Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

SCHOOL OF COMPUTER SCIENCE AND STATISTICS

CENTER FOR RESEARCH IN ARTIFICIAL INTELLIGENCE (CRT-AI)

ADAPT CENTER FOR DIGITAL CONTENT

KNOWLEDGE AND DATA ENGINEERING GROUP (KDEG)

Quality Improvement in the Mapping Process required for Linked Data publication

ALEX RANDLES

SUPERVISED BY PROF. DECLAN O’SULLIVAN

2023

A THESIS SUBMITTED TO THE

UNIVERSITY OF DUBLIN, TRINITY COLLEGE

IN FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE OF

DOCTOR OF PHILOSOPHY

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Signed: ______________________ Date: ______________________

Alex Randles

Acknowledgements

First and foremost, I would like to thank my excellent supervisor Prof. Declan O’Sullivan who provided invaluable support, guidance and motivation throughout my PhD journey. Thanks to Dr. Ademar Crotti Junior who provided great direction at the beginning of my journey. Thanks also goes to my colleagues at the ADAPT centre, especially Albert Navarro Gallinad, who made my journey an extremely enjoyable and unforgettable experience. Finally, I would like to thank my family and friends for their continuous support throughout my PhD.

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. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant No. 13/RC/2106.

Abstract

This thesis presents a quality improvement approach named the **M**apping **Q**uality **I**mprovement (**MQI**) Framework designed to improve and maintain quality in the publication process involved in the creation of linked data.

Linked data is described as a set of best practices used for publishing and interlinking data on the web. A linked data dataset is structured information encoded using the Resource Description Framework (RDF) and provides information, which is interoperable, extensible, and machine-readable. Resources are identified by and linked with other datasets using HTTP URIs which enables the accessibility of resources through the HTTP protocol. Statements within RDF are referred to as triples (subject-predicate-object), which represent the nodes and edges within a data graph. Linked data is referred to as one of the most efficient and effective knowledge integration and discovery approach. The Semantic web represents the web of data which is an extension of the current web where data is stored in machine-readable and standardized formats such as the World Wide Web Consortium linked data recommendation. Transforming heterogeneous data into linked data representation is an essential prerequisite for evolving the semantic web and requires the definition of declarative uplift mappings. *"Uplift"* mappings are used in the publication process in order to define rules for transforming data from non-RDF format to RDF format. However, defining these mappings is a complex and often error prone task, often resulting in propagation of quality issues into the resulting linked data dataset. In addition, linked data is highly dynamic in nature with frequent changes, often resulting in alignment issues between the linked data, the mappings and the underlying data sources. Oftentimes, the burden of quality assessment is on third parties after a linked data dataset has been published, greatly decreasing the trustworthiness of data on the semantic web.

A literature review was conducted in order to define requirements for the MQI framework, which was designed to resolve limitations discovered in the state of the art. The literature review consisted of two phases focused on approaches to support the creation and maintenance of declarative mappings used during the publication process of linked data. The review indicated a lack of approaches to support the relevant processes. Therefore, the MQI framework was designed to support users by providing a suite of quality metrics in order to identify diverse issues early in the linked data publication process, specifically the mapping stage. The framework guides the process of removing detected quality issues in the mappings by suggesting providing semi-automatic refinements. In addition, the framework supports preservation of alignment of the uplift mappings with underlying data sources by detecting source data changes after publication. Importantly, the relevant quality process information captured by the framework is itself represented as semantic data, which enables its association with resulting linked data datasets in order to improve downstream maintenance and reuse.

The proposed approach was evaluated through five experiments and one application study categorized by five aspects: accuracy, usability, understanding, effectiveness and validation. The application study was designed to evaluate the approach when applied in real world settings. The experiments involved over one hundred participants with varying background knowledge, including knowledge engineer students, uplift mapping specialists and ontology design specialists. The varying background knowledge aided retrieval of diverse insightful feedback. The accuracy of the framework was tested by detecting quality issues in real world mappings supplied by others from their projects. The first usability experiment tested the effectiveness of the MQI framework in supporting quality assessment and refinement of uplift mappings by users. The second usability experiment tested the understanding by users of changes detected by the framework in source data of mappings. Finally, ontology design specialist validated the ontologies that underpin the MQI framework processing. Overall the evaluations indicated that the MQI framework provides effective and understandable information to users, facilitates the creation and maintenance of high-quality uplift mappings with downstream impact on the quality of the resultant linked data dataset. Finally, the validation of both ontologies indicated that they are of sufficient design quality.

The research described in this thesis resulted in one major contribution and three minor contributions. The major contribution is the design and development of the **MQI Framework**. The first minor contribution is the **M**apping **Q**uality **I**mprovement **O**ntology (**MQIO**) designed to represent mapping quality assessment, refinement and validation information. The second minor contribution is the **O**ntology for **S**ource **C**hange **D**etection (**OSCD**) designed to represent information about changes in source data of mappings and their alignment with associated mappings. Finally, the third minor contribution is the **evaluation results** from the 5 experiments and 1 application study conducted in order to evaluate the proposed approach.

Table of Contents

List of Figures

List of Tables

List of Listings

List of Abbreviations

Abbreviations	Definition
LD	Linked Data
RDF	Resource Description Framework
URI	Uniform Resource Identifier
IRI	Internationalized Resource Identifier
TURTLE	Terse RDF Triple Language
OWL	Web Ontology Language
RDB	Relational Database
RDFS	RDF Schema
SHACL	Shapes Constraint Language
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
R2RML	RDB to RDF Mapping Language
MQIO	Mapping Quality Improvement Ontology
OSCD	Ontology for Source Change Detection
PROV-O	Provenance Ontology
DQV	Data Quality Vocabulary
FOAF	Friend of a Friend Ontology
LODE	Linking Open Descriptions of Events Ontology
RML	RDF Mapping Language
YARRRML	Human-Readable Mapping Language
CSV	Comma-Separated Values
XML	Extensible Markup Language
JSON	JavaScript Object Notation
SoA	State of the Art
PSSUQ	Post-Study System Usability Questionnaire
W3C	World Wide Web Consortium
ISO	International Organization for Standardization
IEC	International Electrotechnical Commission
FAIR	Findable, Accessible, Interoperable and Reusable
CLI	Command Line Interface
GUI	Graphical User Interface

Chapter 1: Introduction

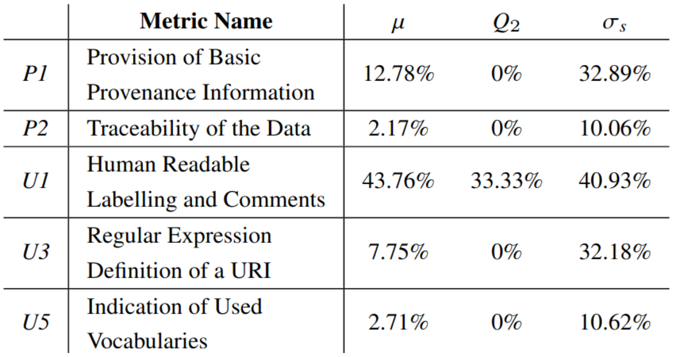
1.1 Motivation

Linked data (LD) is described as a set of best practices used for publishing and interlinking data on the web. A LD dataset is structured information encoded using the Resource Description Framework (RDF) [34] and provides information, which is interoperable, extensible, and machine-readable. Resources are identified by and linked with other datasets using Hypertext Transfer Protocol (HTTP) Uniform Resource Identifiers (URIs) [13] which enables the accessibility of resources through the HTTP protocol. Statements within RDF are referred to as triples (subject-predicate-object), which represent the nodes and edges within a data graph. LD is referred to as one of the most efficient and effective knowledge integration and discovery approaches [145]. The Semantic web (SW) represents the web of data which is an extension of the current web where data is stored in machine-readable and standardized formats such as the World Wide Web Consortium (W3C) linked data recommendation[1]. Oftentimes, LD datasets are published on the Linked Open Data (LOD) cloud[2] which provides a collaborative platform where datasets can be openly published according to LD principles. The number of datasets on the LOD cloud has grown significantly, from 203 (2010) to over 1200 (2022) and contains over 62 billion RDF triples [153]. Datasets published contain information related to various knowledge domains which may result in an overlap of knowledge.

Previous research [38] which evaluated the quality of datasets on the LOD cloud in 2018 demonstrated highly varying level of quality. The results of the evaluation included an aggregated conformance of 60% of the datasets with respect to the 27 implemented metrics which covered quality aspects such as accessibility, licensing, provenance, amongst others. In particular, the metrics that related to provenance and licensing scored worse in terms of quality conformance. Less than half of the datasets had a valid access point (42%), a valid license (40%) or human readable labels and comments (43%). While slightly over half of the datasets contained undefined classes or properties (54%) and incorrect domain or range types (60%). **Table 1** presents an overview of the results of metrics in the contextual category of quality, which relates to the trustworthiness and understanding of the data. The results show poor conformance for the contextual category.

[1] https://www.w3.org/wiki/LinkedData
[2] https://lod-cloud.net/

Table 1: Mean value (μ), Median value (Q_2), and Standard deviation (σ_s) of metrics in Contextual Category



	Metric Name	μ	Q_2	σ_s
P1	Provision of Basic Provenance Information	12.78%	0%	32.89%
P2	Traceability of the Data	2.17%	0%	10.06%
U1	Human Readable Labelling and Comments	43.76%	33.33%	40.93%
U3	Regular Expression Definition of a URI	7.75%	0%	32.18%
U5	Indication of Used Vocabularies	2.71%	0%	10.62%

In addition, other research [6] in 2015 has assessed the quality of the metadata of datasets on the LOD cloud which is an important aspect for consumers to understand and use the datasets. The results indicated that the quality of metadata was extremely poor with 80% of the resources with missing or undefined provenance information. More recent research (2022) [4,70,128,129,152] and (2021) [23,46,144,148] has demonstrated the quality of LD datasets published remains poor. Overall, it can be concluded from these studies and other previous work [66,67] that quality of LD being published has traditionally been poor. Furthermore, limited efforts [150] have been in place to standardize how quality tracking and assurance should be implemented for LD. No consensus exists on how the data quality dimensions and metrics should be defined. However, a common method used to detect relevant quality issues in LD datasets is the Shapes Constraint Language (SHACL) [80]. SHACL is a popular W3C recommendation designed to validate RDF graphs against a set of constraints. It is commonly used to assess dataset quality, however, it can also be applied to assess mappings represented in RDF format.

While the LOD cloud provides a large amount of knowledge, only a subset of knowledge contained on the web is available for consumption [38]. Oftentimes, data is stored in silos on the web in heterogenous formats such as Comma-Separated Values (CSV), JavaScript Object Notation (JSON), Extensible Markup Language (XML) and relational databases [133]. Therefore, the conversion of the data into RDF format is an essential prerequisite for evolving the web of data [79]. Transforming non-RDF data to RDF data typically requires the usage of declarative mappings which contain rules designed to convert the source data into an RDF representation. For instance, a rule could state that a column within a relational database is the object generated within a triple pattern. *'Uplift mappings'* are concerned with the transformation from non-RDF to RDF data [29]. While *'semantic mappings'* define rules which are used to link semantically similar concepts within already existing LD datasets [130]. Uplift mappings are responsible for generating a large amount of datasets on the LOD cloud [141]. Inherently, mapping quality will significantly influence the quality of LD datasets published [75].

The rules in the declarative languages for mapping are syntactically heavy and not intuitive however [75]. The creation of declarative mappings involves multiple stakeholders who define requirements and create rules. which can be defined using one of a number of languages, such as RDB to RDF Mapping Language (R2RML) [35] and RDF Mapping Language (RML) [44]. Furthermore, a high level of domain knowledge is required to create high-quality mappings [32]. Previous research on mapping quality [42,75,95,97,101] has demonstrated the difficulties of producing high-quality mappings due to their complexity, and often the process is error prone [75]. Mapping quality issues can result in quality issues exponentially appearing within the resulting dataset, thus resulting in a huge decrease of data quality [75]. Quality in this context is defined as the *"fitness for use"* of the data [38].

Metrics and dimensions [38,150] in categories such as Contextual (CT), Representational (RP), Accessibility (AC) and Intrinsic (IR) will be directly impacted by the quality of declarative mappings. In total, 52 out of 64 (82%) of quality metrics defined in previous work [38,150] on LD quality are directly impacted by the quality of associated uplift mappings. The LD quality metrics **impacted** by the quality of mappings are in 14 quality dimensions, which include *Completeness (IR), Conciseness (IR), Data Consistency (IR), Interoperability (RP), Interpretability (RP), Interlinking (AC), Licensing (AC), Representational conciseness (RP), Relevancy (CT), Semantic accuracy (IR), Syntactic validity (IR), Timeliness (CT), Trustworthiness (CT) and Understandability (CT).* Metrics **not impacted** by the quality of the mapping are in 4 quality dimensions, which include *Performance (AC), Availability (AC), Security (AC) and Versatility (RP).* For instance, Trustworthiness and Understandability of a LD dataset as presented In **Table 2** (see section 1.3**)**, will be positively impacted by a focus on quality of mapping metadata, as additional provenance related to the data is defined by the mapping engineer who created a mapping to uplift the data. Furthermore, metrics related to the Interoperability dimension (RP) of quality of LD will be positively impacted as undefined classes/properties and blank nodes in the definitions in the uplift mappings are addressed. Moreover, LD quality metrics related to the Data Consistency dimension (IR) including usage of deprecated classes/properties, usage of incorrect range/domain and disjoint classes will be positively impacted as the classes and properties used in the dataset originate in uplift mapping definitions. However, dimensions which will not be impacted by improvement of the quality of a mapping include Performance (AC), as it is determined by the host of the LD access point. In addition, Versatility (RP) will not be necessarily improved as the LD data generated by the mapping depends directly on the quality of the source data. Another benefit of improving quality at the mapping stage of the LD publication pipeline rather than just improving quality of the resultant LD dataset, relates to computation and memory intensity that is required, which is far less for a mapping quality checking compared to an entire dataset [42,75]. In addition, removing quality issues in LD datasets needs to be completed each time a new version is generated [42,75].

Most of existing approaches [36,81,96,123,150] in LD quality assessment focus on the resulting dataset and are independent of the mapping process, and typically are executed by third parties rather than the data publishers [75]. A separate important issue to note is that those approaches [42,75,95,97,101] that are designed to assess and refine the quality of mappings use different methods to represent the captured quality information. The quality process related information in these approaches is either defined in human-readable representation [65,97] or expressed in ontologies designed for dataset quality [42,75] and provenance [101]. For instance, one approach [101] represents the information using the W3C recommendation Provenance Ontology (PROV-O) [84], which was designed to represent general provenance information. However, the approach identified a limitation of using

PROV-O for this purposes and states that *“Richer semantics are needed to describe the results (e.g., violations) of quality assessment.”*.

The quality of mappings has been acknowledged as an extremely important factor for creating high-quality LD datasets [42,65,75]. However, in addition, it is important to maintain the quality of mappings after the publication of the resulting dataset to prevent a decrease in quality [141] and provide better reuse and discovery [3]. Oftentimes, the alignment between LD datasets and respective uplift mappings can become misaligned due to the highly dynamic nature of the source data [43,141], with resources and usage of vocabularies continuously being changed in an attempt to improve the quality of the data [76,109,138]. Alignment issues between mappings and source data used to create LD datasets, therefore, influences the quality of mappings and resulting LD [43,141]. In addition, changes in the source data should be promptly propagated into the resulting data to prevent a decrease in the *“timeliness”* [141] of the data, defined as the *“a comparison of the date the annotation was updated with the consumer’s requirement”* [150]. The dimension has been described as one of the most important aspects of LD quality [18]. Poor conformance to the dimension results in an inaccurate representation of the underlying data sources being provided to consumers [141].

Despite the issue being identified in the State of the Art (SoA) over a decade ago, no standard as yet exists to address the dynamics of LD [138]. Existing approaches [76,104,109,131] proposed to address the dynamics of LD are predominantly designed to capture changes which occur in resources and interlinks, however, one approach [141] targets the dynamics of source data used during the publication process. The approach is limited to relational data and focuses on R2RML mappings, and cannot target heterogenous formats, such as XML, CSV and JSON. Most of the approaches [76,104,109,131] provide the ability to automatically monitor LD in order to detect new changes periodically, therefore, providing regular updates on relevant changes to data maintainers. Another feature provided by some of the approaches [104,109] is a notification mechanism which enables maintainers to be informed of when resources and interlinks are changed in a timely manner. An approach [109] has created an ontology to represent changes in resources of LD datasets, named the DSNotify EventSet Vocabulary. Other approaches represent the changes as changesets [141] and SPARQL Protocol and RDF Query Language (SPARQL) [104] query results. Hereinafter, the term “**SPARQL**” refers to SPARQL 1.1 Query Language [104].

The dynamics of LD and related mapping artefacts will directly impact the metrics in the timeliness quality dimension (CT), which measures how up to date the LD is based on underlying data sources. In addition, changes in data sources can influence compatibility with mappings [43,141]. For instance, a column referenced in a mapping no longer exists in the respective source data, therefore, the mapping should be updated to reflect the change [43].

Capturing quality information related to the publication process of LD datasets is extremely important [42,65,75] as it provides indications on how suitable published data is for the use cases of consumers [38,87]. In addition, research [3] has shown that capturing information about quality of mappings will provide an indication to mapping engineers of when the quality is sufficient for execution. Furthermore, it is expected that the information will benefit the discovery, maintenance and reuse of mappings [3]. Moreover, captured information about the quality of a mapping can be linked with the generated datasets, providing an extensive lineage of quality-oriented provenance, which will benefit the maintenance and reuse of the data [42,75,95,97,101]. However, difficulties exist in processing such information at the moment, as most of the time it is stored in unstructured formats, such as HTML markup, which focuses on presentation of data rather than understanding [136]. Therefore, it is difficult for software agents to understand the meaning of the information, therefore, greatly limiting interoperability [9,45,69,136]. In addition, the linking of such mapping quality information is vastly limited as associated software agents cannot identify similarities between resources [9,45,136]. A common solution to incorporate semantics into data is to create ontologies in order to address such limitations. Ontologies are the core of the semantic web [9] and are defined as "*a formal explicit specification of a shared conceptualization*", which consists of concepts, instances, relations and axioms used to provide an agreed meaning of various aspects of a knowledge domain [136]. In addition, ontologies enhance the functionality of the web in many ways [9], allowing human agents and software agents to exchange knowledge and fulfill collaboration goals [45,136]. Existing approaches [19,29,34,51,57] on LD quality recommend capturing relevant information in a machine-readable format, such as RDF, which allows easier processing and linking of relevant information [77]. Therefore, using ontologies implemented in Web Ontology Language (OWL2) [92] provides an ideal solution for representing the relevant quality information associated with a mapping. In this context, ontologies provide a method to define shared terminology of data captured during the various activities involved in the publication process of LD. Therefore, enabling data publishers and software agents to fulfill collaboration goals involving processing meaningful provenance in order to improve and maintain a high level of quality during publication. Furthermore, representing quality information as OWL2 ontologies will enable for reuse within other semantic frameworks and tools, such as editing and visualization tools involved in the publication process [97]. Moreover, it enables the linking of relevant information with the resulting dataset, providing an extensive lineage of quality-oriented provenance, which will benefit maintenance and reuse of the data [68]. Importantly, mapping quality can be automatically improved by software agents who can easily process machine-readable quality information and add/delete triples or suggest actions to publishers [2,3]. Ontologies are extensible, which enables additional concepts to be integrated as they emerge from various activities in the publication process. In addition, ontologies support powerful logical inference on facts through axiomatization [9,45,136], which could be used to discover relationships between different aspects that impact LD quality. Finally, ontologies are effective methods for answering complicated questions [9], such as an important question in this context, *"Is this mapping sufficient quality for execution?"*. While ontologies exist to represent quality information [1,18] and provenance [84] related to LD datasets, however, an ontology to represent quality of mappings was not found in the state of the art.

1.2 Research Question

The research question addressed in this thesis is:

"To what extent can the detection of declarative mapping quality issues and source data changes, facilitate the creation and maintenance of high-quality Linked Data (LD) datasets?"

Terms used within the research question are defined as follows:

- **Detection:** The action or process of identifying the presence of something concealed [41].
- **Facilitate:** Make (an action or process) easy or easier [41].
- **High quality:** High level of conformance to LD quality metrics impacted by mapping quality as outlined in **Table 2** (see next section).
- **Quality:** *"Fitness for use"* for a specific application or use case [38,150].

1.3 Research Objectives

The following research objectives were identified in order to address the research question:

- **RO1:** Establish the State-of-the-Art of existing approaches which are designed to:
  a) Improve quality of mappings in the LD domain.
  b) Address dynamics of LD datasets.
- **RO2:** Develop the following:
  a) OWL2 ontology to represent LD information related to mapping quality.
  b) An approach to enable the identification and removal of issues related to quality of uplift mappings (see **Table 2** below).
- **RO3:** Develop the following:
  a) OWL2 ontology to represent changes to source data associated with the LD dataset.
  b) An approach to preserve alignment between source data changes and the respective uplift mappings.
- **RO4:** Implement and evaluate the approaches defined in **RO2** and **RO3**.

Table 2 presents quality metrics and dimensions commonly used to assess the quality of LD datasets. The first four columns of the table has been retrieved from a prominent and heavily cited survey [150] discussing LD quality metrics, dimensions and categories. The table has been added to by the author of this thesis (the fifth column) to include the level of potential impact that the quality of declarative mappings will have upon the particular LD quality metric, by proposing a mapping impact (**MI**) for each quality metric. For the MI column, the following keywords and codes have been defined by the author of this thesis in order to indicate the level of impact that the quality of a mapping is expected to have on the LD quality metric presented.

- *Keywords related to LD quality metrics potentially **impacted** by quality of the mapping:*
  - **Classes and Properties** (**CP**): Classes and properties measured by the metric originate in the mapping definitions.
  - **Logical Inconsistencies (LI):** Incorrect semantics in the data measured by the metric are a result of poor mapping design decisions.
  - **Maintenance of Mappings (MM):** Maintenance and re-execution of mappings impacts the metric.
- *Keyword related to LD quality metrics potentially **not impacted** by the quality of the mapping:*
  - **Source Data (SD):** The quality aspect is impacted by the quality of the source data.
  - **Hosting (HS):** The quality aspect is out of scope when it comes to quality of the mapping, such as the security and performance of the server where the data is hosted.

Table 2: List of common LD quality metrics and impact of mapping quality on them (Derived from [150])

Dimension	**Abr**	**Metric**	**Description**	**MI**
Availability	AV1	accessibility of the SPARQL end-point and the server	checking whether the server responds to a SPARQL query [48]	**HS**
	AV2	accessibility of the RDF dumps	checking whether an RDF dump is provided and can be downloaded [48]	**HS**
	AV3	dereferenceability of the URI	checking (i) for dead or broken links i.e. when an HTTP-GET request is sent, the status code 404 Not Found is not be re- turned (ii) that useful data (particularly RDF) is returned upon lookup of a URI, (iii) for changes in the URI i.e the compliance with the recommended way of implementing redirections using the status code 303 See Other [48,67]	**HS**
	AV4	no misreported content types	detect whether the HTTP response contains the header field stating the appropriate content type of the returned file e.g. application/rdf+xml [67]	**HS**
	AV5	dereferenced forward-links	dereferenceability of all forward links: all available triples where the local URI is mentioned in the subject (i.e. the description of the resource) [68]	**HS**
Completeness	CM1	schema completeness	$\frac{\text{no.of classes and properties represented}}{\text{total no.of classes and properties}}$ [51,67]	**CP**
	CM2	property completeness	(i) $\frac{\text{no.of values represented for a specific property}}{\text{total no.of values for a specific property}}$ [47,51] (ii) exploiting statistical distributions of properties and types to characterize the property and then detect completeness [105]	**CP**
Conciseness	CN1	high intensional conciseness	$\frac{\text{no.of unique properties/classes of a dataset}}{\text{total no.of properties/classes in a target schema}}$ [96]	**CP**
	CN2	high extensional conciseness	(i) $\frac{\text{no.of unique objects of a dataset}}{\text{total number of objects representations in the dataset}}$ [96] (ii) $1 - \frac{\text{total no.of instances that violate the uniqueness rule}}{\text{total no.of relevant instances}}$ [51,81,85]	**CP**
	CN3	usage of unambiguous annotations/labels	$1 - \frac{\text{no.of ambiguous instances}}{\text{no.of instances contained in the semantic metadata set}}$ [85,123]	**CP**
Data Consistency	DC1	no use of entities as members of disjoint classes	$\frac{\text{no.of entities described as members of disjoint classes}}{\text{total no.of entities described in the dataset}}$[48,67,81]	**CP**
	DC2	no misplaced classes or properties	using entailment rules that indicate the position of a term in a triple [47,67]	**CP**

	DC3	no misuse of `owl:DatatypeProperty` or `owl:ObjectProperty`	detection of misuse of `owl:DatatypeProperty` or `owl:ObjectProperty` through the ontology maintainer [67]	**CP**
	DC4	members of `owl:DeprecatedClass` or `owl:DeprecatedProperty` not used	detection of use of members of `owl:DeprecatedClass` or `owl:DeprecatedProperty` through the ontology maintainer or by specifying manual mappings from deprecated terms to compatible terms [47,67]	**CP**
	DC5	valid usage of inverse functional properties	(i) by checking the uniqueness and validity of the inverse-functional values [67], (ii) by defining a SPARQL query as a constraint [81]	**CP**
	DC6	absence of ontology hijacking	detection of the re-definition by third parties of external classes/properties such that reasoning over data using those external terms is not affected [67]	**CP**
	DC7	no negative dependencies/correlation among properties	using association rules [16]	**CP**
	DC8	no inconsistencies in spatial data	through semantic and geometric constraints [98]	**LI**
	DC9	correct domain and range definition	the attribution of a resource's property (with a certain value) is only valid if the resource (domain), value (range) or literal value (rdfs ranged) is of a certain type – detected by use of SPARQL queries as a constraint [81]	**LI**
	DC10	no inconsistent values	detection by the generation of a particular set of schema axioms for all properties in a dataset and the manual verification of these axioms [149]	**LI**
Interoperability	IO1	re-use of existing terms	detection of whether existing terms from all relevant vocabularies for that particular domain have been reused [68]	**CP**
	IO2	re-use of existing vocabularies	usage of relevant vocabularies for that particular domain [48]	**CP**
Interpretability	IN1	use of self-descriptive formats	identifying objects and terms used to define these objects with globally unique identifiers [47]	**CP**
	IN2	detecting the interpretability of data	detecting the use of appropriate language, symbols, units, datatypes and clear definitions [48,108]	**CP**
	IN3	invalid usage of undefined classes and properties	detection of invalid usage of undefined classes and properties (i.e. those without any formal definition) [67]	**CP**
	IN4	no misinterpretation of missing values	detecting the use of blank nodes [68]	**CP**
Interlinking	IL1	detection of good quality interlinks	(i) detection of (a) interlinking degree, (b) clustering coefficient, (c) centrality, (d) open sameAs chains and I descriptionrichness through sameAs by using network measures [60], (ii)via crowdsourcing [1,149]	**CP**
	IL2	existence of links to external data providers	detection of the existence and usage of external URIs (e.g. using `owl:sameAs`links) [68]	**CP**
	IL3	dereferenced back-links	detection of all local in-links or back-links: all triples from a dataset that have the resource's URI as the object [68]	**CP**
Licensing	LI1	machine-readable indication of a license	detection of the indication of a license in the VoID description or in the dataset itself [48,68]	**CP**
	LI2	human-readable indication of a license	detection of a license in the documentation of the dataset [48,68]	**CP**
	LI3	specifying the correct license	detection of whether the dataset is attributed under the same license as the original [48]	**CP**
Performance	PE1	usage of slash-URIs	checking for usage of slash-URIs where large amounts of datais provided [48]	**CP**
	PE2	low latency	(minimum) delay between submission of a request by the userand reception of the response from the system [48]	**HS**
	PE3	high throughput	(maximum) no. of answered HTTP-requests per second [48]	**HS**
	PE4	scalability of a data source	detection of whether the time to answer an amount of ten requests divided by ten is not longer than the time it takes to answer one request [48]	**HS**
Representational conciseness	RC1	keeping URIs short	detection of long URIs or those that contain query parameters [47,68]	**CP**
	RC2	no use of prolix RDF features	detection of RDF primitives i.e. RDF reification, RDF containers and RDF collections [47,68]	**CP**
Relevancy	RV1	relevant terms within meta-information attributes	obtaining relevant data by (i) ranking (a numerical valuesimilar to PageRank), which determines the centralityof RDF documents and statements [17], (ii) via crowd- sourcing [1,149]	**CP**

	RV2	coverage	measuring the coverage (i.e. number of entities described in a dataset) and level of detail (i.e. number of properties) in a dataset to ensure that the data retrieved is appropriate for the task at hand [48]	**CP**
Semantic accuracy	SA1	no outliers	by (i) using distance-based, deviation-based and distribution-based methods [12,47], (ii) using the statistical distributions of a certain type to assess the statement's correctness [105]	**LI**
	SA2	no inaccurate values	by (i) using functional dependencies between the values of two or more different properties [51], (ii) comparison be- tween two literal values of a resource [81], (iii) via crowd- sourcing [1,149]	**LI**
	SA3	no inaccurate annotations, labellings or classifications	$1 - \frac{\text{inaccurate instances}}{\text{total no.of instances}} * \frac{\text{balanced distance metric}}{\text{total no.of instances}}$ [85]	**LI**
	SA4	no misuse of properties	by using profiling statistics, which support the detection of discordant values or misused properties and facilitate to find valid formats for specific properties [16]	**LI**
	SA5	detection of valid rules	ratio of the number of semantically valid rules to the number of nontrivial rules [28]	**LI**
Security	SC1	usage of digital signatures	by signing a document containing an RDF serialization, a SPARQL result set or signing an RDF graph [24,48]	**HS**
	SC2	authenticity of the dataset	verifying authenticity of the dataset based on a provenance vocabulary such as author and his contributors, the publisher of the data and its sources (if present in the dataset) [48]	**HS**
Syntactic validity	SV1	no syntax errors of the documents	detecting syntax errors using (i) validators [48,67] (ii) via crowdsourcing [1,149]	**CP**
	SV2	syntactically accurate values	by (i) use of explicit definition of the allowed values for a datatype, (ii) syntactic rules [51] (iii) detecting whether the data conforms to the specific RDF pattern and that the "types" are defined for specific resources [81] (iv) use of different outlier techniques and clustering for detecting wrong values [146]	**CP**
	SV3	no malformed datatype literals	detection of ill-typed literals, which do not abide by the lexical syntax for their respective datatype that can occur if a value is (i) malformed, (ii) is a member of an incompatible datatype [47,67]	**CP**
Timeliness	TI1	freshness of datasets based on currency and volatility	$max\{0, 1 - \frac{currency}{volatility}\}$ [63] which gives a value in a continuous scale from 0 to 1, where score of 1 implies that the data is timely and 0 means it is completely outdated thus unacceptable. In the formula, volatility is the length of time the data remains valid [51] and currency is the age of the data when delivered to the user [47,96,124]	**MM**
	TI2	freshness of datasets based on their data source	detecting freshness of datasets based on their data source by measuring the distance between last modified time of the data source and last modified time of the dataset [14,51]	**MM**
Trustworthiness	TR1	trustworthiness of statements	computing statement trust values based on: (i) provenance information which can be either unknown or a value in the interval [−1,1] where 1: absolute belief, −1: absolute disbelief and 0: lack of belief/disbelief (ii) opinion-based method, which use trust annotations made by several individuals [47,62] (iii) provenance information and trust annotations in Semantic Web-based social-networks [57] (iv) annotating triples with provenance data and usage of provenance history to evaluate the trustworthiness of facts [40]	**CP**
	TR2	trustworthiness through reasoning	using annotations for data to encode two facets of information [17]: (i) blacklists (indicates that the referent data is known to be harmful) (ii) authority (a boolean value which uses the Linked Data principles to conservatively determine whether or not information can be trusted)	**CP**
	TR3	trustworthiness of statements, datasets and rules	using trust ontologies that assigns trust values that can be transferred from known to unknown data using: (i) content-based methods (from content or rules) and (ii) metadata-based methods (based on reputation assignments, user ratings, and provenance, rather than the content itself) [74]	**CP**
	TR4	trustworthiness of a resource	computing trust values between two entities through a path by using: (i) a propagation algorithm based on statistical techniques (ii) in case there are several paths, trust values from all paths are aggregated based on a weighting mechanism [127]	**CP**

	TR5	trustworthiness of the information provider	computing trustworthiness of the information provider by: (i) construction of decision networks informed by provenance graphs [52] (ii) checking whether the provider/contributor is contained in a list of trustedproviders [12] (iii) indicating the level of trust for the publisher on ascale of 1−9 [55,56]	**CP**
	TR6	trustworthiness of information provided (content trust)	checking content trust based on associations (e.g. any- thing having a relationship to a resource such as author of the dataset) that transfers trust from content to resources [55]	**CP**
	TR7	reputation of the dataset	assignment of explicit trust ratings to the dataset by humans or analyzing external links or page ranks [96]	**CP**
Understandability	UT1	human-readable labelling of classes, properties and entities as well as presence of metadata	detection of human-readable labelling of classes, properties and entities as well as indication of metadata (e.g.name, description, website) of a dataset [47,48,68]	**CP**
	UT2	indication of one or more exemplary URIs	detect whether the pattern of the URIs is provided [48]	**CP**
	UT3	indication of a regular expression thatmatches the URIs of a dataset	detect whether a regular expression that matches the URIs is present [48]	**CP**
Versatility	VT1	provision of the data in different serialization formats	checking whether data is available in different serialization formats [48]	**HS**
	VT2	provision of the data in various languages	checking whether data is available in different languages [48]	**SD**

As can be seen, 52/64 (82%) of the LD quality metrics presented in **Table 2** are potentially directly impacted by the quality of declarative mappings. Most of the **impacted** metrics are potentially as result of the presence of classes and properties (**CP**) defined in a mapping (81%), while others (15%) are potentially as a result of incorrect semantics (**LI**) in the definitions of a mapping and associated maintenance (**MM**) of the artefacts (4%). The other 12/64 metrics (18%) potentially **not impacted** by quality of a mapping are as a result of the related aspect being out of scope (**HS**) (91%) or related to the quality of the source data (**SD**) (9%). In summary, these results indicate that a large proportion of LD quality metrics in scope of the publication process is influenced by the quality of the associated declarative uplift mappings.

1.4 Research Methodology

The research methodology involved the identification of relevant information from the state of the art which was used to guide the design and development of the proposed approach. The analysis of the state of the art indicated a requirement for one technical framework and two ontologies which could be utilized to process and link quality related information in a machine-readable format.

1.4.1 State-of-the-Art Review

The research commenced with the review of related work to obtain appropriate knowledge about different aspects that impact the quality of mappings. A two-phase state of the art review was conducted to fulfill the research objectives **RO1(a)** and **RO1(b)**.

The first phase involved a review of existing approaches designed to improve the quality of mappings in the LD domain and fulfilled research objective **RO1(a).** The initial review was used to identify approaches currently utilized to improve the quality of mappings and to understand the requirements and limitations of the approaches. The information was used to identify a limitation in state of the art of a LD representation of mapping quality information. Mapping quality greatly impacts the quality of LD datasets, and the availability of expressive related information is argued will be beneficial. In addition, the review was used to identify quality metrics which are commonly used to assess the quality of mappings. The first phase of the review informed the research undertaken for the development of the approach in support of **RO2(a)** and **RO2(b)** related to support for mapping quality improvement. See **section 2.1**.

The second phase involved a review of existing approaches designed to address the dynamics of LD datasets and fulfilled research objective **RO1(b).** The reviewed approaches targeted the dynamics of the data in a LD dataset and links between them. No existing approach was identified which targets the dynamics of the source data in a manner that will preserve alignment of source data with its associated LD dataset, such that dynamics of the source data is mirrored in the dynamics of the LD dataset. Nonetheless, the review was used to identify the requirements and limitations of the existing approaches towards addressing dynamics in datasets. The second phase of the review informed the research undertaken for the development of the approach in support of **RO3(a)** and **RO3(b)** related to mapping and LD dynamics. See **section 2.2**.

1.4.2 Technical Approach

A **M**apping **Q**uality **I**mprovement framework named the **MQI** framework was developed to fulfill research objectives **RO2** and **RO3**. First, a state of the art review of existing approaches designed to improve the quality of mappings was undertaken. It was identified that that none of the approaches reviewed defined mapping quality information in an ontology designed for representing mapping quality. This is a limitation as a domain specific ontology enables the highest form of expressiveness [143], therefore, allowing easier understanding, processing and linking of the data [73]. Requirements were derived from the review and resulted in an approach that consisted of two components of the framework: the **mapping quality assessment and refinement component** was designed to improve the quality of mappings, and the **source change detection component** extended the framework to detect changes and alignment between source data and their respective mappings. The results of a user evaluation at end of each design iteration (Section 1.4.3) were used to inform subsequent steps.

The **mapping quality assessment and refinement component** of the MQI framework was developed to assess and refine the quality of R2RML mappings. The quality of the mapping was assessed by quality metrics informed by quality assessment of LD datasets in the state of the art. Users were guided through the process of identifying an appropriate value in order to resolve an identified mapping quality issue and associated value(s) being automatically inserted into the mapping by the framework. The **source change detection component** of the MQI framework included the initial functionality of the mapping quality assessment and refinement component with additional functionality integrated. The component was developed to detect changes and alignment between source data and respective mappings. In addition, the framework was extended to include support for RML mappings, which support source data represented in heterogenous formats such as CSV, JSON and relational databases. As a result of these iterations of development **R02(b)** and **RO3(b)** were achieved.

Two ontologies were created to represent the information in each component. The design of each followed best practices involving reusing existing reputable methodologies. The NeON methodology [134], UPON Lite [100], Ontology 101 development book [103] and LOT methodology [110] were reused during the development. The ontology development methodology involved an iterative process where requirements were defined in the form of natural language (non-functional) and ontology competency questions (functional). The defined requirements were translated to concepts and relationships, then constructed and assessed to ensure no logical inconsistencies were identified within the design. Feedback received from peer reviewed publications and dissemination of the documentation was used to iteratively refine the design. The methods used to improve the quality of Ontology for Source Change Detection (OSCD) and Mapping Quality Improvement Ontology (MQIO) are summarized in **Table 3.**

Table 3: Summary of methods used in development and improvement of ontologies

Method	**MQIO**	**OSCD**
User Experiment (Mapping Expert)	✓	✓
User Experiment (Ontology Expert)	✓	✓
Fulfilment of Competency Questions	✓	✓
Semantic Reasoners	✓	✓
OOPS! Common Pitfall Detection	✓	✓
Documentation	✓	✓
Demonstrate application to use-case	✓	✓
External use-case	✓	✓
Comparison with SoA	✓	✓
Analysis of citations	✓	✓
Dissemination of work	✓	✓
Peer reviewed publications	✓	✓
Reproducibility	✓	✓

The evaluation results of the ontology quality and improvement methods when applied to MQIO and OSCD are described in **Section 5.6.3**. The development and evaluation of the ontologies fulfilled research objectives **RO2(a)** and **RO3(a).**

1.4.3 Evaluation Strategy

The evaluation strategy involved 5 experiments and an application study. Experiments 1-3 focused on the design and implementation of the MQI framework. Experiments 4-5 focused on the design and implementation of the MQIO and OSCD ontologies, where feedback was gathered from experts who specialize in the domain of LD mappings and ontology development. In addition, an application study was conducted to demonstrate the applicability of the design in the real world using two use cases. A brief overview of the evaluations now follows.

1. **Evaluate the accuracy of the MQI Framework:** This evaluation involved inputting 30 R2RML-expressed uplift mappings into the framework. The quality information automatically generated by the framework was manually examined by the author of this thesis to ensure no quality issues were incorrectly identified. The mappings were collected from real world independent projects which involved postgraduate students (10 mappings) and research projects (20 mappings). The objective of the experiment was to test if the framework was capable of accurately detecting quality issues within real-world mappings. The results indicated sufficient detection of quality issues in mappings currently being deployed with 228 quality issues detected. The evaluation is described in **Section 5.3**.
2. **Evaluate user satisfaction and effectiveness of the MQI Framework:** This evaluation tested the usability and effectiveness of the MQI framework in relation to the detection and removal of quality issues in mappings. The experiment consisted of 58 participants interacting with the MQI framework. Each was provided with an uplift mapping containing three quality issues and was asked to assess and refine the quality of the mapping using the framework. The user satisfaction was measured through a standardized usability questionnaire (The Post-Study System Usability Questionnaire (PSSUQ) [86]) and effectiveness through comparison of quality of the original mapping with the mapping created as a result of the refinement undertaken by the participant in the experiment. Analysis was completed on the qualitative data to discover patterns which could guide design improvements. This evaluation is described in **Section 5.4**
3. **Evaluate user understanding of alignment between source data changes and mappings provided by the MQI Framework:** This evaluation involved a similar sample size and questionnaire to experiment 2, however, it was extended to include questions designed to test the understanding of the changes in source data detected by the source change detection component of the MQI framework. Similar analysis to experiment 2 was conducted on the collected data during the experiment. This evaluation is described in **Section 5.5**
4. **Evaluate the design of MQIO.** This evaluation involved 5 ontology engineers with 10+ years' experience to review the design/development of MQIO with respect to the following: 1) The design methodology followed during the design of MQIO and 2) The current version of the developed ontology, including documentation. The feedback was implemented which resulted in a refined ontology design. This evaluation is described in **Section 5.6**.
5. **Evaluate the design of OSCD.** This evaluation followed the identical structure of experiment 4, however, the questions were posed with respect to OSCD. This evaluation is described in **Section 5.6**.

6. **Apply the framework in real world.** An application study was conducted which involved applying the implementation of the framework to two use cases which used mappings to uplift diverse knowledge. This evaluation is described in **Section 5.7**.

1.5 Thesis Contributions

The **major contribution** of this thesis is the Mapping Quality Improvement (MQI) Framework. The **three minor contributions** include the two ontologies (MQIO, OSCD) which were developed for the framework to capture expressive mapping quality related information in an ontology-based format, and the final minor contribution is the evaluation results from the 5 experiments conducted on the proposed approach.

1.5.1 Mapping Quality Improvement (MQI) Framework

The **major contribution** of this thesis is the design and development of the MQI framework. The framework includes two core components which are designed to facilitate the 1) assessment and refinement of uplift mappings 2) detection of source data changes and alignment with respective mappings. Unlike existing LD quality approaches, the proposed approach was designed specifically to support the validating of quality early in the LD publication process, specifically during the uplift mapping process in mind. It is anticipated that the MQI framework will benefit the semantic web community who are continuously evolving the web of data by publishing LD datasets generated with uplift mappings. In addition, the approach will benefit the industry and research communities which utilize LD by facilitating the generation and maintenance of high-quality uplift mappings. The approach will enable the removal of mapping quality issues and prevent problems caused by poor quality mappings from exponential propagation into the resulting LD dataset. Capturing quality issues early in the publication stage should result in an improvement of the quality of any LD produced by mappings, and so, hopefully over time improve the quality of data available on the semantic web. In addition, the framework allows quality-oriented provenance information to be provided in an ontology-based format which can be linked with the resulting LD dataset, and so promote trustworthiness of the dataset for the consumers in allowing them to easily assess the suitability of the data for their application based on quality information related to the mappings used for the dataset generation. The framework fulfills research objectives **RO2(b)** and **RO3(b).**

The following peer reviewed publications are associated with this contribution:

- Randles, A., Junior, A. C., & O'Sullivan, D. (2020). **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 and Services (iiWAS '20), Virtual Event, 2020 (pp. 347–351).
    - This publication presents the first mapping quality framework developed. The framework used SHACL constraints to assess the quality of mappings and SPARQL queries to refine the mapping. A

demonstration walkthrough of the framework applied to a real-world use case is described. This publication led to the discovery of limitations with an approach reliant on SHACL.

- Randles, A., & O'Sullivan, D. (2021). **Assessing quality of R2RML mappings for OSi's Linked Open Data portal.** In Proceedings of the 4th International Workshop on Geospatial Linked Data (GeoLD) co-located with the 18th Extended Semantic Web Conference (ESWC 2021), Virtual event, 2021 (Vol. 2977, pp. 51–58).
    - This publication describes the application of the MQI framework to mappings used in Ireland's National Geospatial Data Hub (GeoHive)[3] project and provides a demonstration of how a quality issue could be identified and removed.
- Randles, A., & O'Sullivan, D. (2022). **Evaluating Quality Improvement Techniques Within the Linked Data Generation Process.** In Proceedings of the 18th International Conference on Semantic Systems (SEMANTiCS 2022), Austria, 2022 (Vol. 55, pp. 21–35).
    - This publication describes the design of the component within the MQI framework used to assess and refine the quality of mappings. A detail discussion of the first usability experiment (Section 1.4.3) carried out on the framework is presented along with the results and suggested improvements.
- Randles, A., O'Sullivan, D., Keeney, J., & Fallon, L. (2022). **Applying a Mapping Quality Framework in Cloud Native Monitoring.** In Proceedings of the 18th International Conference on Semantic Systems (SEMANTiCS 2022), Austria, 2022 (Vol. 3235).
    - This publication presents a summary of the source change detection component of the framework applied to mappings and source data which were used to uplift information in a cloud native monitoring use case in Ericsson.

1.5.2 Mapping Quality Improvement Ontology (MQIO)

The **first minor contribution** of this thesis is the MQIO. MQIO was designed to represent and interchange information related to the creation, quality assessment, refinement and validation of uplift mappings. MQIO defines concepts and relationships to model activities and agents involved in the detection and removal of quality issues in mappings. With the increasing amount of LD produced through mappings, an ontology to model related quality information would promote trustworthiness and maintenance for consumers of LD by providing additional quality-oriented provenance in an ontology-based format. Currently, no ontology exists to model information related to the creation, quality assessment, refinement and validation of LD mappings. Therefore, consumers cannot easily assess if the resulting dataset is sufficient for usage in their application. In addition, the information

[3] https://www.geohive.ie/

will benefit the maintenance and reuse of the mappings by providing data lineage. MQIO fulfills research objective **RO2(a)** and enables the fulfillment of **RO2(b).**

The following peer reviewed publications are associated with this contribution:

- Randles, A., Junior, A. C., & O'Sullivan, D. (2020). **Towards a vocabulary for mapping quality assessment**. In Proceedings of the 15th International Workshop on Ontology Matching co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual conference, 2020 (Vol. 2788, pp. 241–242).
    - This publication presents a summary of the concepts and relationships which were defined in the initial version of MQIO. The interactions between the concepts were also outlined. This publication allowed for early feedback on the design of the ontology.
- Randles, A., Junior, A. C., & O'Sullivan, D. (2021). **A Vocabulary for Describing Mapping Quality Assessment, Refinement and Validation.** In Proceedings of 15th IEEE International Conference on Semantic Computing (ICSC 2021), USA, 2021, 425–430.
    - This publication presents a detailed discussion of the development of MQIO, including the requirements, competency questions and reuse of existing ontologies. A demonstration walkthrough is also discussed which demonstrates the quality assessment, refinement and validation of a real-world use case mapping. Finally, related work on mapping quality assessment and refinement frameworks and provenance and metadata models.

1.5.3 Ontology for Source Change Detection (OSCD)

The **second minor contribution** of this thesis is the OSCD. OSCD was designed to represent and interchange information related to changes which have occurred in the source data used by mappings. OSCD defines concepts and relationships to model activities and agents involved in the source data used by mappings to produce LD. LD has been identified as extremely dynamic data, therefore, the probability of the alignment between mappings and the underlying data sources becoming out of sync is high. Oftentimes, resulting in data that does not accurately represent the data sources, therefore, decreasing the overall quality of the data. It is expected that capturing the information related to the changes which could impact the quality of the LD and respective mappings in an ontology-based format will help to promote the generation and maintenance of high-quality data. OSCD fulfills research objective **RO3(a)** and enables the fulfillment of **RO3(b)**.

The following peer reviewed publications are associated with this contribution:

- Randles, A., & O'Sullivan, D. (2022). **Modeling & Analyzing Changes within LD Source Data.** In Proceedings of the 8th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW) co-located with the 21st International Semantic Web Conference (ISWC 2022), Virtual event, 2022, 19–27.
    - This publication presents the design of OSCD, including the concepts, relationships and interaction between them. The evaluation methods used during the development of OSCD are

also mentioned. Furthermore, the design of the source data source change detection component of the MQI framework is outlined along with the application of OSCD within the component. Finally, the application of the ontology and framework in a cloud native monitoring use case is discussed.

- Randles, A., & O'Sullivan, D. (2023). **Preserving the Alignment of LD with Source Data.** In Proceedings of the 4th International Workshop on Knowledge Graph Construction (KGCW) co-located with the 20th Extended Semantic Web Conference (ESWC 2023), Greece, 2023.
  - This publication describes the design of the component within the MQI framework used to preserve alignment between source data and respective mappings using the OSCD. A detail discussion of the second usability experiment (Section 1.4.3) carried out on the framework is presented along with the results and suggested improvements. In addition, the publication presents a high-level overview of additional functionality designed to support the existing functionality in automatically improving alignment, which was added after the evaluation was completed.

1.5.4 Evaluation Results

The **third minor contribution** is the evaluation results from the 5 experiments outlined in this thesis. Unlike, the majority of existing quality related state of the art approaches, the MQI framework has been evaluated through the use of system and user experiments which involved standardized methods. Similar approaches in the state of the art are evaluated through system only evaluations. While one of the approaches reviewed did perform a demonstration with users, no evaluation methods and metrics were published arising from this demonstration. In addition, the 5 experiments conducted on the MQI framework involved over 100 participants and were designed to collect feedback from participants with varying backgrounds, including knowledge engineering students, mapping specialist and ontology design specialist. Therefore, allowing the level of background knowledge required to successfully interact with the framework to be measured. In addition, the experiments enable the approach to be refined with respect to diverse and extensive feedback. It is hoped the combination of experiments which involved evaluating the accuracy, effectiveness and validation of the approach can be applied to similar approaches to consolidate them, therefore, ensuring a suitable level of usability for the intended respective end users. The evaluation results fulfill research objective **RO4.**

The following peer reviewed publications are associated with this contribution:

- Randles, A., & O'Sullivan, D. (2022). **Evaluating Quality Improvement Techniques Within the Linked Data Generation Process.** In Proceedings of the 18th International Conference on Semantic Systems (SEMANTiCS 2022), Austria, 2022 (Vol. 55, pp. 21–35).
  - This publication describes the design of the component within the MQI framework used to assess and refine the quality of mappings. A detail discussion of the first usability experiment (Section

    - 1.4.3) carried out on the framework is presented along with the results and suggested improvements.

- Randles, A., & O'Sullivan, D. (2023). **Preserving the Alignment of LD with Source Data.** In Proceedings of the 4th International Workshop on Knowledge Graph Construction (KGCW) co-located with the 20th Extended Semantic Web Conference (ESWC 2023), Greece, 2023.
    - This publication describes the design of the component within the MQI framework used to preserve alignment between source data and respective mappings. A detail discussion of the second usability experiment (Section 1.4.3) carried out on the framework is presented along with the results and suggested improvements.
- Randles, A., O'Sullivan, D., Keeney, J., & Fallon, L. (2023). **Ontology Driven Closed Control Loop Automation.** In Proceedings of the 3rd International Workshop on Intent-Based Networking (WIN) co-located with the 9th IEEE International Conference on Network Softwarization (NetSoft 2023). Spain, 2023.
    - This publication describes the Prometheus RDF Generator framework and associated ontology, which were designed to transform monitoring data from a popular network/cloud monitoring system (Prometheus) into RDF representation. The MQI framework was applied to the mappings and source data involved in the use case (Section 1.4.3) in order to improve and maintain the quality of the artefacts.

1.6 Thesis Overview

The thesis is structured as follows:

Chapter 2: State of the Art

First, this chapter discusses existing approaches designed to assess and refine the quality of mappings. Thereafter, ontologies used by the approaches to capture the quality information of a mapping are discussed. Existing approaches which are targeted at the dynamics of LD datasets are also discussed. Finally, ontologies designed to represent information related to the dynamics of LD datasets are discussed. This chapter concludes with a discussion of limitations of the approaches and proposed resolutions to them as described in this thesis.

Chapter 3: Background

This chapter provides background information related to the mapping process in the LD domain. The objective is to provide useful introductory information for the readers to gain an understanding of the agents, processes and artefacts involved in the mapping process. The types of mappings used in the LD domain: semantic and uplift mappings are mentioned. Finally, prominent languages used to define a representation of the mappings are described.

Chapter 4: MQI Framework

First, this chapter presents the requirements for the framework which were inspired by the state of the art. In addition, support of requirements by existing approaches are mentioned. Thereafter, the chapter discusses the design and implementation of the mapping quality assessment and refinement component of the MQI framework and MQIO, which were designed to improve the quality of uplift mappings. Thereafter, this chapter discusses the design and implementation of the source change detection component of the MQI framework and OSCD, which were designed to preserve the alignment between source data and respective mappings.

Chapter 5: Evaluations

This chapter describes the 5 experiments and 1 application study which were outlined in Section 1.4.3.

Chapter 6: Conclusion

This chapter concludes the thesis and presents a summary of important findings which were discovered throughout the lifecycle of the conducted research. The extent to which the research described fulfills the research objectives is also outlined. Finally, proposed future work is described.

Appendices: The appendices provide supplementary documentation related to the design and evaluation of the MQI framework, MQIO and OSCD.

Supplementary Information:

Additional supplementary information related to the software in this thesis is stored in this **GitHub repository:**

GitHub Repository
https://github.com/alex-randles/Thesis-Supplementary-Information/

Chapter 2: State of the Art

This chapter presents related work discovered through the state of the art review. The reviewed approaches are related to the research question and objectives of this thesis, which propose approaches to target two main aspects that have been shown to impact LD dataset quality. The **first aspect** focuses on the quality of the mapping artefact itself and how quality issues can be identified and removed from them. Therefore, approaches which support mapping engineers in detecting and resolving mapping quality issues are reviewed. The **second aspect** focuses on approaches that handle changes which have occurred in the source data associated with uplift mappings, after the LD dataset has been published, which impacts the level of alignment between the source data and LD dataset, and which may result in decreased quality in the LD dataset. Therefore, approaches which support the detection of changes in source data and changes in mappings are reviewed. The reviewed approaches were discovered by searching Google Scholar[4], which indexes multiple other sources, such as ACM Digital Library[5], IEEE explorer[6] and Semantic Scholar[7], among others. **The retrieval process initiated with a seed paper, which was used to guide the process and cascaded from there (following references from this and other papers as they were retrieved and filtered).**

2.1 Approaches that support Mapping Quality Improvement

The **first phase** of the review involved reviewing the state of the art of approaches designed to support LD mapping quality improvement. The analysis of these approaches resulted in the identification of limitations in existing approaches. This phase was completed prior to the design and development of the **mapping quality assessment and refinement component** of the MQI framework. By reviewing the approaches relevant information was gathered and definition of requirements for the mapping quality assessment and refinement component of the framework (designed to resolve identified limitations within the state of the art) was undertaken. The retrieval

[4] https://scholar.google.com/
[5] https://dl.acm.org/
[6] https://ieeexplore.ieee.org/Xplore/home.jsp
[7] https://www.semanticscholar.org/

process commenced with a search for papers detailing literature reviews related to LD mapping quality, which is the focus of the research in this thesis. A literature review is defined as a "*survey of scholarly sources on a specific topic. It provides an overview of current knowledge, allowing you to identify relevant theories, methods, and gaps in the existing research that you can later apply to your paper, thesis, or dissertation topic*" [41]. It was decided to search for a seed paper in order to initiate the discovery process. A survey paper was selected as it provides a broad overview of existing research related to the domain, use cases and importantly related work. Therefore, a search of Google Scholar was completed in order to identify a suitable survey, however, at the time no survey was directly identified related to the quality of declarative mappings. It was decided to search for a survey detailing the quality of the resulting LD datasets. Another search was completed using Google Scholar and a survey [150] published in a semantic web journal was identified. The survey provided an in-depth discussion on various aspects which impact LD quality and includes a full description of quality metrics, dimensions and categories. In addition, the survey was prominent with over 550 citations as of 2023, which was significantly more than any other work discovered during the process. Each of the 550 citations were manually examined by reading the abstracts in order to identify the relevance of the work to quality assessment and refinement of mappings. The process resulted in the discovery of several approaches designed to support uplift mapping quality assessment and refinement, however, the majority of the discovered research focused on the quality of LD datasets and were not directly relevant (although heavily influenced motivation for design issues to consider). In addition, several of the works related to the quality of semantic mappings, however, these approaches were not considered as the focus of the research in this thesis is on uplift mappings. Each related work section in the papers of the approaches was examined in order to identify further relevant work. In addition, each of the 98 works which cited these approaches was manually examined by reading each abstract in order to identify further relevant work. Finally, a search of the proceedings of popular semantic web conferences, such as ESWC[8], ISWC[9], ICSC[10], iiWAS[11] and SEMANTiCS[12] was conducted in order to identify further relevant work. The final approaches were consolidated by recommendations from domain experts through various publications related to the MQI framework, which included descriptions of relevant work.

From this state of art review process 6 key approaches emerged, which focused specifically on the quality of declarative mappings and which cited each other identifying similarities between them. These key approaches are presented in **subsections 2.1.1** to **2.1.6**. As previously stated, the research of this thesis is focused on detection of

[8] https://2023.eswc-conferences.org/
[9] https://iswc2023.semanticweb.org/
[10] https://www.ieee-icsc.org/
[11] https://www.iiwas.org/conferences/iiwas2023/
[12] https://2023-eu.semantics.cc/

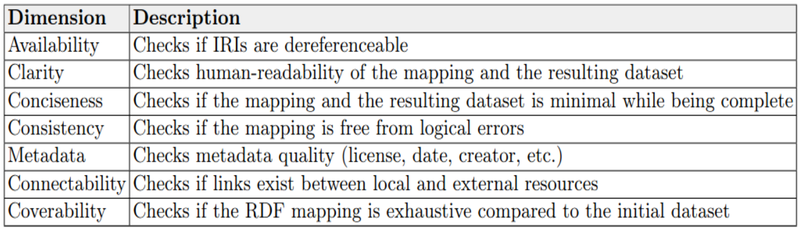
quality issues in declarative mappings in order to facilitate high-quality LD. An overview of each selected approach is presented in terms of the key characteristics which are of interest for this thesis research: target mapping language, required input data, representation of relevant quality information, improvement capabilities and interface functionalities. In addition, experiments completed on them are discussed in order to identify the level of validation of the proposed approaches.

In **section 2.1.7**, these 6 key approaches are then analyzed together with respect to the important characteristics of interest for this thesis research.

2.1.1 EvaMap

Evaluate RDF **Map**pings (EvaMap) [97] is a framework designed to assess the quality of YARRRML (Human-Readable Mapping Language) [94] mappings. The framework uses quality metrics which are organized into 7 quality dimensions. The metrics are executed on the mapping itself or an extract of the resulting RDF dataset when instances are required. Therefore, taking into account the dataset too. The dimensions targeted are presented in **Table 4**.

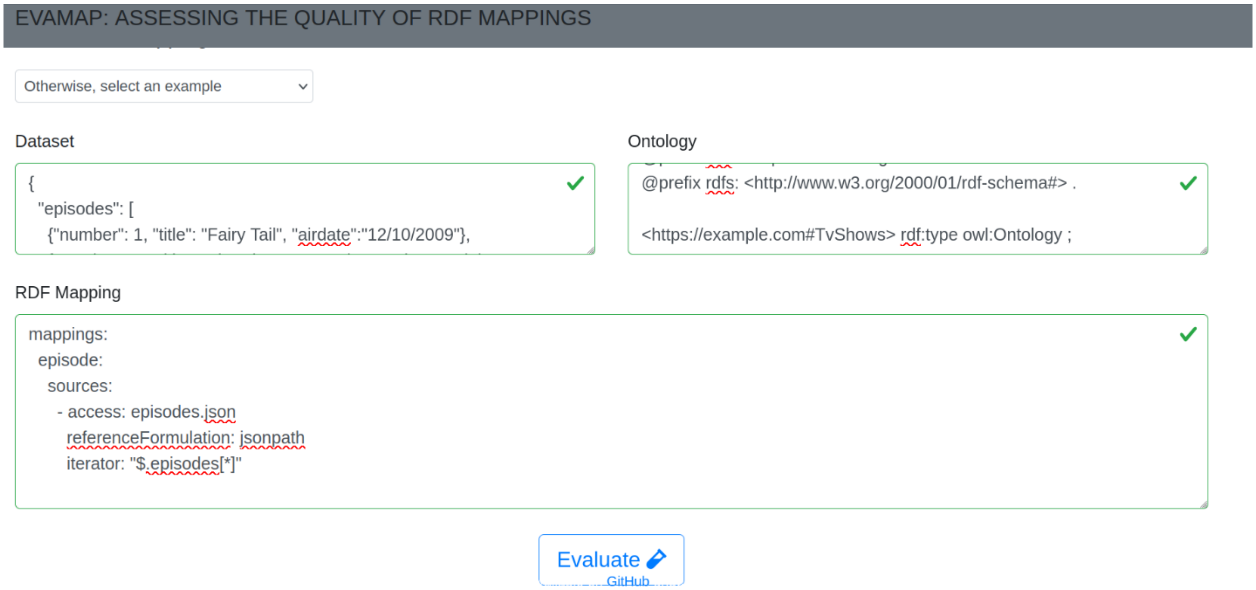
Table 4: Quality Dimensions used by EvaMap



Dimension	Description
Availability	Checks if IRIs are dereferenceable
Clarity	Checks human-readability of the mapping and the resulting dataset
Conciseness	Checks if the mapping and the resulting dataset is minimal while being complete
Consistency	Checks if the mapping is free from logical errors
Metadata	Checks metadata quality (license, date, creator, etc.)
Connectability	Checks if links exist between local and external resources
Coverability	Checks if the RDF mapping is exhaustive compared to the initial dataset

The approach proposes a function which can be used to compute the quality of the mapping by computing a weighted score for the dimensions. The weights can be adjusted to represent the importance of a dimension in the particular use case. The framework has been published online with example data[13]. The result of the assessment of a mapping includes a global quality score and human-readable feedback on how to improve it. EvaMap was demonstrated in action as part of demo track at ESWC, which involved students interacting with the framework without specific tasks to complete and no feedback to provide. As such, no formal evaluations have been published as yet. **Figure 1** shows a screenshot of the interface of EvaMap displaying sample input.

[13] https://evamap.herokuapp.com/


EVAMAP: ASSESSING THE QUALITY OF RDF MAPPINGS
Otherwise, select an example
Dataset
Ontology
RDF Mapping
Evaluate
GitHub

```
{
  "episodes": [
    {"number": 1, "title": "Fairy Tail", "airdate":"12/10/2009"},
```

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<https://example.com#TvShows> rdf:type owl:Ontology ;
```

