Harnessing User Input to Improve Linked Data Fidelity

Caroline Burgess, B.Sc.

Submitted to the University of Dublin,
in partial fulfillment of the requirements for the degree of
Master of Science in Computer Science

2010

Declaration

I, the undersigned, declare that the work described in this dissertation is, except where otherwise stated, entirely my own work and has not been submitted as an exercise for a degree at this or any other university.

Signed: ____________________

Caroline Burgess

13th September, 2010

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this dissertation upon request.

Signed: ____________________

Caroline Burgess
13th September, 2010

Acknowledgements

Firstly, I would like to thank my supervisor Declan O'Sullivan and Postdoc advisors Rob Brennan and Melike Sah for all the support, encouragement and immense interest in my progress all the way through the journey of this dissertation.

I would like to acknowledge all those who participated in my user study. To give up 50 minutes of your valuable time for no personal gain is very much appreciated.

I would like to thank my fellow NDS classmates for a very memorable year. In particular, I would like to thank Ann Marie for the numerous enjoyable chats over coffee and Paul Gildea, whose infectious laughter lightened the strain experienced by the class so often, as we battled through the notorious NDS workload.

I would also like to thank my childminder Mags Dunne, for welcoming my daughter into her home and loving family outside of school term time. Her support enabled me to continue my education without any worry for Devyn's wellbeing and happiness. I could not have attained my undergraduate degree or completed this dissertation without her.

Lastly, but definitely not least importantly, I would like to thank my loving daughter Devyn who has sacrificed **so** much quality time during my time as a mature student — six years in total! Thank you so much Devyn for your kind understanding, endless patience and also for motivating me to achieve more in life. I promise to return that support to you during your journey through secondary school and when you get your dream college placement in NCAD studying Fashion Design.

Caroline Burgess
University of Dublin, Trinity College
2010

Abstract

Tim Berners-Lee has a great vision for the Web of Linked Data. He has suggested that linking of data which has not previously been linked may even lead to a cure for cancer and Alzheimer's. His vision is viewed by many as being a long way off. Most interlinks between data sets on the Web of Linked Data are generated automatically at present and are considered 'best effort' matches only. The quality of data retrieved from Linked Data sets is an area which needs much research if Berners-Lee's vision of the semantic web is to become a reality. Current estimates suggest that there are 150 million interlinks between data sets on the Web of Linked Data. Defining manual interlinks on such a large scale is impractical.

This dissertation investigates a more dual approach to interlinking of data sets with the addition of a human 'in the loop', contributing to existing automatic interlinks with the aim of improving interlink quality. Recent research suggests that users need cognitive support for interlinking activities. The aim of cognitive support for User Contributed Interlinking (UCI) processes is to provide software support that simplifies performing an interlinking task for the user.

In order to ascertain if the quality of Linked Data interlinks can be improved by cognitive tool support for user contributed interlinking processes, a tool which enables a user to create interlinks between data sets from the Web of Linked Data was developed. Cognitive tool support for users was implemented through a clear user interface, decision support for interlinking processes, access to metadata and filtering to display interlinks.

A user evaluation which involved users with various levels of Linked Data experience was performed. Participants achieved an F-Measure of 26 % (average) with cognitive support for UCI processes. The evaluation also indicated that there is an extremely weak positive correlation between F-Measure of participants and participant Linked Data experience. This would suggest that the tool provided enough cognitive support to enable participants with little or no Linked Data/Ontology experience to achieve an F-Measure similar to experienced participants. If participants with no experience can be supported to achieve an F-Measure similar to experienced participants, one might conclude that anyone can contribute to the realisation of Berners-Lee's vision.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

There has been much hype surrounding the Semantic Web in recent years. The Semantic Web is considered a 'Web of data' which is easily processable by machines, on a global scale. It is often likened to 'a globally linked database' [43].

1.1 Motivation

Tim Berners-Lee — "The inventor of the internet" has a great vision for the Web of Linked Data. He has suggested that it may lead to a cure for cancer and Alzheimer's[9]. His vision is viewed by many as being a long way off. The quality of interlinks between data sets is questionable. Most interlinks between data sets on the Web of Linked Data are generated automatically at present and are considered 'best effort' matches only. The quality of data retrieved from Linked Data sets is an area which needs much research if Berners-Lee's vision of the Semantic Web is to become a reality.

The investigation into improving the quality of interlinks by the introduction of a human 'in the loop' and the constraints posed by cognitive support requirements for user contributed interlinking have been the motivating factors for this research.

The overall objective of this dissertation involves the investigation and evaluation into two main areas which are outlined in sections 1.1.1 and 1.1.2.

1.1.1 User Contributed Interlinking (UCI)

User Contributed Interlinking (UCI) is a relatively new way of creating semantic links. It is an approach to improving quality of interlinks between data sets through the inclusion of a human user in the interlinking process.

1.1.2 Cognitive Support

The aim of cognitive support for UCI processes is to provide software support that simplifies performing an interlinking task for the user.

1.2 Dissertation Goal

The research question is:

To ascertain if the quality of the interlinks between data sets on the "Web of data" can be improved by cognitive tool support for the User Contributed Interlinking (UCI) processes.

This leads to the following research objectives:

1. Build a tool which enables a user to create interlinks between two Linked Data sets. The tool should reduce the cognitive load for users whilst performing interlinking tasks.
2. Conduct a user evaluation which involves getting users to use the tool performing interlinking tasks, evaluating what each user did and how effective the tool was in supporting each user.

1.3 Dissertation Outline

This chapter outlines the motivating factors and research objectives for this dissertation.

Chapter 2 gives an insight into the background of Linked Data. The chapter discusses the growth of Linked Data and the benefits of interlinking data on the Web.

Chapter 3 investigates the current situation in key areas of interest influencing this research. The state of the art chapter discusses current research in the areas of UCI, UCI tools, User Cognitive Support, and User Involvement.

Chapter 4 outlines the major design decisions that contributed to the final implementation of the demonstrator.

Chapter 5 details the work carried out during the implementation phase of the dissertation.

Chapter 6 evaluates the extent to which users could improve the quality of Linked Data interlinks and also the level of cognitive support offered by the 'Improv-A-link' demonstrator.

Chapter 7 discusses the main conclusions of this dissertation and suggests future work in the field.

Chapter 2

Background

2.1 Introduction

This chapter gives an insight into the background of Linked Data. The chapter discusses the growth of Linked Data and the benefits of interlinking data on the Web.

2.2 Growth of Linked Data

Early contributors to the Linked Data cloud were mainly researchers and developers but there has been considerable growth over the past few years, with organisations such as the BBC and Reuters getting involved, publishing their data. Figure 2.1 shows the growth of the Linked Open Data (LOD) cloud from 2007 to 2009.

Figure 2.1: The Growth of the Linking Open Data Cloud 2007–2009

Tom Heath, an active member of the Linked Data Community, recently stated:

> “I have no concerns that the Linked Data cloud will explode or dissipate through a lack of interlinking, for a number of reasons. Firstly, a prerequisite for joining this Web is the creation of links between new data sets and those that already exist – a data set being available on the Web is not enough, it must also be *in* the Web. Secondly, I perceive an increased or renewed understanding within the Semantic Web community of the power of networks effects, and the value that linking to existing hubs such as DBpedia and Geonames can bring. This value will ensure the Web remains a Web, not a series of isolated data islands” [19].

2.3 Benefits of Interlinking Data on the Web

At the 2009 Technology, Entertainment, Design (TED) talks, Berners-Lee spoke about the huge unlocked potential that is the Web of Linked Data. He talked of the benefits of putting data on the Web as opposed to documents, and of useful information which can be retrieved from related data. Related data on the Web can be more useful for analysis purposes and for finding patterns in data. A prime example of the enormous potential benefit of Linked Data given by Berners-Lee is in relation to the current challenges facing society, in the quest for cures for illnesses such as cancer and Alzheimer’s. He spoke of how until recently important scientific data was kept locked away in ‘silos’, and how many scientists have now embraced Linked Data as they can see the potential of sharing data with fellow scientists in a new way. He gave a specific example of a question that could be considered by a scientist which may unlock some answers in a way the current Web cannot. The question he asked was “What proteins are involved in signal transductions and are related to pyramidal neurons?” This query, when entered into Google returned 223,000 hits but no results, because no one has asked that question before. Linked healthcare data, on the other hand, returned 32 hits, 32 results. Interlinked data on the Web, brings with it the ability to bridge across different disciplines, increasing the potential for discovery of volumes of potentially useful and previously unknown information from data. [9]

He also spoke of social networking sites, and how all the massive amount of data which is linked to you and which you are linked to is generally locked in to a particular social networking site. Linked Data, according to Berners-Lee is *the* way to bring interoperability to social networking sites. Linked Data allows ‘applications to operate on top of an unbounded set of data sources, via standardised access mechanisms’ [13].

