Quality Driven Linked Data Generation

Alex Randles
ADAPT Centre for Digital Content,
Trinity College Dublin,
Dublin, Ireland
alex.randles@adaptcentre.ie

Declan O'Sullivan
ADAPT Centre for Digital Content,
Trinity College Dublin,
Dublin, Ireland
declan.osullivan@adaptcentre.ie

ABSTRACT

Linked Data datasets are continuously being published with varying levels of quality. Resolving quality issues within the mapping artefacts which produce these datasets will positively impact the quality, reuse and maintenance of these datasets. A process for ensuring quality of the mapping artefacts themselves needs to be tested thoroughly with respective end users to ensure optimal usability is offered. Testing of the mappings will enable the design and requirements for the mappings to be refined and improved.

CCS Concepts

• **Human-centered computing** → **Usability testing** • **Social and professional topics** → Quality assurance • **Information Systems** → Resource Description Framework (RDF)

Keywords

Mapping Quality; Dataset Quality; Linked data generation.

ACM Reference Format:

Alex Randles and Declan O'Sullivan. 2022. Quality Driven Linked Data Generation. In Proceedings of The 1st ADAPT Annual Scientific Conference (ADAPT ASC'22). ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/x.

1. INTRODUCTION

Data quality is a multidimensional concept which is determined by the stakeholders and factors involved in the creation of the data [3]. The quality of the data will affect how useful data consumers find the data for their application. Oftentimes, it is the responsibility of each data consumer to consistently check the level of quality.

The objective of our research is to assist data providers in producing high quality linked data (represented in the graph based Resource Description RDF format) by bringing quality improvement procedures earlier into the publication process, thus resolving limitations that exist in the state of the art.

The limitations include that the state of the art in linked data quality focuses on the quality of the published dataset and not on the mapping artefacts that produce them. The mapping artefacts typically define transformation rules for converting non-graph based data (e.g. excel or relational data) into graph based RDF data. Creating these mappings is a complex, time-consuming task, which is frequently error prone [1]. Furthermore, creating high quality mappings requires a high level of background knowledge. Oftentimes, quality issues within these mappings are not detected until the dataset has been published. Quality improvement procedures which focus on these mapping artefacts will allow a significant number of root causes for published dataset quality issues to be identified and resolved.

In this paper, we provide a description of our mapping quality assessment framework that is used to generate Linked Data datasets. Furthermore, the structure of the usability experiment, the analysis of the results and suggested improvements are summarized.

2. THE MQV FRAMEWORK

The Mapping Quality Vocabulary (MQV)[1] framework [8] is a framework designed for the assessment and refinement of uplift mappings. Uplift mappings are those that specify how to transform non-graph based data into RDF graph based data. The objective of the framework is to improve the quality of mappings and the resulting dataset, while promoting mapping maintenance and reuse. The framework represents the quality information generated during the assessment process in RDF format using the Mapping Quality Vocabulary [6,7]. MQV is used to represent and allow interchanging of provenance information relating to the creation, quality assessment and quality refinement of mapping specifications.

Assessment process. The framework uses a suite of quality metrics to assess the quality of the mapping. These metrics assess quality aspects such as the extent to which the mapping correctly conforms to the mapping specification, the quality of the dataset output generated by mapping processor and the quality of the vocabularies used within the mapping definition.

Refinement process. The framework offers semi-automatic refinements which involve a human in the loop, where the framework guides the user in refining the mappings to resolve detected mapping quality issues. Refinements have been designed for each quality metric used to detect an issue. Once a refinement option has been selected, the framework can fetch and suggest changes which could be used to resolve these quality issues. Once confirmed by the user, the changes are made to the mappings.

3. USABILITY EXPERIMENT

A usability experiment has been conducted with an implementation of the MQV framework[2]. The usability experiment involved participants interacting with the interface of the framework using a mapping provided. The tasks were designed to test the main functionality of the framework (resolution of issues with a given mapping), followed by the examination of the reports generated.

Participants. The participants were grouped into two cohorts. These cohorts included a student and expert cohort, with a view to characterizing the cohorts based on background knowledge.

[1] Mapping Quality Vocabulary Specification at https://alex-randles.github.io/MQV/

[2] Video demonstration of framework at https://drive.google.com/file/d/1LzO-2CuVv8WLSGE6VaNKqmh3B6Q-osPv/view?usp=sharing

Student cohort. These participants were recruited from the MSc Knowledge and Data Engineering module in Trinity College Dublin. These participants had limited experience in creating R2RML mappings. 59 students were invited to participate in the study, however, 48 were included in the results, after inclusion/exclusion criteria was applied.

Expert cohort. The expert cohort included 10 semantic web researchers who had experience in creating R2RML mappings in a research environment. These participants were recruited by agreement with the study's supervisor with respect to known expertise in the creation of uplift mappings.

Experimental setup. The set up for the experiment differed slightly for participants in each cohort.

Student cohort. The student cohort completed the experiment asynchronously by accessing the framework using provided login details. The think-aloud protocol was not used for the cohort as it would not be feasible to transcribe/analyze each of their think-aloud statements.

Expert cohort. The expert cohort used the think-aloud protocol [2], which involves verbalizing their thoughts as they complete each task. The experiment was conducted synchronously using the zoom video conferencing platform.

Data collected. The data collected during the study was in quantitative and qualitative data format.

Quantitative data. The Post-Study System Usability Questionnaire (PSSUQ) [4] was used to measure the participants perceived satisfaction of their usage of the framework. The questionnaire uses a 7-point Likert scale to measure the user's satisfaction. The time spent to complete the experiment and each task was also recorded. The number of mapping quality issues resolved was calculated.