```
mappings:
  episode:
    sources:
      - access: episodes.json
        referenceFormulation: jsonpath
        iterator: "$.episodes[*]"
```

Figure 1: Screenshot of EvaMap interface

Users are required to input the YARRRML mapping, ontologies defined in the mapping and the resulting dataset, as can be seen in the screenshot.

2.1.2 Resglass

Resglass [65] extends an existing rule driven method [42] to consider refinements for transformation rule as well as used ontology terms. The approach provides rankings for the rules to indicate which should be inspected first by the mapping engineer. The approach focuses on two types of inconsistencies, which include 1) raw data inconsistencies and 2) rules that introduce inconsistencies. The previous rule driven method proposed refinements for rules only and assumed used ontologies provide an accurate representation of the users intended semantic model. In addition, the method did not rank the rules. The research question proposed by the work is *"How can we score and rank rules and ontology terms for inspection to improve the manual resolution of inconsistencies?".*

The work provides algorithms to rank rules and rank ontology terms in order to detect inconsistencies. An implementation of approach and an evaluation, which compares the rankings generated with expert rankings. The approach includes the following steps:

1. **Rules inconsistency detection:** This step involves detecting rules which are inconsistent with respect to the ontologies used. First, axiom constraints are generated on the condition that the axioms can be interpreted as constraints. Rules that result in failed constraints are grouped. These groups are analysed to assess if they respect the related constraint, otherwise an inconsistency is detected.
2. **Rules and ontology definitions refinement:** This step involves three sub steps, which include rules clustering, rules and ontology terms ranking and rules and ontology definitions refinement. Rules clustering involves grouping rules related to inconsistencies with respect to the contribution to the

resulting graph. The grouping is based on the related record, such as rows in a table. Rules and ontology term ranking involves scoring the rules using a formula designed for the use case. The formula is based on the number of inconsistencies related to specific rules and ontology terms when compared to the total number of inconsistencies. Equal scores are ranked randomly. Rules and ontology definitions refinement involves selecting the clusters of the highest ranked rules and ontology terms and identifying necessary refinements. The inspection is completed manually by an expert. Thereafter, the refined rules can be re-assessed to identify remaining inconsistencies. In addition, the assessment identifies inconsistencies introduced by the changes to the rules.

3. **Knowledge graph generation:** The graph is generated using the source data, refined rules and ontology terms.
4. **Knowledge graph inconsistency detection:** The result graph is assessed for inconsistencies.
5. **Rules and ontology definitions refinement:** Rules and ontology terms could be further refined based on inconsistencies detected in the resulting graph.

Figure 2 presents an overview of the processes involved in the mapping quality assessment and refinement in the Resglass design.

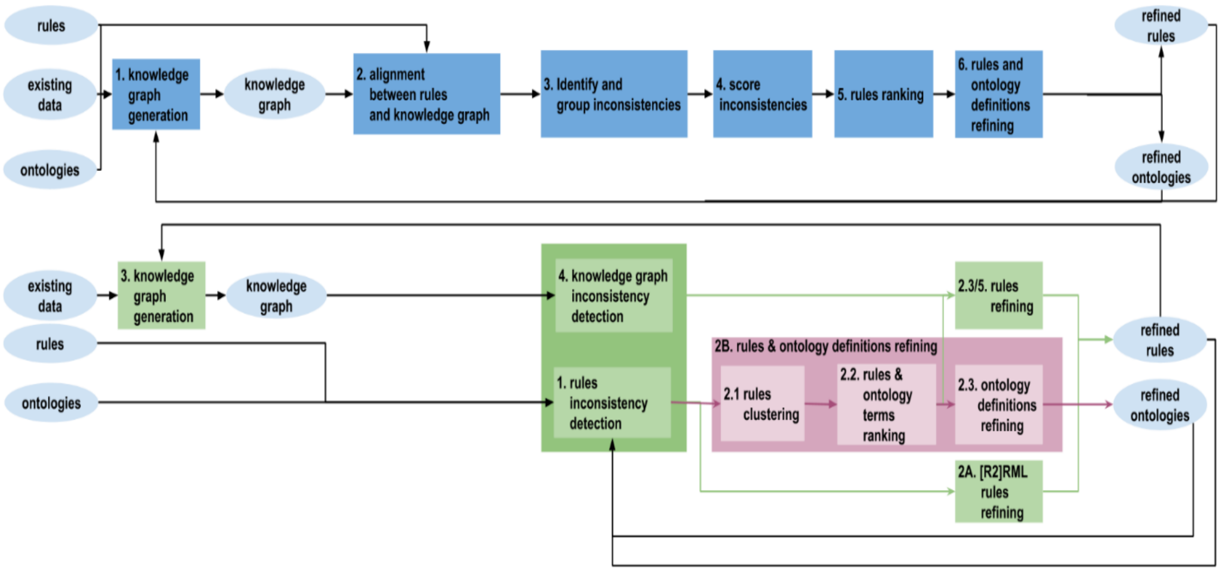

rules
existing data
ontologies
1. knowledge graph generation
knowledge graph
2. alignment between rules and knowledge graph
3. Identify and group inconsistencies
4. score inconsistencies
5. rules ranking
6. rules and ontology definitions refining
refined rules
refined ontologies
existing data
rules
ontologies
3. knowledge graph generation
knowledge graph
4. knowledge graph inconsistency detection
1. rules inconsistency detection
2B. rules & ontology definitions refining
2.1 rules clustering
2.2. rules & ontology terms ranking
2.3. ontology definitions refining
2.3/5. rules refining
2A. [R2]RML rules refining
refined rules
refined ontologies

Figure 2: Processes involved mapping quality assessment and refinement using the Resglass approach

The implementation is designed to rank rules expressed in RML [44]. Some steps are described further as the others are independent of the language. The implementation steps are outlined below:

1. **Rules inconsistency detection:** A rule-based reasoning system named RDFUnit [81] is used to execute constraints when the RML rules, ontologies and inference rules are input. The system outputs details of detected inconsistencies, such as associated RML rules and ontology terms.

2. **RML rules clustering:** The rules are clustered based on related triple map. The rationale for choosing triple maps is that each term map is related to one.
3. **RML rules ranking:** Scores are calculated for the inconsistencies in each triple map cluster. Unique inconsistencies are only counted once for each cluster.
4. **RML rules refinement:** The refinement of rules and ontology terms are completed by experts rather than the implementation.
5. **Knowledge graph generation:** The refined RML rules are executed with the input source data, resulting in the generation of a refined graph.
6. **Knowledge graph inconsistency detection:** Rule based reasoning is used to detect inconsistencies in the graph.

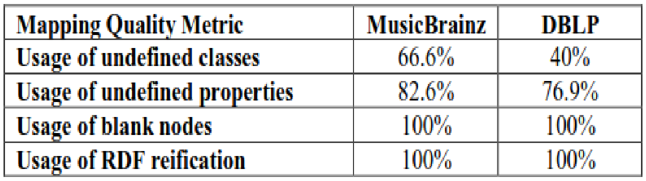
   RML rules refinement: Rules are further refined by experts based on inconsistencies detected in the resulting graph.

The evaluation involved automatically ranking 100 triple maps and respective ontology terms using Resglass. Thereafter, they were manually ranked by experts. The mean overlap between the scores is 81% for the rules and 79% for ontology terms, indicating a large overlap between the scores. The results indicated that Resglass ranking can improve expert overlap with 41% for rules and 23% for ontology terms.

2.1.3 Luzzu-Extension

The mapping quality assessment approach [75] involved the author of this thesis in its development, and uses a set of metrics commonly used to assess the quality of RDF datasets in order to evaluate the mappings which were used to generate them. The implementation was designed to assess R2RML mappings and was implemented by extending an existing quality assessment framework named Luzzu. The assessment process results in two machine-readable reports containing information on detected mapping quality issues and other relevant quality metadata. In addition, the work discusses integrating the report into a mapping editing framework named the Jigsaw Puzzle for Representing Mappings (Juma) [30] to allow use of the graphical user interface in order to alter the mapping. No user evaluation was completed on the approach, however, it has been evaluated using real world use cases. The use cases included the MusicBrainz project and the Computer Science bibliography (DBLP) dataset. The MusicBrainz project included 12 R2RML mappings which were used to uplift data containing music metadata related to artist, releases, works, labels, recordings, among others. DBLP included D2RQ [15] mappings designed to uplift information related to computer science proceedings. The mappings were converted into R2RML representation. **Table 5** presents an overview of the evaluation results collected from the quality assessment of these mappings using the implementation.

Table 5: Evaluation results from Luzzu-Extension



Mapping Quality Metric	MusicBrainz	DBLP
Usage of undefined classes	66.6%	40%
Usage of undefined properties	82.6%	76.9%
Usage of blank nodes	100%	100%
Usage of RDF reification	100%	100%

The results shown are outlined below.

- No use of RDF reification model or blank nodes in mappings.
- *MusicBrainz mappings*
  - 33.4% contained undefined classes.
  - 17.4% contained undefined properties.
  - All originated in one mapping as the ontology could not be returned.
- *DBLP mappings*
  - 60% contained undefined classes.
  - 23.1% contained undefined properties.
  - Most issues related to typos such as `dcterms:partOf` and `dcterms:isPartOf`.

The preliminary results indicated that the approach is capable of identify quality issues for certain cases.

2.1.4 RML-Validator

RML-Validator [42] uses a test-driven approach for the quality assessment of mappings and suggests semi-automatic refinements based on the assessment results. The approach focuses on the intrinsic dimension of data quality and applies test cases to mappings to calculate conformance. The assessment workflow involves six steps designed to be uniform, iterative and incremental. **Figure 3** presents an overview of the workflow of the approach.

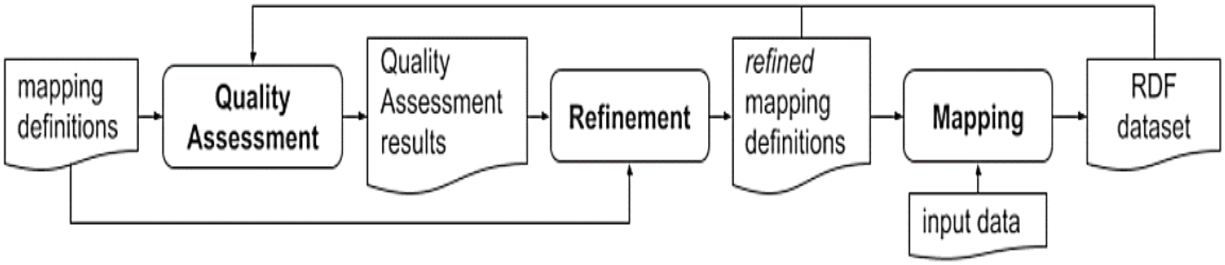


Figure 3: Overview of the workflow of the RML-Validator

The workflow presented in **Figure 3** is outlined below.

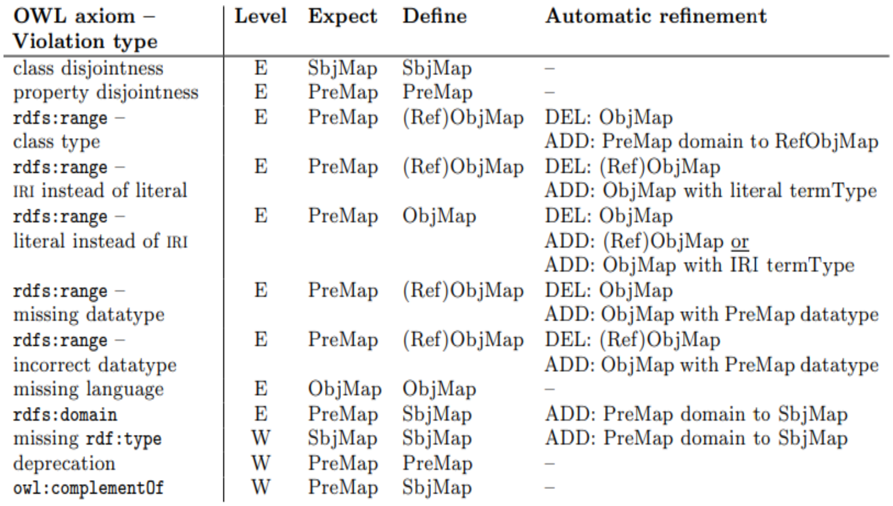
1. The mapping is assessed using different quality metrics.
2. The quality report is output.

3. The quality assessment report is used to refine the mapping until it cannot be further refined.
4. The refined mapping is used to generate a sample of the RDF data.
5. The resulting RDF data is assessed and used to further refine the mapping (if possible).
6. Finally, the final refined mapping is used to generate a better-quality dataset.

The approach has been implemented as a command line tool which targets RML [44] mappings and extends an existing test-based LD quality assessment framework named RDFUnit [81]. The assessment targets 1) conformance to R2RML and RML schema and 2) validation of datasets to be generated against the schema defined in the mapping. Conformance was tested by extending RDFUnit to automatically generate 78 test cases from the OWL axioms in the [R2]RML schema. The validation of datasets to be generated requires the source data to generate a sample of the data. The ontologies used to define properties and namespaces in the mapping are retrieved to generate test cases, similar to datasets. The test cases in RDFUnit are represented as SPARQL queries which have been extended to include queries designed to target mappings.

The results of the mapping quality assessment process are represented in the Test-Driven RDF Validation Ontology [81], which provides concepts and relationships to represent test cases and associated results. Refinements are suggested for quality issues and are designed to provide the minimum required actions to resolve an issue, which provide a proof-of-concept of automated quality improvement. The approach describes various refinements which could be used to automatically refine a violation detected by a specific quality metric. Checks are required to ensure that the selected refinement does not result in additional violations when executed. For instance, adding a domain class to the triple map may result in disjoint classes. **Table 6** presents an overview of violation types and associated refinements.

Table 6: Quality violations detected and associated refinements in the RML-Validator



OWL axiom – Violation type	Level	Expect	Define	Automatic refinement
class disjointness	E	SbjMap	SbjMap	–
property disjointness	E	PreMap	PreMap	–
rdfs:range – class type	E	PreMap	(Ref)ObjMap	DEL: ObjMap ADD: PreMap domain to RefObjMap
rdfs:range – IRI instead of literal	E	PreMap	(Ref)ObjMap	DEL: (Ref)ObjMap ADD: ObjMap with literal termType
rdfs:range – literal instead of IRI	E	PreMap	ObjMap	DEL: ObjMap ADD: (Ref)ObjMap or ADD: ObjMap with IRI termType
rdfs:range – missing datatype	E	PreMap	(Ref)ObjMap	DEL: ObjMap ADD: ObjMap with PreMap datatype
rdfs:range – incorrect datatype	E	PreMap	(Ref)ObjMap	DEL: (Ref)ObjMap ADD: ObjMap with PreMap datatype
missing language	E	ObjMap	ObjMap	–
rdfs:domain	E	PreMap	SbjMap	ADD: PreMap domain to SbjMap
missing rdf:type	W	SbjMap	SbjMap	ADD: PreMap domain to SbjMap
deprecation	W	PreMap	PreMap	–
owl:complementOf	W	PreMap	SbjMap	–

The approach has been applied to use cases, which included the DBpedia (from "DB" for "database") community, DBLP, Contact Details of Flemish Local Governments Dataset (CDFLG), CEUR-WS and iLastic. The mappings and datasets of these use cases were used to evaluate the approach. The results show it is far more efficient to assess the quality of the mapping when compared to the dataset, with 700 DBpedia mappings assessed in 11 seconds. The evaluation of DBpedia resulted in 1316 domain level violations. DBLP had 7 individual violations, resulting in 8.1M violated triples, however, 98% of them could be refined by the approach. CDFLG had four violations and some of the range violations (7%) could not be refined. iLastic used the approach during iterations on the primary version of mappings until they became free of violations. CEUR-WS contained 12 violations in the RML mappings used for the ESWC 2014 challenge and most of them were domain-level violations.

2.1.5 PROV-O Based

The approach presented in [101] represents provenance related to mapping quality assessment and refinement in the ontology PROV-O [84]. The objective of the approach is to provide trust assessment in order to indicate the level of quality in RDF data and respective mappings. The provenance is captured in four main steps shown in **Figure 4**.

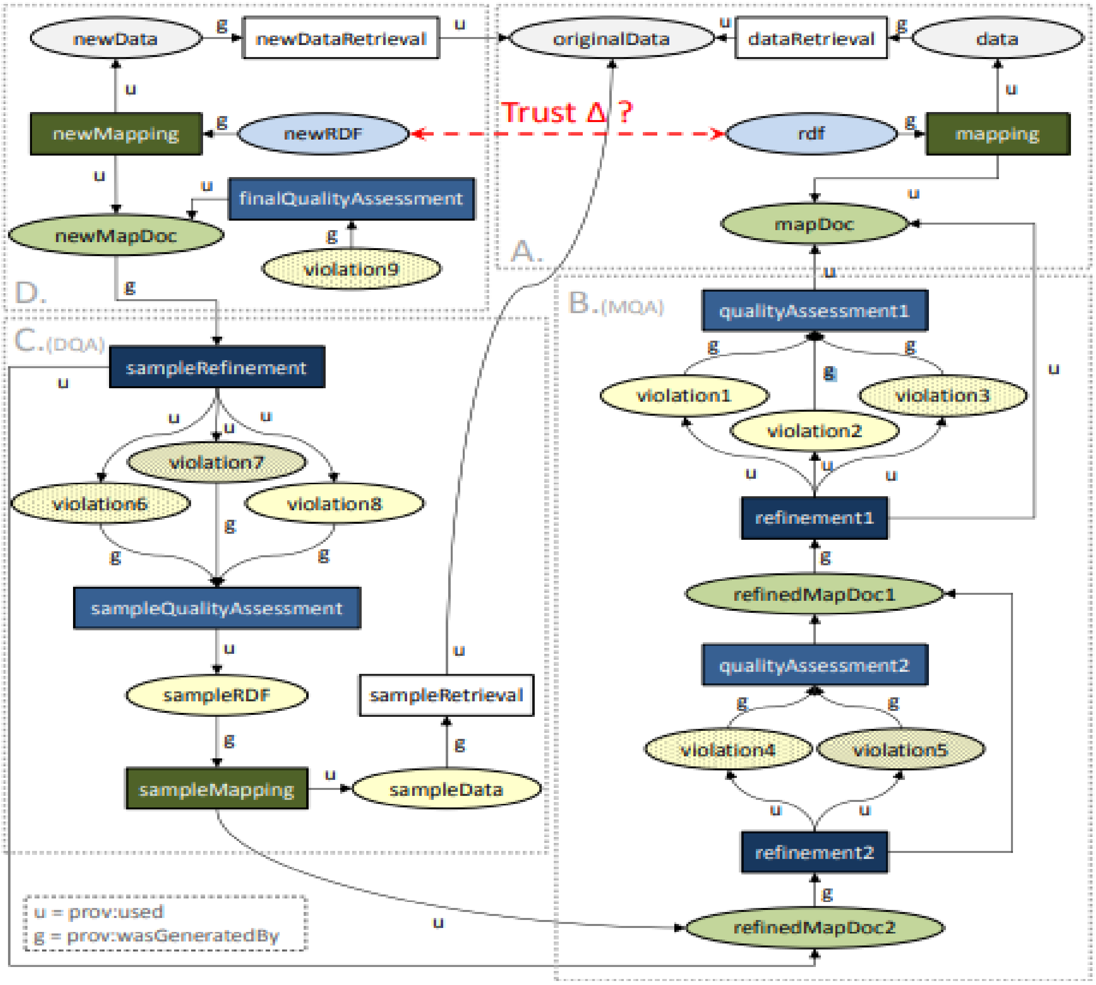


Figure 4: Overview of the provenance of data generated using while assessing and refining a mapping document

The four steps shown are outlined below.

- **A) Provenance of Original Mapping Document:** First, provenance related to the original mapping is captured. The original data (`prov:Entity`) is retrieved and used by the mapping activity (`prov:Activity`) which generates an extract of the resulting RDF data (`prov:Entity`).
- **B) Provenance of Mappings Quality Assessment and Refinement:** The quality of the mapping is assessed (`prov:Activty`) to identify quality violations (`prov:Entity`). Thereafter, the violations are refined (`prov:Activity`) resulting in a new mapping document (`prov:Entity`). The process is completed iteratively in order to remove the most amount of violations possible. The violations can also include an additional type apart from `prov:Entity`.
- **C) Provenance of Dataset Quality Assessment:** This step involves retrieving (`prov:Activity`) a sample of the original dataset (`prov:Entity`) and use it in a sample mapping activity. The mapping activity involves executing (`prov:Activity`) the original mapping resulting in a extract of RDF data. The quality assessment (`prov:Entity`) of the data helps to identify further violations. The result of the step is the final refined mapping.
- **D) Final Mapping:** This step involves executing the refined mapping (`prov:Entity)` resulting in a new RDF representation. Finally, step A is completed on the refined mapping in order to identify which violations have not been resolved.

No evaluation methods are mentioned in the work, however, it concludes that *"Richer semantics are needed to describe the results (e.g., violations) of quality assessments"*. Therefore, indicating an ontology designed for mapping quality improvement could benefit such an ontology-based approach.

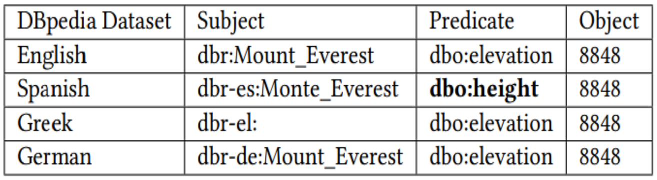
2.1.6 Predictive Model

The approach presented in [122] proposes a data-driven method designed to automatically detect incorrect mappings in the English, Spanish, Greek, and Dutch DBpedia instances using a predictive model. The research question studied in the work is *"is it possible to automatically detect incorrect mappings by analyzing two knowledge graphs created using two sets of different mappings?"*. As DBpedia mappings are crowdsourced and created by a diverse community, inaccuracies and inconsistencies are common. The work specifically targets quality related to the data conciseness quality dimension of the intrinsic category (Section 4.2.2.1). The work presented uses a data-driven approach involving multiple graphs and the problem addressed in it cannot be addressed only by analyzing the mapping definitions. The model includes a machine learning based approach to detect incorrect mappings by automatically analyzing the information contained in instance data and related ontology axioms. A feature extraction method is used in order to extract 22 numeric features for each mapping language-pair in order to compare the instance data in two respective graphs. A classifier is used to analyze these features in order to identify inconsistencies and classify respective mappings as *"correct"* or *"incorrect"*. The approach assumes that a resource which has the same subject-object value for two distinct properties, there is a high probability of having a mapping inconsistency, i.e., the same relation is mapped to two distinct properties. Exclusions are applied to this assumption for certain cases where similar relations are likely, such as a person's birth (`dbo:birthPlace`) and death (`dbo:deathPlace`) place. The generation of required descriptions of features involves the following two steps:

1. **Instance-based features:** No schema information is required for this step. The instances are analyzed in order to identify the number of occurrences of triples with the same subject-object and two distinct properties.
2. **Schema-based features:** Extraction of schema features involves attempting to identify properties which have been incorrectly reused or redundant duplicate properties for the same relation.

Training data was required in order to train the supervised classifiers used by the approach. Therefore, experts in 4 DBpedia chapters were asked to manually inspect and classify DBpedia mappings. Mappings were selected from 4 combinations of DBpedia datasets (English-Spanish, English-German, English-Dutch, English-Greek) and the language-pair experts asked to annotate the mappings as correct or incorrect. The resulting training data consisted of 226 mappings in total, with 182 mappings annotated as *"Correct"* and 44 as *"Incorrect"*. Therefore, the simplest classifier (known as ZeroR), which assigns the most popular class value, has an accuracy of 64.29%. This classifier establishes the baseline value that must be enhanced by our model. **Table 7** presents examples of inconsistencies identified in DBpedia mappings.

Table 7: Examples of data from incorrect mappings



DBpedia Dataset	Subject	Predicate	Object
English	dbr:Mount_Everest	dbo:elevation	8848
Spanish	dbr-es:Monte_Everest	**dbo:height**	8848
Greek	dbr-el:	dbo:elevation	8848
German	dbr-de:Mount_Everest	dbo:elevation	8848

As can be seen, different language versions of DBpedia datasets contain incorrect usage of certain properties, such as `dbo:height`, which should be used to represent a human's height rather than a mountains height (`dbr:Mount_Everest`). The approach is demonstrated by analyzing: (1) instance data from distinct language-specific datasets, and (2) the ontological axioms of the DBpedia ontology.

The approach was evaluated by comparing several supervised learning algorithms. The ROC curve (receiver operating characteristic curve) is a commonly used graph to compare classifiers, which shows the performance of a classification model at all classification thresholds, including the true and false positive rates. **Figure 5** presents the ROC curver for the tested classifiers.

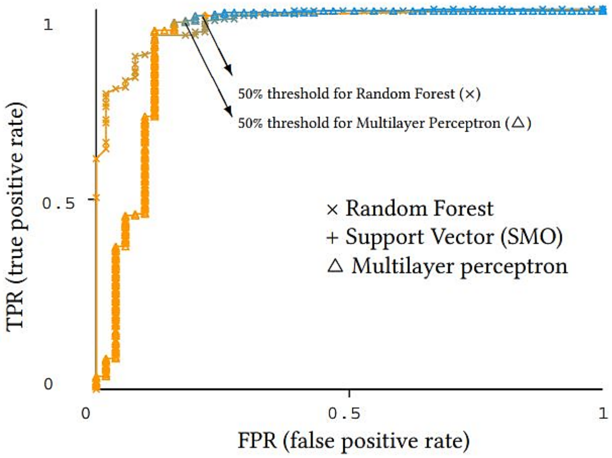
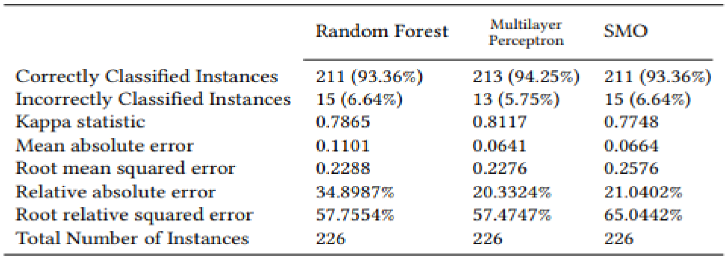


Figure 5: ROC curve for tested classifiers in the Predictive Model

Table 8 presents a summary of statistics collected when different supervised classification models were tasked with classifying the 226 instances of DBpedia mappings.

Table 8: Summary of supervised classifiers output for the Predictive Model



	Random Forest	Multilayer Perceptron	SMO
Correctly Classified Instances	211 (93.36%)	213 (94.25%)	211 (93.36%)
Incorrectly Classified Instances	15 (6.64%)	13 (5.75%)	15 (6.64%)
Kappa statistic	0.7865	0.8117	0.7748
Mean absolute error	0.1101	0.0641	0.0664
Root mean squared error	0.2288	0.2276	0.2576
Relative absolute error	34.8987%	20.3324%	21.0402%
Root relative squared error	57.7554%	57.4747%	65.0442%
Total Number of Instances	226	226	226

As can be seen, the best results for accuracy (correctly classified instances) for different supervised classification algorithms were Random Forest (93.36%), Multilayer Perceptron (94.25%) and Support Vector (SMO) classifier (93.36%). However, Random Forest can be considered the best classifier as it has the highest ROC AUC (Area Under the Curve) among the aforementioned tested, which indicates the best cost and least false positives.

The results showed that the model is better at predicting mappings for certain language pairs. For instance, the English-German language pairs had an accuracy of 87%, while English-Dutch had an accuracy of 61%. In addition, it was concluded that a general predictive model for all language-pairs performs better than a unique models for specific pairs. The quality issues which resulted in respective classifications of mappings were not discussed in the work. Therefore, it is difficult to identify why mappings were classified correct or incorrect and provide resolutions to these quality issues.

Two main limitations were identified with the proposed approach. The model is likely to produce incorrect classifications with triples using the same subject-object pairs and two distinct properties. For instance, the `dbo:deathDate` property is normally mapped to only the year of death and so is the `dbo:deathYear` property. The approach has difficulty identifying that these properties are correctly used and is likely to classify the mapping as incorrect. Another limitation is the high level of domain knowledge required to annotate mappings in the heterogeneous domains on DBpedia.

2.1.7 Analysis of Mapping Quality Improvement SoA

The design and implementation factors were chosen based on key characteristics noted during the state of the art review. This section compares the reviewed approaches presented in **sections 2.1.1** to **2.1.6**, based on characteristics related to the research question of this thesis. The characteristics were defined to represent common features, which were noted during the state of the art review. The characteristics provide a basis for comparison between them. In addition, the experiments conducted on the various approaches were also compared to provide insights into the design of the evaluation strategy of the MQI framework. Thereafter, identified limitations with the approaches were used to inform requirements for the design and implementation of the **mapping quality assessment and refinement component** of the MQI framework. The characteristics used for the analysis are outlined below.

- **Design characteristics**
    - **Mapping Language**: Representation of the mapping targeted by the approach.
    - **Input Data:** The data required by the approach to initiate quality assessment.
    - **Quality Information:** The format and model used to represent mapping quality information.
    - **Refinement Capabilities:** The ability of the approach to support the resolution of detected quality issues.
    - **Interface:** Characteristics related to the interface of the approach.
- **Experiments characteristics:** Details of experiments conducted by the approach, such as metrics and data used.

The comparison of the characteristics and limitations discovered as a result of the comparisons are discussed in the following **subsections**.

2.1.7.1 Comparison of Design characteristics

Input Data: **Table 9** presents the input required by each approach in order to assess the quality of a mapping.

Table 9: Comparison of **input data** required in reviewed approaches related to mapping quality

Approach related to Input Data	**EvaMap**	**RML-Validator**	**Resglass**	**Luzzu-Extension**	**PROV-O model**	**Predictive Model**
Reused Ontologies	✓	✓	✓	✓	✓	✓
Source Data	✗	✓	✓	✗	✓	✗
Result Dataset	✓	✗	✗	✓	✗	✓

As can be seen, all approaches require the used ontologies to be input into the approach for completing the mapping quality assessment process. To achieve this, an additional step is required for users to identify and locate respective ontologies, decreasing the intuitiveness of the approaches. In addition, all of the approaches require either the source data of input mappings (3 out of 6) or an extract of the resulting dataset (3 out of 6) in order to complete the assessment process. Most input datasets are used to calculate metrics related to instances generated by the mapping. Nonetheless, requiring users to upload additional information other than a mapping increases the workload required and decreases the effectiveness of these approaches. A single piece of input should result in a more straightforward workflow. As a note, the Predictive Model is limited to mappings defined using the DBpedia ontology, as the classifier has been trained using instances defined using the ontology. Therefore, it cannot be used to classify mappings defined using other ontologies, resulting in limitations in the applicability of the approach.

Mapping language: **Table 10** presents the number and count of mapping representations targeted by the approaches.

Table 10: Mapping Languages targeted by reviewed approaches

Mapping Language	Reviewed Approaches
RML	Resglass and RML-Validator
R2RML	Luzzu-Extension
YARRRML	EvaMap
Generic	PROV-O model and Predictive Model

The approaches target various RDF-based mapping representations, which includes R2RML (1) [35], RML (2) [44] and YARRRML (1) [94]. R2RML is designed to convert relational data to RDF and is the only W3C recommendation. RML is a superset of the R2RML vocabulary and allows source data in other formats such as JSON and XML. YARRRML is the human-readable representation of RML. Sometimes approaches which target RML have support for R2RML due to vocabulary reuse. However, the PROV-O model is independent of a mapping language and can be applied to any representation. In addition, the Predictive model does not assess the mapping itself, rather it assesses two similar graphs produced by the mappings in order to detect inconsistencies.

Quality information: EvaMap, Predictive Model and Resglass represent the mapping quality information in a human readable format including a report, overall score or classification (correct or incorrect). While the other approaches (3 out of 6) represented the quality information in an ontology-based format. Of course, ontology-based formats can easily be provided in a human readable format, whereas the opposite (human readable to machine readable) is not necessarily straightforward. **Table 11** presents an overview of the format and ontologies (if applicable) used to represent captured quality information.

Table 11: Comparison of **quality information** modelling in reviewed approaches related to mapping quality

Approach	**EvaMap**	**RML-Validator**	**Resglass**	**Luzzu-Extension**	**PROV-O model**	**Predictive Model**
Ontology-Based	✕	✓	✕	✓	✓	✕
Ontology Used	-	Test-Driven	-	QPRO	PROV-O	-

The following ontologies are used to represent mapping quality information in the reviewed approaches.

- **PROV-O** [84]**:** The PROV-O based approach uses the ontology to represent mapping quality information, however, the ontology is not designed for mapping related information. Furthermore, the approach which uses it states that richer semantics are required for modelling mapping quality information. However, the resulting information is machine processable.
- **Test-Driven RDF Validation Ontology** [81]: The approach extends an existing quality assessment framework named RDFUnit [81]. Therefore, the quality reports are represented using an ontology designed to represent test cases and associated results in an RDF representation**.**
- **The Quality Problem Report Ontology (QPRO)** [36]**:** Similarly, the approach extends an existing quality assessment framework named Luzzu [36]. Therefore, the quality reports are represented using an ontology designed to represent fine-grained descriptions of quality problems found in datasets.

As can be seen, none of the approaches use an ontology designed specifically for representation of mapping quality information. Limitations of reusing other ontologies has been stated by the developers of one of the existing approaches [101], that being the ontology not being designed with the requirements of the use case in mind.

Refinement Capabilities: **Table 12** presents the capabilities provided by the approaches to resolve detected mapping quality issues.

Table 12: Comparison of **refinement capabilities** in reviewed approaches related to mapping quality

Approach	**EvaMap**	**RML-Validator**	**Resglass**	**Luzzu-Extension**	**PROV-O model**	**Predictive Model**
Refinements discussed	✕	✓	✓	✓	✓	✕
Refinement type	-	Semi-automatic	Semi-automatic	Manual	Manual	-
Implemented	-	✕	✕	✕	✕	-
Tested	-	✕	✕	✕	✕	-

The papers for 4 out of 6 of the reviewed approaches discuss the refinements that were designed to resolve the issues detected during the quality assessment process. The refinements mentioned are semi-automatic methods, which involve a human in the loop. These refinements provided insights for the design of the mapping quality assessment and refinement component of the MQI framework. None of the papers discuss fully automatic refinements (that is involving no human interaction). Interestingly, none of the approaches are reported as implementing the refinements, and so it has not possible to identify how effective they are at resolving quality issues that are identified.

User Interface: **Table 13** presents the user interface functionality of each approach.

Table 13: Comparison of **interface functionality** in reviewed approaches related to mapping quality

Approach	EvaMap	RML-Validator	Resglass	Luzzu-Extension	PROV-O model	Predictive Model
GUI	✓	✕	✕	✓	✕	✕
Custom GUI	✓	-	-	✕	-	-

As can be seen most (4 out of 6) approaches provide a Command Line Interface (CLI) rather than a Graphical User Interface (GUI) – marked with an X in the table. None of the approaches provide a GUI and CLI. EvaMap [97] provides a custom GUI while the Luzzu-Extension [75] reuses the existing GUI of Luzzu [36]. GUIs are known to improve the usability of software. However, none of the approaches with a GUI have evaluated the usability of the interface with respective end users. Therefore, none of them have provided evidence that the GUI can be effectively used.

2.1.7.2 Comparison of Experiment characteristics

Table 14 presents an overview of the experiments conducted by the approaches. The experiment characteristics outlined are the number of experiments, the type of experiment, whether they used qualitative or quantitative metrics and reproducibility of associated results.

Table 14: Comparison of **experiments conducted** for the reviewed approaches related to mapping quality

Approach	EvaMap	RML-Validator	Resglass	Luzzu-Extension	PROV-O Model	Predictive Model
Experiments Completed	✕	✓	✓	✓	✕	✓
Number of Experiments	-	1	1	1	-	1
Experiment Type	-	System	System	System	-	System
Experimental Metrics Used	-	Quantitative	Quantitative	Quantitative	-	Quantitative
Data Used	-	30 RML Mappings	100 Triple Maps	22 R2RML Mappings	-	226 DBpedia Mappings
Reproducibility	-	✓	✕	✓	-	✓

Most approaches (4 out of 6) conducted experiments which involved system testing designed to evaluate the quality assessment capabilities of them. These system experiments involving testing mappings for quality issues, which used quantitative metrics in order to measure the number of issues detected. However, none of them completed a manual examination of quality issues detected in order to ensure issues were correctly identified, and so, no qualitative metrics were used during the experiments. The experiments completed by the RML-Validator, Luzzu-Extension and Predictive Model provided the data and results to facilitate reproducibility. However, Resglass does not provide any data associated with the experiment completed. The evaluation completed by the Predictive Model involved comparing the accuracy of several supervised classification algorithms when tasked with classifying 226 DBpedia mappings as correct and incorrect. However, quality issues which resulted in the respective classifications were not discussed in the evaluation. The PROV-O model does not mention any validation methods. EvaMap completed an in-action demonstration during a demo track at ESWC, which involved participants freely interacting with the framework using a provided YARRRML [94] mapping. However, no feedback was collected from participants. As mentioned previously, no user evaluation has been published related to the reviewed approaches. In addition, the use of standardized methods is not used in any of the experiments.

2.1.7.3 Limitations Identified

Two imitations were identified through the analysis of state of art approaches of relevance to the problem of improving the quality of uplift mappings.

- **Limitation 1:** No reviewed approach modelled captured information using an ontology that was designed specifically for mapping quality information (Table 11). The limitation resulted in the formalization of research objectives **RO2(a)** and **R02(b)**, which are outlined in **Section 1.3.**
- **Limitation 2:** No reviewed approach published information on usability testing (Table 14). The limitation informed research objective **RO4**, which related to user testing of the proposed approach.

These limitations were used in the definition of the requirements for the mapping quality assessment and refinement component of the MQI framework. These requirements related to the need to propose an approach designed to assess and refine uplift mappings, while capturing associated information using an ontology specifically designed for capturing mapping quality information. In addition, the proposed approach should be evaluated with respective end users.

2.2 Approaches for supporting Linked Data Dynamics

The **second phase** of the review involved reviewing the state of the art of approaches designed to address the dynamics of LD. The analysis of these approaches resulted in the identification of limitations in existing approaches. This phase was completed prior to the design and development of the **source change detection component** of the MQI framework. By reviewing the approaches relevant information was gathered and definition of requirements for the mapping quality assessment and refinement component of the framework (designed to resolve identified limitations within the state of the art) was undertaken. The retrieval process commenced with a search for papers detailing literature reviews related to addressing the dynamics of LD, which is a focus of the research in this thesis. A literature review is defined as a "*survey of scholarly sources on a specific topic. It provides an overview of current knowledge, allowing you to identify relevant theories, methods, and gaps in the existing research that you can later apply to your paper, thesis, or dissertation topic*" [41]. It was decided to search for a seed paper in order to initiate the discovery process. A survey paper was selected as it provides a broad overview of existing research related to the domain, use cases and importantly related work. Therefore, a search of Google Scholar was completed in order to identify a suitable survey. A comparative survey was identified, which provided useful background information on the dynamics of LD and included a comparison of relevant existing approaches based on use cases derived from the community. Each of the 43 citations of the seed paper were manually examined by reading the abstracts in order to identify the relevance of the work to addressing the dynamics of LD. The process resulted in the discovery of several approaches designed to address the dynamics of LD, however, the majority of the discovered research focused on the dynamics of semantic mapping, which were not directly relevant (although heavily influenced motivation for design issues to consider). Another survey was discovered during the search [120], which detailed existing approaches for mapping maintenance, however, these approaches targeted semantic mappings. Therefore, these approaches were not considered as the focus of the research in this thesis is on uplift mappings. Each related work section in the papers of the approaches was examined in order to identify further relevant work. In addition, each of the 323 works which cited these approaches was manually examined by reading each abstract in order to identify further relevant work. Finally, a search of the proceedings of the aforementioned popular semantic web conferences was conducted in order to identify further relevant work. The final approaches were consolidated by recommendations from domain experts through various publications related to the MQI framework, which included descriptions of relevant work.

From this state of art review process 5 key approaches emerged, which focused on addressing the dynamics of LD datasets. One of these approaches (See Section 2.2.5) specifically targeted the dynamics of the source data of uplift mappings. The related work section of this paper stated *"To the best of our knowledge, the computation of changeset for RDB-RDF views has not yet been addressed in any framework."*. In addition, the citations of this paper did not provide additional existing approaches for targeting dynamics of the source data of LD. These key approaches are presented in **subsections 2.2.1** to **2.2.5**. As previously stated, the research of this thesis is focused on the preservation of alignment between uplift mappings and underlying data sources. An overview of each selected approach is presented in terms of the key characteristics which are of interest for this thesis research: target data, representation of relevant change information, notification mechanisms and interface functionalities. In addition, experiments completed on them are discussed in order to identify the level of validation of the proposed approaches.

In **section 2.2.6**, these 5 key approaches are then analyzed together with respect to the important characteristics of interest for this thesis research.

2.2.1 DyLDO

Dynamic Linked Data Observatory (DyLDO) [76] is a long-term experiment to monitor the two-hop neighborhood of 80,000 diverse LD datasets on a weekly basis. The work presents the results from the first six months of the experiment. The focus of the analysis is on the changes in the RDF resources and interlinks between the datasets and intends to address the lack of understanding related to LD dynamics.

First, use cases related to LD dynamics are presented to show why it is an important topic to study. The implementation includes a monitoring system, which has been set up for an indefinite period to capture changes on a weekly period in LD datasets. The analysis of the data collected so far included 29 weekly snapshots. These snapshots were analyzed to capture the stability of the data, frequency of changes and related details. The datasets involved in the experiment included 220 URIs available on the Datahub site in the section containing datasets retrieved from the LOD cloud. In addition, the top 220 datasets from the Billion Triple Challenge (BTC) 2011 dataset[14] were included. Thereafter, the sample size of datasets was expanded using a 2-hop breadth first search and repeating the crawl 10 times, which resulted in 95,737 dereferenceable URIs spanning 652 pay-level domains, therefore, providing a mean of 146.8 dereferenceable URIs per domain.

Weekly the content of the 95,737 URIs were retrieved and downloaded. The resulting 29 weekly snapshots consisted of the content retrieved from the core kernel URIs, the content of the expanded crawl, set of redirects, and access logs for URIs used. The URIs resulted in a mean of 68,998 RDF documents and most unique documents appearing in at least one kernel snapshot was 86,696. The documents in each snapshot were retrieved from a mean of 573.6 domains, therefore, providing diverse data. The total number of triples in the kernel snapshots was 464 million quadruples. The analysis of dynamics of the RDF documents were discussed based on the following aspects.

- **Availability:** 26% of the 86,696 documents were available for all 29 weeks of the monitoring period. 55% of the documents were available for 27 weeks or more. Mean availability was 23.1 snapshots (79.7% availability). Most error codes (32%) were 500 (Server Error) while the 96% of the rest were 404 Not found. The 1/5 unavailability rates suggests that traversing of documents can result in 20% missed content.
- **Death rate:** Dead documents refer to ones that have gone permanently offline. 95% of documents have appeared once since the 14th snapshot. It was observed that most documents that went offline were for temporary issues. However, 98.3% of URIs that returned a 404 error code never returned content again in

[14] https://km.aifb.kit.edu/projects/btc-2011/

the monitoring period. 5% of documents have returned a trailing sequence of five or more 404s or have been offline for more than 14 weeks, strongly indicating death.

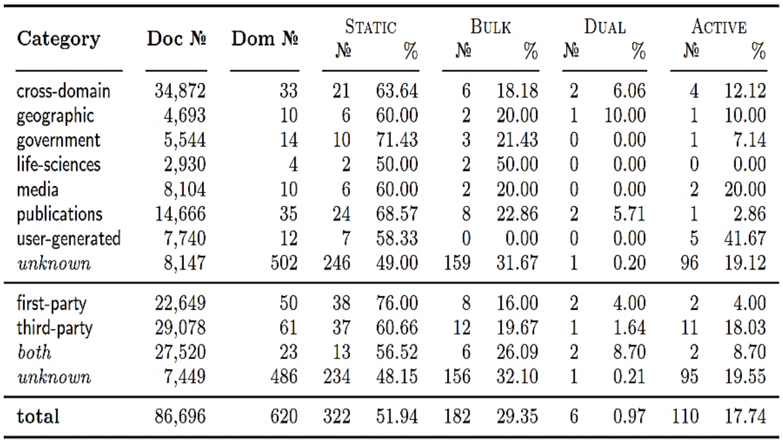
- **Change Ratio:** The change ratio was calculated by comparing the RDF content of the 28 sequential version pairs. Unavailable documents were compared with latest available version. The results showed that 62.2% of the documents did not change over the 29 weeks. The changes in the other documents were infrequently or very frequently. 23.2% were classified into the slightly dynamic interval, 8.4% were classified into the highly dynamic interval and 6.2% remaining middle interval.

Changes were also characterized based on documents within same pay-level-domain. The domains are classified as follows:

- Static domains (51.9%) contain a low ratio of documents that infrequently change.
- Bulk domains (29.4%) contain a high ratio of documents that infrequently change.
- Dual domains (1%) contain a low ratio of documents that infrequently change.
- Active domains (17.7%) contain a high ratio of documents that frequently change.

The work has observed that past dynamicity can be used to predict future dynamicity. **Table 15** presents an overview of the experiment results.

Table 15: Experiment results from DyLDO which show dynamicity of LD domains per topic and per party



Category	Doc №	Dom №	Static		Bulk		Dual		Active	
			№	%	№	%	№	%	№	%
cross-domain	34,872	33	21	63.64	6	18.18	2	6.06	4	12.12
geographic	4,693	10	6	60.00	2	20.00	1	10.00	1	10.00
government	5,544	14	10	71.43	3	21.43	0	0.00	1	7.14
life-sciences	2,930	4	2	50.00	2	50.00	0	0.00	0	0.00
media	8,104	10	6	60.00	2	20.00	0	0.00	2	20.00
publications	14,666	35	24	68.57	8	22.86	2	5.71	1	2.86
user-generated	7,740	12	7	58.33	0	0.00	0	0.00	5	41.67
unknown	8,147	502	246	49.00	159	31.67	1	0.20	96	19.12
first-party	22,649	50	38	76.00	8	16.00	2	4.00	2	4.00
third-party	29,078	61	37	60.66	12	19.67	1	1.64	11	18.03
both	27,520	23	13	56.52	6	26.09	2	8.70	2	8.70
unknown	7,449	486	234	48.15	156	32.10	1	0.21	95	19.55
total	86,696	620	322	51.94	182	29.35	6	0.97	110	17.74

Thereafter, the type of changes on the RDF level of these documents was analyzed and the results over the 29 weeks indicated:

- 27.6% of documents only updated values for terms (one per triple) in the RDF graph, such as updating a literal value and therefore, keeping the number of triples static.
- 24.0% of documents only added triples.
- The remaining (48.4%) changes involved a mix of additions, such as single term updates and deletions.

Interestingly, the bio2rdf.org domain was noted as shrinking, 52% of documents had additions and 85% had deletions. Thereafter, the types of terms changing were analyzed. The following key points were noted:

- Deletions and additions are at a similar level
- Most dynamic position of an RDF triple is the object
- Predicates occasionally added, however, rarely deleted
- Class terms rarely added and removed

Finally, the state of the links throughout the snapshots were analyzed:

- The number of links fluctuate based on availability of documents
- Small number of fresh URI links are added
- Outward link structure of the kernel remains relatively static
- Overall links are on a downward trend

The results demonstrated the dynamic nature of LD data on the web.

2.2.2 DSNotify

DSNotify (DataSet Notify) [109] is a generic change detection framework designed to detect events (create, remove, move, update) in LD datasets in order to inform data maintainers of changes, which may result in broken links between resources. **Figure 6** presents an overview of components in the DSNotify framework.

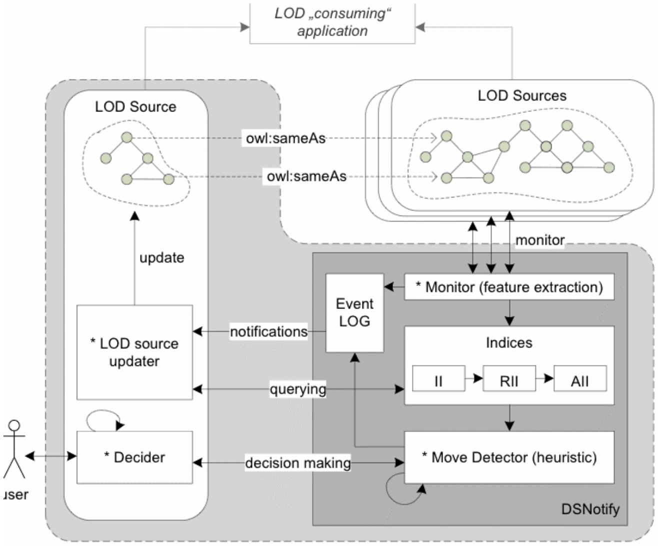


Figure 6: Components of DSNotify framework

The monitoring component periodically executes a SPARQL query on the dataset which can define a target instance type. A feature vector is created for each triple in the data retrieved from the query which is used later for detecting change events. The vectors are compared using string similarity measures with certain thresholds used to categorize data that has been added, removed and when the event has occurred. The identification of a new feature vector indicates a create event has occurred. Detected changes are stored in a central event log, which is analyzed periodically in order to identify when the framework should send a notification to subscribers. The log consists of 3 indices, which include 1) Index that represents the current state of monitored data 2) Index which stores resources that became recently unavailable and 3) index which stores old feature vectors. The process involves periodically accessing indexes 1, 2 and logging of detected events. Thereafter, the indices 1-3 are updated accordingly. Move events are detected by a housekeeper component which uses a heuristic in order to compare the representations of resources. The information captured by activities of the framework is modeled using the DSNotify EventSets vocabulary, which provides concepts and relationships to represent triples in a dataset. **Figure 7** presents the properties and classes in the DSNotify EventSets vocabulary.

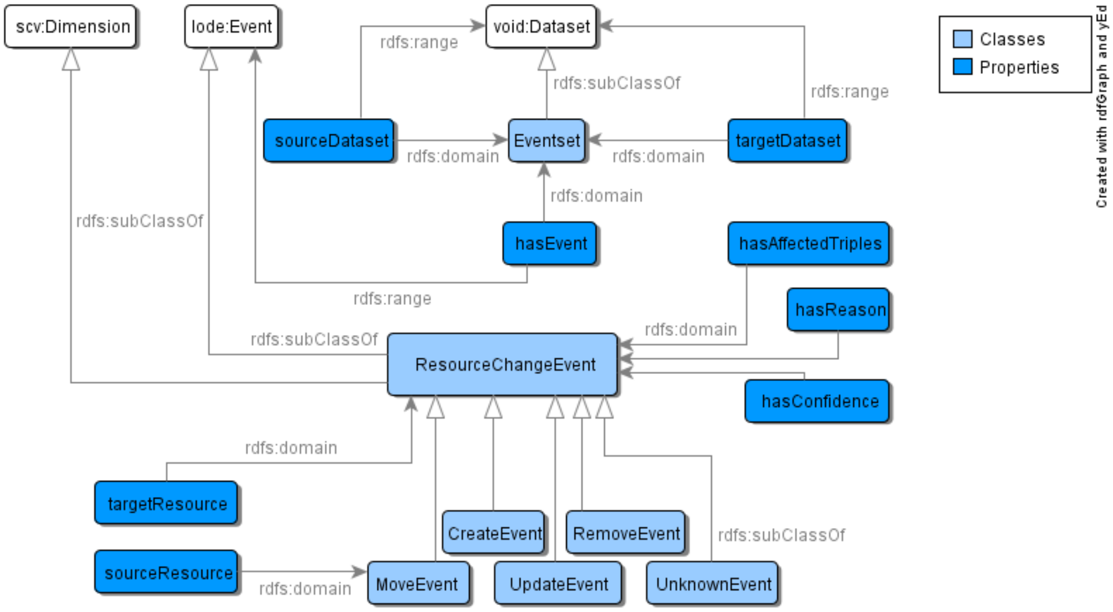


Figure 7: Concepts and properties of DSNotify EventSets Vocabulary

The vocabulary extends the LODE ontology [126] to represent the change events (`lode:Event`). Additional information related to events is captured including the changed triples, reason for the change and confidence a change has occurred. All changes related to a dataset are grouped into a set (`EventSet`). Representing changes using the vocabulary is hoped to improve the maintenance of semantic mappings rather than uplift, through the identification of changes to mapped resources. The open-source implementation of the framework is configurable, allowing plug-ins for most components in order to customize the change detection process.

2.2.3 DELTA-LD

DELTA-LD [131] is an approach which detects and classifies changes in resources and interlinks between two versions of LD datasets. The approach classifies resources that have both their IRI and representation changed. In addition, the approach aids in selecting the same resource in a different version of the data which can be used to update the dataset. The approach proposes the DELTA-LD change model (**Figure 8**) used to represent the detected changes which includes an ontology with two levels of granularity.

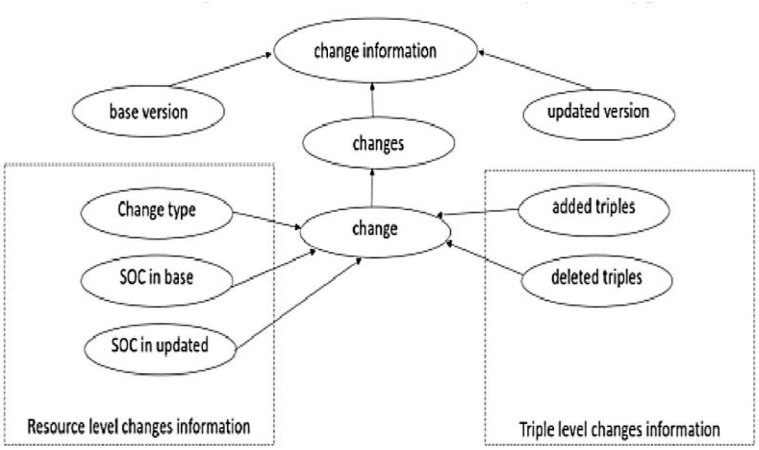


Figure 8: DELTA-LD change model

The base version refers to the original LD, while the updated version refers to the current version of the data. Change type represents the actions which resulted in the resource change (create, remove, update, move). The approach has been implemented using a triple store and database which involved four main activities outlined below:

1. **Ingestion Activity:** This activity involves uploading the RDF data into a triple store in order to allow the different datasets to be identified.
2. **Feature Extraction Activity:** This activity involves creating features for each of the resources, which are added and deleted from the uploaded datasets. Thereafter, the features are used to detect differences between the same resources with different IRIs and representations in different versions of the dataset.
3. **Change detection and Classification Activity:** This activity involves classifying the resources as updated, move or renew based on whether the representation and IRIs have changed.
4. **Transformation Activity**: This activity involves transforming the detected changes into the DELTA-LD change model.

An accuracy experiment has been conducted on the implementation and compared to existing approaches. The experiment involved applying the approach to two versions of a LD dataset. Thereafter, the results were compared to a defined gold standard. The dataset used was derived from the person snapshots of DBpedia[15] involving over 20,000 resources. The gold standard was retrieved from existing work and contained 179 move type of changes. Interestingly, the implementation discovered an additional change when compared to the gold standard. The second part of the experiment involved a larger snapshot containing over 200,000 resources, however, a gold standard was not available, therefore, it was decided to create one. The results show that changes to 296 instances were correctly identified. Thresholds were defined to categorize the probability of a detected change being correct. The results indicate that F-measure of the approach performs better than existing approaches by 4% to 6%. A case study is discussed where 100% of the invalid links were repaired.

2.2.4 SparqlPuSH

SparqlPuSH [104] is a flexible approach designed to enable the real-time notification and broadcasting of changes in RDF stores. The approach utilizes a push mechanism rather than pull where the users have to identify the new data themselves. Notifications are sent in real-time to any RSS or Atom reader. SPARQL query results are delivered through PubSubHubbub (PuSH) protocol when new RDF data is detected by the system. The approach allows users to subscribe to a subset of the content in an RDF store. The users will receive a notification message each time content in the subset has changed. The approach includes an open-source implementation which can be applied to any RDF stores supporting SPARQL. The goal of the approach is to enable proactive notification of changes in RDF stores. Initially, the setup of the approach involves the following two steps:

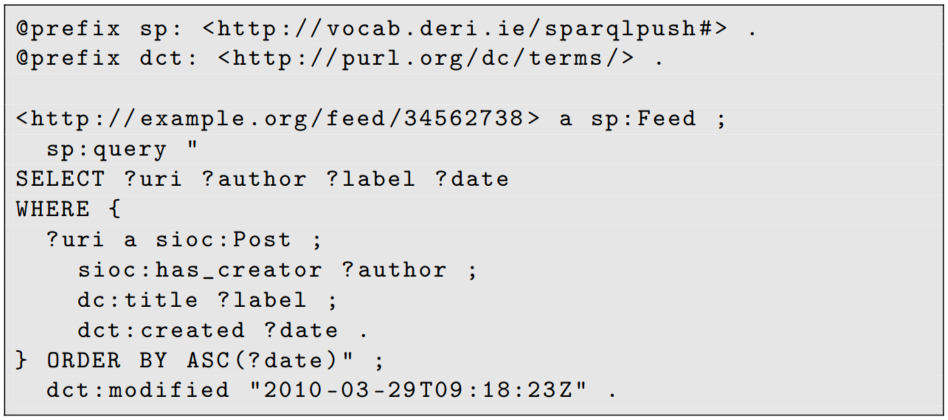
- Registering of SPARQL queries used to retrieve the relevant sub content of the RDF stores
- Broadcasting of changes related to related data

Registration of SPARQL queries involves the following steps:

- SPARQL query input into the interface or using a HTTP post query with parameters
- The query is mapped to a new feedback and information stored in a specific graph e.g. http://example.org/sparqlPuSH/feeds
- The related feedback is registered to a PuSH hub
- The PuSH hub URL is sent to client

Thereafter, the client registers its interest in the feed. **Figure 9** presents a mapping between RSS feedbacks to SPARQL query results.

[15] https://www.dbpedia.org/blog/dbpedia-snapshot-2022-03-release/s



```
@prefix sp: <http://vocab.deri.ie/sparqlpush#> .
@prefix dct: <http://purl.org/dc/terms/> .

<http://example.org/feed/34562738> a sp:Feed ;
  sp:query "
SELECT ?uri ?author ?label ?date
WHERE {
  ?uri a sioc:Post ;
    sioc:has_creator ?author ;
    dc:title ?label ;
    dct:created ?date .
} ORDER BY ASC(?date)" ;
  dct:modified "2010-03-29T09:18:23Z" .
```

Figure 9: SparqlPuSH Mapping RSS feeds to SPARQL query results

RSS or Atoms feeds to transfer the RDF from the triple to the client(s) who request change notifications. **Figure 10** presents an overview of the notification mechanism of the approach.

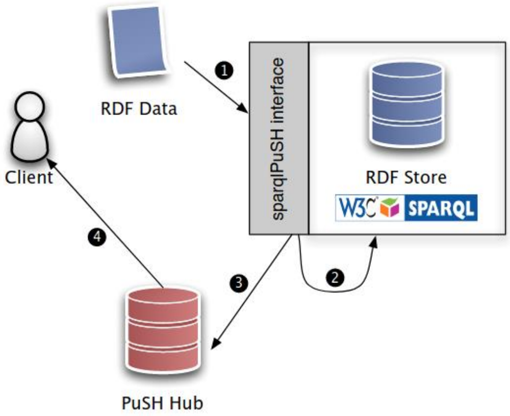
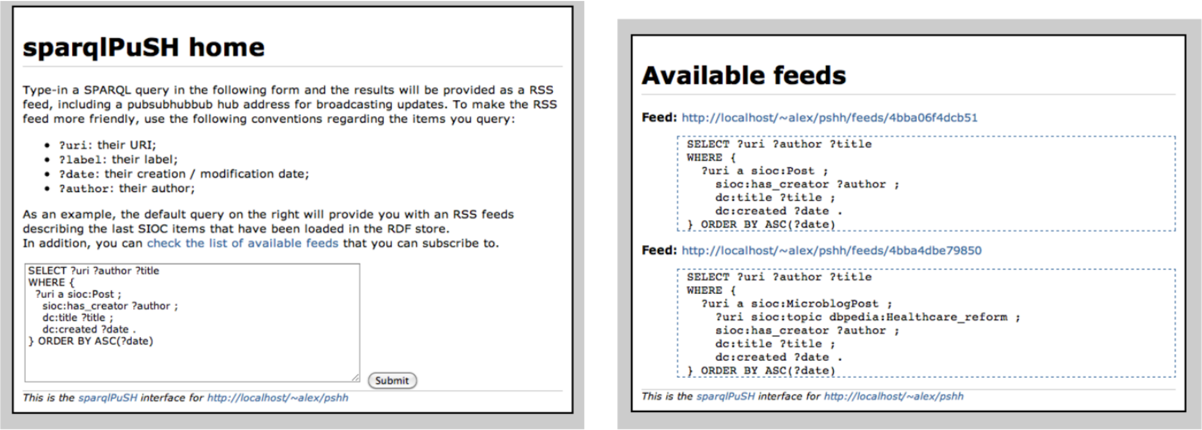


Figure 10: Overview of SparqlPuSH notification mechanism

The following steps are completed by the system once a query has been registered:

- RDF data can be added to the RDF store using the SPARQL update query via the sparqlPuSH interface
- Thereafter, all registered SPARQL queries are executed
- A notification is sent to the PuSH hub for each respective updated hub
- Thereafter, notifications are sent to all clients registered for the feedback

Experiments conducted indicated that client receives the data only a few seconds after it is loaded. The approach has been implemented in PHP, integrated with any RDF store, which can execute SPARQL queries. The GUI of the implementation includes an interface to allow 1) Query registration 2) Listing of available RSS feeds and associated SPARQL queries. **Figure 11** presents a screenshot of the related interfaces.



sparqlPuSH home

Type-in a SPARQL query in the following form and the results will be provided as a RSS feed, including a pubsubhubbub hub address for broadcasting updates. To make the RSS feed more friendly, use the following conventions regarding the items you query:

- `?uri`: their URI;
- `?label`: their label;
- `?date`: their creation / modification date;
- `?author`: their author;

As an example, the default query on the right will provide you with an RSS feeds describing the last SIOC items that have been loaded in the RDF store.
In addition, you can check the list of available feeds that you can subscribe to.

```
SELECT ?uri ?author ?title
WHERE {
  ?uri a sioc:Post ;
    sioc:has_creator ?author ;
    dc:title ?title ;
    dc:created ?date .
} ORDER BY ASC(?date)
```

This is the sparqlPuSH interface for http://localhost/~alex/pshh

(a)

Available feeds

Feed: http://localhost/~alex/pshh/feeds/4bba06f4dcb51

```
SELECT ?uri ?author ?title
WHERE {
  ?uri a sioc:Post ;
    sioc:has_creator ?author ;
    dc:title ?title ;
    dc:created ?date .
} ORDER BY ASC(?date)
```

Feed: http://localhost/~alex/pshh/feeds/4bba4dbe79850

```
SELECT ?uri ?author ?title
WHERE {
  ?uri a sioc:MicroblogPost ;
    ?uri sioc:topic dbpedia:Healthcare_reform ;
    sioc:has_creator ?author ;
    dc:title ?title ;
    dc:created ?date .
} ORDER BY ASC(?date)
```

This is the sparqlPuSH interface for http://localhost/~alex/pshh

(b)

Figure 11: SparqlPuSH interface to (a) register queries (b) list available queries

No user testing has been completed on the implementation.

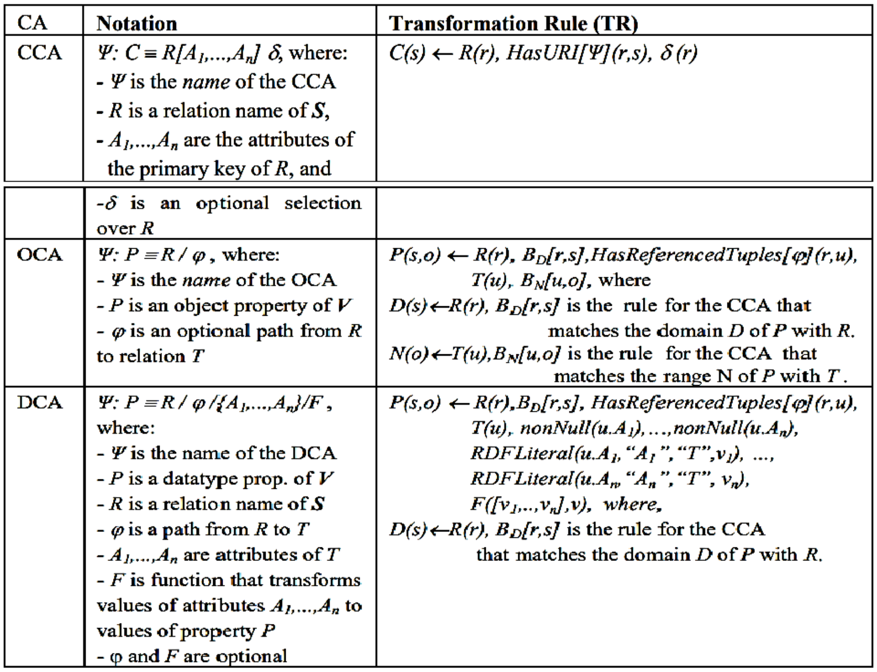
2.2.5 Mapping Changeset

An approach [141] which proposes a framework for supporting alignment between relational databases and RDF views. The approach focuses on R2RML mappings [35], which are designed to transform relational data. Changesets which contain information used to detect differences between two versions of datasets are computed by the framework to support alignment. The changesets are automatically computed using mappings, which transform instance data from a relational database into a target ontology. The formalism has been described as a much simpler language than R2RML [35].

A more concise abstract syntax, based on correspondence assertions (CA) are also described to transform a relational database to a target ontology. The RDB-RDF views addressed focus on schema-directed RDF publishing. Therefore, the correspondence assertions are used to induce schema mappings defined by the class of queries. The CA's are capable of capturing all R2RML mapping patterns found in the state of the art. The three types of CA's which exist are presented in **Table 16** and included:

- **Class Correspondence Assertions (CCA):** Maps relations to class instances.
- **Object Property Correspondence Assertions (OCA):** Maps relationships, through a path, to property instances.
- **Datatype Property Correspondence Assertions (DCA):** Maps attribute values to values of datatype properties.

Table 16: Transformation Rules used to compute changesets

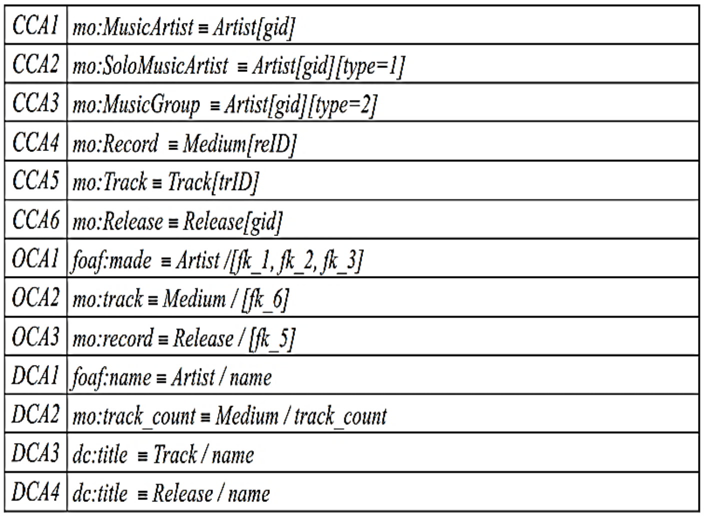


CA	Notation	Transformation Rule (TR)
CCA	Ψ: $C \equiv R[A_1,\ldots,A_n]\ \delta$, where: - Ψ is the *name* of the CCA - R is a relation name of $\boldsymbol{S}$, - $A_1,\ldots,A_n$ are the attributes of the primary key of R, and -δ is an optional selection over R	$C(s) \leftarrow R(r),\ HasURI[\Psi](r,s),\ \delta(r)$
OCA	Ψ: $P \equiv R\ /\ \varphi$, where: - Ψ is the *name* of the OCA - P is an object property of V - φ is an optional path from R to relation T	$P(s,o) \leftarrow R(r),\ B_D[r,s], HasReferencedTuples[\varphi](r,u),$ $T(u),\ B_N[u,o]$, where $D(s) \leftarrow R(r),\ B_D[r,s]$ is the rule for the CCA that matches the domain D of P with R. $N(o) \leftarrow T(u), B_N[u,o]$ is the rule for the CCA that matches the range N of P with T.
DCA	Ψ: $P \equiv R\ /\ \varphi\ /\{A_1,\ldots,A_n\}/F$, where: - Ψ is the name of the DCA - P is a datatype prop. of V - R is a relation name of $\boldsymbol{S}$ - φ is a path from R to T - $A_1,\ldots,A_n$ are attributes of T - F is function that transforms values of attributes $A_1,\ldots,A_n$ to values of property P - φ and F are optional	$P(s,o) \leftarrow R(r), B_D[r,s],\ HasReferencedTuples[\varphi](r,u),$ $T(u),\ nonNull(u.A_1),\ldots,nonNull(u.A_n),$ $RDFLiteral(u.A_1, \text{“}A_1\text{”}, \text{“}T\text{”}, v_1),\ \ldots,$ $RDFLiteral(u.A_n, \text{“}A_n\text{”}, \text{“}T\text{”}, v_n),$ $F([v_1,..,v_n],v)$, *where,* $D(s) \leftarrow R(r),\ B_D[r,s]$ is the rule for the CCA that matches the domain D of P with R.

The rule is induced by a CA Ψ, where R is a Pivot relation of Ψ and r the pivot variable. The approach proposes to automatically generate R2RML mappings based on correspondence assertions with the relational views as a middle layer and has been applied to the relational data and respective R2RML mappings used in the MusicBrainz project, which is an open encyclopedia containing music metadata. CA's were created that specify a mapping between the MusicBrainz relational schema and target ontology.

Table 17 presents the CA generated for the use case.

Table 17: Correspondence Assertions used by mapping changeset approach

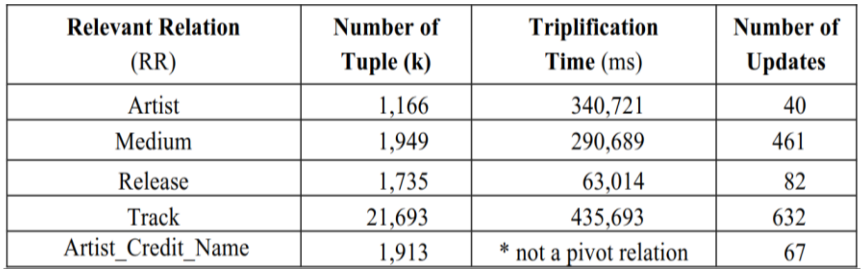


CCA1	mo:MusicArtist ≡ Artist[gid]
CCA2	mo:SoloMusicArtist ≡ Artist[gid][type=1]
CCA3	mo:MusicGroup ≡ Artist[gid][type=2]
CCA4	mo:Record ≡ Medium[reID]
CCA5	mo:Track ≡ Track[trID]
CCA6	mo:Release ≡ Release[gid]
OCA1	foaf:made ≡ Artist /[fk_1, fk_2, fk_3]
OCA2	mo:track ≡ Medium / [fk_6]
OCA3	mo:record ≡ Release / [fk_5]
DCA1	foaf:name ≡ Artist / name
DCA2	mo:track_count ≡ Medium / track_count
DCA3	dc:title ≡ Track / name
DCA4	dc:title ≡ Release / name

The CA’s were used to generate a materialization of the MusicBrainz RDF view from the current relations. Thereafter, the strategy is based on rules, which are used for computation of changesets when updates are detected in the source database using defined triggers.

The approach has been implemented as the LinkedBrainz Live tool (LBL tool). The tool is designed to propagate updates in the MusicBrainz database into the LinkedMusicBrainz view, which publishes music metadata in LD format. A local database is used to install a replica of the MusicBrainz database. R2RML mappings are used for transforming the data into a LD representation expressed in the Music ontology. An update extractor is used to monitor for changes in the database. The changes are propagated into the LD using INSERT/DELETE statements. An experiment has been conducted on the implementation, which involved computing the changesets for 4,069 updates in the data. It took 16 minutes in total to compute these changesets. **Table 18** presents an overview of the experiment results.

Table 18: Overview of experiment results of the mapping changeset approach



Relevant Relation (RR)	**Number of Tuple (k)**	**Triplification Time** (ms)	**Number of Updates**
Artist	1,166	340,721	40
Medium	1,949	290,689	461
Release	1,735	63,014	82
Track	21,693	435,693	632
Artist_Credit_Name	1,913	* not a pivot relation	67

The results indicated an incremental strategy outperforms full re-materialization in cases where changes are frequent.

2.2.6 Analysis of LD Dynamics SoA

This section compares the reviewed approaches presented in **sections 2.2.1** to **2.2.5**, based on characteristics related to the research question of this thesis. The characteristics represent common features, which were noted during the review of the approaches. The characteristics provide a basis for comparison between them. In addition, experiments conducted on the approaches were compared to provide insights into the design of the evaluation strategy of the MQI framework. Thereafter, identified limitations with the approaches were used to inform requirements for the design and implementation of the **source change detection component** of the MQI framework. The characteristics used for the comparison are outlined below.

- **Design characteristics**
    - **Target Data**: Data related to LD targeted by the approach.
    - **Change Information:** The format and model used to represent mapping quality information.
    - **Notification Mechanism:** The ability of the approach to send a notification detailing change information.

- **Interface:** Characteristics related to the interface of the approach.
- **Experiments characteristics:** Details of experiments conducted by the approach, such as metrics and data used.

The comparison of the characteristics and limitations discovered as a result of reviewing the approaches are discussed in the following **subsections**.

2.2.6.1 Comparison of Design characteristics

Target Data: **Table 19** presents the data related to LD datasets which was targeted by the approaches.

Table 19: Data targeted by reviewed approaches

Targe Data	Reviewed Approaches
Resources only	SparqlPuSH
Interlinks only	None
Resources and Interlinks	DSNotify, DyLDO and DELTA-LD
Source Data	Mapping Changeset

Most (3 out of 5) approaches target resources and interlinks in LD. Only one approach specifically targets resources, while none specifically target interlinks between LD datasets. Only one approach takes into account the source data. These results indicated there is a lack of approaches to address the dynamics of the source data.

Change Information: **Table 20** presents an overview of the change information representation of the approaches.

Table 20: Comparison of **change modelling** in reviewed approaches related to LD dataset dynamics

Approach	DyLDO	DSNotify	DELTA-LD	SparqlPuSH	Mapping Changeset
Ontology based	✕	✓	✓	✕	✕
Ontology Used	-	DSNotify EventSets	DELTA-LD Model	-	-

Some approaches (2 out of 5) capture change information in an ontology-based format. All of these approaches use ontologies designed specifically for the use case. The mapping changeset approach [141] which targets the dynamics of the source does not capture information in an ontology-based format. No ontology could be found to represent source data changes.

Notification Mechanism: **Table 21** presents an overview of the notification mechanism of the approaches.

Table 21: Comparison of **notification mechanism** in reviewed approaches related to LD dataset dynamics

Approach	DyLDO	DSNotify	DELTA-LD	SparqlPuSH	Mapping Changeset
Notification Mechanism	✕	✓	✕	✓	✕
Communication Method	-	System	-	RSS Feeds	-

Most approaches (3 out of 5) do not support notifications of detected changes. The approaches (2 out of 5) that support notifications, send them via feeds or custom methods. The notifications provide information on resources which have changed in LD datasets. Most approaches can be seen to be limited in the timely delivery of relevant change information and rather require the users themselves to seek out the information manually.

Interface: **Table 22** presents an overview of the user interfaces of the approaches.

Table 22: Comparison of **interface functionality** in reviewed approaches related to LD dataset dynamics

Approach	**DyLDO**	**DSNotify**	**DELTA-LD**	**SparqlPuSH**	**Mapping Changeset**
GUI	✕	✕	✕	✓	✕
Custom	-	-	-	✓	-

SparqlPuSH includes a basic GUI (Figure 11) to support user interaction, which has not been reused from an existing system. All other approaches support user interaction through a CLI. Therefore, users are required to be familiar with commands to setup and run the implementation.

2.2.6.2 Comparison of Experiment characteristics

Table 23 presents an overview of the experiments completed on the approaches.

Table 23: Comparison of **experiments conducted** in reviewed approaches related to LD dataset dynamics

Approach	**DyLDO**	**DSNotify**	**DELTA-LD**	**SparqlPuSH**	**Mapping Changeset**
Experiments Completed	✓	✓	✓	✓	✓
Number of Experiments	1	1	1	1	1
Experiment Type	System	System	System	System	System
Experiment Metrics Used	Quantitative	Quantitative	Quantitative	None	Quantitative
Data Used	Real world	Sample	Real world	Sample	Real world
Reproducibility	✓	✓	✓	✕	✓

An experiment has been reported on all of the reviewed approaches. Most of the approaches (4 out of 5) used experiment metrics to measure different aspects, such as accuracy which was measured using F-score[16] by one of the approaches (DELTA-LD). Some (2 out of 5) of the approaches used simulated data during the experiments, while others (3 out of 5) used real world data, which was collected from sources such as the LOD cloud. Real world data has the benefit of demonstrating an approach is applicable to real world scenarios. However, none of the experiments involved user testing, therefore, they have not demonstrated their usability with respective end users.

[16] https://deepai.org/machine-learning-glossary-and-terms/f-score

2.2.6.3 Limitations identified

Two limitations were identified through the analysis of state of art approaches of relevance to the problem of LD dataset dynamics.

- **Limitation 1:** No reviewed approach models the changes in heterogeneous source data formats. In addition, no approach captures information using an ontology designed specifically for source data changes (Table 20) and supports notifications (Table 21) for these changes. This limitation informed the formulation of research objectives **RO3(a)** and **R03(b)**, which are outlined in **Section 1.3.**
- **Limitation 2:** No reviewed approach has conducted usability testing (Table 23). The limitation informed the formulation of research objective **R04**, which related to user testing of the proposed approach.

The limitations were used to inform the requirements for the source change detection component of the MQI framework. These requirements related to a proposed approach designed to detect changes in the source data associated with a LD dataset through uplift mappings, while also capturing associated information using an ontology designed for source data changes and providing a notification mechanism for a data maintainer to prompt possible updating of the uplift mapping artefacts. In addition, the proposed approach which has been tested with respective end users.

2.3 Summary findings

This chapter presented the results of the two phases of state of the art review. The first phase reviewed existing approaches designed to support the quality assessment and refinement of uplift mappings. The second phase reviewed existing approaches designed to address the dynamics of LD datasets and underlying data sources. The key findings from the state of the art review are summarized below.

- (***Limitation 1***) **Representation of Mapping Quality:** None of the reviewed approaches designed to support mapping quality assessment and refinement captured provenance from the associated activities in an ontology designed for representing information related to quality of mappings, and so the semantic representation of mapping quality information may be considered as limited.
- (***Limitation 2***) **Lack of User Testing #1:** None of the reviewed approaches designed to support mapping quality assessment and refinement published details of what (if any) formal user evaluation undertaken, and it is not clear if they have been validated with respective end users.
- (***Limitation 3***) **Representation of Source Data Changes:** None of the reviewed approaches designed to support preservation of alignment between mappings and respective source data captured provenance from the associated activities in an ontology designed for source data changes, and so the semantic representation of change information may be considered as limited.

- (***Limitation 4***) **Lack of User Testing #2:** None of the reviewed approaches designed to support preservation of alignment between mappings and source data published details on what (if any) user testing has been conducted, and so it is not clear if they have not been validated with respective end users.

The findings arising from the first phase of the state of the art review were further explored in a preliminary study, which involved the design and implementation of an initial mapping quality assessment and refinement framework. This preliminary study is presented in **Section 3.3**.

The next chapter (Chapter 3) presents background information on the uplift mapping lifecycle, including representations commonly used to define uplift mappings. In addition, the design and limitations of the framework developed in the preliminary study are discussed.

Chapter 3: Background

This chapter describes the stages involved in the mapping lifecycle, including the software agents and artefacts involved. **Section 3.1** discusses the uplift mapping lifecycle. **Section 3.2** presents popular mapping representations in the state of the art designed to represent uplift mappings. **Section 3.3** presents a preliminary study undertaken to explore mapping quality area, which involved the creation of a framework designed to improve the quality of uplift mappings.

3.1 Uplift Mapping Lifecycle

A mapping lifecycle breaks down the mapping process into phases [3,39]. The uplift mapping lifecycle involves the definition of uplift mappings, which allows one to declaratively express the transformations required to convert non-RDF data to RDF. Mappings also detach mapping definitions from the implementations that executes them, which facilitate the reproduction of the process responsible for the generation of datasets (when updating datasets, or creating new ones, where mappings – or parts of – can be reused). The mapping lifecycle is concerned with all the activities executed by stakeholders with the goal of producing a set of relations, or mappings. For instance, stakeholders should identify and analyze requirements taking into consideration inputs to be mapped, including relevant data and vocabularies. **Figure 12** presents the activities involved in the five high level phases [39] of the uplift mapping lifecycle, which includes **stage**, **characterize**, **reuse**, **mapping** and **execution**, involving multiple processes and stakeholders with varying background knowledge. These five phases are discussed below.

Stage: This phase consists of the definition of the project, stakeholders, scope and requirements, among others. Defining the **scope** of the project involves the title, goal and involved stakeholders, such as semantic web experts and business analysts, among others. **Requirements** are designed to represent the objectives and expected functionality of the project. The defined requirements are used to guide the progression of the project. The **outcome** of this phase is artefacts detailing the scope and requirements. Poor requirements defined at this stage will cascade quality issues throughout the following phases.

Characterize: This phase involves analysis of relevant information in order to identify the feasibility of the project. The rationale for analyzing relevant data, vocabularies and tools early in the process is to ensure the feasibility before the definition of mappings commences. The **data analysis** involves generating an artefact with the data to be used during the mapping process. **Vocabulary analysis** involves generating an artefact with the vocabularies which will be used to define a mapping. New vocabularies will be defined if required. **Tool analysis** will generate an artefact detailing a set of relevant tools. These tools will have a variety of functions, such as cleaning, data transformations and/or for the actual process of mapping source data. **Feasibility analysis** is completed by comparing the artefacts created in this stage as well as the requirements to ensure the project is viable. The phase results in one of **two outcomes**, which includes 1) project is feasible and proceeds to the next phase and 2) project is not feasible and must be redefined in the **stage** phase. Poor decisions related to the data, vocabularies and tools will drastically decrease the workflow involved the creation of high-quality RDF data.

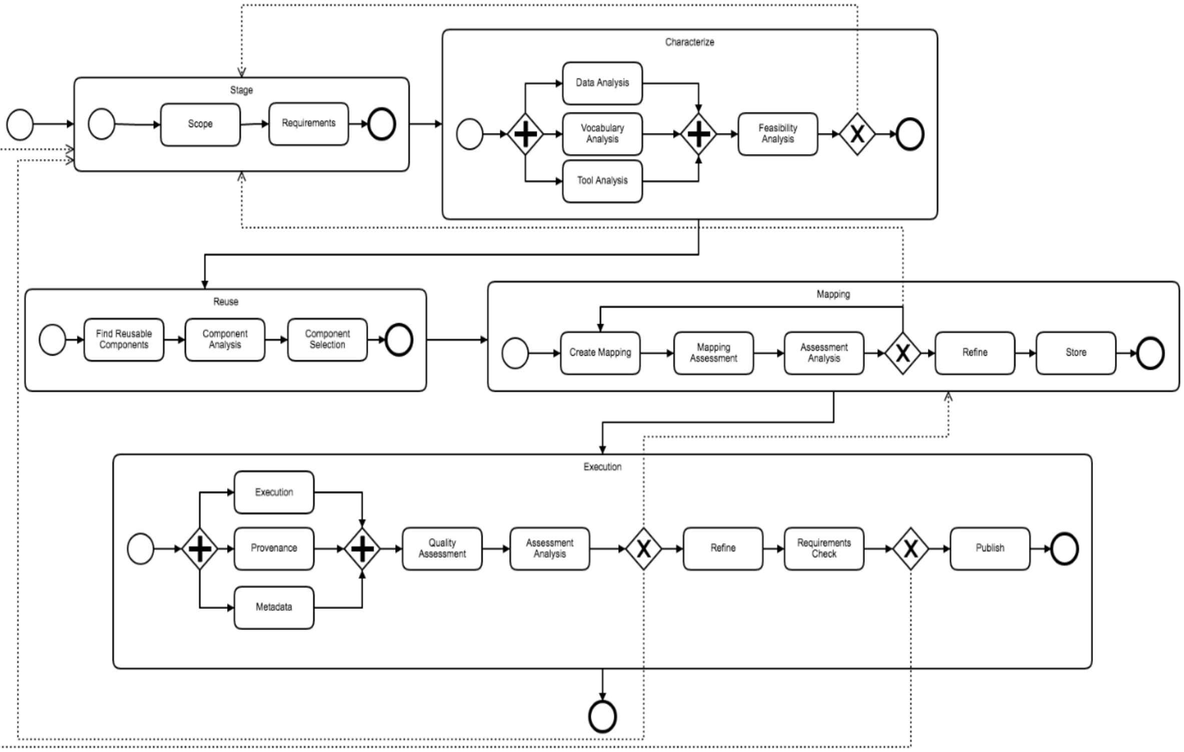


Figure 12: **Uplift mapping lifecycle** (Retrieved from [31])

Reuse: This phase is responsible for finding, analyzing and selecting reusable components by researching related mapping projects. The requirements will guide the search for related projects. The mapping engineer is tasked to **find reusable components**, which involves identifying a set of components to be used in the next phase. Potentially, no reusable components could be found, however, the analysis must be completed in order to satisfy open world assumption. **Component analysis** involves creating an artefact detailing suitability of reuse for each identified component. The result of the analysis is **component selection** used to identify the components which will be reused. The **outcome** of this phase is the set of selected components to be used in the next phase. Poor design decisions during this phase will result in inconsistencies in semantics due to reuse of unsuitable components and vocabularies, therefore, resulting in quality issues in the following phases.

Mapping: The **create mapping** process of this phase involves defining a declarative mapping represented in formats such as those described in **Section 3.2**. The defined mapping will reuse components identified in the previous phase. Thereafter, **mapping assessment** is completed in two procedures. First, the syntax of the mapping is validated against conformance to the mapping specification. Second, the semantics of the mapping are validated in order to identify logical inconsistencies. Quality assessment can be completed by frameworks designed to target uplift mappings, which are described in **Section 2.1**. The result of the assessment process is a report detailing potential quality issues. An **assessment analysis** is completed on the report in order to identify suitable refinements. Thereafter, the **refine** process results in the creation of a refined mapping, which is generated as a result of quality issues being removed from the original mapping. Three **outcomes** are possible for this phase, which includes 1) mapping quality is sufficient and executed 2) mapping quality is not sufficient for execution and is refined in create mapping process and 3) project is redefined in stage phase. Thereafter, the RDF data created from execution of the uplift mapping is shared with the community of stakeholders. The execution of poor-quality issues results in quality issues exponentially multiplying in the resulting RDF dataset, therefore, greatly decreasing the quality of LD (See **Table 2**) [42,65,75,97]. An effective mapping quality assessment and refinement is essential in order to eradicate the root cause of quality, resulting in higher quality data for involved stakeholders.

Execution: This phase is responsible for generating, validating and publishing the RDF dataset generated as a result of **execution** of the declarative uplift mapping defined in the previous phases. **Provenance** and **metadata** related to the process is captured in order to improve traceability. In this sense, provenance is a subset of metadata, captured separately, which relates to trustworthiness of the dataset. **Quality assessment** of the initial dataset is completed in order to detect quality issues, which provide indications of unfulfilled requirements. Thereafter, **assessment analysis** of the detected issues results in three possible **outcomes**, which includes: 1) the dataset can be refined and published 2) problems must be corrected during mapping phase 3) or in stage phase. If the problems should be addressed in this phase, the process **refine** starts, resulting in the creation of the final dataset. Before publishing the dataset, requirements of the project defined in stage must be validated in the **requirement check** process. The **outcome** of this phase is either 1) the dataset is published in the **publish** process on the condition that all project requirements have been fulfilled or 2) unfulfilled requirements identified are redefined in the stage phase. It is noted that publishing a dataset involves more tasks that are out of scope of the uplift mapping lifecycle. Quality issues removed as a result of dataset refinement, which originate in the mapping definitions could be present when the mapping is regenerated, as the root cause of the issue has not been addressed. Therefore, it highlights the importance of eradicating quality issues early in the publication process. In addition, it is important to maintain mapping quality after publication in order to prevent a decrease of quality and alignment with underlying data sources [141]. The uplift mapping lifecycle involves several complex and error prone tasks [65,75], involving multiple stakeholders. Therefore, it is essential to validate artefacts resulting from the process, which decreases the likelihood of quality issues being present in published datasets.

3.2 Uplift Mapping representations

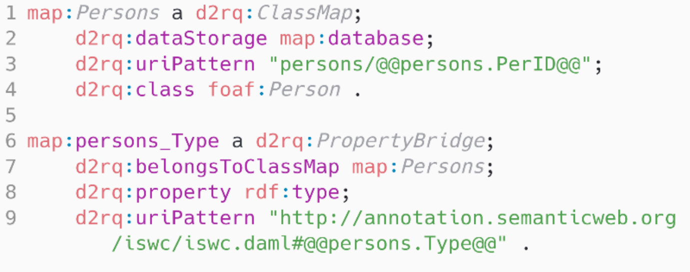
The uplift mapping defined as a result of the activities involved in the mapping process can be expressed in various representations. These mappings contain transformation rules designed to convert input (non-RDF and vocabularies) into RDF representation. This section presents prominent uplift mapping languages, which are relevant to the research described in this thesis.

3.2.1 Direct Mapping

Direct mapping [5] is a W3C recommendation for creating a direct RDF representation of data stored in a relational database (data and schema). The companion of direct mapping is R2RML [35] described in the **Section 3.2.3**. The main difference is that direct mapping is a default mapping language, which does not allow the definition of customized transformation rules, therefore, existing vocabularies cannot be reused as a vocabulary is derived from the schema of the relational database. Direct mapping supports foreign keys in databases by including the reference of the key and values from respective rows.

3.2.2 D2R

The Database to RDF Mapping Language (D2R) [11] is an approach designed for publishing relational databases in RDF format. The D2RQ mapping language is used in the approach for the transformation of relational data. **Listing 1** presents an uplift mapping represented in the D2RQ language.



```
map:Persons a d2rq:ClassMap;
    d2rq:dataStorage map:database;
    d2rq:uriPattern "persons/@@persons.PerID@@";
    d2rq:class foaf:Person .

map:persons_Type a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:Persons;
    d2rq:property rdf:type;
    d2rq:uriPattern "http://annotation.semanticweb.org
       /iswc/iswc.daml#@@persons.Type@@" .
```

Listing 1: Sample mapping represented in **D2RQ** (Retrieved from ISWC examples[17])

The mapping consists of a class map (`d2rq:ClassMap`) which defines the input relational database (`d2rq:dataStorage`), types of the subject of the RDF triples (`d2rq:class`) and a pattern to generate its URI (`d2rq:uriPattern`). A predicate property (`d2rq:PropertyBridge`) is used to relate a column (`d2rq:column`), which represents the object to the predicate (`d2rq:property`) of the triple. In this case an instance of type `foaf:Person` is generated by the mapping.

3.2.3 R2RML

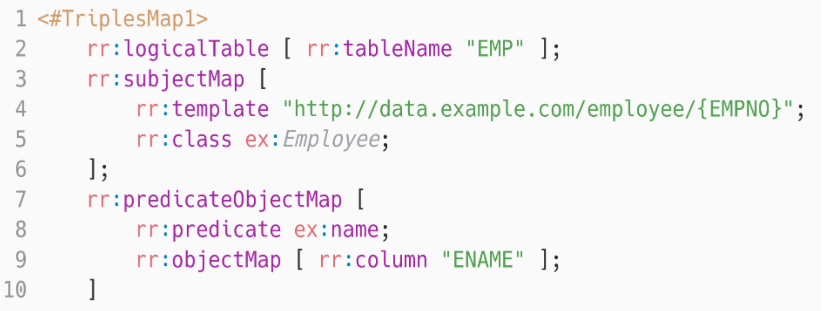
The RDB to RDF Mapping Language (R2RML) [35] is the W3C recommendation for transformation of relational data into RDF representation. Unlike direct mapping, R2RML enables the definition of customized transformation rules,

[17] Sample D2RQ uplift mapping retrieved from https://github.com/d2rq/d2rq/blob/master/doc/example/mapping-iswc.ttl

therefore, allowing the reuse of existing ontologies. In addition, the language is RDF based, allowing them to be automatically processed by machines.

- **Logical Table:** A valid schema-quality name referencing an existing base table, SQL query view in the input relational database is defined.
- **Subject Map:** The subject of the RDF triples, which will be generated can be IRIs or blank nodes and include zero or more class types. Templates are used to define unique IRIs.
- **Predicate Object Map:** The predicates and objects of the RDF triples are defined as zero or more predicate object maps in each triple map. Predicate maps must contain one or more predicates with valid IRIs. Objects can be defined as resources, blank nodes or literal values. Literals may be assigned a language tag or data type.
- **Graph Map:** A subject map or predicate-object map may have zero or more associated graph maps, which are term maps designed to insert RDF triples into a named graph.

Listing 2 presents an uplift mapping represented in R2RML designed to uplift employee information, including their number and name.



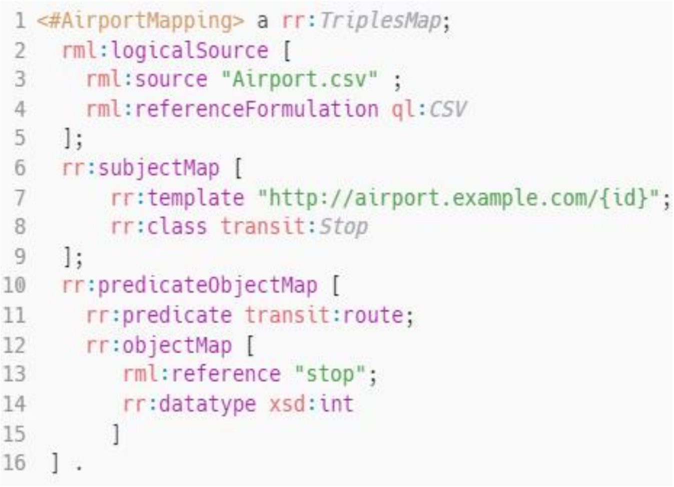
```
<#TriplesMap1>
    rr:logicalTable [ rr:tableName "EMP" ];
    rr:subjectMap [
        rr:template "http://data.example.com/employee/{EMPNO}";
        rr:class ex:Employee;
    ];
    rr:predicateObjectMap [
        rr:predicate ex:name;
        rr:objectMap [ rr:column "ENAME" ];
    ]
```

Listing 2: Sample mapping represented in **R2RML** (Retrieved from [35])

The source data (`rr:logicalTable`) of the mapping is a table (`rr:tableName`) in a relational database, which contains the employee information. The subjects of the RDF triples generated by the mapping includes a class type (`rr:class`). One predicate object map is defined, which includes the predicate (`rr:predicate`) used to relate the employees name to the column (`rr:column`) in the database.

3.2.4 RML and YARRRML

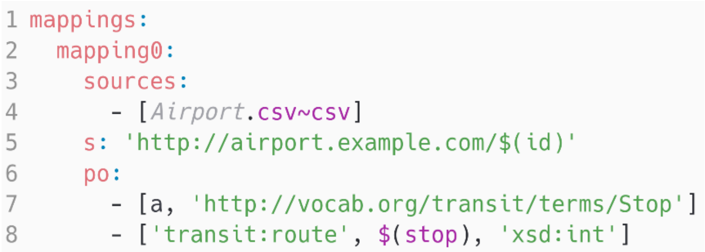
The RDF Mapping Language (RML) [35] extends R2RML by providing support for additional source data formats, such as CSV, XML, JSON, among others. RML has extended the R2RML vocabulary to include five additional properties designed to represent these formats and methods to reference the data, including predefined iterators. **Listing 3** presents an uplift mapping represented in RML designed to transform information related to transport routes for airports.



