Towards a Knowledge-Aware AI
A. Dimou et al. (Eds.)

doi:10.3233/SSW220006

Evaluating Quality Improvement Techniques Within the Linked Data Generation Process

Alex Randles[a1] and Declan O'Sullivan[a]
[a]*ADAPT Centre for Digital Content, Trinity College Dublin, Ireland*

Abstract. Linked Data datasets when they are published typically have varying levels of quality. These datasets are created using mapping artefacts, which define the transformation rules from non-graph based data into graph based RDF data. Currently, quality issues are detected after the mapping artefact has been executed and the Linked Data has already been published. It is argued in this paper that addressing quality issues within the mapping artefacts will positively improve the quality of the resulting dataset that is generated. Furthermore, we suggest that an explicit quality process for mappings will improve quality, maintenance, and reuse. This paper describes the evaluation of the Mapping Quality Vocabulary (MQV) Framework, which aims to guide linked data producers in producing high quality datasets, by enabling the quality assessment and subsequent improvement of the mapping artefacts. The evaluation of the MQV framework consisted of 58 participants with varying level of background knowledge.

Keywords. Semantic Web; Mapping Quality; Dataset Quality; Linked data generation.

1. Introduction

Data quality is often referred to as "fitness for use" [1] and is a multidimensional concept which is determined by the stakeholders and factors involved in the creation of the data [2] . The quality of the data will affect how useful data consumers find the data for their application. Currently, quality assessment within the linked data domain is performed on published data and is the responsibility of data consumers rather than the producers [3]. This paper presents the evaluation of the MQV framework [4] that is designed to address the problem of quality earlier in the linked data publication lifecycle. The objective of the framework is to assist data providers in producing high quality linked data by bringing quality improvement procedures earlier into the publication process, thus resolving limitations that exist in the state of the art, where the focus typically is on the quality of the published dataset and not on quality of the mapping artefacts that produce them. The mapping artefacts typically define transformation rules for converting non-RDF data (e.g. excel or relational data) into RDF data. The W3C recommendation for transforming relational databases to RDF data, R2RML [5] is one example of an uplift mapping language. R2RML is used to express customized

[1] Alex Randles, ADAPT Centre for Digital Content, Trinity College Dublin, Dublin 2, Ireland; E-mail:alex.randles@adaptcentre.ie.

transformation rules. Creating these mappings is a complex, time-consuming task, which is frequently error prone [2]. 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. In our research, we argue that introducing 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. Furthermore, removing quality issues from the rules which generate the dataset will ensure these issues do not appear if the dataset is regenerated. In this paper, we provide a discussion of the structure and results of a usability evaluation of the MQV framework[2] which was conducted with 58 participants. The paper is structured as follows: Section 2 describes the related work within the state of the art; Section 3 presents the MQV framework; Section 4 presents an evaluation of the MQV framework and discusses the results and Section 5 presents final remarks.

2. State of the Art

The state of the art in mapping quality frameworks for linked data has been reviewed. We argue that evaluating the quality of linked data tools with potential end users should be undertaken to demonstrate the usefulness of the design [6]. While several of the approaches in the state of the art have been adopted by users within the community, none of the approaches described have conducted an evaluation which studies user interaction. Most of these approaches have been evaluated using a system evaluation, while the evaluation described within this paper has used a large sample size of users and standardized usability methods.

EvaMap [7] is a mapping quality framework used to assess and improve the quality RDF mappings. The work uses YARRRML mappings, which are a human readable representation of RDF mappings. The framework uses a set of metrics organized into 7 dimensions to assess the quality of the mappings or the resulting datasets when instances are required. Weights can be associated with metrics to provide different importance. Furthermore, a global quality score is generated to represent the overall quality of the mapping. Moreover, feedback is provided to users on how to improve the quality. The reports generated by the framework are human-readable and not machine-readable. An evaluation has not been completed on the framework.

The approach [2] designed by the researchers extends an existing linked data quality assessment framework named Luzzu framework [8]. Noteworthy, the approach focuses only on quality assessment and does not concern quality improvement. R2RML mappings [5] are assessed using metrics which are commonly used to assess dataset quality. Luzzu is extensible which allows the users to add additional metrics to the framework. Four metrics have been implemented by the framework which relate to the representational category [1] of data quality. Luzzu generates two machine-readable reports, however, the problem report is the focus of the work. An evaluation was completed on mappings from a real world uses case. The results show the potential to identify quality issues in certain cases. The approach was found to be reasonably accurate at identify quality issues, however, there was certain cases where ontologies could not be retrieved and queried.

[2] MQV framework at https://mqv-framework.adaptcentre.ie/

Resglass [3] provides a rule-driven methodology to detect inconsistencies within the rules used to generate linked data datasets. The approach ranks rules and ontology terms in order that should be inspected by an expert based on a score. Refinements are completed by an expert. Inconsistencies within the dataset are used to refine the rules and ontologies again. The work provides an implementation which targets RML mappings. The inconsistencies are detected using a rule-based reasoning system [9]. The methodology has been applied to two real-life use cases DBpedia and Computer Science bibliograph (DBLP). The researchers discuss manual refinements which could potentially be used to remove these inconsistencies.