2.4 Linked Data

According to linkeddata.org — ‘The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web’ [39]. Tim Berners-Lee outlined a set of ‘rules’ for publishing data on the Web in a way that all published data becomes part of a single global data space.
The four linked data principles were described as follows:

Linked Data Principles

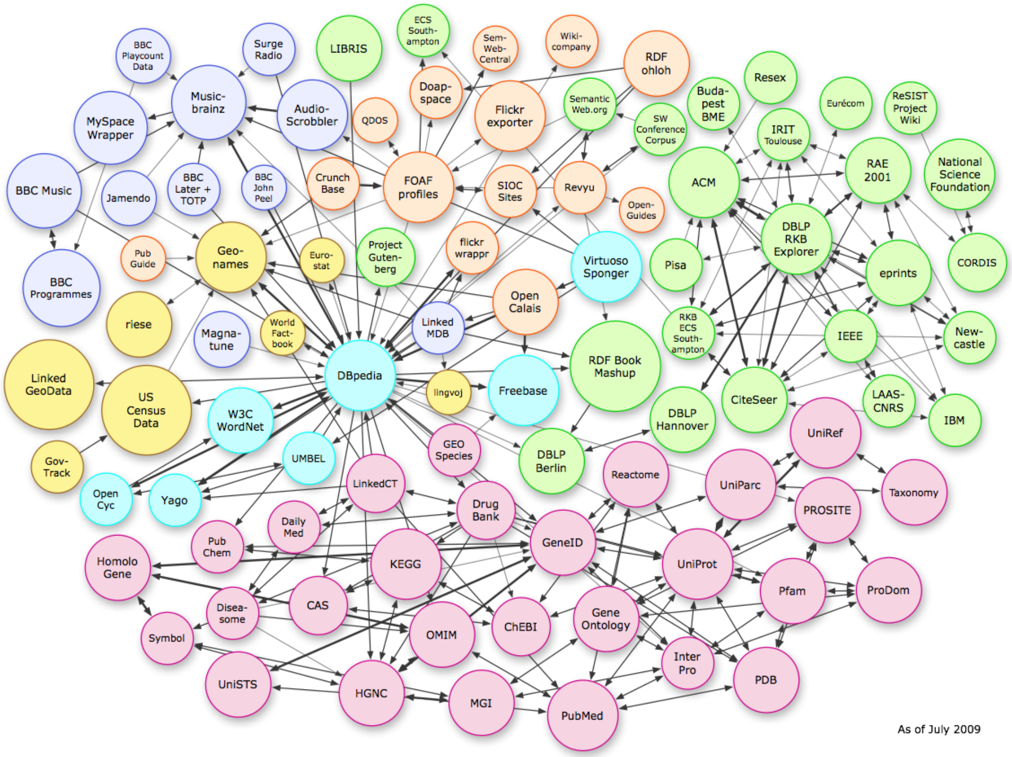
1. All items should be identified using URIs;
2. All URIs should be dereferenceable, that is, using HTTP URIs allows looking up an item identified through the URI;
3. When looking up a URI, that is, a RDF property interpreted as a hyperlink which leads to more data, usually referred to as the ‘follow-your-nose’ principle;
4. Links to other URIs should be included in order to enable the discovery of more data. [8]

It is the view of Berners-Lee that ‘Linked Data is essential to actually connect the Semantic Web’ [8]. He states that although many ontologies and significant data stores have been produced in recent years, much of that data is not accessible as Linked Data on the Web, and is often ‘buried in a zip archive somewhere’ instead. Auer et al. [7] suggest that ‘stitching together the world’s structured information and knowledge to answer semantically rich queries is one of the key challenges of computer science, and one that is likely to have tremendous impact on the world as a whole’.

In 2007, the Linking Open Data Community Project was launched within the W3C Semantic Web Education and Outreach Group [37]. Unlike the current ‘hypertext Web’, semantic links are used to interlink data from different sources. These links can be set manually or generated by the use of interlinking algorithms, the latter being the most common as individual data sets can contain millions of triples.

Figure 2.2 shows the data sets which were published on the Web as of July, 2009 and the interlinks that exist between those data sets. Arcs between data sets indicate that a link exists between items in the data sets joined by the arc. The heavier the arc, the greater the number of links exist between data sets. Bidirectional arcs show outward links to the other exist in each data set [13]. Current estimates suggest that there are 13 billion triples and 150 million interlinks on the Linked Data cloud [14]. Data on the LOD cloud spans a diverse range of topics, data sets exist which relate to music [44], statistical data [31], movies [21] and much more.

Bizer et al. [12] view the existence of large amounts of meaningfully interlinked Resource Description Framework (RDF) data on the Web as being a fundamental prerequisite to the



As of July 2009

Figure 2.2: The LOD Cloud (coloured by topic) [10]

much anticipated 'Semantic Web'. They state that to date this fundamental prerequisite has not been met. They give credit to the Open Data movement for making royalty free interlinked RDF data sets such as Geonames, Wikipedia, dbpedia and the Digital Bibliography & Library Project (DBLP) bibliography available. They state that by doing so the Open Data Movement have prevented postponement of advancements in Semantic Web application development until the 'fundamental prerequisite' for the Semantic Web has been met. With regard to the associated benefit of interlinked RDF data existing on the Web, the authors highlight that "using these links, one can navigate from a computer scientist in dbpedia to her publications in the DBLP database, from a dbpedia book to reviews and sales offers for this book provided by the RDF Book Mashup (data set) or from a band in dbpedia to a list of their songs provided by Musicbrainz or dbtune".

RDF links are in the form of RDF triples (subject, predicate, object), where the subject is a URI reference in the namespace in one data set, the object is the Uniform Resource Identifier (URI) reference in the another data set and the predicate refers to the description of the link between the subject and the object [13]. Popular predicates for interlinking include *owl:sameAs*, which could be used to create a number of different names that refer to the same individual. For example, the individual Bob in one data set may be stated to be the same individual as Bob Smith in another data set. *foaf:based_near* links 'a "spatial thing" (anything that can be somewhere) to a point in space' [50]. Raimond et al.[44] use *foaf:based_near* to interlink a particular artist in the DBTune project to a location in the GeoNames data set, meaning the authors did not need to provide a location name along with latitude and longitude coordinates for each artist.

2.5 Summary

In this chapter, we reviewed the background of Linked Data, which acts as an introduction for Chapter 3.

Chapter 3

State of the Art Review

3.1 Introduction

The state of the art investigates the current situation in key areas of interest influencing this research. The state of the art chapter discusses current research in the areas of UCI, UCI tools, User Cognitive Support, and User Involvement.

3.2 Interlinking Data

3.2.1 Ontology Mapping

Ontology matching is considered a solution to the semantic heterogeneity problem [27]. It is concerned with finding correspondences between ontology entities and is an important operation in traditional applications including data warehousing, data integration and ontology integration. In such applications, matching is a prerequisite step which must be performed to running the actual system. Many emerging applications require the performing of matching operations at run time.

Interlinking data sets is analogous to ontology mapping, particularly in terms of the challenges which remain in relation to achievable accuracy of automated semantic interlinks/mappings and challenges which arise when involving humans with the aim of improving that accuracy. According to Euzenat, 'There remains an interesting path to follow concerning user involvement: relying on the application users in order to learn from them what is useful in the alignments under consideration' [27]. His view of the challenges facing ontology mapping are similar to the challenges facing the interlinking of data sets, consequently much research in this area is related to the topic of the research presented here.

Challenges facing Ontology Mapping include:

Discovering missing background knowledge - Lack of background knowledge about data can result in ambiguities where it comes to entities matched.

Uncertainty in Ontology Matching - Problems arise in applications where matches are performed dynamically as oftentimes, there is no precise correspondence or a correspondence identified by the matching algorithm is not specific enough. The application may require a 'best effort' match in these circumstances but there is a lot of uncertainty with regard to accuracy of such matches.

Matcher selection and self-configuration - Many matchers are available. Matchers can be accurate in some circumstances but not in others. In dynamic situations, there often arises the need to adapt and perform 'run time reconfiguration of a matcher by finding its most appropriate parameters, such as thresholds, weights, and coefficients'. [22]

User involvement - Automatic matching does not yield high quality data, research has indicated that user interaction (at design time) can improve the quality of ontology mappings. Challenges exist in relation to getting users to participate in the performing of tasks which improve mappings. [27]

Gold Standard - The Gold standard is considered to be the most effective evaluation technique for comparing mappings. It is 'a complete set of correct mappings as built by domain experts, in order to measure precision, recall, and F-measure', but is usually not available. [22] The Gold Standard is discussed with regard to this research in Chapter 6.

3.2.2 Automatic Interlinking

To date, the interlinking process has been largely automated. Linking algorithms may interlink resources based on string matching, common key matching or property based matching. There are a multitude of mapping/matching approaches available [48]

Mapping algorithms such as Quick Ontology Mapping (QOM) compare the set similarity between concept properties to help determine concept equality [29].

QOM is proposed by Ehrig et al. [25] as a way to 'trade off between efficiency and effectiveness of the mapping generation algorithms'. The authors claim that while it has a much lower runtime complexity than other approaches, loss of quality is 'marginal'. It uses a number of heuristics to calculate label similarity; edit distance (The edit distance between two strings is the number of operations required to transform one string into the other), substring matches, and exact string matches. QOM also compares internal and external structure similarity and the five heuristics are 'combined using a weighted sum and normalized into a single metric' [29].

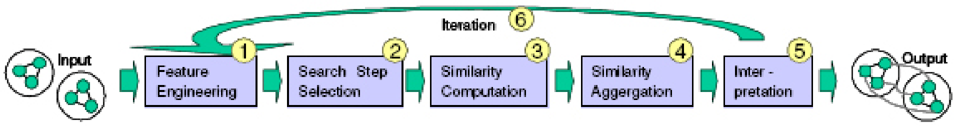


Figure 3.1: Quick Ontology Mapping (QOM) Process [25]

However, there exists a problem with the quality of interlinks/mappings generated by automated means. Attempts have proved to be only 'best-effort' mappings thus far in relation the quality of the interlinks. Automated allocation of interlinking between data sets can result in interlinking of data which may have some shared semantic meaning which may be deemed relevant by the machine interlinking the data. That interlinking however, may not be deemed to be truly relevant to a human user.