```
<#AirportMapping> a rr:TriplesMap;
  rml:logicalSource [
    rml:source "Airport.csv" ;
    rml:referenceFormulation ql:CSV
  ];
  rr:subjectMap [
      rr:template "http://airport.example.com/{id}";
      rr:class transit:Stop
  ];
  rr:predicateObjectMap [
    rr:predicate transit:route;
    rr:objectMap [
       rml:reference "stop";
       rr:datatype xsd:int
      ]
 ] .
```

Listing 3: Sample mapping represented in **RML** (Retrieved from [44])

As can be seen, the main difference compared to an R2RML mapping is the method of referencing the input source data (`rml:logicalSource`) and data to be transformed (`rml:reference`). The mapping shown transforms source data in CSV format (`rml:source`), which contains information related to transport routes. A human readable text-based representation of RML is available, named YARRRML [94]. The representation was created using YAML Ain't Markup Language (YAML) [8], which is a prominent serialization language designed to be easily read by humans. The objective of this language is to provide a method for non-experts to define uplift mappings. **Listing 4** presents the YARRRML generated mapping as a result of converting the RML mapping presented in Listing 3.



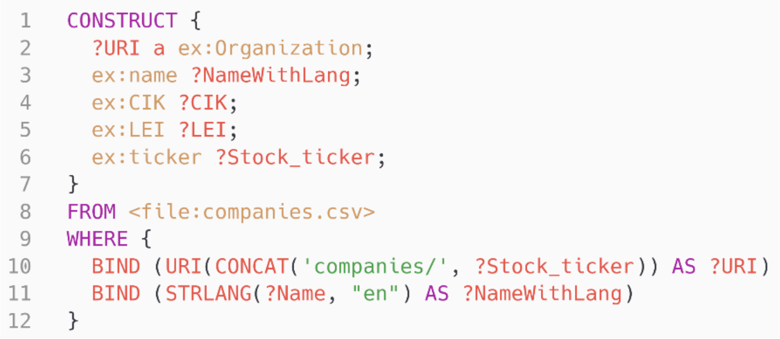
```
mappings:
  mapping0:
    sources:
      - [Airport.csv~csv]
    s: 'http://airport.example.com/$(id)'
    po:
      - [a, 'http://vocab.org/transit/terms/Stop']
      - ['transit:route', $(stop), 'xsd:int']
```

Listing 4: Sample mapping represented in **YARRRML** (Retrieved from [94])

As can be seen, the location of the source data (`sources`) is defined in a list. The RDF triples are represented similar in the format of subject (`s`) and grouped predicate object maps (`po`). While the representation is more easily understood by humans, it is not ontology-based similar to R2RML and RML. Therefore, these mappings cannot be linked with the resulting dataset in order to improve trustworthiness by providing indications of suitable for the applications of consumers.

3.2.5 Tarql

Unlike the aforementioned approaches, the Tables for SPARQL (Tarql) [157] mapping language is SPARQL based rather than RDF based [71], which is designed to transform CSV files into RDF representation. The language supports the full manipulation of the input data using SPARQL functions, such as comparison operators and string manipulation. **Listing 5** presents an uplift mapping represented in Tarql designed to transform the stock prices of companies.



```
CONSTRUCT {
  ?URI a ex:Organization;
  ex:name ?NameWithLang;
  ex:CIK ?CIK;
  ex:LEI ?LEI;
  ex:ticker ?Stock_ticker;
}
FROM <file:companies.csv>
WHERE {
  BIND (URI(CONCAT('companies/', ?Stock_ticker)) AS ?URI)
  BIND (STRLANG(?Name, "en") AS ?NameWithLang)
}
```

Listing 5: Sample mapping represented in **Tarql** (Retrieved from [157])

A SPARQL construct (`CONSTRUCT`) query is used to generate RDF triples related to a company (`ex:Organization`) which are generated from a CSV file (`<file:companies.csv>`). In addition, two bind (`BIND`) statements are used to manipulate the URI (`?URI`) of the subject of the generated triples.

3.3 Preliminary Study

A preliminary study was completed prior to the development of the mapping quality assessment and refinement component of the MQI framework. The study provided useful insights and findings that informed the subsequent design and implementation of the MQI framework. In addition, the study provided relevant background information on common quality issues, which impact mapping and LD dataset quality. The preliminary study also helped to identify requirements necessary for an effective mapping quality improvement approach.

The preliminary study was published: *Randles, A., Junior, A. C., & O'Sullivan, D. (2020).* ***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 and Services (iiWAS '20), Virtual Event, 2020 (pp. 347–351).*

The publication presents an overview of the design and implementation of the framework. In addition, a demonstration walkthrough, which describes the quality assessment and refinement of a sample uplift mapping is presented. The preliminary study involved the design and implementation of a mapping quality assessment and refinement framework, named the SHACL-based framework. The framework utilized SHACL [80] for quality assessment, which is a W3C recommendation for validation of RDF graphs. SHACL constraints are used for mapping quality validation and captured quality information is expressed in the SHACL validation report vocabulary[18]. It was decided to target R2RML [35] mappings as it is the W3C recommendation for transforming relational data into RDF. In addition, the mapping representation is RDF based [7], which enables SHACL to validate the mapping graph. The framework facilitated semi-automatic refinements by providing suggestions and values designed to resolve specific issues. SPARQL [61] update queries were used to refine the mapping as SHACL does not facilitate updates to RDF graphs. **Figure 13** presents an overview of the SHACL-based framework design.

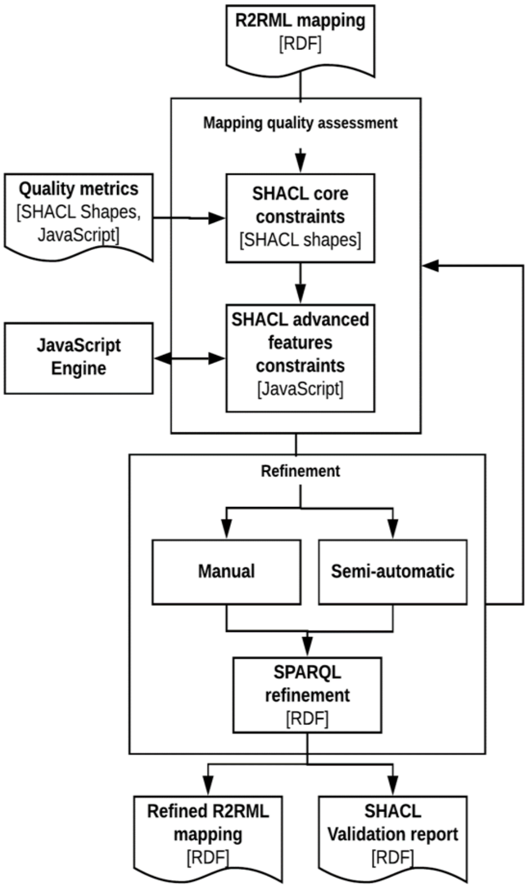
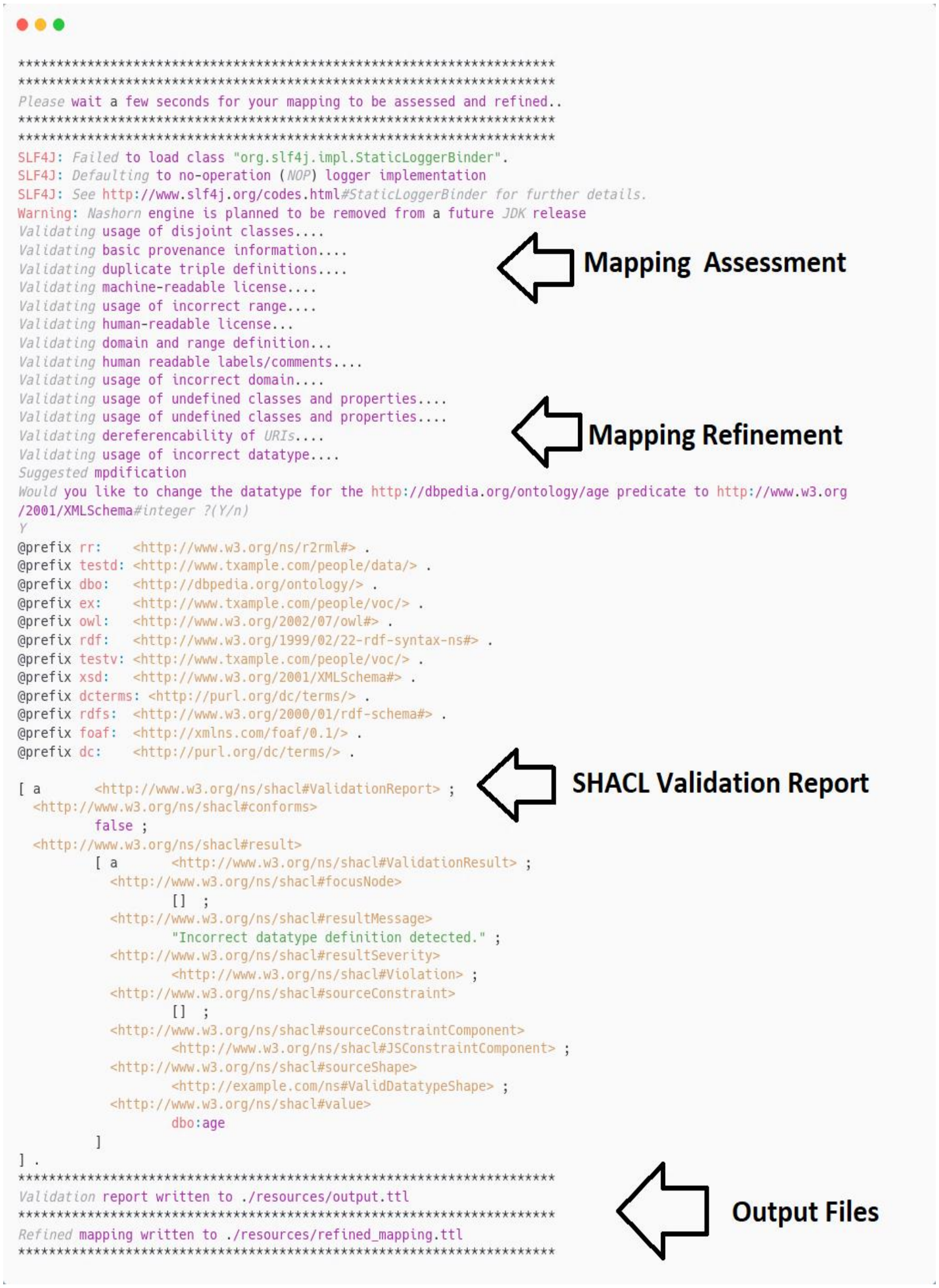


Figure 13: SHACL-based mapping quality assessment and refinement framework

The framework was implemented (**Figure 14**) as a command line interface tool, which provided prompts to users for the refinement process. Thereafter, the validation report generated by the framework, detailing mapping assessment information is displayed.

[18] https://www.w3.org/TR/shacl/#validation-report



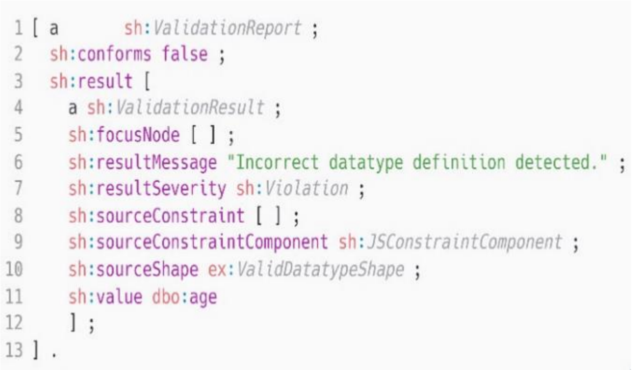
```
***********************************************************************
***********************************************************************
Please wait a few seconds for your mapping to be assessed and refined..
***********************************************************************
***********************************************************************
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
Warning: Nashorn engine is planned to be removed from a future JDK release
Validating usage of disjoint classes....
Validating basic provenance information....
Validating duplicate triple definitions....
Validating machine-readable license....
Validating usage of incorrect range....
Validating human-readable license...
Validating domain and range definition...
Validating human readable labels/comments....
Validating usage of incorrect domain....
Validating usage of undefined classes and properties....
Validating usage of undefined classes and properties....
Validating dereferencability of URIs....
Validating usage of incorrect datatype....
Suggested mpdification
Would you like to change the datatype for the http://dbpedia.org/ontology/age predicate to http://www.w3.org
/2001/XMLSchema#integer ?(Y/n)
Y
@prefix rr:    <http://www.w3.org/ns/r2rml#> .
@prefix testd: <http://www.txample.com/people/data/> .
@prefix dbo:   <http://dbpedia.org/ontology/> .
@prefix ex:    <http://www.txample.com/people/voc/> .
@prefix owl:   <http://www.w3.org/2002/07/owl#> .
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix testv: <http://www.txample.com/people/voc/> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf:  <http://xmlns.com/foaf/0.1/> .
@prefix dc:    <http://purl.org/dc/terms/> .

[ a       <http://www.w3.org/ns/shacl#ValidationReport> ;
  <http://www.w3.org/ns/shacl#conforms>
          false ;
  <http://www.w3.org/ns/shacl#result>
          [ a       <http://www.w3.org/ns/shacl#ValidationResult> ;
            <http://www.w3.org/ns/shacl#focusNode>
                    []  ;
            <http://www.w3.org/ns/shacl#resultMessage>
                    "Incorrect datatype definition detected." ;
            <http://www.w3.org/ns/shacl#resultSeverity>
                    <http://www.w3.org/ns/shacl#Violation> ;
            <http://www.w3.org/ns/shacl#sourceConstraint>
                    []  ;
            <http://www.w3.org/ns/shacl#sourceConstraintComponent>
                    <http://www.w3.org/ns/shacl#JSConstraintComponent> ;
            <http://www.w3.org/ns/shacl#sourceShape>
                    <http://example.com/ns#ValidDatatypeShape> ;
            <http://www.w3.org/ns/shacl#value>
                    dbo:age
          ]
] .
***********************************************************************
Validation report written to ./resources/output.ttl
***********************************************************************
Refined mapping written to ./resources/refined_mapping.ttl
***********************************************************************
```

Figure 14: Screenshot of SHACL-based implementation

The quality assessment and refinement of an R2RML mapping using the framework involves the following steps.

1. An R2RML mapping is input into the framework.
2. SHACL core and advanced constraints are used to validate the mapping. Core constraints are defined using SHACL concepts and relationships alone. However, the advanced constraints[19] also include the use of JavaScript[20] and SPARQL. These constraints represent metrics used to measure the quality of the mapping. The mapping assessment of the framework involves 20 metrics, which assess characteristics related to undefined terms and correct ontology reuse of terms, among others.
3. The quality report generated by the framework is expressed in SHACL validation report vocabulary. The report is queried using SPARQL in order to retrieve relevant information, such as number of quality issues detected. Thereafter, the users can resolve the issues manually, using a text-editor or semi-automatically using suggested refinements. The refinements provided the ability for users to select one of them and the associated queries and values would be inserted by the framework using SPARQL. For instance, a data type is incorrectly identified in a mapping. The framework will retrieve the correct data type from the respective ontology and replace the value in the mapping using a SPARQL update query[21].
4. The result of the process is a refined mapping, which was generated as a result of the refinement process on the original mapping. In addition, an associated validation report is output. The report details the quality of the generated refined mapping rather than the original.

A sample quality report generated by the framework is presented in **Listing 6**.



```
[ a       sh:ValidationReport ;
  sh:conforms false ;
  sh:result [
    a sh:ValidationResult ;
    sh:focusNode [ ] ;
    sh:resultMessage "Incorrect datatype definition detected." ;
    sh:resultSeverity sh:Violation ;
    sh:sourceConstraint [ ] ;
    sh:sourceConstraintComponent sh:JSConstraintComponent ;
    sh:sourceShape ex:ValidDatatypeShape ;
    sh:value dbo:age
    ] ;
] .
```

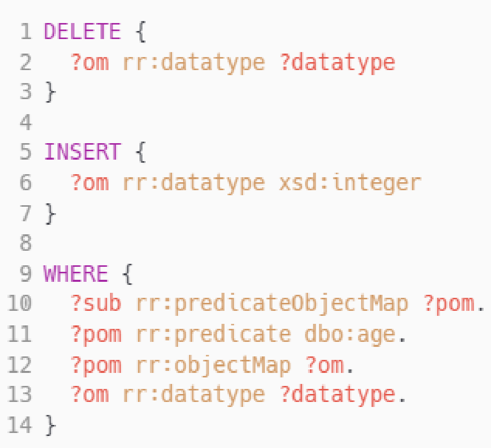
Listing 6: Sample quality report generated by the SHACL-based framework

[19] https://www.w3.org/TR/shacl/#dfn-shacl-sparql

[20] https://www.javascript.com/

[21] https://raw.githubusercontent.com/alex-randles/Mapping-quality-assessment-framework/master/paper_demo/datatype_refinement.rq

The extract (**Listing 6**) of the SHACL validation report (`sh:ValidationReport`) presents that metrics have detected at least one quality issue (`sh:conforms`) in a mapping. The detected quality issue (`sh:ValidationResult`) was detected by the “Usage of Incorrect Datatype” metric (`sh:sourceShape`) and indicates that object of the triple with the `dbo:age` predicate has been assigned an incorrect data type in the mapping. The violation was resolved by a semi-automatic refinement, which retrieves the correct data type from the respective ontology and updates the mapping definition. **Listing 7** presents the SPARQL update query which is executed by the framework on the original mapping if the user accepts the suggestion using the CLI (**Figure 14**).



```
DELETE {
  ?om rr:datatype ?datatype
}

INSERT {
  ?om rr:datatype xsd:integer
}

WHERE {
  ?sub rr:predicateObjectMap ?pom.
  ?pom rr:predicate dbo:age.
  ?pom rr:objectMap ?om.
  ?om rr:datatype ?datatype.
}
```

Listing 7: SPARQL update query executed by SHACL-based framework on sample mapping

The current datatype (`?datatype`) is replaced by the data type defined in the DBpedia ontology[22] (`xsd:integer`) when the update query is executed. Therefore, resulting in the removal of the root cause of the quality issue and preventing it from exponentially multiplying in the resulting dataset when created [42,65,75].

The quality metrics and refinements designed to be utilized by the framework provided inspiration for those in the mapping quality assessment and refinement component of the MQI framework. The implementation demonstrated that semi-automatic refinements are beneficial for the resolution of mapping quality issues, with them resolving issues during sample use cases.

In addition, the implementation of the SHACL-based framework, allowed limitations to be identified with an approach reliant on SHACL constraints. These limitations are outlined below.

1. **Expressiveness of Quality Information:** The SHACL validation report vocabulary is designed to represent quality information related to RDF data rather than mappings used to generate it. Therefore, the

[22] https://www.dbpedia.org/

vocabulary does not provide concepts and relationships suitable to model all of the information required for an effective mapping quality assessment and refinement approach. For instance, the triple map in a mapping where a quality issue is identified, cannot be represented using the vocabulary as it does not provide suitable concepts. In addition, the vocabulary does not contain suitable concepts to represent important metadata captured as a result of quality refinement activities, such as queries executed.

2. **Refinement Capabilities:** The SHACL advanced features support the execution of constraints defined in SPARQL. However, SHACL does not support SPARQL update queries and cannot directly execute refinements on mappings, therefore, limiting the refinement capabilities of the framework.
3. **Retrieval of Relevant Information:** Specific information from a mapping needs to be retrieved to guide the assessment and refinement process. For instance, the term map where a quality issue is located must be retrieved in order to execute refinement queries. In addition, the triple map location must be retrieved in order to ensure that another triple maps in the mapping are not accidentally updated by an executed refinement.
4. **Traceability:** R2RML mappings contain large numbers of blank nodes. Each blank node in a mapping, which relates to a detected quality issue is assigned an artificial identifier in the SHACL validation report generated by the framework. However, these identifiers cannot be traced back to the original mapping as they will be assigned a new identifier when loaded into a triple store in order to update the graph [27]. Therefore, a SHACL-based approach is limited by the traceability of mapping quality issues contained in generated quality reports.

These limitations of the developed SHACL-based framework indicated that a non SHACL-based basis for a mapping quality framework would be needed. The SHACL-based approach was discontinued due to these limitations and the MQI framework was developed with requirements designed to resolve the limitations.

3.4 Chapter Summary

The chapter was intended to provide useful background information related to the phases involved in the uplift of non-RDF to RDF data, while reporting on various quality issues which can arise in the lifecycle. In addition, prominent representations used to define declarative uplift mappings are presented. It is hoped the information will help readers to gain a deeper understanding of the uplift mapping process and factors which can impact quality. Thereafter, a preliminary study that was undertaken is presented. This study involved the development of a SHACL-based mapping quality improvement framework by the author of this thesis. Useful insights arising from the study related to various aspects limiting mapping quality assessment and refinement processes and were used to inform the design of the MQI framework, which is presented next.

Chapter 4: Design and Implementation of MQI Framework

This chapter presents the design and implementation of the MQI framework consisting of the development of two core components, which include the mapping quality assessment and refinement and source data source change detection component. **Supplementary information for this section is stored in a GitHub folder ("/Chapter-4").** A video demonstration[23] has been created, which demonstrates the functionality of the final version of the MQI framework.

4.1 Requirements

The requirements for the framework were defined as a result of limitations identified through the review of existing approaches in the two phases of the state of the art review discussed in **Chapter 2**. Requirement 1 (R1) and 2 (R2) were inspired from the limitations of existing approaches outlined in **Section 2.1.7** and **2.2.6**, respectively. Requirement 3 (R3) was inspired from the identified limitations and also the benefits of capturing related information in an ontology-based format, which were outlined in **Section 1.1**.

- **R1) The framework should facilitate the quality assessment and refinement of uplift mappings.** The framework should provide users with quality metrics which assess various mapping quality aspects. In addition, semi-automatic refinements should be presented to users, which guide them through the process of choosing the most appropriate refinement and associated values in order to resolve quality issues. Finally, the assessment and refinement process should be independent of the mapping language.
- **R2) The framework should facilitate the identification of changes in the source data of mappings.** Users who are responsible for maintaining LD datasets should be provided with information related to changes which have occurred in the source data and their impact on respective uplift mappings. The framework should be capable of sending notifications to users detailing relevant changes and potential mapping alignment issues. The change detection should be independent of the source data format.
- **R3) The framework should provide provenance on related quality of mapping process activities in an ontology-based format.** The framework should provide provenance information in an ontology-based format to allow easier processing of captured information and further knowledge discovery. Furthermore, software agents can automatically process the data and execute actions to improve quality of mappings.

[23] https://drive.google.com/file/d/14-AQloiL9WnQ1iUwzsqYk6cQaKo0u_IZ

Moreover, the information can be linked with associated mappings and resulting LD datasets to improve their trustworthiness.

Table 24 presents how well the approaches reviewed in the state of the art currently satisfy the requirements.

Table 24: **Support of requirements by existing approaches**

(Partial support (✓) Full support ✓ No support ✕)

Approach	**R1 Assessment & Refinement**	**R2 Source Change detection**	**R3 Ontology based Quality related information**
EvaMap	(✓)	✕	✕
Resglass	✓	✕	✕
Luzzu-Extension	(✓)	✕	✓
PROV-O Based	✓	✕	✓
RML-Validator	(✓)	✕	✓
DSNotify	✕	(✓)	✓
DELTA-LD	✕	(✓)	✓
sparqlPuSH	✕	(✓)	✕
DyLDO	✕	(✓)	✕

As can be seen none of the approaches provide full support for each requirement, but this is understandable given that they were not designed with the requirements in mind.

4.2 Mapping Quality Assessment and Refinement Component

The development of the **Mapping Quality Assessment and Refinement Component** involved the creation of an ontology designed to represent information related to quality of mappings, captured by the framework. Thereafter, the framework was designed and implemented which utilizes the ontology to represent mapping quality assessment and refinement information.

4.2.1 Design of MQIO

The **M**apping **Q**uality **I**mprovement **O**ntology (**MQIO**) was designed to express information related to the quality assessment, refinement and validation of LD mappings. The goal is to make relevant mapping quality information easier to publish, exchange and consume, thus providing important provenance information related to the publication process of LD dataset [42,65,75]. The ontology provides concepts and relationships in order to validate defined quality requirements, thus enabling consumers to assess the suitability of the mapping and resulting dataset for their use case (See Section 1.1) [38]. Moreover, an ontology-based representation of quality information enables software agents to automatically update a mapping in order to improve quality [42,75]. The **documentation** describing the MQIO is available online (https://w3id.org/MQIO). This section discusses the design methodology of the ontology. Thereafter, the concepts and relationships defined as a result of the development process are presented.

4.2.1.1 Design Methodology

The design of the MQIO followed best practices as recommended by the semantic web community. Ontology design practices were reused from the most prominent ontology design methodologies. The methodologies included the NeON methodology [134], UPON Lite [100], Ontology development 101: A guide to creating your first ontology [103] and LOT: An industrial oriented ontology engineering framework [110].

1. **Identification of aims, objectives, scope:** The design process commenced with the identification of the aims, objectives and scope of the ontology, which are outlined in **Appendix T** of this document. The template used for the table was retrieved from the methodologies and used to define the ontology requirements specification document. The document outlines requirements and among other things, the aims, objectives and scope of the ontology.
2. **Identify and analyze relevant information:** A review of publications in the state of the art was conducted to identify relevant information. Publications within the state of the art which related to topics within the defined scope were reviewed to facilitate the retrieval of relevant information. Thereafter, the retrieved information was used to formalize competency questions. References to publications which inspired the creation of each competency question are included.
3. **Create Use-cases and Competency questions:** Competency questions were created during the design process of the ontology. The questions define the functional requirements of the ontology and were iteratively refined until an accurate representation of the requirements and objectives was conceived. The final iteration of the questions is presented in **Appendix S**. Use cases were devised in order to refine the requirements of the ontology. Uses cases where MQIO was applied are described in the application study of the MQI framework in **Section 5.7**.
4. **Identify Concepts and Relationships:** It was decided to commence the process by identifying concepts and relationships from the Data Quality Vocabulary (DQV) [2], which could be reused to represent relevant information in the MQIO. DQV is a prominent ontology designed to represent quality information related to LD datasets. The ontology contains concepts and relationships to represent quality measurements, metrics, dimensions, categories and associated metadata. However, it was discovered that DQV reuses the PROV Ontology (PROV-O) [16] in order to represent activities and entities associated with the quality metadata, such as the quality assessment process of a dataset, as DQV does not provide suitable concepts for representing related activities and entities. PROV-O is a W3C recommendation designed to represent LD provenance and contains concepts to represent generalized provenance related activities and entities captured in the LD domain. Therefore, it was decided to reuse PROV-O in a similar manner to the DQV. However, it was discovered that it would be required to extend the ontology in order to capture domain specific information required for the MQIO. For instance, the quality assessment and refinement of a mapping are two separate activities, which should be distinguished between in order to support sufficient

interoperability. PROV-O includes a single concept to represent general activities (`prov:Activity`) and it would not be possible for a machine to automatically determine differences between information related to quality assessment and refinement activities, as these would be represented by the same concept. Similar constraints apply to the entities (`prov:Entity`) involved in the MQIO model, such as the mapping itself (`mqio:MappingArtefact`) and associated quality reports (`mqio:MappingValidationReport`), which must have a distinguishable different meaning. Therefore, it was decided to extend PROV-O in order to capture the domain specific concepts required in an ontology designed to represent information related to the quality improvement of mappings. An ontology fully reliant on concepts in the DQV and PROV-O would result in a semantic inoperable model, drastically limiting the ability to support quality assessment and refinement of mappings by agents. The concepts and relationships were iteratively defined until the information modeling provided by the ontology satisfied each of the competency questions. In addition, concepts and relationships were reused from existing vocabularies as recommended by the methodologies and the W3C recommendation on Data on the Web Best Practices [87]. The reuse in MQIO is demonstrated in the competency questions and ontology documentation.

5. **Progressive iterations:** Steps 2-4 were iteratively repeated until the point when the proposed concepts/relationships provided information which satisfied each requirement defined in the form of a competency question.
6. **Create Ontology:** The ontology was implemented in OWL 2 Web Ontology Language [92]. Concepts and relationships which were defined in the previous step were constructed using Protégé ontology development tool [154]. Furthermore, semantic reasoners were also utilized to detect logical inconsistencies within the ontology.
7. **Evaluate:** The ontology was evaluated with respect of the ability for the defined concepts and relationships to fulfill each competency question. The usage of a semantic reasoner within Protege ensured logical inconsistencies were identified and removed. Ontology Pitfall Scanner (OOPS!) [111] was used to detect common ontology design issues. The quality of metadata and documentation was evaluated through presentation within peer reviewed publications. Feedback received from reviewers allowed us to identify areas for improvement. Peer reviewed publications related to MQIO are outlined in **Section 1.5.2**. A sample graph (“mqio_sample_graph.ttl”) represented in the MQIO is available in the GitHub.
8. **Publication:** Ontology documentation was created using a Wizard for DOCumenting Ontologies (WIDOCO) [54] which is a tool designed to use ontology metadata to create HTML documents listing descriptions of the classes and properties. Thereafter, the ontology and human readable documentation were published with a permanent identifier as a Findable, Accessible, Interoperable and Reusable (FAIR) resource including an open and permissive license. The documentation contains information about the creation, design, usage, class interaction diagrams and provides various serializations.

The resulting ontology from this design methodology was **version 1.0**.

4.2.1.2 Concepts and Relationships of MQIO

Figure 15 presents the reuse of PROV-O in the design of MQIO. The **yellow** boxes shown represent concepts introduced by MQIO and **blue** boxes represent concepts reused from ontologies from the state of the art.

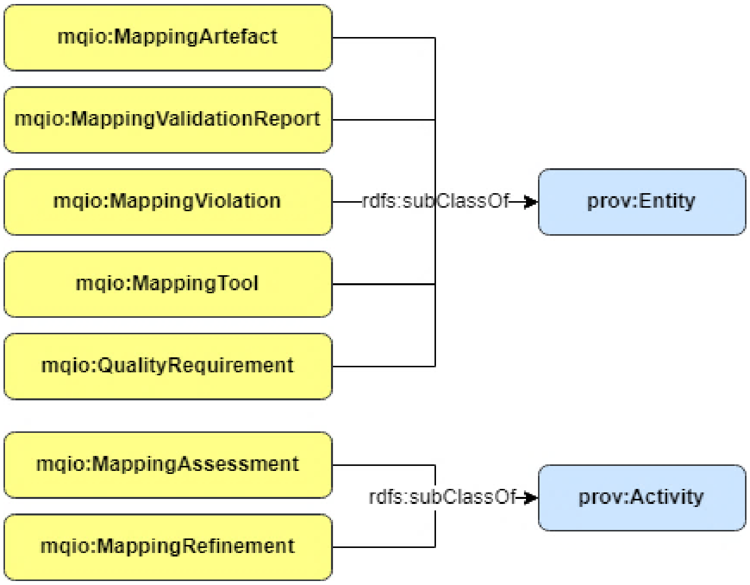


Figure 15: Reuse of existing vocabularies in MQIO

As can be seen, entities designed to represent mapping artefacts and associated quality reports and tools are represented as specialized provenance related entities (`prov:Entity`). The activities involved in the quality assessment and refinement of mappings are represented as specialized provenance related activities (`prov:Activity`). Therefore, providing semantically rich concepts in the MQIO.

Figure 16 presents the ontology classes and properties of MQIO used during the quality improvement of a mapping.

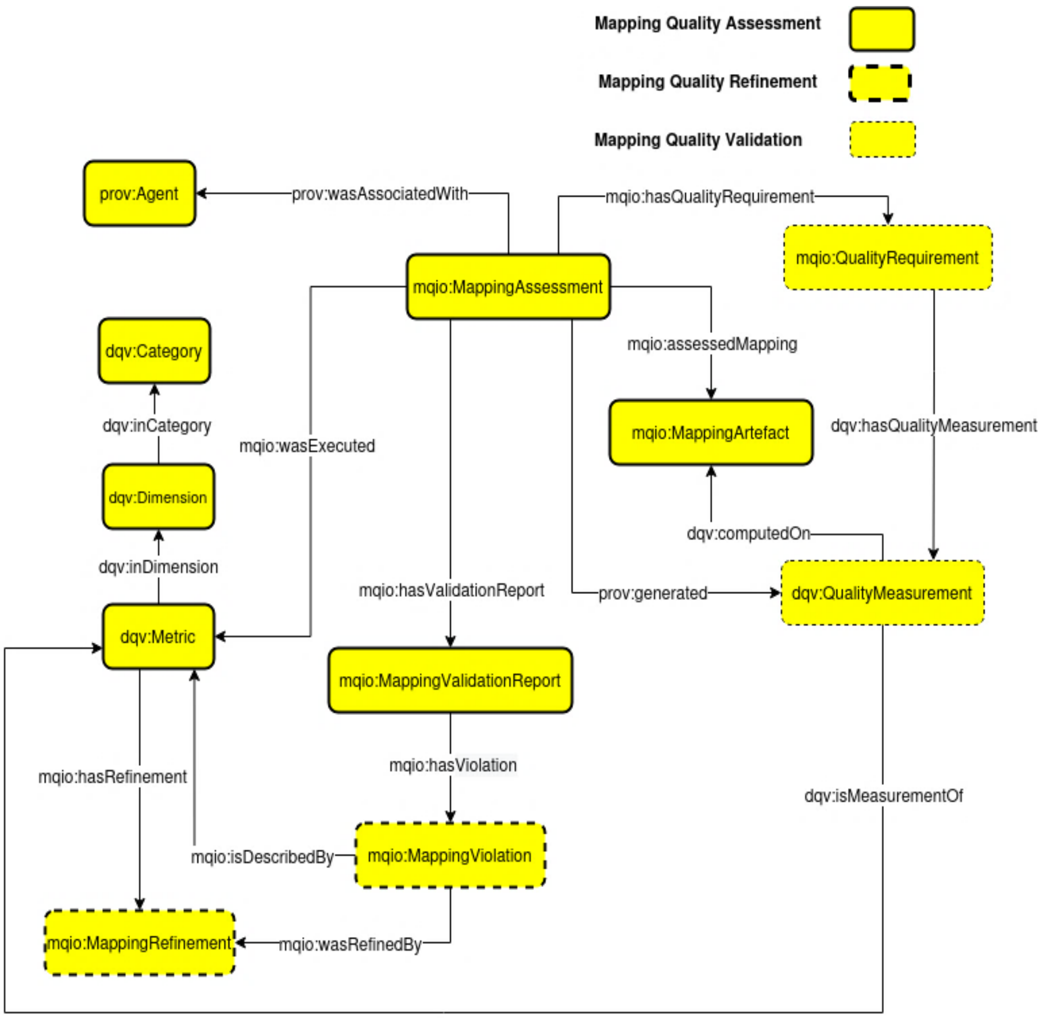


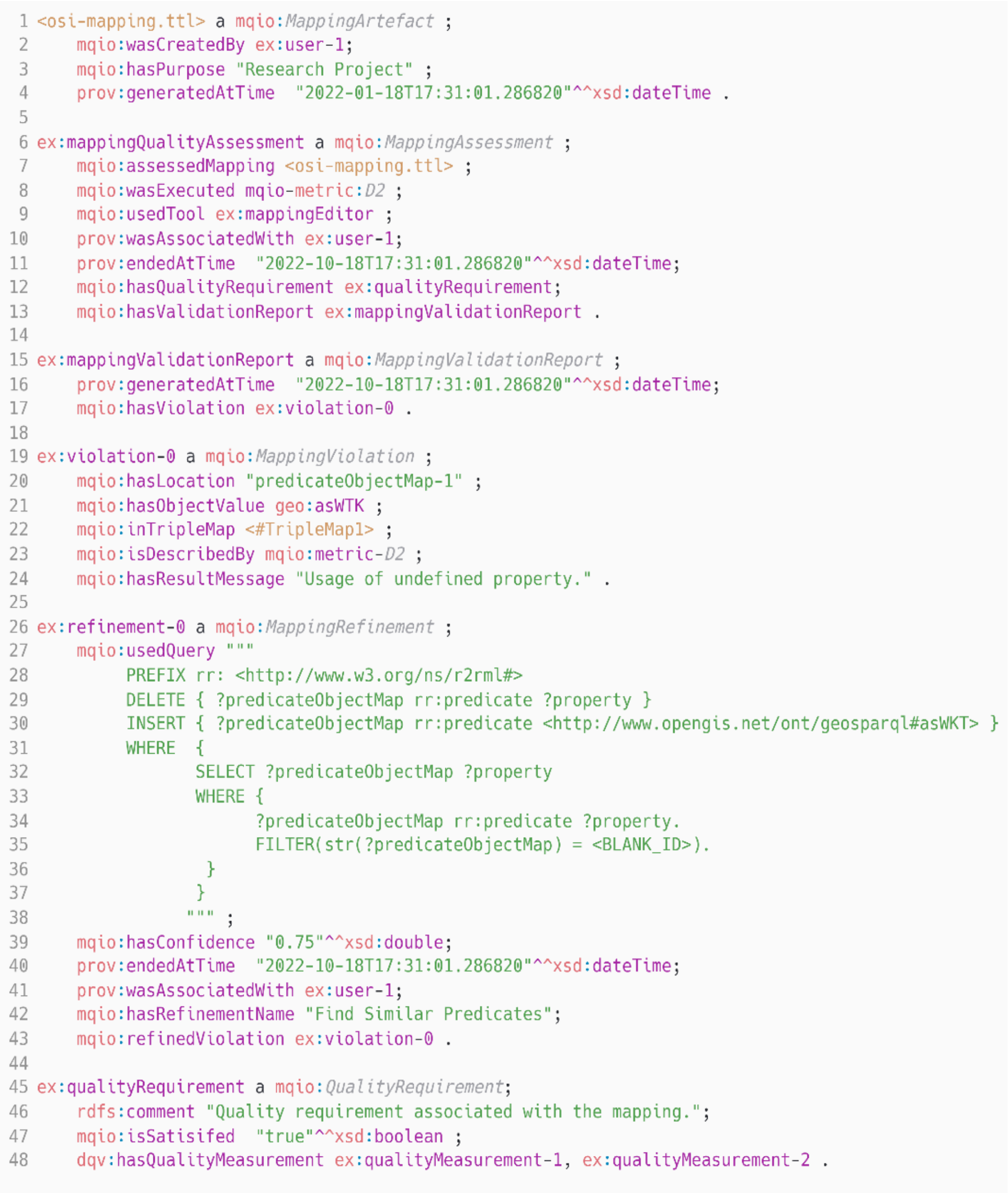
Figure 16: MQIO classes and properties usage during mapping quality improvement

There are three stages in the quality improvement of a mapping, which includes mapping quality assessment, mapping quality refinement and mapping quality validation. These stages are presented in **Figure 16**.

- **Mapping Quality Assessment:** In this stage, the mapping artefact (`mqio:MappingArtefact`) is assessed. The assessment activity is captured through the `mqio:MappingAssessment` class. The agent who initiated the process is also captured (`prov:Agent`). A mapping assessment activity may have quality requirements which are captured through the `mqio:QualityRequirement` class. This information allows the ontology to validate whether such quality requirements have been satisfied in the assessment and refinement stage. The ontology draws inspiration from DQV where data quality was classified into categories (`dqv:Category`), dimensions (`dqv:Dimension`) and metrics (`dqv:Metric`). The ontology uses the information in such classes to generate a mapping validation report (`mqio:MappingValidationReport`). Each violation identified is then represented with the `mqio:MappingViolation` class.

- **Mapping Quality Refinement:** This stage captures mapping refinements executed on the mapping. Each metric described using the ontology may have multiple refinements (`mqio:MappingRefinement`) associated with it depending on the quality aspect being measured. The refinement executed in the mapping is associated with the identified violation through the property `mqio:wasRefinedBy`. In addition, refinements have scores representing the likelihood the respective violation will be resolved.
- **Mapping Quality Validation:** Finally, the ontology provides quality information on the original mapping being assessed and the mapping which has been generated as a result of the refinement. As mentioned, each mapping assessment process may have quality requirements which can be validated at this stage. For instance, one may define a quality requirement for understandability related quality metrics to be of a particular value. In this stage, this can be validated by comparing the defined quality requirement and the resulting one. Requirements can be validated by identifying quality measurements (`dqv:QualityMeasurement`) associated with the quality assessment of the mapping.

Listing 8 presents an extract of a quality report expressed in MQIO.



```
<osi-mapping.ttl> a mqio:MappingArtefact ;
    mqio:wasCreatedBy ex:user-1;
    mqio:hasPurpose "Research Project" ;
    prov:generatedAtTime  "2022-01-18T17:31:01.286820"^^xsd:dateTime .

ex:mappingQualityAssessment a mqio:MappingAssessment ;
    mqio:assessedMapping <osi-mapping.ttl> ;
    mqio:wasExecuted mqio-metric:D2 ;
    mqio:usedTool ex:mappingEditor ;
    prov:wasAssociatedWith ex:user-1;
    prov:endedAtTime  "2022-10-18T17:31:01.286820"^^xsd:dateTime;
    mqio:hasQualityRequirement ex:qualityRequirement;
    mqio:hasValidationReport ex:mappingValidationReport .

ex:mappingValidationReport a mqio:MappingValidationReport ;
    prov:generatedAtTime  "2022-10-18T17:31:01.286820"^^xsd:dateTime;
    mqio:hasViolation ex:violation-0 .

ex:violation-0 a mqio:MappingViolation ;
    mqio:hasLocation "predicateObjectMap-1" ;
    mqio:hasObjectValue geo:asWTK ;
    mqio:inTripleMap <#TripleMap1> ;
    mqio:isDescribedBy mqio:metric-D2 ;
    mqio:hasResultMessage "Usage of undefined property." .

ex:refinement-0 a mqio:MappingRefinement ;
    mqio:usedQuery """
         PREFIX rr: <http://www.w3.org/ns/r2rml#>
         DELETE { ?predicateObjectMap rr:predicate ?property }
         INSERT { ?predicateObjectMap rr:predicate <http://www.opengis.net/ont/geosparql#asWKT> }
         WHERE  {
                SELECT ?predicateObjectMap ?property
                WHERE {
                      ?predicateObjectMap rr:predicate ?property.
                      FILTER(str(?predicateObjectMap) = <BLANK_ID>).
                 }
                }
               """ ;
    mqio:hasConfidence "0.75"^^xsd:double;
    prov:endedAtTime  "2022-10-18T17:31:01.286820"^^xsd:dateTime;
    prov:wasAssociatedWith ex:user-1;
    mqio:hasRefinementName "Find Similar Predicates";
    mqio:refinedViolation ex:violation-0 .

ex:qualityRequirement a mqio:QualityRequirement;
    rdfs:comment "Quality requirement associated with the mapping.";
    mqio:isSatisifed  "true"^^xsd:boolean ;
    dqv:hasQualityMeasurement ex:qualityMeasurement-1, ex:qualityMeasurement-2 .
```

Listing 8: Sample graph expressed in the MQIO

The sample graph presents mapping quality assessment (ex:mappingQualityAssessment) which has been executed by an agent (ex:user-1) and resulted in the detection of one quality violation (ex:violation-0). A refinement (ex:refinement-1) was executed which resolved the violation. The quality requirement (ex:qualityRequirement) was validated in order to identify if the refined mapping was sufficient quality for execution. Specialized metadata can be inserted into quality reports expressed in MQIO, by reusing PROV-O [84] similarly.

4.2.2 Design of Mapping Quality Assessment and Refinement Component

This section discusses the design and implementation of the mapping quality assessment and refinement component of the MQI framework. First, an overview of the design is described, including the metrics and refinements used by the framework. Thereafter, the implementation of the design is discussed. **R2RML mappings** [35] were chosen as the target language for the mapping quality assessment and refinement component of the framework as it is the W3C recommendation for creating customized transformation rules, unlike the direct mapping representation [5], which does not allow customized rules. Furthermore, the language is RDF based, which enables these mappings to be linked with relevant RDF graphs, such as the resulting data in order to improve trustworthiness [42,65,75]. Moreover, the mappings are extensively used in the semantic web community [29], which enables mappings to be easily gathered for testing purposes.

4.2.2.1 Mapping Quality Assessment

The following subsections discuss the quality metrics that were designed by the author of this thesis specifically to target uplift mappings. The design of these metrics has been inspired by existing metrics [38,150] in the state of the art, which are commonly used to assess the quality of the LD datasets produced by the mappings**. Table 25** presents an overview of LD quality dimensions, which represent a grouping of metrics designed to target a specific aspect of quality. These quality dimensions represent the grouping of metrics shown in **Table 2**. The respective quality categories of each quality dimension are shown in brackets below the name. A description of each dimension is shown. In addition, the number (**%**) of LD quality metrics that are potentially impacted by the quality of a mapping is indicated in the "**Description of Impact**" column. For example, in the table, you can see that (10 of the 10 (10/10) LD quality metrics) (100%) in the Data Consistency Dimension are potentially impacted by the quality of a mapping. Metrics and dimensions [38,150] in categories such as Contextual (CT), Representational (RP), Accessibility (AC) and Intrinsic (IR) will be directly impacted by quality of a mapping. A description of each related quality category is presented in **Appendix E**. An abbreviation (**Abr**) has been created in order to reference each quality dimension presented.

Table 25: Description and number of metrics impacted by mapping quality in each quality dimension

Dimension	Abr	%	Description of Dimension	Description of Impact
Availability (AC)	**AV**	**0/5** (0%)	*Access methods of the data*	Hosting is out of scope of the mapping itself
Completeness (IR)	**CM**	**2/2** (100%)	*Extent to which data is complete with respect to the real world*	Classes and properties used in the data originate in the mapping
Conciseness (IR)	**CN**	**3/3** (100%)	*Degree of redundancy in the dataset*	Classes and properties used in the data originate in the mapping
Data Consistency (IR)	**DC**	**10/10** (100%)	*Level of coherence in a dataset with respect to the knowledge it represents*	Classes and properties used in the data originate in the mapping
Interoperability	**IN**	**2/2**	*Degree to which the format and structure of*	Reuse of vocabularies

(RP)		(100%)	*the information conforms to previously returned information*	demonstrated in mapping definitions
Interpretability (RP)	**IO**	**4/4** (100%)	*Technical aspects of the data, that is, whether information is represented using an appropriate notation and whether the machine is able to process the data*	Classes and properties used in the data originate in the mapping
Interlinking (AC)	**IL**	**3/3** (100%)	*Degree of internal and external interlinks between data sources*	Links in the dataset are defined in mappings
Licensing (AC)	**LI**	**3/3** (100%)	*Permissions (if defined) to re-use a dataset*	Licensing information originates in mapping definitions.
Performance *(AC)*	**PE**	**1/4** (25%)	*Efficiency of a system, that is, the more performant a data source is the more efficiently a system can process data*	Hosting is out of scope of the mapping itself, however, format of URIs is defined in mappings
Relevancy (CT)	**RL**	**2/2** (100%)	*Provision of information which is in accordance with the task at hand and important to the users' query*	Relevant information is provided through additional mapping definitions
Representational conciseness (RP)	**RC**	**2/2** (100%)	*Extent of the representation of the data, which is compact and well formatted, clear and complete*	Classes and properties used in the data originate in the mapping
Semantic accuracy (IR)	**SA**	**5/5** (100%)	*Degree to which data values correctly represent the real-world facts.*	Logical inconsistencies originate from incorrect mapping definitions
Security (AC)	**SC**	**0/2** (0%)	*Extent to which data is protected against alteration and misuse.*	Hosting is out of scope of the mapping itself
Syntactic validity (IR)	**SV**	**3/3** (100%)	*The conformance of an RDF graph with the RDF standard*	Syntax of resulting data originates in mapping definitions
Timeliness (CT)	**TI**	**2/2** (100%)	*How up-to-date data is relative to a specific task*	Reuse and maintenance of mapping will ensure up-date data
Trustworthiness (CT)	**TR**	**7/7** (100%)	*Degree to which the information is accepted to be correct, true, real and credible*	Provenance is provided through additional mapping definitions
Understandability (CT)	**UT**	**3/3** (100%)	*Ease with which data can be comprehended without ambiguity and be used by a human information consumer*	Vocabularies used in linked data datasets originate in the mapping definitions
Versatility (RP)	**VT**	**0/2** (0%)	*Availability of the data in different representations and in an internationalized way*	Serialization formats and languages of source data are out of scope

As can be seen most (14 out of 18) of the quality dimensions shown are potentially directly impacted by mapping quality. These results indicate the quality of the published datasets is heavily influenced by the quality of the mapping artefacts used during the publication process. Three of the quality dimensions (Availability, Security, Performance) which are not directly impacted by mapping quality are related to the accessibility category. The category groups dimensions related to the hosting of the published LD dataset, which is out of scope of the mapping process. Versatility is related to the representational category and is influenced by the mapping processor used rather than the quality of mapping. Therefore, it can be concluded mappings are an extremely important aspect with regards to the quality of LD as most quality dimensions are potentially impacted.

4.2.2.1.1 Categorizing Mapping Quality Metrics

The mapping quality metrics developed are grouped into three quality **aspects** which include mapping quality, data quality and vocabulary quality outlined below. The aspects were inspired by existing work [38,99,150] in LD quality, which focuses on the quality of data and vocabularies used for representation. However, the existing work did not take quality of a mapping artefact into account, therefore, an additional aspect to capture related quality has been defined.

- **Mapping Quality Aspect (MP):** First, the quality of the mapping itself is considered by assessing, for instance, the extent to which the mapping correctly conforms to the specification of the R2RML [35] mapping language used and also, whether there are any redundant definitions within the mapping (which would affect the performance of R2RML mapping engines).
- **Data Quality Aspect (D):** The second aspect relates to the quality of the output generated by an engine processing the input data and the R2RML mapping. Poor design decisions made at the stage of defining an R2RML mapping, such as using non-dereferenceable classes and properties, or deprecated ones, will decrease the quality of the dataset. This aspect focuses on the quality of the output data which can be identified and fixed during mapping design-time.
- **Vocabulary Quality Aspect (VOC):** Finally, the third aspect considered relates to the quality of the vocabularies used within the R2RML mapping, by assessing for instance, that these classes and properties contain human readable labels or comments in the vocabulary. The rationale for including this aspect is to ensure the quality of the resulting datasets by making quality information related to the vocabularies being used in the mapping transparent to mapping engineers.

In addition, mapping quality metrics are categorized (**Table 26**) based on the quality dimension related to them, which were inspired by previous work in LD quality [38,99,150]. However, one quality dimension (**Table 27**) was newly defined rather than reused as a similar dimension could not be found in the state of the art.

Table 26: Newly defined quality dimension related to metrics designed to assess quality of mappings

Dimension	**Abr**	**Description**
Mapping Consistency (IR)	MC	This dimension refers to the extent to which a mapping is conformant to its mapping language.

The mapping consistency dimension was inspired by data consistency, however, it relates to the consistency of the mapping rather than the resulting data. The following subsections describe the mapping quality metrics grouped based on their related mapping quality aspect.

4.2.2.1.2 Mapping Quality Metrics

Table 27 presents the quality metrics and respective refinements proposed to resolve the quality issues in the artefact (Section 4.2.2.2), related to the **mapping quality aspect** of a mapping. In addition, the quality dimension of each metric is shown in brackets below the ID. The abbreviations for the quality dimensions shown have been taken from **Table 25** and **Table 26**.

Table 27: Metrics related to **mapping quality aspect** of mappings

ID	Metric	Description	Refinements
MP1 (MC)	Valid logical table definition	A logical table exist and references either a table (or view) or an SQL query [30,35,75].	• Add Logical Table • Add Logical Source
MP2 (MC)	Valid subject map definition	One subject map, which may have zero or more class definitions [30,35,75,107].	• Add Subject Map
MP3 (MC)	Valid predicate object map definition	There must exist at least one predicate map and one object map. These are used to generate the predicates and objects of the triples [30,35,75,107].	• Add Predicate • Add Object Map
MP4 (MC)	Valid parent triples map definition	The triples map being referenced must exist in the mapping and, when defined, join conditions must have both parent and child column definitions [30,35,75].	• Add Parent Colum • Add Child Column
MP5 (MC)	Valid language datatype definition	Object maps with literal values may refer to only one language tag or one datatype definition (not both) [30,35,75].	• Remove Datatype • Remove Language Tag
MP6 (MC)	Valid term type definition	Terms maps are assigned the correct term types. Subject maps may be IRIs or blank nodes, predicate maps must be IRIs, and object maps may be IRIs, blank nodes, or literal [30,35,75].	• Add Correct Term Type • Choose Term Type
MP7 (SV)	Valid subject definition	Subject definitions must be valid URIs unless its type is defined as blank node [30,35,75].	• Change URI
MP8 (SV)	Valid predicate definition	The predicate definition is a valid URI [30,35,75].	• Change URI
MP9 (SV)	Valid named graph definition	The named graph definition is a valid URI [30,35,75].	• Change URI
MP10 (SV)	Valid datatype definition	The datatype definition is a valid URI. In R2RML, this would involve validating object maps associated with datatypes [30,35,75].	• Change URI • Remove Datatype
MP11 (SV)	Valid literal language tags	Valid language tags are defined as per RFC 5646 (BCP 47)[24] [30,35,75].	• Choose Valid Language Tag • Change Language Tag
MP12 (RC)	Duplicate triples defined	Mappings which generate the same triple more than once [26].	• Change Predicate • Change Object • Remove Triple Definition

[24] http://www.rfc-editor.org/info/rfc5646

4.2.2.1.3 Data Quality Metrics

Table 28 presents the quality metrics and respective refinements related to the **data quality aspect** of mappings.

Table 28: Metrics related to **data quality aspect** of mappings

ID	Metric	Description	Refinements
D1 (IN)	Usage of undefined classes	Classes are considered undefined when it is not possible to dereference them against their namespace [38,150,151].	• Find Similar Classes • Change Class
D2 (IN)	Usage of undefined properties	Properties are considered undefined when it is not possible to dereference them against their namespace [38,150,151].	• Find Similar Predicates • Change Predicate
D3 (DC)	Usage of incorrect domain	A class defined in the domain is not included in the mapping [38,150,151].	• Add Domain Class • Change Predicate
D4 (RC)	No query parameters in URI's	The use of query parameters is not recommended by the W3C best practices for URIs [38,137,150,151].	• Change URI
D5 (DC)	No use of entities as members of disjoint classes	Individuals of one class cannot be simultaneously members of another class [38,150,151].	• Change Class • Remove Class
D6 (DC)	Usage of incorrect range	This will validate that the correct domain and range are being used in the mapping definition [38,150,151].	• Add Correct Term Type • Choose Term Type • Remove Term Type
D7 (DC)	Usage of incorrect data type	This will validate that all objects have been assigned the correct datatype [38,150,151].	• Add Correct Datatype • Change Datatype • Remove Datatype

4.2.2.1.4 Vocabulary Quality Metrics

Table 29 presents the quality metrics and respective refinements related to the **vocabulary quality aspect** of mappings.

Table 29: Metrics related to **vocabulary quality aspect** of mappings

ID	Metric	Description	Refinements
VOC1 (UT)	Human readable labels/comments	Humans consuming the information should be able to understand the linked data resource [38,87,150,151].	• Add Comment • Add Label
VOC2 (UT)	Domain and range definitions	A property should have a range and/or domain definition [22,38,87,150,151].	• Add Domain Definition • Add Range Definition
VOC3 (TR)	Basic Provenance Information	Consumers need to understand where the data has originated [22,38,87,150,151].	• Add Provenance
VOC4 (LI)	Machine readable licensing	Allows the licensing information to be queried by machines [38,87,112,150,151].	• Add Machine License
VOC5 (LI)	Human readable licensing	Allows humans to read and understand license in a textual format [38,87,150,151].	• Add Human License

Documentation[25] detailing the RDF representation of these metrics and related dimensions and categories is available online, enabling linking with relevant RDF graphs. For example **Figure 17** presents the concepts and relationships used to define an instance of the D2 metric in RDF format using MQIO. The **yellow** boxes shown represent concepts introduced by MQIO and **blue** boxes represent concepts reused from ontologies in the state of the art.

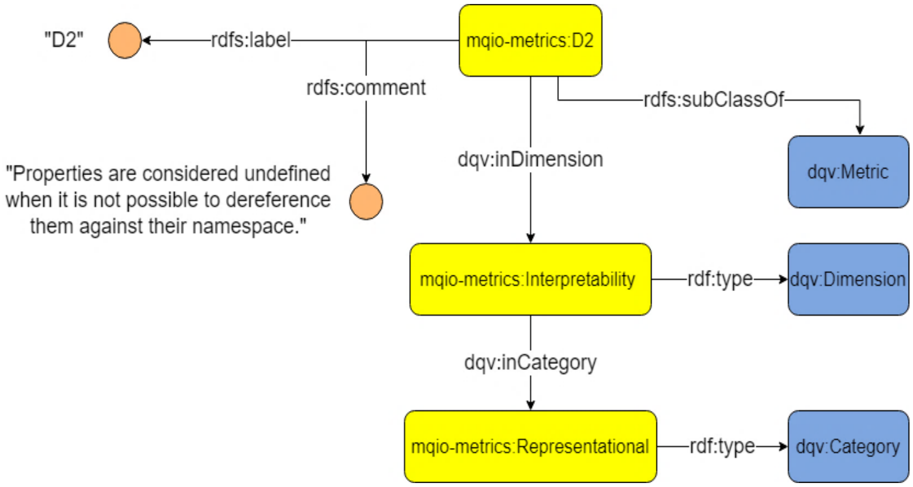

"D2"
rdfs:label
mqio-metrics:D2
rdfs:comment
rdfs:subClassOf
dqv:Metric
dqv:inDimension
"Properties are considered undefined when it is not possible to dereference them against their namespace."
mqio-metrics:Interpretability
rdf:type
dqv:Dimension
dqv:inCategory
mqio-metrics:Representational
rdf:type
dqv:Category

Figure 17: Concepts and relationships used to represent quality metrics expressed in MQIO

The D2 quality metric (`mqio-metrics:D2`) includes a label (`rdfs:label`) and description (`rdfs:comment`) of the instance. In addition, the respective quality dimension (`dqv:inDimension`) and category (`dqv:inCategory`) of the metric is represented.

4.2.2.2 Mapping Quality Refinement

Semi-automatic refinements which involve a human-in-the-loop were designed in order to resolve quality issues detected in uplift mappings. These refinements are grouped based on quality aspect of the refined metric and are outlined below.

Mapping Quality Aspect Refinements:

- **Add Logical Table:** Values involved in the definition of a logical table will be inserted into the mapping in the respective location. Logical tables refer to tables in a relational database.
- **Add Logical Source:** Values involved in the definition of a logical source will be inserted into the mapping in the respective location. Logical sources refer to other source data formats available.

[25] https://w3id.org/MQIO-metrics

- **Add Subject Map:** Values involved in the definition of a subject map will be inserted into the mapping in the respective location.
- **Add Predicate:** A property identifier will be inserted into the mapping into the respective location.
- **Add Object Map:** Values involved in the definition of an object map will be inserted into the mapping in the respective location.
- **Add Parent Column:** The name of the parent column will be inserted into the respective location.
- **Add Child Column:** The name of the child column will be inserted into the respective location.
- **Remove Datatype:** The respective datatype will be removed from the mapping.
- **Remove Language Tag:** The language tag will be removed from the mapping.
- **Add Correct Term Type:** The ontology of the respective property will be queried to retrieve the type of the range (object or literal) and the corresponding term type will be suggested. The selected value will replace the current term type.
- **Choose Term Type:** A defined term type will replace the current value.
- **Change URI:** A URI can be entered or created using a predefined prefix. The entered value will replace the respective URI.
- **Choose Valid Language Tag:** Valid language tags will be retrieved and the selected tag will replace the current tag.
- **Change Language Tag:** Any language tag can replace the current tag.
- **Change Predicate:** A URI can be entered or created using a predefined prefix. The entered value will replace the respective predicate.
- **Change Object:** A URI can be entered or created using a predefined prefix. The entered value will replace the respective object.
- **Remove Triple Definition:** One of the duplicate term maps will be removed from the mapping.

Data Quality Aspect Refinements:

- **Find Similar Classes:** Classes within the same namespace will be retrieved and the selected class will be inserted into the respective location.
- **Change Class:** The URI of a class replaces the existing class URI.
- **Find Similar Properties:** Properties within the same namespace will be retrieved and the selected property will be inserted into the respective location.
- **Change Predicate:** URI of a property replaces the existing property URI.
- **Add Domain Class:** The domain(s) of the respective property is retrieved from the ontology. Thereafter, a retrieved class URI can be inserted.
- **Remove Class:** The class will be removed.

- **Add Correct Term Type:** Term type(s) which represent the range of the property retrieved from the ontology will replace the current term type.
- **Choose Term Type:** The ontology of the mapping representation is queried to identify available term types. The selected term type will be insert into the respective location.
- **Remove Term Type:** The term type will be removed.
- **Add Correct Datatype:** The range of the datatype property is retrieved from the ontology and replaces the defined data type.
- **Change Datatype:** The URI of a datatype replaces the existing URI of the data type.
- **Remove Datatype:** The datatype will be removed.

Vocabulary Quality Aspect Refinements:

- **Add Comment:** A comment will be inserted using a suitable comment property into the respective location.
- **Add Label:** A label will be inserted using a suitable label property into the respective location.
- **Add Domain Definition:** The URI of the domain(s) is inserted using a suitable domain property into the respective location.
- **Add Range Definition:** The URI of the range(s) is inserted using a suitable range property into the respective location.
- **Add Provenance:** Provenance will be inserted using a suitable provenance related property into the respective location.
- **Add Machine License:** A license will be inserted using a suitable machine-readable license property into the respective location.
- **Add Human License:** A license will be inserted using a suitable human-readable license property into the respective location.

Refinements can also be completed manually directly by the mapping engineer guided by the quality assessment information that has been generated by the framework.

4.2.2.3 Overview of Design

Figure 18 presents the component diagram of the mapping quality assessment and refinement component of the MQI framework.

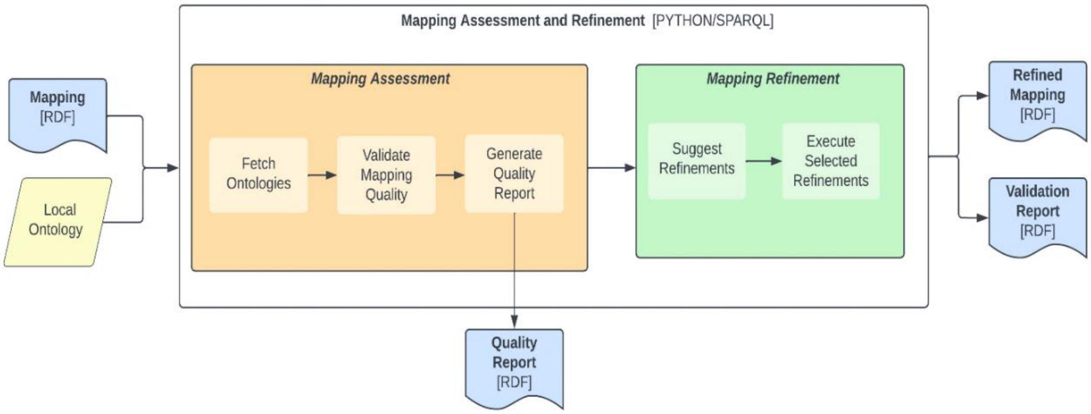


Figure 18: Component diagram of the mapping quality assessment and refinement component of the MQI framework

The process presented in **Figure 18** is outlined below.

- **Input:** An uplift mapping is input into the framework by users. In addition, they can upload a local ontology which is used in the mapping. A local ontology refers to an ontology which has not been published online. The functionality allows users to test ontologies prior to publication, such as those under development.
- **Mapping Assessment:** Each online ontology used in the mapping is retrieved in order to compute quality metrics. The retrieved ontologies are stored in a cache to improve efficiency when querying. Thereafter, the quality of the mapping is assessed by executing the metrics described in **Section 4.2.2.1**, which assess quality aspects related to syntactic, semantics and vocabulary reuse of the mappings. The results of the metrics are uplifted into a quality report represented in RDF format, which allows it to be automatically processed by agents in order to suggest refinements to resolve detected issues. In addition, the report can be linked with the resulting dataset to provide indications to consumers on the suitability of the data for their use case.
- **Mapping Refinement:** Refinements related to addressing (see last section) detected issues in the quality report are suggested to the users. The semi-automatic refinement process involves a human in the loop where the user is guided through the process of choosing the most appropriate refinement and associated values. The goal is to provide methods to easily resolve quality issues in the mappings. The information generated by the refinement activity are the refined mapping and validation report. The refined mapping relates to the mapping generated by executing the refinements on the original mapping which was input.

The validation report contains information related to the quality refinement activity and also related to the assessment of the refined mapping, and as such it an extension of the quality report.

4.2.2.4 Implementation

This section discusses the implementation of the mapping quality assessment and refinement component of the MQI framework. The implementation of the final version of the framework has been published as open-source[26]. **Supplementary information related to this section is stored in a folder ("/First-Iteration-Implementation") of the GitHub.**

4.2.2.4.1 Overview of Implementation

The MQI framework was implemented using SPARQL [61], Python [140], Apache Jena Fuseki[27], R2RML [35], Hyper Text Markup Language (HTML) [113] and Cascading Style Sheets (CSS) [88]. Python was used to create a web application using the Flask library [58] . It was decided to not use Java[28] as Python provides the latest RDF based libraries. The GUI was created using HTML[29] and CSS[30]. Apache Jena Fuseki was used as it is a prominent open-source triple store. The RDFLib library [82] in Python was used to execute SPARQL in order to query and update mapping graphs. R2RML was used to uplift data into RDF format. It was decided to use R2RML as it is the W3C recommendation for transforming relational data into RDF. Apache Jena Fuseki was used to store RDF data. **Figure 19** presents how the technologies are used within the implementation of the mapping quality assessment and refinement component of the MQI framework.

[26] https://github.com/alex-randles/MQI-Framework/
[27] https://jena.apache.org/documentation/fuseki2/
[28] https://www.java.com/en/
[29] https://www.w3.org/html/
[30] https://www.w3.org/Style/CSS/

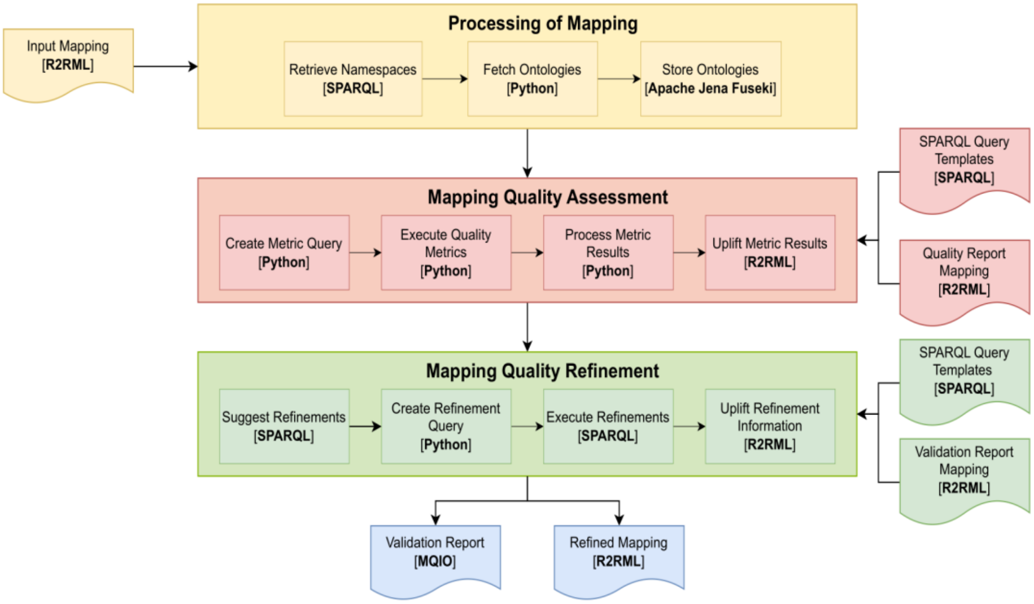

Input Mapping [R2RML]
Processing of Mapping
Retrieve Namespaces [SPARQL]
Fetch Ontologies [Python]
Store Ontologies [Apache Jena Fuseki]
Mapping Quality Assessment
Create Metric Query [Python]
Execute Quality Metrics [Python]
Process Metric Results [Python]
Uplift Metric Results [R2RML]
SPARQL Query Templates [SPARQL]
Quality Report Mapping [R2RML]
Mapping Quality Refinement
Suggest Refinements [SPARQL]
Create Refinement Query [Python]
Execute Refinements [SPARQL]
Uplift Refinement Information [R2RML]
SPARQL Query Templates [SPARQL]
Validation Report Mapping [R2RML]
Validation Report [MQIO]
Refined Mapping [R2RML]

Figure 19: Overview of technology used in the implementation of the mapping quality assessment and refinement component of MQI framework

The process presented in **Figure 19** is outlined below.

- **Processing of Mapping:** An uplift mapping defined in R2RML is input into the framework for quality assessment and refinement. The ontologies used in the mapping are required to calculate certain metrics, such as undefined classes and properties. The namespaces are used to retrieve respective ontologies, which are fetched using RDFLib library in Python and stored in the local Apache Jena Fuseki triple store.
- **Mapping Quality Assessment:** The SPARQL queries required for calculating the metrics are created using SPARQL query templates ("/Quality-Metrics"), where the respective values being measured are inserted using built-in Python functions. Thereafter, the queries are executed using the RDFLib library in Python. The results are processed using Python and stored as relational data which can be uplifted into RDF format using an R2RML engine[31]. The resulting RDF data is stored in a local cache which allows the users to export them.
- **Mapping Quality Refinement:** Refinements are suggested to the users by executing a SPARQL query ("retrieve_refinements.rq"), which retrieves refinements associated with metrics that have detected quality issues. The refinement query used to update the mapping is created using SPARQL query templates ("/Quality-Refinements"), where values input by the user are inserted. The refined mapping is created by executing the update query on the original mapping using RDFLib. Quality information related to the

[31] https://github.com/chrdebru/r2rml-tutorial

refinement activities that have been undertaken is uplifted into RDF format using R2RML. The result of the process is two files containing RDF format data, which includes a validation report expressed in MQIO and a refined mapping expressed in R2RML.

The steps are demonstrated in the following **subsections**.

4.2.2.4.2 Demonstration of Implementation

A running example is provided in the following subsections in order to describe how the implementation assesses and refines the quality of mappings. For the example, an R2RML mapping was retrieved from the MusicBrainz project[32], which includes mappings designed to uplift data related to their music encyclopedia such as artists and associated songs. The original mapping used for the demonstration represents song lyrics in the Music Ontology[33] . The Music Ontology (`mo`) provides concepts and properties for describing music (i.e. artists, albums and tracks) on the Semantic Web. For purpose of illustration for this section, the original mapping[34] has been altered ("demo_mapping.ttl") to include 2 quality violations. These violations are outlined below:

- **Usage of undefined class (D1):** The `mo:AudioFiles` class is not defined in the Music Ontology and, therefore, undefined.
- **Usage of incorrect range (D6):** The `mo:lyrics` property is defined as an object property in the Music Ontology and, therefore, requires an object to be defined as a resource rather than a literal.

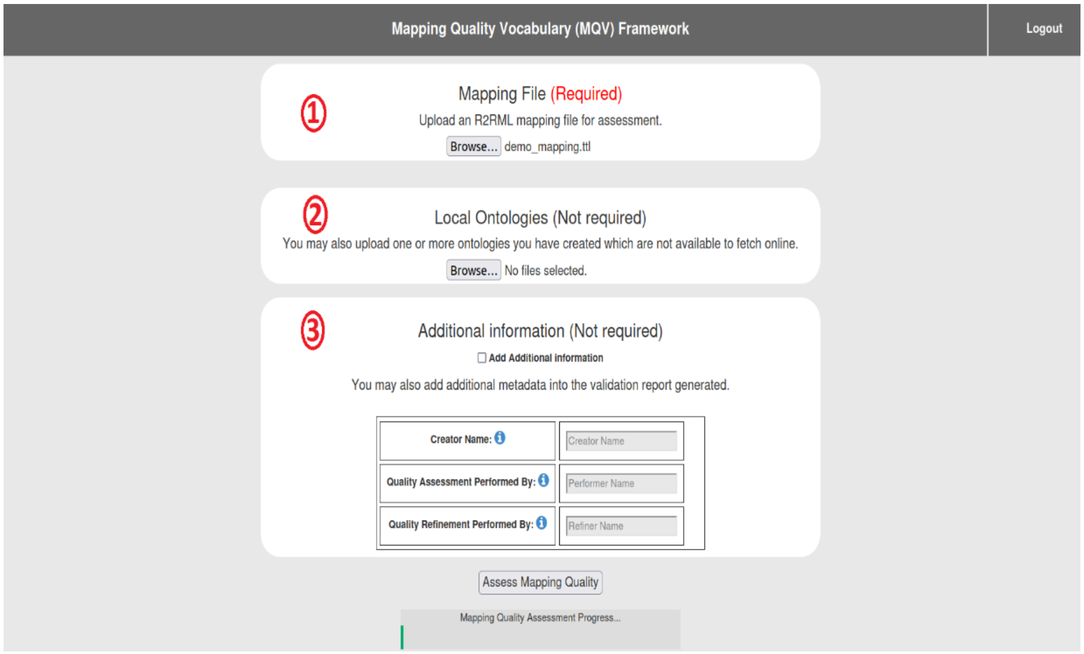
4.2.2.4.2.1 Mapping Assessment

First, users input information into the GUI presented in **Figure 20.**

[32] https://musicbrainz.org/

[33] http://musicontology.com/specification/

[34] https://github.com/metabrainz/MusicBrainz-R2RML/blob/master/mappings/work.ttl



Mapping Quality Vocabulary (MQV) Framework

Mapping File (Required)

Upload an R2RML mapping file for assessment.

Browse... demo_mapping.ttl

Local Ontologies (Not required)

You may also upload one or more ontologies you have created which are not available to fetch online.

Browse... No files selected.

Additional information (Not required)

□ Add Additional information

You may also add additional metadata into the validation report generated.

Creator Name: ℹ	Creator Name
Quality Assessment Performed By: ℹ	Performer Name
Quality Refinement Performed By: ℹ	Refiner Name

Mapping Quality Assessment Progress...

Figure 20: Screenshot of initiation of mapping quality assessment and refinement process

The information presented in **Figure 20** is outlined below:

1. An R2RML mapping needs to be uploaded to the framework for quality assessment and refinement.
2. Local ontologies in formats such as Web Ontology Language (OWL), Terse RDF Triple Language (TURTLE), XML can be uploaded if they are not published online.
3. Additional metadata relating to the agents who created the mapping and conducted the quality assessment and refinement can be input which will be included in the reports generated. The additional metadata can be used to identify which agents have interacted with the mapping, therefore, providing additional provenance information.

The published ontologies used in the mapping are needed to calculate certain metrics such as usage of incorrect property and class. Ontologies not already in the local cache are retrieved and stored in a named graph where the graph name is the namespace of the respective ontology. Storing the ontologies locally improves the efficiency of quality assessment when compared to retrieving the ontology when each metric is executed. **Figure 21** presents an overview of the process of retrieving and storing the Music ontology.

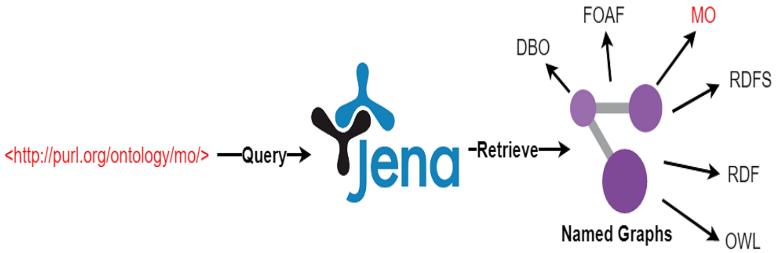


Figure 21: Process of retrieving the ontologies for calculating quality metrics

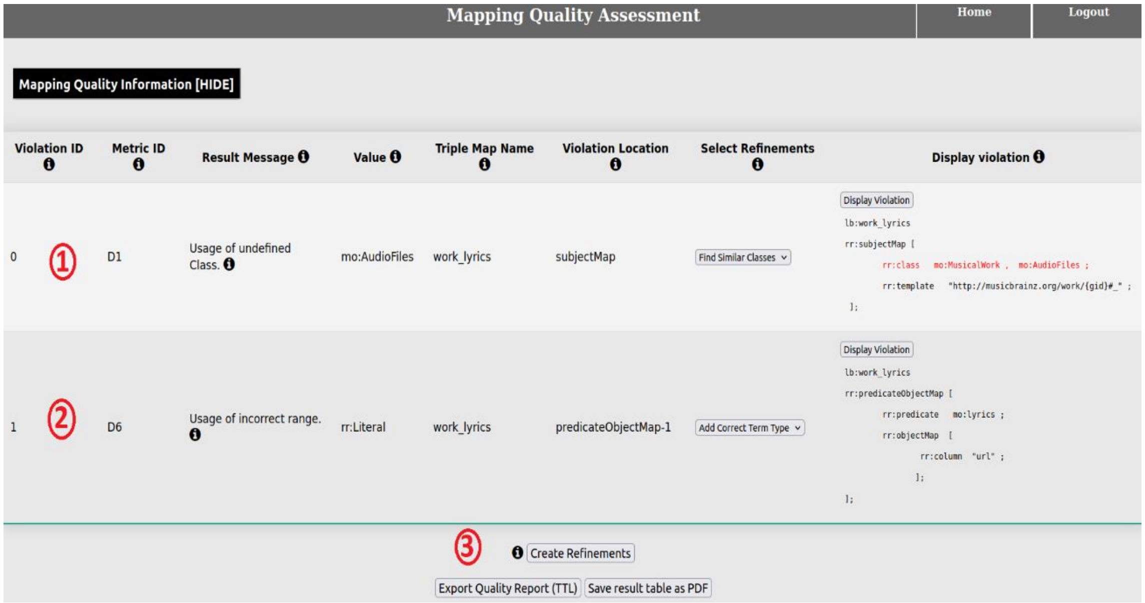
The named graph for each ontology can be retrieved by executing a SPARQL query ("retrieve_ontologies.rq") on the endpoint of the Apache Jena Fuseki triple store in order to retrieve relevant information for calculating metrics. SPARQL query templates are used to generate each query executed for each metric. **Listing 9** presents the queries which were executed to detect the quality issues in the sample mapping.