The approach [10] provides a test-driven approach for mapping assessment and semi-automatic refinements based on the quality assessment. The implementation targets RML mapping language and extends RDFUnit [11] which is an RDF test-case-based architecture. The RDFUnit test cases are extended to apply to mappings by adjusting the assessment queries. The semi-automatic refinements query the RDFUnit serializations of the quality information which enables triples to be add/delete or suggest actions to the user. The evaluation was applied to diverse use cases which included DBpbedia and iLastic. The mappings collected were assessed which detected a large number of quality issues and a discussion of possible semi-automatic refinements. The results indicated that assessing mappings is more efficient in terms of computational complexity and requires significantly less time compared to assessing the dataset.

3. MQV Framework

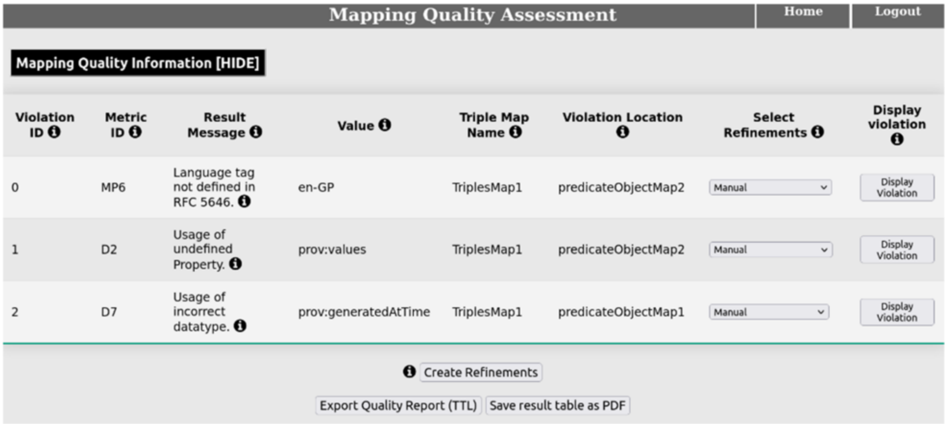
The Mapping Quality Vocabulary (MQV) framework[3] [4] is a framework designed for the assessment and refinement of uplift mappings. Uplift mappings specify how to transform non-RDF data into RDF data. The objective of the framework is to improve the quality of these mappings, which will improve quality of 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[4] [12,13]. MQV is used to represent and allow interchanging of provenance information relating to the creation, quality assessment and quality refinement of mapping specifications.

3.1. Design

A screenshot of the user interface of the MQV framework displaying the quality information for the mapping used during the evaluation is shown in **Figure 1.**

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

[4] Mapping Quality Vocabulary (MQV) specification at https://alex-randles.github.io/MQV/


Mapping Quality Assessment
Home
Logout
Mapping Quality Information [HIDE]
Create Refinements
Export Quality Report (TTL)
Save result table as PDF

Violation ID	Metric ID	Result Message	Value	Triple Map Name	Violation Location	Select Refinements	Display violation
0	MP6	Language tag not defined in RFC 5646.	en-GP	TriplesMap1	predicateObjectMap2	Manual	Display Violation
1	D2	Usage of undefined Property.	prov:values	TriplesMap1	predicateObjectMap2	Manual	Display Violation
2	D7	Usage of incorrect datatype.	prov:generatedAtTime	TriplesMap1	predicateObjectMap1	Manual	Display Violation

Figure 1: Screenshot of the user interface of the MQV framework

Figure 2 shows the component diagram of the MQV framework, which is designed using a Python web application. The application uses the RDFLib library [14] to query and update the mapping graph using SPARQL queries.

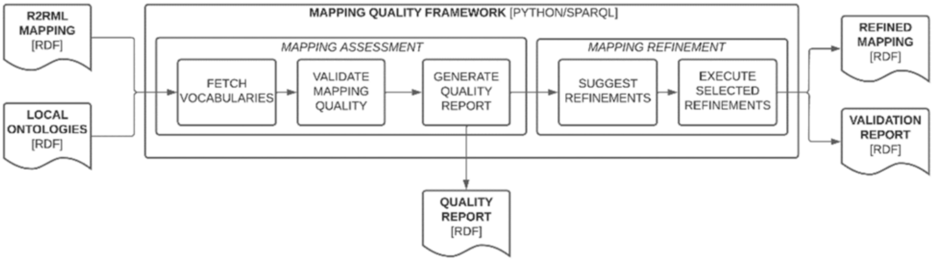


Figure 2: Component diagram of the MQV framework

The process starts with an R2RML [5] mapping and optional local ontology input into the framework. A local ontology refers to an ontology which is not available online, which could be currently being used for testing purposes. The vocabularies used within the mappings are fetched from online and stored in a local cache, which helps improve performance by querying the local copy. These vocabularies are queried using metrics defined as SPARQL queries to validate the quality of the mapping.

