ADEQUATe - Analytics and Data Enrichment to Improve the Quality of Open Data

Johann Höchtl*, Thomas J. Lampoltshammer**

** Danube University Krems, johann.hoechtl@donau-uni.ac.at*
*** Danube University Krems, thomas.lampoltshammer@donau-uni.ac.at*

Abstract: Over the past decade, open data became an important economic resources and therefore rapidly found their way into data warehouses and data management systems of companies and organizations world-wide. Yet providing these data and the associated data curation still presents a challenging task for all involved stakeholders. The ADEQUATe project therefore aims to develop and evaluate mechanisms to measure, monitor, and improve data quality of open data. In specific, statistical and algorithmic methods, data linkage, as well as crowdsourcing approaches will be employed to boost data quality. The combined outcomes of the project will lead to a Quality Monitoring and Evaluation Framework that is deployed in two real-world us-cases (namely data.gv.at and opendataportal.at) for evaluation and refinements, following a user- and data-driven development approach.

Keywords: Open data, data quality, quality metrics, data enrichment, crowd sourcing

Acknowledgement: The ADEQUATe project is funded by the Austrian Research Promotion Agency (FFG) under grant no. 849982; the authors would also like to thank their project partners Semantic Web Company GmbH as well as the Vienna University of Economics and Business.

1. Introduction

Over the past decade, open data became an important economic resource and therefore rapidly found their way into data warehouses and data management systems of companies and organizations world-wide. Yet providing these data and the associated data curation still presents a challenging task for all involved stakeholders. But not only the economy sector faces challenges regarding open data, the public sector does as well. Due to legal binding such as the PSI directive[1], governmental bodies in Europe face challenges for publishing open data to foster transparency and economy in the European market. In this course, the ADEQUATe project identifies two main critical issues that have to be overcome in order to unleash the full potential of open data in the

[1] https://ec.europa.eu/digital-agenda/en/european-legislation-reuse-public-sector-information

before-mentioned domains: i) existing overall quality issues regarding data and the associated meta data, and ii) the missing interoperability between existing data sources.

In order to solve these main challenges, the project pursues a data and community-driven working approach, consisting of these three main steps: i) constant data quality monitoring of the two use-case portals, ii) (semi-) automated identification of potential quality issues via smart algorithms and community-based input, and iii) the application of semantically-enabled Web technologies to convert legacy data towards linked data.

2. Publishing and Consuming Open Data

Currently, two major frameworks position themselves world-wide for publishing and hosting Open data: the commercial portal Socrata[2] and the open source framework CKAN[3]. Both systems provide a content management system to search within the hosted datasets, as well as an API for external services to connect to the platform. Furthermore, both systems offer possibilities for extensions and plugins to improve the overall connectivity and productivity.

The Socrata system positions itself as a cloud-based Software-as-a-Service (SaaS) platform regarding data publishing and visualization. The largest platform installations can be found in the US, specifically in New York, Austin, or Maryland. The most prominent feature that comes with Socrata is its rich API environment.

CKAN is used by over 100 public installations, covering application fields such as European governmental portals, but also in countries of North and South America, and the Middle East. One of the biggest strengths of CKAN is its open architecture, which enables the integration of community-contributed extensions, the connection to external CMs, provision of an environment for linked data, as well as the option to serve as a meta portal by providing a unified view onto other CKAN portals.

While the commercial and the community world does not always "play well together", there exists many examples where CKAN data portals have been successfully integrated into publicly endorsed open data installations and monitoring solutions, which has already been demonstrated by projects towards the monitoring of data quality, e.g. the Open Data Institute[4], or Fraunhofer FOKUS[5].

Data Quality assessment and methodologies towards the improvement of data quality can be found throughout various research areas such as information systems, data warehouses, databases, and (linked) data pools. Quality metrics and techniques for measuring (meta) data

---

2 https://www.socrata.com

3 http://ckan.org

4 http://theodi.org/blog/how-important-is-open-data-quality

5 http://open-data.fokus.fraunhofer.de/?page_id=125

quality have been developed to keep pace with the ever increasing amount and complexity of associated tasks (Zhu et al., 2014). In general, existing approaches can be divided into four steps: i) defining the metrics, ii) measuring the quality of data sets according to these metrics, iii) analyze the outcomes, and iv) finally reveal possible improvement scenarios with optional feedback loops. Out of these research works, numerous quality assessment methodologies for Linked Data have arisen (Hogan et al., 2012). For example, Kontokostas et al. (2014) present an approach emerging from test-driven software development towards test-driven Linked Data quality evaluation. The core idea behind this concept is to provide test cases via defined SPARQL queries, similar to unit tests in software development. This enables a high level of flexibility in terms of adapting to various metrics as required and, at the same time, provide an integrity check as all necessary attributes have to be available in order to complete the test.