```
ASK {
 GRAPH <http://purl.org/ontology/mo/> {
   mo:AudioFiles a ?type .
   FILTER (?type IN (owl:Class, rdfs:Class))
 }
}
```

```
SELECT ?range
WHERE {
  GRAPH <http://purl.org/ontology/mo/> {
    mo:lyrics rdfs:range ?range .
  }
}
```

Listing 9: Query used to detect the D1 violation (left) and D6 violation (right) in the mapping

The queries are generated using the SPARQL templates where the values being tested (`mo:AudioFiles`, `mo:lyrics`) are inserted before execution. The quality assessment results are uplifted in RDF format using the framework's quality information R2RML mapping ("quality_report_mapping.ttl") which uses MQIO in generating the quality report for the input uplift mapping. The framework mapping is executed using the R2RML engine and the quality report output in TURTLE format. **Figure 22** presents a screenshot of the quality assessment information captured by the framework when the sample mapping was input.


Mapping Quality Assessment
Home
Logout
Mapping Quality Information [HIDE]
①
②
③
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	D1	Usage of undefined Class.	mo:AudioFiles	work_lyrics	subjectMap	Find Similar Classes	Display Violation lb:work_lyrics rr:subjectMap [rr:class mo:MusicalWork , mo:AudioFiles ; rr:template "http://musicbrainz.org/work/{gid}#_" ;];
1	D6	Usage of incorrect range.	rr:Literal	work_lyrics	predicateObjectMap-1	Add Correct Term Type	Display Violation lb:work_lyrics rr:predicateObjectMap [rr:predicate mo:lyrics ; rr:objectMap [rr:column "url" ;];];

Figure 22: Screenshot of overview of mapping quality assessment information and refinement selection

The information presented in **Figure 22** is outlined below:

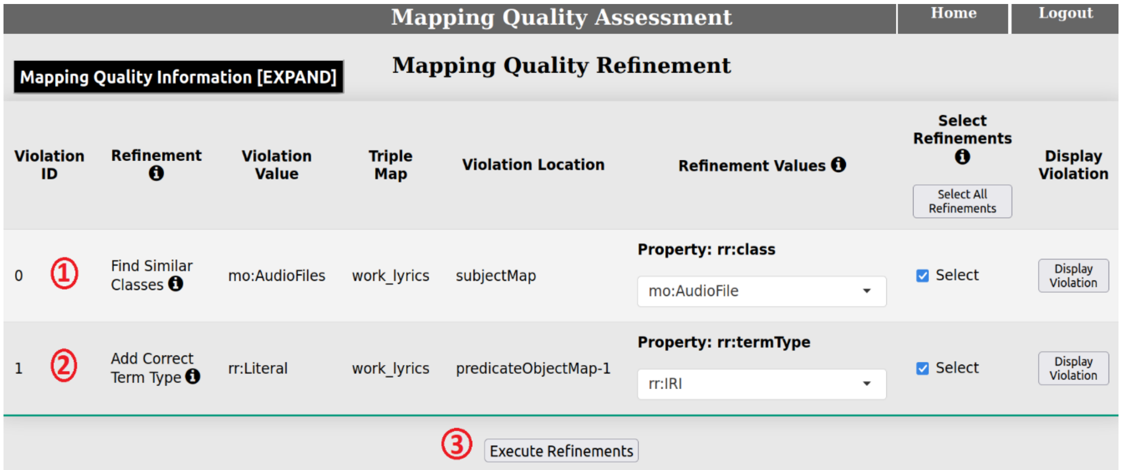
1. A description ("Usage of undefined class") associated values (mo:AudioFiles) and location ("subjectMap") of the D1 violation are shown. In addition, the extract of input uplift mapping where the violation is located is shown with the associated values highlighted in red. The hope is that the extract improves traceability of violations for users. A semi-automatic refinement ("Find Similar Classes") has been chosen to resolve the violation.
2. A description ("Usage of incorrect range") associated value (rr:Literal) and location ("predicateObjectMap-1") of the D6 violation are shown. In addition, the extract of the input uplift mapping where the violation is located is shown. However, the rr:Literal term type is inferred from the R2RML specification[35] rather than explicitly defined in the mapping. The value was inferred as no term type has been defined and it is a column-based term map. A semi-automatic refinement ("Add Correct Term Type") has been chosen to resolve the violation.
3. The quality report detailing the assessment information expressed in MQIO can be exported. In addition, the table shown can be exported to PDF format. The quality report in RDF can be linked with the resulting data to improve trustworthiness. PDF format provides a human-readable representation for human agents to exchange. Thereafter, the selected refinements can be created. For this demonstration, a semi-

[35] https://www.w3.org/TR/r2rml/#termtype

automatic refinement was selected for the D1 (“Find Similar Classes”) and D6 (“Add Correct Term Type”) violation.

4.2.2.4.2.2 Mapping Refinement

Figure 23 presents a screenshot of the refinements selected for the violations.



Mapping Quality Assessment | Home | Logout

Mapping Quality Information [EXPAND]

Mapping Quality Refinement

Violation ID		Refinement ℹ	Violation Value	Triple Map	Violation Location	Refinement Values ℹ	Select Refinements ℹ Select All Refinements	Display Violation
0	①	Find Similar Classes ℹ	mo:AudioFiles	work_lyrics	subjectMap	Property: rr:class mo:AudioFile ▾	☑ Select	Display Violation
1	②	Add Correct Term Type ℹ	rr:Literal	work_lyrics	predicateObjectMap-1	Property: rr:termType rr:IRI ▾	☑ Select	Display Violation

③ Execute Refinements

Figure 23: Screenshot of refinement execution

The information presented in **Figure 23** is outlined below:

1. **Find Similar Classes:** Defined classes within the same namespace are retrieved using a SPARQL query (“retrieve_ontology_classes.rq”). The retrieved classes are displayed in a drop down menu to the users where they can choose the desired class. The chosen class is `mo:AudioFile` which is a misspelling of the `mo:AudioFiles` class defined in the mapping. Thereafter, a SPARQL update query (“update_violation_1.rq”) is executed on the input uplift mapping, which deletes the existing class and inserts the selected class.
2. **Add Correct Term Type:** The range of the `mo:lyrics` property is retrieved by executing a query (“retrieve_range.rq”). Thereafter, the corresponding term type is suggested to the users. The term types for R2RML are `rr:IRI`, `rr:BlankNode` (Object property) and `rr:Literal` (Data type property). The user’s selection (`rr:IRI`) is inserted into the mapping using an update query (“update_violation_2.rq). As the current value (`rr:Literal`) for the term type was inferred rather than explicitly defined in the mapping, the query does not need to delete the current value. However, the update query supports the deletion of existing term types and insertion of users selected type.
3. The mentioned SPARQL update queries are created by inserting respective values into query templates. The execution of the update queries on the input uplift mapping results in the refined mapping, validation report and validation bar chart. The validation report generated by the framework, which is expressed in the MQIO includes the information in the quality report and additional uplifted information (“validation_report_mapping.ttl”) related to the executed refinements.

Figure 24 presents a screenshot of the information displayed after the refinements are executed.

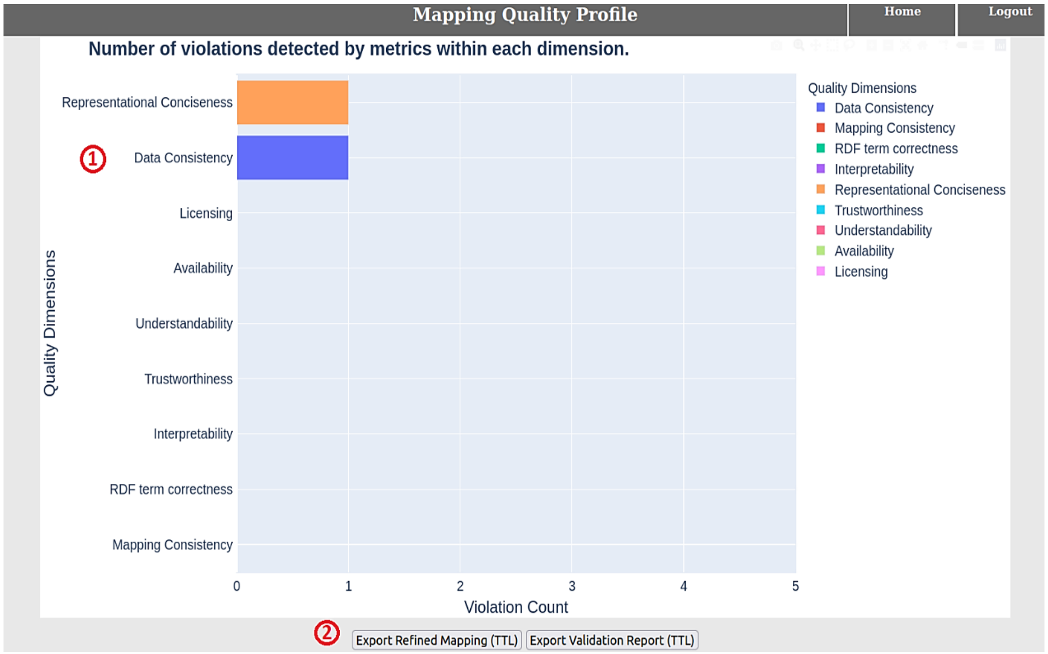


Figure 24: Screenshot of mapping quality validation bar chart

The information presented in **Figure 24** is outlined below:

1. The validation bar chart shows the relationship between each detected violation and respective quality dimensions, therefore, providing insights into the quality measurements of each dimension.
2. The refined mapping ("demo_refined_mapping.ttl") and validation ("demo_validation_report.ttl") can be exported. The refined mapping for the example contains no detectable violations.

4.3 Source Change Detection Component

The development of the **source change detection component** involved creating an ontology designed to represent source data changes in heterogeneous formats, which are captured by the framework. The functionality of the framework was extended to include support for **RML mappings** [44], which allows more diverse data source formats, such as XML, JSON and CSV. Thus, the framework can be applied to detect changes in source data represented in various formats, resulting in improved coverage.

4.3.1 Design of OSCD

The **O**ntology for **S**ource **C**hange **D**etection (**OSCD**) provides a set of classes, properties, and restrictions that can be used to represent and interchange provenance and metadata information relating to changes that occur within source data that has been used to create a LD dataset. It can also be specialized to create new classes and properties to model provenance and metadata information for domain specific changes. Representing source data changes in an ontology-based format enables software agents to automatically process and propagate them appropriately into respective LD. The **documentation** describing OSCD is available online (https://w3id.org/OSCD). This section discusses the design methodology of the ontology. Thereafter, the concepts and relationships defined as a result of the development process are presented.

4.3.1.1 Design Methodology

The design of the OSCD followed best practices as recommended by the semantic web community. Ontology design practices were reused from the most prominent ontology design methodologies. The methodologies included the NeON methodology [134], UPON Lite [100], Ontology development 101: A guide to creating your first ontology [103] and LOT: An industrial oriented ontology engineering framework [110].

The design methodology followed is as follows:

1. **Identify aims, objectives, scope:** The design process commenced with the identification of the aims, objectives and scope of the ontology, which are outlined in **Appendix S** of this document. The template used for the table was retrieved from the methodologies and used to define the ontology requirements specification document. The document outlines requirements and among other things, the aims, objectives and scope of the ontology.
2. **Identify and analyze relevant information:** A review of publications in the state of the art was conducted to identify relevant information. Publications within the state of the art which related to topics within the defined scope were reviewed to facilitate the retrieval of relevant information. Thereafter, the retrieved information was used to formalize competency questions. References to publications which inspired the creation of each competency question are defined.
3. **Create use-cases and competency questions:** Competency questions were created during the design process of the ontology. The questions define the functional requirements of the ontology and were iteratively refined until an accurate representation of the requirements and objectives was conceived. The final iteration of the questions is presented in **Appendix T**. Use cases were devised in order to refine the requirements of the ontology and were retrieved from the RML test case files[36]. The test case files provided a diverse set of source data in formats such as XML, JSON, relational databases and CSV as well as respective RML mappings. In addition, the R2RML test case files[37] were used, however, the source data is only represented in relational format. The test cases facilitated the creation of use cases through the generation of graphs defined in OSCD when changes were detected between the file versions. The use case has been documented in a publication [116]. A use case graph generated by the RML test cases is available ("oscd_sample_graph.ttl") in the GitHub. In addition, OSCD was applied to a network management use case described in **Section 5.7.2**.

[36] https://rml.io/test-cases/

[37] https://www.w3.org/2001/sw/rdb2rdf/test-cases/

4. **Identify concepts and relationships:** Concepts and relationships were identified through the state of the art review and the researchers previous experience in the creation of LD. The concepts and relationships were iteratively defined until the information modeling provided by the ontology satisfied each of the competency questions. In addition, concepts and relationships were reused from existing vocabularies as recommended by the methodologies and the W3C recommendation on Data on the Web Best Practices [87]. The reused ontologies included an ontology for Linking Open Descriptions of Events (LODE) [126], which is designed to represent events in the LD domain. LODE was extended to model changes as specialized events which have occurred in source data. LODE is a prominent ontology, which has been cited over 300 times as of 2023. In addition, the ontology has been reused by a prominent existing related approach, named DSNotify (see Section 2.2.2) in order to represent changes in resources of LD datasets. Therefore, it was decided to represent changes detected in source data similarly. The Rei ontology [77] is a universal policy ontology designed to model policies for various domains. The ontology was reused to represent the details of the notification policy, which is used to inform maintainers of relevant changes in a timely manner. Rei was reused as it is a prominent domain independent policy ontology used in the LD domain. The FOAF [57] ontology was reused similar to represent the agents involved in the activities. It was decided to reuse FOAF as it is a prominent ontology, which is often used to represent agents in various LD domains. The RDF Schema (RDFS) [21] vocabulary was reused to describe the changed data resource. It was decided to reuse as RDFS as it is a prominent vocabulary, which was designed to represent heterogenous information related to RDF resources. The reuse in OSCD is demonstrated in the competency questions and ontology documentation.
5. **Progressive iterations:** Steps 2-4 were iteratively repeated until the point when the proposed concepts/relationships provided information which satisfied each requirement defined in the form of a competency question.
6. **Create Ontology:** The ontology was implemented in OWL 2 Web Ontology Language [92]. Concepts and relationships which were defined in the previous step were constructed using Protégé ontology development tool [154]. In addition, semantic reasoners were also utilized to detect and remove logical inconsistencies within the ontology.
7. **Evaluate:** The ontology was evaluated for sufficiency to provide information to fulfill each competency question. The usage of a semantic reasoner within Protege ensured logical inconsistencies were identified. In addition, OOPS! [111] was used to detect common ontology design issues. The quality of metadata and documentation was evaluated through presentation in peer reviewed publications. Feedback received from reviewers allowed areas for improvement to be identified. Peer reviewed publications related to OSCD are outlined in **Section 1.5.3**. Further expert feedback was received in a previous user evaluation where they were asked to provide feedback on the design and application of the ontology. In addition, the ontology was presented to a panel of semantic web experts at the semantic interoperability conference (SEMIC2022)[38] organized by the European commission. Each graph generated in the use cases was assessed with the RDFUnit [81] quality assessment framework which provides test driven validation of RDF data.
8. **Publication:** The ontology documentation was created using WIDOCO [54] which is a tool designed to use ontology metadata to create HTML documents listing its classes and properties. Thereafter, the ontology

[38] https://joinup.ec.europa.eu/collection/semic-support-centre/semic-conference

and human readable documentation were published with a permanent identifier as a FAIR resource including an open and permissive license. The documentation contains information about the creation, design, usage, class interaction diagrams and provides various serializations.

The resulting ontology from this design methodology was **version 1.0**.

4.3.1.2 Concepts and Relationships of OSCD

Figure 25 presents the ontology classes and properties for OSCD. The **yellow** boxes shown represent concepts introduced by OSCD and **blue** boxes represent concepts reused from ontologies in the state of the art.

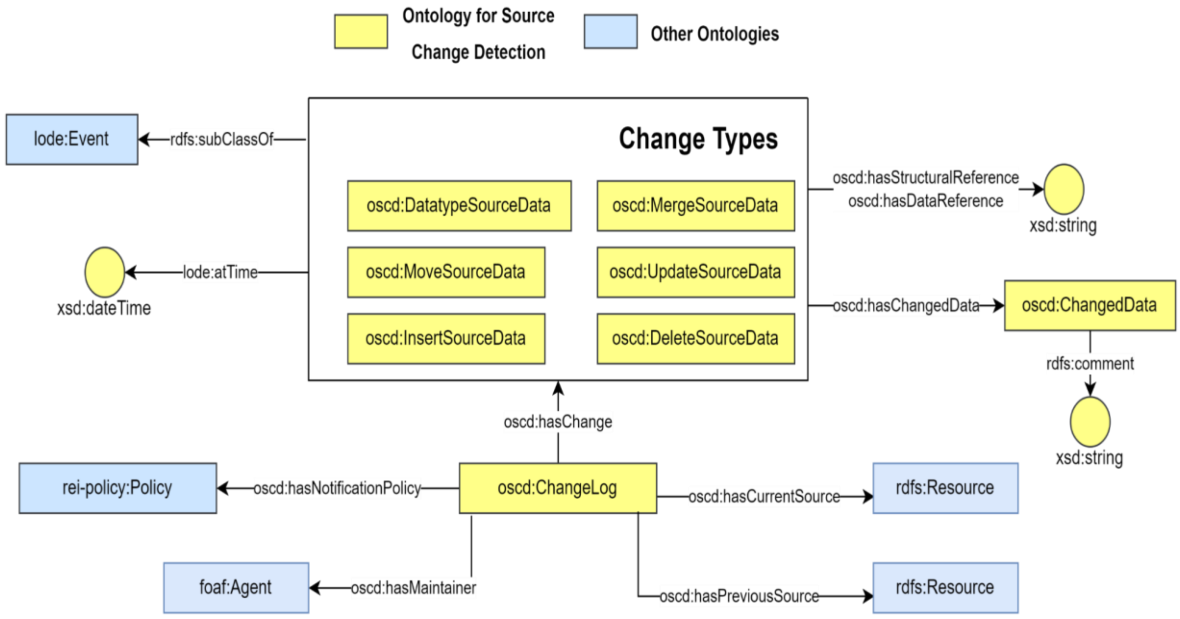
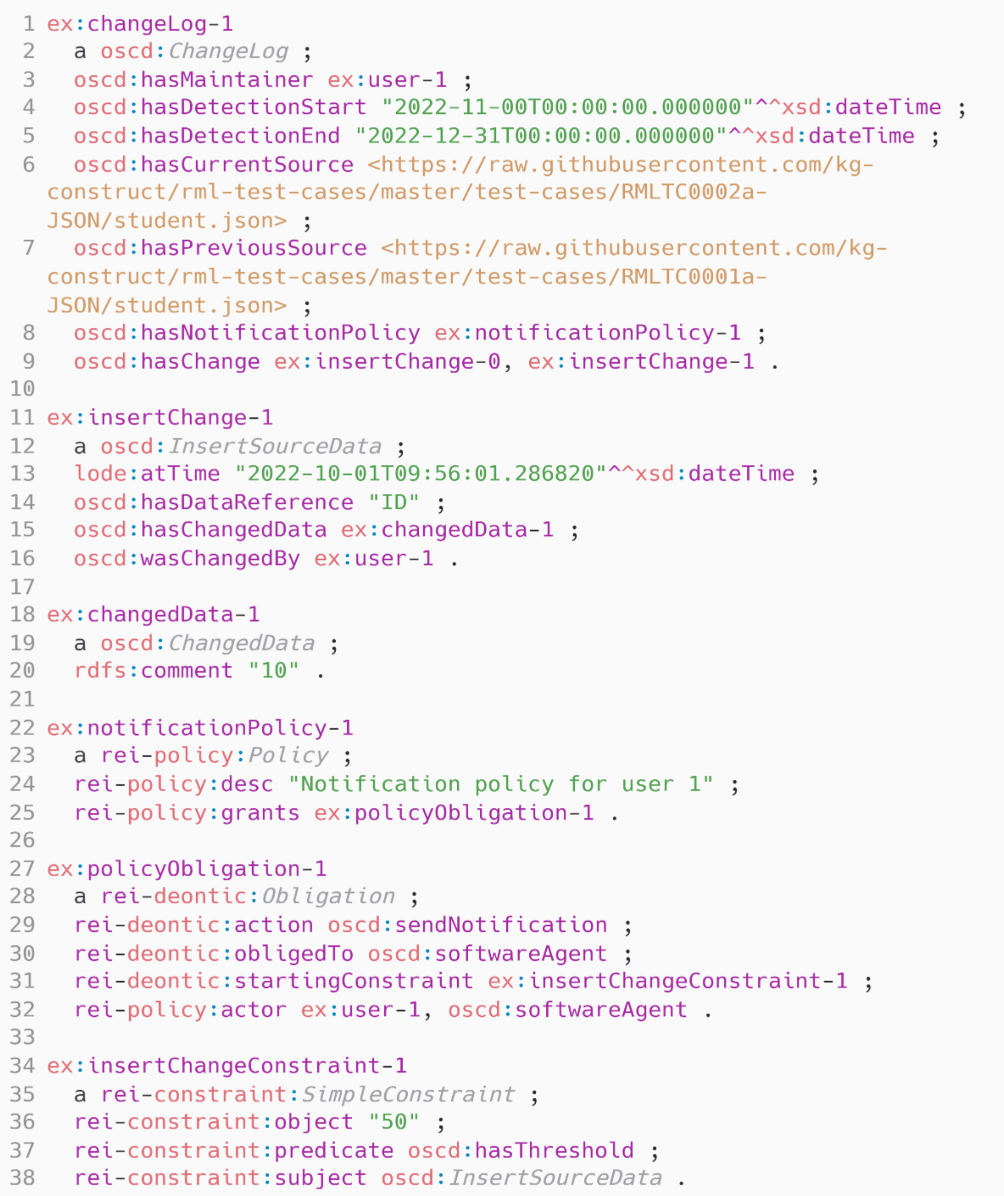


Figure 25: OSCD classes and properties usage during change detection

The detection of changes within the source data is initiated by an agent (`foaf:Agent`) who is responsible for maintaining the data (`oscd:hasMaintainer`). Changes will be detected between a previous version (`oscd:hasPreviousSource`) and the current version (`oscd:hasCurrentSource`). Once the change detection process has been initiated, changes are detected between when the process began (`oscd:hasDetectionStart`) and a predetermined end date (`oscd:hasDetectionEnd`) which is defined by the maintainer. A notification policy (`oscd:hasNotificationPolicy`) can be defined which is represented with the Rei policy ontology. The notification policy allows agents to be notified of changes when a certain amount has occurred. Therefore, ensuring the maintainers are informed of changes when required.

Changes which occur in a specific source data are grouped into a log (`oscd:ChangeLog`). Each change (`oscd:hasChange`) within a change log is represented as a specialized event (`lode:Event`) which has resulted in a change in the source data used to generate LD. Different change types exist within the ontology which allow the cause of the change to be captured. For instance, a value is inserted (`oscd:InsertSourceData`) in the source data. Additional information related to the change such as the data inserted (oscd:hasChangedData) or its location within the source data (oscd:hasDataReference). The model can be extended to capture domain specific changes. **Listing 10** presents an extract of a sample graph expressed in the OSCD.



```
ex:changeLog-1
  a oscd:ChangeLog ;
  oscd:hasMaintainer ex:user-1 ;
  oscd:hasDetectionStart "2022-11-00T00:00:00.000000"^^xsd:dateTime ;
  oscd:hasDetectionEnd "2022-12-31T00:00:00.000000"^^xsd:dateTime ;
  oscd:hasCurrentSource <https://raw.githubusercontent.com/kg-construct/rml-test-cases/master/test-cases/RMLTC0002a-JSON/student.json> ;
  oscd:hasPreviousSource <https://raw.githubusercontent.com/kg-construct/rml-test-cases/master/test-cases/RMLTC0001a-JSON/student.json> ;
  oscd:hasNotificationPolicy ex:notificationPolicy-1 ;
  oscd:hasChange ex:insertChange-0, ex:insertChange-1 .

ex:insertChange-1
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-01T09:56:01.286820"^^xsd:dateTime ;
  oscd:hasDataReference "ID" ;
  oscd:hasChangedData ex:changedData-1 ;
  oscd:wasChangedBy ex:user-1 .

ex:changedData-1
  a oscd:ChangedData ;
  rdfs:comment "10" .

ex:notificationPolicy-1
  a rei-policy:Policy ;
  rei-policy:desc "Notification policy for user 1" ;
  rei-policy:grants ex:policyObligation-1 .

ex:policyObligation-1
  a rei-deontic:Obligation ;
  rei-deontic:action oscd:sendNotification ;
  rei-deontic:obligedTo oscd:softwareAgent ;
  rei-deontic:startingConstraint ex:insertChangeConstraint-1 ;
  rei-policy:actor ex:user-1, oscd:softwareAgent .

ex:insertChangeConstraint-1
  a rei-constraint:SimpleConstraint ;
  rei-constraint:object "50" ;
  rei-constraint:predicate oscd:hasThreshold ;
  rei-constraint:subject oscd:InsertSourceData .
```

Listing 10: Sample graph expressed in the OSCD

The sample graph shows a grouping of changes (ex:changeLog-1) which have been detected between the original (oscd:hasPreviousSource) and current (oscd:hasCurrentSource) version of source data. One change has been detected (ex:insertChange-1). Data references (oscd:hasDataReference) and changed data (oscd:hasChangedData) are associated with the change. In addition, the agent responsible (oscd:wasChangedBy) for the change is represented. A notification policy (ex:notificationPolicy-1) has been defined with a threshold (`ex:insertChangeConstraint-1`) of a value of 50, therefore, a notification will be sent once that number of changes type occurs.

4.3.2 Design of Source Change Detection Component

First, the detection of source data changes is discussed. Thereafter, the linking of data source changes with their respective mappings and then the validation of the notification policy is described.

4.3.2.1 Overview of Design

Figure 26 presents a component diagram of the source change detection component of the framework.

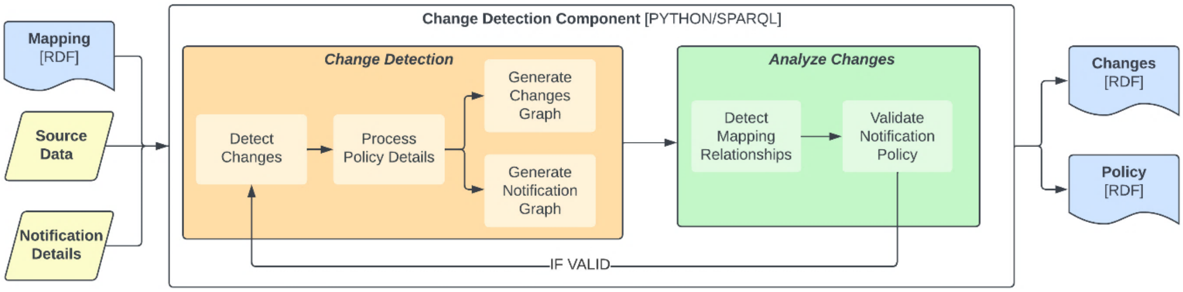


Figure 26: Component diagram of the source change detection component of the MQI framework

The components presented in **Figure 26** are described below:

- **Input:** Two versions of source data and respective uplift mappings are input into the framework. Oftentimes, the mapping will be already uploaded to the mapping quality assessment and refinement component. Related mappings are needed to link detected changes. In addition, a notification policy can be created to define when users will be notified of changes. Notifications allow users to be informed of changes in a timely manner.
- **Change Detection:** Changes are detected between the versions using existing methods. Thereafter, the changes and notification details are uplifted in RDF format, resulting in two named graphs, which are used to link changes with respective mappings and to validate notification policy constraints.
- **Analyze Changes:** The detected changes are linked with the input uplift mappings in order to identify data source changes which could impact them. In addition, the changes can be used to indicate that the current source does not accurately represent the underlying data sources. The notification policy is periodically queried to validate the defined thresholds and end date.
- **Output:** The results of the process are the named graphs which detail the changes detected and notification policy details. The changes are periodically detected until the notification policy becomes invalid by the fulfillment of a threshold or end date. Thereafter, a notification will be sent detailing

detected changes and links with respective mappings. The details in the notification can be used by data maintainers to take appropriate actions to preserve the alignment between source data and respective mappings.

4.3.2.2 Change Detection

Figure 27 presents the components involved in detecting changes in the source data.

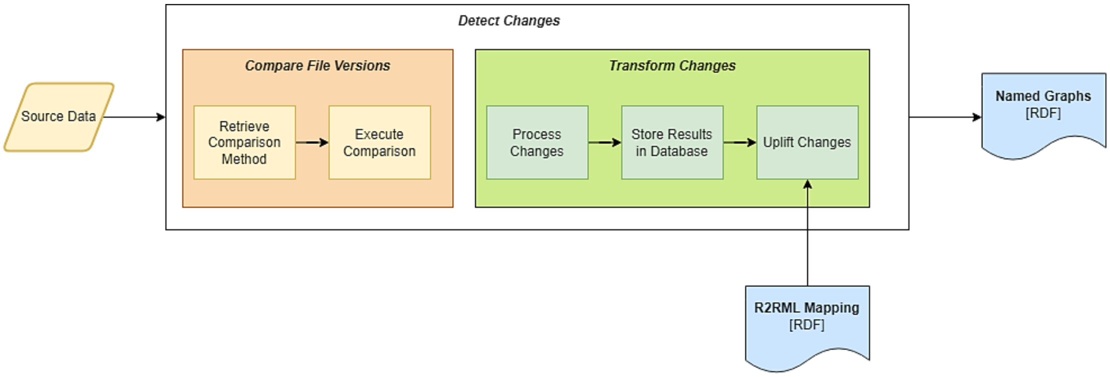


Figure 27: Components involved in detection of changes

The changes which have occurred between the original and current version of the source data are detected through existing file comparison methods[39] which are designed to calculate the differences and similarities between data objects. Existing methods include diff, cmp, Beyond Compare, and File Compare. It was decided to reuse these existing approaches rather than design a new one, as comparison methods are out of scope of the research objectives of this thesis. Thereafter, the detected changes and notification policy details are converted into a format suitable for uplift to RDF using an R2RML mapping stored by the framework. The uplift process is executed and the resulting graph is output.

[39] https://blog.pics.io/11-best-file-comparison-software/

4.3.2.3 Analyzing Changes

Figure 28 presents the components involved in analyzing the detected changes.

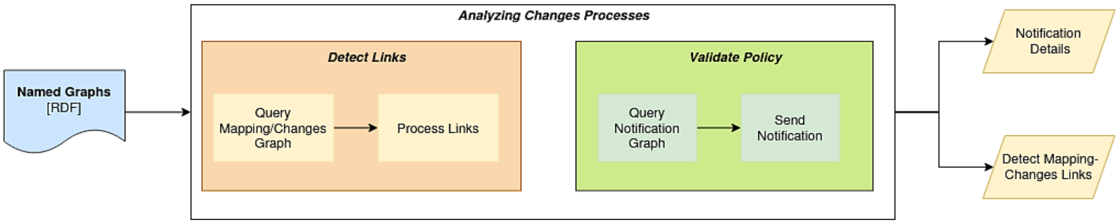


Figure 28: Components involved in analyzing detected changes

The changes are linked with respective uplift mappings by executing a query which targets both the changes detected named graph and the uplift mappings named graph . In addition, the changes are compared to thresholds to validate the notification policy with a similar query on the notification named graph. The result of the process is notification details (if applicable) and detected links between changes and their respective mappings.

4.3.2.4 Implementation

This section discusses the implementation of the source change detection component of the MQI framework. **Supplementary information related to this section is stored in a folder ("/Second-Iteration-Implementation") of the GitHub.**

4.3.2.4.1 Overview of Implementation

Figure 29 presents how the technologies are used within the implementation of the source change detection component of the MQI framework.

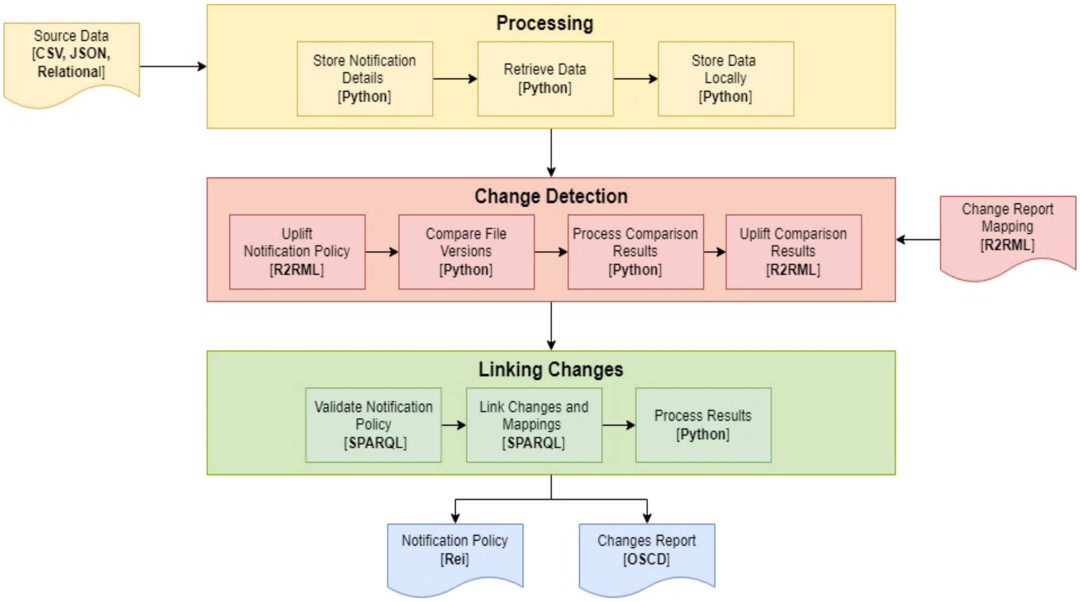


Figure 29: Overview of technologies used in the implementation of the source change detection component

The components presented in **Figure 29** are outlined below.

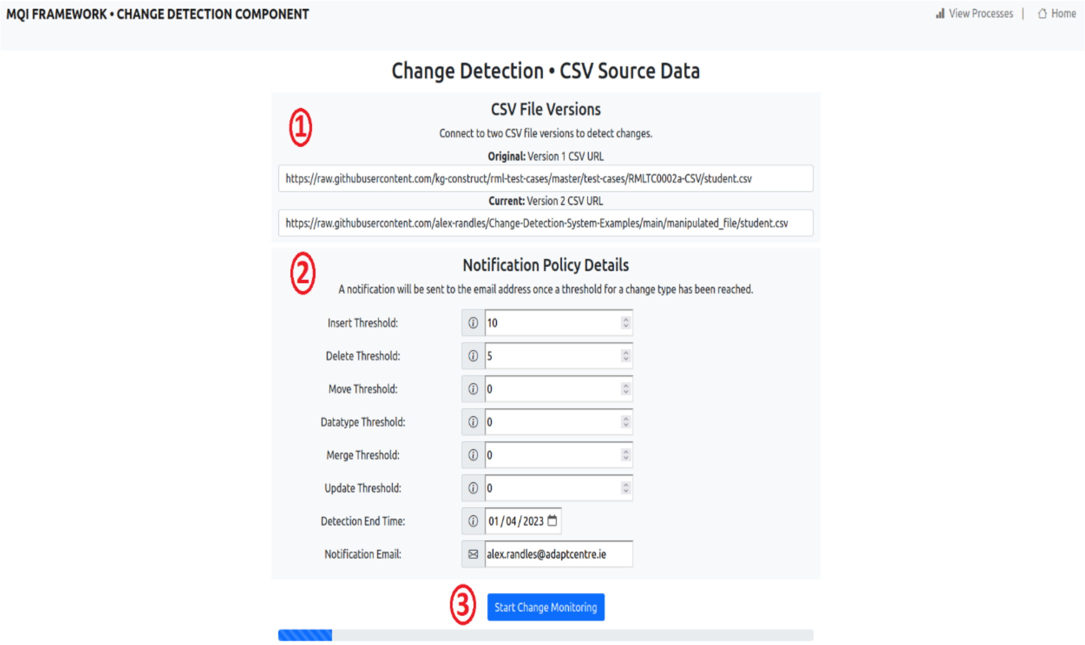
- **Processing:** First, the notification details input into the framework are stored using Python built-in functions. Thereafter, the previous and current version of source data are retrieved using Pythons built-in requests library[40]. The library allows the data to be fetched using a HTTP request and stored locally for comparison. HTTP is used as it allows remote data to be retrieved rather than requiring users to upload large data stores.
- **Change Detection:** The stored notification policy details are uplifted by executing an R2RML mapping expressed in the Rei policy ontology [77] using an R2RML engine[41] and the resulting RDF data stored locally. Thereafter, the versions of source data which were retrieved are compared using diff libraries in Python, such as XMLDiff and CSVDiff. The results of the comparison are processed using built-in Python functions to retrieve relevant information. Thereafter, an R2RML mapping expressed in OSCD [116] is used by the framework to uplift the information into RDF format.
- **Linking Changes:** First, the notification policy is validated by linking changes with defined thresholds using a SPARQL query executed by the RDFLib library [82] in Python. A notification can be sent using the smtplib library[42] in Python, which can be used to send an email to the address retrieved from the notification policy using a SPARQL query. Source data changes are linked with respective mappings using a query in order to identify changes which may impact the alignment between them. The result of the process is two files containing RDF data, which were generated by the framework and contain the notification policy expressed in Rei policy ontology and detected changes report expressed in the OSCD.

The steps involved in the process are outlined in the following **subsections.**

4.3.2.4.2 Change Detection

First, the details of the versions of the source data are input into the framework. **Figure 30** presents a screenshot of the input fields.

[40] https://pypi.org/project/requests/
[41] https://github.com/chrdebru/r2rml-tutorial
[42] https://docs.python.org/3/library/smtplib.html



Change Detection • CSV Source Data

CSV File Versions

Connect to two CSV file versions to detect changes.

Original: Version 1 CSV URL

https://raw.githubusercontent.com/kg-construct/rml-test-cases/master/test-cases/RMLTC0002a-CSV/student.csv

Current: Version 2 CSV URL

https://raw.githubusercontent.com/alex-randles/Change-Detection-System-Examples/main/manipulated_file/student.csv

Notification Policy Details

A notification will be sent to the email address once a threshold for a change type has been reached.

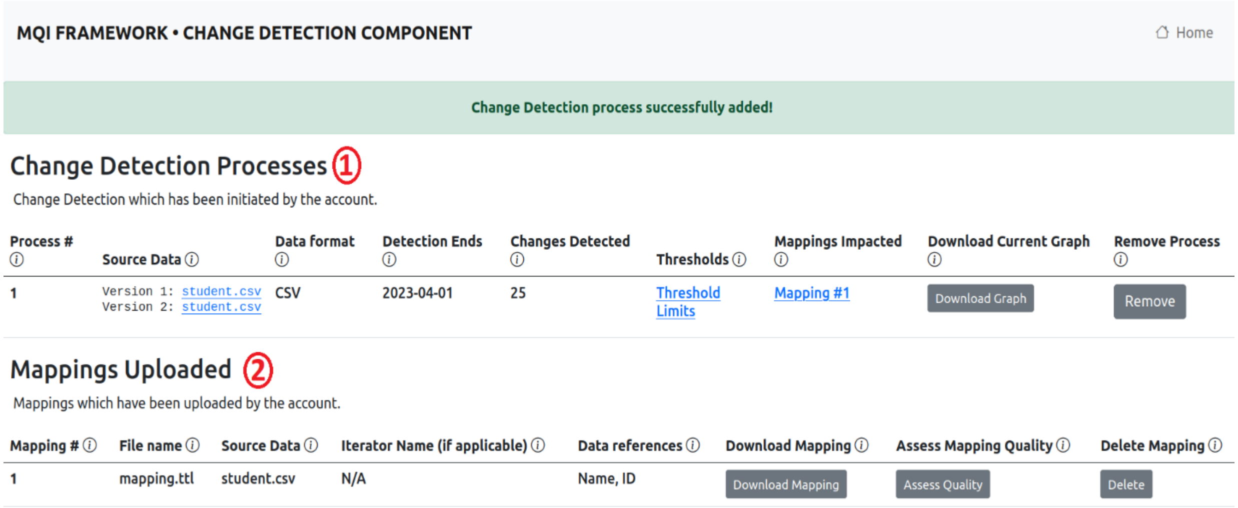
Insert Threshold:	10
Delete Threshold:	5
Move Threshold:	0
Datatype Threshold:	0
Merge Threshold:	0
Update Threshold:	0
Detection End Time:	01/04/2023
Notification Email:	alex.randles@adaptcentre.ie

Figure 30: Screenshot of initiation of change detection processes

The information presented in **Figure 30** is outlined below:

1. The original and current version are input into framework which allows changes to be detected.
2. The details of the notification policy are input which includes the thresholds, detection end date and an email address where the notification will be sent. The details are used to generate the RDF representation of the policy.
3. Changes are detected using third party libraries such as XMLDiff (XML format)[43] and CSVDiff (CSV format)[44] and uplifted into RDF format using an R2RML mapping (“change_detection_mapping.ttl”). Thereafter, SPARQL queries (“retrieve_overview.rq”) are executed on the graph to display an overview (**Figure 31**) of the different change detection processes and mappings uploaded to the framework.

[43] https://pypi.org/project/xmldiff/
[44] https://pypi.org/project/csvdiff/



MQI FRAMEWORK • CHANGE DETECTION COMPONENT

⌂ Home

Change Detection process successfully added!

Change Detection Processes ①

Change Detection which has been initiated by the account.

Process # ⓘ	Source Data ⓘ	Data format ⓘ	Detection Ends ⓘ	Changes Detected ⓘ	Thresholds ⓘ	Mappings Impacted ⓘ	Download Current Graph ⓘ	Remove Process ⓘ
1	Version 1: student.csv Version 2: student.csv	CSV	2023-04-01	25	Threshold Limits	Mapping #1	Download Graph	Remove

Mappings Uploaded ②

Mappings which have been uploaded by the account.

Mapping # ⓘ	File name ⓘ	Source Data ⓘ	Iterator Name (if applicable) ⓘ	Data references ⓘ	Download Mapping ⓘ	Assess Mapping Quality ⓘ	Delete Mapping ⓘ
1	mapping.ttl	student.csv	N/A	Name, ID	Download Mapping	Assess Quality	Delete

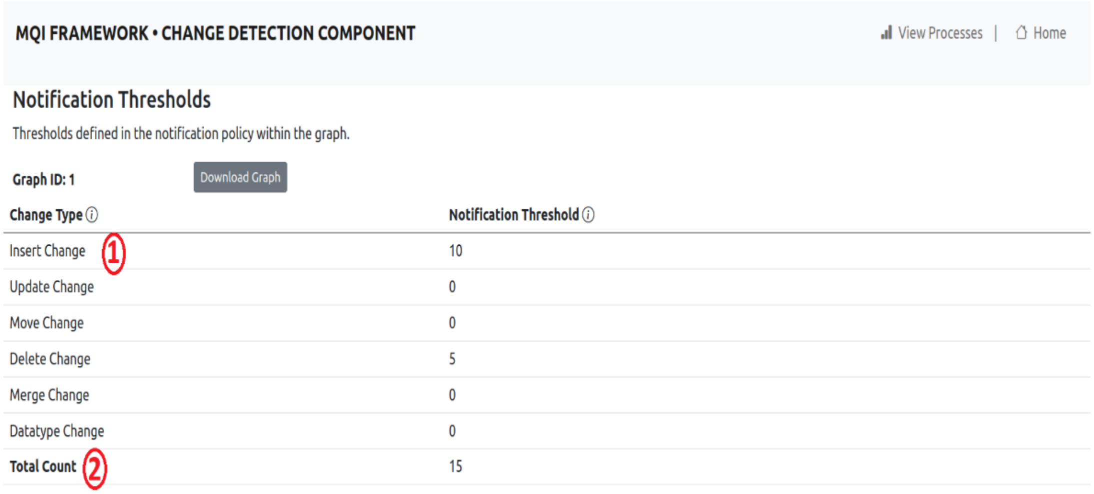
Figure 31: Screenshot of overview of change detection processes and mappings uploaded to the framework

The information presented in **Figure 31** is outlined below:

1. The overview of the **change detection processes** presents hyperlinks to the versions of the source data, format of data, number of changes detected, thresholds related to defined notification policies and uplift mappings impacted by respective source data changes. In addition, buttons to download the generated graph and ability to remove the process in order to terminate the change detection in a source data. The overview can be used to identify which change detection processes are still running.
2. The overview of the **mappings uploaded** presents information related to mappings which have been uploaded to the assessment and refinement component of the framework. These mappings have been linked with respective source data and are identified through an ID. The overview shows the name, source data and data references in each mapping. In addition, relevant quality reports generated by the mapping quality assessment and refinement component can be downloaded.

4.3.2.4.3 Notification Policy

The details of the notification policy which were input into the interface presented in **Figure 30** are uplifted by the framework into RDF format using an R2RML mapping ("notification_policy_mapping.ttl") expressed in Rei policy ontology [34]. Thereafter, the graph is queried ("retrieve_thresholds.rq") to generate an overview of the thresholds defined (**Figure 32**).



MQI FRAMEWORK • CHANGE DETECTION COMPONENT

Notification Thresholds

Thresholds defined in the notification policy within the graph.

Graph ID: 1

Change Type	Notification Threshold
Insert Change	10
Update Change	0
Move Change	0
Delete Change	5
Merge Change	0
Datatype Change	0
Total Count	15

Figure 32: Screenshot of overview of thresholds defined in notification policy

The information presented in **Figure 32** is outlined below:

1. The thresholds defined in the notification policy are displayed in respect to the associated change type. The information provides an insight into how many changes have occurred and whether the notification policy needs to be updated.
2. The total number of changes defined in the thresholds is shown to provide an indication of when a notification will be sent.

4.3.2.4.4 Analyzing Changes

This section discusses the linking of source data changes and respective mappings and the validation of associated notification policies.

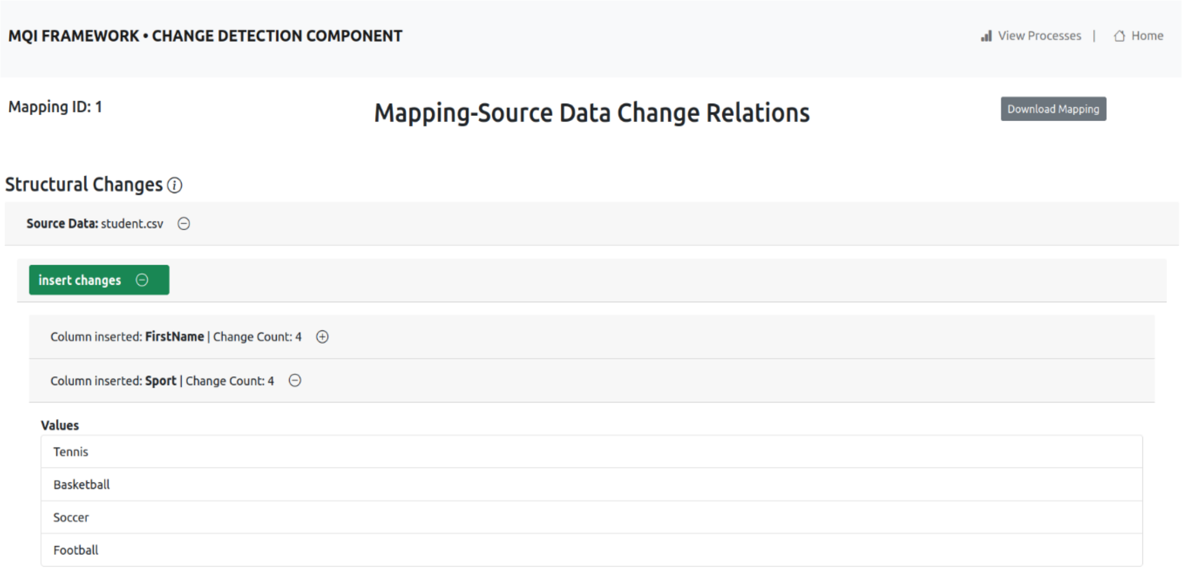
4.3.2.4.4.1 Linking with Mappings

The details of changes are linked with respective mappings using the SPARQL query presented in **Listing 11**. The retrieved links provide indications of changes which could impact the alignment between respective source data and mappings.

```
SELECT ?tripleMap ?source ?change ?reference ?changeDescription
WHERE {
  GRAPH ?changesGraph {
    ?changeLog a oscd:ChangeLog;
            oscd:hasChange ?change;
            oscd:hasCurrentSource ?currentSource .
    ?change oscd:hasDataReference ?reference;
            oscd:hasChangedData ?changedData .
    ?changedData rdfs:comment ?changeDescription .
    BIND (REPLACE(STR(?currentSource), "^.*/([^/]*)$", "$1") as ?source)
  }
  GRAPH ?mappingGraph {
    ?tripleMap rml:logicalSource|rr:logicalTable ?logicalSource;
               rr:predicateObjectMap ?pom .
    ?logicalSource rml:source|rr:tableName ?source .
    ?pom rr:objectMap ?objectMap .
    ?objectMap rml:reference|rr:column ?reference .
  }
}
```

Listing 11: Query used to detect links between changes and respective mappings

The query retrieves values from two graphs which represent the source data changes (`?changesGraph`) and respective uplift mapping (`?mappingGraph`) graph. The values include the source defined in the mapping (`?source`) and the data referenced in it (`?reference`). These values are compared with changes detected (`?changeLog`) in the same source data (`?source`). The references defined in the mapping are compared with the data references (`?reference`) related to the changes which can indicate that the change could impact that mapping. For instance, the name of a column referenced in the uplift mapping is updated in the source data and not in the mapping. **Figure 33** presents a screenshot of the framework displaying the information retrieved from the query executed on data used in evaluation described in **Section 5.5.2.4**.



MQI FRAMEWORK • CHANGE DETECTION COMPONENT

Mapping ID: 1

Mapping-Source Data Change Relations

Structural Changes

Source Data: student.csv

Column inserted: **FirstName** | Change Count: 4

Column inserted: **Sport** | Change Count: 4

Values

Tennis

Basketball

Soccer

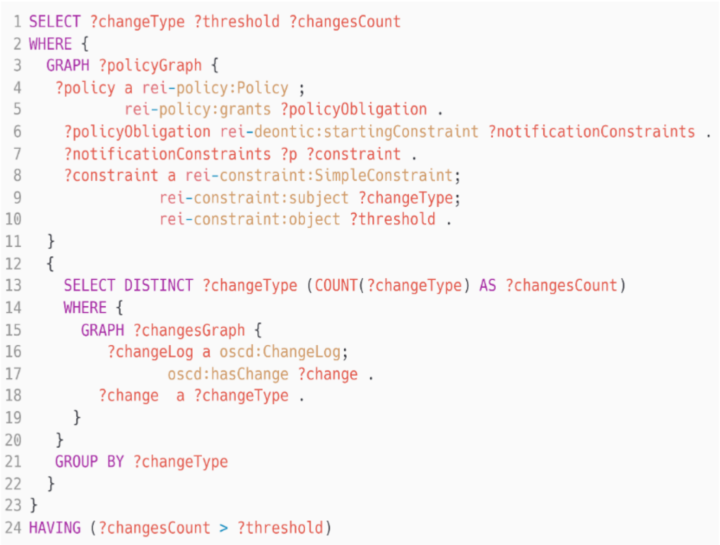
Football

Figure 33: Overview of detected links between source data changes and respective mapping

The screenshot (**Figure 33**) presents changes which have occurred to source data (“student.csv”). The changes include 4 values being inserted into two new columns (“Sport”, “FirstName”). In addition, the respective mapping can be downloaded for examination. The change information shown can be used by data maintainers in order to identify if the LD should be regenerated, therefore, preserving the freshness of the data, by providing the most up to date representation of the underlying data sources [141]. In addition, changes related to the structure of the data, such as the column insertion shown, should be examined by maintainers in order to preserve alignment between mappings and source data [43,141].

4.3.2.4.4.2 Validation of Notification Policy

The notification policy graph is validated using the SPARQL query presented in **Listing 12**. The query allows thresholds in the notification policy to be compared with detected changes in order to identify thresholds which have been fulfilled, therefore, indicating a notification should be sent.



```
SELECT ?changeType ?threshold ?changesCount
WHERE {
  GRAPH ?policyGraph {
   ?policy a rei-policy:Policy ;
           rei-policy:grants ?policyObligation .
    ?policyObligation rei-deontic:startingConstraint ?notificationConstraints .
    ?notificationConstraints ?p ?constraint .
    ?constraint a rei-constraint:SimpleConstraint;
               rei-constraint:subject ?changeType;
               rei-constraint:object ?threshold .
  }
  {
    SELECT DISTINCT ?changeType (COUNT(?changeType) AS ?changesCount)
    WHERE {
      GRAPH ?changesGraph {
         ?changeLog a oscd:ChangeLog;
                oscd:hasChange ?change .
        ?change  a ?changeType .
     }
   }
   GROUP BY ?changeType
  }
}
HAVING (?changesCount > ?threshold)
```

Listing 12: Query used to validate notification policy

The query retrieves values from the graphs containing the source data changes (?changesGraph) and the notification policy (?policyGraph), which have been generated by the framework. The retrieved values from the notification policy details (?policy) include the current number of changes for a specific type (?changeCount) and respective threshold (?threshold) which are compared (HAVING) in order to identify if a notification should be sent. Thereafter, an additional query ("retrieve_contact_details.rq") is executed to retrieve the email address where the notification should be sent. The process is periodically repeated following the same steps until the detection end date is reached which can be validated by executing a different query ("retrieve_detection_period.rq").

4.4 Interaction of Framework Components

Figure 34 presents the interaction between the components of the MQI framework.

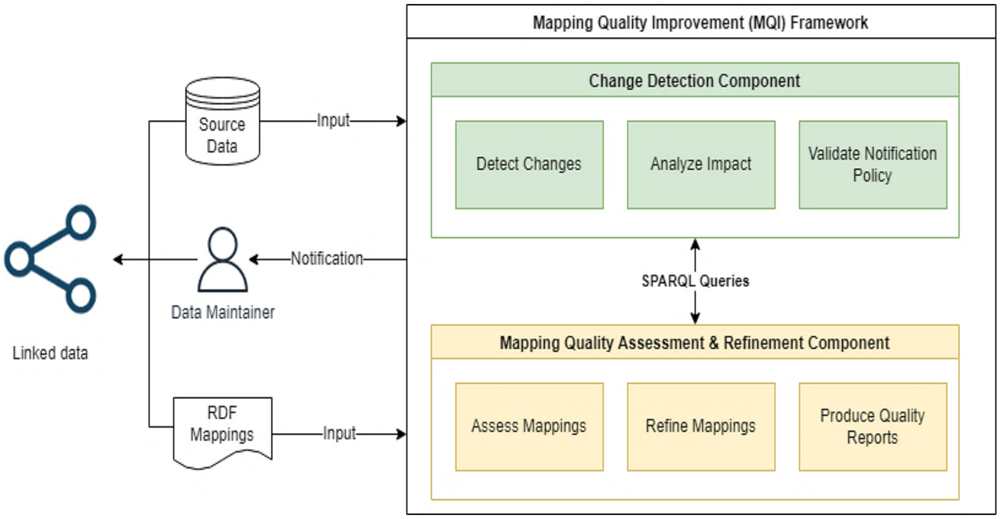
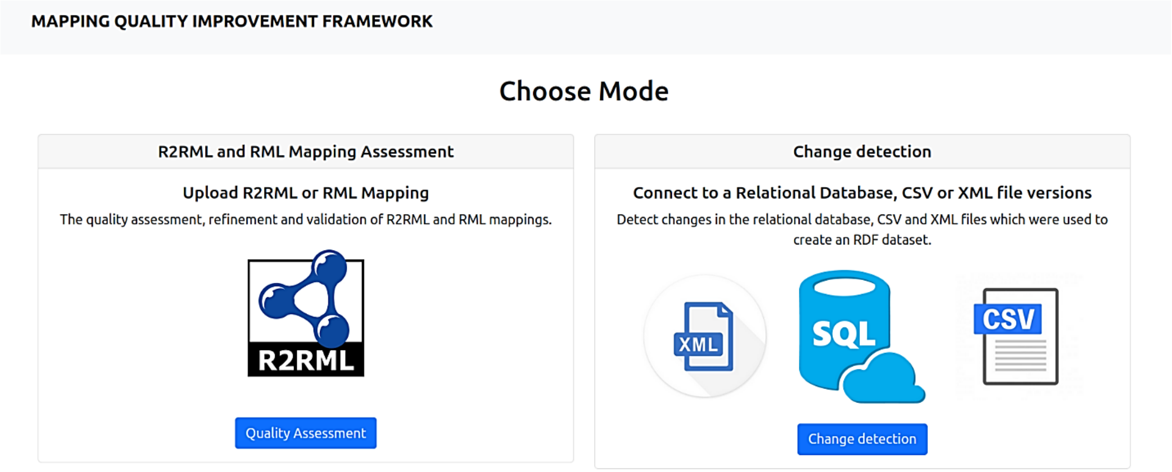


Figure 34: Interaction between components of MQI Framework

Figure 35 presents a screenshot of the implementation of the final version of the framework.


MAPPING QUALITY IMPROVEMENT FRAMEWORK

Choose Mode

R2RML and RML Mapping Assessment

Upload R2RML or RML Mapping

The quality assessment, refinement and validation of R2RML and RML mappings.

Change detection

Connect to a Relational Database, CSV or XML file versions

Detect changes in the relational database, CSV and XML files which were used to create an RDF dataset.

Figure 35: Screenshot of implementation of mapping quality assessment (left) and source change detection component (right)

Redirection to both components is presented as an option when users initially access the framework. Pressing the *"Quality Assessment"* and *"Change Detection"* button will redirect them to the screens presented in **Figure 20** and **Figure 30**, respectively. SPARQL queries are used by the framework to link information captured by the components, which is hoped to aid users in preserving alignment and quality conformance to metrics described in

Section 4.2.2.1. The RDF graphs provided to users who interact with both components include three named graphs and represent information on the 1) quality assessment and refinement of the mapping 2) changes detected in the source data of the mapping and 3) details of defined notification policies associated with change detection. The graphs generated by the framework can be linked with the mapping to improve discovery, maintenance and reuse [3]. In addition, they can be linked with resulting dataset to improve trustworthiness for consumers by providing additional detailed provenance in machine-readable format [42,75].

4.5 Chapter Summary

The fulfillment of each defined requirement is described below.

- **R1) The framework should facilitate the quality assessment and refinement of uplift mappings.** The framework provides metrics which cover aspects relating to the quality of a mapping with respect to conformance to mapping specifications, data being mapped and reuse of ontologies, which are executed during the assessment process. As a note, the mapping specifications targeted were RML and R2RML, however, similar metrics can be defined for other mapping representations. Mapping quality improvement is facilitated by providing semi-automatic refinements for each metric, which are designed to guide users through the process of selecting the refinement and associated values most likely to resolve detected quality issues.
- **R2) The framework should facilitate the identification of changes in the source data of mappings.** The framework facilitates the identification of source data changes by executing comparison methods on the different versions of the data. The results of the comparison are uplifted and linked with respective mappings using SPARQL queries. Notification details are uplifted and queried periodically to compare thresholds with their respective count in order to identify when a notification should be sent. Contact details such as email addresses are stored in the notification policy in order to enable notifications to be sent to users when required.
- **R3) The framework should provide provenance on related quality of mapping process activities in an ontology-based format.** The detailed provenance related to activities in the framework is captured using MQIO and OSCD. MQIO is used by the mapping quality assessment and refinement component of the framework in order to represent captured mapping quality assessment and refinement information. OSCD is used by the source change detection component to capture information related to changes detected in the source data of LD mappings. The ontologies are implemented in OWL2 ontology language [92] which is an ontology-based format, therefore, enabling linking of relevant information using SPARQL queries.

The fulfillment of these three requirements and testing of the resulting design, resolved the limitations identified in **Section 2.1.7.3** and **2.2.6.3**.

Chapter 5: Evaluations

This chapter describes and presents the key findings of 5 experiments and 1 application study undertaken to evaluate the MQI framework. **Section 5.1** summarizes the 5 experiments. **Section 5.2** discusses the instruments used to measure the different aspects within each of the experiments. Thereafter, the **subsections** discuss each experiment and their results. **Supplementary information related to the evaluations described in this chapter are stored in a folder ("/Chapter-5") of the GitHub and each subdirectory relates a specific subsection.**

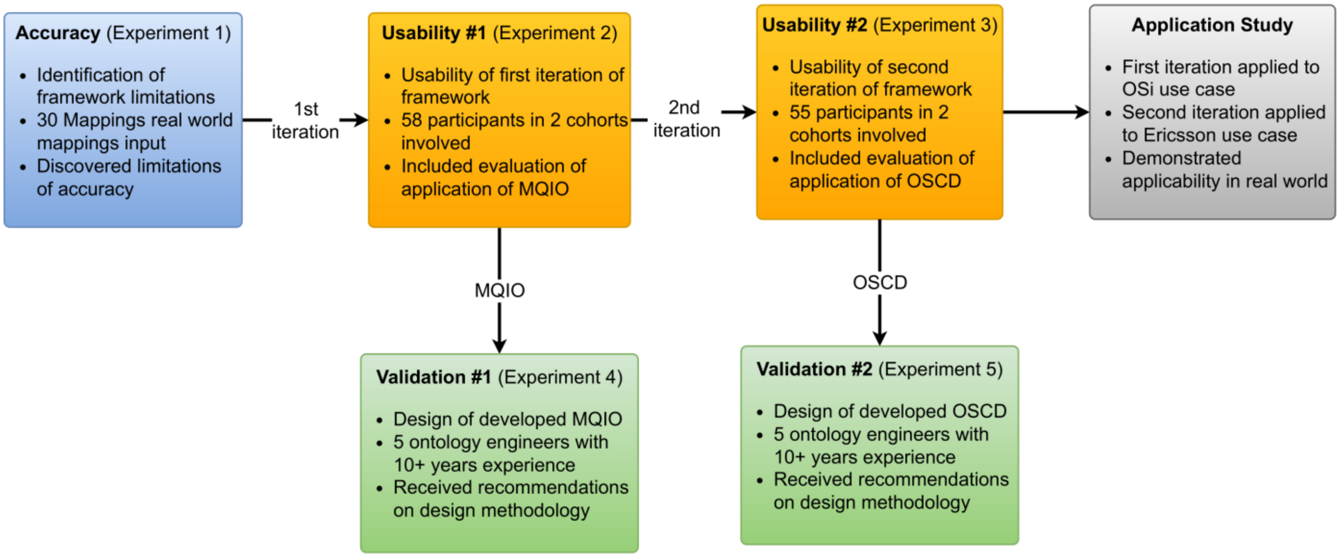
5.1 Experiment Summaries

The following definitions related to aspects of the experiments were retrieved from the ISO 9241-11:2018 standard (Usability: Definitions and concepts) [72], ISO/IEC 25000:2005 (Software product Quality Requirements and Evaluation) [50] and ISO 9001:2015 (Quality management systems) [20]. However, no ISO definitions were found for *"Understanding"* and *"Application"*, therefore, definitions were retrieved from the Oxford dictionary [41]. The *"Ontology Validation"* definition was derived from the ISO definitions as the standard refers to the validation of applications rather than specifically ontologies as stated.

- **Accuracy:** The closeness of the data values to a set of values defined in a domain considered semantically correct.
- **Usability**: Extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use
- **Satisfaction**: Extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet the user's needs and expectations.
- **Effectiveness:** Accuracy and completeness with which users achieve specified goals.
- **Understanding:** To have deep understanding is to be able to put the pieces together and to use such understanding to do things (e.g., solve problems, create new ideas etc).
- **Ontology Validation:** Confirmation, through the provision of objective evidence, that the requirements for an ontology have been fulfilled.
- **Application:** The action of putting something into operation.

The aforementioned aspects were chosen in order to evaluate the requirements (See Section 4.1) posed for the MQI framework. First, it was decided to test the accuracy of the framework in order to evaluate the framework in terms of its capability to undertake quality assessment of mappings. Accuracy allows for quality issues in mappings detected by the framework to be tested for semantic correctness. Usability is tested as the framework is designed to support users in facilitating them to achieve the creation and maintenance of high quality uplift mappings. Satisfaction is tested in order to identify user's responses and fulfillment of their expectations while interacting with the framework. Effectiveness is tested in order to identify limitations of the framework while supporting users in their quality refinement of mappings. Understanding is tested in order to validate that users were capable of using the data source change detection information displayed on the framework in their maintenance of mapping quality. The evaluation of MQIO and OSCD provides evidence that their respective ontology design is of appropriate quality. Finally, application of the framework provides an opportunity to evaluate the approach when put into operation in real world use cases.

The accuracy, usability and effectiveness of the mapping quality assessment and refinement component of the MQI framework was tested. Usability was measured using standardized satisfaction questionnaires and effectiveness metrics. In addition, qualitative data analysis was conducted to identify statements related to usability. Usability and understanding were tested for the Source Change Detection component of the framework with the usability measured in a similar manner to the experiment for the mapping quality assessment and refinement component of the framework. Understanding of the change detection information was measured using a questionnaire which required participants to use the information provided on the framework to identify links between detected changes and respective mappings. The validation of both ontologies involved assessing the fulfillment of requirements through software tools and feedback from ontology design experts, who reviewed the design methodologies followed by both ontologies. In addition, participants in the usability experiments were asked to provide feedback on the application of each ontology in graphs used during the experiment in order to identify issues when the ontologies are used. Finally, an application study of the framework was conducted involving real world use cases, which helped to consolidate the design. **Figure 36** summarizes the 5 experiments and 1 application study which were used to evaluate the proposed MQI framework.



Accuracy (Experiment 1)

- Identification of framework limitations
- 30 Mappings real world mappings input
- Discovered limitations of accuracy

Usability #1 (Experiment 2)

- Usability of first iteration of framework
- 58 participants in 2 cohorts involved
- Included evaluation of application of MQIO

Usability #2 (Experiment 3)

- Usability of second iteration of framework
- 55 participants in 2 cohorts involved
- Included evaluation of application of OSCD

Application Study

- First iteration applied to OSi use case
- Second iteration applied to Ericsson use case
- Demonstrated applicability in real world

Validation #1 (Experiment 4)

- Design of developed MQIO
- 5 ontology engineers with 10+ years experience
- Received recommendations on design methodology

Validation #2 (Experiment 5)

- Design of developed OSCD
- 5 ontology engineers with 10+ years experience
- Received recommendations on design methodology

Figure 36: Summary of Evaluations

The evaluations consisted of an accuracy experiment (Experiment 1), usability experiments (Experiment 2 and 3), ontology validation experiments (Experiment 4 and 5), which in total involved over 100 participants, consisting of knowledge engineering students, mapping specialists and ontology design specialists. In addition, an application study was conducted which demonstrated the applicability of the framework in mapping quality processes of two real world use cases. Accuracy was measured in experiment 1 for the initial protype of the mapping quality assessment and refinement component to test if the software code created was capable of accurately detecting quality issues in real world mappings. In addition, the testing was used to determine if the issues were correctly identified, which is essential for a high-quality assessment process. The results of the accuracy experiment provided evidence that the framework was capable of quality assessment of uplift mappings, which is one of the main functionalities of the design. Accuracy was not tested for the source change detection component as changes between file versions of the source data were initially detected using third party libraries. Effectiveness was measured in experiment 2 for the mapping quality assessment and refinement component to test the refinement capabilities of the framework, which is the other main functionality of the design. The effectiveness of the refinements provided evidence that the design is capable of removing quality issues from mappings, therefore, demonstrating the framework facilitates mapping quality improvement. In combination with the accuracy experiment, these experiments provided evidence that the framework is capable of accurately detecting quality issues and provides effective guidance on how to resolve them. Effectiveness was not tested for the source change detection component as the design was not intended to directly improve the alignment. Instead, the design is intended to provide information which is understandable by data maintainers in order to guide them in taking appropriate actions to preserve the level of alignment. The results of experiment 3 provided evidence that the change related information was understandable and could be utilized to preserve alignment, resulting in improved mapping and LD quality.

It was decided to validate the ontologies after the user experiments were completed as it allowed feedback related to the application to be collected from domain experts (mapping specialists) and addressed prior to validation. In addition, system testing enabled the identification of limitations when representing required quality information in diverse real world mappings. Both ontologies went through several iterations of refinement as a result of these experiments. The accuracy experiment (Experiment 1) helped to identify issues with the initial version of the MQIO when tasked with representing diverse mapping quality assessment and refinement information and resulted in the creation of version 1.1 of the MQIO. The first usability experiment (Experiment 2) involved graphs representing mapping quality assessment and refinement information expressed in the MQIO, which experts were asked to provide feedback on various aspects related to the application and resulted in version 1.2. The second usability experiment (Experiment 3) involved graphs expressed in the OSCD, which represented changes detected in a related source data in the OSCD, where experts were asked to provide feedback on aspects related to the application and resulted in version 1.1. The conformance of the MQIO (version 1.2) and OSCD (version 1.1) was measured in experiment 4 and 5, respectively, in order to validate that the developed ontologies were designed as recommended by ontology design best practices. Therefore, the final versions of the MQIO (version 1.3) and OSCD (version 1.2) addressed all of the recommendations received from mapping specialists and ontology design specialists as a result of the 5 experiments. Finally, the application study helped to identify issues when the framework and ontologies were applied to real world situations.

5.2 Experiment Instruments

Figure 37 presents each instrument used in the experiments.

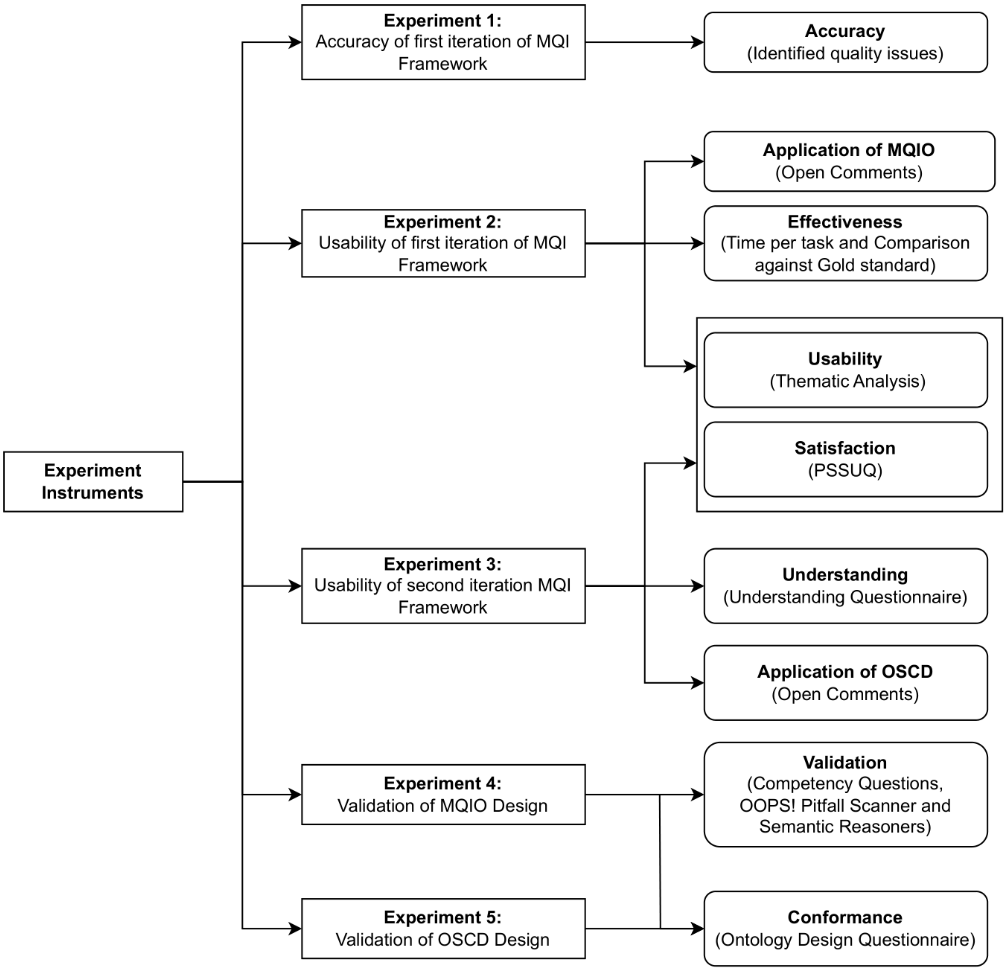

Experiment Instruments
Experiment 1: Accuracy of first iteration of MQI Framework
Experiment 2: Usability of first iteration of MQI Framework
Experiment 3: Usability of second iteration MQI Framework
Experiment 4: Validation of MQIO Design
Experiment 5: Validation of OSCD Design
Accuracy (Identified quality issues)
Application of MQIO (Open Comments)
Effectiveness (Time per task and Comparison against Gold standard)
Usability (Thematic Analysis)
Satisfaction (PSSUQ)
Understanding (Understanding Questionnaire)
Application of OSCD (Open Comments)
Validation (Competency Questions, OOPS! Pitfall Scanner and Semantic Reasoners)
Conformance (Ontology Design Questionnaire)

Figure 37: Instruments used in the experiments

An overview for each of the metrics in **Figure 37** is presented in the following subsections.

5.2.1 Accuracy Metric: Identified Quality Issues

Mapping quality issues which were identified in experiment 1 by assessing real-world mappings using the framework were manually examined by the author of this thesis to ensure that the issues were correctly identified by the framework. A checklist (see **Appendix A**) was created in order to complete the manual examination of the mapping quality issues. The checklist outlined methods used by approaches in the state of the art to identify a specific mapping quality issue. The rationale for doing the assessment was that if mapping quality issues which were incorrectly identified or not identified by the framework, it may indicate the framework does not accurately perform quality assessment of mappings. The results of the examination would therefore be used to improve the design of mapping quality metrics of the framework to ensure that the framework accurately facilitates the identification of quality issues in mappings. In addition, a checklist of mapping quality issues could be used to determine how common they are in real world mappings and may indicate a possible benefit of the approach to those mappings. Within this thesis the term **quality violation** relates to a quality issue within a mapping.

5.2.2 Effectiveness Metric: Comparison against Gold Standard

The gold standard mapping used in experiment 2 for comparison of participant results was derived from the examples shown in the PROV-O [84]. The mapping had to include a structure which would allow three mapping quality issues to be introduced. Introducing, three quality issues allowed various semi-automatic refinements available on the framework to be utilized by participants to improve the quality of the mapping. The mapping was designed to uplift provenance information and once created was input into the framework to ensure no quality issues were present which resulted in the gold standard mapping. Thereafter, the mapping was slightly altered to introduce quality issues. Each refined mapping generated from participants interaction with the framework could be assessed by comparing it to the gold standard. The framework assesses mappings using the metrics outlined in **Section 4.2.2.1**.

5.2.3 Effectiveness Metric: Time taken

The time taken for each task by participants in both cohorts of experiment 2 was recorded. The times for both cohorts were used to provide indications of how effective the framework was for completing a specific task. A longer task time would indicate that certain participants could less effectively use a certain area of the framework and indicate areas for improvements [125]. In addition, the task times could be correlated with other metrics to identify relationships impacting effectiveness.

5.2.4 Metrics measured by Questionnaires

Three post study questionnaires were used in the evaluations for different purposes in different experiments.

5.2.4.1 Satisfaction Metric: Post-Study System Usability Questionnaire (PSSUQ)

The Post-Study System Usability Questionnaire (PSSUQ) [86] is a standardized questionnaire used to test usability in the second and third experiment. The PSSUQ was designed by IBM to assess the overall satisfaction of the system usability. The PSSUQ consists of 19 positive statements (See **Appendix B**) which the user rates on a seven-point

Likert scale. The questionnaire was chosen over other usability questionnaires such as the System Usability Scale (SUS) as extensive psychometric evaluation has been completed on the PSSUQ. The questionnaire uses a Likert scale from 1 (Strongly Agree) to 7 (Strongly Disagree) and has a not applicable option (N/A). In addition, each question has an open comment area. The PSSUQ provides satisfaction scores on three sub scales and overall usability score:

- System usefulness (**SysUse**): Average the responses to questions 1 to 8
- Information quality (**InfoQual**): Average the responses to questions 9 to 15;
- Interface quality (**IntQual**): Average the responses to questions 16 to 18;
- **Overall**: Average the responses to questions 1 to 19. Question 19 asks participants about their overall impression of the system.

The sub scales are often referred to as the **metrics** of the questionnaire. A **lower score** on the Likert scale of the PSSUQ is considered better satisfaction.

5.2.4.2 Understanding Metric: Change Detection Understanding Questionnaire

A change detection understanding questionnaire (See **Appendix C**) was created by the author of this thesis, to test if users could interpret the data changes alignments with mappings which were detected by the framework. The understanding questionnaire contains two sections which include: Section 1 (S1) which asks questions about the change detection processes and Section 2 (S2) which asks questions about the changes which have been detected and their relationship with the respective mapping. The questions can be summarized as follows:

- Number of total changes which occurred in the source data (S1.Q1)
- Number of mappings impacted by the source data changes (S1.Q2)
- Notification thresholds definition (S1.Q3)
- Total thresholds defined (S1.Q4)
- Values assigned to notification thresholds (S1.Q5)
- Data references in the mapping (S1.Q6)
- Type of change (S2.Q1)
- Columns which have been inserted into the source data (S2.Q2)
- Values inserted into a specific column (S2.Q3)
- Columns which have been deleted from the source data (S2.Q4)
- Name of column deleted from the source data (S2.Q5)
- Total number of values inserted into a column (S2.Q6)

The questions were designed to test if the users could understand the information presented to them on the framework.

5.2.4.3 Conformance Metric: Ontology Design Questionnaire

No user evaluation questionnaire of ontologies was found within the ontology design methodologies followed during the development of MQIO and OSCD. Consequently, an Ontology Design questionnaire (See **Appendix D**) was created by the author of this thesis to allow feedback from ontology design experts to be gathered. The questionnaire was flexible and asked for open comment feedback on the following aspects of each ontology:

- The conformance of the design with ontology design best practices (Q1)
- Design methodology followed by the ontology (Q2)
- Concepts and relationships in the ontology (Q3)
- Documentation of the ontology (Q4)
- Other related feedback (Q5)

The questionnaire allowed the design of each ontology to be validated in line with best practices as recommended by specialists.

5.2.5 Metrics measured by Application

The application of both ontologies in a graph used in the usability experiments was assessed by receiving expert feedback.

5.2.5.1 Application Metric: Application of MQIO

The application of MQIO was measured through examination of a validation report ("evaluation_validation_report.ttl") which detailed three quality violations, respective quality metrics and associated refinements which resolved them. The mapping specialist who participated in experiment 2 which tested the mapping quality assessment and refinement component of the MQI framework were asked to examine the graph and provide feedback on the application of MQIO. The quality information was generated as a result of the quality assessment and refinement of the provided mapping. The feedback was provided through their statements and open comments in the PSSUQ.

5.2.5.2 Application Metric: Application of OSCD

The application of OSCD was measured through examination of a change log ("evaluation_change_log.ttl"). In addition, associated documentation listing the defined concepts/relationships defined in OSCD was reviewed. The graph detailed information related to 25 changes which were detected in the provided source data. In addition, the graph included information related to the agents involved in maintaining the data. The mapping specialist who participated in the experiment provided feedback on the application of OSCD in the graph through a number of questions rather than their statements as the evaluation was conducted asynchronously. The questions asked for feedback on the following aspects:

1. Concepts and relationships contained in OSCD.
2. Presentation of the graph.
3. Open comments about application.

In addition, participants could provide additional feedback through the open comments in the PSSUQ.

5.2.6 Usability Metric: Thematic Analysis

Thematic analysis was completed on qualitative data collected. It provides a method for identifying, analyzing and reporting patterns within data [102] and is often used to analysis user study data. Initially, the analysis involves generating codes and categories which represent the data. Thereafter, the categories are combined which results in the generation of themes. A *"bottom-up"* approach was adopted to identify patterns and themes as they emerged in the data. Unlike, a *"top-down"* approach where existing codes are applied to the data [19]. Key themes were derived from the data that encapsulated common patterns identified in the study. A six-step process was followed during the thematic analysis [102]:

1. **Familiarizing Yourself with the Data:** First, the data was transcribed. Thereafter, all qualitative data was examined extensively to become familiar.
2. **Generation of Initial Codes:** The examination of the data resulted in codes which described patterns discovered in the data.
3. **Searching for Themes:** Associated codes were grouped in order to identify themes.
4. **Reviewing Themes:** Each code was reviewed in order to ensure that it had been assigned to an associated theme.
5. **Defining and Naming Themes:** The theme names were defined in a manner which would describe assigned codes.
6. **Producing the Report:** The report containing the tagged data was generated using Taguette [114] which is an open-source tool designed for text tagging tool of qualitative data.

Thematic analysis was conducted on the qualitative data which was collected related to usability from experiments 2 and 3.

5.2.7 Ontology Validation

The following instruments were used to validate the design of MQIO and OSCD in experiments 4 and 5, respectively. First, the software tools used to validate the ontologies are described. Thereafter, the feedback received from experts is described.

5.2.7.1 Software Validation

The following methods were used to validate the design of each ontology during development.

- **Ontology Pitfall Scanner** [111]**:** OOPS! [18] is a web-based tool which is used to detect issues within the design of an ontology. The tool provides a method to validate and verify the design of an ontology. Furthermore, the tool is independent of any ontology development environment. Moreover, the tool provides recommendations for how issues can be repaired. The issues are represented in different severity which include minor, important, and critical. Critical issues must be repaired to ensure an adequate quality level. The tool has been applied to both ontologies.
- **Ontology Competency Questions** [10]**:** Competency Questions are used to represent the requirements of an ontology. These questions state information which the ontology should contain and are often defined in natural language. These questions ensure that design requirements have been satisfied. Furthermore, these questions can be answered by using SPARQL queries to query the ontology or instances. Competency questions have been designed for the ontologies and have been answered using SPARQL queries which query sample graphs. The fulfillment of the questions is used to validate the design requirement of the ontologies.
- **Semantic Reasoners** [154]**:** Protégé is an open source tool which provides the capability to create and edit ontologies using an intuitive GUI. Plugins are also available which allow extra functionality such as alternative visualization. Furthermore, reasoners are available which can be used to detect inconsistencies within the ontology design. Both ontologies were constructed and reasoned overusing Protégé.
- **RDFUnit** [81]: RDFUnit is a test-driven framework designed to assess the quality of RDF datasets which can automatically generate schema axioms. The results are represented in the Test-Driven Data Validation Ontology[45]. Graphs during the development process of both ontologies were assessed using the test-cases in RDFUnit.

5.2.7.2 Expert Validation

Feedback has been received from semantic web experts during the lifecycle of both ontologies.

- **Publication:** Publishing the ontologies in peer reviewed publications provided feedback from experts in the semantic web domain. In addition, publishing documentation allowed feedback through dissemination.

[45] http://rdfunit.aksw.org/ns/core

- **Application of Ontology:** Participants in the expert cohort of usability experiments 2 and 3 were asked to provide feedback on the application of each ontology in the graphs used during the experiments. The student cohort was not asked for feedback as their knowledge of ontology design is limited. The experts in experiment 2 provide feedback through think-aloud statements and open comments in the PSSUQ. The experts in experiment 3 provided feedback through additional questions in the questionnaire and open comments.
- **Presentation:** The ontologies were presented to a panel of semantic web experts at the semantic interoperability conference (SEMIC2022)[46] organized by the European commission.

5.2.8 Statistical Tests

The following statistical methods were used to analyze the results collected from the respective instruments in each experiment.

- **Spearman Rank-Order Correlation Coefficient:** Spearman's correlation test [132] is a nonparametric test used to measure correlation between variables. The null hypothesis is that there is no correlation between the variables. The alternative hypothesis is that there is a correlation between the variables. Spearman's test is less sensitive to outliers when compared to the Pearson test.
- **Pearson Correlation Coefficient:** Pearson's correlation test is parametric [106] and used to determine the strength of a relationship between linear variables. The null hypothesis of the test is that there exist no correlations between the variables. The alternative hypothesis is that a correlation exists. The test assumes the normality of the data.

Spearman's and Pearson's test have been applied to the data collected in experiments 2 and 3 to identify if correlations exist between effectiveness, understanding and usability. The coefficient of test measures the strength and direction of association between two ranked variables. The p-value of the test is a measure of how likely or probable it is that any observed correlation is due to chance. A confidence level of 0.05 was applied to the p-value of both tests to indicate a statically signification correlation. In addition, the standard deviation has been applied to data to determine how closely grouped values are around the mean. A standard deviation of less than 1 is considered to be low for the purpose of the study.

[46] https://joinup.ec.europa.eu/collection/semic-support-centre/semic-conference

5.3 Experiment 1: Accuracy of MQI Quality Assessment and Refinement Component

This section presents a system evaluation of the **Mapping Quality Assessment and Refinement Component** which evaluated the perceived accuracy when applied to real world mappings. **Supplementary information related this experiment is stored in a folder ("/Experiment-1") of the GitHub.** The version of the MQIO used during this experiment was **version 1.0**.

5.3.1 Hypothesis

The hypothesis related to this study was:

- **Hypothesis H1:** The framework facilitates the identification of quality issues in real world mappings.

5.3.2 Experimental Setup

5.3.2.1 Methodology

A system evaluation was conducted to test the hypothesis related to the study. The evaluation involved inputting real world mappings into the framework. First, each mapping was manually assessed by following a checklist (**Appendix A**). Each violation detected by the framework was compared to the manual examination results in order to identify incorrectly detected violations or correct violations which were not identified. Furthermore, if a violation was incorrectly identified, the problem was resolved, and the mapping reassessed. The mappings were stored in accordance with the data protection policy at Trinity College Dublin[47]. **Figure 38** presents a flow chart of the steps applied to each mapping in the experiment.

[47] https://www.tcd.ie/dataprotection/assets/docs/dataprotectionpolicy/Trinity_College_Dublin_Data_Protection_Policy_16122020.pdf

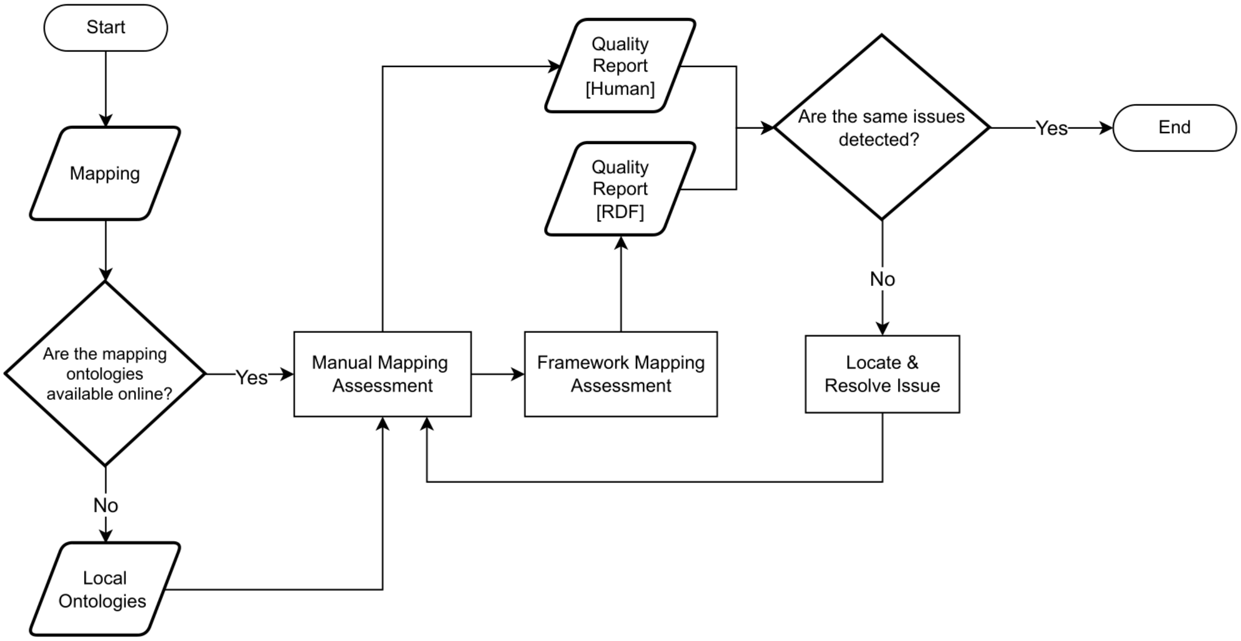

Start
Mapping
Are the mapping ontologies available online?
Yes
No
Local Ontologies
Manual Mapping Assessment
Framework Mapping Assessment
Quality Report [Human]
Quality Report [RDF]
Are the same issues detected?
Yes
No
End
Locate & Resolve Issue

Figure 38: Flow chart of steps in Experiment 1

Table 30 presents the steps which were applied to each mapping in the experiment.

Table 30: Experiment 1 Steps

#	Step description
1	Upload the mapping
2	Upload the local ontologies (if applicable)
3	Add additional metadata relating to creators
4	Manual assessment of mapping
5	Framework assessment of mapping
6	Compare issues in quality reports
7	Repeat of manual examination of issues detected
8	Export the quality report

Steps 1-2 involved inputting a mapping and respective ontologies. Step 3 involved the input of metadata related to the mapping. Steps 4-7 involved the assessment of the mapping and examination of quality issues detected. Step 8 involved the exporting of the quality report expressed in MQIO, which was generated as a result of the assessment process. The report details important metadata and provenance relating to the creation, quality assessment and refinement of a mapping.

5.3.2.2 Mappings

30 R2RML-based mappings were collected from various research projects and semantic web students who designed uplift mappings as part of their group project in the Knowledge and Data Engineering (CS7IS1) MSc module at Trinity College Dublin[48]. It was hoped that by selecting mappings from different sources, the framework would demonstrate it is capable of assessing mapping quality from a variety of author's and project contexts. The projects included Beyond2022 project, FAIRVASC project, OSi-ADAPT project and data.geohive.ie project (**Table 31**). Mappings in those projects were created by post-doctoral researchers. The mappings are stored in a folder ("/Mappings") of the GitHub.

Table 31: Research projects which mappings were retrieved from for the experiment

Name	URL	Description	Data used
Beyond2022	https://beyond2022.ie/	Designing a virtual reconstruction of the Record Treasury destroyed in 1922.	Historical data
data.geohive.ie	http://data.geohive.ie/	Publishing Irish geospatial data in linked data format.	Geospatial data
OSI-ADAPT	https://www.osi.ie/	Exposing OSi's new Prime2 spatial object model as linked data.	Geospatial data
FAIRVASC	https://fairvasc.eu/	Linking disease data using semantic web technologies.	Medical data

The students created ontologies designed to represent information related to fitness and cooking. The fitness related information included exercises, goals, equipment and workout sessions. The cooking related information included recipes, nutrients and origin of foods. Thereafter, mappings were created which uplifted sample data expressed in these ontologies. The number of mappings collected from each source is presented in **Table 32**. In addition, the total number of lines, triple maps, distinct data sources and ontologies present in the mappings are shown.

Table 32: Number of mappings from each source (30 total)

Mapping source	Mappings	Lines	Triple Maps	Sources	Distinct Ontologies
Semantic web MSc students	10	5133	135	13	25
data.geohive.ie	10	1182	43	24	9
OSi-ADAPT	4	523	21	9	7
FAIRVASC	3	997	15	6	2
Beyond2022	3	447	12	13	3

[48] https://tcdlibrary.rl.talis.com/modules/cs7is1.html

5.3.3 Results

Since the mapping quality validation reports generated by the framework are in RDF format, a decision was made to convert the reports from the experiment into a collection of named graphs ("experiment_1_reports.trig"). The resulting RDF graphs enabled the experiment results to be processed and automatically analyzed using a SPARQL [61] query ("analyze_quality_reports.rq"). **Table 33** presents the number of issues detected by each quality metric when the 30 R2RML mappings were input.

Table 33: Violation counts associated with each metric (228 total)

Quality Aspect	Metric ID	Violation count	Metric	Metric Description
Mapping Quality Aspect	***MP11***	5	Language tag not defined in RFC 5646.	Valid language tags are defined as per RFC 5646(BCP 47)[49].
Data Quality Aspect	***D1***	14	Usage of undefined class	Classes are considered undefined when it is not possible to dereference them against their namespace
	D2	57	Usage of undefined property	Properties are considered undefined when it is not possible to dereference them against their namespace.
	D3	36	Usage of incorrect domain	The class type defined in the subject map does not include the domain for this property.
	D6	1	Usage of range	This will validate that the correct domain and range are being used in the mapping definition.
	D7	3	Usage of incorrect datatype	The datatype assigned to the object does not match the datatype range for this property.
Vocabulary Quality Aspect	***VOC1***	1	Human readable labels/comments	Humans consuming the information should be able to understand the linked data resource.
	VOC2	14	Domain and range definitions	A property should have a range and/or domain definition.
	VOC3	4	Basic Provenance Information	Consumers need to understand where the data has originated.
	VOC4	48	Machine readable licensing	Allows the licensing information to be queried by machines.
	VOC5	45	Human readable licensing	Allows humans to read and understand license in a textual format.

The metrics used by the framework for quality assessment are further described in Section 4.2.2.1. In total 228 mapping quality violations were detected in the experiment. **Table 34** presents the quality issues which were

[49] https://www.rfc-editor.org/rfc/rfc5646.html

detected in each mapping. Each issue in the "**Description of Issue**" column has in brackets the metric which detected the issue.

Table 34: Quality issues detected in Experiment 2

#	Source	Total Issues	Description of Issue
1	Geohive-Mapping-1	6	•ont:logainmId (D2) • rdfs: (VOC5) • rdfs: (VOC4) • geo: (VOC5) • geo: (VOC4) • ga (MP11)
2	Geohive-Mapping-2	4	•rdfs: (VOC5) • geo: (VOC4) • geo: (VOC5) • rdfs: (VOC4)
3	Geohive-Mapping-3	5	•ga (MP11) • geo: (VOC4) • rdfs: (VOC5) • rdfs: (VOC4) • geo: (VOC5)
4	Geohive-Mapping-4	4	•rdfs: (VOC4) • geo: (VOC5) • geo: (VOC4) • rdfs: (VOC5)
5	Geohive-Mapping-5	6	•geo: (VOC4) • ont:logainmId (D2) • ga (MP11) • geo: (VOC5) • rdfs: (VOC5) • rdfs: (VOC4)
6	Geohive-Mapping-6	4	•rdfs: (VOC5) • geo: (VOC5) • rdfs: (VOC4) • geo: (VOC4)
7	Geohive-Mapping-7	4	•rdfs: (VOC4) • rdfs: (VOC5) • geo: (VOC4) • geo: (VOC5)
8	Geohive-Mapping-8	5	•ga (MP11) • geo: (VOC5) • rdfs: (VOC5) • geo: (VOC4) • rdfs: (VOC4)
9	Geohive-Mapping-9	4	•rdfs: (VOC5) • rdfs: (VOC4) • geo: (VOC5) • geo: (VOC4)
10	Geohive-Mapping-10	6	•geo: (VOC4) • geo: (VOC5) • rdfs: (VOC5) • ont:logainmId (D2) • ga (MP11)

			• rdfs: (VOC4)
11	Student-Mapping-1	3	•recipes:serving (D7) • recipes:stepNumber (D7) • recipes:stepNumber (D7)
12	Student-Mapping-2	4	• www:Step (D1) • www:isProvidedBy (D3) • www:Nutrients (D1) • www:isContainedBy (D3)
13	Student-Mapping-3	22	•recipes:isServedIn (D3) • recipes:inSeason (D3) • recipes:costs (D3) • recipes:energy (D3) • recipes:preparationTime (D3) • recipes:containsIngredient (D3) • recipes:originCountry (D3) • recipes:dietryFiber (D3) • rdfs: (VOC4) • rdfs: (VOC5) • recipes:protein (D3) • recipes:has (D3) • recipes:quantityPerson (D3) • recipes:servedAs (D3) • foaf: (VOC5) • recipes:uses (D3) • recipes:fat (D3) • recipes:glucide (D3) • recipes:sal (D3) • recipes:contains (D3) • foaf: (VOC4) • recipes:tastes (D3)
14	Student-Mapping-4	4	• owl: (VOC4) • rdfs: (VOC5) • owl: (VOC5) • rdfs: (VOC4)
15	Student-Mapping-5	3	•foodreport:appliesToIngredient (D2) • foodreport:manufacturedFrom (D2) • foodreport:lawText (D2)
16	Student-Mapping-6	20	•gym:name (D2) • gym:helpstobuild (D3) • gym:type (D2) • gym:achieve (D3) • gym:calories (D2) • gym:workouttypes (D1) • gym:exercisetypes (D1) • gym:hastypes (D2) • gym:tool (D3) • gym:name (D2) • gym:type (D2) • gym:achieve (D3) • gym:name (D2) • gym:workout (D1) • gym:hastypes (D2) • gym:calories (D2)

			• gym:exercise (D1) • gym:goal (D1) • gym:name (D2) • gym:bodyparts (D1)
17	Student-Mapping-7	17	•flavourtown:tastesLike (D2) • flavourtown:fat (D3) • flavourtown:recipeServes (D2) • flavourtown:preparationStep (D1) • flavourtown:provideIngredient (D2) • flavourtown:belongsTo (D2) • flavourtown:ingredientName (D2) • flavourtown:protein (D3) • flavourtown:flavourName (D2) • flavourtown:consistsOf (D2) • flavourtown:instructionSet (D2) • flavourtown:stepNumber (D2) • flavourtown:recipeName (D2) • flavourtown:provideFlavour (D2) • flavourtown:calories (D3) • flavourtown:cuisineName (D2) • flavourtown:hasIngredient (D2)
18	Student-Mapping-8	1	•wclass:cabrohydrates (D2)
19	Student-Mapping-9	4	• foaf: (VOC5) • rdfs: (VOC4) • rdfs: (VOC5) • foaf: (VOC4)
20	Student-Mapping-10	15	•openfit:goal_id (D2) • openfit:exercise_id (D2) • openfit:has_target_area (D3) • openfit:MuscleGain (D1) • openfit:strength_id (D2) • openfit:CardioExercise (D1) • openfit:weigth (D2) • openfit:sets (D3) • openfit:repetitions (D3) • openfit:avg_cals (D2) • openfit:duration (D3) • openfit:name (D3) • openfit:has_target_muscle (D3) • openfit:description (D3) • openfit:has_goal (D3)
21	Beyond2022-Mapping-1	19	•cidoc:P81a_end_of_the_begin (D2) • cidoc:P82b_end_of_the_end (D2) • b2022:reports_to (VOC2) • cidoc:P81b_begin_of_the_end (D2) • cidoc:E81_Actor_Appellation (D1) • b2022: (VOC4) • rdfs: (VOC5) • b2022:subjects (VOC2) • b2022:part_of (VOC2) • cidoc:P81a_end_of_the_begin (D2) • cidoc:P82a_begin_of_the_begin (D2) • cidoc:P82a_begin_of_the_begin (D2)

			• rdfs: (VOC4) • cidoc:P81b_begin_of_the_end (D2) • cidoc:P82b_end_of_the_end (D2) • b2022: (VOC3) • b2022:composed_of (VOC2) • b2022: (VOC5) • cidoc: (VOC4)
22	Beyond2022-Mapping-2	10	•cidoc:P82a_begin_of_the_begin (D2) • cidoc:P81a_end_of_the_begin (D2) • b2022: (VOC3) • rdfs: (VOC4) • b2022: (VOC5) • cidoc: (VOC4) • cidoc:P81b_begin_of_the_end (D2) • rdfs: (VOC5) • cidoc:P82b_end_of_the_end (D2) • b2022: (VOC4)
23	Beyond2022-Mapping-3	18	•b2022:object (VOC2) • cidoc: (VOC4) • b2022:subjects (VOC2) • b2022: (VOC3) • b2022:subject (VOC2) • b2022:object (VOC2) • b2022:counselor (VOC2) • rdfs: (VOC4) • b2022: (VOC4) • b2022: (VOC5) • rdfs: (VOC5) • b2022:counsellee (VOC2) • cidoc:E81_Actor_Appellation (D1) • b2022:represented_by (VOC1) • b2022:object (VOC2) • b2022:subject (VOC2) • b2022:subject (VOC2) • b2022:represented_by (VOC2)
24	FAIRVASC-Mapping1	5	•fvc:creatinineRelationToLab (D2) • fvc:hasANCA (D2) • fvc:lastVisit (D2) • fvc:dateOfEncounter (D2) • fvc:hasOutcomes (D2)
25	FAIRVASC-Mapping-2	0	N/a
26	FAIRVASC-Mapping-3	4	•fvc:hasANCA (D2) • fvc:hasOutcomes (D2) • fvc:lastVisit (D2) • fvc:dateOfEncounter (D2)
27	OSi-Mapping-1	13	•sdmx-dimension: (VOC3) • prov:generated (D3) • sdmx-dimension:timePeriod (D6) • daq:MetricProfile (D1) • rdf: (VOC4) • prov: (VOC5) • daq:totalDatasetTriplesAssessed (D2) • prov: (VOC4)

			• daq: (VOC4) • sdmx-dimension: (VOC4) • rdf: (VOC5) • sdmx-dimension: (VOC5) • daq: (VOC5)
28	OSi-Mapping-2	6	•rdf: (VOC5) • daq: (VOC5) • rdf: (VOC4) • daq: (VOC4) • rdfs: (VOC4) • rdfs: (VOC5)
29	OSi-Mapping-3	6	•rdf: (VOC4) • rdf: (VOC5) • daq: (VOC5) • rdfs: (VOC5) • rdfs: (VOC4) • daq: (VOC4)
30	OSi-Mapping-4	6	•rdf: (VOC4) • rdfs: (VOC5) • rdfs: (VOC4) • rdf: (VOC5) • daq: (VOC5) • daq: (VOC4)

The mean number of violations per mapping was 8. The violations were grouped by mapping source (**Table 35**). The grouping shows the relationship between the number of violations detected in each source.

Table 35: Violations per mapping source

Mapping source	Violation count
Semantic web students	**93** (41%)
FAIRVASC	**9** (4%)
data.geohive.ie	**48** (21%)
OSi-ADAPT	**31** (14%)
Beyond2022	**47** (20%)

None of the mapping sources were free of violations, indicating quality issues are common. Metrics related to the vocabulary aspect (See Section 4.2.2.1.4) of quality were not considered for the vocabularies designed by the students, as they were not asked to include information, such as licensing and provenance. Nonetheless, mappings created by the semantic web students accounted for a similar amount of detected violations, while accounting for far less of mappings available (**Table 36**). The high number of violations within these mappings is likely caused by their inexperience of creating and using mappings outside the context of a postgraduate module project. However, it is clear from the results from the research projects, that even experienced postdoctoral researchers had difficulty creating error-free mappings for their research projects. The results reinforced the view of the author of this thesis that mapping violations are common and a tool which can detect and refine these violations is likely to improve mapping and dataset quality.

Table 36: Comparison of mapping sources

Mapping Source	Total mappings	Total violations
Semantic web students	10	93
Research projects	20	135

Interestingly, a large number of violations (112) related to the quality of ontologies used in the mappings. These violations mostly (83%) related to lack of licensing information, which included:

- **RDF Schema (RDFS)**[50]**:** This ontology is commonly used by the semantic web community [13], however, it does not contain human-readable and machine-readable licenses. The ontology was used in OSI and date.geohive.ie mappings to define metadata related to the uplifted data, such as human-readable labels (`rdfs:label`) and comments (`rdfs:comment`). These mappings accounted for a large amount of mappings available (33.3%), therefore, resulting in many violations (38).
- **GeoSPARQL Ontology**[51]: 20 violations were detected as a result of using the GeoSPARQL ontology, as it does not contain human-readable and machine-readable licenses. The ontology was used by the data.geohive.ie mappings, which accounted for 33% of mappings available, therefore, resulting in a large number of quality issues.
- **Beyond 2022 Ontology**[52]: 24 violations were detected as a result of mappings using the Beyond 2022 Ontology, as it does not contain human-readable and machine-readable licenses. In addition, it does not contain domain and range definitions for any of the properties in the ontology. However, the Beyond 2022 mappings only accounted for 10% of mappings available, therefore, resulting in less violations.
- **Dataset Quality Vocabulary (daQ)** [37]: 8 violations were detected as a result of the OSi mappings using daQ, which does not contain human-readable and machine-readable license.
- **The RDF Concepts Vocabulary (RDF)**[53]**:** Interestingly, despite the common reuse of this ontology by the semantic web community [13], the RDF concepts vocabulary does not contain human-readable and machine-readable licenses. The ontology was used in the OSi mappings to define properties (`rdf:Property`) for representing measurements related to geospatial data.

These results indicated that there is a low conformance to licensing in the ontologies used in the mappings. In addition, a number of violations (4) were detected by the VOC3 metric, which measured the presence of basic

[50] https://www.w3.org/2000/01/rdf-schema.ttl
[51] http://www.opengis.net/ont/geosparql
[52] http://ont.virtualtreasury.ie/ontology
[53] https://www.w3.org/1999/02/22-rdf-syntax-ns

provenance information in ontologies, such as creators, publishers and descriptions. These violations all related to usage of the Beyond 2022 Ontology as it does not contain provenance such as creators, licenses and descriptions. In addition, all (14) of the violations detected by the metric which measures missing domain and range of properties (VOC2) related to the Beyond 2022 Ontology. The only information included for each property in the ontology was a human readable description (`rdfs:comment`). These results could indicate that the ontology is still under development as the manual examination confirmed it was lacking detail. Only one concept and property tested during the experiment was missing human readable labels/comments (VOC1), which indicated high conformance to this metric. The `b2022:represented_by` property triggered the violation as it has no respective descriptive properties (See **Appendix A**). The property definition in the ontology only contains the type of the property (`owl:ObjectProperty`) and no other information.

The D2 metric in the data quality aspect detected the most violations (25%), which measured usage of undefined properties in mapping definitions. A violation is generated by the metric when a property is not defined in the respective namespace ontology. Examples of undefined properties/classes which were detected by the framework during the experiment are presented in **Table 37**.

Table 37: Incorrect classes/properties detected in the experiment

Incorrect Term	**Correct Term**
`gym:goals`	`gym:goal`
`gym:exercise`	`gym:Exercise`
`foodreport:manufactoredFrom`	`foodreport:manufacturedFrom`
`foodreport:lawText`	`foodreport:lawName`
`openfit:avg_cals`	`openfit:avg_calories_burnt`

Beyond 2022 mappings accounted for a large number (50%) of the undefined terms detected in the research project mappings available. The undefined terms used the CIDOC Conceptual Reference Model namespace, however, are undefined in the CIDOC ontology[54]. Examples of undefined properties included `cidoc:P81b_begin_of_the_end`, `cidoc:P82b_end_of_the_end` and `P81b_begin_of_the_end`. In addition, the `cidoc:E81_Actor_Appellation` class is undefined. These undefined terms could have been used during the development phase of the mappings in order to capture relevant information until defined terms were identified. Violations which related to the usage of undefined classes did not occur as often, which could be a result of mappings containing less class definitions compared to properties.

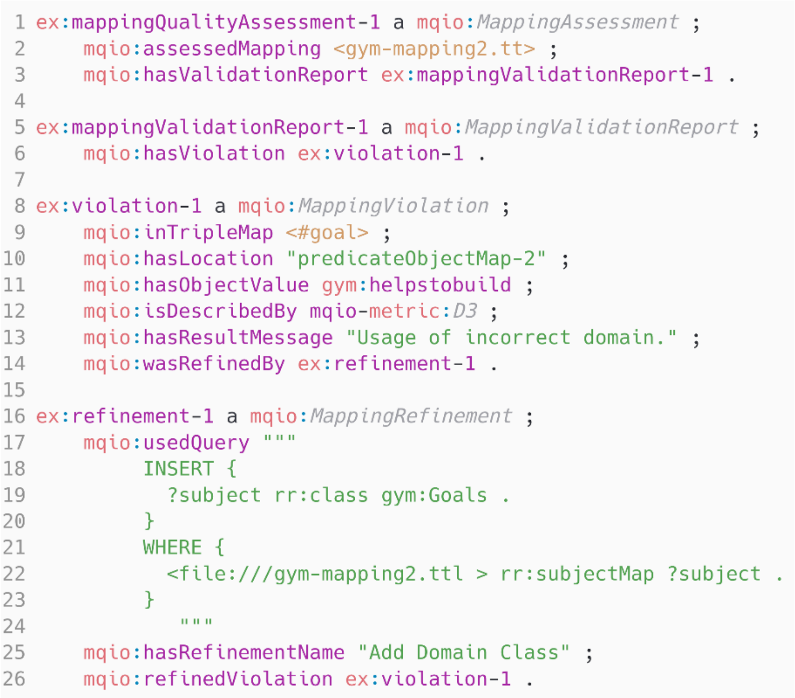
[54] http://erlangen-crm.org/current/

The second most common violation detected in the data quality aspect related to the D3 metric, which accounted for 16% of violations and relates to the usage of incorrect domain[55]. Most of these mappings were missing the domain class and others included incorrect class definitions. **Table 38** presents examples of properties/classes in the mappings which triggered the metric and the respective correct domain class.

Table 38: Incorrect domain violations detected in the experiment

Property	**Domain Defined in Mapping**	**Domain of Property**
prov:generated	**prov:Entity**	**prov:Activity**
www:isContainedBy	**www:Nutients**	**Ingredient:Nutrients**
www:isProvidedBy	**www:Step**	**Recipe:Step**
gym:achieve	**gym:exercise**	**gym:Exercises**
gym:helpstobuild	**gym:goal**	**gym:Goals**

Listing 13 presents an extract of a validation report generated ("sample_validation_report.ttl") by mapping #16 during the experiment.