For example, the datatype range defined for a predicate within the mapping can be compared against the range within the vocabulary to ensure it is correct. A quality report is generated in MQV format after the mapping quality has been assessed. Refinements are suggested to the user based on quality issues within the mapping. These refinements are semi-automatic refinements which guide the users through the selection and execution. Each refinement has specifically been created for the quality issue. Once the refinements have been executed by the framework, a refined mapping is generated, which is a result of these refinements. Furthermore, a validation report is generated in MQV format which details the quality issues and the refinements which have been executed to resolve these issues. Moreover, the SPARQL query which was executed on the mapping during the refinement process is contained within the report.

3.2. Quality Assessment

The framework assesses the quality of mappings using domain specific metrics. These metrics assess different quality aspects within the mapping, which include Vocabulary, Mapping and Data quality aspects [15]. These metrics and aspects have been inspired from the state of the art in Linked Data quality [1,2,5,15–17].

Mapping Quality Aspect. The aspect ensures the concepts defined within the mapping conform to the specification of mapping language. For example, a join condition must have exactly one parent and child column.

Data Quality Aspect. The aspect focuses on the quality of the output which will be generated when the mapping processor has executed the mapping. For example, a non-dereferenceable class definition within the mapping will result in an exponential number of non-dereferenceable classes within the dataset, thus decreasing the quality of the data.

Vocabulary Quality Aspect. The quality of the vocabularies which are used within the mapping. For example, a class defined within the mapping should contain human-readable labels within the vocabulary**.**

A **quality violation** is generated when a metric related to one of these quality aspects detects a quality issue within the mapping. Thereafter, the violation can be refined by using the frameworks semi-automatic refinements.

3.3. Quality Refinement

The semi-automatic refinements which involve a human-in-the-loop can be described as three different methods, which have been inspired by previous research [3]. These methods are outlined below.

Insert custom value. Inserting a custom value involves the user entering an IRI or Literal value within a text box. Thereafter, the framework will replace the violation value within the mapping with the value entered. Prefixes are provided on the framework which could help to create the IRI. For example, if an undefined property is used within the mapping, users can select a prefix and enter the remaining IRI within the text box, which will replace the undefined property.

Select from suggested values. Selecting a value from suggested values involves the users browsing a drop-down menu and selecting a value. These values are designed to resolve the quality issues. Thereafter, the framework will replace the quality violation value with the selected value. For example, if an undefined property is used within the mapping, defined properties within the same namespace will be suggested to the user. Thereafter, users will select one of the values and the framework will replace the undefined property.

Insert suggested value. Inserting a suggested value involves the framework suggesting only one value to the user, which could hopefully resolve the quality issue. If the users are satisfied with the suggestion, the value will replace the violation value. For example, if a datatype defined within the mapping does not match the datatype defined within the vocabulary. The datatype from the vocabulary will be suggested to the user.

Once the refinements have been executed, a validation bar chart is displayed which shows the relationship between each quality violation and their corresponding quality dimension. These dimensions have been inspired from previous research [1] in linked data quality. The refined mapping generated by the refinements is available to download on the framework by pressing a button.

4. Evaluation

A usability experiment has been conducted with an implementation of the MQV framework. 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.

4.1. Experiment cohorts

The participants were grouped into two cohorts. These cohorts included an expert and student cohort. These cohorts' recruitment process and background knowledge differed. Grouping participants into two cohorts allow them to be characterized based on background knowledge.

Recruitment. The expert participants were recruited based on a discussion with the supervisor of the study who would meet the inclusion/exclusion criteria. These participants were recruited individually through email invitation. These participants completed the experiment to contribute to the research objectives. The participants from the student cohort were recruited from the Knowledge and Data Engineering (CS7IS1) module in Trinity College Dublin. Each member of the class had the option to complete the experiment as a portfolio task for the course.

Background. The participants within the expert cohort are Semantic web researchers who are very knowledgeable with RDF and the R2RML mapping language. These participants have previous experience in creating and executing R2RML mappings. Participants from the student cohort have little knowledge of the theory of the R2RML mapping language. Furthermore, these participants have little experience with creating R2RML mappings, however, they have basic knowledge of semantic web technologies. Each cohort's background knowledge is further described within each of their respective sections.

Number of participants. The expert cohort consisted of 10 participants after the inclusion/exclusion was applied. The student cohort consists of 59 students from the Knowledge and Data Engineering (CS7IS1) module in Trinity College Dublin. The cohort was reduced to 48 participants after the inclusion/exclusion criteria was applied to the cohort.

4.2. Experiment Setup

The experiment setup for each cohort was identical with relation to the information and mapping provided prior to the experiment. However, the setup differed slightly with relation to the completion of the experiment and metrics used to measure the usability. The difference was due to the large sample size of the student cohort. It would not be feasible to arrange a video call with each participant in the cohort and transcribe/analyze their statements.