Much research has been undertaken with regard to the role of humans in improving the quality of said interlinks, however Raimond et al. [45] state that whilst defining manual interlinks may a viable solution for small data sets, it is impractical for large data sets. The authors use a similarity metric for automatic interlinking between artists in the Jamendo and Musicbrainz data sets, comparing the names of artists as well as the titles of their albums and songs and setting RDF links based on the similarity of artist resources. Several different methods are used for interlinking; naive approaches initially followed by a more elaborate algorithm, which not only examined similarity of the resources themselves, but also similarity of their neighbours. The authors concluded that the techniques they presented were 'far from perfect'. Evaluation of the automatic interlinking experiments showed that in many cases the quality of the interlinks generated suffered due to ambiguity of artist resources in either data set and/or due to the fact the algorithm used did not specify any heuristics to use if the ontologies differ.

Halb et al. [31] state that current interlinking algorithms in LOD data sets are largely based on templates and the quality of the interlinks generated by the algorithms is questionable with regard to their 'semantic strength' as a result. Hausenblas et al. [37] view automatically generated interlinks as being unreliable stating that 'there is no guarantee that the automatically generated interlinks are truly relevant'. They also suggest that a problem exists whereby the process of automatically generating interlinks is restricted to predefined data sets which means only a small portion of information published on the Semantic Web is actually considered when interlinks are being generated. Bürger et al. [17] claim that automatic interlinking methods yield fair results in cases where the resources being interlinked are textual and quality metadata is available to identify objects and their associated semantics. However, they do not deem automatic interlinking methods as being adequate for the interlinking of fine-grained multimedia content.

Further work by Bürger and Hausenblas focusses on automatic interlinking of multimedia resources. The authors point out that most automatic interlinking to date uses heuristics to determine when resources in data sets identify the same object, and that this interlinking, which is based on an RDF descriptions cannot be applied directly to multimedia objects. They claim that it is possible to adapt automated interlinking algorithms in order to add an analysis step where some information could be extracted from the multimedia object prior to the derivement of interlinks using software such as 'Henry', which they describe as 'An application of automatic interlinking of media fragments in the music domain'. However, they found the accuracy of derived interlinks using Henry debatable and largely dependent on the underlying analysis algorithms used. [38]

Falconer and Storey [30] state that most ontology mapping processes require user involvement. The authors along with Breslin et al. [15] cite problems stemming from local ambiguity with regard to language coupled with constraints on available data formats as issues which complicate the 'heterogeneous data mapping problem'. In the opinion of [30] owing to the above issues, it is highly unlikely that the mapping procedures will ever be completely automated and consequently they feel it is critical to have a human user 'in the loop' when it comes to said procedures. They state that much research to date has ignored user involvement focussing instead on algorithms to compute candidate matchings. They feel that this focus has resulted in tools and interfaces which require tremendous patience, as often a mass of data is returned, something which would support the view of Halpin [32], who states that for the majority of queries, the Semantic Web returns too much data. [30] also claim that expert understanding of the ontology domain, terminology, and semantics is often required to use existing tools effectively.

Much current research attempts to address the many shortcomings of completely automatic approaches to interlinking by way of human involvement in interlinking.

3.2.3 User Contributed Interlinking

User Contributed Interlinking (UCI), was introduced in [31, 37] as a new way to create semantic links. It is an approach to improving quality of interlinks between data sets through the inclusion of a human user in the interlinking process. The UCI approach to interlinking data sets is viewed by Halb et al. [31], [35] as 'enriching the data set with high-quality links'. [31] describe their experience of building the *riese* data set. The riese data set is an interlinked RDF version of the Eurostat data, which contains statistical data pertaining to the European Union. The authors implement the data set as a Semantic Web application which uses the 'follow-your-nose principle', a term used in [34] to describe an RDF property; a URI, interpreted as a hyperlink which leads to more data. The application is interlinked with other data sets in order to increase the usefulness of the data. The view of the authors is that a completely automated approach to interlinking would not be sufficient for the creation of quality interlinks and so, they aim to increase interlink quality by adding human users into the loop, allowing them to add semantic links to other data sets. The UCI is carried out in a 'Wikipedia-style' way, in terms of openness and ease of use, with the users specifying interlinks between resources using *rdfs:seeAlso*, *owl:sameAs* and *foaf:topic*. Whilst [45] consider the manual interlinking of data sets to be impractical, [31] only consider the manual interlinking of data sets to be impractical 'at first sight'. They feel that users can be encouraged to contribute to linked data, performing what is referred to as 'collaborative interlinking' processes in a way not too disimilar to the way Wikipedia has successfully evolved in recent years. One such 'semantic wiki' is MetaVidWiki (MVW) [38], which supports users in the addressing and interlinking between temporal multimedia fragments.

Conroy et al. [20] state that semantic mapping has been identified as one of 'the most time-consuming data management problems'. The authors state that 'new user-centric semantic mapping approaches, processes and tools are required' [20] and they emphasize the importance of usability evaluation over matching algorithm optimisation. They present a user driven approach to the process of semantic mapping. Three groups of users with varying levels of knowledge with regard to ontologies performed tagging based mapping tasks. The role of the users was to categorise automatically generated candidate matches between their personal ontologies and an ontology which described sports news content from Really Simple Syndication (RSS) feeds. The generated candidate matches were presented to the user using a natural language over a long time period. In contrast, this research needs to be carried out over a relatively short period of time. The authors conclude that the results of the study were positive and that users can achieve rich mappings using a user contributed tagging approach but that there was still a need for improvement and refinement. They also suggested mapping tasks may be deemed less intrusive by users if they were context-sensitive to what the user is doing and did not prompt the user to interact in a mapping process during periods where the user is busy.

With regard to UCI based interlinking methods and multimedia resources, [17] claims the first steps have already been taken and gives the example of 'Catch Me If You Can' (CaMiCatzee) [4]. CaMiCatzee is a multimedia interlinking concept demonstrator which uses Flickr as a base, allowing users to interlink still images, and also to query for persons using their FOAF documents, URIs or person names. According to [36] the objective of CaMiCatzee 'is to show how images from flickr can be interlinked with other data, such as person-related (FOAF) data, locations, and related topics'. At present images can only annotated using Flickr but there is no functionality for interlinking those images with other data, nor can the image be used by other Web applications.

3.2.4 User Contributed Interlinking Tools

Most current research with regard to use of UCI for the improvement of quality of interlinks between data sets describes mapping via annotations in order to add semantics between resources [6, 12, 15, 17, 30, 33, 36, 38, 44, 52].

3.2.4.1 irs

[35] implement a demonstrator which enables UCI. The demonstrator which is named ‘interlinking of resources with semantics’ (irs) enables the listing, adding, and removal of user-contributed semantic links between any resources identified through URIs. The riese Web application [31] also allows users to add and remove links to certain page data. Users specify the type of link e.g. “Same as” or “See also” via a drop down menu. The user contributed interlinks are stored in a named graph. It is possible to ‘understand *who* stated *what*’ as named graphs enable signing of graphs. The authors describe this ability to see who stated what as ’a simple version of provenance tracking’. Provenance tracking with RDF is important as where information comes from is considered as important as the information itself. Carroll et al. [18] view provenance tracking highly important for ascertaining the accuracy of information published on the Semantic Web. The signing of graphs may also influence user behaviour during the interlinking process in two ways; the user who perceives that that the interlinking information they are supplying may be analysed and is traceable back to them may think more ‘semantically’ before creating interlinks, which may lead to higher quality interlinks. In contrast, the user may be reluctant to perform interlinking tasks for those same reasons which may lead to reduced participation in interlinking tasks.

3.2.4.2 ODDLinker

Consens [21] uses ‘ODDLinker’, for the creation of interlinks in the LinkedMDB data set. The LinkedMBD data set contains over 3.5 million RDF triples pertaining to movies including interlinks to several other data sets on the LOD project cloud. It is considered to be a data set which has ‘dense interlinking’ and was awarded first prize in the Triplification Challenge having been deemed as making a ‘significant contribution’ to the Linked Open Data Community. ODDLinker allows ‘administrator’ users to create and maintain interlinks and is described as a ‘toolset under development which supports state-of-the-art join and linkage techniques’.

3.2.4.3 DART

Zhou et al. [52] present a browser based semantic mapping tool called ‘Dynamic, Adaptive, RDF-mediated and Transparent (DART)’ [51] for converting Linked Data on the Web so that it can be used by Semantic Web applications. The purpose of the tool is to enable users to define a semantic mapping from relational databases to RDF schemas. The tool enables the user to perform mappings using drag and drop functionality. Users can also perform visualisation mappings, annotate data sources and manage ontologies with this ‘easy-to-use’ tool. The tool can handle complex relational schema and is developed based on OAT (Openlink Ajax Toolkit) framework and Virtuoso Universal Server.

3.2.4.4 Active Learning Framework for Ontology Matching

In contrast, Shi et al. [46] consider the annotation step to be time-consuming. In their opinion ‘users are usually not patient enough to label thousands of concept pairs for the relevance feedback’ and consequently they focus their research on finding a way to minimise the number of user interactions required in a UCI based process whilst at the same time improving quality of the interlinks between ontologies. The authors propose an active learning framework for ontology matching. The aim of framework is to find the most informative candidate matches to query and to propagate user correction back to the ontology structure, using a ‘correct propagation algorithm’, thus improving matching accuracy.