Besides unit-testing-related approaches, there exist movements towards minimal information models. These RDF-based models enable automated conformance evaluation of data objects or Web service based on pre-defined evaluation criteria (Gamble et al., 2012).

Another research direction is focusing on data repositories, e.g. CKAN, and the challenge of identifying suitable technologies for data quality improvements (Kučera et al., 2013). One option for the improvement of data consistency regarding meta data comes in form of controlled vocabularies. This is an important step as at the moment, portals can and do define own, non-standardized meta data keys (Gamble et al., 2012).

One of the main challenges associated with data quality metrics is the high rate of human involvement, which presents an obstacle regarding distributed large-scaled systems. Yet, first steps have already been taken, using a rule-based system for automated quality monitoring (Cappiello et al., 2005). This monitoring system is based on a two-way approach by performing necessary assessment and refinement actions automatically and demanding for human interaction where necessary. Comparable to this approach is the rsine framework[6] for real-time quality checks in Semantic Web environments, originating from the EU project LOD2 (Mader et al., 2014).

3. Open Issues and Challenges in ADEQUATe

The overall availability of open data continually rises as does its business and economical value. Forecasts state open data will boost the US economy in several sectors by $3 trillion[7]. At the same time, forecasts predict for the European realm a GDP increase of € 200 billion until the end of the current decade[8]. Yet, open data still faces quality issues regarding the data themselves and the

[6] https://github.com/rsine/rsine

[7] http://www.mckinsey.com/~/media/McKinsey/Business%20Functions/Business%20Technology/Our%20Insights/Open%20data%20Unlocking%20innovation%20and%20performance%20with%20liquid%20information/MGI_Open_data_FullReport_Oct2013.ashx

[8] https://www.microsoft.com/global/eu/RenderingAssets/pdf/2014%20Jan%2028%20EMEA%20Big%20and%20Open%20Data%20Report%20-%20Final%20Report.pdf

associated meta-data. The two main problems are i) the overall quality issues, and ii) the limited or missing interoperability of the data. Based on these problems, the authors identify the following challenges within the ADEQUATe project:

Challenge 1: To establish sustainable quality metrics for open data

The importance of measuring quality of data was already identified and pursued by other research projects (e.g., the OpenDataMonitor[9]). Yet, the heterogeneity of data producers, data sources, and data consumers require the metrics integrated in open data portals to be specifically tailored towards stakeholder needs. While there exist metrics, which are meeting standard requirements of open data sets, others have to be created or adapted to meet expectations in present scenarios. For example, the absence of geoinformation in a geo-tagged data set should definitely affect its quality ranking, while it does not make sense to include this metric in non-geoinformation data sets, as it would falsify the overall quality ranking. Once suitable metrics where identified and implemented, the overall quality score across these metrics has to be monitored to counter decreasing quality trends in time.

Challenge 2: To automatically improve (meta) data quality

While in a perfect world, all necessary data and associated meta data would be already in place during the upload of the data set to the portal, reality looks different. As it is not economically reasonable to perform cleaning and enrichment procedures manually, flexible and generic algorithms have to be implemented for an automated correction of affected data sets. While context-related information can be acquired with relatively low effort (e.g., the file extension or the language of the content), missing or erroneous provenance information is hard to correct or recreate e.g., via crosschecks with other similar data sets or collating associated time-value series. While temporal information within and related to a particular data set can help to improve the overall quality of the data set itself and – if possible – of associated data sets as well, dealing with the heterogeneous structures of the data sets is non-trivial. To set a step into the right direction, the ADEQUATe project aims at the improvement of CSV files in terms of syntax clearing, encoding, and data fusion (unification).

Challenge 3: To understand CSV-based data sources

Due to the high grade of diversity regarding data providers and data processing chains, understanding the inherent structure and logic over several "versions of CSV dialects" is challenging. For example, CSV can be manually exported from spreadsheet software such as Excel and features multiple tables or descriptive information inside the CSV file, which invalidates the original syntax proposed by the standard. Furthermore, organization-related internal meta data descriptors are often hardly useable without suitable "meta-meta" data for their description, which can render the entire data set challenging to be usable by external data users. Finally, uploaded

[9] http://opendatamonitor.eu/frontend/web/index.php?r=dashboard%2Findex

CSV files are usually way larger than other forms, e.g., Web tables, which can negatively impact processing algorithms in terms of computational resources and timely availability.