4.2.1. Experiment Preliminaries

The information sheet/informed consent was provided to all the participants prior to the experiment. Furthermore, the task sheet/mapping used during the experiment interaction were available on the framework. The information sheet and informed consent outlined the procedures and motivation for the experiment. These were provided to participants prior to the completion of the experiment, which would enable them to make an informed decision on whether to participate in the experiment. Furthermore, the participants could withdraw at any time prior to start of the experiment. These documents were reviewed and approved by the School's ethics committee within Trinity College Dublin. Following, the participants signing the informed consent document, a presentation was physically presented to the participants which outlines the motivation of the framework, the objectives of the study, its main contribution to research and an explanation of the mapping which will be used during the participants interaction. Noteworthy, the participants had no prior interaction with the framework before the experiment commenced.

The tasks which the participants completed during the experiment were designed to test the main functionality of the framework, which is the assessment and refinement of mappings. The process involves using a suite of quality metrics and related refinements while capturing information related to these processes in RDF format using MQV. The tasks outlined in the task sheet enable each of these characteristics to be evaluated. Twelve tasks were included within the task sheet. Tasks 1-3 involve the quality assessment of a mapping. Tasks 4-7 involve the selection and execution of refinements to remove quality issues within the mapping. Tasks 8-12 involve the examination of quality assessment information in MQV format and also visually.

The sample R2RML mapping which the participants used to interact with the framework was designed as a realistic use case. The use case of the sample mapping involves provenance information relating to datasets being uplifted to RDF, which can be easily understood by both cohorts as they both have knowledge about datasets. The use case is realistic as the PROV-O [18] documentation includes similar examples. PROV-O was chosen to represent the information as it is the W3C recommendation for capturing provenance information and is widely known. Furthermore, PROV-O includes the necessary data type restrictions to introduce a data type violation into the mapping. Three violations were introduced into the mapping. A quality violation relates to a quality issue within a mapping. The violations introduced into the mapping were chosen from the violations detected in Experiment 1, which indicates these violations occur in real-world mappings. Experiment 1 involved assessing the quality of 30 R2RML mappings, which were collected from semantic web research projects and students. The violations introduced allow the participants to evaluate the various refinement options available on the framework. These refinements involve semi-automatic refinements where the participant can enter a custom value, choose from a drop-down list of restricted values or select a suggested value. The three violations within the mapping are outlined below.

Usage of undefined property. `prov:values` predicate is undefined within PROV-O. The participants can choose a refinement which finds predicates within the same namespace or enter a new predicate within a text box. The predicate must be replaced by the participants with a valid defined predicate to resolve the violation.

Incorrect data type. The `xsd:time` assigned to the predicate object map with predicate `prov:generatedAtTime` is incorrect. The correct data type for the `prov:generatedAtTime` property is `xsd:dateTime`. The participants can choose from a refinement which suggests the correct data type or allows them to enter a data type in a text box. The participants must replace the invalid data type (`xsd:time`) within the mapping with the correct data type (`xsd:dateTime`) to resolve the violation.

Invalid language tag. The language tag "`en-GP`" is invalid. The participants can choose a refinement which is a drop-down menu with valid language tags. The language tag must be replaced by a valid English language tag to resolve the violation.

4.2.2. Experiment execution

Assistance was available to participants if they were unable to complete an experiment task. The assistance provided and completion of the experiment differed slightly for both cohorts due to the aforementioned reasons.

Completion of experiment. The participants in the expert cohort completed the experiment synchronously using zoom video conferencing platform while their think aloud statements were being recorded. The participants from the student cohort completed the experiment asynchronously by accessing the framework using provided login details. Furthermore, the cohort did not require the use of a video conferencing platform as the think-aloud protocol was not used because it would not be feasible to arrange a zoom meeting for each student and to transcribe/analyze their think-aloud statements.

Experiment Assistance. Each cohort could avail of assistance if they were unable to complete the experiment. The expert cohort was informed at the start of the experiment that assistance could be provided during the call if they are unable to complete a task. The student cohort was informed that assistance could be provided via email if they are unable to complete a task.

4.2.3. Data collected

Data was collected during the experiment from both cohorts in a quantitative and qualitative format.

Quantitative data. The Post-Study System Usability Questionnaire (PSSUQ) [19] was completed by both cohorts. The violation counts, which refers to the number of quality issues present after refined and time taken to complete the experiment was calculated for each cohort.

Qualitative data. The open comment section of the PSSUQ served as a basis of qualitative analysis for both cohorts. The main difference between the qualitative data collected was the use of the Think-aloud protocol [20]. The protocol was used to collect think-aloud statements, where participants verbalize their thoughts while completing the tasks. Only think-aloud statements were collected from the expert cohort as it would not be feasible to collect think-aloud statements from each participant in the student cohort.

4.2.4. Experiment metrics

The experiment metrics used include the usability questionnaire, deriving themes from the qualitative data, time taken to complete each task and count of quality issues remaining in the mapping after the completion of the experiment.

PSSUQ. The Post-Study System Usability Questionnaire (PSSUQ) [19] is widely used to measure users perceived satisfaction of a software system. The questionnaire provides the ability to do standardized comparison with other systems or evolutions of the system. The PSSUQ uses a 7-point Likert Scale where the lower score results in higher satisfaction. The second version of this questionnaire was used for the study, which includes 19 questions.