3.3 User Cognitive Support

A vast amount of research has been undertaken in relation to cognitive support for users of software in the fields of computer science and psychology. For the purpose of my research I will focus solely on cognitive support for users of software which enables them to perform mappings based on semantic relations.

Falconer et al. [28] emphasize the importance of understanding decision making processes in order to satisfy user needs. They feel that understanding this process will enable them to introduce *cognitive support*, defined as ‘the introduction of external aids to support cognitive processes’ to mapping tools which will reduce the *cognitive load*, defined as ‘the load on working memory during problem solving’ experienced by users [29].

According to Falconer [30], “(a tools) usefulness is ultimately dependent upon (its) utility relating to cognition: i.e. to thinking, reasoning, and creating’. Cognitive support within a system used to perform interlinking would be to offload some of the users cognitive processes involved in performing a mapping task to the software.

[29] states that in the domain of ontology engineering, the main task areas which require cognitive support are *navigation*, *modeling*, and *verification*. Interfaces used for the creation of mappings generally contain a lot of information which can affect usability of a system. A cluttered screen may overwhelm a user, particularly a new user or user with little or no knowledge of ontologies, decreasing likelihood of user participation in performing of tasks. Navigation of ontologies and potential mapping correspondences between those ontologies needs to be as straightforward as can be, with the user assisted as much as is possible with regard to verification that their mappings are correct.

Downey [24], in part of her research into a semantic mediation system (SMS) used for automatic and semi-automatic merging of heterogenous scientific data sets, conducted a paper prototyping activity in order to gain a better understanding of expectations of users involved in annotation activities, determine the steps users take during the annotation process and understand the information required by that user to support decision making whilst annotating. The aim of the research was the superior design of annotation mechanisms. A small number of scientists were involved in the study and results of design activities 'revealed a trend toward simplicity'. According to the author 'users wanted to examine and understand the data set before annotating' (*navigation*). 'They also wanted to be able to access metadata about the data set' (*verification*). 'After examining the data set they wanted to keep the data displayed and launch the annotation dialog' (*modeling*), which would support the view of [29], with regard to tasks which require cognitive support.

[28] believe that user interfaces that offer more effective cognitive support will result in greater productivity gains where it comes to users performing mapping tasks than improvements to precision and recall in matching algorithms. The authors observe users performing mapping processes using two different tools, COMA++ and PROMPT, and based on analysis and user concerns propose preliminary cognitive support requirements for ontology mapping tools. Key findings of the study were that the problems users experience go beyond the processing of the algorithms. They found that users of the two tools had 'trouble remembering what they have looked at and executed, understanding output from the algorithm, remembering why they performed an operation, reversing their decisions, and gathering evidence to support their decisions'. They later use 'data from the observational study and further research into cognitive psychology to develop a theoretical framework describing mapping concepts relating to cognitive support'. The framework has four conceptual dimensions - user analysis and decision making, interaction, analysis and generation and representation. The authors use the above considerations in the design of their ontology mapping tool. They developed a PROMPT plugin called COGZ (Cognitive Support and Visualization for Human-Guided Mapping Systems), described by the authors as 'a user-interface plugin for the ontology management suite PROMPT. This tool introduces visualizations to support user cognition, filters to reduce mapping scope, mapping annotations, and the novel cognitive aid of a temporary mapping'.

Research by [24, 30], suggests that cognitive support of mapping tasks can be as simple as allowing the user to alter the level of granularity of information they are viewing (e.g. provision of a component which allows for expansion of a tree, thus providing more or less information as required), provision of non-modal dialogs and/or tabbed panes so that the user can easily toggle between screens displaying all information needed to make informed decisions, whilst also reducing the likelihood of cluttered screens.

3.4 User Involvement

The success of the UCI approach is highly dependent on user uptake, participation levels and easy-to-use tools. Interlinking processes must be non-complicated if the user is to perform interlinking tasks multiple times and if the interlinks created are to be accurate. In relation to the riese project [31] suggest that a more ‘appealing’ data set (compared to the statistical one used) may increase user uptake and participation levels with regard to UCI.

[38] suggest that manual methods for interlinking could be combined with Game Based Interlinking (GBI), in order to make interlinking of resources “fun”, increasing user participation and consequently also increasing the quality of interlinks between data sets. The authors list OntoTube, Peekaboom, or ListenGame which they describe as ‘exemplary’ games which ‘hide the complexity of the annotation process of videos, images or audio files respectively’. This is an interesting suggestion as repeat user involvement/participation is often a major challenge faced undertaking this type of research.

[6] feels that user tagging behaviour may be influenced by the design of the tagging system and he suggests trend prediction in folksonomy systems as a topic for future work. [20] suggest that mapping tasks may be deemed less intrusive by users if they were context-sensitive to what the user is doing and did not prompt the user to interact in a mapping process during periods where the user is busy. The user annotation study conducted in [24] concluded that there was a wide variation with regard to annotation terms selected by the participants and participants had differing ideas as to the scope to which they needed to annotate. The study also revealed that participants whom were very familiar with the annotation target selected terms differently to those whom were less familiar with an annotated target.

3.5 Summary

The current situation with regard to key areas of interest influencing this research have been investigated. Various current research in the areas of UCI, UCI tools, User Cognitive Support, and User Involvement have been reviewed. Reviewing of this work has served to identify a research gap for this dissertation.

Chapter 4

Design

4.1 Introduction

The following chapter outlines the major design decisions that contributed to the final implementation of the Improv-A-link demonstrator.

4.2 User Cognitive Support

As discussed in section 3.3, Downey [24], having evaluated users performing annotations, concluded that:

- Users wanted to examine and understand the data set before annotating.
- Users wanted to be able to access metadata about the data set.
- After examining the data set users wanted to keep the data displayed and launch the annotation dialog.

Similar research by Falconer [28], who observed users performing mapping tasks stated that:

- Users had trouble remembering what they had looked at and executed, understanding output from the algorithm, remembering why they performed an operation, reversing their decisions, and gathering evidence to support their decisions.
- Users could feel overwhelmed by a cluttered screen, particularly a new user or user with little or no knowledge of ontologies.

4.3 Requirements

This section outlines the User Interface (UI) requirements, the functional and non-functional requirements of the Improv-A-link demonstrator.

4.3.1 UI Requirements

Having considered the findings of [24] and [28], whom are considered leaders in the field of cognitive support for mapping tasks, the UI requirements for the demonstrator became apparent. It was decided that the main UI requirements should be as follows:

- Ease of Use
- Uncluttered Interface
- Decision Support
- Filtering of Information

4.3.2 Functional Requirements

Functional requirements capture intended use of a system. It is important that those requirements have been given adequate consideration prior to designing the architecture. Figure 4.1 depicts intended use of the Improv-A-link demonstrator.

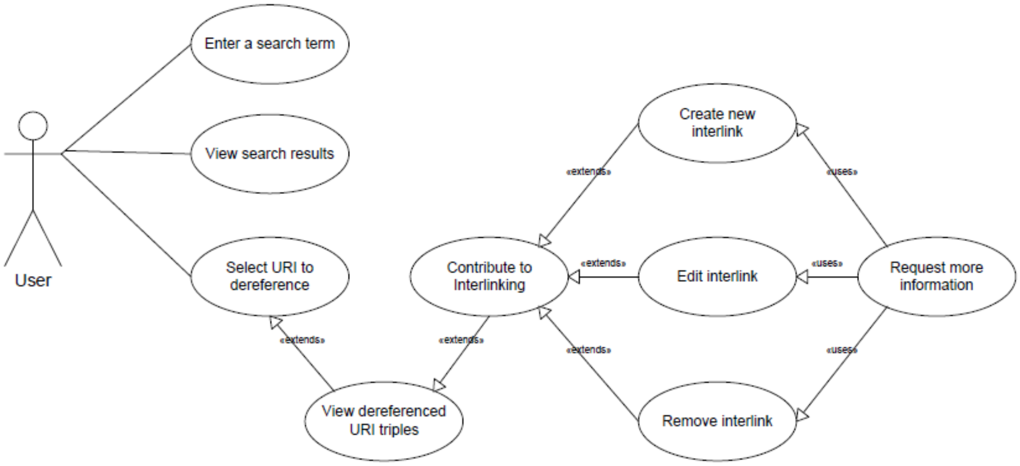


Figure 4.1: Use Case For a User Contributing Interlinks with the Improv-A-link demonstrator

4.3.3 Non-Functional Requirements

Non-Functional requirements are considered to be constraints in relation to the functional requirements. The two main non-functional requirements of the Improv-A-link demonstrator are:

- Performance — Semantic Web queries and reading and writing of RDF data can be slow. A slow system will inevitably lead to a frustrated user.
- Reliability — Reliability with regard to data storage is of huge importance. User evaluation participants can be difficult to recruit so it is vital to have reliable data storage.

4.4 Architecture

Design with regard to the architecture of the system was carried out in two phases. Firstly, a high level approach was chosen in order to define a structure which would support the requirements of the Improv-A-link demonstrator. Definition of the structure of the architecture during this phase was independent of programming languages and other technologies which might potentially be used. Secondly, a lower level approach was undertaken in order to define the actual programming language and technologies which support the structure designed in phase one.