```
ex:mappingQualityAssessment-1 a mqio:MappingAssessment ;
    mqio:assessedMapping <gym-mapping2.tt> ;
    mqio:hasValidationReport ex:mappingValidationReport-1 .

ex:mappingValidationReport-1 a mqio:MappingValidationReport ;
    mqio:hasViolation ex:violation-1 .

ex:violation-1 a mqio:MappingViolation ;
    mqio:inTripleMap <#goal> ;
    mqio:hasLocation "predicateObjectMap-2" ;
    mqio:hasObjectValue gym:helpstobuild ;
    mqio:isDescribedBy mqio-metric:D3 ;
    mqio:hasResultMessage "Usage of incorrect domain." ;
    mqio:wasRefinedBy ex:refinement-1 .

ex:refinement-1 a mqio:MappingRefinement ;
    mqio:usedQuery """
         INSERT {
           ?subject rr:class gym:Goals .
         }
         WHERE {
           <file:///gym-mapping2.ttl > rr:subjectMap ?subject .
         }
            """
    mqio:hasRefinementName "Add Domain Class" ;
    mqio:refinedViolation ex:violation-1 .
```

Listing 13: Extract of the validation report generated during experiment for Mapping #16

[55] https://www.w3.org/TR/owl-ref/#domain-def

The violation (`ex:violation-1`) presented in **Listing 13** was detected by the D3 metric (`mqio-metric:D3`) which identified that domain of the property (`gym:helpstobuild`) was not defined in the mapping. The violation was refined (`ex:refinement-1`) using a semi-automatic refinement (`"Add Domain Class")` outlined below.

- The framework queries ("find_domain.rq") the namespace of the property to find classes defined in the domain (`rdfs:domain`) of the property.
- The result of the query could be one or more classes (`owl:unionOf`).
- The retrieved classes are displayed in a drop down menu. The result for the violation shown is the `gym:Goals` class (**Appendix G**).
- The users select the most appropriate class for their use case.
- The selected class will be inserted into the respective location in the mapping using a SPARQL update query ("insert_domain.rq"), therefore, resolving the issue. The refinement query (`mqio:usedQuery`) executed by the framework is represented in the validation report.

Violations related to ontologies used in the mappings cannot be directly updated by the framework as updates must be completed by maintainers, who guided by the information, can resolve the issue by using semi-automatic refinements suggested by the framework in order to update the ontology accordingly. The refinements are further described in **Section 4.2.2.2**.

5.3.4 Discussion

The results indicated that spelling mistakes of ontology class and property names are common within mappings. This is an important finding, since R2RML engines only validate the syntax of the mappings, the users most likely would not detect spelling violations in their mappings before a dataset is generated. Therefore, poor quality LD could potentially be continuously generated, published, and consumed. As mentioned before, semantic concepts can often be overlooked, however, the RDF data will still be generated as the syntax of the mapping is still correct. In addition, the results indicated that the mapping creators had a good understanding of R2RML as the framework did not detect violations relating to incorrect R2RML syntax. These results would be expected as ontologies used within mappings contain many properties, classes and restrictions relating to the usage of them.

However, the ontologies used in the mappings showed varying levels of quality, as metrics in the vocabulary quality aspect detected 49% of violations. Most (83%) of these violations related to the lack of licensing information, which is important as they provide permission for consumers to reuse the models. These results were expected as previous research which assessed the quality of LD datasets on the LOD cloud [38], identified that only 8% of them contained a human-readable license. Licensing information could be overlooked as the ontology still provides a sufficient model for representing relevant knowledge, however, it does not provide explicit permission for reuse in a mapping. Conformance to human-readable descriptions or domain/range definitions for ontology terms was better, with 14% of violations related to these metrics, which all originated from the use of the Beyond 2022 ontology. The manual examination of the ontology observed overall lacking detail, which could be as a result of the ontology being still under development. No other ontology contained terms missing descriptions or domain/range definitions, which indicated high overall conformance for these metrics.

5.3.5 Conclusion

Hypothesis H1 (The framework facilitates the identification of quality issues in real world mappings) has been shown to be supported. The framework accurately identified quality issues in the mappings with 228 total violations identified. Each of the 30 mappings were manually assessed prior to input into the framework following the checklist outlined in **Appendix A** and results were compared. Initially, 1 incorrect violation was identified. The violation related to the metric which tested for usage of incorrect domain (D3). The metric did not take into account when the class definition in the mapping is a sub class of the domain (`rdfs:domain`) rather than the class defined in the ontology. For instance, mapping #27 contained a triple map with the class defined as `daq:QualityGraph` [37] and a predicate defined as `qb:structure` [33] . However, the domain of the property in the ontology is `qb:DataSet` which triggered a violation as the class did not match the definition (`daq:QualityGraph)` in the mapping. The issue was resolved by introducing functionality to fetch the subclass (`rdfs:subClassOf`) of class definitions in a mapping. Thereafter, retrieved classes are added to the query used to calculate the metric. No other quality issues were incorrectly identified by the metrics, therefore, indicating the framework is capable of accurately identifying quality issues in real world mappings. The semantic quality reports expressed in MQIO, which were generated for each mapping could be linked in order to improve maintenance and reuse [42,65,75].

5.4 Experiment 2: Effectiveness of MQI Framework

This section presents a user evaluation of the **mapping quality assessment and refinement component** of the MQI framework which evaluated the perceived usability and effectiveness with respective end users. **Supplementary information related this experiment is stored in a folder ("/Experiment-2") of the GitHub.** The version of the MQIO used during this experiment was **version 1.1**.

5.4.1 Hypotheses

The hypotheses related to this study were:

- **Hypothesis H2:** The MQI framework facilitates the assessment and refinement of uplift mappings.
- **Hypothesis H3:** The participants background knowledge influences the successful completion of the mapping tasks with the MQI framework.

5.4.2 **Experimental Setup**

5.4.2.1 Methodology

A user evaluation was conducted to test the hypotheses related to the study. The experiment involved participants using the framework to assess and refine the quality of a provided uplift mapping expressed in R2RML. Rather than asking participants to create an uplift mapping which may result in inconsistent results, an identical mapping (created by the author of this thesis) was provided to all participants. The mapping needed to contain detectable quality issues in order to exercise the refinement functionality of the framework. Therefore, the uplift mapping that was provided to participants contained a number of quality issues which were inspired from the results of experiment 1. The participants were grouped into a student and expert cohort to discover if the framework was usable by both experienced and inexperienced mapping engineers. The student cohort contained participants in a third level class, while the expert cohort contained researchers who had experience creating and operating LD mappings. Participants in the student cohort completed the experiment asynchronously by accessing the framework using provided login details. As participants in the expert cohort used the think aloud test a conference call was required. Therefore, a conference call was organized for each of the participants. Hypothesis **H2** was tested by collecting and examining the refined mappings generated by both cohorts as a result of the experiment. The examination allowed the number of quality violations that were detected by a participant to be calculated and to validate that the provided mapping had been refined appropriately. In addition, the participants' time per task was recorded, for the expert cohort a think aloud test was also recorded, and all participants answered the PUSSQ. Hypothesis **H3** was tested by comparing the results collected from participants in each cohort. Ethical approval has been received from the Research Ethics Committee at Trinity College Dublin.

5.4.2.2 Think-Aloud Test

Think-Aloud tests [49] are commonly used to evaluate the usability of software. The participants are asked to verbalize their thoughts and actions while completing scenario-based tasks. The objective is to identify difficulties which are encountered in the usability of the software by examining the resulting narrative. The scenario (**Appendix F**) of the test for experiment 2 involved a mapping engineer who wants to assess and refine the quality of an uplift mapping. The transcript from the conference call for each expert participant was stored in order to complete an analysis of their statements. As recommended by the think-aloud test, assistance was not provided to participants unless they were unable to proceed with the tasks. Thematic analysis (See **Section 5.2.6**) was conducted on the transcripts collected as a result of the think aloud test which allowed patterns in the data to be discovered.

5.4.2.3 Experiment Layout

The structure of the experiment is outlined below.

- **Inclusion/Exclusion Criteria:** Participants in the expert cohort satisfied each of the following criteria: 1) Semantic web researchers 2) Knowledgeable in RDF and R2RML 3) Previous experience creating R2RML mappings and 4) Previous experience executing R2RML mappings. Participants in the student cohort satisfied each of the following criteria: 1) Third level student 2) Attempted Experiment and 3) Provided answers to the PSSUQ.
- **Recruitment:** The participants in the student cohort were recruited from the CS7IS1 module at Trinity College Dublin. Each member of the class had the option to complete the experiment as a portfolio task for the course. Each member of the class was sent an email invitation using a template (**Appendix H**) with the Gmail mail merge function[56]. The expert participants were recruited based on a discussion with the supervisor of the study as to who satisfied the inclusion/exclusion criteria. These participants were recruited individually through email invitation generated by the template.
- **Completion of Experiment:** First, participants were required to provide consent to participate in the experiment. The informed consent for this experiment is shown in **Appendix I**. The participants in the expert cohort completed the experiment synchronously using zoom video conferencing platform [89], while their think aloud statements were being recorded using the platform. The participants from the student cohort completed the experiment asynchronously by accessing the framework using provided login details. These participants did not require the use of a video conferencing platform as the think-aloud protocol was not used. It was decided it would not be feasible to arrange a zoom meeting for each student, then manually correct, analyze and tag the resulting 48 transcripts.
- **Experiment Assistance**: Assistance was available to participants in both cohorts if they were unable to complete the tasks. The student cohort were informed that assistance could be provided via email if required. The expert cohort was informed at the start of the experiment that assistance could be provided during the call.
- **Information Provided:** A presentation[57] was presented to participants prior to the experiment which provided background information on the framework and experiment.

[56] https://developers.google.com/apps-script/samples/automations/mail-merge

[57] https://docs.google.com/presentation/d/1_hU2GAIk3Fu5YNyQvmMCAcXZQDuxxvsD

5.4.2.4 Sample Size

The sample sizes used in the experiment differed in several ways which are outlined below.

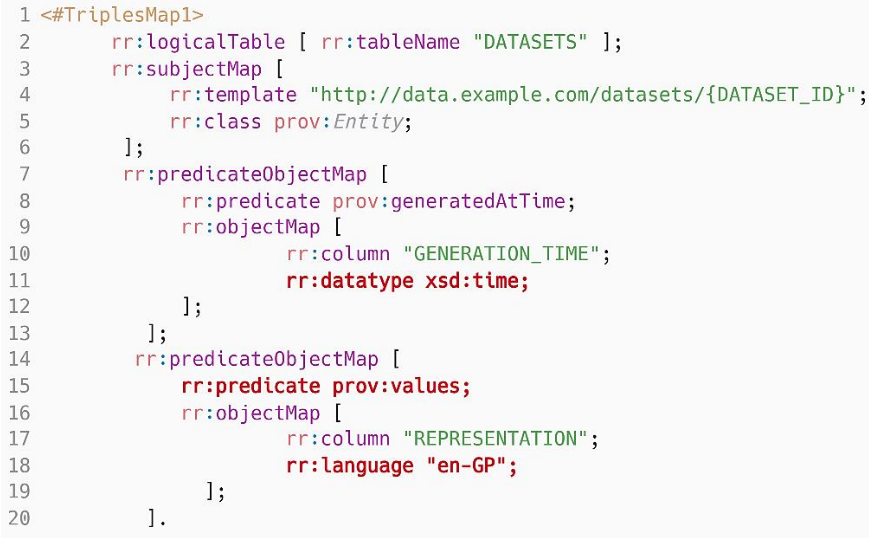
- **Background:** Participants from the student cohort have little knowledge of the theory of the R2RML mapping language. Furthermore, these participants have little experience with creating R2RML mappings, however, they have basic knowledge of semantic web technologies. The participants within the expert 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. Each cohort's background knowledge is further described in each of their respective sections.
- **Number of Participants:** Initially, the student cohort consisted of 58 students. The cohort was reduced to 48 participants after the inclusion/exclusion criteria was applied. The expert cohort consisted of 10 participants. The participants recruited were 80% post-doctoral researchers who hold a PhD degree in semantic web technologies and 20% PhD students who have at least 1 year experience in researching semantic web technologies. All of the expert participants have previous experience in creating and executing R2RML mappings in a research environment.

5.4.2.5 Experiment Tasks

The experiment tasks involved the assessment and refinement of a mapping. Therefore, a sample R2RML mapping was provided to participants to allow them to carry out the required interaction. Furthermore, the mapping needed to contain quality issues in order to ensure the refinement process is initiated.

5.4.2.5.1 Sample Mapping

Participants in both cohorts were provided with an identical sample R2RML mapping (**Listing 14**) which was used to interact with the framework.



```
<#TriplesMap1>
      rr:logicalTable [ rr:tableName "DATASETS" ];
      rr:subjectMap [
           rr:template "http://data.example.com/datasets/{DATASET_ID}";
           rr:class prov:Entity;
       ];
       rr:predicateObjectMap [
            rr:predicate prov:generatedAtTime;
            rr:objectMap [
                     rr:column "GENERATION_TIME";
                     rr:datatype xsd:time;
            ];
         ];
        rr:predicateObjectMap [
            rr:predicate prov:values;
            rr:objectMap [
                     rr:column "REPRESENTATION";
                     rr:language "en-GP";
              ];
         ].
```

Listing 14: R2RML mapping used in Experiment 2, introduced violations in bold type

The use case of the sample mapping involves provenance information relating to datasets being uplifted to RDF, which is hoped can be easily understood by both cohorts as they both have knowledge about datasets. The use case is realistic as the PROV-O documentation includes similar examples. Furthermore, PROV-O was chosen to represent the information as it is the W3C recommendation for capturing provenance information and is widely known. Moreover, PROV-O includes the necessary data type restrictions to introduce a data type violation into the mapping. The provenance information being captured includes the time the dataset was generated (GENERATION_TIME) and a representation (REPRESENTATION) of the data. Three violations were introduced into the mapping. The violations introduced into the mapping were chosen from the violations detected in experiment 1, which indicated these violations occur in real-world mappings. The violations introduced allow the participants to evaluate the various refinement options available in 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 uplift mapping, which was provided to participants are outlined below, together with the associated metric ID.

- **Incorrect data type (D7):** 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.
- **Usage of undefined property (D2):** `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.
- **Invalid language tag (MP11):** 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.

5.4.2.5.2 Task Sheet

Table 39 presents the task sheet used during the experiment. The scenario for the participants is described in **Appendix F**.

Table 39: Task sheet used in Experiment 2

You will be asked to complete the following tasks after you have downloaded and saved the sample mapping to your computer:

1. Upload the provided mapping file to the framework. You will be done when the mapping file has been successfully uploaded and you have pressed the “Assess Mapping Quality” button.
2. Explore the mapping quality assessment information generated by the framework. You will be done when you have acknowledged the number of violations, their result message, value,

location, refinements and display the violation.

3. Explore (by hovering) more information related to each violation result message. You will be done when you have acknowledged the usefulness of the information.
4. Select a refinement for each violation which you consider the most appropriate. The refinement must not be Manual. You will be done when you have selected a refinement for each violation and pressed the “Create Refinements” button.
5. Enter values for each refinement if required. You will be done when you have entered values for each refinement.
6. Select all refinements to be executed. You will be done once you have pressed the “Select All Refinements” button.
7. Execute the refinements. You will be done once you have pressed the “Execute Refinements” button.
8. Explore the mapping quality profile bar chart generated. You will be done when you have acknowledged the violation count for each quality dimension.
9. Export the refined mapping. You will be done when you have successfully exported the refined mapping.
10. Open and examine the refined mapping. You will be done when you have acknowledged that the refined mapping has been generated correctly.
11. Export the validation report. You will be done when you have successfully exported the validation report.
12. Open and examine the validation report. You will be done when you have acknowledged the number of violations and their associated refinements.

The tasks allowed the main characteristics of the mapping quality assessment and refinement component of the MQI framework to be evaluated. Tasks 1-3 involved the quality assessment of a mapping. Tasks 4-7 involved the selection and execution of refinements to remove quality issues within the mapping. Tasks 8-12 involved the examination of quality assessment information in MQIO format and visually.

5.4.3 Results of Student Cohort

5.4.3.1 Usability

Table 40 presents the PSSUQ scores of each participant in the student cohort. The response number relates to the participant ID, therefore, numbers which are not present in the table indicate the invited student was excluded from the results when the inclusion/exclusion criteria in **Section 5.4.2.3** were applied. As a note, PSSUQ scores are graded on a scale of 1 (Best case) and 7 (Worst case). In addition, scores marked as null (∅) in the table indicate the participant did not provide a response to the metric as each question is optional.

Table 40: PSSUQ scores for the **student** cohort

#	SysUse	InfoQual	IntQual	Overall
1	1.38	1.29	1.0	1.26
2	1.38	1.43	4.5	1.74
3	1.75	2.0	1.0	1.68
4	1.62	1.43	2.0	1.58
6	2.0	2.0	2.0	2.0
10	1.0	1.29	3.0	1.32

11	3.12	3.0	3.0	3.0
12	5.25	5.43	4.5	5.26
13	2.12	1.71	2.0	1.84
15	2.0	2.29	4.0	2.47
16	5.62	5.29	6.5	5.63
17	1.0	1.0	1.0	1.0
18	1.0	1.43	3.0	1.37
19	6.38	6.43	5.5	6.37
20	2.0	2.0	2.0	2.0
21	2.25	1.86	4.0	2.32
22	1.88	3.29	1.0	2.37
23	1.0	1.43	2.0	1.32
24	1.75	2.14	2.0	1.89
25	5.0	5.43	4.0	5.16
26	2.25	∅	∅	2.25
27	2.0	2.14	1.5	2.0
28	1.0	1.0	1.5	1.05
29	1.75	2.43	2.5	2.21
30	1.0	1.86	1.0	1.32
31	1.88	2.14	2.0	2.0
33	1.25	1.14	4.5	1.53
35	1.88	1.86	2.0	1.84
36	2.5	1.71	3.0	2.21
37	1.88	1.71	3.0	1.89
38	6.25	4.86	6.0	5.58
39	1.88	2.14	2.5	2.05
40	1.5	1.57	2.0	1.63
41	2.5	2.71	3.0	2.68
42	1.38	2.0	2.5	1.79
43	2.5	2.29	2.0	2.37
44	5.75	5.57	5.5	5.68
45	2.25	2.0	1.5	2.11
48	1.38	1.57	2.0	1.58
50	1.0	1.0	1.0	1.0
51	1.0	1.57	2.0	1.32
53	4.12	4.14	6.0	4.16
54	1.0	1.29	1.5	1.16
55	6.29	6.43	6.5	6.44
56	1.88	1.57	3.0	1.84
57	1.88	2.14	3.0	2.11
58	1.0	1.0	1.0	1.0
59	2.0	2.0	2.0	2.0
Mean score	2.34	2.42	2.8	2.42
± Threshold (%)	+ 19.66	+ 24.28	-11.39	+ 16.53

Figure 39 presents a box plot [83] of the PSSUQ scores for the student cohort. The rectangle of the plot represents 50% of the data points. The position of the line in the rectangle represents how skewed the data is. The line indicates if the data is normally distributed (center), positive skew (left) and negative skew (right). The points outside of the rectangle indicate outliers.

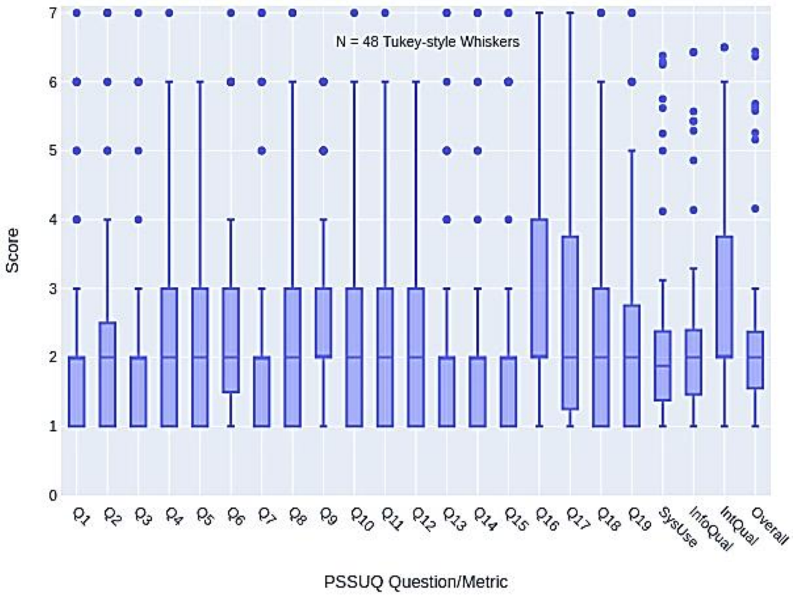


Figure 39: Box plot of PSSUQ scores of **student** cohort

The results indicated sufficient satisfaction for the system usefulness, information quality and overall, with them scoring better than 20% to the acceptable thresholds. However, interface quality scored worse than the threshold by 11.39% indicating the participants were not satisfied with the user interface of the framework.

5.4.3.2 Effectiveness

Table 41 presents the results related to metrics which assess the perceived effectiveness of the framework in undertaking the uplift mapping assessment and refinement tasks.

Table 41: Overview of results for the **student** cohort

#	Time taken to complete experiment (minutes)	Number of violations remaining after refinement complete (original mapping had 3 violations)
1	12	0
2	13	0
3	7	0
4	3	0
6	18	0
10	11	0
11	13	1
12	6	2
13	7	0
15	6	3
16	12	0
17	2	0
18	4	0
19	18	1
20	4	3
21	9	3

22	5	2
23	5	0
24	13	1
25	9	3
26	6	0
27	11	3
28	15	0
29	9	0
30	10	1
31	9	0
33	8	3
35	16	0
36	21	2
37	13	0
38	5	1
39	13	1
40	16	0
41	6	1
42	9	0
43	8	2
44	7	0
45	4	3
48	10	0
50	22	1
51	14	2
53	12	3
54	5	0
55	13	1
56	11	1
57	6	3
58	4	0
59	23	0

Table 42 presents the percentage of violations remaining in the refined mappings from the experiment.

Table 42: Overview of violation counts for the **student** cohort

Number of violations remaining after refinement complete	**Number of participants**
0 violations (best case)	24 participants (50%)
1 violation	10 participants (21%)
2 violations	5 participants (10%)
3 violations (worst case)	9 participants (19%)

Figure 40 presents a box plot of the time taken to complete each task for the student cohort.

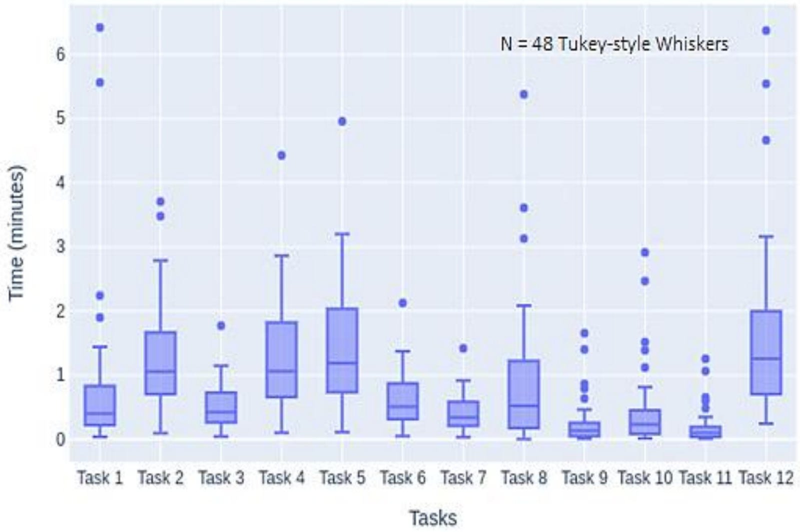


Figure 40: Time spent to complete each task during the usability experiment in minutes for the **student** cohort

The original mapping contained 3 violations. 50% of participants have 0 violations remaining after refining the mapping. 71% have 1 or 0 violations. 30% have 2 or 3 violations. The results indicated certain students encountered difficulties in either detecting or removing the quality issues of the provided mapping. The mean time for the student cohort to complete the tasks was 10.06 minutes. The maximum time was 23 minutes and minimum time was 2 minutes. The high standard deviation (5.04 minutes) of the times indicated that the data is widely spread around the mean, indicating certain students were able to complete the task more efficiently.

5.4.3.3 Discussion

The results of the student cohort are discussed as follows:

- **Interface Quality:** The mean score for the interface quality metric is 2.8 which is the worst scoring metric. Furthermore, the framework scores 10% worse than the threshold as defined in [86]. Furthermore, Q16 and Q17 of the PSSUQ, which relate 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 indicated that the participants found the interface poor quality. In particular, the aesthetics of the framework.
- **Information Quality:** The framework scores 20% better than the information quality threshold. These results indicated that the information provided by the framework is sufficient for the participants to complete the experiment, however, the qualitative analysis indicated 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 to more easily select and execute the refinements.
- **System Usefulness and Overall Usability:** 48 out of the 59 (81%) students invited successfully completed the experiment. These results indicated that 81% of the students could successfully interact with the

framework. The overall usability scored 10% better than the threshold. These results indicated 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.

- **Timing:** The minimum time of 2 minutes based on the experience of the researcher indicated certain students were not careful when completing the experiment. Most of the task times of the student cohort had a median less than 1 minute (8 out of 12). The other tasks had a median time of more than 1 minute but less than 1 minute and a half (4 out of 12). The tasks with the highest median time related to the selection of refinements and the examination of the validation report. These results indicated 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 more easily interpret the report.
- **Violation Count:** Several refined 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.

5.4.3.3.1 Correlations of Usability and Violations

Table 43 presents the correlations between the number of violations remaining and the scores of each metric in the PSSUQ for the student cohort. Statistically significant relationships are marked with an asterisk (*) I.e., p-value is below the confidence level (0.05).

Table 43: Correlation between violation count and PSSUQ scores for the **student** cohort

Violation Count and PSSUQ	Pearson's Test (α = .05)		Spearman's Test (α = .05)	
	Coefficient	p-value	Coefficient	p-value
SysUse	0.19	0.202	0.413	0.004*
InfoQual	0.199	0.18	0.389	0.007*
IntQual	0.211	0.155	0.245	0.096
Overall	0.203	0.171	0.45	0.002*

The correlation for interface quality and violation count is **not statistically significant**. The correlations for system usefulness, information quality and overall usability are **statistically significant**. These correlations show positive values which indicate when effectiveness goes up, the usability score decreases since small values in the PSSUQ indicate better usability. The correlations are monotonic and not linear as Pearson's test did not identify any statistically significant relationships.

5.4.3.3.2 Correlations of Usability and Time

Table 44 presents the correlations between the time for completion of the tasks and the scores of each metric in the PSSUQ for the student cohort.

Table 44: Correlation between time for completion and PSSUQ scores for the **student** cohort

Time and PSSUQ	**Pearson's Test (α = .05)**		**Spearman's Test (α = .05)**	
	Coefficient	**p-value**	**Coefficient**	**p-value**
SysUse	0.051	0.731	0.077	0.604
InfoQual	0.018	0.904	0.027	0.855
IntQual	0.051	0.734	0.116	0.439
Overall	0.038	0.798	0.034	0.816

The correlations shown are **not statically significant**.

5.4.4 Results of Expert Cohort

5.4.4.1 Usability

Table 45 presents the score of each PSSUQ metric for the participants in the expert cohort.

Table 45: PSSUQ scores for the **expert** cohort

#	**SysUse**	**InfoQual**	**IntQual**	**Overall**
60	1.63	3.29	5.5	2.74
61	1.13	1.14	3	1.32
62	1.63	1.71	2.5	1.79
63	2.38	3.14	2	2.58
64	2.25	3.14	2.5	2.68
65	1	1.71	1.5	1.32
66	1.88	3.14	4.5	2.68
67	2	2.86	2	2.32
68	1.75	1.86	2	1.89
69	1.29	2.29	2	1.78
Mean score	1.69	2.43	2.75	2.11
± Threshold (%)	+ 65.7	+ 24.3	- 9.5	+ 33.6

Figure 41 presents a box plot of the PSSUQ scores for the expert cohort.

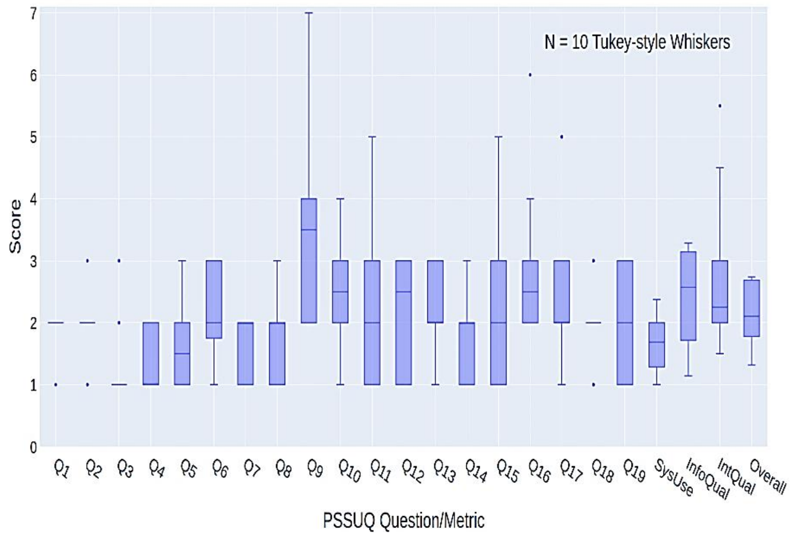


Figure 41: Box plot of PSSUQ scores for the **expert** cohort

The results indicated sufficient satisfaction for the system usefulness, information quality and overall, with them scoring better than 20% to the acceptable thresholds. However, interface quality scored worse than the threshold by 9%, therefore, indicating the participants were not satisfied with the interface of the framework.

5.4.4.2 Effectiveness

Table 46 presents the violation count of each participants refined mapping and the time it took them to complete the experiment.

Table 46: Overview of results for the **expert** cohort

#	Time taken to complete experiment (minutes)	Number of violations remaining after refinement complete (original mapping had 3 violations)
60	17:12	0
61	12:28	0
62	12:54	0
63	12:07	0
64	11:22	0
65	20:41	0
66	11:05	0
67	20:52	1
68	11:20	0
69	24:05	0

Table 47 presents the percentage of violations remaining in the refined mappings from the experiment for the expert cohort.

Table 47: Overview of violation counts for the **expert** cohort

Number of violations remaining after refinement complete	Number of participants
0 violations (best case)	9 participants (90%)
1 violation	1 participant (10%)
2 violations	0 participants (0%)
3 violations (worst case)	0 participants (0%)

Figure 42 presents a box plot of the time taken to complete each task for the expert cohort.

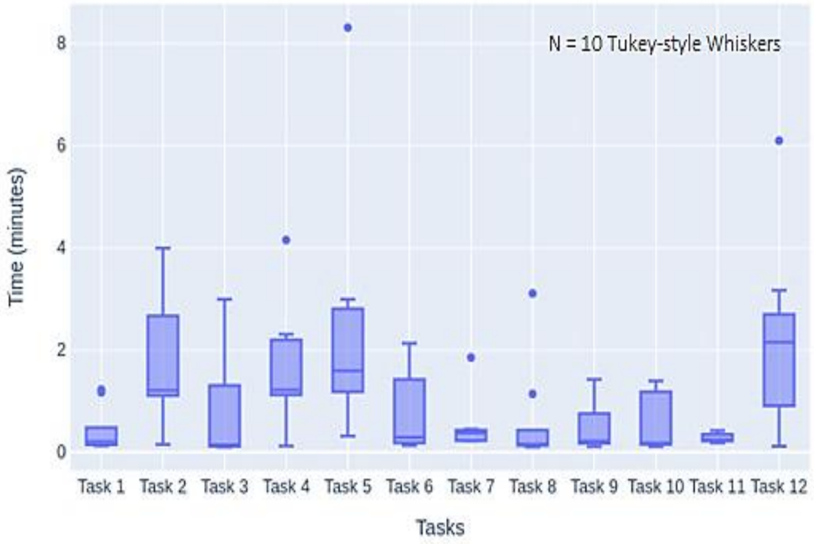


Figure 42: Time spent to complete each task during the usability experiment in minutes for the **expert** cohort

90% of participants have 0 violations in the refined mapping, while 10% have 1 violation in the refined mapping. No participants have 2 or 3 violations in the refined mapping. The mean time is 15.4 minutes. The maximum time is 24.05 minutes and minimum time is 11.05 minutes. The high standard deviation (4.6) of the times means that the data is widely spread around the mean, indicating certain students completed the tasks more efficiently.

5.4.4.3 Discussion

The results of the expert cohort are discussed as follows:

- **Interface Quality:** The interface quality is the worst scoring metric within the PSSUQ with a mean score of 2.75. Furthermore, previous research states that a score of 2.49 or less for the interface quality metric is sufficient, however, the framework scores 10% below the threshold. These results indicated that the interface quality should be improved in future versions. In particular, the aesthetics of the framework need to be improved.

- **Information Quality:** Previous research indicated a mean score of 3.02 or less for the information quality metric is sufficient, with the framework scoring 2.42 which is better than the threshold. These results indicated that the information provided to the users is sufficient.
- **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 and overall usability with a mean of 1.69 and 2.11 respectively. Furthermore, these metrics both scored more than 20% better than the threshold. The results indicated the participants found the system useful with an overall positive user experience.
- **Timing:** The results indicated 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. Most of the task times of the expert cohort had a median less than 1 minute (7 out of 12). The other tasks had a median time of more than 1 minute but less than 1 minute and a half (3 out of 12). The longest tasks had 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. The 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 simplified in future versions to improve the time it takes for participants to interpret the report.
- **Violation Count:** The low violation counts in the refined mappings indicated that the framework could be an effective tool for helping an expert user to identify and remove quality violations with 90% of participants having 0 violations remaining in the refined mapping.

5.4.4.3.1 Correlations of Usability and Violations

Table 48 presents the correlations between the number of violations remaining and the scores of each metric in the PSSUQ for the expert cohort.

Table 48: Correlation between violation count and PSSUQ scores for the **expert** cohort

Violation Count and PSSUQ	**Pearson's Test (α = .05)**		**Spearman's Test (α = .05)**	
	Coefficient	**p-value**	**Coefficient**	**p-value**
SysUse	-0.433	0.211	-0.407	0.243
InfoQual	-0.58	0.079	-0.53	0.115
IntQual	0.069	0.85	0.3	0.399
Overall	-0.496	0.145	-0.467	0.174

The correlations shown are **not statically significant**.

5.4.4.3.2 Correlations of Usability and Time

Table 49 presents the correlations between the time for completion of the tasks and the scores of each metric in the PSSUQ for the expert cohort.

Table 49: Correlation between time for completion and PSSUQ scores for the **expert** cohort

Violation Count and Time	Pearson's Test (α = .05)		Spearman's Test (α = .05)	
	Coefficient	p-value	Coefficient	p-value
SysUse	-0.462	0.179	-0.462	0.179
InfoQual	-0.032	0.931	-0.203	0.574
IntQual	-0.231	0.521	-0.37	0.292
Overall	-0.277	0.438	-0.402	0.249

The correlations shown are **not statically significant**.

5.4.4.4 Thematic Analysis

Appendix L contains the comments which were collected by the PSSUQ open comment section for both cohorts. The themes and codes which were defined as a result of the thematic analysis are described in **Appendix K. Figure 43** presents the number of references of each theme which was defined as a result of the thematic analysis conducted on the comments.

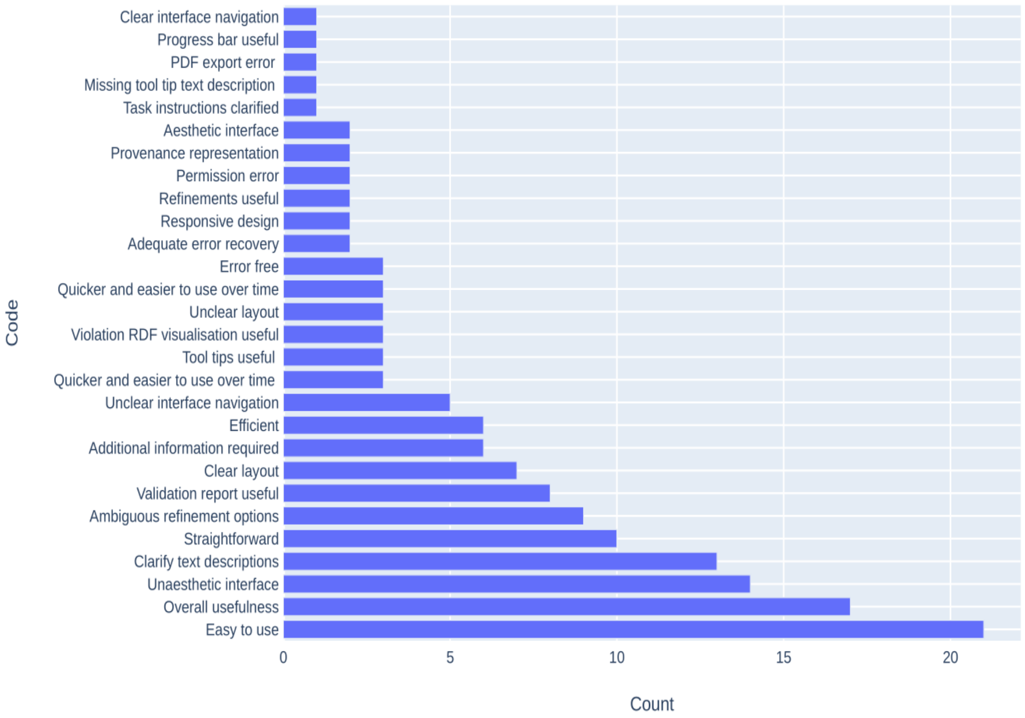


Figure 43: Bar chart of number of references of each code in the data

The codes are discussed as follows:

- **Easy to use** (21 references): The most common code indicated that participants found the tasks straightforward to complete using the framework. Related comments include *"The system was simple to use."* and *"it corrected all the mistakes and helped fix them easily"*.
- **Unaesthetic interface** (14 references) and **Aesthetic interface** (2 references): The second most common code indicated that the interface of the framework is not sufficient. Furthermore, the low occurrences of the code which indicated the aesthetics of the interface are sufficient further reinforces this observation.
- **Overall usefulness** (17 references), **Straightforward** (10 references), **Quicker and easier to use over time** (3 references) and **Error free** (3 references): A total of 33 references for these codes indicated that participants had an overall positive experience when using the framework to complete the tasks and did not encounter any errors.
- **Additional information required** (6 references) and **Ambiguous refinement options** (9 references): These codes mainly related to the lack of textual descriptions for selecting the refinements and values used by them. Related comments include *"Could have additional explanation for the 'type of refinement' options."* and *"Next time I would be, needed to understand better what the drop-down for modification really means."*.

Figure 44 presents the occurrences of the defined themes in the data. The themes were defined as a result of discussion between the author and supervisor of this thesis. It was decided to define themes based on relation to positive and negative patterns discovered as a result of the data analysis. For instance, codes in the *"Positive GUI requirements"* theme indicated sufficient GUI usability, such as an aesthetic interface and clear layout. However, codes in the *"Negative GUI requirements"* theme indicated insufficient GUI usability, such as an unaesthetic interface and unclear layout. Therefore, the themes would provide groupings of codes and provide indications of areas in the framework which fulfill the user's expectations or require improvements.

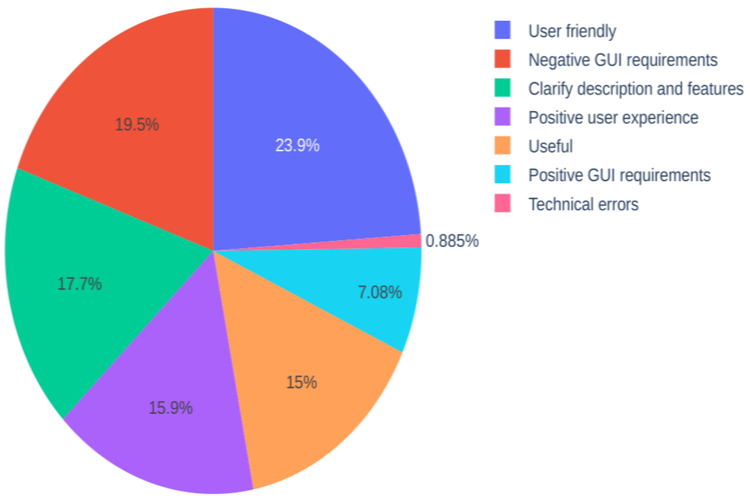


Figure 44: Pie chart of themes in the qualitative data

The themes are discussed as follows:

- **User friendly** (23.9%), **Positive user experience** (15.9%), **Useful** (15%): These themes accounted for nearly 60% of the codes. Therefore, indicating that participants found the overall experience provided by the framework positive with the most common codes in these themes being *"Easy to use"* and *"Overall Usefulness"*.
- **Negative GUI Requirements** (19.5%) and **Positive GUI Requirements** (7.08%): The results indicated that more participants found the interface insufficient. Most code references related to the aesthetics of the framework rather than the layout and navigation.
- **Clarify Textual Descriptions and Features** (17.7%): Refinement options were mentioned most often in this theme. The lack of additional text describing the refinements and associated value was not intuitive for certain participants when completing the refinement of the mapping.
- **Technical Errors** (0.885%): Pop up blocked resulted in the frequency of technical errors. However, most participants did not encounter this problem.

5.4.5 Recommended Improvements

This section discusses recommended improvements which were discovered through the analysis of the results. In addition, the solutions for the recommendations are mentioned.

- **Aesthetics of Framework:** 13 references mentioned the aesthetics of the interface of the framework as unacceptable. Related comments include *"Not very esthetically pleasing"* and *"Interface was responsive/easy to use, but could look nicer."* Therefore, indicating the aesthetics should be improved to provide a better user experience. The aesthetics of the framework were improved by using bootstrap templates which are designed by user experience specialist.

- **Textual Descriptions of Refinements:** 9 references mentioned difficulties when choosing a refinement as they lacked further textual descriptions. Related comments included *"Could have additional explanation for the 'type of refinement' options."* and *"Next time I would be, needed to understand better what the drop-down for modification really meant"*. Additional text was added to the refinements. For instance, the refinement named *"Choose a new predicate"* had the following sub text add to the option which stated, *"This refinement allows you to enter a URI or create one using a list of prefixes"*. It is hoped the additional information will allow the users to choose a refinement more straightforward.
- **Structure of Provenance Information:** While most feedback received on the validation reports represented in MQIO was positive. 2 references mentioned that inclusion of the blank node identifier in the SPARQL queries used for refinements was not intuitive as these are dynamic values. The issue was resolved by replacing the blank node identifier with a placeholder.
- **Additional Information on Metrics:** 13 references related to the *"Clarify text descriptions"* code. Majority of these references were as a result of participants being unable to understand the metric ID for each violation, which were displayed on the framework. Related comments included *"Not quite sure what this metric ID is right now."* and *"Okay, we scan over this first. So, three violations, random metric IDs i don't understand"*. Therefore, it was decided to include a hyperlink to the documentation where these IDs are defined in hopes of improving understanding of them.
- **Pop up Blocked:** 2 references mentioned that the pop up to export files on the framework was blocked by the browser. The participants who encountered the problem resolved the issue by providing permission to the framework. However, the issue was resolved by changing the protocol used by the framework from HTTP to HTTPS which provides additional security.
- **Ordering of Components:** The ordering in which information is shown could be improved. The main area of improvement involves the following. Displaying validation bar chart before refinements are executed. The bar chart has been moved from the refinements page to below the quality assessment table to resolve the issue.

The results indicated that participants in both cohorts would benefit from these improvements.

5.4.6 Discussion

This section discusses 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 expert cohort scored the interface quality (2.75) slightly better than the student cohort (2.91). However, these are the worst scoring metrics for both cohorts, which indicated that they were not satisfied by the interface. The expert cohort could have found the interface easier to use due to their previous experience in using semantic web related interfaces. Furthermore, both cohorts scored worse than the acceptable threshold. The results indicated that the aesthetics of the interface should be improved for both cohorts.

- **Information Quality:** The student cohort scored the same information quality (2.42) as the expert cohort. However, 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 indicated 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, both cohorts scored better than the threshold indicating overall sufficient information quality.
- **System Usefulness and Overall Usability:** Only one participant in both cohorts required assistance during the completion of the tasks. Experts scored these metrics the best, while students scored the information quality best. However, both cohorts scored each metric better than their corresponding thresholds by at least 15%. The results indicated the participants found the system useful with an overall positive user experience.
- **Analysis of each cohort's PSSUQ question scores:** 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. The question related to error messages has been noted as an outlier in the questionnaire as none may be shown to participants. 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 indicated that the aesthetics of the interface should be improved for both cohorts in future versions of the framework.
- **Violation Count:** 71% of the student cohort have 0 or 1 violations in the refined mapping, while 90% of the expert cohort have 0 violations, which could indicate that the background knowledge of the expert cohort helped them to identify and remove the quality violations. Furthermore, no expert had 3 violations, while 10% of the student cohort had 3 violations. These results indicated that the expert cohort found the framework more effective.
- **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 third quartile of participants completed the experiment in 20 minutes or less. However, 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 of 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 of 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 indicated 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. 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 background knowledge.

- **Correlations:** No correlations were identified in the expert cohort, however, the student cohort had correlations between the system usefulness, information quality and overall usability and violation count. Therefore, indicating students benefitted more from the information displayed on the framework during the experiment which could be as a result of experts not requiring additional information due to their background knowledge.

5.4.7 Conclusion

Hypothesis H2 (The MQI framework facilitates the assessment and refinement of uplift mappings) is **supported** as 90% of the participants of the expert cohort could complete the experiment tasks without assistance. 1 participant required assistance after skipping a task and pressing the incorrect button, which is not a problem relating to the framework design. Furthermore, 90% of the expert cohort participants have 0 violations, which indicated that they could successfully identify and remove quality issues using the framework. 81% of the students invited completed the experiment. No student required assistance to complete the tasks. Furthermore, 50% of the students identified and removed all violations, however, 18% of participants failed to remove any violations from the mapping. These results indicated that the framework facilitates the assessment and refinement of mappings for certain participants. Several students' refined mappings contained violations such as including a data type named `admingeo:a`, `date:xsd` and `rr:generatedAtTime`, which are not data types. Other examples of violations include a property named `aair:http://www.w3.org/r2rml#` and `example:ex` which are undefined. These are simple violations and could indicate that only students with very little background knowledge were influenced. The framework requires a basic understanding of semantic web technologies and is not designed to teach participants these technologies. The overall consensus gathered from the analysis is that the framework facilitates the assessment and refinement of mappings for participants with little background knowledge of semantic web technologies as 85% of all participants could remove at least 1 violation from the mapping. Furthermore, only one participant required assistance to complete the experiment. Moreover, the most common themes (*"Positive user experience", "User friendly")* discovered by thematic analysis identified patterns related to overall positive usability.

Hypothesis H3 (The participants background knowledge influences the successful completion of the mapping tasks with the MQI framework) is **supported**. The results indicated that the background knowledge of the participants influenced the effectiveness of the framework, as no expert failed to remove a violation from the mapping, while 19% of the students failed to remove a violation. Furthermore, none of the expert mappings contained simple violations similar to those mentioned. These results could indicate that participants who have more background knowledge found the framework more effective. The PSSUQ results indicated that each cohort rated the information and interface quality similarly. However, the faster mean time for the student cohort could be as a result of the expert cohort spending more time analyzing the framework when completing the experiment tasks. Furthermore, the use of the think-aloud protocol for the expert cohort could have slowed them when completing the experiment as they had to verbalize each of their actions.

5.5 Experiment 3: Understanding of Change Detection Information

This section presents a user evaluation of the **source change detection component** of the MQI framework which evaluated the perceived usability and understanding of the information related to detected source data changes and respective mappings provided by the Source change detection component of the framework. **Supplementary information related this experiment is stored in a folder ("/Experiment-3") of the GitHub.** The version of the OSCD used during this experiment was **version 1.0**.

5.5.1 Hypotheses

The hypotheses related to this study were:

- **Hypothesis H4:** The framework facilitates the identification of changes in source data and relationships with respective mappings;
- **Hypothesis H5:** The participants background knowledge influences the successful completion of the tasks.

5.5.2 Experimental Setup

5.5.2.1 Methodology

A user evaluation was conducted to test the hypotheses of the study. The experiment involved participants interacting with the framework using two versions of source data and respective mapping which was collected from the RML test case files[58]. The two versions of source data allowed the framework to detect changes between them, therefore, allowing participants to use the framework to identify links between detected changes and respective mappings. Participants were grouped into a student and expert cohort. The student cohort contained participants in a third level class, while the expert cohort contained researchers who had experience creating and operating LD mappings. Participants in both cohorts were provided with a document containing background information on the framework and all other information necessary to complete the experiment such as the task sheet and URL. No other information was provided to participants. Thereafter, they could access the framework at any time. To test hypothesis **H4** the participants answered the understanding questionnaire which asks them to identify the changes to the structure and content of the source data used by the mapping. In addition, the participants answered the PSSUQ, which provided indications of user satisfaction. Hypothesis **H5** was tested by comparing the results collected from participants in each cohort. Ethical approval has been received from the Research Ethics Committee at Trinity College Dublin.

5.5.2.2 Experiment Layout

The structure of the experiment is outlined below.

- **Inclusion/Exclusion Criteria:** Participants in the expert cohort satisfied each of the following criteria: 1) Researchers 2) Knowledgeable in RDF and R2RML 3) Previous experience creating R2RML mappings and 4) Previous experience executing R2RML mappings. Participants in the student cohort satisfied each of the following criteria: 1) Third level student 2) Attempted Experiment and 3) Provided answers to the questionnaires.
- **Recruitment:** The participants from the student cohort were recruited for the CS7IS1 module at Trinity College Dublin. Each member of the class had the option to complete the experiment as a portfolio task for the course. Each student was sent an email invitation using a template (**Appendix N**) with the Gmail mail merge function. The expert participants were recruited based on a discussion with the supervisor of the study who would satisfy the inclusion/exclusion criteria. These participants were recruited individually

[58] https://rml.io/test-cases/

through email invitation. These participants completed the experiment to contribute to the research objectives.

- **Completion of Experiment:** First, participants were required to provide consent to participate in the experiment. The informed consent for this experiment is presented in **Appendix M**. The participants in both cohorts completed the experiment asynchronously by accessing the framework using a provided URL.
- **Experiment Assistance**: Assistance was available to participants in both cohorts if they were unable to complete the tasks. The participants were informed that assistance could be provided via email if required.
- **Information Provided:** The information provided to each cohort during the experiment was identical. The participants were provided with a document[59], which contained background information on the framework and experiment, links to the experiment data, task sheet and access details for the framework. No additional information was provided.

5.5.2.3 Sample Size

The sample size of the experiment differed in several ways which are outlined below.

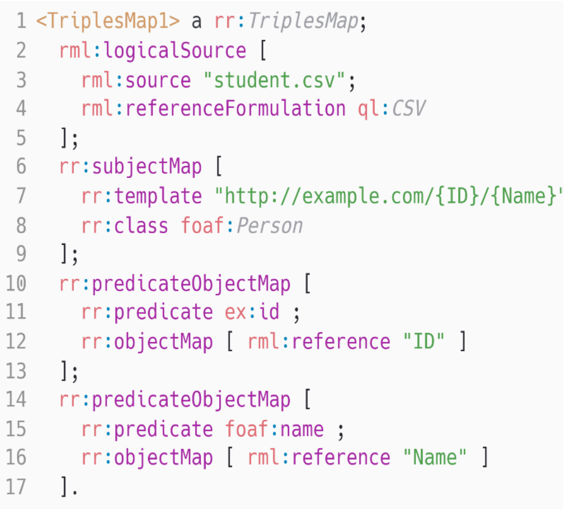
- **Background:** Participants in the student cohort have little knowledge of the mapping process involved in creating LD datasets. The participants have little experience in creating and operating mappings, however, they have a basic knowledge of semantic web technologies. Participants in the expert cohort are researchers who are very knowledgeable with RDF and mapping languages. The participants have experience in creating and operating mappings in a research environment.
- **Number of Participants:** The student cohort consists of 48 students, which was reduced to 45 participants after the inclusion/exclusion criteria was applied. The expert cohort consisted of 10 participants.

5.5.2.4 Experiment Tasks

5.5.2.4.1 Source Data

Source data and respective mappings were supplied to the participants to allow them to detect changes and relationships between them using the framework. An RML [44] mapping (**Listing 15**) and two versions of source data in CSV format were retrieved from the RML test case files.

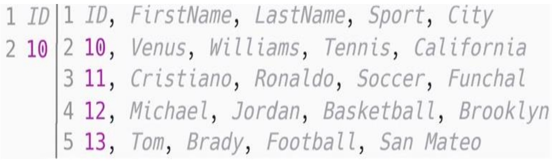
[59] https://drive.google.com/file/d/1_g9ATvqQbr7M3W2PTnwBm1sxR9P2yO0g



```
<TriplesMap1> a rr:TriplesMap;
  rml:logicalSource [
    rml:source "student.csv";
    rml:referenceFormulation ql:CSV
  ];
  rr:subjectMap [
    rr:template "http://example.com/{ID}/{Name}"
    rr:class foaf:Person
  ];
  rr:predicateObjectMap [
    rr:predicate ex:id ;
    rr:objectMap [ rml:reference "ID" ]
  ];
  rr:predicateObjectMap [
    rr:predicate foaf:name ;
    rr:objectMap [ rml:reference "Name" ]
  ].
```

Listing 15: RML mapping used during experiment 3

Listing 16 presents the original version (left) and changed version (right) of the source data.