Thematic analysis. Thematic analysis [21] is designed to analyze qualitative data. The method involves deriving themes from the data. These themes are used to identify patterns within the data. Each theme consists of codes which relate to specific areas within a theme. The frequency of each code is calculated to identify the most commonly occurring themes. Thematic analysis is widely used within the qualitative research field.

Time per task. The time per task can be used as a comparative measure to determine if certain factors such as a worse PSSUQ score have a relationship with their timing.

Violation count. The violation count refers to the number of quality issues which have been resolved by the participant during the experiment. Three violations are present within the mapping provided to participants. The number of violations within the refined mapping generated was used to determine how effective the framework is at improving the quality of mappings.

4.3. Experiment Results

The data collected[5] was analyzed to identify usability issues within the framework. The analysis of data from both cohorts was completed separately and then the results of each cohort were compared. The comparison identifies patterns between both cohorts and determines which cohort found the framework more usable. **Table 1** shows the summary of results for the **expert cohort**.

Table 1: Summary of results for expert cohort

Time taken to complete experiment	***Mean time***	15.4 minutes
	Median time	12.8 minutes
PSSUQ mean metric score (lower number considered better)	***System usefulness (SysUse)***	1.69
	Information quality (InfoQual)	2.43
	Interface quality (IntQual)	2.75
	Overall usability (Overall)	2.11
Number of violations remaining after refinement complete (original mapping had 3 violations)	***0 violations (Best case)***	9 participants (90%)
	1 violation	1 participant (10%)
	2 violations	0 participants
	3 violations (Worst case)	0 participants

The analysis starts by discussing the PSSUQ results, followed by the other quantitative data. The qualitative data is discussed in parallel with the quantitative data. The provenance requirements heading does not directly relate to a metric, however, the heading is included to capture important qualitative data noted during the analysis. The PSSUQ scores have been compared against norms within a previous research study [19] as no previous scores exist for the framework.

Interface & Information quality. The interface quality relates to the quality of the items used to interact with the framework. The interface quality (**IntQual**) metric is the worst scoring metric within the PSSUQ with a mean score of 2.75. Furthermore, previous research [19] states that a score of 2.49 or less for the interface quality metric is sufficient, with the framework scoring lower, which indicates the interface needs to be improved. The qualitative data also indicates that the interface quality as the "Unaesthetic Interface" theme occurs commonly. The information quality relates to the quality of the information which is provided to users by the framework. Previous research indicates a

[5] Evaluation results at https://drive.google.com/drive/folders/1DvkYoKWwFAjke1MuBN76MoMHnR3CImFD?usp=sharing

mean score of 3.02 or less for the information quality metric is sufficient, with the framework scoring better than the threshold in the research study. The qualitative data indicates that additional information should be added to describe the refinements.

System usefulness and Overall usability. Only one participant required assistance during the completion of the tasks. The participant skipped a task within the task sheet, which resulted in them being redirected to the incorrect page on the framework. The best scoring metrics related to system usefulness (**SysUse**) and overall usability (**Overall**) with a mean of 1.69 and 2.11 respectively. Furthermore, these metrics both score more than 20% better than the thresholds within the research study. The metric scores and qualitative data indicate the participants found the system useful with an overall positive user experience.

Timing. The mean time for completing the experiment is 15.4 minutes with the fastest time being 11.05 minutes and the slowest time being 24.05 minutes. These results could indicate that not all experts could use the framework equally. Furthermore, noted during the experiment that some experts spent more time exploring the framework while others spent less time. The fastest tasks to complete were related to the assessment process. The participants took longer to choose and execute refinements. Furthermore, the slowest task related to examination of the patterns within the validation report. These results could indicate that the information provided relating to refinements could be improved to enable participants to select a refinement more easily. Furthermore, the layout of the validation report should be improved in future versions to improve the time it takes for participants to interpret the report.

Violation count. 90% of participants have 0 violations in the refined mapping, while 10% have 1 violation in the refined mapping. No participants have 3 violations in the refined mapping. The low violation count within the refined mapping indicates that the framework could be an effective tool for helping an expert user to identify and remove quality violations.

Provenance requirements. The provenance requirements of the framework refer to the quality assessment and refinement information provided by the validation bar chart and validation report. These areas relate to the information quality, however, these areas are more specifically highlighted within the qualitative analysis. The qualitative analysis of the participants' think-aloud statements and questionnaire open comments, indicate that the information provided by these items could be improved. **Table 2** shows a summary of quantitative data results for the **student cohort**. The time for completion, PSSUQ metric mean scores and violation count within refined mapping were calculated.