4.4.1 Components

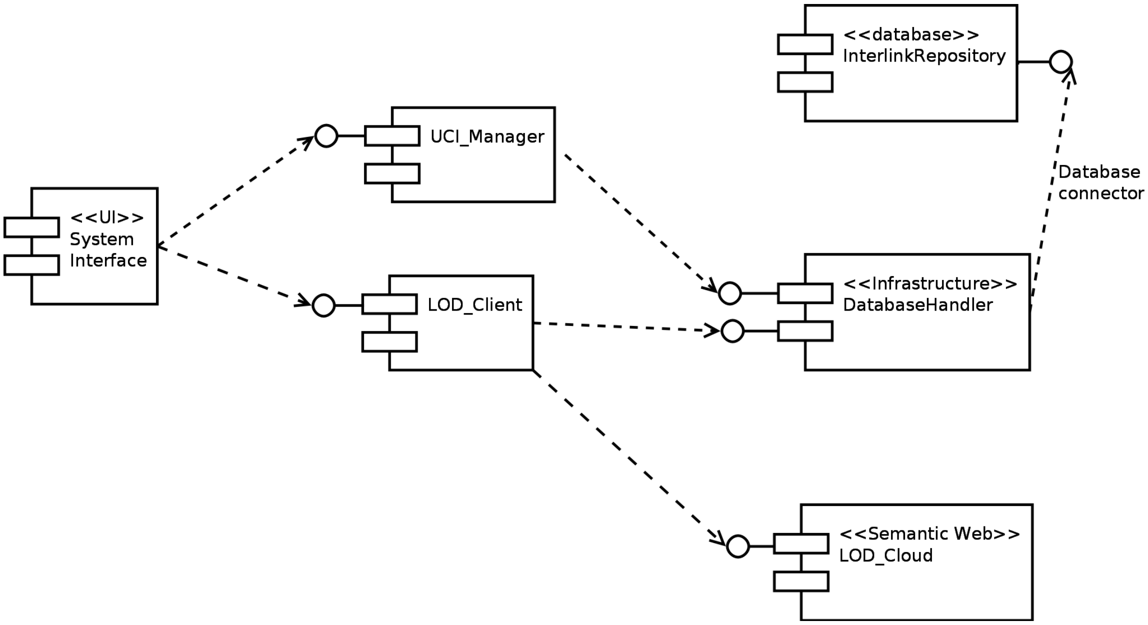


Figure 4.2: High Level View of the Components

Figure 4.2 shows the high level architecture design. The components consist of:

- System Interface - UI component which allows the user to interact with the system.
- UCI_Manager - a component for managing user's new, edit and remove actions.

- LOD_Client - a component for accessing interlinks from LOD data sets.
- InterlinkRepository - a component for persistent storage of interlinks.
- DatabaseHandler - a component which communicates with the InterlinkRepository
- LOD_Cloud - Refers to existing Linked Data data sets. Added to the component diagram for completeness purposes.

4.4.2 System Architecture

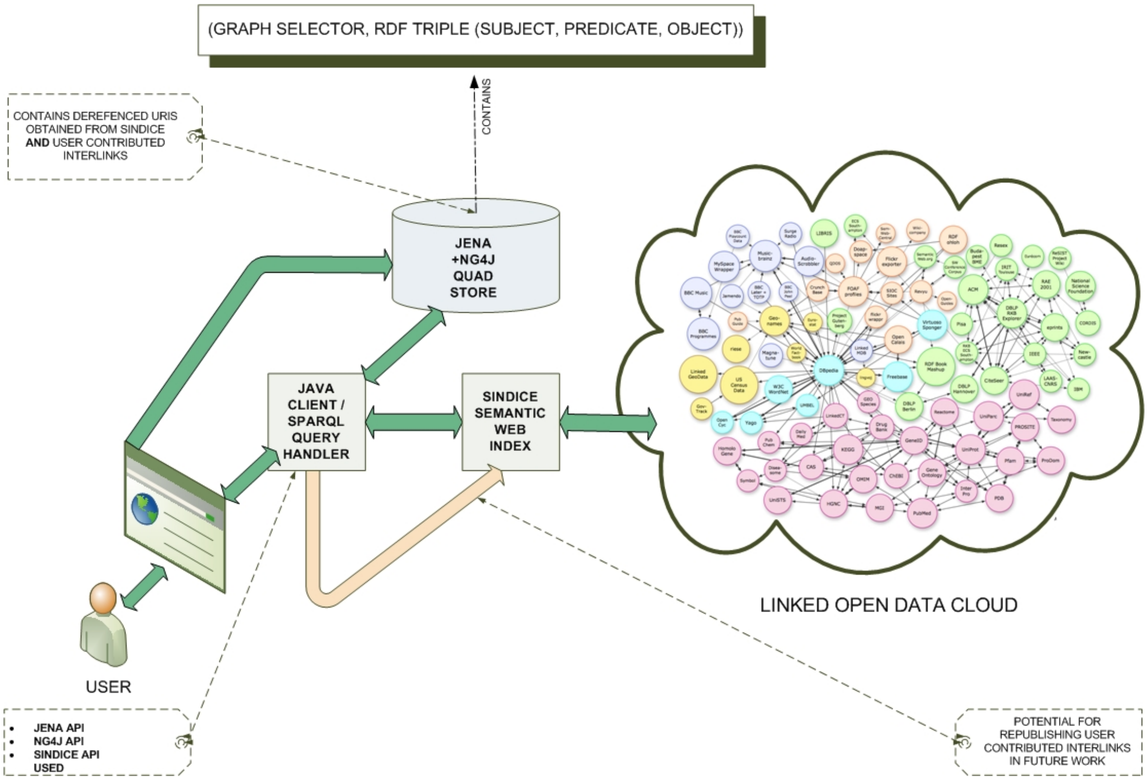


Figure 4.3: Architecture of the Improv-A-link demonstrator

Figure 4.3 shows the low level view of components which supports the structure shown in figure 4.2. The architecture consists of the following components.

4.4.2.1 Java Swing Interface

A Java Swing interface would afford a lot of flexibility with regard to the design of an interface where cognitive support requirements are key. It would also be easily integrated with the rest of the system.

4.4.2.2 Jena

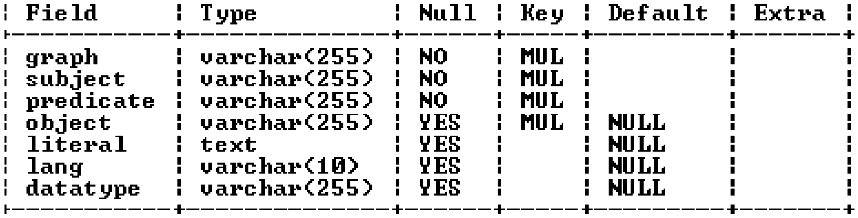
Jena is an open source (under the liberal BSD-style license) Java framework for building Semantic Web applications [47]. It was developed by Hewlett Packard and allows users to manipulate and query RDF graphs. The framework enables Sematic Web developers to do the following:

- Retrieve and parse an RDF file containing one or more graphs.
- Store the deserialized file in memory.
- Examine a dereferenced URI, iterating over each triple in turn, potentially accessing subject, predicate, object values.
- Examine only triples that meet specified criteria.
- Write a serialized version of a graph to a file.
- Persist a Jena model to a database.

Using Jena, a RDF graph is stored in a model which is essentially a set of Statements. “Methods are provided for creating resources, properties and literals”. Jena also provides methods for “adding statements to and removing statements from a model, and for querying a model” [40]. Jena 2.6.2 was the latest version of Jena at the time the demonstrator was designed.

4.4.2.3 NG4J - Named Graphs for Jena

NG4J is an open source extension to the Jena Semantic Web toolkit discussed in section 4.4.2.2 for parsing, manipulating and serializing sets of Named Graphs [3]. It is available under the same open source licence as Jena. According to [5], with NG4J, a NamedGraph-Set can can be manipulated by adding and removing entire Graphs, or by working with individual Quads”. Bizer et al. [11] state with regard to manipulating graphsets, that for some Semantic Web applications, a quad-centric view is more practical. The authors also state that NG4J supports such a view. Figure 4.4 shows the structure of the NG4J_quads table.



```
+-----------+--------------+------+-----+---------+-------+
| Field     | Type         | Null | Key | Default | Extra |
+-----------+--------------+------+-----+---------+-------+
| graph     | varchar(255) | NO   | MUL |         |       |
| subject   | varchar(255) | NO   | MUL |         |       |
| predicate | varchar(255) | NO   | MUL |         |       |
| object    | varchar(255) | YES  | MUL | NULL    |       |
| literal   | text         | YES  |     | NULL    |       |
| lang      | varchar(10)  | YES  |     | NULL    |       |
| datatype  | varchar(255) | YES  |     | NULL    |       |
+-----------+--------------+------+-----+---------+-------+
```

Figure 4.4: Default NG4J Quad Table

4.4.2.4 Sindice

Sindice is a Semantic Web crawler and indexer developed by the Digital Enterprise Research Institute (DERI). Oren and Tummarello [42] present a solution to the problem faced by Semantic Web developers with regard to 'the decentralised publication model'. Sindice aims to allow developers of Semantic Web applications find relevant sources of information automatically. According to Tummarello et al.[49], Sindice "crawls the Semantic Web and indexes the resources encountered in each source". Sindice provides an API to Semantic Web developers. This API allows developers to locate relevant data sources automatically and to integrate data from said sources into their applications. Sindice provides all results as RDF data and so is considered a "good citizen of the Linked Data web" [49]. Sindice would enable access to interlinks from a wide variety of domains. It would also provide the ability to map between arbitary strings and semantic web/linked data content/URIs. In this way it forms a human-centric entry point to the system.