Qualitative data. The session transcripts were recorded for the expert cohort, which was used to analyze their think-aloud statements. Although the think-aloud protocol was not used for the student cohort, the PSSUQ open-comment section allowed for both students and experts to provide textual statements relating to usability.

Data analysis. Thematic analysis [5] was conducted on the *Qualitative data*, in order to discover emerging themes that will guide improvements to the framework. The scores of each PSSUQ metric were calculated for each participant and averaged for each cohort. Thereafter, these scores were compared against norms from the state of the art as no previous version scores for the MQV framework exists.

Experiment Results. The PSSUQ consists of four quality metrics[3], which include Interface quality, Information quality, System usefulness and Overall usability metric. The percentage values related to these metrics within **Table 1** are the difference with the state of the art norms [4]. "Timing" relates to the time spent completing the experiment and "At least one quality issue resolved" refers to the percentage of participants who resolved at least one quality issue.

Table 1. Comparison of each cohort results

Cohort	Student cohort	Expert cohort
Interface quality	-11.07%	-9.45%
Information quality	+24.79%	+24.27%
System usefulness	+19.65%	+65.68%
Overall usability	+16.52%	+33.64%
Timing	10.06 minutes	15.4 minutes
At least one quality issue resolved	81.25%	100%

The results show the expert cohort scores better for all metrics except the information quality metric, which could be as a result of the student cohort being less critical of the information. The faster completion time for student cohort could be as result of them taking less consideration when completing each task. Furthermore, expert participants could have spent more time analyzing the information on the framework.

High Level Summary of Findings. Taking into account the data analysis, it was indicated that the MQV framework facilitates the assessment and refinement of mappings as more than 90% of participants could resolve at least one quality issue. Furthermore, only one participant required assistance to complete the tasks. Moreover, the participants in the expert cohort, who have significantly more semantic web knowledge, scored the framework more usable in the PSSUQ and had a lower unresolved issue count in refined mapping generated.

Improvements. The feedback provided by participants is being used to refine the requirements of the framework and to guide improvements necessary for the revised version. The most mentioned feedback related to the aesthetics of the interface, textual descriptions of refinements, representation of the provenance information.

Interface aesthetics. The most common theme within the thematic was the "Unaesthetic Interface" theme. Furthermore, the interface quality metric was the worst scoring within the PSSUQ for both cohorts. The aesthetics of the interface has been improved in the revised version of the framework.

Refinement descriptions. The refinements allow users to resolve quality issues semi-automatically. However, it was discovered through thematic analysis that textual descriptions for these refinements was unclear. Additional textual descriptions have been added to refinements to ensure users can easily select and execute an appropriate refinement.

Provenance information. The provenance information is represented visually in a chart and machine-readable format using MQV. The chart represents the relationship between quality issues and their corresponding quality dimensions, however, the visual representation was unclear for certain participants. Participants were unsure which quality issue related to each dimension. The issue has been resolved by adding the ID of the quality issue to the chart.

Once the refinements to the framework based on what we have learnt from this evaluation are finished, focus will shift to how the framework can be triggered by detecting quality issues arising in either source or linked data datasets that might prompt a user to refine their mappings.

[3] PSSUQ metric information at https://uiuxtrend.com/pssuq-post-study-system-usability-questionnaire/

4. ACKNOWLEDGMENTS

This research was conducted with the financial support of the SFI AI Centre for Research Training under Grant Agreement No. 18/CRT/6223 at the ADAPT SFI Research Centre at Trinity College Dublin.

5. REFERENCES

[1] Ademar Crotti Junior, Christophe Debruyne, and Declan O'Sullivan. 2017. Juma: An Editor that Uses a Block Metaphor to Facilitate the Creation and Editing of R2RML Mappings. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 87–92. DOI:https://doi.org/10.1007/978-3-319-70407-4_17

[2] Marsha E Fonteyn, Benjamin Kuipers, and Susan J Grobe. 1993. A description of think aloud method and protocol analysis. *Qual. Health Res.* 3, 4 (1993), 430–441.

[3] Ademar Crotti Junior, Jeremy Debattista, and Declan O'Sullivan. 2019. Assessing the Quality of R2RML Mappings. In *Joint Proceedings of the 1st International Workshop On Semantics ForTransport and the 1st International Workshop on Approaches for MakingData Interoperable co-located with 15th Semantics Conference (SEMANTiCS2019), Karlsruhe, Germany, September 9, 2019* (CEUR Workshop Proceedings), CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-2447/paper2.pdf

[4] James R. Lewis. 2002. Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies. *Int. J. Hum. Comput. Interact.* 14, 3–4 (September 2002), 463–488. DOI:https://doi.org/10.1080/10447318.2002.9669130

[5] Lorelli S Nowell, Jill M Norris, Deborah E White, and Nancy J Moules. 2017. Thematic analysis: Striving to meet the trustworthiness criteria. *Int. J. Qual. methods* 16, 1 (2017), 1609406917733847.

[6] Alex Randles, Ademar Crotti Junior, and Declan O'Sullivan. 2020. Towards a vocabulary for mapping quality assessment. *To Appear Proc. 15th Int. Work. Ontol. Matching 19th Int. Semant. Web Conf. (ISWC), 2020.* (2020).

[7] Alex Randles, Ademar Crotti Junior, and Declan O'Sullivan. 2021. A Vocabulary for Describing Mapping Quality Assessment, Refinement and Validation. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, 425–430. DOI:https://doi.org/10.1109/ICSC50631.2021.00076

[8] Alex Randles and Declan O'Sullivan. 2021. Assessing quality of R2RML mappings for OSi's Linked Open Data portal. *4th Int. Work. Geospatial Linked Data ESWC 2021* (2021).