disambiguations of cell values, and (3) discovered relations between the columns. Odalic provides annotations based on the ADEQUATe knowledge base (and optionally based on an external SPARQL knowledge base, such as DBpedia), its internal semantic table interpretation algorithm, and also user feedback – the user may further fine-tune the suggested annotations, e.g., by marking certain disambiguation as wrong, manually setting column classification, or proposing new relations. Finally, the semantically interpreted data can be exported as RDF/Linked Data, saved back to the ADEQUATe platform, and also can further improve the search engine.

The *Orchestrator* in Figure 1 ensures that all the datasets, as they are harvested from their original locations, are also annotated with the entities from the ADEQUATe knowledge base. For those automatic annotations, we prepared UnifiedViews [8] pipelines using the "PoolParty extractor", a tool for automatic annotation of unstructured data.[5]

In order to resolve the users requirements w.r.t. dataset search, a search engine, implemented using the "PoolParty GraphSearch"[6], allows users to search datasets not only via full-text search, but also via facets based on the classes and entities in the ADEQUATe knowledge base. In the final version of the platform, the search engine will also use Odalic annotations produced by the community; however, this feature is not yet fully integrated.

Community Portal. The platform's *Community Portal* is based on Gitlab[7]: a web-based Git repository manager with several collaboration features. For each dataset there is a dedicated project in our ADEQUATe Gitlab instance where we re-publish improved versions of these datasets and metadata descriptions to make them available to the community. Additionally, we provide a landing page for all datasets, where we display the quality assessment results, the improved version of the resources, and provide users a way to interact with the datasets, e.g. to leverage CSV files to Linked data using Odalic (cf. *Linkage*). In order to enable the community to work on datasets users can fork and publish improved/changed versions of a dataset on the portal.

4 EXEMPLARY USER JOURNEYS ON THE PLATFORM

In this section, we provide three user journeys – an unexperienced rookie, an intermediate and a pro user – to exemplify the process of finding and working with datasets on the ADEQUATe platform. These user journeys represent, how the requirements and the skill levels of the members of the initial focus groups (see Section 2) are reflected by the developed functionality of the ADEQUATe platform.

User journey 1: The Rookie

Meet Michael, who is proficient in daily work with tabular data (e.g., CSV, Excel), but has no experience with open data at all. Thus, he has no knowledge about potential shortcomings, maintenance of open data, limitations of liability, or related quality and format issues. Furthermore, Michael has not heard of Open Data portals and their functionalities before. At a point in time, Michael recognizes that in order to be able to continue with his project, he has to get additional data. First thing for Michael to do is to search for data online. At this point, there are two alternative scenarios how Michael might get in touch with the ADEQUATe platform. In the first scenario, Michael discovers the data he is looking for on an open data portal, i.e., on opendataportal.at. While looking at the dataset, he discovers the "ADEQUATe button",[8] indicating that there is an alternative version of this dataset available. The button refers Michael to the detail page about the dataset on the ADEQUATe platform.

The second scenario brings Michael directly to the ADEQUATe platform through a search engine result. From here[9] he starts a search for interesting data, which provides an improved search experience through the underlying use of semantic technologies and Linked Data approaches, including advanced search options, such as faceted search. Both scenarios offer Michael a basic overview of the properties of the dataset of interest, covering all relevant information such as general descriptions of the dataset (e.g., release data, last time updated, related keywords, entities within the dataset etc.). Furthermore, he can see all available distributions of the dataset (i.e. data being available as CSV files, Shapefiles, etc.), paired with a list of assessed quality attributes of the data and its associated metadata, including suggestions of how to improve existing shortcomings. The list of downloads also includes an improved version, created automatically by the ADEQUATe platform. The improvements cover formatting and CSV standardization issues (e.g. standardizing inconsistent separators). In addition, Michael also sees the activity of the community regarding discussions and work on this particular dataset in regard to suggestions how to further improve the dataset.

While having a closer look at the dataset, Michael discovers that the description is only available in German language. As the dataset contains assets that could be of interest for an international audience, Michael would like to request an English version of the description. In order to file the request, he clicks on the discussion section within the dataset page, registers an account free of charge at the ADEQUATe platform and is immediately able to file his request to be picked-up by the community, including the owner of the dataset.

User journey 2: The Intermediate

Meanwhile, Margarita is working on her data project. She is experienced in working with different types of data, including open data. Due to the setting of her project, Margarita has distinct requirements and expectations regarding the format of the data she works on, including necessities originating out of technology-related compatibility issues. At one point in her work, she realizes that the

[5] The main difference between Odalic and "PoolParty extractor" is that Odalic runs on demand, takes into account table structure by employing unique semantic table interpretation algorithm, incorporates user feedback, and also allows customized Linked Data exports.
[6] https://www.poolparty.biz/poolparty-semantic-graph-search-server/
[7] https://about.gitlab.com/
[8] Cf. button "ADEQUATe Checked" at http://data.opendataportal.at/dataset/kunstler-der-sammlung-des-mumok
[9] Search datasets directly via the ADEQUATe page: https://www.adequate.at/

dataset she downloaded recently from opendataportal.at is missing entries, which should be in place, based on the description and metadata of the dataset. Thus, she heads back to the download page of the portal, to contact the authors/owner of the dataset. On the download page, she sees the "ADEQUATe button", indicating that there is an alternative version of this dataset available. She clicks on the button and arrives at detail page about the dataset on the ADEQUATe platform.