Table 2: Summary of results for student cohort

Time taken to complete experiment	***Mean time***	10.06 minutes
	Median time	9 minutes
PSSUQ mean metric score (lower number considered better)	***System usefulness (SysUse)*** ***Information quality (InfoQual)*** ***Interface quality (IntQual)*** ***Overall usability (Overall)***	2.34 2.42 2.8 2.42
Number of violations remaining after refinement complete (original mapping had 3 violations)	***0 violations (Best case)*** ***1 violation*** ***2 violations*** ***3 violations (Worst case)***	24 participants (50%) 10 participants (21%) 5 participants (10%) 9 participants (19%)

The analysis starts by discussing the PSSUQ results, followed by the other quantitative data. The qualitative data is discussed in parallel with the quantitative data.

Interface & Information quality. The interface quality relates to the quality of the items used to interact with the framework. The mean score for the interface quality (**IntQual**) metric is 2.8 which is the worst scoring metric. Furthermore, previous research states that a score of 2.49 or less for the interface quality metric is sufficient, however, the framework scores more than 10% worse than the threshold in the research study [19]. Furthermore, questions related to interface quality have the worst scoring third quartile (Q3), with a score of 4 and 3.75 respectively. The poor scoring of the interface quality within the PSSUQ results and the qualitative data indicates that the participants found the interface poor quality. In particular, the aesthetics of the framework needs to be improved in future versions of the framework. The information quality relates to the quality of the information which is provided to users by the framework. Previous research states that a score of 3.02 is sufficient for the information quality metric, with the framework scoring more than 20% better than the threshold. Moreover, the qualitative data was analysed to find data relating to the information displayed on the framework. These results indicate that the information provided by the framework is sufficient for the participants to complete the experiment, however, the qualitative analysis indicates that certain information provided by the framework needs to be improved in future versions. In particular, the information provided for the refinement needs to be improved. The PSSUQ results and qualitative data indicate that the information provided by the framework is sufficient, however, additional information should be added to the refinements to allow users to select and execute the refinements easier.

System usefulness and Overall usability. 48 out of the 59 (81%) students successfully completed the experiment. These results indicate that 81% of the students could successfully interact with the framework. Furthermore, previous research states that a mean score of 2.82 or less is sufficient for the overall usefulness metric and the framework scored 2.42, which is more than 15% better. Moreover, the qualitative data indicates these results also. These results indicate that the framework is fit for purpose and the participants are satisfied by the overall usability. Furthermore, the best scoring metric is the system usefulness with a mean of 2.34. The improvements previously mentioned could further improve the overall usability of the framework.

Timing. The mean time for the student cohort is 10.06 minutes. The maximum time is 23 minutes and the minimum time is 2 minutes. The minimum time of 2 minutes based on the experience of the researcher could indicate certain students were not careful when completing the experiment. The fastest tasks related to the assessment process. The slowest tasks related to the selection/executing of refinements and the examination of the patterns within the validation report. These results indicate that the participants struggled to select refinements and interpret the validation report. The additional information previously mentioned could improve the time taken to select and execute refinements. The patterns within the validation report could be simplified to allow the participants to interpret the report more easily.

Violation count. The original mapping contained 3 violations. 50% of participants have 0 violations. 70% have 1 or 0 violations. 30% have 2 or 3 violations. These results indicate that several students struggled to remove quality issues from the mapping. Several mappings contained violations such as including a data type named `admingeo:a` or `date:xsd`, which are not data types. Other examples of violations include a property named `aair:http://www.w3.org/r2rml#,` which is undefined. These are simple violations and could indicate students who gained more knowledge about semantic technologies during the module were able to remove quality issues easier, as 50% of them had no violations remaining.

Thematic analysis was completed following the six-step process [21] which includes data familiarization, generating initial codes, searching for themes, reviewing the themes, and producing the report. The most common themes and codes within the qualitative data that relate to improvements are shown in **Table 3**.

Table 3: Most common themes and codes discovered through thematic analysis

Themes	**Codes**
GUI Requirements. The layout and aesthetics of the framework are inadequate.	***Unaesthetic interface.*** The look and feel of the interface are inadequate.
	Unclear interface navigation. Guidance provided by the framework interface is hard to understand.
Clarify description and features. Overly complicated and ambiguous text displayed on the framework.	***Clarify text descriptions.*** Text descriptions need to be further described.
	Ambiguous refinement options. The refinement options for violations are not described adequately.

The GUI and textual descriptions need to be improved in the next version of the framework. The improvements will focus improving the aesthetics of the framework and adding additional text to describe different components.

4.4. Comparison of each Cohorts Results

The following section compares the main differences between the results of the **student** and **expert** cohorts. The results of the analysis of each cohort's data were compared based on the PSSUQ results, followed by the other quantitative data. The thematic analysis of the qualitative data and a summary of the overall analysis is then discussed.

Interface quality. The mean score for the interface quality (**IntQual**) metric for the expert cohort is 2.75 while the student cohort has a mean score of 2.8 which shows that the expert cohort rated higher satisfaction from the interface. These are the worst scoring metrics for both cohorts, which indicates that the interface needs to be improved for both cohorts. However, the expert cohort could have found the interface easier to use due to their previous experience in using semantic web related interfaces. Furthermore, previous research indicates that a mean score of 2.49 or less is sufficient for the interface quality metric, with both cohorts scoring worse than the threshold. Moreover, the "Unaesthetic Interface" code from the thematic analysis occurs frequently within the qualitative data of both cohorts, which further demonstrates that the aesthetics of the interface need to be improved for both cohorts.