```
ID
10
```

```
ID, FirstName, LastName, Sport, City
10, Venus, Williams, Tennis, California
11, Cristiano, Ronaldo, Soccer, Funchal
12, Michael, Jordan, Basketball, Brooklyn
13, Tom, Brady, Football, San Mateo
```

Listing 16: Original (left) and Latest (right) CSV source data used in Experiment 3

The source data contains data about sports personalities such as their name, sport and city of birth. The graph generated by the framework as a result of detecting changes between the versions of the source data is presented in **Appendix O**.

5.5.2.4.2 Task Sheet

Table 50 presents the task sheet used during the experiment.

Table 50: Task sheet used in Experiment 3

1. Press the “Quality Assessment” button. You will be done once you have been redirected to the “Mapping Quality Assessment” page.
2. Upload the provided RML mapping to the page. Do not enter any additional information. You will be done once the mapping has been uploaded to the framework.
3. Press the “Home” button on the menu bar of the page. You will be done once you have been redirected to the “Choose Mode” page.
4. Press the “Change Detection” button. You will be done once you have been redirected to the “Choose Data Format” page.
5. Choose the correct data format for the data provided to you. **Two data formats have been disabled for the experiment.** You will be done once you have selected a data format and been redirected to another page.
6. Enter the URL of both of the source data files provided to you. **The details of the notification policy have been fixed for the experiment.** You will be done once you have entered the URLs for the source data.

7. Press the "Start Change Monitoring" button. You will be done once you have been redirected to the "Change Detection Processes" page.
8. Examine the information provided by the framework. You will be done once you have examined the information including hover text and hyperlinks. **You should keep this tab open as the information shown will be required to complete the questionnaire.**
9. ***(Expert Only)*** Download the graph generated by clicking the "Download Changes Graph" button. You will be done once you have successfully downloaded the graph to your computer.
10. ***(Expert Only)*** Examine the information contained within the downloaded graph. You will be done once you have examined the graph with respect to the application of the Ontology for Source Change Detection (https://w3id.org/OSCD). **You will be asked to provide feedback on the graph and ontology in the questionnaire.**
11. Press the hyperlink for the mapping under the "Mappings Impacted" column of the table. You will be done once you have been redirected to the "Mapping-Source Data Change Relations" page.
12. Examine the information displayed by the framework. You will be done once you have expanded each drop down and examined the information within them. **You should keep this tab open as the information shown will be required to complete the questionnaire.**
13. Press the "Complete Questionnaire" button. You will be done once you have successfully completed the questionnaire.

The tasks allowed the main characteristics of the source change detection component of the MQI framework to be evaluated. Tasks 1-2 involved the quality assessment of the mapping. Tasks 3-7 involved initiation of the change detection process on the source data. Task 8 involved the examination of the overview of the change detection processes. **Task 9 and 10 only applied to the expert cohort.** The two tasks were designed to retrieve expert feedback on the application of OSCD within the graphs generated. The participants in the student cohort were not asked for feedback as their knowledge of ontology design and application are limited. Task 11-12 involved the examination of relationships between changes in the source data and mapping. Task 13 involved the completion of the questionnaires which measured usability and understanding.

5.5.3 Results of Student Cohort

5.5.3.1 Usability

Table 51 presents the scores of each PSSUQ metric for the student cohort. As a note, PSSUQ scores are graded on a scale of 1 (Best case) and 7 (Worst case).

Table 51: PSSUQ scores for the **student** cohort

#	**SysUse**	**InfoQual**	**IntQual**	**Overall**
0	2.25	2.00	2.33	2.11
1	1.88	3.17	2.00	2.28
2	1.00	1.00	1.00	1.00
3	1.88	1.33	2.33	1.83
4	2.25	4.00	2.67	2.94
5	1.00	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00
7	1.12	1.86	1.33	1.42

8	2.00	2.00	2.00	2.00
9	1.12	2.14	1.33	1.53
10	1.00	2.29	2.00	1.68
11	1.38	1.57	1.00	1.37
12	1.00	1.00	1.00	1.00
13	1.25	1.40	1.33	1.29
14	1.00	2.14	1.33	1.47
15	1.00	1.29	1.00	1.11
16	1.00	1.00	1.00	1.00
17	1.00	1.00	1.00	1.00
18	1.00	1.57	1.00	1.21
19	2.00	3.43	3.00	2.74
20	1.88	2.29	1.67	1.95
21	1.75	1.86	1.67	1.74
22	2.50	2.14	2.33	2.32
23	1.00	1.33	2.67	1.57
24	1.25	2.33	1.00	1.89
25	1.88	2.40	2.33	2.35
26	2.25	2.14	1.67	2.11
27	1.75	1.86	2.67	2.00
28	1.00	1.00	4.00	1.44
29	2.75	2.71	2.33	2.63
30	4.50	4.71	4.67	4.58
31	1.00	1.86	1.67	1.42
32	1.12	1.14	2.00	1.32
33	1.50	1.60	2.67	1.71
34	3.38	3.00	3.00	3.11
35	2.00	2.14	2.67	2.16
36	2.00	3.86	2.00	2.74
37	1.00	1.00	1.00	1.00
38	2.75	2.86	2.33	2.74
39	3.00	2.71	1.67	2.68
40	3.00	3.00	4.00	3.16
41	1.25	1.43	2.00	1.47
42	3.75	3.14	3.00	3.37
43	1.25	1.50	2.33	1.56
44	1.25	3.57	2.33	2.32
45	1.00	1.14	4.00	1.74
Mean score	1.71	2.06	2.05	1.91
± Threshold (%)	+ 63.30	+ 46.35	+ 21.42	+ 47.38

Figure 45 presents a box plot of the PSSUQ scores for the student cohort. The rectangle of the plot represents 50% of the data points. The position of the line in the rectangle represents how skewed the data is. The line indicates if the data is normally distributed (center), positive skew (left) and negative skew (right). The points outside of the rectangle indicate outliers.

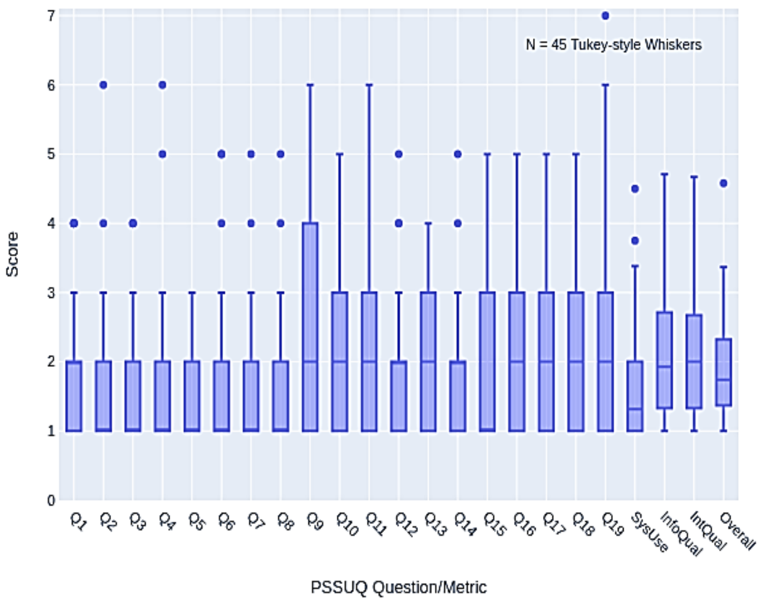


Figure 45: Box plot of PSSUQ scores for the **student** cohort

The mean score for each metric has been compared with acceptable thresholds found in research [86]. The results indicated sufficient satisfaction with each metric scoring between 21% - 63% better.

5.5.3.2 Understanding

Table 52 presents the results from the understanding questionnaire for the student cohort, including the mean scores and standard deviation (SD) of each section.

Table 52: Understanding scores for the **student** cohort

#	Section 1	Section 2	Section 1 and 2
0	1.00	0.83	0.92
1	0.33	1.00	0.67
2	1.00	1.00	1.00
3	0.50	0.67	0.58
4	0.83	1.00	0.92
5	0.67	0.67	0.67
6	1.00	1.00	1.00
7	0.83	1.00	0.92
8	0.50	0.33	0.42
9	1.00	1.00	1.00
10	0.67	1.00	0.83
11	0.50	1.00	0.75
12	0.83	1.00	0.92
13	0.67	1.00	0.83
14	0.83	0.83	0.83
15	0.83	1.00	0.92
16	0.83	1.00	0.92
17	0.67	1.00	0.83
18	0.83	0.92	0.88

19	0.83	0.83	0.83
20	0.67	0.83	0.75
21	1.00	0.83	0.92
22	0.83	1.00	0.92
23	0.83	1.00	0.92
24	0.83	0.83	0.83
25	0.83	0.83	0.83
26	1.00	1.00	1.00
27	0.83	0.83	0.83
28	1.00	1.00	1.00
29	0.83	0.83	0.83
30	0.83	0.83	0.83
31	0.67	0.33	0.50
32	0.83	1.00	0.92
33	0.83	0.83	0.83
34	1.00	0.67	0.83
35	1.00	1.00	1.00
36	0.83	1.00	0.92
37	1.00	0.83	0.92
38	0.83	1.00	0.92
39	0.75	1.00	0.88
40	0.83	1.00	0.92
41	0.67	0.67	0.67
42	0.83	1.00	0.92
43	1.00	1.00	1.00
44	0.83	0.83	0.83
45	0.67	0.83	0.75
Mean scores	0.81	0.89	0.85
SD	0.154	0.160	0.127

The results indicated that participants in the student cohort were able to understand the information provided by the framework as the mean score of both sections in the questionnaire was 85%. In addition, the low standard deviation of each section indicated the scores are clustered around the mean.

5.5.3.3 Correlation of Usability and Understanding

Table 53 presents the correlations between scores of each metric in the PSSUQ and the understanding questionnaire for the student cohort.

Table 53: Correlation between understanding and PSSUQ scores for the **student** cohort

Understanding and PSSUQ	**Pearson's Test (α = .05)**		**Spearman's Test (α = .05)**	
	Coefficient	**p-value**	**Coefficient**	**p-value**
SysUse	0.019	0.9	-0.037	0.807
InfoQual	-0.014	0.924	-0.136	0.367
IntQual	0.035	0.816	-0.008	0.956
Overall	0.0	0.997	-0.083	0.582

The correlations shown are **not statically significant**.

5.5.3.4 Discussion

The results of the student cohort are discussed as follows:

- **Interface Quality:** The interface quality is the worst scoring metric within the PSSUQ with a mean score of 2.05. Previous research states that a score of 2.49 or less for the interface quality metric is sufficient. The framework scored 21% better than the threshold. These results indicated that the interface of the framework is sufficient for users.
- **Information Quality:** The research indicated a mean score of 3.02 or less for the information quality metric is sufficient, with the framework scoring 2.06 which is 46% better than the threshold. These results indicated that the information provided to the students is sufficient.
- **System usefulness and Overall usability:** No participants required assistance during the completion of the tasks. The best scoring metrics related to system usefulness and overall usability with a mean of 1.72 and 1.91 respectively. Furthermore, these metrics both scored more than 45% better than the threshold. The results indicated the participants found the system useful with an overall positive user experience.
- **Understanding Questionnaire:** Questions in section 2 scored 8% better than section 1 in the understanding questionnaire. Therefore, indicating the participants could understand the information easier on the second page. The page provides an overview of the changes which occurred in the source data. Whereas the first page showed an overview of the different aspects of the change detection processes currently running. The difference could be as a result of the amount of information displayed on the first page which contained far more information. However, the scores indicated students could understand both pages with a high degree of accuracy as the mean score of both sections was 85%.

5.5.4 Results of Expert Cohort

5.5.4.1 Usability

Table 54 presents the scores of each PSSUQ metric for the expert cohort.

Table 54: PSSUQ scores for the **expert** cohort

#	SysUse	InfoQual	IntQual	Overall
0	1.62	1.71	1.33	1.63
1	1.00	1.57	1.67	1.32
2	1.75	1.60	2.00	1.76
3	2.75	2.43	1.33	2.32
4	1.25	4.14	2.33	2.53
5	3.62	5.00	5.00	4.07
6	2.12	1.86	2.00	2.00
7	1.12	1.57	2.33	1.53

8	1.25	1.71	2.00	1.53
9	1.38	2.14	1.67	1.68
Mean scores	1.79	2.37	2.17	2.04
± Threshold (%)	+ 56.64	+ 27.20	+ 14.92	+ 38.49

Figure 46 presents a box plot of the PSSUQ scores for the expert cohort.

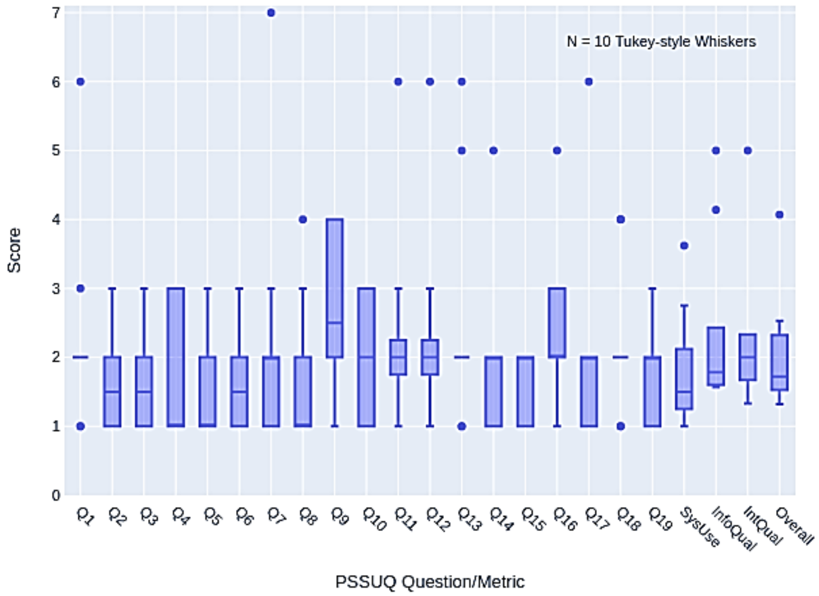


Figure 46: Box plot of PSSUQ scores for the **expert** cohort

The mean score for each metric has been compared with acceptable thresholds found in research [86]. The results indicated sufficient satisfaction with each metric scoring between 14% - 56% better.

5.5.4.2 Understanding

Table 55 presents the results from the understanding questionnaire for the expert cohort, including the mean scores and standard deviation (SD) of each section.

Table 55: Understanding scores for the **expert** cohort

#	**Section 1**	**Section 2**	**Section 1 and 2**
0	1.00	1.00	1.00
1	1.00	1.00	1.00
2	0.83	1.00	0.92
3	1.00	1.00	1.00
4	0.83	0.42	0.62
5	0.83	1.00	0.92
6	1.00	0.83	0.92
7	1.00	1.00	1.00
8	1.00	1.00	1.00
9	0.83	0.67	0.75
Mean score	0.93	0.89	0.91

SD	0.083	0.189	0.123

The results indicated that participants in the expert cohort were able to understand the information provided by the framework as the mean score of both sections in the questionnaire was 91%. In addition, the low standard deviation of each section indicated the scores are clustered around the mean.

5.5.4.3 Correlation of Usability and Understanding

Table 56 presents the correlations between scores of each metric in the PSSUQ and the understanding questionnaire for the expert cohort.

Table 56: Correlation between understanding and PSSUQ scores for the **expert** cohort

Understanding and PSSUQ	**Pearson's Test (α = .05)**		**Spearman's Test (α = .05)**	
	Coefficient	**p-value**	**Coefficient**	**p-value**
SysUse	0.139	0.702	-0.224	0.535
InfoQual	-0.513	0.129	-0.574	0.083
IntQual	-0.173	0.633	-0.375	0.285
Overall	-0.251	0.484	-0.625	0.053

The correlations shown are **not statically significant**.

5.5.4.4 Recommendations on OSCD

Table 57 presents each response to the questions which asked for expert feedback on the application of OSCD in the graphs used in the experiment. The questions which were posed to the experts are shown below.

- **Q1)** Do you think the OSCD should be altered to include new concepts/relationships?
- **Q2)** Do you think the graph of changes detected generated by the application based on the OSCD (as a vocabulary) could be better organized or presented to the user?
- **Q3)** Any additional comments?

Each response was reviewed. The method to address each recommendation is mentioned. Not applicable (N/A) is stated if the response does not include any recommendations on the application of OSCD.

Table 57: Responses for questions related to the application of OSCD from the **expert** cohort

#	Response	Method to Address
Q1	*"I would change the "event" class to "change" since you do not mention events but changes in your UI and documentation even the ontology is named "Ontology for Source Change Detection". Would it be possible to add also the responsible of the change like the mantainer? This would be useful in case the mantainer changes in long projects."*	The change class is subclass of an event rather than an event itself. Multiple maintainers are allowed as the predicate which represents them in the ontology is not defined as a functional property.
	"No"	N/A

	“I can't judge that yet, I got to know OSCD for the first time during this experiment.”	N/A
	“This could be useful to have as an option”	N/A
	“should define subclass/subproperties of prov-o”	PROV-O was found to not be suitable to model changes.
	“No, I don't”	N/A
	“No, but I have not used it properly so there may be something that would need to be added in the future.”	N/A
	“provenance data related to who made the changes but it might be difficult to find that info in the ontology metadata. Also the time period the change has been made (after how long the change was made). But these are only minor things and only some suggestions to consider.”	Additional property named “wasChangedBy” was added to the ontology.
	“Maybe by including the previous value for UpdateSourceData”	A new predicate named “hasPreviousValue” was added to the ontology.
	“No”	N/A
Q2	*“You could summarise the changes in a table format instead of a dropdown. I would add a summary section with the number of Insert/Deletion/... changes and then a table underneath with the actual changes to the data like a DIFF table. You could also make it available as an exportable CSV and HTML files. This could be useful if the data engineer needs to share the results with a wider group of non-tech people.”*	N/A
	“Yes. Sorted results by insert change number or by column and data reference”	N/A
	“I can't judge that yet, I got to know OSCD for the first time during this experiment.”	N/A
	“Seems clear to me”	N/A
	“No, I don't”	N/A
	“No”	N/A
	“It seems to be well presented.”	N/A
	“Prefixing data properties by "has" (e.g. hasStructureReference, hasChangedData, ...) make it seems like they are boolean values or object properties to me”	The properties are already prefixed with “has” as shown in the ontology documentation (https://w3id.org/OSCD).
	“The notification policy could also be presented to the user in the graph of changes as RDF.”	The notification policy was removed from the graph for the purposes of the experiment to focus on OSCD.
Q3	*“This might be out of the scope of the framework ... but would it be possible to show the parts of the RML/R2RML mapping that need to be changed based on the new data source? Could you add example sections in the updated mapping so the user can change them to whatever property and shape that they want? For example, a "lastName" column has been added, so add a* *rr:predicateObjectMap [* *rr:predicate example:lastName ;* *rr:objectMap [rml:reference "lastName"]* *].*	The functionality stated is out of scope of this iteration of the framework, however, it could be considered for future work.

	Something like this to highlight that there is new data that could be included. You can even add a free text input where the user can edit in your UI the mapping based on your suggestions. Then, you validate the mapping with your other component so the user can make sure the mapping is valid."	
	"No"	N/A
	"Seems like a useful tool"	N/A
	"None"	N/A
	"No"	N/A

5.5.4.5 Discussion

The results of the expert cohort are discussed as follows:

- **Interface Quality:** The interface quality is the worst scoring metric within the PSSUQ with a mean score of 2.17. However, previous research states that a score of 2.49 or less for the interface quality metric is sufficient. The framework scores 14% better than the threshold. These results indicated that the interface of the framework is sufficient for users.
- **Information Quality:** The research indicated a mean score of 3.02 or less for the information quality metric is sufficient, with the framework scoring 2.37 which is 27% better than the threshold. These results indicated that the information provided to the users is sufficient.
- **System Usefulness and Overall Usability:** No participants required assistance during the completion of the tasks. The best scoring metrics related to system usefulness and overall usability with a mean of 1.79 and 2.04 respectively. Furthermore, these metrics both scored more than 35% better than the threshold. The results indicated the participants found the system useful with an overall positive user experience.
- **Understanding Questionnaire:** Questions in section 1 scored 4% better than section 2 in the understanding questionnaire. Therefore, indicating the participants were able to understand the overview of the change detection processes easier than the detected changes. However, 85% of the questions in both sections of the questionnaire were answered correctly, therefore, indicating the participants could successfully understand the overall information provided by the framework.

5.5.5 Thematic Analysis

Appendix P contains the comments which were collected by the PSSUQ open comment section for both cohorts. The themes and codes which were defined as a result of the thematic analysis are described in **Appendix K**. **Figure 47** presents the number of references of each code which was defined as a result of the thematic analysis conducted on the comments.

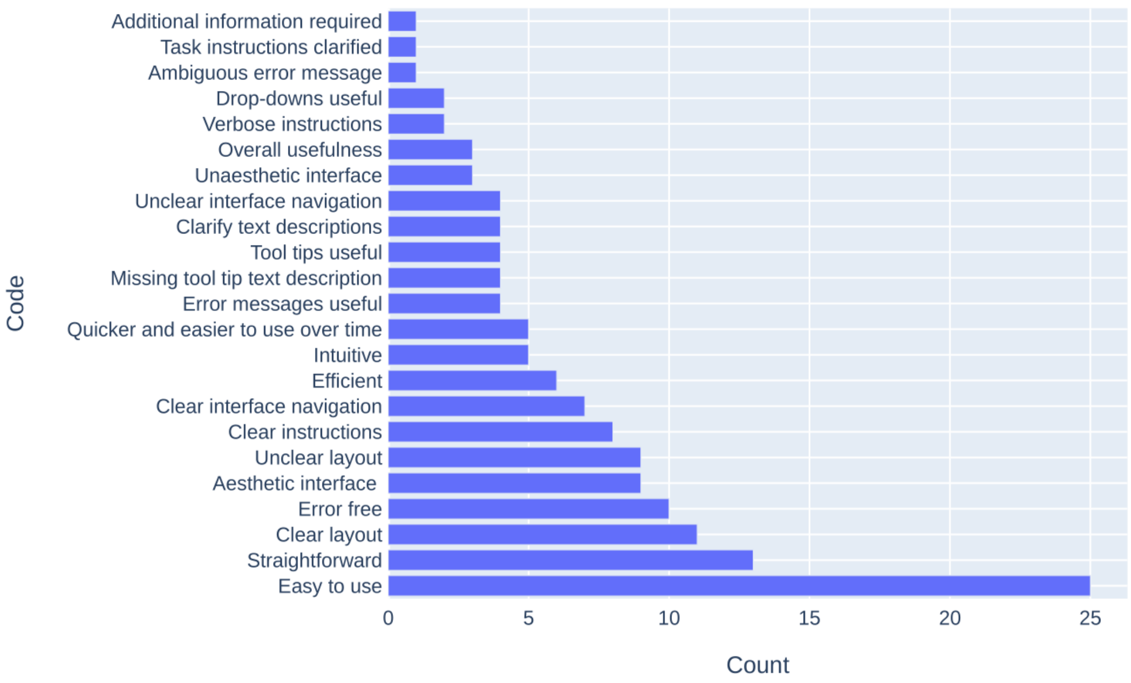


Figure 47: Bar chart of number of references of each code in the data

The codes are discussed below.

- **Easy to use** (25 references): The most common code identifies the framework as easy-to-use. Therefore, indicating participants were able to easily complete the required tasks using the framework. Comments related to this code included *"Overall, very easy to use."*, *"its easy to use the system"* and *"So easy, very intuitive!"*.
- **Aesthetic interface** (9 references) and **Unaesthetic interface** (3 references): The higher number of occurrences related to the framework including an aesthetic GUI indicated that the interface quality is sufficient. Comments related to this code included *"I really like the UI"*, *"I think the UI is presented in a manner that makes it feel simple to use which makes me feel comfortable"* and *"I really like the UI"*.
- **Tool tips useful** (4 references) and **Missing tool tip text description** (4 references): Interestingly, codes related to the tool tips being useful and tool tips which did not work occurred the same number of times. Therefore, the results indicated that the tool tips worked for specific platforms, while others could not utilize them.

Clear layout (11 references), **Straightforward** (13 references) and **Clear interface navigation** (7 references): Each of these codes indicate that the user experience was positive. Comments related to these codes included *"So easy, very intuitive!"*, *"Easy to use. Other tools in RDF (for example for data validation, like SHACL) are usually very hard to understand" and "The system is intuitively usable."*. Therefore, a total of 31 occurrences of these codes indicated that the framework provides sufficient usability to end users.

Figure 48 presents the occurrences of the defined themes in the data. The themes were defined as a result of discussion between the author and supervisor of this thesis. It was decided to define themes based on relation to positive and negative patterns discovered as a result of the data analysis. For instance, codes in the *"Positive GUI requirements"* theme indicated sufficient GUI usability, such as an aesthetic interface and clear layout. However, codes in the *"Negative GUI requirements"* theme indicated insufficient GUI usability, such as an unaesthetic interface and unclear layout. Therefore, the themes would provide groupings of codes and provide indications of areas in the framework which fulfill the user's expectations or require improvements.

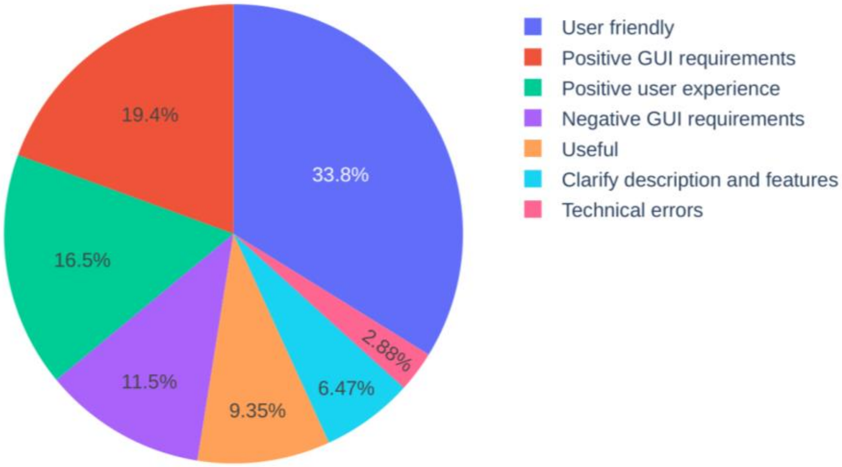


Figure 48: Pie chart of themes in the qualitative data

The themes are discussed as follows:

- **User Friendly** (33.8%), **Positive user experience** (16.5%), **Useful** (9.35%) and **Positive GUI Requirements** (19.4%): These themes accounted for nearly 80% of the codes. Therefore, the analysis indicated participants found the overall experience provided by the framework positive.
- **Positive GUI Requirements** (19.4%) and **Negative GUI Requirements** (11.5%): The analysis indicated that more participants found the GUI sufficient. However, the occurrences of codes related to a negative GUI experience indicated that the interface somewhat needs improvement. Most comments related to negative GUI requirements identified the number of tabs as an issue such as *"new tabs made it a little cluttered"*.
- **Clarify Textual Descriptions and Features** (6.47%): Related codes identified ambiguous information on the framework, however, not a lot of participants encountered problems. The references for the codes did not provide much elaboration on the clarification required. Related comments included *"The descriptions are vague"* and *"Although the system is easy to navigate, there were some ambiguities."*. Therefore, it was difficult to identify which information was ambiguous.
- **Technical Errors** (2.88%)**:** Technical errors occurred infrequently, which indicated most participants did not encounter any. The main errors related to the missing tool tip text, which a few participants encountered.

5.5.6 Recommended Improvements

This section discusses recommended improvements which were discovered through the analysis of the results. In addition, the solutions for the recommendations are mentioned.

- **Verbose Instructions:** 2 references described the instructions as verbose. The related comments are *"Cumbersome instructions page"* and *"Could've made stuff more clear and in single document".* Therefore, indicating the instructions could be simplified. However, since the comments related to the experimental setup, no further action was taken.
- **Tool Tips Not Working:** 4 references mentioned that the tool tips did not reveal information as expected. Interestingly, 4 references described the tool tip information as useful with comments such as *"There are help tool tips which make life easier.".* Therefore, the contradictory comments indicated that the tool tip issue is dependent on the machine used. The issue was resolved by implementing cross platform compatible tool tips.
- **Multiple Tabs:** 3 references mentioned that the number of tabs which opened made it difficult to navigate the framework. Related comments include *"new tabs made it a little cluttered"*. The framework opens each page in a new tab which was hoped to provide easier navigation during the experiment, however, participants found the number of tabs difficult to navigate. Therefore, the frameworks functionality was altered to open new pages in the same tab.

5.5.7 Discussion

This section discusses 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 scores of the understanding questionnaire.

- **Interface Quality:** The student cohort scored the interface quality (2.05) better than the expert cohort (2.17), however, the difference is not statically significant. Both cohorts scored better than the acceptable threshold by more than 14%, therefore, indicating all participants found the interface quality sufficient. The expert cohort could have been more critical of the interface due to previous experience in using advanced semantic web tools.
- **Information Quality:** The student cohort scored the information quality (2.06) better than the expert cohort (2.37), however, the difference is not statically significant. Both cohorts scored better than threshold by more than 25%. The expert cohort could have been more critical of the information as they have a higher level of related background knowledge.
- **System Usefulness and Overall Usability:** No participants in both cohorts required assistance during the completion of the tasks. Participants in both cohorts scored system usefulness and overall usability metrics

the best. The scores of each metric were better than their corresponding thresholds by at least 35%. The results indicated the participants found the system useful with an overall positive user experience.

- **Analysis of each Cohort's PSSUQ Question Scores:** Most median scores of the PSSUQ for the **expert** cohort have a median of 2 or less (18 out of 19 questions) and one had a median of 2.5 (Q9). The best scoring third quartile related to efficiency (Q8) which scored 1.75. The worst scoring question related to error messages (Q9) which scored 4. Most median scores of the PSSUQ for the **student** cohort have a median of 1 (11 out of 19 questions) with the other questions having a median of 2 (8 out of 19 questions). The ease of use (Q1) and completion of tasks (Q3) scored best. The worst scoring question related to error messages (Q9) with a score of 4 for both cohorts. However, the question has been identified as an outlier in the questionnaire as no error messages may be shown to participants.
- **Understanding Questionnaire:** An overall mean score of 88% for participants in both cohorts indicated that they were able to successfully understand the information provided by the framework. The experts mean score (91%) was slightly higher than the students (85%), however, both scores indicated sufficient understanding. In addition, the small difference (0.006) between the standard deviation indicated both cohort's scores were clustered around the mean. Most likely, the missing tool tip text would have impacted the scores of both cohorts as it was required to answer certain questions.
- **Correlations:** No correlations between usability and understanding in both cohorts were statistically significant.

5.5.8 Conclusion

Hypothesis H4 (The framework facilitates the identification of changes in source data and relationships with respective mappings) was supported for participants in both cohorts. The PSSUQ indicated that the usability provided by the framework was sufficient for completing the tasks with both scoring better than acceptable thresholds by at least 14%. The understanding questionnaire which provides evidence that the relationships between source data changes and respective mappings could be understood scored with high numbers of questions correct in both sections. The results of section 1 for both cohorts scored a mean of 87% correct. The results of section 2 for both cohorts scored a mean of 89% correct. Therefore, the mean score of both sections in the questionnaire for both cohorts is 88% correct. The results indicated that participants with varying levels of knowledge were able to understand information related to changes in the source data of respective mappings. In addition, the most common themes (Positive user experience, Positive GUI requirements, User friendly) discovered by thematic analysis identified patterns related to positive overall usability.

Hypothesis H5 (The participants background knowledge influences the successful completion of the tasks) was **not supported** as the satisfaction of the usability which was measured through the PSSUQ indicated that participants in both cohorts had similar levels of satisfaction. The scores of PSSUQ for both cohorts scored similarly better than acceptable thresholds found in research with a mean of 44% better for students and 34% for experts. Furthermore, the results of the understanding questionnaire indicated that the participants in both cohorts could similarly understand the information provided by the framework. The scores of the understanding questionnaire were similar with a difference of 6% between their mean scores. The small difference of 0.04 between the standard deviations indicated that the scores of both cohorts are clustered close to the mean. Moreover, no participants in both cohorts required assistance to complete the experiment. Therefore, it can be concluded that participants with limited knowledge of semantic web technologies can successfully interact with the framework to complete the tasks.

5.6 Validation of Ontology Design

This section discusses the validation of the MQIO and OSCD. A study was conducted in order to validate the design of both ontologies with experts in ontology design. First, an overview of the evaluation structure is presented. Thereafter, the results for each ontology are discussed. The version of the MQIO used during this experiment was **version 1.3**. The version of the OSCD used during this experiment was **version 1.2**. The designs of these versions implemented feedback collected from the previous experiments. **All versions of both ontologies are stored in a folder ("/Ontology-Versions") of the GitHub.**

5.6.1 Hypotheses

The hypotheses related to this study were:

- **Hypothesis H6:** The design of the MQIO conforms to best practices in ontology design.
- **Hypothesis H7:** The design of the OSCD conforms to best practices in ontology design.

5.6.2 Experimental Setup

5.6.2.1 Methodology

A user evaluation was conducted to test the hypotheses of the study. The experiment involved providing participants with documents which detailed the design methodology, implementation, evaluation methods and documentation associated with the MQIO and OSCD. First, participants were asked to review the documents provided with respect to the design methodology and development process of both ontologies. Thereafter, feedback was provided through a questionnaire requesting information on various aspects related to the provided documents. In addition, an open comment section was included in the questionnaire to capture feedback not explicitly requested. The participants were semantic web researchers who are extremely knowledgeable about ontology design practices. In addition, the participants have experience creating and operating mappings in a research environment. Hypothesis **H6** was tested by asking participants to provide feedback after reviewing a document detailing the MQIO. Hypothesis **H7** was tested by asking participants to provide feedback after reviewing a document detailing the OSCD. Ethical approval has been received from the Research Ethics Committee at Trinity College Dublin.

5.6.2.2 Experiment Layout

The structure of the experiment is outlined below.

- **Inclusion/Exclusion Criteria:** Participants satisfied each of the following criteria: 1) Semantic web post-doctoral researcher or Professor 2) Extremely knowledgeable in RDF and R2RML 3) Previous experience in ontology development and 4) Previous experience as mapping engineers.
- **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 (**Appendix Q**). These participants completed the experiment to contribute to the research objectives.
- **Completion of Experiment:** First, participants were required to provide consent to participate in the experiment. The informed consent for this experiment is presented in **Appendix R**. The participants completed the experiment using the provided documents, which included links to the questionnaires.
- **Experiment Assistance**: Assistance was available to participants if they were unable to complete the tasks. The participants were informed that assistance could be provided via email if required.
- **Information Provided:** Two identical documents were provided to the participants. The documents detailed the design process of each ontology, resulting implementation and associated evaluation methods. No additional information was provided. The document is further described in **Section 5.6.2.4**.

5.6.2.3 Sample Size

The sample size of the experiment is outlined below.

- **Background:** Participants are semantic web researchers who are very knowledgeable with semantic web technologies, ontology design practices and mapping creation. The participants have experience in developing ontologies in a research environment.
- **Number of Participants:** The study involved 5 expert participants.

5.6.2.4 Experiment Structure

Figure 49 presents an overview of the experiment structure.

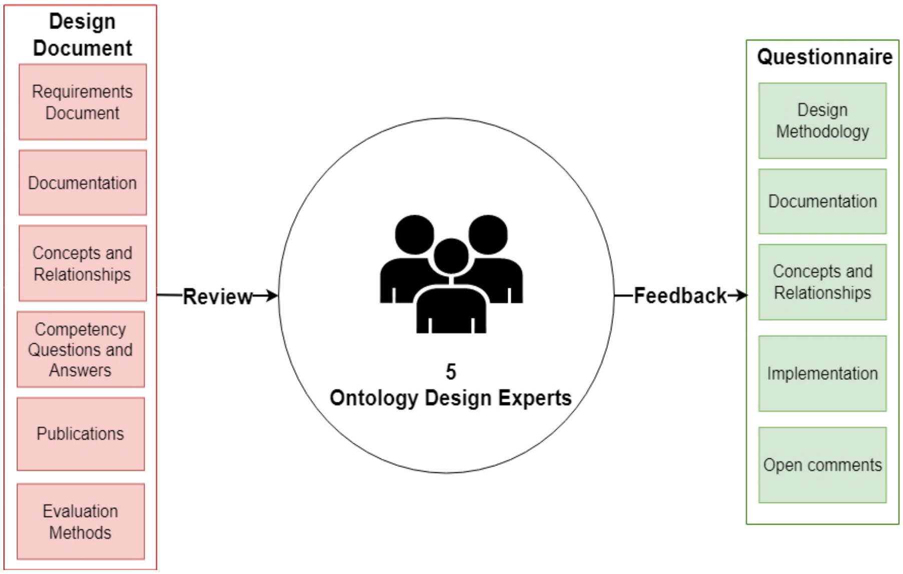


Figure 49: Overview of experiment structure for ontology design evaluation

Documents detailing design, implementation and access to online documentation for both ontologies were provided to 5 ontology design experts. The online documentation for each ontology includes a listing of concepts and relationships, descriptions, guidance on usage, sample SPARQL queries, sample graphs and the implementation in various serializations. After reviewing the document, they were asked to provide feedback through a questionnaire. The questionnaire was described in **Section 5.2.4.3** and requested descriptive feedback on aspects, such as design methodology, conformance to ontology design best practices [53], documentation, concepts and relationships, implementation and evaluation methods. In addition, an open comment section was included to capture feedback not covered by the other questions.

5.6.3 Experiment 4 and 5: Validation of MQIO and OSCD Design

This section discusses the results of both ontologies. The document provided to participants which described the development of the MQIO is presented in **Appendix S** and the OSCD is presented in **Appendix T**. The design methodology of MQIO is described in **Section 4.2.1.1** and OSCD is described in **Section 4.3.1.1**. The results of the expert's feedback for each question posed are discussed in the following **subsections**.

5.6.3.1 Results on Design Practices of MQIO and OSCD

The following question was posed in relation to design practices: ***"1a) In your opinion, does the design of MQIO/OSCD correctly follow best practices in ontology design? 1b) Can you provide a reason for your response?".*** **Table 58** presents the responses to this question from each ("**#**") of the 5 expert participants for the MQIO. Part a) of the question provides an initial response ("***Response***"). This initial response can be positive ("Yes") or negative ("No"). Part b) of the question provides a further elaboration ("***Rationale for Response***") on the initial response. **Table 59** presents the response for OSCD to this question.

Table 58: Participants responses to design practices of **MQIO**

#	***Response***	***Rationale for Response***
1	Yes	*"It is explained and declared that they followed the best practices."*
2	Yes	*"the ontology seems to be following the best practices in methodologies relevant to the ontology space."*
3	Yes	*"It follows well established methods for developing and iteratively improving the concepts based on application."*
4	Yes	*"The documentation is comprehensive and detailed and the most relevant methodologies in the state of the art have been considered and reused."*
5	Yes	*"Use of both methodologies and state of the art ontology validation tools"*

Table 59: Participants responses to design practices of **OSCD**

#	***Response***	***Rationale for Response***
1	Yes	*"It is declared and explained that the methodologies in the best practice documents and books are applied.".*
2	Yes	*"OSCD is using well known recommendations in the ontology building space".*
3	Yes	*"It follows well established methods for developing and iteratively improving the concepts based on application."*
4	Yes	*"the documentation is comprehensive and detailed and the most relevant methodologies in the state of the art have been considered and reused."*
5	Yes	*"In addition to using common methodologies it incorporates state of the art practical tools like OOPs and Widoco which impose additional requirements on ontologies."*

As can be seen, all of the participants provided a positive answer to this question, which indicated that the MQIO and the OSCD were designed in conformance with ontology design best practices in the state of the art. In addition, the rationales for each response supported these findings.

5.6.3.2 Results on Design Methodology of MQIO and OSCD

The following question was posed in relation to design practices: ***"2a) Do you suggest any alterations to the design methodology followed by MQIO/OSCD? 2b) Can you provide a reason for your response?".*** **Table 60** presents the responses to this question from the 5 expert participants for the MQIO. **Table 61** presents the response for OSCD to this question.

Table 60: Participants responses to design methodology of **MQIO**

#	***Response***	***Rationale for Response***
1	No	None
2	No	*"sound methodology"*
3	No	*"I think the current iteration is sufficient for the task."*
4	No	*"it looks all complete, sound and comprehensive"*
5	Yes	*"Apply FAIR checker software to the ontology metadata, check against Gruber design principles"*

Table 61: Participants responses to design methodology of **OSCD**

#	*Response*	*Rationale for Response*
1	No	None
2	No	*"methodology is sound"*
3	No	*"I think the current iteration is sufficient for the task."*
4	No	*"it looks all complete, sound and comprehensive"*
5	Yes	*"Run the ontology through a FAIR metadata checker, test if the ontology meets Gruber's design principles https://doi.org/10.1006/ijhc.1995.1081"*

As can be seen, 4 out of the 5 participants provided a negative response to this question for both ontologies, which indicated the design methodology followed by MQIO (See Section 4.2.1.1) and OSCD (See Section 4.3.1.1) are sufficient. In addition, the rationales provided for each response supported these findings. Participant #5 recommended completing two assessments on both ontologies, which included inputting them into a FAIR metadata validator and assessing conformance to the Grubers design principles. The FAIR data principles [147] were published in 2016 and were proposed by a consortium of scientists and organizations. These principles are designed to guide data publishers in supporting the reusability of published data assets. Therefore, it was decided to input both ontologies into an online FAIR metadata validator[60]. The results (**Appendix U**) indicated that MQIO and OSCD conformed to FAIR metadata principles with all concepts and relationships registered in the registries used during validation. The Grubers design principles (proposed in 1995) [59] provide comprehensive objective criteria's designed to guide the development of ontologies. The five principles are 1) Clarity 2) Coherence 3) Extendibility 4) Minimal encoding bias and 5) Minimal ontological commitment, which encapsulate key criteria's an ontology should satisfy. These principles are described in **Appendix V**. In addition, the conformance of the MQIO and OSCD to these principles are outlined. The final iteration of the design methodology followed by both ontologies conformed to the expert's guidance after these tasks were included in the methodology.

5.6.3.3 Results on Concepts and Relationships of MQIO and OSCD

The following question was posed in relation to design practices: ***"3a) Do you suggest any alterations to the concepts/relationships in MQIO/OSCD? 3b) Can you provide a reason for your response?"***. **Table 62** presents the responses from the 5 expert participants.

Table 62: Participants responses to concepts and relationships of **MQIO**

#	*Response*	*Rationale for Response*
1	Yes	*"I didn't understand the MappingRefinement relation with Agent. The relation is not on the ontology or floating representation does not show it as well. So in general I would check the*

[60] https://fair-checker.france-bioinformatique.fr/inspect

		relations and concepts there."
2	Yes	*"I'd suggest that SHACL is considered instead of RDF constructs to define and capture violations. SHACL is being widely used in industry for this aspect."*
3	Yes	*"Some relationships are probable to have more information associated with the context. In such cases, the target of these should not be a data property, but instead should be a concept that can have more information associated with it. For example, used Query relates to a Query which may have an author or a version, and requirements satisfaction could be variances in states (e.g. none, somewhat, almost, fully), or used Tool which can be associated with more information for the tool."*
4	No	*"MQIO looks comprehensive and fit for purpose"*
5	Yes	*"Philosophically I view a set of mappings as a dataset. We already have 2 vocabularies for expressing the relationships between a dataset and its quality assessment metadata (DQV, DAQ) and I am not 100% convinced we need another specialised vocabulary that focuses on mapping datasets. It is true there are mapping specific metadata captured here but again perhaps this could simply reuse other vocabs or be defined alone (ie without incorporating the quality metadata)."*

Four of the participants provided recommendations for the concepts and relationships of the MQIO. Therefore, the following modifications were completed on the ontology in order to address these recommendations.

- *(Participant #1)* The description (`rdfs:comment`) of the property (`mqio:wasCreatedBy`), which relates agents to refinements was further clarified with additional text.
- *(Participant #2)* Several sample SHACL constraints were added to the documentation. These constraints can be applied to quality report graphs expressed in MQIO in order to validate certain aspects, such as whether each detected quality issue has a related refinement, thus providing indications of the number of issues which have been resolved.
- *(Participant #3)* It was decided to not change the type of the property (`mqio:usedQuery`) used to represent queries as these are linked to a refinement (`mqio:MappingRefinement`) resource where additional metadata can be added. However, the property representing mapping tools (`mqio:usedTool`) was changed from a data type property (`owl:DatatypeProperty`) to an object property (`owl:ObjectProperty`). In addition, a new concept (`mqio:MappingTool`) was added, which relates relevant tools used during the construction of a mapping and is related using the `mqio:usedTool` property. Therefore, changed data is identified as a resource rather than a literal, which allows additional information to be added.
- *(Participant #5)* The rationale for reusing certain concepts and relationships in DQV [2] and DAQ [37] rather than reusing them in their entirety was outlined in **Section 4.2.1.1**. DQV was reused to represent quality metrics, measurements, dimensions and categories, however, it does not provide suitable concepts to represent activities involved in mapping creation, quality assessment and refinement. Therefore, it was decided to reuse PROV-O [84] similar to DQV and DAQ, which represent activities and entities using PROV-O. Therefore, it was decided to represent the activities involved in the definition and quality improvement of mappings using concepts and relationships in PROV-O. The rationale for the design decision was not

stated in the documentation of MQIO, which was provided to the participants. Therefore, it was decided to add the rationale into the description section of the documentation in order to clarify the design decision.

Participant #4 did not provide recommendations and indicated the current concepts and relationships of the MQIO are sufficient. **Table 63** presents the responses received to this question for OSCD.

Table 63: Participants responses to concepts and relationships of **OSCD**

#	*Response*	*Rationale for Response*
1	Yes	*"It's a minor issue however I would use foaf:Agent instead of dul:agent. You're not describing the person in detail why would you use an unknown or less known ontology which might not be in use in the future?."*
2	Yes	*"hasChangeData is losing the data semantics when storing as a string. In semantic applications, it is important that the data's semantics are not lost (be it a datatype e.g integer) or a URI."*
3	Yes	*"Some relationships are probable to have more information associated with the context. In such cases, the target of these should not be a data property, but instead should be a concept that can have more information associated with it. For example, data change being represented by a string, or by a concept called ChangedData or similar."*
4	No	*"OSCD looks comprehensive and fit for purpose"*
5	No	*"The competency question answers seem appropriate"*

Three of the participants provided recommendations for the concepts and relationships of the OSCD. Therefore, the following modifications were completed on OSCD in order to address these recommendations.

- *(Participant #1)* The agents in the OSCD were changed from a resource type of `dul:Agent` to `foaf:Agent`, which involved updating the respective definition (`rdf:type`) in the ontology.
- *(Participant #2 and #3)* The `oscd:hasChangedData` property was changed from a data type property (`owl:DatatypeProperty`) to an object property (`owl:ObjectProperty`). Therefore, changed data is identified as a resource rather than a literal, which allows additional information to be added.

Participant #2 recommended adding SHACL constraints to the documentation of the MQIO in order to validate relevant information in graphs expressed using the ontology. The participant did not make the same recommendation for the OSCD. However, it was decided to add SHACL constraints to the documentation of the OSCD in order to allow users to easily validate their respective graphs. The constraints were designed to validate aspects in relevant graphs detailing change reports and notification policies. For instance, a constraint was added to validate that relevant graphs contained a previous and current version of source data. In addition, a constraint was added to validate that a maintainer has been defined for the source data. Two participants (#4 and #5) did not provide recommendations and indicated the current concepts and relationships of the ontology are sufficient.

5.6.3.4 Results on Documentation of MQIO and OSCD

The following question was posed in relation to design practices: ***"4a) Do you suggest any alterations to the documentation of MQIO/OSCD? 4b) Can you provide a reason for your response?".*** **Table 64** presents the responses to this question from the 5 expert participants.

Table 64: Participants responses to documentation of **MQIO**

#	*Response*	*Rationale for Response*
1	Yes	*"The relation is not on the ontology or floating representation does not show it as well. So in general I would check the relations and concepts there."*
2	No	*"documentation is well written"*
3	No	*"I think the current iteration is sufficient for the task."*
4	No	*"very good documentation"*
5	Yes	*"Add an appendix with sample instance data and use case"*

Two participants (#1 and #5) provided a recommendation for the documentation of MQIO. Therefore, the following modifications were completed on MQIO in order to address these recommendations.

- *(Participant #1)* The diagram of the ontology in the documentation was updated in order to resolve unconnected concepts and relationships.
- (*Participant #5*) Hyperlinks to graphs expressed in the MQIO containing sample instance data and use case data were included in the documentation rather than including the RDF representation itself. Therefore, it was decided to add an Appendix section (See **Appendix W**) to the documentation, which includes extracts of these RDF graphs in order to the address the recommendation.

Three participants (#2, #3 and #4) did not provide recommendations and indicated that the documentation of MQIO was sufficient quality by providing positive comments to the question. **Table 65** presents the responses received to this question for OSCD.

Table 65: Participants responses to documentation of **OSCD**

#	*Response*	*Rationale for Response*
1	Yes	*"It would be nice to see more type of examples such as MoveSourceData".*
2	Yes	*"The description states that this ontology is for identifying changes in an RDF datasets, however, the example shown is related to CSV changes. Whilst this might be used for both, it would be important to show how RDF changes are captured with this ontology, as it this topic is a well known problem in both academia and industry"*
3	No	*"I think the current iteration is sufficient for the task."*
4	No	*"very good documentation"*
5	Yes	*"It would be good to have an appendix with an example of use (instances and a use case)"*

Three of the participants provided recommendations for the documentation of the OSCD. The following modifications were completed on OSCD in order to address these recommendations.

- *(Participant #1)* Two additional examples detailing move changes (`oscd:MoveSourceData)` concept were added to the documentation.
- *(Participant #2)* The description of the ontology in the documentation was clarified in order to ensure readers understood that it is designed to represent only source data changes and not changes in RDF datasets.
- *(Participant #5)* Similar to the MQIO, an Appendix section was added to the documentation of the ontology. Extracts of graphs expressed in the OSCD which were linked in the previous version of the documentation were included in the section.

These results indicated that three modifications should be completed on OSCD, which related to the addition of additional examples to the documentation. Two participants (#3 and #4) did not provide recommendations and indicated the current documentation of the ontology was sufficient.

5.6.3.5 Results on Open Comments of MQIO and OSCD

The following question was posed in relation to design practices: ***"5) Any additional comments?"***. **Table 66** presents the open comments provided by the 5 expert participants related to MQIO.

Table 66: Participants open comments related to **MQIO**

#	***Open Comments***
1	None
2	None
3	None
4	*"well done!"*
5	*"no"*

None of the participants provided recommendations through open comments, which indicated sufficient quality from MQIO apart from the aforementioned recommendations. Furthermore, the *"well done!"* comment supports these findings. **Table 67** presents the responses received to this question for OSCD.

Table 67: Participants open comments related to **OSCD**

#	***Open Comments***
1	None
2	None
3	None
4	*"well done!"*
5	*"Nice and lean. I would prefer a link to Prov:Agent instead of DUL:Agent but I suppose you inherit this from LODE and I acknowledge it is a hard choice."*

Only one of the participants provided recommendations through open comments, which indicated sufficient quality from OSCD apart from the aforementioned recommendations. Similar to the MQIO, the *"well done!"* comment supported these findings. The recommendation from Participant #5 was addressed in the previous section (Section 5.6.3.3) by changing the agent class to type `foaf:Agent`, which is a more prominent concept used to represent agents. However, the participant provided a positive comment related to the ontology, stating it was *"Nice and lean."*, which indicated positive characteristics related to the design.

5.6.4 Summary of Results

All participants indicated that the design practices followed during the development of the MQIO (**Table 58**) and the OSCD (**Table 59**) conformed to best practices in the state of the art. In addition, positive comments, such as *"It is explained and declared that they followed the best practices."* and *"the ontology seems to be following the best practices in methodologies relevant to the ontology space."* supported these findings. 4 out of the 5 participants did not recommend any alterations to the design methodology followed by the MQIO (**Table 60**) and the OSCD (**Table 61**), which indicated overall that the followed methodology was sufficient. In addition, positive comments, such as *"sound methodology"* and *"it looks all complete, sound and comprehensive"* supported these findings. However, one participant (#5) recommended including validation of FAIR metadata principles [147] and Grubers ontology design principles [59] into the methodology of both ontologies The final iteration of the design methodology of both ontologies included these two assessments as previously described. The results of these assessments validated successful conformance to them. Therefore, it can be concluded that the final design process of the MQIO and the OSCD conformed to the level of quality as recommended by these ontology design experts.

While the expert feedback provided evidence which supported the design process, minimal improvements were recommended. The improvements which were recommended by participants for **both ontologies** are outlined below.

- Sample SHACL constraints (See **Figure 54**) were added to the documentation of both ontologies. Constraints added to the MQIO were designed to validate aspects related to the quality reports. For instance, a constraint added can be used to validate that each quality violation in a mapping is associated with a refinement. Constraints added to the OSCD were designed to validate aspects related to change reports and notification policies. For instance, constraints were added to validate that the reports contained a previous and current source data and validate the policies include a threshold for each change type.
- An Appendix section (See **Appendix W**) was added to the documentation. The section contains sample instance data and use case graphs expressed in the respective ontology which were previously referenced using hyperlinks.

In total 3 modifications were individually recommended for the MQIO. These recommendations related to the concepts/relationships (**Table 62**) and documentation (**Table 64**) of the **MQIO**, which are outlined below.

- The description (`rdfs:comment`) of a property (`mqio:wasCreatedBy`) in the implementation of the ontology was updated.

- The diagram of the ontology in the documentation was updated in order to resolve unconnected concepts.
- The mqio:usedTool property was changed from a data type property (owl:DatatypeProperty) to an object property (owl:ObjectProperty).

The lack of recommendations through the open comments (**Table 66**) indicated no further modifications were required according to these experts. Therefore, it was concluded that the final version of MQIO, which resulted from addressing the participants recommendations, provides a sufficient model to represent mapping quality information. The final version of MQIO, which resulted from implementing the feedback from participants in this experiment was **version 1.4**. This version of the ontology is available online (https://w3id.org/MQIO).

In total 4 modifications were individually recommended for the OSCD. These recommendations related to the concepts/relationships (**Table 63**) and documentation (**Table 65)** of the **OSCD**, which are outlined below.

- Additional examples of the move changes (oscd:MoveSourceData) concept were added to the documentation (See **Figure 54**) by initially creating sample RDF data in TURTLE format and inserting it into the documentation and republishing it.
- The foaf:Agent concept replaced the current concept (dul:Agent) used to represent agents in the ontology, which have similar meanings. It was recommended to use a concept from the FOAF ontology as it is more prominent than the DOLCE+DnS Ultralite (DUL) ontology [25] .
- Additional text was added to the documentation of OSCD in order to clarify that the ontology is designed for representing source data changes and not changes in the resulting LD datasets as well.
- The oscd:hasChangedData property was changed from a data type property (owl:DatatypeProperty) to an object property (owl:ObjectProperty) to allow linking of relevant information with resources representing changes.

The lack of recommendations through the open comments (**Table 67**) from 4 out of the 5 participants indicated no further modifications were required according to these experts. However, one participant (#5) provided an open comment stating that the ontology was *"Nice and lean."* and recommended the following, *"I would prefer a link to Prov:Agent instead of DUL:Agent but I suppose you inherit this from LODE and I acknowledge it is a hard choice."*. The recommendation was addressed by using a more prominent concept (foaf:Agent) in order to represent agents as previously stated. Therefore, it can be concluded that the final version of OSCD, which resulted from addressing the participants recommendations, provides a sufficient model to represent changes detected in the source data of mappings. The final version of the OSCD, which resulted from implementing the feedback from participants in this experiment was **version 1.3**. This version of the ontology is available online (https://w3id.org/OSCD).

5.6.5 Discussion

The results indicated that the development process of MQIO and OSCD followed best practices as recommended in the state of the art, as no participant provided a negative response to the respective question (Q1). The positive comments provided by participants, which related to the question, such as *"It follows well established methods for developing and iteratively improving the concepts based on application."* and *"It is explained and declared that they followed the best practices."* supported these findings. The results indicated that the design methodologies followed during the development of both ontologies were overall sufficient as only one of the participants (#5) recommended alterations to the methodology. In addition, positive comments, such as *"sound methodology"* and *"it looks all complete, sound and comprehensive"* supported these findings. The alterations recommended by a participant to the design methodologies were minimal and involved including conformance validation to FAIR metadata principles and Grubers ontology design principles as previously discussed. Similar results were anticipated for both ontologies as they followed a similar methodology, which was strongly inspired by prominent approaches (See Section 4.2.1.1 and 4.3.1) in the state of the art. Therefore, it was concluded that the design process of both ontologies was sufficient and conformed to best practices in the state of the art.

The participants provided useful insights which were used to refine the current version of both ontologies. Both ontologies required similar modification, which mainly related to changing a small number of properties from data type to object, allowing additional information to be related to the values. For instance, a value changed in a source data may include further background information on the change. In addition, additional examples of graphs expressed using the ontologies were added to the documentation to provide useful demonstrative information to users. It was concluded that the final iteration of both ontologies conformed to the level of quality as recommended by these 5 ontology design experts as all recommendations were successfully addressed, which provided strong indications of sufficient overall quality.

5.6.6 Conclusion

Hypothesis H6 (The design of the MQIO conforms to best practices in ontology design) was **supported** as all participants provided positive responses and comments related to questions about the design practices. Comments from experts such as *"It follows well established methods for developing and iteratively improving the concepts based on application."* and *"It is explained and declared that they followed the best practices."* supported these findings. Minimal recommendations were provided by experts, related to the design methodology, implementation and documentation, which were all addressed in order to include the views of the ontology design experts. Therefore, the final iteration of the MQIO provides a sufficient ontology in the opinion of the experts to represent quality information related to LD mappings.

Hypothesis H7 (The design of the OSCD conforms to best practices in ontology design) was **supported** as all participants provided positive responses and comments related to questions about the design practices. Comments from experts such as *"OSCD is using well known recommendations in the ontology building space"* and *"the documentation is comprehensive and detailed and the most relevant methodologies in the state of the art have been considered and reused."* supported these findings. Minimal recommendations were provided by experts, related to the design methodology, implementation and documentation, which were all addressed in order to further consolidate the views of the ontology design experts. Therefore, the final iteration of the OSCD provides a sufficient ontology in the opinion of the experts to represent changes in the source data of LD mappings.

5.7 Application Study

This section discusses the application of the MQI framework in two real world use cases which included research projects involving Ordnance Survey Ireland (OSi)[61] and Ericsson software technology[62]. **Supplementary information related to the study is stored in a folder ("/Application-Study") of the GitHub.**

5.7.1 Use Case: Ordnance Survey Ireland (OSi)

The mapping quality assessment and refinement component of the MQI framework was applied to a use case was for a project named data.geohive.ie [90] which involved the national geospatial agency of Ireland OSi. The data.geohive.ie project involved creating uplift mappings which transformed some of Ireland's national geospatial data into RDF representation and making it available as a publicly available LD endpoint. These uplift mappings were created independently of the author of this thesis. Supplementary information related to the use case is stored in a folder ("/OSi-Use-Case") of the GitHub. **Figure 50** presents an overview of the interaction between the data.geohive.ie project and the MQI framework. As seen in **Figure 50**, the MQI framework was used to assess and refine the quality of each R2RML mapping used to uplift from the OSi PRIME2 geospatial database into the LD endpoint.

[61] https://osi.ie/

[62] https://www.est.tech/

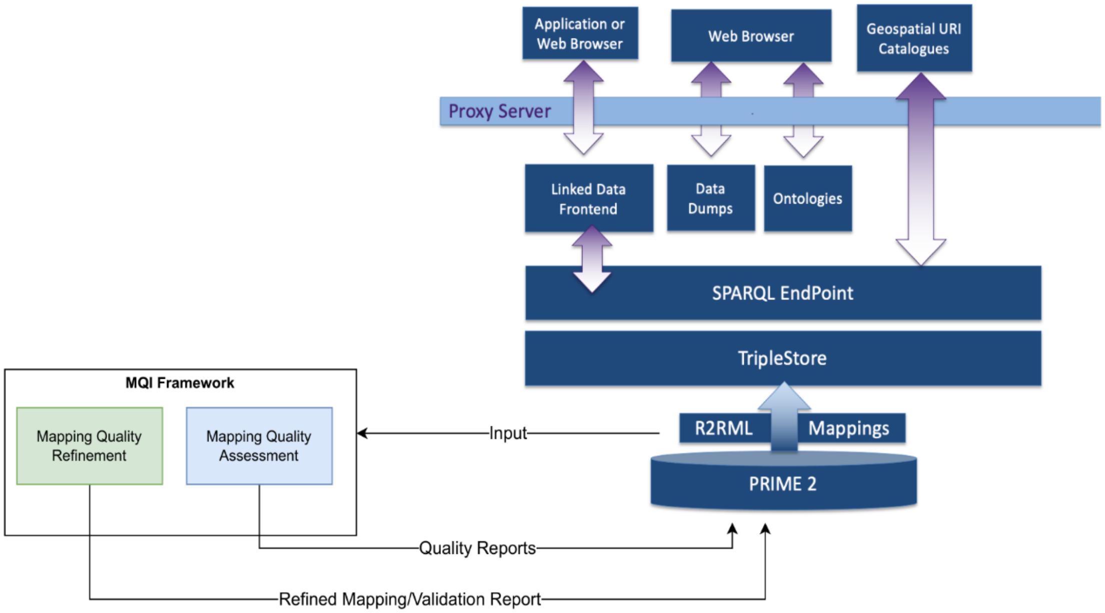


Figure 50: Interaction between MQI framework and OSi architecture [91]

The mapping quality assessment information is described in the results of experiment 1 (See **Section 5.3.3**). Four quality violations were assessed and refined in the OSi mapping (#25).

- **Usage of undefined classes (D1):** `daq:MetricProfile` is undefined in the Dataset Quality Vocabulary (daQ) [37]. The issue was resolved by using a semi-automatic refinement ("Find Similar Classes") which suggests classes from the same namespace as the undefined class. `daq:MetricProfile` was replaced with one of the suggested classes, `daq:Metric`, therefore, resolving the issue.
- **Usage of undefined properties (D2):** `daq:totalDatasetTriplesAssessed` is undefined in daQ. The issue was resolved using a semi-automatic refinement ("Find Similar Predicates") which suggests properties from the same namespace as the undefined property. `daq:totalDatasetTriplesAssessed` was replaced with `daq:value`, therefore, resolving the issue.
- **Usage of incorrect domain (D3):** The domain of the `prov:generated` is defined in PROV-O [84] as `prov:Activity`, however, the class defined in the mapping is `daq:Observation` which is a subclass of `prov:Entity` and results in an incorrect domain definition. While the framework identified the correct domain is `prov:Activity`, it did not suggest inserting it as the class is disjoint with `prov:Entity`, therefore, it would result in an additional inconsistency. Nonetheless, the mapping engineer guided with the information decided to add an additional triple map with the property and `prov:Activity` class as the domain.
- **Usage of incorrect range (D6):** `sdmx-dimension:timerPeriod` is a property used in the mapping to define when an observation was measured by the PRIME2 database. The range of this property in the

SDXM ontology[63] is `rdfs:Resource`, therefore, the mapping should define an object for term maps in the mapping with the property. However, the range in the mapping was defined as a literal, which was identified through the term type definition. The term type for the term map used to generate respective triples was `rr:Literal`[64] . Therefore, literal values will be generated as the range in the triples of the resulting LD dataset. The issue was resolved by using a semi-automatic refinement ("Add Correct Term Type") which suggests the correct term type(s) for the range defined in the ontology. The framework for this violation suggests two term types for IRIs (`rr:IRI`) and Blank nodes (`rr:BlankNode`). The IRI term type (`rr:IRI`) was selected and the term type (`rr:termType`) updated, resulting in the resolution of the violation as the refined mapping generates resources for the range.

In addition, metrics which measured the quality of ontologies used in the mappings detected 10 violations. These violations are outlined below:

- **RDF**: This ontology was used to represent properties in order to define measurements in the mappings, however, it does not contain human-readable and machine-readable licenses.
- **DaQ**: This ontology was used to represent data quality metrics, dimensions and categories, however, it does not contain human-readable and machine-readable licenses.
- **SDXM**: This ontology was used to represent time periods related to measurements, however, it does not contain basic provenance, human-readable and machine-readable licenses.
- **RDFS**: This ontology was used to represent metadata related to measurements, such as human-readable labels and comments, however, it does not contain human-readable and machine-readable licenses.

It is important to provide transparent mapping quality information, which allows mapping engineers to consider the reuse of existing ontologies for their particular use case. For instance, the lack of licensing could present an issue if the published dataset is used for commercial purposes without explicit permission provided by ontology maintainers for reuse. Semi-automatic refinements (See **Section 4.2.2.2**) were used to resolve three violations and facilitated the resolution of the fourth by informing the mapping engineer of the violation and correct domain class. In addition, the framework informed the engineer of the aforementioned quality issues related to the reused ontologies, which allowed them to consider using different terms. 1725 triples in the resulting graph were positively impacted by the refinement of these violations. The use case demonstrated the applicability of the framework in mappings transforming geospatial data. Details of the use case have been published [115].

[63] https://lov.linkeddata.es/dataset/lov/vocabs/sdmx

[64] https://www.w3.org/TR/r2rml/#termtype

5.7.2 Use Case: Ericsson

Both components of the MQI framework were applied to a cloud native monitoring use case in Ericsson. In this case, the author of this thesis was involved in developing and refining the uplift mappings used in the project. The use case involved uplifting monitoring information collected by metrics used in the Prometheus [155] cloud native monitoring system. The data was represented in an ontology named the Intent Based Closed Loop Ontology (IBCLO) [118] designed to automatically define and valid intent in a control loop. Details related to the Ericsson architecture and the associated IBCLO, which were involved in this use case have been published [119]. Supplementary information related to the use case is stored in a folder ("/Ericsson-Use-Case") of the GitHub. **Figure 51** presents an overview of the interaction between the framework and the Ericsson architecture.

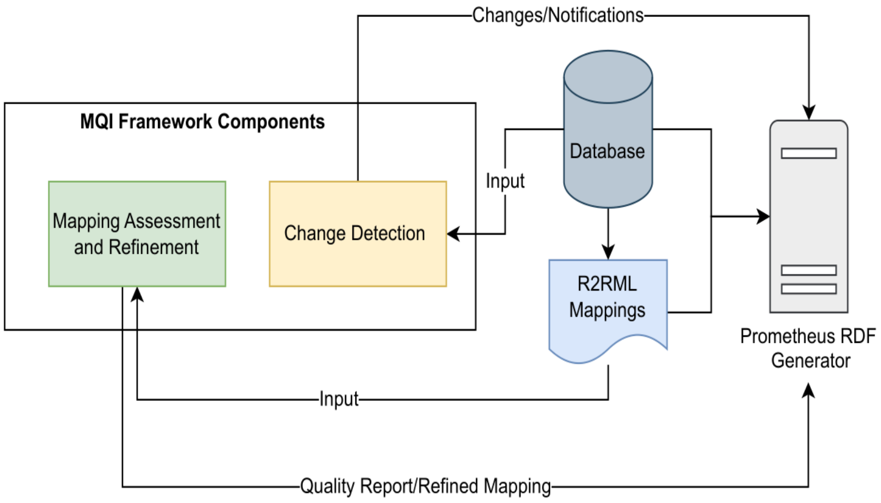


Figure 51: Interaction between MQI framework and Ericsson architecture

The Prometheus RDF Generator framework captures time series metric data from Prometheus and uplifts it into RDF format expressed in the IBCDLO which was under development at the time. The application of the framework aided two main aspects of the use case: 1) Assessment and refinement of uplift mappings created and 2) Preservation of mapping alignment with the local database. These two aspects are outlined below:

Mapping Assessment and Refinement: The mappings ("/Mappings") were input into the mapping assessment and refinement component before execution to capture and remove quality violations. Each violation was resolved by the semi-automatic refinements or mapping engineer guided by the assessment information provided. Quality requirements were defined on each mapping that ensured no mapping was executed which contained detectable issues. The resulting validation report expressed in MQIO was linked with the resulting dataset ("link_reports.rq") to provide evidence of the suitability of the mapping for execution. However, not all violations detected by the framework originated in the mapping rather the definitions of IBCLO which were incomplete leading to conflicts.

The most common quality issues captured by the framework during the use case related to metrics in the data quality aspect (See **Section 4.2.2.1.3**). Several instances of quality issues detected from applying the mapping quality assessment and refinement component as a result of the use case are outlined below.

- **Usage of incorrect range (D6):** `ibclo:hasMetricQuery` is a property used to represent the query, which was used to retrieve the metric data from Prometheus. This property is an object property, therefore, requiring a resource as the range. However, the initial mapping defined the range as a literal. A semi-automatic refinement ("Add Correct Term Type") was used to resolve the issue in the mapping by changing the range from a literal (`rr:Literal`) to a resource (`rr:IRI`). However, it was later discovered when the mapping was executed that the property was incorrectly defined in the IBCLO. Therefore, the framework helped to identify conflicts between mapping definitions and respective ontologies, which are under development.
- **Usage of incorrect domain (D3):** The `ibclo:hasResultStatus` is a property used to represent the status returned from the Prometheus server when a query is executed. The status returned can be a valid result ("True") or not valid ("False"). The domain of this property is `ibclo:MetricQuery`, however, a mapping incorrectly defined the domain as `ibclo:MetricResult`, therefore, leading to inconsistencies. A semi-automatic refinement ("Add Domain Class") was used to resolve the issues in the mapping by retrieving and updating the class definition in the mapping with the correct domain class.
- **Domain and range definitions (VOC2):** The properties in the IBCLO should contain domain (`rdfs:domain`) and range (`rdfs:range`) definitions, which inform users on the correct usage of the properties. However, the framework identified that the ontology did not contain a range definition for the `ibclo:effectsMetric`, which is used to represent relationships between actions and metrics. A semi-automatic refinement ("Add Range") was used to add the correct range definition (`ibclo:Metric`) to the ontology.

Quality issues detected during the application are detailed in quality reports ("quality_report.ttl") represented in the MQIO in the folder of the GitHub. The final quality reports expressed in MQIO were stored to improve trustworthiness for consumers.

Source Change Detection: The database was connected to the source change detection component and periodically queried using SQL queries which retrieved relevant information and uplifted into RDF format expressed in the OSCD. A notification policy was defined to inform when 10 insert changes were detected to provide a high level of timeliness. Information related to links between the mappings previously uploaded and changes were examined once a week to identify the potential impacts on them. The examination ensured that new target servers or metrics were captured by the resulting RDF graphs ("/Change-Logs").

Key insights from applying the source change detection component during the use case are outlined below.

- **Occurrences of Changes:** The most common changes detected during the application were insert and delete changes. Other types of changes such as move, merge and datatype rarely occurred. One move change was detected when the source data was relocated to another table in the database as result of modifications to the schema. These results would be expected as previous work addressing the dynamics of LD (See Section 2.2.6) has observed similar change frequencies.
- **Notifications are Helpful:** The notification policy was triggered several times during the application, which proved useful for identifying changes in a timely manner. Notifications were sent using email and the layout contained a listing of changes and their respective count and threshold, which were retrieved using the SPARQL query presented in **Listing 12**. However, the email did not contain the RDF representation of the change report expressed in the OSCD. Retrieving the report required accessing the framework using the GUI and identifying/pressing the respective download button. Therefore, it was decided to include the RDF graph file as an attachment in the email in order to streamline the process. An example email address notification was shown in the video demonstration of the framework[65] .
- **Detection is Performance Intensive:** Capturing large numbers of changes, uplifting them into RDF representation, executing multiple SPARQL queries ("/Chapter-4/Second-Iteration-Implementation") and rendering the results at once, when initiating the change detection process was discovered to be performance intensive. The processing time observed was longer than expected. Therefore, it was decided to incorporate threading into this component using the multiprocessing library[66] in Python. Threading allows tasks to be run in concurrently rather than consecutively. Threading enabled the component to execute each SPARQL query in parallel and join the results when all threads are finished executing. The processing time observed after implementation of threading was faster, which is hoped to improve usability for respective end users when detecting changes in large data sources.

The final iteration of the mappings is available ("refined_mapping.ttl"). Details of the use case have been published in [118].

5.7.3 Discussion

The application of the framework to OSi provided evidence that it could facilitate the assessment and refinement of real-world mappings with 3 out of 4 violations removed using semi-automatic refinements. The fourth violation was referred to the mapping engineer who resolved it using the assessment information provided by the framework which identified the correct domain. Interestingly, applying the framework to the Ericsson use case

[65] https://drive.google.com/file/d/14-AQloiL9WnQ1iUwzsqYk6cQaKo0u_IZ
[66] https://pypi.org/project/multiprocess/

identified a new possible usage of the mapping quality assessment and refinement component. The mappings expressed in the IBCLO were created in parallel to the development of the ontology. Therefore, the framework not only aided in identifying mapping quality issues, in addition, conflicts between the intended mapped data and ontology were discovered. For instance, the data mapped with the `ibclo:hasMetricQuery` included an literal range, however, the ontology defined it as an object property. The issue was identified by the framework when the mapping was assessed and was resolved by updating the property definition in IBCLO.

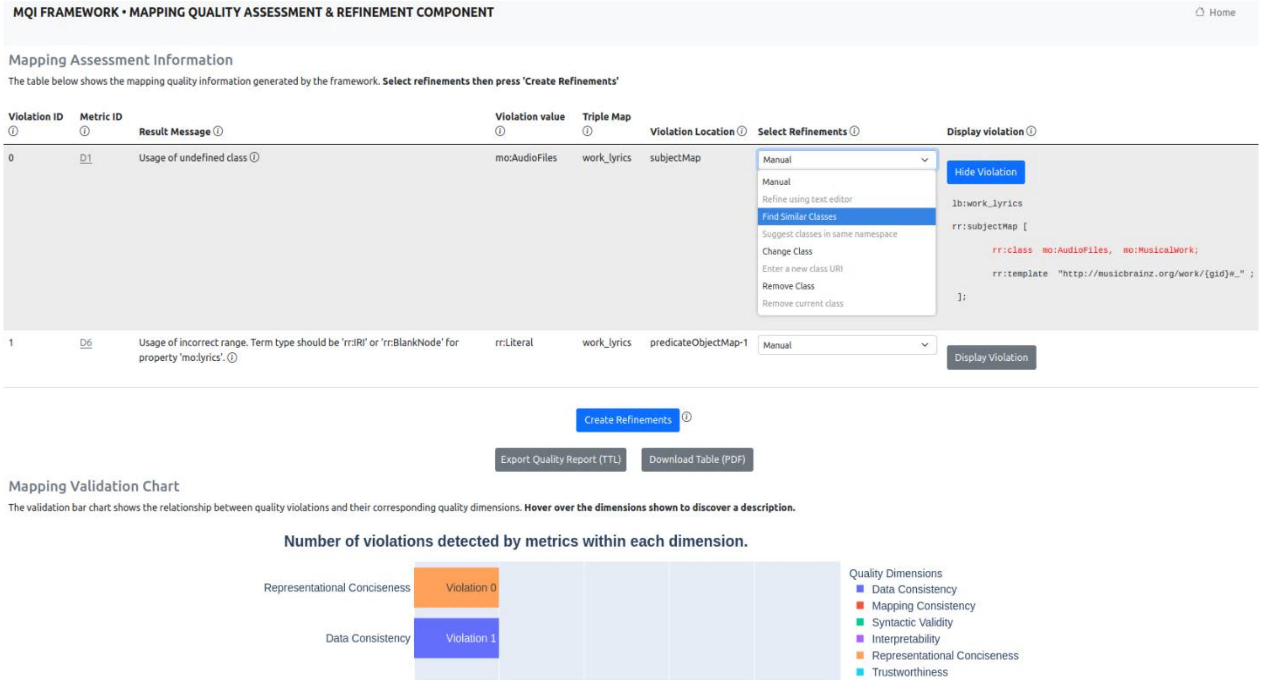
The application of the change detection proved beneficial as the network metric data being mapping were diverse and originated in various systems which were constantly evolving, and tracking changes enabled a high level of alignment. The identification of limitations related to performance and notifications provided useful insights which were used to inform the final version of the framework. The use cases consolidated the design by demonstrating the applicability of the framework in real world. In addition, demonstrating the framework on mappings which uplifted different knowledge including geospatial and network metrics provided evidence that the approach is domain agnostic.

5.8 Improvements made since completion of Evaluations

This section discusses key improvements which were completed on the MQI framework, MQIO and OSCD as a result of feedback collected in the aforementioned evaluations. **Supplementary information related to this section is stored in a folder ("/Improvements") of the GitHub.**

Improvements as a result of **experiment 1** were iteratively completed throughout the lifecycle of the experiment. The improvements involved debugging and refactoring of code in order to resolve incorrectly detected quality issues. Each improvement was thoroughly tested to ensure that the root cause of the issue was identified and resolved correctly. In addition, other mappings were retested to ensure that the modifications did not introduce new identification problems. These improvements are demonstrated in the framework used in experiment 2.

It was identified through the feedback of **experiment 2** that the previous interface (See **Figure 21**) of the mapping quality assessment and refinement component of the MQI framework required improvements. Therefore, the interface was modified in order to address the participants recommendations. **Figure 52** presents the latest interface of the mapping quality assessment and refinement component of the MQI framework.



MQI FRAMEWORK • MAPPING QUALITY ASSESSMENT & REFINEMENT COMPONENT

Mapping Assessment Information

The table below shows the mapping quality information generated by the framework. **Select refinements then press 'Create Refinements'**

Violation ID ⓘ	Metric ID ⓘ	Result Message ⓘ	Violation value ⓘ	Triple Map ⓘ	Violation Location ⓘ	Select Refinements ⓘ	Display violation ⓘ
0	D1	Usage of undefined class ⓘ	mo:AudioFiles	work_lyrics	subjectMap	Manual	Hide Violation
1	D6	Usage of incorrect range. Term type should be 'rr:IRI' or 'rr:BlankNode' for property 'mo:lyrics'. ⓘ	rr:Literal	work_lyrics	predicateObjectMap-1	Manual	Display Violation

```
lb:work_lyrics
rr:subjectMap [
        rr:class  mo:AudioFiles,  mo:MusicalWork;
        rr:template  "http://musicbrainz.org/work/{gid}#_" ;
 ];
```

Mapping Validation Chart

The validation bar chart shows the relationship between quality violations and their corresponding quality dimensions. **Hover over the dimensions shown to discover a description.**

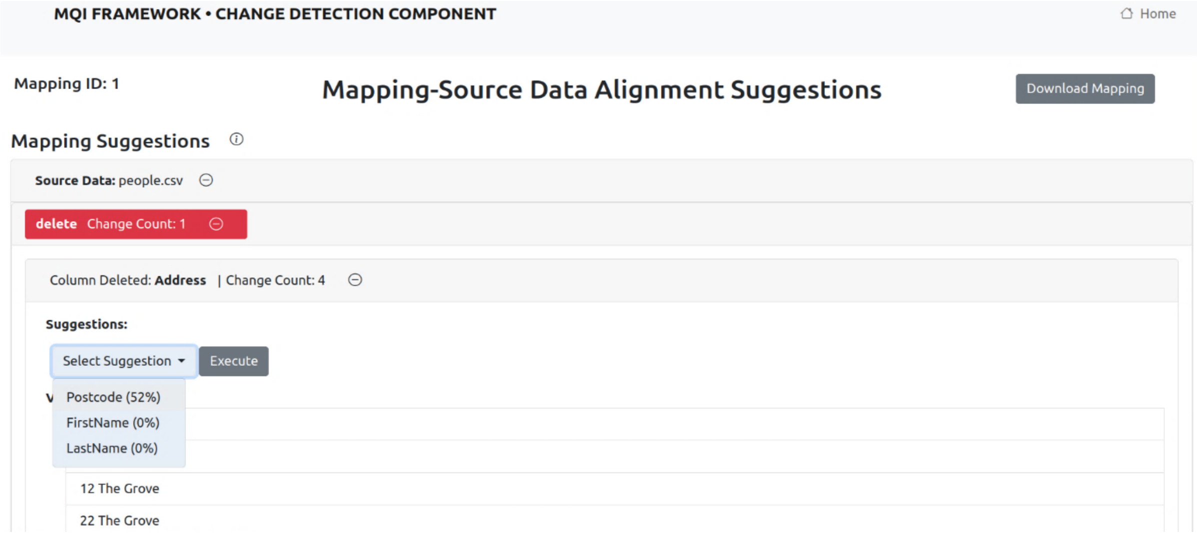
Figure 52: Latest interface of the mapping quality assessment and refinement component of the MQI framework

Apart from general restyling of the interface, hyperlinks ("Metric ID") have been added, which reference the documentation[67] detailing each of the respective quality metrics in RDF format. Furthermore, textual descriptions have been added below the name of each refinement in order to clarify their functionality. Moreover, the validation bar chart shown was moved from a different page to below the mapping quality assessment information shown in hopes of improving navigation. The quality violation identifier ("Violation ID") was added to the bars of the respective quality dimension in the chart in hopes of improving traceability of detected quality issues.

It was identified through the feedback of **experiment 3** that the source change detection component of the MQI framework required a small number of minimal modifications, such as adding tool tip text and opening hyperlinks in the current tab. These modifications were implemented by modifying several instances of source code and are reflected in the final version of the framework. In addition, it was decided to implement functionality into the final version which provided alignment recommendations based on the change detection information of the framework. **Figure 53** presents a screenshot of the suggestions provided by the source change detection component for the replacement of incompatible data references. A similarity measurement is used to rank similarities between the

[67] https://w3id.org/MQIO-metrics

data references and compatible references. The similarity is measured using the WordNet Similarity[68] library in Python, which is designed to measure semantic similarity between a pair of concepts.



Mapping ID: 1

Mapping-Source Data Alignment Suggestions

Mapping Suggestions

Source Data: people.csv

delete Change Count: 1

Column Deleted: **Address** | Change Count: 4

Suggestions:

12 The Grove

22 The Grove

Figure 53: Screenshot of source change detection component displaying suggestions to improve alignment

The drop down shows data references which have been inserted into the source data and a similarity score with the deleted column (“Address”). As can be seen, the “Postcode” column is most similar with a similarity score of 52%. The *“Execute”* button triggers a SPARQL query (“update_data_reference.rq”), which is designed to update the respective data reference in the mapping with the selected reference. In addition, it was decided to implement functionality to generate SHACL [80] constraints from the original source data which could detection of alignment issues when applied to associated mappings represented in RDF format. **Table 68** presents the pseudocode for generation of respective constraints and a sample SHACL constraint generated as a result.

Table 68: Pseudocode for Shape Generation (**A**) and SHACL Shape generated (**B**)

A

```
Input: columns of original source data
column-count ← count total number of columns
Output: SHACL shape to assess alignment
Initialization of Variables: Assign zero to variable i and empty list to columns
while (i < column-count) do // Iterate column names
        column-name ← retrieve current column name using i
        append column-name to columns list
end
compute remaining shape targeting rr:objectMap
compute SHACL list using columns, rr:column and rml:reference
```

B

```
schema:PersonShape
   a sh:NodeShape ;
   sh:targetObjectsOf  rr:objectMap ;
   sh:property [
     sh:path rr:column, rml:reference;
     sh:in ("ID" Address");
     sh:message """Data reference no
             longer in source data.""" ;
   ] .
```

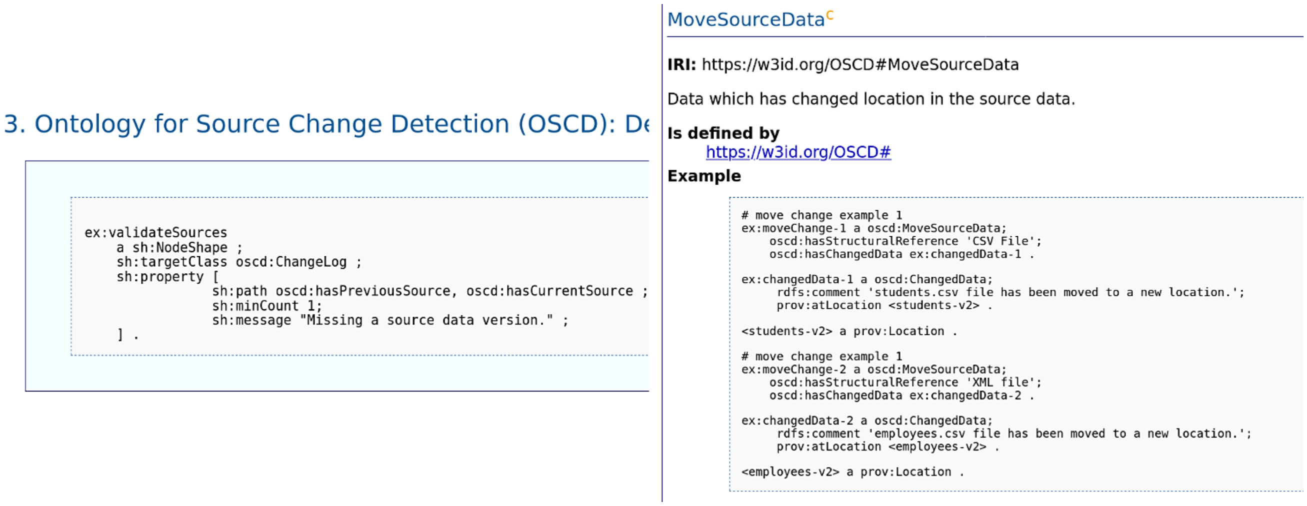
[68] https://pypi.org/project/sematch/

The pseudocode (**A**) can be used to map the columns in the original source into a SHACL constraint (“sample_shacl_shape.ttl”), which validates the existence of each data reference in a respective mapping. The sample constraint (**B**) shown validates that the data references in a mapping are one of the existing references (`“ID”, “Address”`) from the sample source data (“sample_source_data.csv”).

The constraints can be applied to mappings represented in RDF format, such as RML [44] and R2RML [35] mappings, throughout their lifecycle in order to detect potential alignment issues. The shape constraint when executed will result in a validation report (“sample_shacl_report.ttl”) represented in RDF format. Therefore, it can be linked with respective mappings to provide indications on the current level of alignment.

While the new functionality was not tested, it is hoped it provides insights into possible future work for autonomic improvement of alignment between mappings and underlying data sources. The functionality was discussed in the latest peer review publication [117] associated with the research in this thesis and received positive feedback from reviewers.

It was identified through feedback collected from **experiment 4 and 5** that MQIO and OSCD required a small number of modifications to the implementation and associated documentation. The implementation was modified using tools described in the design methodologies of both ontologies (See Section 4.3.1.1). Protégé [154] is a GUI tool, which was used to modify the concepts and relationships of both ontologies in line with respective feedback. Widoco [54] is a tool which was used to regenerate the new documentation of both ontologies, reflecting the final versions of the ontologies. **Figure 54** presents how the modifications were implemented on the documentation of the OSCD.



3. Ontology for Source Change Detection (OSCD): De

```
ex:validateSources
    a sh:NodeShape ;
    sh:targetClass oscd:ChangeLog ;
    sh:property [
                sh:path oscd:hasPreviousSource, oscd:hasCurrentSource ;
                sh:minCount 1;
                sh:message "Missing a source data version." ;
    ] .
```

MoveSourceData

IRI: https://w3id.org/OSCD#MoveSourceData

Data which has changed location in the source data.

Is defined by
https://w3id.org/OSCD#

Example

```
# move change example 1
ex:moveChange-1 a oscd:MoveSourceData;
    oscd:hasStructuralReference 'CSV File';
    oscd:hasChangedData ex:changedData-1 .

ex:changedData-1 a oscd:ChangedData;
     rdfs:comment 'students.csv file has been moved to a new location.';
     prov:atLocation <students-v2> .