While browsing through the page, she discovers the discussion section on the platform. After she registered an account free of charge at the ADEQUATe platform, she dives right into the discussions. One of the first comments she comes across is Michael's request regarding an English description, to which she replies that this would be a good idea, as she also see the added value for an international audience.

Further down the line she learns that there exists a discussion about the exact issue she came here for in the first place, namely the missing entries within the dataset. It seems that this issue was reported some time ago already, and that an ADEQUATe user provided a fixed version of the dataset. She follows the link to the new version and discovers the ADEQUATe versioning history, showing the development path on the platform that led to this improved version, including all previous version in a history line. In addition, the ADEQUATe platform allows her to see the exact changes that have been made to the dataset, informing her about the new entries, which fit her request.

User journey 3: the Pro

And, finally, there is Kate, a seasoned data scientist with hands-on experience in dealing with data of any kind, ranging from closed proprietary data up to standardized open government data. Kate's current projects include the work with Linked Data, especially using RDF datasets. Unfortunately, the recent dataset she downloaded directly from the opendataportal.at did not included any semantic annotations. Yet, she knows of the ADEQUATe platform, she successfully used for her projects in the past, including its provision of additional tools for semantic enrichment.

She directly enters the platform and searches for the dataset via the enhanced search functionality of ADEQUATe. She quickly finds the dataset and proceeds to the overview page. Starting from there, she checks the available versions on ADEQUATe and recognizes that no enriched version is yet available. So, she initiates the ADEQUATe-integrated tool Odalic, which provides semantic enrichment capabilities. After she refines the classes and entities suggested by Odalic, Kate exports the enriched CSV file, together with an RDF version of the data. She likes the results and decides to provide the enriched version back to the community. Via the versioning functionality of ADEQUATe, Kate forks the dataset, adds the newly enriched version, and submits a pull-request to the dataset owner for the new version to be added to the repository.

In addition, she also decides to post a message about the new version on the discussion thread. While being there, she sees a request of a user named Michael regarding an English version of the description of the very same dataset. As this request is also endorsed by second user called Margarita, she decides to quickly translate the description and submits a second pull-request for this new version as well.

5 CONCLUSIONS

Based on focus group interviews and surveys we have identified three major challenges/needs for Open Data consumers: Firstly, the participants have reported overall quality issues with metadata and data itself; second, a lack of interoperability between data sources; and third, limited search functionalities at the data portals.

The ADEQUATe platform combines data and community driven approaches to address these issues. It consists of a framework which 1) continuously assesses the datasets' quality based on a comprehensive list of quality metrics; 2) it applies a set of heuristic algorithms to improve identified quality issues; 3) it uses Semantic Web technologies, semantic table interpretation tool Odalic and ADEQUATe knowledge base to transform legacy Open Data sources (CSV files) into Linked Data by detecting the underlying entities, classes, and relations, and enabling semantic search over those annotations; and 4) it allows community collaboration by forking and re-publishing datasets at the platform. In order to showcase how users with different technical/data-processing know-how can interact with the platform we have presented three user journeys.

As future work, we plan to integrate more tools, such as OpenRefine,[10] and to improve the search results by using annotations produced by Odalic. To guarantee a high level of exploitation and continuation of the outcomes of the ADEQUATe project, we plan to take over several parts of the existing platform in the substantively related Austrian project CommuniData.[11] The goal of the CommuniData project is to enhance usability and accessibility for non-expert users and it therefore is complementary to the ADEQUATe outcomes: the usability and community involvement aspects will be an ideal continuation and extension of the existing platform.

ACKNOWLEDGEMENT

ADEQUATe is funded by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) under the program "ICT of the Future" (grant no. 849982) between October 2015 and June 2018.

REFERENCES

[1] Thomas J Lampoltshammer and Johannes Scholz. *Open Data as Social Capital in a Digital Society*, pages 137–150. Cambridge Scholars Publishing, Newcastle upon Tyne, 2017.

[2] Martin Beno, Kathrin Figl, Jürgen Umbrich, and Axel Polleres. Perception of key barriers in using and publishing open data. *JeDEM-eJournal of eDemocracy and Open Government*, 9(2):134–165, 2017.

[3] Thomas Lampoltshammer and Johann Höchtl. Requirements Specification. Technical report, ADEQUATe Deliverable D1.2, 2016. https://www.adequate.at/wp-content/uploads/2016/07/D1.2RequirementsSpecification_del.pdf.

[4] Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. Automated quality assessment of metadata across open data portals. *Journal of Data and Information Quality (JDIQ)*, 8(1):2, 2016.

[5] Fadi Maali and John Erickson. Data Catalog Vocabulary (DCAT). http://www.w3.org/TR/vocab-dcat/, January 2014.

[6] Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. Lifting data portals to the web of data. In *WWW2017 Workshop on Linked Data on the Web (LDOW2017), Perth, Australia, April 3-7, 2017*, 2017.

[7] Tomáš Knap. Towards Odalic, a Semantic Table Interpretation Tool in the ADEQUATe Project. In *Proceedings of the 5th International Workshop on Linked Data for Information Extraction co-located with the 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 22, 2017.*, pages 26–37, 2017.

[8] Tomáš Knap et al. UnifiedViews: An ETL Tool for RDF Data Management. *Semantic Web Journal*, 2018 (to appear). http://www.semantic-web-journal.net/content/unifiedviews-etl-tool-rdf-data-management-0.

[10] http://openrefine.org/

[11] https://www.communidata.at/