Information quality. The mean score for the information quality (**InfoQual**) metric for the expert cohort is 2.43 while the student cohort has a mean score of 2.42 which shows that the student cohort rated slightly higher satisfaction from the information provided by the framework. Furthermore, 40% of experts rated the information quality a score of 3 or more, while only 20% of students rated the information quality metric with a score of 3 or more. The better scores for the information quality metric could indicate that the background knowledge of the expert cohort allowed them to notice information quality issues more easily. Furthermore, their background knowledge could result in them being more critical of the information displayed on the framework. However, previous research indicates that a mean score of 3.02 is sufficient for the information quality metric, with the information quality metric scoring better than the interface quality metric for each cohort. However, the information quality most frequently noted within the thematic analysis of the cohorts relates to the "Clarify text descriptions" and "Ambiguous refinement options" code. These results could indicate that simplified text and clearer refinement options could benefit both cohorts.

Analysis of each cohort's PSSUQ question scores. The worst scoring metric for both cohorts is the interface quality which indicates the interface should be improved for overall better user experience. The information quality metric scored similarly for both cohorts with a difference of less than 1% which could indicate better quality information is needed for both cohorts. Most median scores of the PSSUQ for the **expert** cohort have a median of 2 (10 out of 19 questions) and a spread below 2 points (5 out of 19 questions). The ease of use and (Q1) and efficiency (Q5) score the best. The questions relate to the error messages (Q9) and the aesthetics of the interface (Q16) score worse. All median scores of the PSSUQ for the **student** cohort have a median of 2. However, questions 16 and 17 have the worst third quartile (Q3), with a score of 4 and 3.75 respectively. These questions relate to the quality of the interface. These results indicate that the aesthetics of the interface should be improved for both cohorts in future versions of the framework.

Violation count. 90% of the expert cohort have 0 violations, while 70% of the student cohort have 0 or 1 violations in the refined mapping, which could indicate that the background knowledge of the expert cohort helped them to identify and remove the quality violations. Furthermore, no expert has 3 violations, while 10% of the student cohort had 3 violations. These results indicate that the effectiveness of the framework is influenced by the background knowledge. However, improvements previously mentioned could help students to identify and remove quality issues more easily.

Timing. The mean time for the expert cohort to complete the experiment is 15.4 minutes while the mean time for the student cohort is 10.06, which is about 5 minutes faster. The student and expert cohort have a median time of 13 and 12 minutes, respectively, which is only a difference of 1 minute. The majority of participants (Q3) completed the experiment in 20 minutes or less. However, the main difference is the maximum value. The student and expert cohort have a maximum time of 23 and 24 minutes, respectively, which could indicate that background knowledge does not influence the time taken to interact with the framework. Most of the task times of the **expert** cohort have a median less than 1 minute (7 out of 12). The other tasks have a median time of more than 1 minute but less than 1 minute and a half (3 out of 12). The longest tasks have a median time of more than 1 minute and a half (2 out of 12) which relate to choosing a refinement value and examining the validation report. Most of the task times of the **student** cohort have a median less than 1 minute (8 out of 12). The longest tasks have a median time of more than 1 minute and but less than a 1 minute and a half (4 out of 12) which related to choosing the refinement and examining the validation report. The task times indicate that both cohorts took the longest time to choose the refinement (Task 4, 5) and examine the patterns within the validation report (Task 12) . These areas could not be influenced by background knowledge and could be simplified in future versions. The reason for the student cohort completing the experiment faster than the expert cohort could be as a result of the expert cohort being more careful while completing each task. Furthermore, the expert cohort was using the think-aloud protocol, which could slow the completion of each task. Moreover, the usability of the framework could require a similar background knowledge, however, the effectiveness could be only influenced by the background knowledge.

Thematic analysis. The thematic analysis is used to discover emerging themes within the data which can be used to guide system improvements. Similar themes occurred within both cohorts, these themes include the "MQV Framework usability", the "GUI Requirements" and the "Clarify descriptions and features". The main areas highlighted within these themes in both cohorts are the poor aesthetics of the framework, unclear interface navigation and the textual descriptions of the refinement options. These are areas that should be improved for overall better usability by each cohort. The main difference between each cohort is the "Provenance usability" theme, which relates to the information provided by the validation report and bar chart. The theme was only noted within the expert cohort, where participants highlighted the patterns used to model the provenance information. The background knowledge of the experts in information modelling could have helped them to discover issues in the information modelling. These patterns should be improved to make them easier to understand by both cohorts.

5. Final Remarks

We would argue that the current approach of improving the quality of Linked Data datasets after the publication stage is more inefficient compared to improving the mapping artefacts that create the dataset in the first place. We introduced the MQV framework, designed to detect and address quality issues of mapping artefacts before they are executed. The framework generates machine-readable quality information represented in a domain specific vocabulary by executing metrics specifically designed for mappings. No previous research could be found within the state of the art where a mapping quality framework has been evaluated with a large sample size of users using standardized methods. The analysis of the results from an evaluation using a real-life use case mapping demonstrates the usability and effectiveness of the implementation. Next steps include the refinement of the framework based on the findings from the evaluation. Furthermore, the framework is currently being applied within a network management use case in Ericsson Software Technology.