<students-v2> a prov:Location .

# move change example 1
ex:moveChange-2 a oscd:MoveSourceData;
    oscd:hasStructuralReference 'XML file';
    oscd:hasChangedData ex:changedData-2 .

ex:changedData-2 a oscd:ChangedData;
     rdfs:comment 'employees.csv file has been moved to a new location.';
     prov:atLocation <employees-v2> .

<employees-v2> a prov:Location .
```

Figure 54: Screenshot of modifications to the documentation of the OSCD

The screenshot shows one of the SHACL constraints added (left) and examples of move changes (right) in a source data, which is expressed in the OSCD. The constraint shown can be used to validate that a change report graph expressed in the OSCD includes a previous and current version of related source data. In addition, an Appendix section (See **Appendix W**) was added to the documentation of both ontologies which included sample instance graphs.

5.9 Chapter Summary

This chapter has presented five evaluations designed to assess different aspects related to the MQI framework and associated ontologies, MQIO and OSCD. Experiment 1 evaluated the accuracy of the mapping quality assessment and refinement component of the framework. Experiment 2 evaluated the usability and effectiveness of the mapping quality assessment and refinement component of the framework. Experiment 3 evaluated the usability and understanding of the source change detection component of the framework. Experiment 4 evaluated the design of MQIO. Experiment 5 evaluated the design of OSCD. To the best of the author's knowledge, this thesis presented the first usability experiments evaluating the effectiveness and understanding of a mapping quality framework in the LD domain.

Experiment 1 involved assessing mappings collected from research projects and knowledge engineering students. The mappings were designed for transforming data in different domains and had different levels of authorship. The results from the experiment showed that the framework is capable of accurately identifying quality issues in real world mappings. In addition, the results showed the framework is applicable to domain agnostic mappings.

Experiment 2 evaluated the usability and effectiveness of the mapping quality assessment and refinement component of the MQI framework. Participants with varying background knowledge interacted with the framework in order to resolve quality issues in a mapping. The results showed that all participants were effective in resolving quality issues. However, it was discovered experienced participants were more effective when removing quality issues. Participants in both cohorts were similarly satisfied with the usability of the framework. Nonetheless, the results showed that the framework facilitated effective mapping quality assessment and refinement.

Experiment 3 evaluated the usability and understanding of the source change detection component of the MQI framework. This experiment involved participants interacting with the framework in order to discover changes in the source data of a mapping. The results showed similar levels of understanding and satisfaction by participants in both cohorts, therefore, facilitating the identification of source data changes of respective mappings.

Finally, **experiment 4** and **5** validated the design of the MQIO and OSCD. The experiment involved gathering feedback from ontology design experts on various aspects, such as design methodology, documentation and implementation. Experts suggested minor modifications to the current iterations of both ontologies. However, it was concluded that both ontologies conformed to ontology design best practices in the state of the art as all participants supported the design practices.

The **application** of the framework in two use cases demonstrated the applicability of the approach to real world scenarios. In addition, the use cases helped to identify design issues when applied to various domains.

Overall, it can be concluded that the MQI framework provides an effective and understandable approach to support users in the creation and maintenance of high-quality LD mappings. In addition, it can be concluded that MQIO and OSCD have been designed in conformance with ontology design best practices.

Chapter 6: Conclusion

This section discusses the conclusions which were identified as a result of the research described in this thesis. **Section 6.1** describes the fulfillment of the research objectives. **Section 6.2** discusses the contributions of the research and presents the uptake of the research. **Section 6.3** outlines future work which could be completed. **Section 6.4** provides some final remarks.

6.1 Fulfillment of Research Objectives

The research question investigated in this thesis was:

> *"To what extent can the detection of declarative mapping quality issues and source data changes, facilitate the creation and maintenance of high-quality Linked Data (LD) datasets?"*

The extent to which the research objectives outlined in **Section 1.3** have been fulfilled are presented in the following subsections.

6.1.1 Research Objective 1

The first research objective of this thesis was ***"Establish the State-of-the-Art of existing approaches which are designed to: (a) Improve quality of mappings in LD domain and (b) Address dynamics of LD.".*** This objective was accomplished by conducting a state of the art review involving two phases, which was presented in **Chapter 2**. RO1(a) was accomplished by the first phase of the review which reviewed existing approaches designed to assess and refine the quality of uplift mappings used to generate LD datasets. This phase resulted in the identification of two limitations, which related to the lack of an ontology designed for mapping quality and user testing in a related approach. RO1(b) was accomplished by the second phase of the review, which reviewed approaches to address the dynamics of resources, interlinks and source data of LD. The key characteristics examined during the analysis were derived from the research question of this thesis, being an approach to support the creation and maintenance of high-quality LD datasets, testing with respective end users. The state of the art review resulted in the identification of the lack of user testing with existing approaches. None of the 10 approaches reviewed published details of any formal user testing. The state of the art review was used to identify requirements that informed the design of the MQI framework that is presented in **Chapter 4**.

6.1.2 Research Objective 2

The second research objective of this thesis was ***"Develop the following: (a) OWL2 ontology to represent LD information related to mapping quality. (b) An approach to enable the identification and removal of issues related to quality of uplift mappings (see Table 2 below).".*** This objective was accomplished by the design, implementation and evaluation of the mapping quality assessment and refinement component of the MQI framework and Mapping Quality Improvement Ontology (MQIO) presented in **Section 4.2**, which support requirements derived from limitations discovered during the first phase of the state of the art review described in **Section 2.1**. The framework and ontology facilitate the identification and removal of the root cause of mapping quality issues shown to impact the quality of LD datasets. R02(a) was accomplished by the development of MQIO, which resolved the limitation related to the lack of an ontology to represent information related to LD mapping quality assessment, refinement and validation. RO2(b) was accomplished by the development of the mapping quality assessment and refinement component of the MQI framework designed to identify quality issues using metrics and suggest semi-automatic refinements inspired by the reviewed approaches. MQIO enables detailed provenance to be captured by this component in an interoperable and ontology-based format.

6.1.3 Research Objective 3

The third research objective of this thesis was ***"Develop the following: (a) OWL2 ontology to represent changes to source data associated with the LD dataset. (b) An approach to preserve alignment between source data changes and the respective uplift mappings.".*** This objective was accomplished by the design, implementation and evaluation of the source change detection component of the MQI framework and Ontology for Source Change Detection (OSCD) presented in **Section 4.3**, which support requirements derived from limitations discovered during the second phase of the state of the art review described in **Section 2.2**. The framework and ontology facilitate the identification of relevant source data changes, which have been shown to impact the level of alignment with respective mappings. R03(a) was accomplished by the development of OSCD, which resolved the limitation related to the lack of an ontology to represent changes in the source data of LD mappings. RO3(b) was accomplished by the development of the source change detection component of the MQI framework designed to provide timeliness information to data maintainers on changes which shown to impact the level of alignment with respective mappings. OSCD enables detailed provenance to be captured by this component in an interoperable and ontology-based format.

6.1.4 Research Objective 4

The fourth research objective of this thesis was ***"Implement and evaluate the approaches defined in RO2 and RO3. ".*** This research objective was accomplished by the design of the MQI framework was implemented using several different technologies as described in **Chapter 4**. OSCD and MQIO were implemented as OWL2 ontologies. The implementation provides a user-friendly GUI designed with requirements in mind. 5 evaluations and 1 application study were conducted on the resulting implementation, with results indicating a high-level of satisfaction, effectiveness and understanding as measured by multiple questionnaires and comparison of relevant artefacts. RO2 was evaluated in experiment 1, 2 and 4. Experiment 1 demonstrated that the framework can accurately identify quality issues in real world mappings, with over 200 violations detected, which would have exponentially propagated into the resulting dataset if not captured. Experiment 2 demonstrated that the framework provides an effective method for users to remove quality issues with 70% of all participants removing all three quality issues. Experiment 4 demonstrated that MQIO, which was used by the framework to represent quality information, was designed according to ontology design best practices. In addition, the application of MQIO was measured in the usability tests, which demonstrated that the ontology provides sufficient representation of relevant knowledge. RO3 was evaluated in experiment 3 and 5. Experiment 3 demonstrated that the change information designed to support preservation of alignment between mappings and source data was highly-understandable, with an overall mean score of 88% and standard deviation of 0.125 in the understanding questionnaire (**Section 5.2.4.2**). The results of user testing on the implementation provided evidence that respective end users could successfully interact with the design intended to facilitate the creation and maintenance of high-quality LD datasets. The application study demonstrated the applicability of the framework to real world scenarios, which include a network monitoring use case in Ericsson (**Section 5.7.2**) and geospatial data use case in Ordinance Survey Ireland (**Section 5.7.1**). The evaluation strategy advances the state of the art as none of the reviewed approaches conducted user experiments using standardized methods on the resulting design.

6.2 Research Contributions

This section discusses the one major and three minor contributions of this thesis, which were presented in **Chapter 1**.

The **first major contribution of this research is the MQI framework**, which was described in **Chapter 4**. The framework advances the state of the art by resolving four of the limitations identified through the review of existing state of the art approaches described in **Chapter 2**. These limitations identified a lack of a user tested approach designed to improve quality in the publication process, which captures associated information in ontologies created with the requirements in mind. The framework is designed to support different types of users (experts and non-experts) in the creation and maintenance of high-quality LD datasets. The framework provides linkable and queryable information, which was captured during the publication process in RDF format in order to support the maintenance of the mappings and resulting dataset. Therefore, the MQI framework brings quality assessment earlier into the process and results in the removal of the root issues of many quality issues identified in the state of the art. In addition, the framework supports the maintenance of datasets after publication in order to ensure the most accurate and up to date data is provided to consumers. To the best of the author’s knowledge, the framework is the first to provide push notifications for changes in the source data of LD. Finally, it is hoped that the framework can improve the adoption of linked data by providing methods to support and model the publication processes involved in the creation of high-quality LD datasets.

The first minor contribution of this research is the Mapping Quality Improvement Ontology (MQIO). The ontology advances the state of the art as no ontology was found in the review of existing approaches to represent information related to mapping quality assessment, refinement and validation in the LD domain. The ontology provides an ontology-based model to capture quality-related provenance during the publication process of LD, therefore, improving the trustworthiness of the published data.

The second minor contribution of this research is the Ontology for Source Change Detection (OSCD). The ontology advances the state of the art as no ontology was found in the review of existing approaches to represent information related to changes which have occurred in heterogeneous source data of LD mappings. The ontology provides an ontology-based model in order to capture the changes and support the preservation of alignment between the mappings and underlying data sources. The information will benefit the maintenance of LD by providing indications of source data changes which should be propagated into the resulting dataset.

The third minor contribution of this thesis is the evaluation results of the 5 experiments and 1 application study which were conducted on the MQI framework. To the best of the author's knowledge, the experiments described in this thesis advances the state of the art by conducting a combination of system and user testing, involving standardized methods in order to validate an approach designed to improve the quality and maintenance of LD mappings. The experiments involved over 100 participants with diverse background knowledge, including knowledge engineering students, mapping specialists and ontology design specialists. The results shown provide insights for researchers on how different methods (e.g., synchronized and unsynchronized) and metrics (e.g., questionnaires, verbalization of actions, comparison of artefacts) captured relevant information. Finally, it is hoped that the evaluation results and setup, improves the adoption of linked data by providing methods to validate user-based approaches.

6.2.1 Impact and Uptake

The research described in this thesis has already impacted the community by 9 successful publications describing the design, application and evaluations of the MQI framework, MQIO and OSCD at peer reviewed conferences and workshops. The well-known venues include the 18th International Conference on Semantic (SEMANTiCS) 2022, Extended Semantic Web Conference (ESWC) 2021 and 2023, 21st International Semantic Web Conference (ISWC) 2022 and 5th International Conference on Semantic Computing (ICSC) 2021, as outlined in **Section 1.5**. Furthermore, the MQIO and OSCD were presented to a panel of semantic web experts at the semantic interoperability conference (SEMIC2022)[69] organized by the European commission. Moreover, work related to the Ericsson use case has been published in the 9th International Conference on Network Softwarization (NetSoft 2023). The published work has been cited by 3 other works as of 2023, which include:

[69] https://joinup.ec.europa.eu/collection/semic-support-centre/semic-conference

- A citation from a semantic quality assessment framework focusing on transformed tabular data (*Conference Paper*).
- A citation from a framework which generates semantic clinical data (*PhD Thesis*).
- A citation from a lifecycle model of LD, involving the generation of a German Tourism Knowledge Graph (*Journal Article*).

It is hoped further citations will be received by approaches involved in the various processes of the publication of LD, which could result in a far-reaching impact with increasing uptake in the research community. In addition, the MQI framework has been used in the evaluation strategy of the LD publication process of a number of use cases (**Section 5.7**). The usage of MQI framework is outlined below:

- Applied to existing independently developed mappings, which represent various knowledge domains as presented in **Section 5.3**. The framework enabled the identification of various quality issues covering multiple quality aspects, dimensions and categories in the assessed mappings. In addition, the framework demonstrated potential to support the resolution of most of the detected issues using the proposed semi-automatic refinements.
- **Third levels students completing an uplift project:** This use case involved students in a third level MSc module creating uplift mappings as part of their course work. These mappings were designed to transform data from various domains, such as people and related hobbies. The framework was made available to students during the course in order to allow them to assess and refine the quality of any created mappings. Feedback from students was positive and none of them encountered issues using the framework, which indicated sufficient functionality for them.
- **OSi uplifting project transforming geospatial data:** This use case involved applying the framework in a research project, which transformed Irelands national geospatial data into LD representation. The project involved the creation of several uplift mappings designed to transform information related to metrics and measurements, among others. The framework was used to identify and remove quality issues in each of the mappings created throughout the lifecycle of the project. The application of the framework resulted in a significant quality improvement in the mappings by eradicating the root cause of several quality issues (See Section 5.7.1) and potentially positively impacting the quality of nearly 2000 triples. Therefore, improving the trustworthiness of the resulting dataset. The application of the framework in this use case provided indications that the framework can be applied to geospatial domain related mappings in order to improve mapping quality.
- **Ericsson uplift project transforming network monitoring data:** This use case involved applying the framework in a research project, which involved transforming network monitoring data into LD. The project involved the creation of several uplift mappings designed to transform information related to metrics, goals and actions. The framework was used to identify and remove quality issues (See Section 5.7.2) in each of the mappings created throughout the lifecycle of the project. In addition, the framework

was used to maintain alignment between the mappings and underlying data sources after the data was published. The use case proved beneficial as it also discovered issues in the design of the associated ontology under construction as a result of detected mapping quality issues. In addition, the use case demonstrated the applicability of the framework in real world scenarios. The reports generated during the use case were linked with resulting LD datasets in order to improve trustworthiness for consumers. The application of the framework in this use case provided indications that the framework can be applied to network monitoring domain related mappings in order to improve and maintain mapping quality.

The usage of the framework within Ericsson and OSi use case resulted in the identification and removal of quality issues described in **Section 5.7**, preventing them from exponentially propagating into the resulting dataset and decreasing quality provided to their consumers [42,65,75,97,101]. It is hoped sharing of the open-source implementation of MQI framework will increase the uptake by agents involved in the publication of LD datasets.

6.3 Future Work

This section discusses potential future work for the research outlined in this thesis.

Expanding range of Mapping Languages supported: While the design of the MQI framework is mapping independent and applicable to all representations. The implementation of the design focused on the quality of mappings represented in R2RML [35] and RML [44], however, other mapping representations exist in the LD domain, such as YARRRML [94] and Direct Mapping [5]. Therefore, it would be useful to provide support for these representations, which could involve preprocessing of the mapping, such as conversion of representation using tools, such as the YARRRML parser[70], designed to convert mappings represented in YARRRML into RML.

Expanding types of Mappings supported: The MQI framework was designed to improve the quality of uplift mappings involved in the publication process of LD datasets. However, it is envisaged that aspects of the approach could be reused by an approach to target the quality of semantic mappings designed to link semantically similar concepts in published datasets [130]. For instance, MQIO could be reused to capture quality information on semantic mappings as it provides support for both types (uplift and semantic). It is envisaged similar techniques used by the metrics of the MQI framework could be applied to semantic mappings. For instance, the "Usage of Undefined Classes" metric (See Section 4.2.2.1.3) could be used to test that classes linked by semantic mappings are defined in their respective namespaces. In addition, the human-in-the-loop refinements used by the framework

[70] https://github.com/RMLio/yarrrml-parser

could be adapted to support the resolution of quality issues detected in semantic mappings. For instance, the “Find Similar Classes” semi-automatic refinement (See Section 4.2.2.2) could be used to identify the correctly defined class for the mapping, by suggesting classes from the namespace ontology.

Automating Alignment: Currently, MQI facilitates identification of changes in the source data of mappings and provides information necessary for human data maintainers to preserve alignment between them. However, autonomic alignment support would allow software agents to automatically take actions to preserve alignment, such as periodic regeneration of the published dataset in order to capture recent source data changes. It would be interesting to explore how much of the alignment maintenance could be achieved just through reasoning over the available graphs of mappings, changes detected and change policies.

Integration with Existing Tools: Integrating the functionality of the MQI framework, MQIO and OSCD with similar semantic web technologies involved in the publication process of LD. For instance, integrating the quality reports into existing LD visualization tools, such as Juma [32], designed to create uplift and semantic mappings visually. The integration would allow detailed provenance to be captured during the creation and maintenance of the resulting mapping.

6.4 Final Remarks

It is hoped that with the design, implementation and evaluation of the MQI framework, that the LD community will benefit, by providing an approach to improve and maintain the quality of mappings, therefore, resulting in improved LD dataset quality and maintenance. The publication process of LD can greatly benefit from the MQI framework by capturing quality issues early in order to eradicate the root cause and providing detailed quality-oriented provenance to data publishers and consumers. In addition, MQIO and OSCD can be utilized by other tools involved in the publication process to provide a semantic and linkable representation of associated information.

Researchers involved in the creation and maintenance of LD datasets can benefit from the use of the framework by facilitating the improvement of mapping quality and alignment with underlying data sources. It is hoped the RDF-based reports generated by the framework can improve the trustworthiness of data by providing indications to consumers on the suitability of the data for their use case. Finally, it is hoped the aspects of the evaluation strategy applied to the MQI framework can be reused by similar approaches to promote collaboration between computer scientists and domain experts.

References

[1] Maribel Acosta, Amrapali Zaveri, Elena Simperl, Dimitris Kontokostas, Sören Auer, and Jens Lehmann. 2013. Crowdsourcing linked data quality assessment. In *12th International Semantic Web Conference (ISWC 2013)*, 260–276. DOI:https://doi.org/10.1145/1866029.1866078

[2] Riccardo Albertoni and Antoine Isaac. 2016. Data on the web best practices: Data quality vocabulary. *W3C Working Draft 19*. Retrieved April 1, 2023 from https://www.w3.org/TR/vocab-dqv/

[3] Sarah Alzahrani and Declan O'Sullivan. 2022. Towards a unified metadata model for semantic and data mappings. In *17th International Workshop on Ontology Matching (OM 2022) co-located with the 21th International Semantic Web Conference (ISWC 2022)*, 223–225.

[4] Jennie Andersen, Sylvie Cazalens, and Philippe Lamarre. 2022. Improving KG Completeness Evaluation considering Information Need as First-Class Citizen. Retrieved April 1, 2023 from https://hal.science/hal-03986309

[5] Marcelo Arenas, Alexandre Bertails, Eric Prud'hommeaux, and Juan Sequeda. 2012. A direct mapping of relational data to RDF. *World Wide Web Consortium (W3C) Recommendation 27*, 1–11. Retrieved April 1, 2023 from https://www.w3.org/TR/rdb-direct-mapping/

[6] Ahmad Assaf, Raphaël Troncy, and Aline Senart. 2015. What's up LOD Cloud? Observing the State of Linked Open Data Cloud Metadata. In *12th Extended Semantic Web Conference (ESWC 2015)*, 247–254. DOI:https://doi.org/10.1007/978-3-319-25639-9_40

[7] Dylan Van Assche, Thomas Delva, Gerald Haesendonck, Pieter Heyvaert, Ben De Meester, and Anastasia Dimou. 2023. Declarative RDF graph generation from heterogeneous (semi-)structured data: A systematic literature review. *J. Web Semant.* 75, (2023), 35–60. DOI:https://doi.org/10.1016/j.websem.2022.100753

[8] Oren Ben-Kiki, Clark Evans, and Brian Ingerson. 2009. Yaml ain't markup language (yaml^{TM}) version 1.1. *Work. Draft 2008* 5, (2009), 11. Retrieved April 1, 2023 from https://yaml.org/spec/1.1/

[9] Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Sci. Am.* 284, 5 (2001), 34–43.

[10] Camila Bezerra, Fred Freitas, and Filipe Santana. 2013. Evaluating ontologies with Competency Questions. In *Proceedings - 2013 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IATW 2013*. DOI:https://doi.org/10.1109/WI-IAT.2013.199

[11] Christian Bizer. 2003. D2R MAP - A Database to RDF Mapping Language. In *12th International World Wide Web Conference- Posters (WWW 2003)*, 319–320.

[12] Christian Bizer and Richard Cyganiak. 2009. Quality-driven information filtering using the WIQA policy framework. *J. Web Semant.* 7, 1 (2009), 1–10. DOI:https://doi.org/10.1016/j.websem.2008.02.005

[13] Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee. 2008. Linked data on the web (LDOW2008). In *Proceedings of the 17th international conference on World Wide Web (WWW '08)*, 1265–1266. DOI:https://doi.org/10.1145/1367497.1367760

[14] Christian Bizer, Pablo Mendes, Zoltán Miklos, Jean-Paul Calbimonte, Alexandra Moraru, and Giorgos Flouris. 2012. *D2. 1 conceptual model and best practices for high-quality metadata publishing*. Retrieved April 1, 2023 from https://www.planet-data.eu/results/deliverables.html

[15] Christian Bizer and Andy Seaborne. 2004. D2RQ – Treating Non-RDF Databases as Virtual RDF Graphs. In *Proceedings of the 3rd international semantic web conference (ISWC2004)*, 125–127.

[16] Christoph Böhm, Felix Naumann, Ziawasch Abedjan, Dandy Fenz, Toni Grütze, Daniel Hefenbrock, Matthias Pohl, and David Sonnabend. 2010. Profiling linked open data with ProLOD. In *2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*, 175–178. DOI:https://doi.org/10.1109/ICDEW.2010.5452762

[17] Piero A Bonatti, Aidan Hogan, Axel Polleres, and Luigi Sauro. 2011. Robust and scalable linked data reasoning incorporating provenance and trust annotations. *J. Web Semant.* 9, 2 (2011), 165–201. DOI:https://doi.org/10.1016/j.websem.2011.06.003

[18] Mokrane Bouzeghoub. 2004. A framework for analysis of data freshness. In *Proceedings of the 2004 international workshop on Information quality in information systems*, 59–67. DOI:https://doi.org/10.1145/1012453.1012464

[19] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 2 (2006), 77–101. DOI:https://doi.org/10.1191/1478088706qp063oa

[20] Laura Bravi, Federica Murmura, and Gilberto Santos. 2019. The ISO 9001:2015 Quality Management System Standard: Companies' Drivers, Benefits and Barriers to Its Implementation. *Qual. Innov. Prosper.* 23, (2019), 64. DOI:https://doi.org/10.12776/qip.v23i2.1277

[21] Dan Brickley, Ramanathan V Guha, and Brian McBride. 2014. RDF Schema 1.1. *World Wide Web Consortium (W3C) Recommendation*. Retrieved April 1, 2023 from https://www.w3.org/TR/rdf-schema/

[22] Andrew Burton-Jones, Veda C. Storey, Vijayan Sugumaran, and Punit Ahluwalia. 2005. A semiotic metrics suite for assessing the quality of ontologies. In *Data and Knowledge Engineering*, 84–102. DOI:https://doi.org/10.1016/j.datak.2004.11.010

[23] TANG Caifang, R A O Yuan, Y U Hualei, S U N Ling, CHENG Jiamin, and WANG Yutian. 2021. Improving knowledge graph completion using soft rules and adversarial learning. *Chinese J. Electron.* 30, 4 (2021), 623–633.

[24] Jeremy J Carroll. 2003. Signing RDF graphs. In *2nd International Semantic Web Conference (ISWC 2003)*, 369–384. DOI:https://doi.org/10.1007/978-3-540-39718-2_24

[25] Victor Charpenay, Sebastian Käbisch, and Harald Kosch. 2016. Introducing Thing Descriptions and Interactions: An Ontology for the Web of Things. In *Joint Proceedings of the 3rd Stream Reasoning (SR) and the 1st Semantic Web Technologies for the Internet of Things (SWIT) workshops co-located with 15th International Semantic Web Conference (ISWC)*, 55–66.

[26] David Chaves-Fraga, Kemele M. Endris, Enrique Iglesias, Oscar Corcho, and Maria-Esther Vidal. 2019. What Are the Parameters that Affect the Construction of a Knowledge Graph? In *On the Move to Meaningful Internet Systems: OTM 2019 Conferences- Confederated International Conferences: CoopIS, ODBASE (C&TC 2019)*, 695–713. DOI:https://doi.org/10.1007/978-3-030-33246-4_43

[27] Lei Chen, Haifei Zhang, Ying Chen, and Wenping Guo. 2012. Blank Nodes in RDF. *J. Softw.* (2012), 1993–1999. DOI:https://doi.org/10.4304/jsw.7.9.1993-1999

[28] Ping Chen and Walter Garcia. 2010. Hypothesis generation and data quality assessment through association mining. In *9th IEEE International Conference on Cognitive Informatics (ICCI'10)*, 659–666.

[29] Ademar Crotti, Christophe Debruyne, Rob Brennan, and Declan O'Sullivan. 2017. An evaluation of uplift mapping languages. *Int. J. Web Inf. Syst.* 13, 4 (2017), 405–424. DOI:https://doi.org/10.1108/IJWIS-04-2017-0036

[30] Ademar Crotti, Christophe Debruyne, and Declan O'Sullivan. 2018. Juma Uplift: Using a Block Metaphor for Representing Uplift Mappings. In *Proceedings - 12th IEEE International Conference on Semantic Computing, ICSC 2018*, 211–218. DOI:https://doi.org/10.1109/ICSC.2018.00037

[31] Ademar Crotti Junior. 2019. A Jigsaw Puzzle Metaphor for Representing Linked Data Mappings. PhD Thesis.

Trinity College Dublin. Retrieved April 1, 2023 from http://hdl.handle.net/2262/86157

[32] 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 *14th Extended Semantic Web Conference (ESWC 2017) Posters and Demos*, 87–92. DOI:https://doi.org/10.1007/978-3-319-70407-4_17

[33] Richard Cyganiak, Dave Reynolds, and Jeni Tennison. 2014. The RDF Data Cube Vocabulary. *World Wide Web Consortium (W3C) Recommendation*. Retrieved April 1, 2023 from https://www.w3.org/TR/vocab-data-cube/

[34] Richard Cyganiak, David Wood, Markus Lanthaler, Graham Klyne, Jeremy J Carroll, and Brian McBride. 2014. RDF 1.1 Concepts and Abstract Syntax. *World Wide Web Consortium (W3C) Recommendation 25*. Retrieved April 1, 2023 from https://www.w3.org/TR/rdf11-concepts/

[35] Souripriya Das, Seema Sundara, and Richard Cyganiak. 2012. R2RML: RDB to RDF Mapping Language. *World Wide Web Consortium (W3C) Recommendation*. DOI:https://doi.org/10.1017/CBO9781107415324.004

[36] Jeremy Debattista, Soren Auer, and Christoph Lange. 2016. Luzzu-A Framework for Linked Data Quality Assessment. In *IEEE 10th International Conference on Semantic Computing (ICSC 2016)*, 124–131. DOI:https://doi.org/10.1109/ICSC.2016.48

[37] Jeremy Debattista, Christoph Lange, and Sören Auer. 2014. daQ, an Ontology for Dataset Quality Information. In *Workshop on Linked Data on the Web (LDOW 2014) co-located with the 23rd International World Wide Web Conference (WWW 2014)*, 75–83.

[38] Jeremy Debattista, Christoph Lange, Sören Auer, and Dominic Cortis. 2018. Evaluating the quality of the LOD cloud: An empirical investigation. *Semant. Web* 9, (March 2018), 1–43. DOI:https://doi.org/10.3233/SW-180306

[39] Christophe Debruyne, Brian Walshe, and Declan O'Sullivan. 2015. Towards a project centric metadata model and lifecycle for ontology mapping governance. In *17th International Conference on Information Integration and Web-based Applications & Services (iiWAS 2015)*, 1–10.

[40] Mariangiola Dezani-Ciancaglini, Ross Horne, and Vladimiro Sassone. 2012. Tracing where and who provenance in linked data: A calculus. *Theor. Comput. Sci.* 464, (2012), 113–129.

[41] Oxford English Dictionary. 1989. Oxford english dictionary. *Simpson, Ja Weiner, Esc* 3, (1989).

[42] Anastasia Dimou, Dimitris Kontokostas, Markus Freudenberg, Ruben Verborgh, Jens Lehmann, Erik Mannens, Sebastian Hellmann, and Rik Van de Walle. 2015. Assessing and refining mappings to RDF to improve dataset quality. In *14th International Semantic Web Conference (ISWC 2015)*, 133–149. DOI:https://doi.org/10.1007/978-3-319-25010-6_8

[43] Anastasia Dimou, Tom De Nies, Ruben Verborgh, Erik Mannens, and Rik de Walle. 2016. Automated Metadata Generation for Linked Data Generation and Publishing Workflows. In *Proceedings of the 9th Workshop on Linked Data on the Web (LDOW) co-located with 25th International World Wide Web Conference (WWW)*, 1–10.

[44] Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. 2014. RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In *Proceedings of the Workshop on Linked Data on the Web co-located withthe 23rd International World Wide Web Conference (WWW 2014)*.

[45] Li Ding, Pranam Kolari, Zhongli Ding, Sasikanth Avancha, and others. 2005. Using Ontologies in the Semantic Web: A Survey. *Ontol. Context Inf. Syst.* (2005), 79–113. DOI:https://doi.org/10.1007/978-0-387-37022-4_4

[46] Amy Farrow. 2021. Open Data Quality Is Poor but Slowly Improving. *Cat. Qual. Scores Open Data Toronto* (2021). Retrieved April 1, 2023 from https://tellingstorieswithdata.com/inputs/pdfs/paper_one-2021-Amy_Farrow.pdf

[47] Kevin Chekov Feeney, Declan O'Sullivan, Wei Tai, and Rob Brennan. 2014. Improving curated web-data quality with structured harvesting and assessment. *Int. J. Semant. Web Inf. Syst.* 10, 2 (2014), 35–62. DOI:https://doi.org/10.4018/IJSWIS.2014040103

[48] Annika Flemming. 2010. Quality Characteristics of Linked Data Publishing Datasources. Master's Thesis. Humboldt-Universität zu Berlin. Retrieved April 1, 2023 from https://cs.uwaterloo.ca/~ohartig/files/DiplomarbeitAnnikaFlemming.pdf

[49] 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. DOI:https://doi.org/10.1177/104973239300300403

[50] International Organization for Standardization (ISO). 2005. ISO/IEC 25000:2005, Software Engineering - Software Product Quality Requirements and Evaluation (SQuaRE). Retrieved April 1, 2023 from https://www.iso.org/standard/64764.html

[51] Christian Fürber and Martin Hepp. 2011. SWIQA - A Semantic Web information quality assessment framework. In *19th European Conference on Information Systems (ECIS 2011)*, 19–30.

[52] Matthew Gamble and Carole A Goble. 2011. Quality, trust, and utility of scientific data on the web: towardsa joint model. In *Web Science 2011 (WebSci 2011)*, 15:1–15:8. DOI:https://doi.org/10.1145/2527031.2527048

[53] Aldo Gangemi and Valentina Presutti. 2009. Ontology design patterns. In *Handbook on ontologies*. Springer, 221–243. DOI:https://doi.org/10.1007/978-3-540-92673-3_10

[54] Daniel Garijo. 2017. WIDOCO: A Wizard for Documenting Ontologies. In *16th International Semantic Web Conference (ISWC 2017)*, 94–102. DOI:https://doi.org/10.1007/978-3-319-68204-4_9

[55] Yolanda Gil and Donovan Artz. 2006. Towards content trust of web resources. In *Proceedings of the 15th international conference on World Wide Web*, 565–574. DOI:https://doi.org/10.1145/1135777.1135861

[56] Jennifer Golbeck, Bijan Parsia, and James Hendler. 2003. Trust networks on the semantic web. In *7th International Workshop on Cooperative Information Agents (CIA 2003)*, 238–249. DOI:https://doi.org/10.1007/978-3-540-45217-1_18

[57] Jennifer Golbeck and Matthew Rothstein. 2008. Linking Social Networks on the Web with FOAF: A Semantic Web Case Study. In *23rd AAAI Conference on Artificial Intelligence (AAAI)*, 1138–1143.

[58] Miguel Grinberg. 2018. *Flask web development: developing web applications with python*. O'Reilly Media, Inc. Retrieved April 1, 2023 from https://www.oreilly.com/library/view/flask-web-development/9781491991725/

[59] Thomas R Gruber. 1995. Toward principles for the design of ontologies used for knowledge sharing? *Int. J. Hum. Comput. Stud.* 43, 5 (1995), 907–928. DOI:https://doi.org/10.1006/ijhc.1995.1081

[60] Christophe Guéret, Paul Groth, Claus Stadler, and Jens Lehmann. 2012. Assessing linked data mappings using network measures. In *9th Extended Semantic Web Conference (ESWC 2012)*, 87–102. DOI:https://doi.org/10.1007/978-3-642-30284-8_13

[61] Steve Harris, Andy Seaborne, and Eric Prud'hommeaux. 2013. SPARQL 1.1 Query Language. *World Wide Web Consortium (W3C) Recommendation 21*, 778. Retrieved April 1, 2023 from https://www.w3.org/TR/sparql11-query/

[62] Olaf Hartig. 2008. Trustworthiness of data on the web. In *Proceedings of the STI Berlin & CSW PhD Workshop*, 21–26.

[63] Olaf Hartig and Jun Zhao. 2009. Using Web Data Provenance for Quality Assessment. In *1st Workshop on the role of Semantic Web in Provenance Management (SWPM 2009) co-located with the 8th International*

Semantic Web Conference (ISWC-2009), 526–532.

[64] Pieter Heyvaert, Anastasia Dimou, Aron-Levi Herregodts, Ruben Verborgh, Dimitri Schuurman, Erik Mannens, and Rik de Walle. 2016. RMLEditor: A Graph-based Mapping Editor for Linked Data Mappings. 709–723. DOI:https://doi.org/10.1007/978-3-319-34129-3_43

[65] Pieter Heyvaert, Ben De Meester, Anastasia Dimou, and Ruben Verborgh. 2019. Rule-driven inconsistency resolution for knowledge graph generation rules. *Semant. Web* 10, 6 (2019), 1071–1086. DOI:https://doi.org/10.3233/SW-190358

[66] Pascal Hitzler and Krzysztof Janowicz. 2013. Linked Data, Big Data, and the 4th Paradigm. *Semant. Web* 4, 3 (2013), 233–235. DOI:https://doi.org/10.3233/SW-130117

[67] Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. 2010. Weaving the Pedantic Web. In *19th World Wide Web Conference (WWW 2010) Workshop on Linked Data on the Web (LDOW2010)*, 23–33.

[68] Aidan Hogan, Jürgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, and Stefan Decker. 2012. An empirical survey of Linked Data conformance. *J. Web Semant.* 14, (2012), 14–44. DOI:https://doi.org/10.1016/j.websem.2012.02.001

[69] Ian Horrocks. 2008. Ontologies and the semantic web. *Commun. ACM* 51, 12 (2008), 58–67. DOI:https://doi.org/10.1145/1409360.1409377

[70] Elwin Huaman and Dieter Fensel. 2022. Knowledge Graph Curation: A Practical Framework. In *Proceedings of the 10th International Joint Conference on Knowledge Graphs* (IJCKG '21), 166–171. DOI:https://doi.org/10.1145/3502223.3502247

[71] Ana Iglesias-Molina, Andrea Cimmino, David Chaves-Fraga, Edna Ruckhaus, Raúl García-Castro, and Oscar Corcho. 2022. An Ontological Approach for Representing Declarative Mapping Languages. *Semant. Web Jounal* (2022), 1–31. DOI:https://doi.org/10.3233/SW-223224

[72] International Organization for Standardization. 2018. Ergonomics of Human System Interaction: Usability, Definitions and Concepts. Retrieved April 1, 2023 from https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

[73] Ethan Jackson and Janos Sztipanovits. 2009. Formalizing the Structural Semantics of Domain-Specific Modeling Languages. *Softw. Syst. Model.* 8, (2009), 451–478. DOI:https://doi.org/10.1007/s10270-008-0105-0

[74] Ian Jacobi, Lalana Kagal, and Ankesh Khandelwal. 2011. Rule-based trust assessment on the semantic web. In *Rule-Based Reasoning, Programming, and Applications: 5th International Symposium (RuleML 2011)*, 227–241.

[75] Ademar Crotti Junior, Jeremy Debattista, and Declan O'Sullivan. 2019. 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*, 12–24.

[76] Tobias Käfer, Ahmed Abdelrahman, Jürgen Umbrich, Patrick O'Byrne, and Aidan Hogan. 2013. Observing Linked Data Dynamics. In *10th International Conference (ESWC 2013)*, 213–227. DOI:https://doi.org/10.1007/978-3-642-38288-8_15

[77] Lalana Kagal. 2002. Rei: A policy language for the me-centric project. *Technical Report, HP Labs*. Retrieved April 1, 2023 from http://www.hpl.hp.com/techreports/2002/HPL-2002-270.html

[78] A. M. Khattak, Z. Pervez, W. A. Khan, A. M. Khan, K. Latif, and S. Y. Lee. 2015. Mapping evolution of dynamic web ontologies. *Inf. Sci. (Ny).* 303, (2015), 101–119. DOI:https://doi.org/10.1016/j.ins.2014.12.040

[79] Pavel Klinov and Dmitry Mouromtsev. 2015. Measuring the Quality of Relational-to-RDFMappings. In

International Conference on Knowledge Engineering and the Semantic Web (KESW 2015), 210–224. DOI:https://doi.org/10.1007/978-3-319-24543-0

[80] Holger Knublauch and Dimitris Kontokostas. 2017. Shapes Constraint Language (SHACL). *World Wide Web Consortium (W3C) Recommendation*. Retrieved April 1, 2023 from https://www.w3.org/TR/shacl/

[81] Dimitris Kontokostas, Patrick Westphal, Sören Auer, Sebastian Hellmann, Jens Lehmann, Roland Cornelissen, and Amrapali Zaveri. 2014. Test-driven evaluation of Linked Data quality. In *23rd International Conference on World Wide Web (WWW 2014)*, 747–757. DOI:https://doi.org/10.1145/2566486.2568002

[82] D Krech. 2006. RDFLib: A python library for working with rdf. Retrieved April 1, 2023 from https://rdflib.readthedocs.io/en/stable/

[83] Martin Krzywinski and Naomi Altman. 2014. Visualizing samples with box plots. *Nat. Methods* 11, 2 (2014), 119–120. DOI:https://doi.org/10.1038/nmeth.2813

[84] Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. 2013. The PROV Ontology (PROV-O). *World Wide Web Consortium (W3C) Recommendation*. Retrieved April 1, 2023 from https://www.w3.org/TR/prov-o/

[85] Yuangui Lei, Victoria Uren, and Enrico Motta. 2007. A framework for evaluating semantic metadata. In *Proceedings of the 4th international conference on Knowledge capture*, 135–142. DOI:https://doi.org/10.1145/1298406.1298431

[86] 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 (2002), 463–488. DOI:https://doi.org/10.1080/10447318.2002.9669130

[87] Bernadette Farias Lóscio, Caroline Burle, and Newton Calegari. 2016. Data on the Web Best Practices. *World Wide Web Consortium (W3C) Recommendation*. Retrieved April 1, 2023 from https://www.w3.org/TR/dwbp/

[88] Davood Mazinanian and Nikolaos Tsantalis. 2016. An empirical study on the use of CSS preprocessors. In *2016 IEEE 23rd international conference on Software Analysis, Evolution, and Reengineering (SANER)*, 168–178. DOI:https://doi.org/10.1109/SANER.2016.18

[89] Kimberly McCoy. 2015. Using zoom, cloud based video web conferencing system: To enhance a distance education course and/or program. In *Society for Information Technology & Teacher Education International Conference*, 412–415.

[90] Kris McGlinn, Rob Brennan, Christophe Debruyne, Alan Meehan, Lorraine McNerney, Eamonn Clinton, Philip Kelly, and Declan O'Sullivan. 2021. Publishing authoritative geospatial data to support interlinking of building information models. *Autom. Constr.* 124, (2021), 103534. DOI:https://doi.org/10.1016/j.autcon.2020.103534

[91] Kris McGlinn, Christophe Debruyne, Lorraine McNerney, and Declan O'Sullivan. 2017. Integrating Ireland's Geospatial Information to Provide Authoritative Building Information Models. In *13th International Conference on Semantic Systems* (SEMANTICS 2017), 57–64. DOI:https://doi.org/10.1145/3132218.3132223

[92] Deborah L McGuinness, Frank Van Harmelen, and others. 2004. OWL Web Ontology Language Overview. *World Wide Web Consortium (W3C) Recommendation 10*, 2004. Retrieved April 1, 2023 from https://www.w3.org/TR/owl-ref/

[93] Alan Meehan, Rob Brennan, Dave Lewis, and Declan O'sullivan. 2014. Mapping Representation based on Meta-data and SPIN for Localization Workflows. In *2nd International Workshop on Semantic Web Enterprise Adoption and Best Practice and 2nd International Workshop on Finance and Economics on the Semantic Web co-located with 11th European Semantic Web Conference, WaSABi-FEOSW@ESWC 2014*, 33–41.

[94] B De Meester, P Heyvaert, and A Dimou. YARRRML. Unofficial draft, Ghent University-IDLab (2019). Retrieved April 1, 2023 from https://rml.io/yarrrml/spec/

[95] Ben De Meester. 2018. High Quality Schema and Data Transformations for Linked Data Generation. In *Doctoral Consortium at ISWC 2018 co-located with 17th International Semantic Web Conference (ISWC 2018)*, 18–26.

[96] Pablo Mendes, Hannes Mühleisen, and Christian Bizer. 2012. Sieve: linked data quality assessment and fusion. In *2012 Joint EDBT/ICDT Workshops*, 116–123. DOI:https://doi.org/10.1145/2320765.2320803

[97] Benjamin Moreau and Patricia Serrano-Alvarado. 2020. Assessing the Quality of RDF Mappings with EvaMap. In *17th Extended Semantic Web Conference (ESWC2020)*. 164–167. DOI:https://doi.org/10.1007/978-3-030-62327-2_28

[98] Mir-Abolfazl Mostafavi, Geoffrey Edwards, and Robert Jeansoulin. 2004. An ontology-based method for quality assessment of spatial data bases. In *Third International Symposium on Spatial Data Quality*, 49–66.

[99] Bojanand Longo Luca Nayak Aparnaand Božić. 2021. Linked Data Quality Assessment: A Survey. In *Web Services - ICWS 2021 - 28th International Conference, Held as Part of the Services Conference Federation, SCF 2021*, 63–76. DOI:https://doi.org/10.1007/978-3-030-96140-4_5

[100] Antonio De Nicola and Michele Missikoff. 2016. A Lightweight Methodology for Rapid Ontology Engineering. *Commun. ACM* 59, 3 (February 2016), 79–86. DOI:https://doi.org/10.1145/2818359

[101] Tom De Nies, Anastasia Dimou, Ruben Verborgh, Erik Mannens, and Rik de Walle. 2015. Enabling dataset trustworthiness by exposing the provenance of mapping quality assessment and refinement. In *METHOD 2015: The 4th International Workshop on Methods for Establishing Trust of (Open) Data*, 1–6.

[102] 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* (2017). DOI:https://doi.org/10.1177/1609406917733847

[103] Natalya F Noy, Deborah L McGuinness, and others. 2001. *Ontology development 101: A guide to creating your first ontology*. Stanford knowledge systems laboratory technical report. Retrieved April 1, 2023 from https://protege.stanford.edu/publications/ontology_development/ontology101.pdf

[104] Alexandre Passant and Pablo N Mendes. 2010. sparqlPuSH: Proactive Notification of Data Updates in RDF Stores Using PubSubHubbub. In *6th Workshop on Scripting and Development for the Semantic Web*, 56–66.

[105] Heiko Paulheim and Christian Bizer. 2014. Improving the Quality of Linked Data Using Statistical Distributions. *Int. J. Semant. Web Inf. Syst.* 10, (2014), 63–86. DOI:https://doi.org/10.4018/ijswis.2014040104

[106] Karl Pearson. 1895. Notes on the history of correlation. *Proc. R. Soc. London* 58, (1895), 240–242.

[107] Christoph Pinkel, Carsten Binnig, Peter Haase, Clemens Martin, Kunal Sengupta, and Johannes Trame. 2014. How to Best Find a Partner? An Evaluation of Editing Approaches to Construct R2RML Mappings. In *11th International Extended Semantic Web Conference (ESWC 2014)*, 675–690. DOI:https://doi.org/10.1007/978-3-319-07443-6_45

[108] Leo L Pipino, Yang W Lee, and Richard Y Wang. 2002. Data quality assessment. *Commun. ACM* 45, 4 (2002), 211–218. DOI:https://doi.org/10.1145/505248.506010

[109] Niko Popitsch and Bernhard Haslhofer. 2011. DSNotify - A solution for event detection and link maintenance in dynamic datasets. *J. Web Semant.* 9, 3 (2011), 266–283. DOI:https://doi.org/10.1016/j.websem.2011.05.002

[110] María Poveda-Villalón, Alba Fernández-Izquierdo, Mariano Fernández-López, and Raúl García-Castro. 2022. LOT: An industrial oriented ontology engineering framework. *Eng. Appl. Artif. Intell.* 111, (2022), 104755.

DOI:https://doi.org/10.1016/j.engappai.2022.104755

[111] María Poveda-Villalón, Asunción Gómez-Pérez, and Mari Carmen Suárez-Figueroa. 2014. OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation. *Int. J. Semant. Web Inf. Syst.* (2014), 7–34.

[112] Maria Poveda-Villalón, Bernard Vatant, Mari Carmen Suárez-Figueroa, and Asunción Gómez-Pérez. 2013. Detecting Good Practices and Pitfalls When Publishing Vocabularies on the Web. In *Proceedings of the 4th International Conference on Ontology and Semantic Web Patterns - Volume 1188* (WOP'13), 39–51.

[113] Dave Raggett, Arnaud Le Hors, Ian Jacobs, and others. 1999. HTML 4.01 Specification. *World Wide Web Consortium (W3C) Recommendation 24*. Retrieved April 1, 2023 from https://www.w3.org/TR/html401/

[114] Rémi Rampin and Vicky Rampin. 2021. Taguette: open-source qualitative data analysis. *J. Open Source Softw.* 6, 68 (2021), 3522. DOI:https://doi.org/10.21105/joss.03522

[115] Alex Randles and Declan O'Sullivan. 2021. Assessing quality of R2RML mappings for OSi's Linked Open Data portal. In *4th International Workshop on Geospatial Linked Data (GeoLD) co-located with 18th Extended Semantic Web Conference (ESWC)*, 51–58.

[116] Alex Randles and Declan O'Sullivan. 2022. Modeling & Analyzing Changes within LD Source Data. In *8th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW) co-located with the 21st International Semantic Web Conference (ISWC 2022)*, 19–27.

[117] Alex Randles and Declan O'Sullivan. 2023. Preserving the Alignment of LD with Source Data. In *Proceedings of the 4th International Workshop on Knowledge Graph Construction (KGCW) co-located with the 20th Extended Semantic Web Conference*.

[118] Alex Randles, Declan O'Sullivan, John Keeney, and Liam Fallon. 2022. Applying a Mapping Quality Framework in Cloud Native Monitoring. In *18th International Conference on Semantics Systems (SEMANTiCS)*, 183–188.

[119] Alex Randles, Declan O'Sullivan, John Keeney, and Liam Fallon. 2023. Ontology Driven Closed Control Loop Automation. In *3rd International Workshop on Intent-Based Networking (WIN'2023) co-located with the 9th International Conference on Network Softwarization (NetSoft)*.

[120] Julio Cesar Dos Reis, Cédric Pruski, and Chantal Reynaud-Delaître. 2015. State-of-the-art on mapping maintenance and challenges towards a fully automatic approach. *Expert Syst. Appl.* 42, 3 (2015), 1465–1478. DOI:https://doi.org/10.1016/j.eswa.2014.08.047

[121] Julio Cesar Dos Reis, Cédric Pruski, Marcos Da Silveira, and Chantal Reynaud-Delaître. 2015. DyKOSMap: A framework for mapping adaptation between biomedical knowledge organization systems. *J. Biomed. Inform.* 55, (2015), 153–173. DOI:https://doi.org/10.1016/j.jbi.2015.04.001

[122] Mariano Rico, Nandana Mihindukulasooriya, Dimitris Kontokostas, Heiko Paulheim, Sebastian Hellmann, and Asunción Gómez-Pérez. 2018. Predicting Incorrect Mappings: A Data-Driven Approach Applied to DBpedia. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (SAC '18), 323–330. DOI:https://doi.org/10.1145/3167132.3167164

[123] Edna Ruckhaus, Oriana Baldizán, Maria-Esther Vidal, and Oriana Baldizán. 2014. Analyzing Linked Data Quality with LiQuate. In *11th Extended Semantic Web Conference (ESWC 2014)*, 629–638. DOI:https://doi.org/10.1007/978-3-642-41033-8_80

[124] Anisa Rula, Matteo Palmonari, and Andrea Maurino. 2012. Capturing the age of linked open data: Towards a dataset-independent framework. In *IEEE 6th International Conference on Semantic Computing (ICSC 2012)* , 218–225. DOI:https://doi.org/10.1109/ICSC.2012.17

[125] Jeff Sauro and Erika Kindlund. 2005. A method to standardize usability metrics into a single score. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 401–409. DOI:https://doi.org/10.1145/1054972.1055028

[126] Ryan Shaw, Raphaël Troncy, and Lynda Hardman. 2009. LODE: Linking open descriptions of events. In *Asian semantic web conference*, 153–167. DOI:https://doi.org/10.1007/978-3-642-10871-6_11

[127] Saeedeh Shekarpour and S D Katebi. 2010. Modeling and evaluation of trust with an extension in semantic web. *J. Web Semant.* 8, 1 (2010), 26–36. DOI:https://doi.org/10.1016/j.websem.2009.11.003

[128] Tong Shen, Fu Zhang, and Jingwei Cheng. 2022. A Comprehensive Overview of Knowledge Graph Completion. *Know.-Based Syst.* 255, C (2022). DOI:https://doi.org/10.1016/j.knosys.2022.109597

[129] Kartik Shenoy, Filip Ilievski, Daniel Garijo, Daniel Schwabe, and Pedro Szekely. 2022. A study of the quality of Wikidata. *J. Web Semant.* 72, 100679 (2022). DOI:https://doi.org/https://doi.org/10.1016/j.websem.2021.100679

[130] Pavel Shvaiko. 2014. Ontology Matching. In *Encyclopedia of Social Network Analysis and Mining*. 1209–1211. DOI:https://doi.org/10.1007/978-1-4614-6170-8_123

[131] Anuj Singh, Rob Brennan, and Declan O'Sullivan. 2018. DELTA-LD: A Change Detection Approach for Linked Datasets. In *4th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW) co-located with the 15th Extended Semantic Web Conference (EWSC 2018)*, 1–15.

[132] C Spearman. 1904. The proof and measurement of association between two things. *Am. J. Psychol.* 15, (1904), 72–101. DOI:https://doi.org/10.2307/1412159

[133] Sebastian Speiser and Andreas Harth. 2010. Taking the lids off data silos. In *Proceedings of the 6th International Conference on Semantic Systems (SEMANTICS)*, 1–4. DOI:https://doi.org/10.1145/1839707.1839761

[134] Asunciónand Fernández-López Mariano Suárez-Figueroa Mari Carmenand Gómez-Pérez. 2012. The NeOn Methodology for Ontology Engineering. In *Ontology Engineering in a Networked World*. 9–34. DOI:https://doi.org/10.1007/978-3-642-24794-1_2

[135] Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, and Boris Villazón-Terrazas. 2009. How to write and use the ontology requirements specification document. In *On the Move to Meaningful Internet Systems (OTM) Confederated International Conferences*, 966–982. DOI:https://doi.org/10.1007/978-3-642-05151-7_16

[136] Mohammad Taye. 2010. Understanding Semantic Web and Ontologies: Theory and Applications. *J. Comput.* 2, (2010), 182–192. DOI:https://doi.org/10.48550/arXiv.1006.4567

[137] Olivier Théreaux. 2003. Common HTTP implementation problems. *World Wide Web Consortium (W3C) Note*, 49. Retrieved April 1, 2023 from https://www.w3.org/TR/chips/

[138] Jürgen Umbrich, Boris Villazön-Terrazas, and Michael Hausenblas. 2010. Dataset dynamics compendium: a comparative study. In *1st International Conference on Consuming Linked Data-Volume 665*, 49–60.

[139] J ¨ Urgen Umbrich, Michael Hausenblas, Aidan Hogan, Axel Polleres, and Stefan Decker. 2010. Towards Dataset Dynamics: Change Frequency of Linked Open Data Sources. In *3rd International Workshop on Linked Data on the Web (LDOW2010) co-located with 19th International World Wide Web Conference*, 56–63.

[140] Guido VanRossum and Fred L Drake. 2010. The Python Language Reference. Retrieved April 1, 2023 from http://marvin.cs.uidaho.edu/Teaching/CS515/pythonReference.pdf

[141] Vânia Vidal, Narciso Arruda, Matheus Cruz, Marco Casanova, Carlos Brito, and Valéria Pequeno. 2017. Computing changesets for RDF views of relational data. In *Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW 2017)*, 43–58.

[142] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. 2009. Discovering and Maintaining Links on the Web of Data. In *8th International Semantic Web Conference (ISWC 2009)*, 650–665.

DOI:https://doi.org/10.1007/978-3-642-04930-9_41

[143] Marko Vujasinovic, Nenad Ivezic, and Boonserm Kulvatunyou. 2015. A survey and classification of principles for domain-specific ontology design patterns development. *Appl. Ontol.* 10, 1 (2015), 41–69. DOI:https://doi.org/10.3233/AO-150140

[144] Xiangyu Wang, Lyuzhou Chen, Taiyu Ban, M Usman, Yifeng Guan, Shikang Liu, Tianhao Wu, and Huanhuan Chen. 2021. Knowledge Graph Quality Control: A Survey. *Fundam. Res.* 1, (2021), 607–626. DOI:https://doi.org/10.1016/j.fmre.2021.09.003

[145] Gerhard Weikum. 2021. Knowledge Graphs 2021: A Data Odyssey. *Proc. VLDB Endow.* 14, 12 (2021), 3233–3238. DOI:https://doi.org/10.14778/3476311.3476393

[146] Dominik Wienand and Heiko Paulheim. 2014. Detecting incorrect numerical data in dbpedia. In *The Semantic Web: Trends and Challenges: 11th Extended Semantic Web Conference (ESWC 2014)*, 504–518. DOI:https://doi.org/10.1007/978-3-319-07443-6_34

[147] Mark Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gaby Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Olavo da Silva Santos, Philip Bourne, Jildau Bouwman, Anthony Brookes, Tim Clark, Merce Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris Evelo, Richard Finkers, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, (2016). DOI:https://doi.org/10.1038/s41597-019-0009-6

[148] Peiran Yao and Denilson Barbosa. 2021. Typing Errors in Factual Knowledge Graphs: Severity and Possible Ways Out. In *The Web Conference 2021 (WWW '21)*, 3305–3313. DOI:https://doi.org/10.1145/3442381.3449977

[149] Amrapali Zaveri, Dimitris Kontokostas, Mohamed A Sherif, Lorenz Bühmann, Mohamed Morsey, Sören Auer, and Jens Lehmann. 2013. User-driven Quality Evaluation of DBpedia. In *9th International Conference on Semantic Systems (SEMANTICS 2013)*, 97–104. DOI:https://doi.org/10.1145/2506182.2506195

[150] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. 2015. Quality assessment for Linked Data: A Survey. *Semant. Web* 7, (2015), 63–93. DOI:https://doi.org/10.3233/SW-150175

[151] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, Sören Auer, and Pascal Hitzler. 2013. Quality assessment methodologies for linked open data. *Semant. Web J.* 1, (2013), 1–5. DOI:https://doi.org/10.3233/SW-150175

[152] Baifanand Zhou Dongzhuoranand Cheng Gongand Jiménez-Ruiz Ernestoand Soylu Ahmetand Kharlamov Evgeny Zheng Zhuoxunand Zhou. 2022. Query-Based Industrial Analytics over Knowledge Graphs with Ontology Reshaping. In *19th Extended Semantic Web Conference (ESWC 2022)*, 123–128. DOI:https://doi.org/10.1007/978-3-031-11609-4_23

[153] Mussab Zneika. 2019. Querying Semantic Web/Linked Data Graphs Using Summarization. PhD Thesis. Cergy-Pontoise University, France. Retrieved April 1, 2023 from https://tel.archives-ouvertes.fr/tel-02861761

[154] Protégé. Retrieved January 1, 2022 from https://protege.stanford.edu/

[155] Prometheus. Retrieved January 1, 2022 from https://prometheus.io/

[156] 2010. Dataset Dynamics (dady) Vocabulary. Retrieved June 14, 2022 from http://purl.org/NET/dady

[157] 2019. TARQL: SPARQL for Tables. Retrieved December 8, 2022 from http://tarql.github.io/

Appendix A. Manual Examination Checklist

The checklist used to manually validate the quality of mappings in experiment 1 is presented below.

Mapping Quality Aspect (MP)

- **MP1 Valid logical table definition**
  - Does the mapping contain one of the following properties: `rr:logicalTable` or `rml:logicalSource`?
- **MP2 Valid subject map definition**
  - Does the mapping contain the following property: `rr:subjectMap`?
- **MP3 Valid predicate object map definition**
  - Does the mapping contain a predicate object map (`rr:predicateObjecMap)` ?
  - Does it contain at least one predicate (`rr:predicate`) and one object (`rr:object`) or object map (`rr:objectMap`) ?
- **MP4 Valid language datatype definition**
  - Is a language tag (`rr:language`) and datatype (`rr:datatype`) defined?
  - Are these defined in the same object map (`rr:objectMap`) ?
- **MP5 Valid parent triple maps definition**
  - Is a join (`rr:joinCondition`) defined?
  - Does the join include the `rr:child` property?
  - Does the join include the `rr:parent` property?
- **MP6 Valid term type definition**
  - Does the object map contain a term type?
  - Is the term type one of the following: `rr:IRI`, `rr:BlankNode`, `rr:Literal`?
  - Does the subject map contain a term type?
  - Is the term type one of the following: `rr:IRI`, `rr:BlankNode`?
  - Does the predicate map contain a term type?
  - Is the term type `rr:IRI`?
- **MP7 Valid subject definition**

  - Is the subject identifier (`rr:class`) a valid IRI[71]?
- **MP8 Valid predicate definition**
  - Does each predicate (`rr:predicate`) have a valid IRI?
- **MP9 Valid named graph definition**
  - Does each named graph (`rr:graph`) have a valid IRI?
- **MP10 Valid datatype definition**
  - Are there datatypes defined?
  - Does each defined datatype (`rr:datatype`) have a valid IRI?
- **MP11 Valid literal language tags**
  - Are there language tags defined?
  - Are the tags defined in RFC 5646 (BCP 47)?
- **MP12 Duplicate triples defined**
  - Do term maps exist with the same subjects (`rr:subjectMap`), predicates (`rr:predicateObjectMap`) and objects (`rr:objectMap`)?

Data Quality Aspect (D)

- **D1 Usage of undefined classes**
  - Are the classes ontology accessible?
  - Is each class defined in accessible ontologies?
- **D2 Usage of undefined properties**
  - Are the properties ontology accessible?
  - Is the class defined in the ontology?
- **D3 Usage of incorrect domain**
  - Are the properties ontology accessible?
  - Is the class defined in the ontology?
- **D4 No query parameters in URI's**
  - Does any property or class have a "?" in the URI?
- **D5 No use of entities as members of disjoint classes**
  - Are the classes ontology accessible?
  - Is the class disjoint (`owl:disjointWith`) any other ontologies?

[71] Validation of IRIs was completed by ensuring conformance to syntax outlined at https://www.w3.org/International/iri-edit/draft-duerst-iri.html

  - Is the disjoint class defined in the mapping?
- **D6 Usage of incorrect range**
  - Does the mapping contain a term type?
  - Does the term type correspond to the type of property i.e data type or object
- **D7 Usage of incorrect datatype**
  - Does the mapping contain a data type (`rr:datatype`)?
  - Does the data type match the respective property range (`rdfs:range`)?

Vocabulary Quality Aspect (VOC)

- **VOC1 Human readable labels/comment**
  - Is the property or class defined?
  - Does the respective property contain a triple in the ontology with one of the following properties: `rdfs:label, dcterms:title, dcterms:description, dcterms:alternative, skos:altLabel, skos:prefLabel, powder-s:text, skosxl:altLabel, skosxl:hiddenLabel, skosxl:prefLabel, skosxl:literalForm, rdfs:comment, schema:description, schema:description, foaf:name`
- **VOC2 Domain and range definitions**
  - Is the property defined?
  - Does the respective property contain a triple in the ontology with one of the following properties: `rdfs:domain`, `rdfs:range`
- **VOC3 Machine readable licensing**
  - Is the namespace accessible?
  - Does the ontology associated with the namespace contain a triple with one of the following properties: `dct:license, dct:rights, dc:rights, xhtml:license, cc:license, dc:licence, doap:license, schema:license`
- **VOC4 Human readable licensing**
  - Is the namespace accessible?
  - Does the ontology associated with the namespace contain a triple with one of the following properties: `dct:description, rdfs:comment, rdfs:label, schema:description`
  - And an object matching the following regular expression: `‟.*(licensed?|copyrighte?d?).*(under|grante?d?|rights?).*”`
- **VOC5 Basic provenance information**
  - Is the namespace accessible?

- Does the ontology associated with the namespace contain a triple with one of the following properties: `dc:creator`, `dc:publisher`, `dct:creator`, `dct:contributor`, `dcterms:publisher`, `dc:title`, `dc:description`, `rdfs:comment`, `foaf:maker`

Appendix B. PSSUQ

The table below presents the questions of the Post-Study System Usability Questionnaire (PSSUQ).

#	Question
1	Overall, I am satisfied with how easy it is to use this system
2	It was simple to use this system
3	I could effectively complete the tasks and scenarios using this system
4	I was able to complete the tasks and scenarios quickly using this system
5	I was able to efficiently complete the tasks and scenarios using this system
6	I felt comfortable using this system
7	It was easy to learn to use this system
8	I believe I could become productive quickly using this system
9	The system gave error messages that clearly told me how to fix problems
10	Whenever I made a mistake using the system, I could recover easily and quickly
11	The information (such as on-line help, on-screen messages, and other documentation) provided with this system was clear
12	It was easy to find the information I needed
13	The information provided for the system was easy to understand
14	The information was effective in helping me complete the tasks and scenarios
15	The organization of information on the system screens was clear
16	The interface of this system was pleasant
17	I liked using the interface of this system
18	This system has all the functions and capabilities I expect it to have
19	Overall, I am satisfied with this system

Appendix C. Understanding Questionnaire

The table below presents the questions of the understanding questionnaire.

#	Section 1	Section 2
1	How many total changes were detected between the source data files?	A "Referenced Data Change" is one of the following:
2	How many mappings were impacted from the source data changes?	How many columns have been inserted in the source data?

3	A threshold is one of the following: a) A threshold defines when changes are detected. b) A threshold defines when users will be notified of source data changes. c) A threshold defines where the notification is sent. d) A threshold defines the types of changes which can be detected.	Select two values which have been inserted into the "FirstName" column in the source data. ▪ Tom ▪ Bob ▪ Richard ▪ Michael ▪ Lewis
4	How many total changes were included in the thresholds?	How many columns have been deleted in the source data?
5	What is the threshold for insert changes?	Which column has been deleted in the source data? ▪ FirstName ▪ LastName ▪ Name ▪ Age
6	Select the two data references in Mapping #1: ▪ ID ▪ Age ▪ Gender ▪ Name ▪ Job	How many total values have been inserted into the "ID" data reference?

Appendix D. Ontology Design Questionnaire

The table below presents the questions of the ontology design questionnaire.

#	Question
Q1(a)	In your opinion, does the design of OSCD correctly follow best practices in ontology design?
Q1(b)	Can you provide a reason for your response?
Q2(a)	Do you suggest any alterations to the design methodology followed by OSCD?
Q2(b)	Can you provide a reason for your response?
Q3(a)	Do you suggest any alterations to the concepts/relationships in OSCD?
Q3(b)	Can you provide a reason for your response?
Q4(a)	Do you suggest any alterations to the documentation of OSCD?
Q4(b)	Can you provide a reason for your response?
Q5	Any additional comments?

Appendix E. Categories of Quality metrics

The table below presents quality categories in the linked data domain.

Category	Description
Intrinsic (IR)	The dimensions that are independent of the user's context. In the mapping process context, these dimensions focus on whether the mapping is correct (syntactically and semantically),

	whether the mapping is consistent in itself, and how complete it represents the data being mapped and the vocabularies being used.
Representational (RP)	Concerned with the design of the data. In other words, metrics in this category evaluate how well the data is represented in terms of best practices and guidelines. The dimensions in this category are mostly focused on the quality of the resulting datasets produced by mappings.
Contextual (CT)	Groups dimensions and metrics highly depend on the context of the task at hand. The dimensions for this category deal with trustworthiness and understandability. The dimensions and metrics within this category do not specify how (i.e. which vocabularies, etc.) mappings and resulting datasets must define trustworthiness and understandability information. Nonetheless, existing work has shown how such information can be incorporated into mappings, which is argued to improve trustworthiness and understandability of mappings and datasets. Thus, this category provides quality metrics related to detect whether such information is present.
Accessibility (AC)	Groups dimensions and metrics related to the access, authenticity and retrieval of data. The dimensions for this category deal with licensing and availability and are derived from previous work on capturing mapping provenance and metadata information.

Appendix F. Scenario for Think-Aloud Test

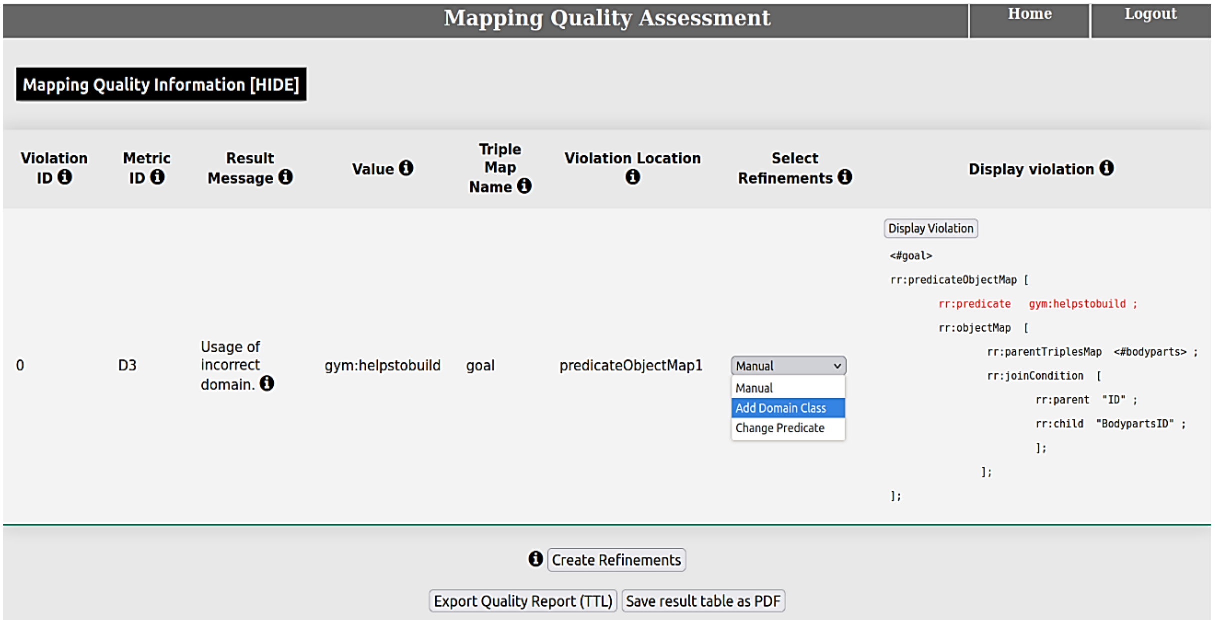
Scenario:

You are a mapping engineer working for a company designed to manage student records. You have created an R2RML mappings to transform data stored in the “DATASETS” table of their relational database. These mappings transform the time the data was generated and a representation of the data using the Provenance Ontology (PROV-O), which contains suitable concepts and properties.

In order to ensure this mapping is suitable for execution you must complete quality assessment and refinement on the artefact. This process will involve using a mapping quality framework for the procedure, named the “Mapping Quality Improvement (MQI) Framework”. First, you will upload the mapping to the framework. Thereafter, you will examine the quality assessment information generated by the framework. Finally, you will refine the quality issues detected in the mapping using semi-automatic refinements, resulting in an improved refined mapping, which will hopefully be free of quality issues.

Appendix G. Refinement Value Suggestions for Experiment 1

The screenshot below presents suggested refinements for the violation detected by the framework.



Violation ID	Metric ID	Result Message	Value	Triple Map Name	Violation Location	Select Refinements	Display violation
0	D3	Usage of incorrect domain.	gym:helpstobuild	goal	predicateObjectMap1	Manual	Display Violation