Challenge 4: To engage the community to cooperate

While sophisticated algorithms can technically and semantically maintain and improve the quality of (meta) data, there are limits due to the lack of expert and domain knowledge, which can be brought into the quality improvement process by the potential users and contributors to open data platforms. To set the foundation for a sustainable and vibrant open data platform, user needs and requirements have to be incorporated from the very beginning. The ADEQUATe project will achieve this setting, via crowd-sourcing approaches to receive direct feedback from the community. This is even more important as reviews have revealed (Prieto-Martín et al., 2011) that projects ignoring these inputs do fail in generating public impact.

4. Outlook

The requirements elicitation process is well under way and will be finished by end of March 2016. During that course, more than 100 users filled out the online survey[10] and 120 people participated during the focus groups which were organized around meetings of Cooperation OGD Austria[11], Hackathons and Barcamps. The next step will be to categorize the input and distil the most frequent requirements and mentioned impediments to data quality. This input will be fused into the architectural blueprint which will guide the subsequent implementation of the data improvement and monitoring framework. As the ADEQUATe project puts its focus on CKAN, many of the to be developed and integrated data quality components will be made available as plugins for this open software stack. Of ongoing and intense discussion are the various possibilities to include the end users into the data quality improvement process. Results of past participation projects will provide guidance on promising participatory elements which will be provided by the projects community portal. Participation in a user-friendly way to actually improving data quality, either by uploading changed datasets according to a fork-and-push model as provided by the Dat-project[12] or by publishing data transformation steps as recorded by OpenRefine[13], are still under intense discussion. The community platform will also serve as a test-bed to determine unobtrusive, promising quality improvement methodologies and processes which will be implemented on the authoritative Austrian data portals data.gv.at and opendataportal.at.

---

[10] http://odsurvey.ai.wu.ac.at/index.php/637325?lang=de

[11] https://www.data.gv.at/infos/cooperation-ogd-oesterreich/ Cooperation OGD Austria is a think tank operated by open data portal operators in Austria and includes academia and business representatives to maximize

[12] http://dat-data.com/

[13] http://openrefine.org/

References

Cappiello, C., Francalanci, C., & Pernici, B. (2005). A self-monitoring system to satisfy data quality requirements. In On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE (pp. 1535-1552). Springer Berlin Heidelberg.

Gamble, M., Goble, C., Klyne, G., & Zhao, J. (2012, October). Mim: A minimum information model vocabulary and framework for scientific linked data. In E-Science (e-Science), 2012 IEEE 8th International Conference on (pp. 1-8). IEEE.

Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., & Decker, S. (2012). An empirical survey of linked data conformance. Web Semantics: Science, Services and Agents on the World Wide Web, 14, 14-44.

Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., & Zaveri, A. (2014, April). Test-driven evaluation of linked data quality. In Proceedings of the 23rd international conference on World Wide Web (pp. 747-758). ACM.

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

Mader, C., Martin, M., & Stadler, C. (2014). Facilitating the exploration and visualization of linked data. In Linked Open Data--Creating Knowledge Out of Interlinked Data (pp. 90-107). Springer International Publishing.

Prieto-Martín, P., de Marcos, L., & Martínez, J. J. (2011). The e-(R) evolution will not be funded. European Journal of ePractice, 15, 62-89.

Zhu, H., Madnick, S. E., Lee, Y. W., & Wang, R. Y. (2014). Data and Information Quality Research: Its Evolution and Future.

About the Authors

Johann Höchtl

Johann Höchtl graduated from University of Vienna and Vienna University of Technology in Business Informatics. He is research fellow at Danube University Krems, Center for E-Governance, Austria. His projects include EU-funded research projects and national grants in the domain of social media application in administration, open data and ICT in public administration. He is former member of OASIS SET TC standardization group. Johann Höchtl is member of OKFN Austria and member of Cooperation Open Government Data Austria, where he is heading data quality sub working group. His current research focus is in the domain of Open Data, the effects of ICT application in a connected society and semantic technologies.

Thomas J. Lampoltshammer

Thomas Lampoltshammer holds a doctoral degree in Applied Geoinformatics in addition to his Master's degrees in the fields of Information and Communication Technology, Embedded and Intelligent Systems, as well as in Adult Education. He currently works as a Senior Researcher (postdoc) at the Department for E-Governance and Administration at the Danube University Krems/Austria. His research experience covers national and EU-funded projects in ICT-related topics, such as Geoinformatics, Semantics, Social Media, Legal Informatics, and E-Health. Furthermore, he acts as reviewer for several SCI-indexed journals.