Acknowledgements

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 (Grant # 13/RC/2106_P2) at Trinity College Dublin.

References

[1] Debattista J, Lange C, Auer S, Cortis D. Evaluating the quality of the LOD cloud: An empirical investigation. Semant Web. 2018 Mar;9:1–43.

[2] Junior AC, Debattista J, O'Sullivan D. Assessing the Quality of R2RML Mappings. In: Joint Proceedings of the International Workshop On Semantics For Transport and on Approaches for

Making Data Interoperable co-located with 15th Semantics Conference, Karlsruhe, Germany. CEUR-WS; 2019. (CEUR Workshop Proceedings; vol. 2447).

[3] Heyvaert P, De Meester B, Dimou A, Verborgh R. Rule-driven inconsistency resolution for knowledge graph generation rules. Semant Web. 2019;10(6).

[4] Randles A, O'Sullivan D. Assessing quality of R2RML mappings for OSi's Linked Open Data portal. 4th Int Work Geospatial Linked Data ESWC 2021. 2021;

[5] Das S, Sundara S, Cyganiak R. R2RML: RDB to RDF Mapping Language. W3C Recomm [Internet]. 2012; Available from: http://www.w3.org/TR/r2rml/

[6] Navarro-Gallinad A, Meehan A, O'Sullivan D. The semantic combining for exploration of environmental and disease data dashboard for clinician researchers. CEUR Workshop Proc. 2020;2778:73–85.

[7] Moreau B, Serrano-Alvarado P. Assessing the Quality of RDF Mappings with EvaMap. In: 17th Extended Semantic Web Conference (ESWC2020) [Internet]. 2020. p. 164–7. Available from: http://link.springer.com/10.1007/978-3-030-62327-2_28

[8] Debattista J, Auer S, Lange C. Luzzu-A Framework for Linked Data Quality Assessment. In: 2016 IEEE 10th International Conference on Semantic Computing, (ICSC). Institute of Electrical and Electronics Engineers Inc.; 2016. p. 124–31.

[9] Arndt D, Meester B De, Dimou A, Verborgh R, Mannens E. Using rule-based reasoning for RDF validation. In: International Joint Conference on Rules and Reasoning. 2017. p. 22–36.

[10] Dimou A, Kontokostas D, Freudenberg M, Verborgh R, Lehmann J, Mannens E, et al. Assessing and refining mappings to RDF to improve dataset quality. In: Lecture Notes in Computer Science. Springer Verlag; 2015. p. 133–49.

[11] Kontokostas D, Westphal P, Auer S, Hellmann S, Lehmann J, Cornelissen R, et al. Test-driven evaluation of Linked Data quality. In: WWW 2014 - Proceedings of the 23rd International Conference on World Wide Web. Association for Computing Machinery, Inc; 2014. p. 747–57.

[12] Randles A, Crotti Junior A, O'Sullivan D. Towards a vocabulary for mapping quality assessment. Proc 15th Int Work Ontol Matching 19th Int Semant Web Conf (ISWC),. 2020;

[13] Randles A, Junior AC, O'Sullivan D. A Vocabulary for Describing Mapping Quality Assessment, Refinement and Validation. In: 2021 IEEE 15th International Conference on Semantic Computing (ICSC). 2021. p. 425–30.

[14] Krech D. Rdflib: A python library for working with rdf. Online https://github com/RDFLib/rdflib. 2006;

[15] Randles A, Crotti Junior A, O'Sullivan D. A Framework for Assessing and Refining the Quality of R2RML mappings. In: Proceedings of the 22nd International Conference on Information Integration and Web-Based Applications & Services. New York, USA: ACM; 2020. (iiWAS2020).

[16] Poveda-Villalón M, Gómez-Pérez A, Suárez-Figueroa MC. OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation. Int J Semant Web Inf Syst. 2014;10(2):7–34.

[17] Zaveri A, Rula A, Maurino A, Pietrobon R, Lehmann J, Auer S, et al. Quality assessment methodologies for linked open data. Submitt to Semant Web J. 2013;1:1–5.

[18] Lebo T, Sahoo S, McGuinness D, Belhajjame K, Cheney J, Corsar D, et al. PROV-O: the prov ontology. w3c recommendation, 30 April 2013. World Wide Web Consort [Internet]. 2013; Available from: https://www.w3.org/TR/prov-o/

[19] Lewis JR. Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies. Int J Hum Comput Interact. 2002 Sep;14(3–4):463–88.

[20] Fonteyn ME, Kuipers B, Grobe SJ. A description of think aloud method and protocol analysis. Qual Health Res. 1993;3(4):430–41.

[21] Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: Striving to meet the trustworthiness criteria. Int J Qual methods. 2017.