```
<#goal>
rr:predicateObjectMap [
        rr:predicate   gym:helpstobuild ;
        rr:objectMap  [
                rr:parentTriplesMap  <#bodyparts> ;
                rr:joinCondition  [
                        rr:parent  "ID" ;
                        rr:child  "BodypartsID" ;
                        ];
                ];
];
```

As can be seen, 2 semi-automatic refinements ("Add Domain Class", "Change Predicate") and one manual refinement ("Manual") are suggested by the framework to resolve the issue.

Appendix H. Email Invitation used for Experiment 2

Hi <Name>,

This usability experiment should take roughly 30 minutes to complete and should be completed all at once.

The details for the usability experiment of the Mapping Quality framework are as follows:

To login to the framework:

Login details

URL: https://mqv-framework.adaptcentre.ie/

Participant ID: [ID]

Password: [Password]

If you have any problems, please contact me at alex.randles@adaptcentre.ie

Thank you for participating in the experiment.

Kind Regards,

Alex Randles.

Appendix I. Experiment 2 Informed Consent

TRINITY COLLEGE DUBLIN

INFORMED CONSENT

LEAD RESEARCHERS: Alex Randles

BACKGROUND OF RESEARCH: The Knowledge and Data Engineering Group (KDEG) is located within Trinity College and focuses on researching challenges posed by knowledge discovery, representation and engineering. This knowledge is often represented as RDF data. Creating RDF data requires the creation of mappings, which are definitions that describe the transformation from source to target data. Our research will lead to finding a better way to improve the quality of the mappings being produced.

Creating these mappings has a steep learning curve, which can result in poor quality mappings being generated. These poor-quality mappings can often be unnoticed as the state of the art in linked data quality focuses on the published datasets which are generated by these mappings. These mapping quality issues can grow exponentially within the resulting dataset. We aim to address the challenge of quality issues by bringing quality assessment earlier into the publication stage. We have designed the Mapping

Quality Vocabulary (MQV) framework to facilitate the quality assessment and refinement of these mappings, which involves a human-in-the-loop, who guides the decisions taken during the quality refinement process within the framework. The framework is hoped to improve the quality of mappings, thus improving the quality of the datasets generated by these mappings.

PROCEDURES OF THIS STUDY:

- You are going to be briefed on the experiment task and what to do in the experiment.
- You will be asked to use the MQV framework and perform some tasks while thinking aloud.
- You will be asked to fill out a usability test survey.

This experiment will take place online over a conference call via a video conferencing platform, with access to the local computer of the lead researcher to test the framework.

The total duration should take less than an hour to perform the tasks and fill in the questionnaire. We will track the time you spent in the completion of each task with a stopwatch. While thinking aloud, you will be recorded with an automatic transcription feature of this video conferencing platform. This transcription will be used to correct the statements that the note taker will write down during the experiment. Audio and video will not be recorded during your session. The resulting data stored will be summary tables with the time per task and the numeric answers of the usability test survey, and texts files with the open comments of the usability test survey and the automatic transcriptions of the experiment session.

Your data (time per task, transcription and the usability test survey) will not be identifiable since it will be coded with a participant ID and stored using the IT services called MyZone Google Drive which complies with GDPR rules. The lead researcher (Alex Randles) and the supervisor (Prof. Declan O'Sullivan) will be the only people with access to the data until its publication in an open data repository.

We will perform a qualitative analysis of the think-aloud data by coding and categorizing the statements, once we have the aggregated data from all the participants. Then, resulting emerging themes will be the ones reported as the results of this experiment. Furthermore, the quantitative results from the time per task and the usability test survey will be analysed with statistical summaries, reporting aggregated results. None of your personal details will be recorded and you are free to stop and leave the experiment at any point if you so choose.

PUBLICATION: The goal will be to publish the results of the usability test at Semantic Web conference and workshops, such as Extended Semantic Web Conference (ESWC) and International Semantic Web Conference (ISWC); and other relevant journals, as well as the PhD thesis of the lead researcher at Trinity College Dublin.

CONFLICTS OF INTEREST:

My supervisors will not take part in the experiment. Potential participants of the experiment will not be provided with any prior information before the experiment.

Individual results will be anonymized and published in open data repositories for reproducibility and research will be reported on the aggregate results.

DECLARATION:

- ○ I am 18 years or older and am competent to provide consent.
- ○ I have read, or had read to me, a document providing information about this research and this consent form. I have had the opportunity to ask questions and all my questions have been answered to my satisfaction and understand the description of the research that is being provided to me.
- ○ I agree that my data is used for scientific purposes and I have no objection that my data is published in scientific publications in a way that does not reveal my identity.
- ○ I understand that if I make illicit activities known, these will be reported to appropriate authorities.
- ○ I understand that I may refuse to answer any question and that I may withdraw at any time without penalty.
- ○ I understand that if the results of the research have been published, <or my data has been fully anonymised so that it can no longer be attributed to me>, then it will no longer be possible to withdraw.
- ○ I understand that I may stop electronic recordings at any time, and that I may at any time, even subsequent to my participation [request to] have such recordings destroyed (except in situations such as above).
- ○ I understand that, subject to the constraints above, no recordings will be replayed in any public forum or made available to any audience other than the current researchers/research team.
- ○ I freely and voluntarily agree to be part of this research study, though without prejudice to my legal and ethical rights.
- ○ I understand that my participation is fully anonymous and that no personal details about me will be recorded.
- ○ I understand that if I or anyone in my family has a history of epilepsy then I am proceeding at my own risk.
- ○ I understand that personal information about me, including the transfer of this personal information about me outside of the EU, will be protected in accordance with the General Data Protection Regulation.
- ○ I have received a copy of this agreement.

By signing this document, I consent to participate in this study, and consent to the data processing necessary to enable my participation and to achieve the research goals of this study.

PARTICIPANT'S NAME:

PARTICIPANT'S SIGNATURE:

Date:

Statement of investigator's responsibility: I have explained the nature and purpose of the research study, the procedures to be undertaken and any risks that may be involved. I have offered to answer any questions and fully answered such questions. I believe that the participant understands my explanation and has freely given informed consent.

RESEARCHERS CONTACT DETAILS: alex.randles@adaptcentre.ie

RESEARCHER'S SIGNATURE: Alex Randles

Date:

Appendix J. Sample Think-Aloud Test

Participant 62 [00:00:00] So first task upload the provided mapping to the framework.

Participant 62 [00:00:11] So you will be done when the mapping file has been successfully uploaded and you have pressed the assess mapping quality button. Ok so uploaded mapping. Assess mapping quality button. Right. So I assume that's task one done.

Participant 62 [00:01:20] So task two explore the mapping quality assessment information generated by the framework you'll be done when you have acknowledged the number of violations, their result message, value, location, refinements and display the violation. Okay, so there appears to be three violations. Okay, so just checking out what each of these mean. So this this one here. So properties are considered undefined when it is not possible to dereference them against the respective namespaces.

Participant 62 [00:03:10] You will be done when you acknowledged the number of violations, their result message, value, location and display it. Okay. So I think I have completed task two.

Participant 62 [00:03:35] So task three. Explore more information related to each violation and result message, you will be done when you acknowledge the usefulness of this information. You know, I think it's very useful information, actually, because just pointing out. Exactly. What the problem is with the violation. So it's even pointing toward saying like this value doesn't exist within the respective namespace or even in the prov ontology. And this is not a valid data type for the range of that property. Pretty cool. I think that's task three done.

Participant 62 [00:04:34] Moving onto task four. Select a refinement for each for violation which you consider the most appropriate. The refinement must not be manual. You'll be doing when you're selected a refinement for each violation and pressed the create refinement button. So you obviously do refinement through here. The problem is that it's not a defined language tag. Therefore, I'm going to say change language. That's where the values is not a prov property. Okay. I am going to select find similar predicates for that one. The data type assigned to the objects does not match the data type for this range property. Right. This, to me, looks like the best refinements to select in each of these different cases. Create refinements button. And what I'm going to do here actually is just say display violation. Ah, okay. So obviously just displays that little snippet of code of the mapping where it occurs that is very cool. So create refinements. Okay. So I think this task four complete.

Participant 62 [00:06:29] Task five. Enter values for each refinement if required and select all refinements to be executed. You'll be done when you have entered the values for each refinement. I am going to leave that as a general English tag. Rather than going for a specific flavor. This one here. This is where it's providing me. Related. I'm actually going to go back and look at this mapping and I am going to choose prov value in this case. Yeah so look I think that's actually task five done. So I've put in the different values for the different refinements. So I'd say that's task five done, let me just read through it again, just to make sure.

Participant 62 [00:08:46] So onto task six select all refinements to be executed. You'll be done once you've pressed the Select Refinement button. Okay, so you tick these boxes to select them and execute refinements. So I obviously jumped through a step there. I didn't press select all refinements, but I did select each of them individually, and then I pressed the execute refinements button. So that's task six and seven done.

Participant 62 [00:09:36] So onto task eight explore the mapping quality profile bar chart generated. You'll be done when you have acknowledged the violation count for each quality dimension. Okay, So each different type of quality dimension is color coded and you hover over it to see the numbers that exist. The violation counts and a bit of a description as well as to what went wrong. Sorry a bit of a description of the quality dimension. Okay, so let's say that's task eight done.

Participant 62 [00:10:33] Moving onto task nine, export to refine mapping. You'll be done when you successfully exported the refined mapping. Export refined mapping. So that was obviously blocked by my browser as a pop up. So you have to allow that. Download the mapping. Have a look at it. Just from having a look at it, it looks like all the changes are made to it. So open it and examine the refined mapping you will be done when you have acknowledged that the refined mapping has been generated correctly. So that's task nine and ten complete.

Participant 62 [00:11:36] Task eleven. Export the validation report. So that is task eleven done. Task twelve. Examine the validation report you'll be done when you've acknowledged the number of violations and their associated refinements. Let's. Okay, very cool. So you obviously have a machine-readable validation report that you can query over yourself and do some analysis. I like this idea written as the query that you use to make the changes to the mappings. Pretty cool. All right. I think that's well done.

Appendix K. Description of Thematic Themes and Codes

The table below presents descriptions of the themes and codes defined as a result of thematic analysis.

Theme	Theme description	Code	Code description
User friendly	The framework was easy to use and understand.	**Easy to use**	The framework was easy to use.
		Efficient	The tasks could be completed with minimal effort.
		Intuitive	The tool provided guidance in completing the tasks.
Positive user experience	Positive user experience while interacting with the framework.	**Straightforward**	The tasks were easy to complete.
		Error free	No errors were encountered during the completion of the tasks.
		Adequate error recovery	The tool provided sufficient ability to recover from errors.
		Quicker and easier to use over time	Quicker and easier to use over time
Positive GUI Requirements	The layout and aesthetics of the framework are sufficient.	**Aesthetic interface**	The interface is aesthetically pleasing.
		Clear interface navigation	Guidance provided by the framework interface is easy to understand.
		Clear layout	The structure of the framework is easy to understand.
		Responsive Design	The framework resizes and adjusts to different screen sizes.
Negative GUI Requirements	The layout and aesthetics of the framework are not sufficient.	**Unclear interface navigation**	Guidance provided by the framework interface was hard to understand.
		Unaesthetic interface	The interface is not aesthetically pleasing.
		Unclear layout	The structure of the framework is hard to understand.
Useful	Functionality of the framework which was useful.	**Overall usefulness**	The framework has a practical purpose.
		Drop-downs useful	Drop downs helped to guide users.
		Validation report useful	The report represented in MQIO provided beneficial quality information.
		Violation RDF visualization useful	The visualization of the RDF extract from the mapping helps explain the mapping violation.

		Tool tips useful	Information provided by tool tips help to guide users.
		Error messages useful	Information provided by error messages helps to guide users.
Clarify description and features	Overly complicated and ambiguous text displayed on the framework.	**Clarify text descriptions**	Text descriptions need to be further described.
		Verbose instructions	The instructions can be condensed.
		Ambiguous error message	The error messages are not described adequately.
		Task instructions clarified	Task wording and structure is difficult to understand.
		Ambiguous refinement options	The refinement options for a violation are not described adequately.
		Additional information required	Information provided by framework alone is not sufficient.
Technical errors	Technical errors which occurred during the completion of the experiment tasks.	**Missing tool tip text description**	Tool tip text description is not present for a feature.
		PDF export error	The PDF version of the validation report could not be exported.
		Permission error	Pop-ups from the framework were blocked.
Provenance usability	Provenance information provided by the framework is not clear and concise.	**Provenance representation**	The provenance information provided by the framework is poorly structured.

Appendix L. PSSUQ Comments from Experiment 2

The table below presents the grouped open comments received from the PSSUQ in experiment 2.

PSSUQ Metric	Comments
System usefulness (Q1-8)	• I was somewhat confused by the third violation as I didn't fully understand why there was a problem. I think this is due to my lack of knowledge though. • easy to use, • The presentation beforehand made it easy to use the system • Everything worked well with the example, would need a larger, more complicated mapping to see if it would work just as well • The presentation made things faster, might have had to spend more time working with it otherwise • This is a very helpful tool for checking mappings • The error regarding the dateTime was slightly confusing in that it highlighted "rr:predicate prov:generatedAtTime;" instead of "rr:datatype xsd:time;", but the textual message cleared up the issue • I didn't have any troubke using the system, it was very intuitive • Again very simpe and easy to use • It was a very quick process!

	• it corrected all the mistakes and helped fix them easily • simple UI • very intuitive • it was very quick, so it would increase the productivity for sure! • while the errors were understandable, if someone has no experience at all. The might not understand what it exaclty means! • The system was simple to use. Just a siggesstion, if the button sizes or the text size for exporting the updatd mappings and validations could be a bit bigger, it may be more visible • The system is not so simple that all functions are self-explanatory. For the same reasons as above. It lacks the necessary design. • The system is not easy to use, but it does get the job done. • Disagree. A prerequisite for efficient performance of tasks with this system is to read the manuals and to be familiar with the system. • Agreed, it does get the job done. • Strongly disagree. This system is not a good and comfortable system in terms of interaction design, interface design or user experience. • Partially agree, as there are still some elements that confuse me, but the amazing thing is that even so, I managed to complete the task. • I don't believe it, the system is lacking in design, and design of the user experience. When I was using this system, I thought I was living in the 1990s. • The system guided me through the process, didn't face any challenges when using it • Found all the options related to the violations very reasonable • Took under 10 mins • took some time to read through all the tooltips, but when you get used to the system it obviously gets faster • when you select a new value for e.g. a datatype you have to scroll to a huge list (maybe some suggestions would help) • don't have working experience in that field, so can't comment • Yeah I think it is, I feel like after using it for an hour or so with better guidelines on its use I would feel much more confident with it. But with one example adn no tutorial video it is a bit daunting at first - a concise <10min tutorial video would be great. • see above • After reading through the relevant ontology site it was quite simple to see what had to be done. The lack of manual input was helpful too • I had to change my intial refinement • I still had to reference the ontolgy but this is expected • yes very simple • Little bit difficult to determine the correct datatype that a property should have without more context, but otherwise pretty straightforward • The error messages showed the line that an error occurred but could be more specific to show exactly what was going wrong • The system guided me through the process, didn't face any challenges when using it • Found all the options related to the violations very reasonable • Took under 10 mins • Same • I appreciate the specifity of the responses... Hope it's like that for other smples • The first time took me several tries because the page layout wasn't clear • Not very esthetically pleasing • every buttons give the straight forward directions • I was actually surprised at how quick it was • For these examples, yes; I feel like it may get more harder on larger mappings • I think so...but issues somehow always come up when larger datasets are in question

	• It was pretty clear where to click etc. • I had no issues completing the tasks. • Yes, I think I would be able to use the framework very quickly if I was using it again it the future. • I didn't necessarily make errors that forced me to see any error messaging. • No error messages that is why I answered with a 7 • For a first time user there were some features which were not always clear. • I did need a little help once or twice. • I was not sure at times how to fix the errors • It was fairly easy for me, but a lot of background knowledge is needed • Some minor issues following instructions but not the systems fault • As a first time usee I had to scan the interface a lot, there is not much emphasis on specific elements • Next time I would be, needed to understand better what the drop-down for modification really meant • Only with a lot of background knowledge • Not sure how well it handles large datasets/numbers of mappings • Some error messages felt generic and mouse-overs sometimes added little • Knowing the RDF model, I was a bit confused by some of the information provided w.r.t. to blanknodes (e.g., the artificial identifiers). I was also confused by the patterns in the validation report. • That depends on the support for triples pattern maps, hence my reservation. • n/a is missing from above, as I had no errors
Information quality (Q9-15)	• One small nitpick is that everything seemed to get larger / zoom in when I displayed the violations. • Takes a second to get used to, but after using it multiple times it would be very clear and easy to use • looks slightly too zoomed in • it is very intituive but could be more "modern" I guess? • Did not make a mistake, so can't accurately assess • The task sheet was a good reference point • Could have been more intuitive • The user interface could have been more intuitive, some tables had to be horizontally scrolled. First time users can be provided with a sequential tutorial using dialog boxes and arrows. • I couldn't clearly understand what mistakes I had made and how to deal with them, probably because the necessary user guidance was missing. • Not enough. • It can work. • The organisation of the information on the system screen looks like it was created by a part-time programmer living in the 1980s. However, I am sure that the designer of the system, did not consider the information architecture before designing the system, so all the information seems to be just spread out on the screen. • I thought this system was just for testing functionality, and hadn't thought about the user interface or UX at all, until I saw the title. This system is primitive in terms of user interface and user experience, and the prototype I spent a lunch hour on could look better than this. I'm sorry, I don't know who designed this, but it's really bad in terms of interface, really. • I didn't encounter making a mistake, or else I didn't realise I did • It looks quite outdated, but it's intuitive. Also, the huge drop-down menus (e.g. when selecting an alternative value for a datatype) are a bit difficult to navigate • since I did not make any mistakes, I'm not sure how to answer this • There was no 'hover' information for the incorrect datatype refinement (xsd:time vs

	xsd:dateTime) • Interface was responsive/easy to use, but could look nicer. • I didn't encounter making a mistake, or else I didn't realise I did • the information hover pop up, donated by the '(i)' icon was very helpful when assessing each problem • The code extract could have had taken up more screen space - but the optional display button was a satisfying feature • They ID's the error, but I do not think identifying an error means knowing how to fix it in all cases. It was here, but I would doubt it would be so in all cases • I didn't really make a mistake, however, I was unsure about the most appropriate 'type of refinement' options that I should select from the dropdown menus so it would be useful to have some guidance there (i.e. hover text etc.) • Yes, there were lots of on screen explanations/hover text • I wasn't able to find information explaining the difference between the refinement options, and I almost missed the hover text that appeared over the diagram/bar chart. Also, it would be useful to see what refinements/errors correlated to which data quality metric. • I found the data quality metric definitions to be slightly complex. Generally, some of the hover text could be simplified. • I liked the table format • Some colouring for different buttons etc. could be useful • I feel like by refreshing the page I could fix most issues • Some of the violation mapping information was hard to understand • The interface was pleasant, clean and neat. However, maybe it could be designed even better to make it more attractive to the eye. • I did need a little help. • Some items, like the use of ids, were not always clear. • The instructions were useful in case I got lost. • It was forgiving, good impl of back button etc, but I was not sure it would be like that • Hard to get much from tabular view, lacks example mappings or data instances • Fairly easy in this very tightly controlled experiment, with less specific instructions it would be easy to get lost • But I have deep knowledge of the topic • Yrs, almost too detailed as I was accidentially skipping steps • Very little is highlighted as important • Well done basic UI • Except for the chart • The page with the two tables was a bit confusing, and the bar chart after corrections would confuse me as well.
Interface quality (Q16-18)	• I may have missed it (or my solutions just didn't fix any of the violations and it remained 3 anyways), but I don't think I seen a way to check if the refinements fixed the violations or not. • The buttons like 'Execute Refinements' , 'Export report' etc were too close to task bar for me. Maybe a little bit of padding would help. I am using windows + chrome. • It is very intresting to be able to validate the mapping and resolve quality issues early. I loved the tool and its applicability. • The HTML components can be responsive and flexible to support all screens • If I need a task done urgently, and it is a task that only this tool can solve, then I will use it, because it is functional. But I would be uncomfortable, because UX ≈ 0 • It works. • I guess, I am unsure what to expect to be honest • I'm a beginner, so I don't know what to expect • When the system saw that there was an incorrect datatype, it should be able to suggest

	suggestions based on the ranges of the predicates used. • I wasnt sure what to expect but turned out useful • Would consider using this verification tool for mapping projects in the future • I think it's a really useful framework. Perhaps there could be more prompts for selecting refinements/corrections e.g. I was unsure what language tag to select because I didn't know what language the dataset was in - perhaps suggestions could be more refined based on the dataset that the mapping will be used with? That being said, the person creating the mapping is likely to be familiar with the dataset but it would still be interesting to see if the dropdown list of refinements/corrections could be refined further in some instances. Also, hover text over the list of suggested corrections would be useful e.g. hover text defining each of the prov predicates etc. might help in selecting the correct/best option. • I would add a couple of functionalities as stated during my session • It was effective • It was pretty streamlined, which I appreciated
Overall (Q19)	• System was good! • The Save as PDF option doesn't work on Chrome. Some issue with the JS library. • user friendly • This is a very useful tool for productivity • The layout is easy to understand, and the hovering messages are good for additional explanations • I really liked the dashboard at the end and the fact that i could export the validation report • It works, but it's not perfect. • The system lacks the necessary user experience design (UX). • It's okay, nice lay out and easy to use. The guidelines in the shared file on google drive could be more concise. Could be useful to provide the user with an example of how to fix the error, or a link to a relevent document, It does provide the error code which is good, but and example of the fix would be even belter if possible • it was pretty easy to use • Good User interface • I thought the system was great. • Made very easy to use with the work instruction • Very easy to use for a first time user, I found the available comments on different sections to be very helpful • It isn't very intuitive to look below to implement changes • it's easy to understand how to use it from its simple interface. • Very useful overall and easy to use • Easy to use, lots of hover/i text. Could have additional explanation for the 'type of refinement' options. • The system was useful and easy to use and I think it would really help mapping designers improving the quality of their mappings. • Simple and clean interface, very easy to use • Overall, I think I would pick up using this system, with a little practice. • On a technical level it is great, to make an "IDE for mappings" it is a long way off. It still adds huge value as is. • First time user, I am confident it will become easier • I did not know exactly what to expect, but it fits its intended purpose well • I am suspicious of how it says "change X to the correct value" ; this seems to me that it "knows" what is correct, which I am doubtful of

Appendix M. Experiment 3 Informed Consent

TRINITY COLLEGE DUBLIN

INFORMED CONSENT

LEAD RESEARCHERS: Alex Randles

BACKGROUND OF RESEARCH: The Knowledge and Data Engineering Group (KDEG) is located within Trinity College and focuses on researching challenges posed by knowledge discovery, representation and engineering. This knowledge is often represented as RDF data. Creating RDF data requires the creation of mappings, which are definitions that describe the transformation from source to target data. Our research will lead to finding a better way to improve the quality of the RDF data being produced by these mappings.

Creating these mappings has a steep learning curve, which can result in poor quality mappings being generated. These poor-quality mappings can often be unnoticed as the state of the art in RDF data quality focuses on the published datasets which are generated by these mappings. These mapping quality issues can grow exponentially within the resulting dataset. Furthermore, changes can occur within the source data used by these mappings after the RDF data has been created, resulting in inaccurate information. We aim to address the challenge by improving the quality of these mappings while capturing information related to source data changes. We have designed the Mapping Quality (MQ) framework to facilitate the quality assessment and refinement of these mappings and change detection within the source data used by these mappings. The framework is hoped to improve the quality of mappings, thus improving the quality of the datasets generated by these mappings, while providing fresh data.

PROCEDURES OF THIS STUDY:

- You are going to be briefed on the experiment task and what to do in the experiment.
- You will be asked to use the MQ framework to complete these tasks.
- You will be asked to fill out a usability test survey.

The total duration could take up to an hour to perform the tasks and fill the questionnaire. The resulting data stored will be summary tables with the numeric answers of the usability test survey and numeric ratings for the information provided by the framework.

Your data (usability test survey and information ratings) will not be identifiable and will be stored using the IT services called MyZone Google Drive which complies with GDPR rules. The lead researcher (Alex Randles) and the supervisor (Prof. Declan O'Sullivan) will be the only people with access to the data until its publication in an open data repository.

We will perform a quantitative analysis of the usability test survey using statistical summaries, reporting aggregated results. None of your personal details will be recorded and you are free to stop and leave the experiment at any point if you so choose.

PUBLICATION: The goal will be to publish the results of the usability test at Semantic Web conference and workshops, such as Extended Semantic Web Conference (ESWC) and International Semantic Web Conference (ISWC); and other relevant journals, as well as the PhD thesis of the lead researcher at Trinity College Dublin.

CONFLICTS OF INTEREST:

My supervisors will not take part in the experiment. Potential participants of the experiment will not be provided with any prior information before the experiment.

Individual results are anonymous and published in open data repositories for reproducibility and research will be reported on the aggregate results.

DECLARATION:

- I am 18 years or older and am competent to provide consent.
- I have read, or had read to me, a document providing information about this research and this consent form. I have had the opportunity to ask questions and all my questions have been answered to my satisfaction and understand the description of the research that is being provided to me.
- I agree that my data is used for scientific purposes, and I have no objection that my data is published in scientific publications in a way that does not reveal my identity.
- I understand that if I make illicit activities known, these will be reported to appropriate authorities.
- I understand that I may refuse to answer any question and that I may withdraw at any time without penalty.
- I understand that if the results of the research have been published, <or my data has been fully anonymised so that it can no longer be attributed to me>, then it will no longer be possible to withdraw.
- I understand that I may stop electronic recordings at any time, and that I may at any time, even subsequent to my participation [request to] have such recordings destroyed (except in situations such as above).

- I understand that, subject to the constraints above, no recordings will be replayed in any public forum or made available to any audience other than the current researchers/research team.
- I freely and voluntarily agree to be part of this research study, though without prejudice to my legal and ethical rights.
- I understand that my participation is fully anonymous and that no personal details about me will be recorded.
- I understand that if I or anyone in my family has a history of epilepsy then I am proceeding at my own risk.
- I understand that personal information about me, including the transfer of this personal information about me outside of the EU, will be protected in accordance with the General Data Protection Regulation.
- I have received a copy of this agreement.

By signing this document, I consent to participate in this study, and consent to the data processing necessary to enable my participation and to achieve the research goals of this study.

PARTICIPANT'S NAME:

PARTICIPANT'S SIGNATURE:

Date:

Statement of investigator's responsibility: I have explained the nature and purpose of the research study, the procedures to be undertaken and any risks that may be involved. I have offered to answer any questions and fully answered such questions. I believe that the participant understands my explanation and has freely given informed consent.

RESEARCHERS CONTACT DETAILS: alex.randles@adaptcentre.ie

RESEARCHER'S SIGNATURE: Alex Randles

Date:

Appendix N. Email Invitation used for Experiment 3

Hi <Name>,

You are invited to participate in the experiment of the Mapping Quality (MQ) framework, which I have developed with my supervisor (Declan O'Sullivan). The experiment is focused on the change detection component of the framework. The component was designed to detect and analyse changes in the source data of mappings. The experiment will involve you interacting with the framework with provided data (source data and mapping), followed by a questionnaire. The **attached document** contains all the information necessary to complete the experiment, including:

- **Section 1.1 -** Additional information on framework
- **Section 1.2 -** Additional information on experiment
- **Section 1.3** - Task sheet and framework access details

The information sheet and informed consent are provided on the framework prior to commencement of the tasks. The questionnaire is provided on the framework at the end of the experiment.

No conference call is required for the experiment, and it can be completed in your own time. It should take roughly 30 minutes to complete.

Please let me know if you have any questions.

Kind Regards,

Alex Randles.

Appendix O. Graph generated in Experiment 3

The listing below presents the graph expressed in the OSCD, which was generated during the second usability experiment (Experiment 3).

```
@prefix oscd: <https://w3id.org/OSCD#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix lode: <http://linkedevents.org/ontology/> .

<http://www.example.com/changeLog/1>
  a oscd:ChangeLog ;
  oscd:hasMaintainer <http://www.example.com/user/1>;
  oscd:hasDetectionStart "2022-11-
00T00:00:00.000000"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDetectionEnd "2022-12-
31T00:00:00.000000"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasCurrentVersion <https://raw.githubusercontent.com/alex-randles/Change-Detection-
System-Examples/main/manipulated_file/student.csv> ;
  oscd:hasPreviousVersion <https://raw.githubusercontent.com/kg-construct/rml-test-
```

```
cases/master/test-cases/RMLTC0002a-CSV/student.csv>;
  oscd:hasNotificationPolicy <http://www.example.com/notificationPolicy/user/1>;
  oscd:hasChange <http://www.example.com/insertChange/15>,
                            <http://www.example.com/insertChange/19>,
                            <http://www.example.com/insertChange/10>,
                            <http://www.example.com/deleteChange/23>,
                            <http://www.example.com/insertChange/20>,
                         <http://www.example.com/insertChange/1>,
                         <http://www.example.com/insertChange/12>,
                         <http://www.example.com/insertChange/5>,
                         <http://www.example.com/insertChange/3>,
                         <http://www.example.com/insertChange/22>,
                         <http://www.example.com/insertChange/14>,
                         <http://www.example.com/insertChange/7>,
                         <http://www.example.com/insertChange/18>,
                         <http://www.example.com/insertChange/0>,
                         <http://www.example.com/insertChange/9>,
                         <http://www.example.com/insertChange/4>,
                         <http://www.example.com/insertChange/11>,
                         <http://www.example.com/insertChange/16>,
                         <http://www.example.com/insertChange/2>,
                         <http://www.example.com/insertChange/13>,
                         <http://www.example.com/insertChange/6>,
                         <http://www.example.com/deleteChange/24>,
                         <http://www.example.com/insertChange/17>,
                         <http://www.example.com/insertChange/21>,
                         <http://www.example.com/insertChange/8> .

<http://www.example.com/insertChange/14>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-18T17:41:01.286820"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "ID";
  oscd:hasChangedData "13" .

<http://www.example.com/insertChange/6>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-18T15:02:01.286820"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "LastName";
  oscd:hasChangedData "Ronaldo" .

<http://www.example.com/insertChange/21>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-20T09:55:01.286844"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasStructuralReference "Column";
  oscd:hasChangedData "Sport" .

<http://www.example.com/insertChange/0>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-18T14:32:01.286820"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "FirstName";
  oscd:hasChangedData "Venus" .

<http://www.example.com/insertChange/19>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-18T15:00:01.286820"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasStructuralReference "Column";
  oscd:hasChangedData "LastName" .

<http://www.example.com/insertChange/13>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-28T16:38:04.139923"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "City";
  oscd:hasChangedData "Brooklyn" .

<http://www.example.com/insertChange/22>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-28T16:07:01.893723"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasStructuralReference "Column";
  oscd:hasChangedData "City" .

<http://www.example.com/insertChange/5>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-18T14:33:01.286820"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "FirstName";
  oscd:hasChangedData "Cristiano" .

<http://www.example.com/insertChange/20>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-18T14:31:01.286820"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasStructuralReference "Column";
  oscd:hasChangedData "FirstName" .

<http://www.example.com/insertChange/18>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-28T16:58:01.148712"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "City";
  oscd:hasChangedData "San Mateo" .

<http://www.example.com/insertChange/12>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-20T10:11:12.137723"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "Sport";
  oscd:hasChangedData "Basketball" .

<http://www.example.com/insertChange/4>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-18T17:35:01.286820"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "ID";
  oscd:hasChangedData "11" .

<http://www.example.com/deleteChange/24>
  a oscd:DeleteSourceData ;
  lode:atTime "2022-10-18T14:05:44.2832120"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasStructuralReference "Column";
  oscd:hasChangedData "Name" .

<http://www.example.com/insertChange/17>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-20T10:12:01.135823"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "Sport";
  oscd:hasChangedData "Football" .

<http://www.example.com/insertChange/9>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-18T17:37:01.286820"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "ID";
  oscd:hasChangedData "12" .

<http://www.example.com/insertChange/11>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-18T15:07:01.286820"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "LastName";
  oscd:hasChangedData "Jordan" .

<http://www.example.com/insertChange/3>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-28T16:12:01.135723"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "City";
  oscd:hasChangedData "California" .

<http://www.example.com/deleteChange/23>
  a oscd:DeleteSourceData ;
  lode:atTime "2022-10-18T14:02:56.9832347"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "Name";
  oscd:hasChangedData "Venus" .

<http://www.example.com/insertChange/16>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-18T15:12:01.286820"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "LastName";
  oscd:hasChangedData "Brady" .

<http://www.example.com/insertChange/8>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-28T16:15:01.187723"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "City";
  oscd:hasChangedData "Funchal" .

<http://www.example.com/insertChange/10>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-18T14:38:01.286820"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "FirstName";
  oscd:hasChangedData "Michael" .

<http://www.example.com/insertChange/2>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-20T09:58:01.526823"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "Sport";
  oscd:hasChangedData "Tennis" .

<http://www.example.com/insertChange/15>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-18T14:41:01.286820"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "FirstName";
  oscd:hasChangedData "Tom" .

<http://www.example.com/insertChange/7>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-20T10:04:01.186823"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "Sport";
  oscd:hasChangedData "Soccer" .

<http://www.example.com/insertChange/1>
  a oscd:InsertSourceData ;
  lode:atTime "2022-10-18T15:01:01.286820"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  oscd:hasDataReference "LastName";
  oscd:hasChangedData "Williams" .
```

Appendix P. PSSUQ Comments from Experiment 3

The table below presents the grouped open comments received from the PSSUQ in experiment 3.

PSSUQ Metric	Comments
System usefulness (Q1-8)	• I think that above all the user interface is easy to understand and to navigate around. It is presented nicely in a manner that makes things easy to find. • Everything is clearly labelled and the use of simple +/- buttons for expand and contract make it easy to go multiple levels deep where necessary. • It was easy to fins what I was looking for. • I had to make sure that I understood the questions correctly which added to my completion time but once I knew what I was looking for I was able to find it

	quickly. • Assuming that what I was looking for was correct I found it efficient to move around the UI and find what I was looking for. • I think the UI is presented in a manner that makes it feel simple to use which makes me feel comfortable. • The use of color and simple button styles makes it easy to learn. Furthermore, the use of button text (eg. 'delete Changes Count: 1') made it easy to learn. • I found that after I answered one or two questions I found I remembered where buttons/info were. • The UX design is simple yet sufficient. I had nor problems to complete the task • As mentioned before, for a user with a bit of experience, the system was almost self-explanatory. • I think I was able to solve almost every task correctly. • Sometimes it took me some time to find the right value since it was hidden and I had to expand a list twice. • see above • Easy to use system enough information provided. • see above • I do not have a lot of experience in uplifting data into RDF, however this system did not give me a hard time doing this. • With PDF instruction it is clear, I can see if you need to use it more than once it will be easy to remember the steps, layout is clear and good flow • Drop down menus make it easily to visualise the changes made • Very Good and clear • It is easy to read • Agree • Very smooth • Very good • So easy, very intuitive! • The system is intuitively usable. • The information button didn't work. Tried hovering and clicking • The information button didnt work again • new tabs made it a little cluttered • Very nice UI • Intuitive • The process for uploading mappings could use some work; ideally a successful data upload should be followed immediately by a link to upload the data, rather than just a congratulations with a home button in the corner. Otherwise, it was fairly intuitive. • The system is very easy to navigate. • Although the system is easy to navigate, there were some ambiguities. For me these consist of the total changes included in the thresholds, and the number of values inserted into the "ID" data reference. • I was not able to tell the total changes included in the thresholds, and the number of values inserted into the "ID" data reference. • Other than the ambiguities, the system gave me the information that I needed quickly. • Other than the ambiguities, the system gave me the information that I needed efficiently. • The UI for the front page containing the change detection processes can be somewhat improved. • Overall, very easy to use.

	• If the ambiguities are removed, then yes, the system provides a lot of productivity. • good • None of the information icons revealed anything to me. I would assume that if I hoover over them they would give me more information on the header or section but there was nothing which was frustrating. • The descrptions are vague • its easy to use the system • I was able to follow the given instruction • I was able to complete it • Was able to figure out • its user intuitive • with instructions its easy to use • yes it would be handy • Could've made stuff more clear and in single document • Given the PDF, yes. • but lots of improvement should be there.. looks more like git diff where they have a better ui • If I started using this daily, it would become easier obviously. It is well designed, but it can be made more intuitive • There are help tool tips which make life easier. But the tool still remains a little technical, and it can be reduced even further, so that it is less conceptually technical to use • The system was easy to use without any glitches but I believe the UI can be made a little bit more self explanatory. • It take me some time to figure out all the information in the hyperlinks from the change detection process page • The mapping-source data change relations took me a bit to fully understand because of the format that they were presented • Easy to use. Other tools in RDF (for example for data validation, like SHACL) are usually very hard to understand • The info buttons wouldn't work for me for some reason. Maybe a video demo could help to learn.
Information quality (Q9-15)	• I don't think I encountered an error. • The task instructions were pretty clear so that was good. However, I don't really recall much on-screen help and I don't think the '?' buttons work. • Provided I understood the questions finding the information was easy. • The documentation was easy to follow. • UI was very nice. • When I uploaded the wrong link for the second dataset, the system told me the exact error, so I could fix it. • s.o. • Good guides for using the system • see above • I didn't make a mistake when using the system so i didn't see this functionality • The PDF instructions were very clear • The hover text labels to gain more information were useful and aided in making the information easy to find, without them I would have been confused, particularly in distinguishing between the two source data fields. • didnt encounter an error • I got an error if two data set are the same. • Got an error that csv URL was incorrect as I had entered the same URL twice, a

	clearer error message might be "the second CSV entered cannot match the first" • No errors • Yes but information button missing • Nice UI • Well named • Meaningful names • Nice UI • I did not encounter any error messages. • Received one error, which gave a good indication of what the problem was. • The one error received due to my mistake, I was able to quickly resolve. • Some ambiguities can be reduced, other than those, the information provided was clear and concise. • good • great • decent • wonderful • excellent • grand • good • The URL I had pasted was wrong. I was given the correct error message and I was able to fix it • Did not always realise drop down options were there. • No error messages appeared as there is no much user input besides the csv files and mapping • No errors made • Table display for the changes might be better instead of the foldable format • I would add the definition for the changes types, the rest is great • For me it was clear but other people might prefer to have everything in the same place instead of clicking links to new pages. I would add the threshold limits information as a popover, so you click and a small table pops with the information needed. • The walk-through was easy to follow. Maybe unguided use had been a better test • Nice hovering info, use of IRIs etc • Easy to follow • no error occurred • did not happen • As mentioned, the info buttons didn't work for me • It was text heavy at times, but not overly complex • pointed to wrong mapping file and had to reload page • I can't really say. I didn't come across a problem. • It wasn't clear if the files were uploaded to the system. For a long time I thought my file was being uploaded. Because the green bar was always at the beginning only blinking. Later I noticed that I had to push the button. Also number of mappings effected was not very clear.
Interface quality (Q16-18)	• I really like the UI. • Good use of colors and simple design was very pleasant. • Simple yet effective • Very clear and comfortable • it's a little simple. • I can tell that React-Bootstrap was used for the UI, therefore, I believe that the UI can be significantly improved.

	• The interface uses bare-bones bootstrap framework which I think can be made more pleasing with a custom design. • wonderful • amazing • Would be nice to have everything working on one web app without the hyperlinks opening to a new tab. • I would say simple but effective
Overall (Q19)	• I wouldn't consider myself and expect in this field so I'm not sure. • Great system, great UI, accessible and easy to use! • Once the ambiguities are removed and the UI is improved, I shall be fully satisfied. • good • really satisfied • Cumbersome instructions page • I would like more information regarding how I should update my mapping as I mentioned previously • I had a positive experience but with the suggestions I think it could be even better • Seems really useful

Appendix Q. Email Invitation for Experiment 4 and 5

Hi <Name>,

You are invited to participate in an evaluation related to research I am conducting with my supervisor (Declan O'Sullivan). The objective of the evaluation is to receive expert feedback on the development process of two ontologies that I have developed related to my research into linked data quality improvement.

The evaluation involves two tasks:

1. You will be asked to review a document containing information on the development of the ontologies, which are designed to capture information related to the publication process of linked data and include:
    - An ontology to model mapping quality
    - An ontology to model changes in the source data of mappings
2. Thereafter, you will be asked to provide feedback through a questionnaire on the following aspects of each ontology
    - Design methodology
    - Implementation
    - Documentation

No conference call is required and the evaluation can be completed without synchronization with me. In total it should take roughly 45 minutes to review the artefacts and fill in the questionnaire.

This is my last experiment before submitting my thesis, and your participation would be greatly appreciated.

Please reply to this email if you would be willing to participate. If you need any clarifications on what is involved, please do not hesitate to ask me.

Kind Regards,
Alex Randles.

Appendix R. Experiment 4 and 5 Informed Consent

TRINITY COLLEGE DUBLIN

INFORMED CONSENT

LEAD RESEARCHERS: Alex Randles

BACKGROUND OF RESEARCH: The Knowledge and Data Engineering Group (KDEG) is located within Trinity College and focuses on researching challenges posed by knowledge discovery, representation and engineering. This knowledge is often represented as RDF data. Creating RDF data requires the creation of mappings, which are definitions that describe the transformation from source to target data. Our research will lead to finding a better way to improve the quality of the mappings being produced.

Creating these mappings has a steep learning curve, which can result in poor quality mappings being generated. These poor-quality mappings can often be unnoticed as the state of the art in linked data quality focuses on the published datasets which are generated by these mappings. These mapping quality issues can grow exponentially within the resulting dataset. We aim to address the challenge of quality issues by bringing quality assessment earlier into the publication stage. We have designed two ontologies: Mapping Quality Improvement Ontology (MQIO) and Ontology for Source Change Detection (OSCD), which are designed to represent information related to the quality of mappings and changes in respective source data. It is hoped the ontologies can be used to improve the quality and alignment of the mappings, thus improving the resulting data quality and maintenance and reuse of those mappings.

PROCEDURES OF THIS STUDY:

- You are going to be provided access to the online documentation of both ontologies.
- You will be asked to review each ontology.
- You will be asked to fill out a feedback questionnaire.

The total duration should take roughly an hour to perform the tasks and fill in the questionnaire.

Your data will not be identifiable since the questionnaire is fully anonymous and results are stored using the IT services called MyZone Google Drive which complies with GDPR rules. The lead researcher (Alex Randles) and the supervisor (Prof. Declan O'Sullivan) will be the only people with access to the data until its publication in an open data repository.

We will perform a qualitative analysis of the questionnaire data by coding and categorising the statements. Then, resulting emerging themes will be the ones reported and used to refine the design of the corresponding ontology. None of your personal details will be recorded and you are free to stop and leave the experiment at any point if you so choose.

PUBLICATION: The goal will be to publish the results of the usability test at Semantic Web conference and workshops, such as Extended Semantic Web Conference (ESWC) and International Semantic Web Conference (ISWC); and other relevant journals, as well as the PhD thesis of the lead researcher at Trinity College Dublin.

CONFLICTS OF INTEREST:

My supervisors will not take part in the experiment. Potential participants of the experiment will not be provided with any prior information before the experiment.

Individual results will be anonymized and published in open data repositories for reproducibility and research will be reported on the aggregate results.

DECLARATION:

- I am 18 years or older and am competent to provide consent.
- I have read, or had read to me, a document providing information about this research and this consent form. I have had the opportunity to ask questions and all my questions have been answered to my satisfaction and understand the description of the research that is being provided to me.
- I agree that my data is used for scientific purposes and I have no objection that my data is published in scientific publications in a way that does not reveal my identity.
- I understand that if I make illicit activities known, these will be reported to appropriate authorities.

- ○ I understand that I may refuse to answer any question and that I may withdraw at any time without penalty.
- ○ I understand that if the results of the research have been published, then it will no longer be possible to withdraw.
- ○ I understand that I may stop electronic recordings at any time, and that I may at any time, even subsequent to my participation request to have such recordings destroyed (except in situations such as above).
- ○ I understand that, subject to the constraints above, no recordings will be replayed in any public forum or made available to any audience other than the current researchers/research team.
- ○ I freely and voluntarily agree to be part of this research study, though without prejudice to my legal and ethical rights.
- ○ I understand that my participation is fully anonymous and that no personal details about me will be recorded.
- ○ I understand that if I or anyone in my family has a history of epilepsy then I am proceeding at my own risk.
- ○ I understand that personal information about me, including the transfer of this personal information about me outside of the EU, will be protected in accordance with the General Data Protection Regulation.
- ○ I have received a copy of this agreement.

By signing this document, I consent to participate in this study, and consent to the data processing necessary to enable my participation and to achieve the research goals of this study.

PARTICIPANT'S NAME:

PARTICIPANT'S SIGNATURE:

Date:

Statement of investigator's responsibility: I have explained the nature and purpose of the research study, the procedures to be undertaken and any risks that may be involved. I have offered to answer any questions and fully answered such questions. I believe that the participant understands my explanation and has freely given informed consent.

RESEARCHERS CONTACT DETAILS: alex.randles@adaptcentre.ie

Alex Randles

RESEARCHER’S SIGNATURE:

Date: 02/01/2022

Appendix S. MQIO Design Document

This document provides information about the design and development of the **M**apping **Q**uality **I**mprovement **O**ntology (**MQIO**).

1. Tasks

You are asked to complete the following tasks:

1. Review the feedback questionnaire which will provide an indication of the requested feedback: https://forms.gle/8DdsSXKJz1eGWhWw6
2. Review this document while considering the design methodology followed by MQIO in comparison to the state of the art.
3. Please complete the feedback questionnaire after you have reviewed the document.

Contact Alex Randles (alex.randles@adaptcentre) if you have any questions.

2. Design Methodology

The design of the MQIO followed best practices as recommended by the semantic web community. Ontology design practices were reused from the most prominent ontology design methodologies. The methodologies included the NeON methodology [134], UPON Lite [100], Ontology development 101: A guide to creating your first ontology [103] and LOT: An industrial oriented ontology engineering framework [110].

1. **Identification of aims, objectives, scope:** The design process commenced with the identification of the aims, objectives and scope of the ontology, which are outlined in Table 1 of this document. The template used for the table was retrieved from the methodologies and used to define the ontology requirements specification document. The document outlines requirements and among other things, the aims, objectives and scope of the ontology.
2. **Identify and analyze relevant information:** A review of publications in the state of the art was conducted to identify relevant information. Publications within the state of the art which related to topics within the defined scope were reviewed to facilitate the retrieval of relevant information. Thereafter, the retrieved information was used to formalize competency questions. Table 2 includes references to publications which inspired the creation of each competency question.

3. **Create Use-cases and Competency questions:** Competency questions were created during the design process of the ontology. The questions define the functional requirements of the ontology and were iteratively refined until an accurate representation of the requirements and objectives was conceived. The final iteration of the questions is shown in Table 2. Use cases were devised in order to refine the requirements of the ontology. The use cases involved projects which uplifted geospatial data (data.geohive.ie) [115] and network monitoring data (Ericsson) [118]. A use case graph generated is available (https://tinyurl.com/2ks5urb9).
4. **Identify Concepts and Relationships:** Concepts and relationships were identified through the state of the art review and the researchers previous experience in the creation of linked data (LD). The concepts and relationships were iteratively defined until the information modeling provided by the ontology satisfied each of the competency questions. In addition, concepts and relationships were reused from existing vocabularies as recommended by the methodologies and the W3C recommendation on Data on the Web Best Practices [87]. **Reused ontologies** included the PROV-O [16] was reused and extended to capture provenance related to mapping information. The Data Quality Vocabulary (DQV) [2] was reused to represent quality metrics utilized during the mapping assessment and validation phase. The reuse in MQIO is demonstrated in the competency questions and ontology documentation.
5. **Progressive iterations:** Steps 2-4 were iteratively repeated until the point when the proposed concepts/relationships provided information which satisfied each requirement defined in the form of a competency question.
6. **Create Ontology:** The ontology was implemented in OWL 2 Web Ontology Language [92]. Concepts and relationships which were defined in the previous step were constructed using Protégé ontology development tool [154]. Furthermore, semantic reasoners were also utilized to detect logical inconsistencies within the ontology.
7. **Evaluate:** The ontology was evaluated with respect of the ability for the defined concepts and relationships to fulfill each competency question. The usage of a semantic reasoner within Protege ensured logical inconsistencies were identified and removed. OOPS! [111] was used to detect common ontology design issues. The quality of metadata and documentation was evaluated through presentation within peer reviewed publications. Feedback received from reviewers allowed us to identify areas for improvement. Peer reviewed publications related to MQIO are outlined in Section 5.
8. **Publication:** Ontology documentation (https://w3id.org/MQIO) was created using WIDOCO [54] which is a tool designed to use ontology metadata to create HTML documents which include descriptions of the classes and properties. Thereafter, the ontology and human readable documentation were published with a permanent identifier as a FAIR resource including an open and permissive license. The documentation contains information about the creation, design, usage, class interaction diagrams and provides various serializations.

3. Background

The following section provides information related to the requirements, design and purpose of MQIO.

3.1 Description

MQIO provides an ontology for expressing information relating to the quality assessment, refinement and validation of declarative mapping definitions. The objective is to make this information easier to publish, exchange and consume, thus improving the overall quality of the resulting LD datasets which are created by these mappings. Furthermore, providing data quality information to the users will allow them to assess the suitability of the mapping for their application. The ontology was designed to resolve the gap in the state of the art in relation to an ontology which represents quality assessment, refinement and validation information of LD mappings.

3.2 Requirements

The development of the ontology follows best practices in ontology development methodologies, such as those mentioned. Creating a specification for the ontology provided additional guidance during the development phase. The requirements have been derived from state of the art review and application of the ontology within a framework and use cases. **Table 1** shows the requirements document for MQIO

Table 1: OSCD Competency Questions

Ontology Requirements Specification Document	
1	**Purpose**
	Capture information related to the quality assessment, refinement and validation of mappings used to generate, relate or interlink RDF datasets. Capturing such information is expected to positively impact the quality of the mappings and datasets as well as improve the reuse and maintenance of those mappings.
2	**Scope**
	In scope: • Mapping quality • Mapping agents • Mapping creation • Mapping validation **Out of scope:** • Source data of the mappings • Resulting dataset
3	**Implementation Language (optional)**
	OWL 2 Web Ontology Language
4	**Intended End-Users (optional)**
	Agents involved in the quality assessment, refinement and validation of LD mappings.
5	**Intended Uses**
	Capturing metadata and provenance relating to the quality assessment, refinement and validation of LD mappings. This metadata also allows for the datasets involved to be assessed in terms of its quality.
6	**Ontology Requirements**
	1. Non-Functional Requirements
	Allow the users of the ontology to define and validate quality requirements related to mappings and capture metadata and provenance related to the quality assessment, refinement and validation of LD mappings.

	2. Functional Requirements: Lists or tables of requirements written as Competency Questions and sentences
	Competency questions in Section 4 (next section).
7	**Pre-Glossary of Terms (optional)**
	1. Terms from Competency Questions
	Competency questions in Section 4.
	2. Terms from Answers
	Competency questions in Section 4.
	3. Objects
	Competency questions in Section 4.

4. Competency Questions

Ontology Competency questions define design requirements in natural language form. These questions state information which should be provided by the ontology. The fulfillment of the questions is accomplished by providing a concept/relationship which represents the required information. Most questions were inspired by literature discovered in the state of the art review. However, certain questions were defined through application to use cases (**DTA**) and feedback from experts (**DTF**). **Table 2** shows the final iteration of competency questions created for MQIO.

The **answer** to each question is structured as <Subject, Relationship, Concept> which represent an RDF triple. A description of each concept and relationship used is available[72]

Table 2: MQIO Competency Questions

#	Question	Relationship	Concept	References
Subject: mqio:MappingArtefact A mapping artefact contains rules which link or create linked data datasets.				
1	Who created the mapping?	mqio:wasCreatedBy	prov:Agent mqio:MappingRefinement	[29,30, 43,75,93]
2	What was the rationale for creating the mapping?	mqio:hasPurpose	xsd:string	[32,43, 93]
3	What instruments were utilized to define the mapping?	mqio:usedTool	xsd:string	[32,43, 64,107]
4	When was the mapping defined?	prov:generatedAtTime	xsd:dateTime	[43,84, 93]

[72] https://drive.google.com/file/d/1fWzhZr7UDCXm86Zo9qpHC8egu76_ZypC

Subject: mqio:MappingAssessment An activity in which the quality of a mapping document is assessed, generating information on quality issues within the mapping.				
5	Who performed the quality assessment of the mapping?	prov:wasAssociatedWith	prov:Agent	[42,43,84]
6	What mapping is associated with the assessment?	mqio:assessedMapping	mqio:MappingArtefact	[35,42,43,65]
7	What quality metrics were executed during the assessment process?	mqio:wasExecuted	dqv:Metric	[2,42,65,75,79]
8	What quality measurements resulted from the assessment?	prov:generated	dqv:QualityMeasurement	[2,37]
9	What quality issues were detected?	mqio:hasValidationReport	mqio:MappingValidationReport	[42,65,75,80,97]
10	What value is associated with the violation?	mqio:hasObjectValue, mqio:hasLiteralValue	rdfs:Resource, xsd:string	[42,65,75,80,97]
11	How are the quality issues described?	mqio:hasResultMessage	mqio:MappingViolation	[42,65,75,80,97]
12	When were the quality issues detected?	prov:endedAtTime	xsd:dateTime	[42,65,75,97]
13	Where can provenance on the issues be accessed?	mqio:hasViolation	mqio:MappingValidationReport	[80,84]
14	What quality metrics were associated with the detected quality issues?	mqio:isDescribedBy	dqv:Metric	[2,38,65,75,97]
15	What quality dimensions represent the metrics?	dqv:inDimension	dqv:Dimension	[2,38,65,75,97]
16	What quality categories represent the dimensions?	dqv:inCategory	dqv:Category	[2,38,65,75,97]
Subject: mqio:MappingRefinement An activity which involves removing quality violations contained within a mapping document.				
17	Who performed the quality refinement of the mapping?	prov:wasAssociatedWith	prov:Agent	[43,84]
18	When was the refinement process completed?	prov:endedAtTime	xsd:dateTime	[42,43,84]

19	What queries are associated with refinements?	mqio:usedQuery	xsd:string	DTA
20	What confidence score did the refinements have?	mqio:hasConfidenceScore	xsd:double	DTA
21	What violations have been refined?	mqio:wasRefinedBy	mqio:Violation	DTA
22	What quality requirements are associated with the mapping?	mqio:hasQualityRequirement	mqio:QualityRequirement	[2,75,97]
Subject: mqio:QualityRequirement A quality requirement is a requirement a mapping should satisfy.				
23	What quality measurements were associated with the requirements?	mqio:hasQualityMeasurement	dqv:QualityMeasurement	[2,37]
24	Are the requirements satisfied?	mqio:isSatisfied	xsd:boolean	[2,37]
Subject: dqv:Metric Represents a standard to measure a quality dimension.				
25	What refinements are associated with the quality metrics?	mqio:hasRefinement	mqio:MappingRefinement	[42,65,97]

SPARQL query answers to the competency questions are available[73]. Further information on the graph used to execute the queries can be found in the “Description” section of the ontology documentation.

5. Publications

The following peer reviewed publications related to the design and usage of MQIO[74].

1) *Randles, A., Junior, A.C. and O'Sullivan, D., 2020. Towards a vocabulary for mapping quality assessment. In OM@ ISWC (pp. 241-242).*

In this publication we presented a brief overview of the design of MQIO.

[73] https://drive.google.com/file/d/1zsgX66NokGc3mxXuaRAL3S3bVmkrhMlU

[74] As a note the ontology was previously called the Mapping Quality Vocabulary (MQV) in the publications

2) *Randles, A., Junior, A.C. and O'Sullivan, D., 2021, January. A vocabulary for describing mapping quality assessment, refinement and validation. In 2021 IEEE 15th International Conference on Semantic Computing (ICSC) (pp. 425-430). IEEE*

In this publication we presented a detailed description of the design process followed by MQIO, use case of the ontology and reuse of existing vocabularies. Furthermore, we discussed an application of the ontology within a demonstration walkthrough. Finally, we mentioned related provenance and metadata models.

3) *Randles, A. and O'Sullivan, D., Assessing Quality of R2RML Mappings for OSi's Linked Open Data Portal.*

In this publication we presented an overview of the MQI framework applied to geospatial R2RML mappings within a current research project. The reports generated during the application were expressed in MQIO.

4) *Randles, A., O'Sullivan, D., Keeney, J. and Fallon, L., Applying a Mapping Quality Framework in Cloud Native Monitoring.*

In this publication we presented an overview of the MQI framework applied to mappings designed to uplift time series metric data utilized within cloud native monitoring. The reports generated during the application were expressed in MQIO.

5) *Randles, A. and O'Sullivan, D., 2022. Evaluating Quality Improvement Techniques Within the Linked Data Generation Process. In Towards a Knowledge-Aware AI (pp. 21-35). IOS Press.*

In this publication we presented a detail description of the framework and the usability evaluation which was conducted on the MQI framework. Furthermore, we discuss realizations of the results and outline respective improvements. The reports generated during the application were expressed in MQIO.

Appendix T. OSCD Design Document

This document contains information on the development process of the **O**ntology for **S**ource **D**ata **C**hange (**OSCD**).

1. Tasks

You are asked to complete the following tasks:

1. Review the feedback questionnaire which will provide an indication of the requested feedback: https://forms.gle/gCSHZn79eTAFVCrz8

2. Review this document while considering the design methodology followed by OSCD in comparison to the state of the art.
3. Please complete the feedback questionnaire after you have reviewed the document.

Contact Alex Randles (alex.randles@adaptcentre) if you have any questions.

2. Design Methodology

The design of the OSCD followed best practices as recommended by the semantic web community. Ontology design practices were reused from the most prominent ontology design methodologies. The methodologies included the NeON methodology [134], UPON Lite [100], Ontology development 101: A guide to creating your first ontology [103] and LOT: An industrial oriented ontology engineering framework [110].

1. **Identification of aims, objectives, scope:** The design process commenced with the identification of the aims, objectives and scope of the ontology, which are outlined in Table 1 of this document. The template used for the table was retrieved from the methodologies and used to define the ontology requirements specification document. The document outlines requirements and among other things, the aims, objectives and scope of the ontology.
2. **Identify and analyze relevant information:** A review of publications in the state of the art was conducted to identify relevant information. Publications within the state of the art which related to topics within the defined scope were reviewed to facilitate the retrieval of relevant information. Thereafter, the retrieved information was used to formalize competency questions. Table 2 includes references to publications which inspired the creation of each competency question.
3. **Create use-cases and competency questions:** Competency questions were created during the design process of the ontology. The questions define the functional requirements of the ontology and were iteratively refined until an accurate representation of the requirements and objectives was conceived. The final iteration of the questions is shown in Table 2. Use cases were devised in order to refine the requirements of the ontology and were retrieved from the RML test case files[75]. The test case files provided a diverse set of source data in formats such as XML, JSON, relational databases and CSV as well as respective RML mappings. In addition, the R2RML test case files[76] were used, however, the source data is only represented in relational format. The test cases facilitated the creation of use cases through the generation of graphs defined in OSCD when changes were detected between the file versions. The use case has been documented in a publication [116]. In

[75] https://rml.io/test-cases/

[76] https://www.w3.org/2001/sw/rdb2rdf/test-cases/

addition, the ontology was applied to a network management use case [118]. A use case graph generated by the RML test cases is available (https://tinyurl.com/hyrmr9aa).

4. **Identify concepts and relationships:** Concepts and relationships were identified through the state of the art review and the researchers previous experience in the creation of linked data (LD). The concepts and relationships were iteratively defined until the information modeling provided by the ontology satisfied each of the competency questions. In addition, concepts and relationships were reused from existing vocabularies as recommended by the methodologies and the W3C recommendation on Data on the Web Best Practices [87]. **Reused ontologies** included an ontology for Linking Open Descriptions of Events (LODE) [126] which is designed to model events. LODE was extended to model changes as specialized events which have occurred in source data. The Rei ontology [77] is designed to model policies for various domains and was reused to represent the details of the notification policy, which is used to inform maintainers of changes. The DUL ontology [25] was reused similar to LODE to represent the agents involved in the activities. The reuse in OSCD is demonstrated in the competency questions and ontology documentation.
5. **Progressive iterations:** Steps 2-4 were iteratively repeated until the point when the proposed concepts/relationships provided information which satisfied each requirement defined in the form of a competency question.
6. **Create Ontology:** The ontology was implemented in OWL 2 Web Ontology Language [92]. Concepts and relationships which were defined in the previous step were constructed using Protégé ontology development tool [154]. In addition, semantic reasoners were also utilized to detect and remove logical inconsistencies within the ontology.
7. **Evaluate:** The ontology was evaluated for sufficiency to provide information to fulfill each competency question. The usage of a semantic reasoner within Protege ensured logical inconsistencies were identified. In addition, OOPS! Pitfall Scanner [111] was used to detect common ontology design issues. The quality of metadata and documentation was evaluated through presentation in peer reviewed publications. Feedback received from reviewers allowed areas for improvement to be identified. Peer reviewed publications related to OSCD are outlined in Section 5. Further expert feedback was received in a previous user evaluation where they were asked to provide feedback on the design and application of the ontology. In addition, the ontology was presented to a panel of semantic web experts at the semantic interoperability conference (SEMIC2022)[77] organized by the European commission. Each graph generated in the use cases were assessed with the RDFUnit [81] quality assessment framework which provides test driven validation of RDF data.

[77] https://joinup.ec.europa.eu/collection/semic-support-centre/semic-conference

8. **Publication:** Ontology documentation (https://w3id.org/OSCD) was created using WIDOCO [54] which is a tool designed to use ontology metadata to create HTML documents listing its classes and properties. Thereafter, the ontology and human readable documentation were published with a permanent identifier as a FAIR resource including an open and permissive license. The documentation contains information about the creation, design, usage, class interaction diagrams and provides various serializations.

3. Background

The following section provides information related to the requirements, design and purpose of OSCD.

3.1 Description

OSCD provides an ontology for expressing information related to changes which occur in source data used by mappings to create a LD dataset. The information will allow notification of changes to data maintainers which will enable them to make appropriate changes to the mappings and data in order to maintain alignment between them. In addition, it is hoped the information will benefit the maintenance and reuse of mappings. The ontology was designed to resolve the gap in the state of the art in relation to an ontology which represents information related to changes in the source data of LD mappings.

3.2 Requirements

The development of the ontology follows best practices in ontology development methodologies, such as those mentioned. Creating a specification for the ontology provided additional guidance during the development phase. The requirements have been derived from state of the art review and application of the ontology within a framework and use cases. **Table 1** shows the requirements document for OSCD.

Table 1: Ontology requirements specification document [135]

Ontology Requirements Specification Document	
1	**Purpose**
	Capturing information related to changes which have occurred in the source data used by mappings to produce LD. The information is expected to positively impact the quality of mappings and datasets by preserving alignment with the underlying data sources as well as facilitate the reuse and maintenance of those mappings.
2	**Scope**
	In scope: ● Source data used by mappings ● Changes within source data ● Agents related to the source data ● Notifications of changes **Out of scope:** • Actions to improve alignment • Resulting published dataset
3	**Implementation Language (optional)**
	OWL 2 Web Ontology Language
4	**Intended End-Users (optional)**
	Agents involved in the transformation of data into LD representation.
5	**Intended Uses**
	Capturing information associated with the agents and activities involved in the generation

	and maintenance of LD.
6	**Ontology Requirements**
	1. Non-Functional Requirements
	Allow the users of the ontology to identify and be notified of changes which have occurred in the source data of LD mappings.
	2. Functional Requirements: Lists or tables of requirements written as Competency Questions and sentences
	Competency questions in Section 4 (next section).
7	**Pre-Glossary of Terms (optional)**
	1. Terms from Competency Questions
	Competency questions in Section 4.
	2. Terms from Answers
	Competency questions in Section 4.
	3. Objects
	Competency questions in Section 4.

4. Competency Questions

Ontology Competency questions define design requirements in natural language form. These questions state information which should be provided by the ontology. The fulfillment of the questions is accomplished by providing a concept/relationship which represents the required information. Most questions were inspired by literature discovered in the state of the art review. However, certain questions were defined through application to use cases (**DTA**) and feedback from the experts (**DTF**). The answers to each question are structured as <Subject, Relationship, Concept> which represent an RDF triple. **Table 2** shows the final iteration of competency questions created for OSCD.

The **answer** to each question is structured as <Subject, Relationship, Concept> which represent an RDF triple. A description of each concept and relationship used is available[78].

Table 2: OSCD Competency Questions

#	Question	Relationship	Concept	References
Subject: cdo:ChangeLog A grouping of changes which have occurred in a source data.				
1	Who maintains the source data?	oscd:hasMaintainer	dul:Agent	[109,126,139]
2	What data is represented in the source data?	oscd:hasCurrentSource oscd:hasPreviousSource	rdfs:Resource	DTA
3	What changes were detected in the source data?	oscd:hasChange	oscd:InsertSourceData ………	[78,109,121,142,156]

[78] https://drive.google.com/file/d/12tQrdzb3xKDmt2VTlKhIzs9Y5FWtkhew

			cdo:DeleteSourceData	
4	What notification policies are associated with the source data?	oscd:hasNotificationPolicy	rei-policy:Policy	[104,109,139, 142,156]
5	What thresholds are associated with the notification policy?	oscd:hasThrehsold	xsd:integer	[104,109]
6	When did the change detection process begin?	oscd:hasDetectionStart	xsd:dateTime	[104,109,131, 141]
7	When did the change detection process finish?	oscd:hasDetectionEnd	xsd:dateTime	[104,109,131, 141]
Subject: oscd:[<Action>]SourceData The change which has occurred. <Action> represents one of the change types.				
8	Who is responsible for the change?	oscd:wasChangedBy	dul:Agent	DTF
9	What data was changed as a result of a specific change?	oscd:hasChangedData	xsd:string	[78,109,131,1 39,141]
10	When did the change occur?	lode:atTime	xsd:dateTime	[78,109,126]
11	What was the original value?	oscd:hasPreviousValue	xsd:string	DTF
12	What data is associated with the change?	oscd:hasStructuralReference, oscd:hasDataReference	xsd:string	DTA

SPARQL query answers to the competency questions are available[79]. Further information on the graph used to execute the queries can be found in the "Description" section of the ontology documentation.

5. Publications

The following peer reviewed publications related to the design and usage of OSCD[80].

1) *Randles, A. and O'Sullivan, D., 2022, October. Modelling & Analyzing Changes within LD source data. In MEPDaW 2022-8th Workshop on Managing the Evolution and Preservation of the Data Web.*

[79] https://drive.google.com/file/d/1rd-xMDkMcPfPry28vmG4MsKuvnXjb6xn

[80] As a note the ontology was previously called the Change Detection Ontology (CDO) in the publications

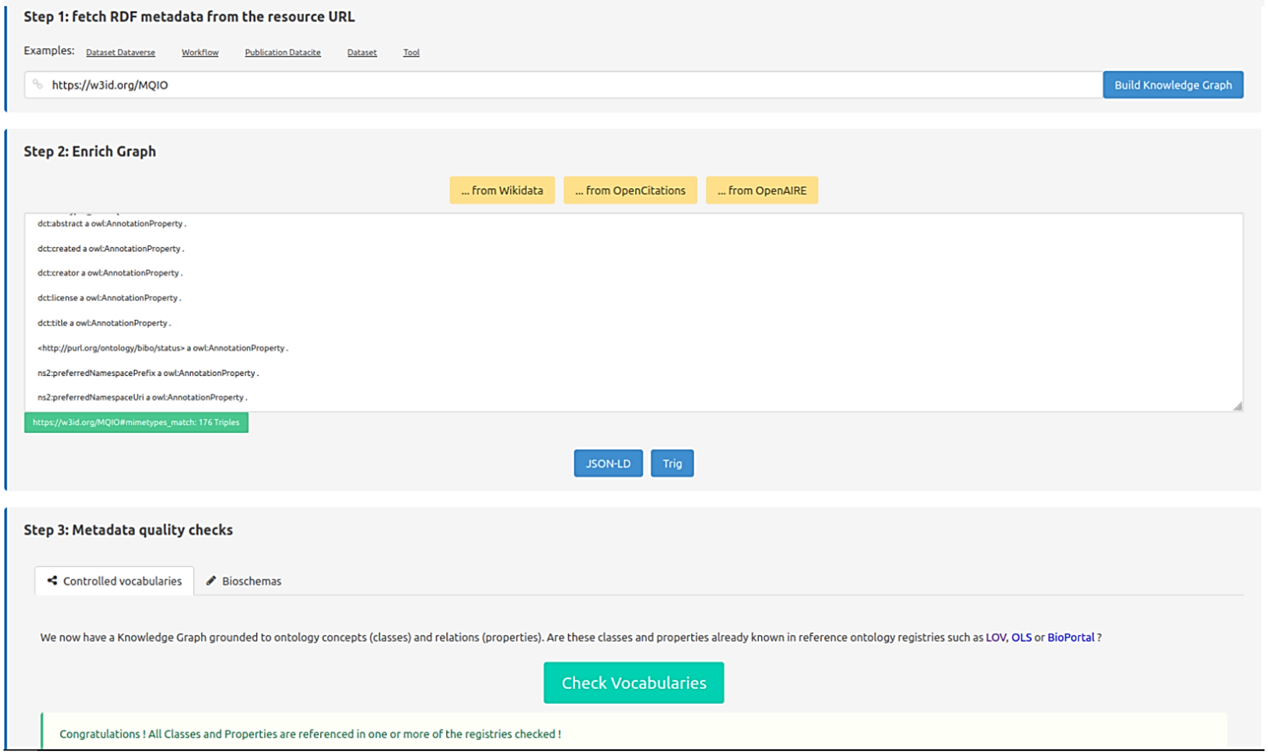
In this publication we presented a description of the design process followed by OSCD. Furthermore, we outlined the application of the ontology within a use case and our quality improvement framework, named the Mapping Quality Improvement (MQI) framework.

2) *Randles, A., O'Sullivan, D., Keeney, J. and Fallon, L., 2022, September.* *Applying a Mapping Quality Framework in Cloud Native Monitoring*. *In Proceedings of the 18th International Conference on Semantic Systems (SEMANTiCS).*

In this publication we briefly described the application of OSCD in a network management use case. The change detection component of the MQI framework was applied to data which represented metrics used in network management.

Appendix U. Screenshot of result of FAIR Metadata Validator

The screenshot below presents the results of the FAIR metadata analysis when the MQIO was input into the validator.



Step 1: fetch RDF metadata from the resource URL

Examples: Dataset Dataverse Workflow Publication Datacite Dataset Tool

https://w3id.org/MQIO

Step 2: Enrich Graph

```
dct:abstract a owl:AnnotationProperty .

dct:created a owl:AnnotationProperty .

dct:creator a owl:AnnotationProperty .

dct:license a owl:AnnotationProperty .

dct:title a owl:AnnotationProperty .

<http://purl.org/ontology/bibo/status> a owl:AnnotationProperty .

ns2:preferredNamespacePrefix a owl:AnnotationProperty .

ns2:preferredNamespaceUri a owl:AnnotationProperty .
```

https://w3id.org/MQIO#mimetypes_match: 176 Triples

Step 3: Metadata quality checks

We now have a Knowledge Graph grounded to ontology concepts (classes) and relations (properties). Are these classes and properties already known in reference ontology registries such as LOV, OLS or BioPortal ?

Congratulations ! All Classes and Properties are referenced in one or more of the registries checked !

As can be seen, the result message shown at the bottom of the screenshot indicated the MQIO conforms to FAIR metadata principles. The same result was observed for the OSCD.

Appendix V. Conformance of the MQIO and OSCD to Gruber Principles

The 5 Gruber principles are described between the quotation marks. Thereafter, the conformance of both ontologies to each principle is outlined.

1. ***“Clarity:*** *An ontology should effectively communicate the intended meaning of defined terms. Definitions should be objective. While the motivation for defining a concept might arise from social situations or computational requirements, the definition should be independent of social or computational context. Formalism is a means to this end. When a definition can be stated in logical axioms, it should be. Where possible, a complete definition (a predicate defined by necessary and sufficient conditions) is preferred over a partial definition (defined by only necessary or sufficient conditions). All definitions should be documented with natural language.”*

   Conformance: All definitions have been documented in the respective online documentation for the ontologies. Definitions are stated in logical axioms in the respective competency questions. In addition, the rationale for the formalization of the concepts is provided in the respective design methodology sections described in this thesis.
2. ***“Coherence:*** *An ontology should be coherent: that is, it should sanction inferences that are consistent with the definitions. At the least, the defining axioms should be logically consistent. Coherence should also apply to the concepts that are defined informally, such as those described in natural language documentation and examples. If a sentence that can be inferred from the axioms contradicts a definition or example given informally, then the ontology is incoherent.”*

   Conformance: Logical consistency of both ontologies was assessed using semantic reasoners in the Protégé development tool. In addition, OOPS! was used to detect additional design inconsistencies in order to determine if the designs were sufficiently coherent.
3. ***“Extendibility:*** *An ontology should be designed to anticipate the uses of the shared vocabulary. It should offer a conceptual foundation for a range of anticipated tasks, and the representation should be crafted so that one can extend and specialize the ontology monotonically. In other words, one should be able to define new terms for special uses based on the existing vocabulary, in a way that does not require the revision of the existing definitions.”*

   Conformance: The definition of a requirements specification documentation for both ontologies helped to identify anticipated tasks and guide the resulting representation. The resulting ontologies provide models which are easily extendable as they reuse well known vocabularies (PROV-O, DQV etc.), which can be extended in a similar manner in order to specialize relevant quality or change information. Specialized concepts can be added to both ontologies as sub classes of existing concepts, which allows them to inherit the attributes without the requirement of revision of existing term definitions.

4. ***"Minimal encoding bias:*** *The conceptualization should be specified at the knowledge level without depending on a particular symbol-level encoding. An encoding bias results when a representation choices are made purely for the convenience of notation or implementation. Encoding bias should be minimized, because knowledge-sharing agents may be implemented in different representation systems and styles of representation."*

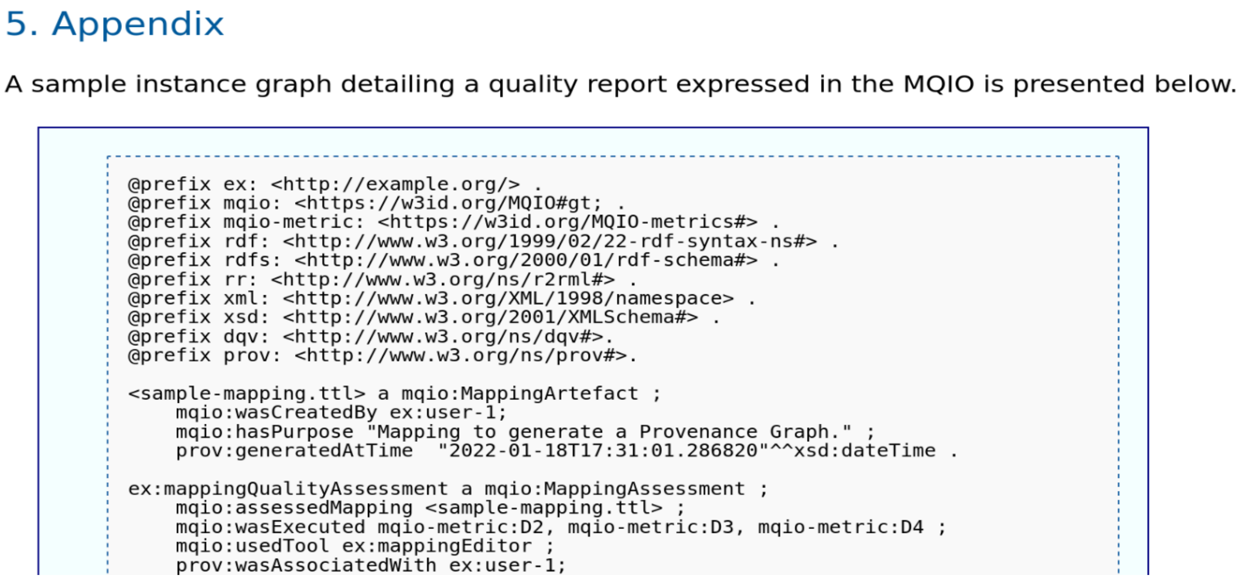
   Conformance: The definition of requirements, use cases and terms in natural language enables different agents to implement the knowledge into their respective systems. While the ontologies are implemented in OWL2 Web Ontology Language, the terms defined in natural language can be translated into required encodings. In addition, it is a straightforward process to automatically transform the OWL2 encodings into other required encodings.

5. ***"Minimal ontological commitment:*** *An ontology should require the minimal ontological commitment sufficient to support the intended knowledge sharing activities. An ontology should make as few claims as possible about the world being modeled, allowing the parties committed to the ontology freedom to specialize and instantiate the ontology as needed. Since ontological commitment is based on consistent use of vocabulary, ontological commitment can be minimized by specifying the weakest theory (allowing the most models) and defining only those terms that are essential to the communication of knowledge consistent with that theory."*

   Conformance: Concepts and relationships defined within both ontologies were kept as minimal as possible. The definition of them was an iterative process where only key concepts were defined in the final versions of both ontologies. Feedback from experts throughout the lifecycle of the ontologies helped to identify terms which were required to communicate the proposed knowledge of them. An expert (Participant #5) in the ontology design evaluation, described the OSCD as "*Nice and lean"*, which indicated the design was minimalist and supported these observations.

Appendix W. Appendix Section added to documentation of MQIO

The screenshot below presents the Appendix section which was added to the documentation of the MQIO.



5. Appendix

A sample instance graph detailing a quality report expressed in the MQIO is presented below.

```
@prefix ex: <http://example.org/> .
@prefix mqio: <https://w3id.org/MQIO#gt; .
@prefix mqio-metric: <https://w3id.org/MQIO-metrics#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix dqv: <http://www.w3.org/ns/dqv#>.
@prefix prov: <http://www.w3.org/ns/prov#>.

<sample-mapping.ttl> a mqio:MappingArtefact ;
    mqio:wasCreatedBy ex:user-1;
    mqio:hasPurpose "Mapping to generate a Provenance Graph." ;
    prov:generatedAtTime  "2022-01-18T17:31:01.286820"^^xsd:dateTime .

ex:mappingQualityAssessment a mqio:MappingAssessment ;
    mqio:assessedMapping <sample-mapping.ttl> ;
    mqio:wasExecuted mqio-metric:D2, mqio-metric:D3, mqio-metric:D4 ;
    mqio:usedTool ex:mappingEditor ;
    prov:wasAssociatedWith ex:user-1;
```

A similar section was added to the documentation of the OSCD.