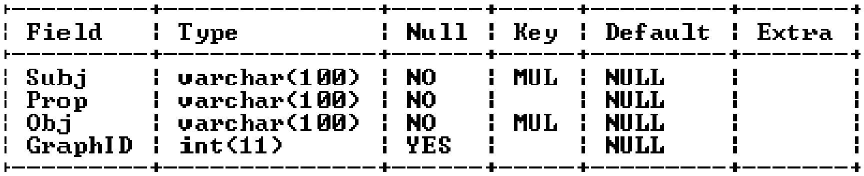
4.4.2.5 Persistent Storage

Jena and NG4J provide persistent storage of RDF data in relational databases.

Implementations are available for:

- Microsoft SQL Server
- MySQL
- PostgreSQL
- Oracle

"The Jena2 persistence subsystem implements the Jena Model interface providing persistence for models through use of a back-end relational database engine" [2]. The default Jena2 database layout uses a denormalized schema in which literals and resource URIs are stored directly in statement tables (see figure 4.5).



```
+---------+--------------+------+-----+---------+-------+
| Field   | Type         | Null | Key | Default | Extra |
+---------+--------------+------+-----+---------+-------+
| Subj    | varchar(100) | NO   | MUL | NULL    |       |
| Prop    | varchar(100) | NO   |     | NULL    |       |
| Obj     | varchar(100) | NO   | MUL | NULL    |       |
| GraphID | int(11)      | YES  |     | NULL    |       |
+---------+--------------+------+-----+---------+-------+
```

Figure 4.5: Default Jena2 Statement Table

The default NG4J quad table contains an additional 'graph' field which acts as a graph selector (see figure 4.4). The ng4j.db.NamedGraphSetDB constructor will create all the necessary tables if they do not exist.

4.4.2.6 MySQL Server 5.1

MySQL Server 5.1 is a *free*, *fast*, *reliable open source* database which *integrates* seamlessly with a number of programming languages and other technologies. MySQL was an ideal choice for a back-end database for the following reasons:

Free No money was available for purchasing a database.

Fast Reading and writing of RDF Graphs can be slow.

Reliable As discussed in section 4.3.3 reliability was considered a non-functional requirement.

Open source Peer reviewing of open source products leads to high quality.

Integrates The database needed to be compatible with Jena 2 and NG4J.

4.4.3 Architecture Verification

Sequence diagrams are used to document the flow of logic in a system and can prove a useful aide to validating logic and verifying that a proposed design can work. Figure 4.6 shows the usage scenario for part of the use case depicted in figure 4.1, whereby the user looks up a search term and proceeds to dereference a URI.

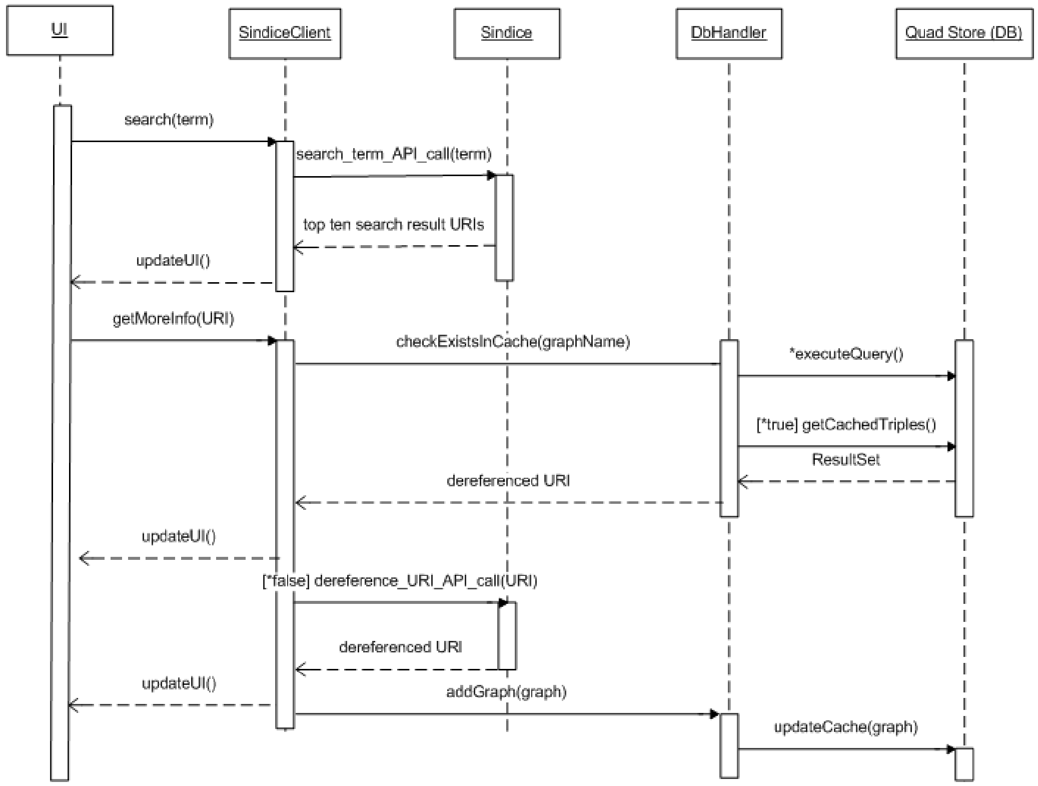


Figure 4.6: Sequence Diagram for Search Term Look Up and URI Dereferencing

4.4.4 Summary

In this chapter, the main areas where users needed cognitive support for interlinking tasks were summarised. Having considered those main areas, UI requirements which would adequately support users were identified. Functional and non-functional requirements of the demonstrator were discussed. Components needed to implement the system, firstly at a high level, and then on a more specific lower level, were outlined.

Chapter 5

Implementation

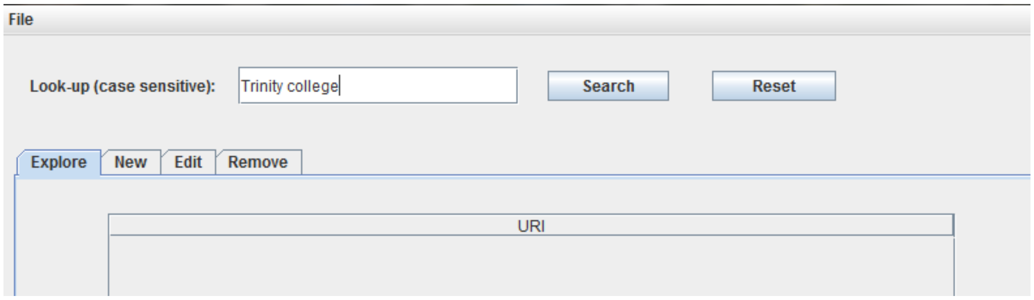
5.1 Introduction

This chapter details the work carried out during the implementation phase of the dissertation. The chapter also details how the main requirements outlined in sections 4.3.1 and 4.3.2 have been implemented.

5.2 Functionality

This section describes how the main functionality of the demonstrator was implemented.

5.2.1 Search



Trinity college

Figure 5.1: Searching for the term "Trinity college"

The search functionality was implemented using Sindice's developer API:

The string search term entered by the user is concatenated to the 'Term Search' REST API string.

```
searchUri = "http://api.sindice.com/v2/search?q=" +
userInput + "&qt=term&page=1";
```

Sindice 'Term Search' allows you to retrieve documents that are related to keywords and/or URIs. The page parameter is used to request the first page only, that is, the first ten results.

The following steps are required to request documents related to the search term in HTML, plain text, RDF/XML or JSON via content negotiation.

- Create an URL object from the 'searchUri' string representation.
- Use the abstract class *URLConnection* to read from the resource referenced by the URL.
- Set the request property to RDF/XML.

```
URL url = new URL(searchuri);
URLConnection urlc = url.openConnection();
urlc.setRequestProperty("Accept", "application/rdf+xml");
```

The statements from the RDF/XML serialization are then stored in a Jena model.

```
InputStream inputStream = urlc.getInputStream();
modelsearch.read(inputStream, searchuri);
```

A *jena.ref.model.StmtIterator* is used to update the UI with the string representation of the URI resources.

5.2.2 Dereferencing a URI

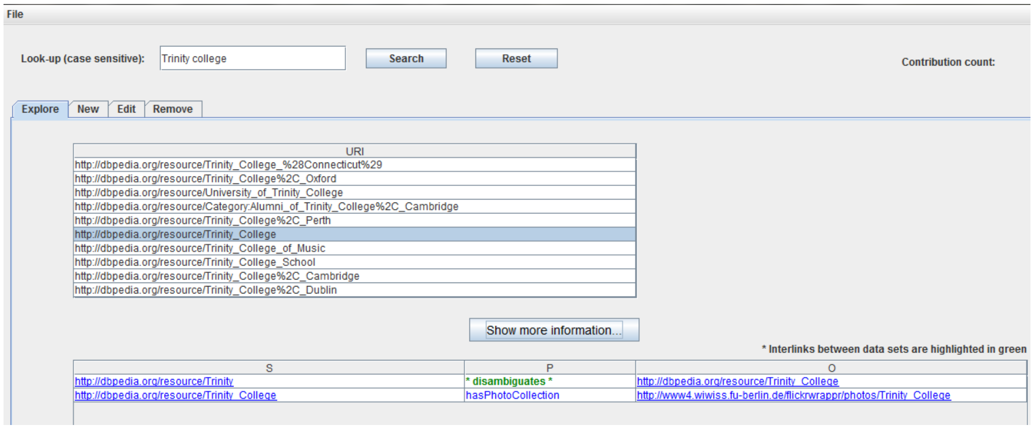
The Sindice developer API was used to dereference a URI selected by the user. The developer API provides two methods for dereferencing:

Cache API V3 to dereference using Sindices cached documents.

Sindice Live API to directly dereference URIs using web documents.

The cached API V3 approach was chosen as the dereferenced data was less likely to change over the course of the evaluation period. With this approach, it is also easy to see when the cached documents were updated. This ensures that all participants access the same data.

To dereference a URI, the selected URI is passed as a parameter to the *uriSearch* method.



Look-up (case sensitive): Trinity college

URI
http://dbpedia.org/resource/Trinity_College_%28Connecticut%29
http://dbpedia.org/resource/Trinity_College%2C_Oxford
http://dbpedia.org/resource/University_of_Trinity_College
http://dbpedia.org/resource/Category:Alumni_of_Trinity_College%2C_Cambridge
http://dbpedia.org/resource/Trinity_College%2C_Perth
http://dbpedia.org/resource/Trinity_College
http://dbpedia.org/resource/Trinity_College_of_Music
http://dbpedia.org/resource/Trinity_College_School
http://dbpedia.org/resource/Trinity_College%2C_Cambridge
http://dbpedia.org/resource/Trinity_College%2C_Dublin

* Interlinks between data sets are highlighted in green

S	P	O
http://dbpedia.org/resource/Trinity	* disambiguates *	http://dbpedia.org/resource/Trinity_College
http://dbpedia.org/resource/Trinity_College	hasPhotoCollection	http://www4.wiwiss.fu-berlin.de/flickrwrappr/photos/Trinity_College

Figure 5.2: Dereferencing the URI "http://dbpedia.org/resource/Trinity_College"

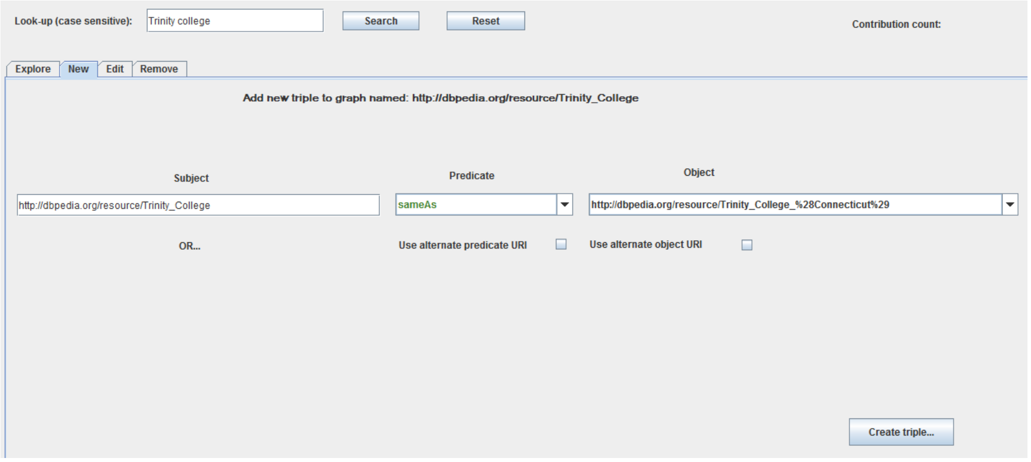
```
public void uriSearch(String uri) throws SQLException {

searchuri = "http://api.sindice.com/v2/cache?url=http://".concat(uri);

...

}
```

The database is first checked to see if a graph with same name as the parameter (URI) exists. If it exists, the UI is updated from the database cache displaying all triples associated with the selected URI to the user. If a graph of the same name does not exist, the URI is added as a parameter to the Sindice API call by concatenating it to the end of the call string. Content negotiation steps are the same as the steps involved for the 'Term Search' method above . The statements from the RDF/XML serialization are stored in a *NamedGraph* created using a factory method called on an *ng4j.impl.NamedGraphSetImpl* object. *The ng4j.db.NamedGraphSetDB* class is used to create a persistent *NamedGraphSet* from a database connection. The *addGraph(NamedGraph graph)* method is used to add the graph to the *graphSet*. The *ng4j.db.NamedGraphSetDB* constructor will create all the necessary tables if they do not exist.

5.2.3 Creating a New Triple

A new interlink can be added to a graph by accessing the *NamedGraphSet*, creating a *ng4j.Quad* object, specifying the URI as the first parameter — graphname, and then adding the quad to the set instance.

```
NamedGraphSet set = new NamedGraphSetDB(dbh.getConn2());
Quad quad = new Quad(Node.createURI(graphName),
Node.createURI(s.toString()),
```



Look-up (case sensitive): Trinity college

Add new triple to graph named: http://dbpedia.org/resource/Trinity_College

Subject: http://dbpedia.org/resource/Trinity_College

Predicate: sameAs

Object: http://dbpedia.org/resource/Trinity_College_%28Connecticut%29

Use alternate predicate URI ☐

Use alternate object URI ☐

Figure 5.3: Adding a New Triple to the Graph Named "http://dbpedia.org/resource/Trinity_College"

```
Node.createURI(p.toString()),
Node.createURI(o.toString()));
set.addQuad(quad);
set.close();
```

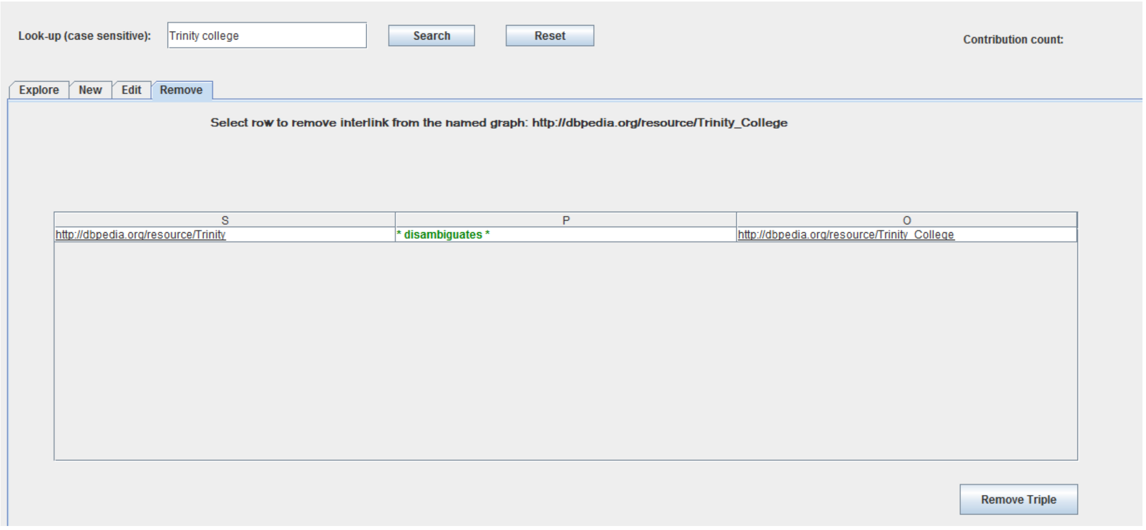
5.2.4 Removing a Triple

A triple can be removed from a graph in a similar way to that outlined in section 5.2.3 by specifying the name of the graph from which the triple should be removed and calling the *removeQuad(Quad quad)* method on the *NamedGraphSet* and passing the quad to be removed as a parameter. Figure 5.4 shows the UI tabbed pane which allows the user select a triple to remove.

```
NamedGraphSet set = new NamedGraphSetDB(db.getConn2());
Quad quad = new Quad(Node.createURI(uri),
Node.createURI(s.toString()),
Node.createURI(p.toString()),
Node.createURI(o.toString()));
set.removeQuad(quad);
set.close();
}
```

5.2.5 Editing an Interlink

Editing of an interlink was implemented in the following way. Figure 5.5 shows the UI tabbed pane which allows the user select a triple in order to edit that triples interlink.



Look-up (case sensitive): Trinity college

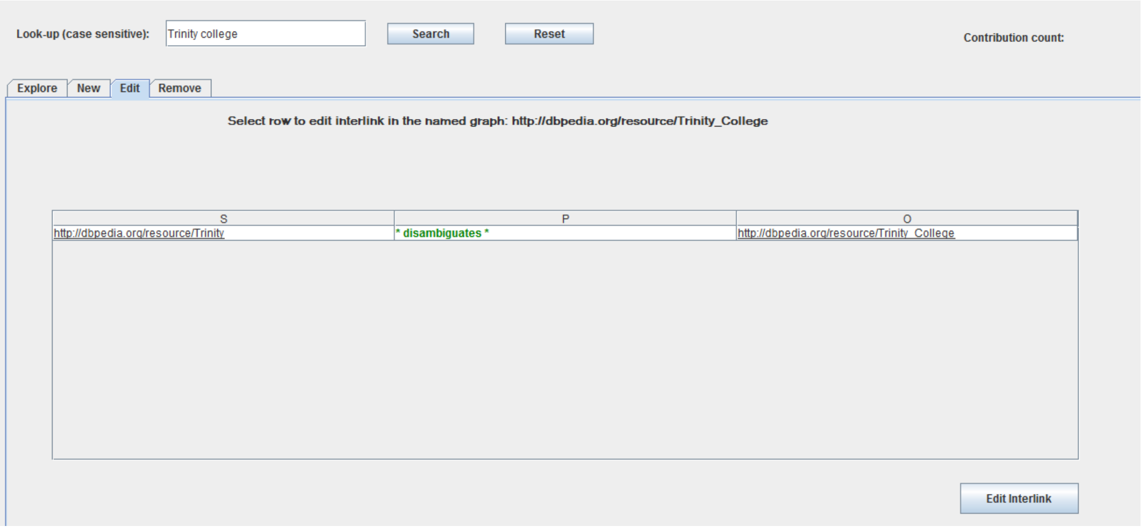
Contribution count:

Select row to remove interlink from the named graph: http://dbpedia.org/resource/Trinity_College

S	P	O
http://dbpedia.org/resource/Trinity	* disambiguates *	http://dbpedia.org/resource/Trinity_College

Figure 5.4: Remove Pane

```
public synchronized void editMapping(String uri, String newInterlink,
String s, String p, String o) {

...

NamedGraphSet set = new NamedGraphSetDB(db.getConn2());

Quad quad = new Quad(Node.createURI(uri),

Node.createURI(s),
Node.createURI(p),
Node.createURI(o));

Quad newQuad = new Quad(Node.createURI(uri),

Node.createURI(s),
Node.createURI(newInterlink),
Node.createURI(o));

set.removeQuad(quad);
set.addQuad(newQuad);
...
set.close();
}
```



Look-up (case sensitive): Trinity college

Select row to edit interlink in the named graph: http://dbpedia.org/resource/Trinity_College

S	P	O
http://dbpedia.org/resource/Trinity	* disambiguates *	http://dbpedia.org/resource/Trinity_College

Figure 5.5: Edit Pane

5.3 Cognitive Support Objectives

As discussed in section 1.2, one of the objectives was to build a demonstrator which enables a user to create interlinks between two Linked Data sets. The other objective was to reduce the cognitive load for users whilst performing interlinking tasks.

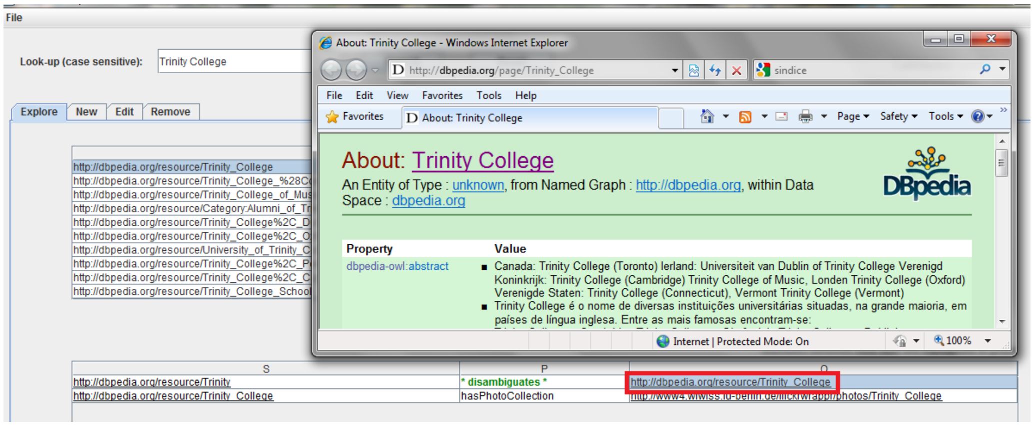
As discussed in section 4.3.1, in order to reduce cognitive load the demonstrator must have a clear interface and support the following:

5.3.1 Ease of Use

In order to ensure the system was easy to use, it was decided that users should be allowed to look up arbitary strings as opposed to URIs. User interaction with the system was "point and click" where possible. Combo boxes were used to allow the user to select a value as opposed to type one. Users were only required to type if they wanted to contribute an alternative predicate or object URI. In these cases, if they selected the radio button to add an alternative URI, the combo box was disabled and if the radio button was not selected the *alternateURI* text box was disabled (see figure 5.3).

5.3.2 Decision Support

Figure 5.6, shows how a users requirement for cognitive support for decision making whilst performing interlinking tasks has been implemented in the Improve-A-link demonstrator. Should a user be unsure as to what a URI relates to, they can click the URI hyperlink which launches a browser window. Having been presented with more information about the URI, the user has been supported with regard to the cognitive need for decision support.



http://dbpedia.org/resource/Trinity_College
http://dbpedia.org/resource/Trinity_College_%28C
http://dbpedia.org/resource/Trinity_College_of_Mus
http://dbpedia.org/resource/Category:Alumni_of_Tr
http://dbpedia.org/resource/Trinity_College%2C_D
http://dbpedia.org/resource/Trinity_College%2C_O
http://dbpedia.org/resource/University_of_Trinity_C
http://dbpedia.org/resource/Trinity_College%2C_P
http://dbpedia.org/resource/Trinity_College%2C_C
http://dbpedia.org/resource/Trinity_College_School

About: Trinity College

An Entity of Type : unknown, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Property	Value
dbpedia-owl:abstract	■ Canada: Trinity College (Toronto) Ierland: Universiteit van Dublin of Trinity College Verenigd Koninkrijk: Trinity College (Cambridge) Trinity College of Music, Londen Trinity College (Oxford) Verenigde Staten: Trinity College (Connecticut), Vermont Trinity College (Vermont) ■ Trinity College é o nome de diversas instituições universitárias situadas, na grande maioria, em países de língua inglesa. Entre as mais famosas encontram-se:

S	P	O
http://dbpedia.org/resource/Trinity	* disambiguates *	http://dbpedia.org/resource/Trinity_College
http://dbpedia.org/resource/Trinity_College	hasPhotoCollection	http://www4.wiwiss.fu-berlin.de/flickrwrappr/photos/Trinity_College

Figure 5.6: Decision Support for Interlinking Tasks

5.3.3 Filtering of Information/Uncluttered Screens

Filtering of information was required so as not to overwhelm the user with a large amount of data. To avoid this, all *sameAs*, *differentFrom*, *seeAlso*, *disambiguates* interlinks were highlighted to the user in a green font with an asterisk and the start and end of the string representation.

With the "explore" pane, users were unavoidably faced with a lot of data as the demonstrator needed to present them with all triples associated with dereferenced URI in order to be in a position to explore. Triples were displayed on a JTableOnScrollPane with the above mentioned interlinks again highlighted in green. The "explore pane" is shown in figure 5.2.2.

In the "edit" and "remove" panes, only *sameAs*, *differentFrom*, *seeAlso*, *disambiguates* interlinks were displayed as these are the predicate interlinks we wanted to focus on in this research. Having to scroll up and down on every pane would prove very cumbersome. Users could explore all triples on the "explore" pane anyway so it was not vital to display all the same data in this pane. Again, these interlinks were highlighted in green with asterisks, thus keeping the system consistent. The "edit" pane is shown in figure 5.5 and the "remove" pane is shown in figure 5.4.

5.4 Summary

This chapter described how the Improv-A-link demonstrator was implemented and how the cognitive needs of users were addressed in our implementation.

Chapter 6

Evaluation

6.1 Introduction

This chapter evaluates the extent to which users could improve the quality of Linked Data interlinks and also the level of cognitive support offered by the Improv-A-link demonstrator. The chapter is broken down as follows: Firstly, the chapter discusses the user study in detail, secondly the chapter focusses on quantitative evaluation in order to 'quantify' improvements made by the users to interlink quality and finally the chapter presents a qualitative evaluation of the experiences reported by users whilst they were using the Improv-A-link demonstrator to contribute to interlink quality.

6.2 User Evaluation

A user evaluation was carried out as part of the overall evaluation of this research. Research ethical approval was attained from the School of Computer Science and Statistics prior to commencement of the study. The recruitment method was an e-mail requesting evaluation participants, which was sent to the Knowledge and Data Engineering Group (KDEG) and to the MSc Computer Science (Networks & Distributed Systems) class at Trinity College, Dublin. Seven participants were recruited from these two groups and a further one participant was recruited from a non-Computer Science background. The eight participants recruited had varying levels of knowledge and familiarity with regard to Linked Data. The evaluations were held in the KDEG meeting room, O'Reilly Institute, at room 0.1 in 8 Westland Square and at home.

The evaluation was carried out on a one-to-one basis. The duration of the evaluation was 50 minutes per participant. The evaluation was broken up as follows:

- Briefing — 10 mins
- Interlinking tasks — 30 mins (10 minutes x 3 tasks)
- Debriefing — 10 mins

6.2.1 Briefing

All evaluation participants were briefed with the following:

- An short overview of Linked Data:
  - The Linked Data Cloud
  - Linked Data Principles
  - Definition of a Triple
- A definition of the meaning of common interlinking predicates:
  - owl#sameAs
  - owl#differentFrom
  - rdfs:seeAlso
  - dbpedia#disambiguates
- A simple example for each of the common interlinking predicates of its usage Example:

  `http://wordnet.rkbexplorer.com/id/word-William Jefferson Clinton`

  `‘sameAs’`

  `http://dbpedia.org/resource/Bill Clinton`
- A short tutorial detailing how the demonstrator is used.

Consent was also obtained from users in the briefing period and users were given a short tutorial on how to use the Improv-A-link demonstrator (see Appendix).

6.2.2 Interlinking Tasks

Users were then asked to carry out interlinking tasks in relation to three different search terms. The three terms were the same for each participant and were selected from different domain areas:

1. **Search Term**: Dublin
   Core Concept: http://dbpedia.org/resource/Dublin
2. **Search Term**: William Shakespeare
   Core Concept: http://dbpedia.org/resource/William_Shakespeare
3. **Search Term**: As We May Think Vannevar Bush 1945
   Core Concept: http://dbpedia.org/resource/As_We_May_Think

Participants were asked to follow the steps outlined in the tutorial for each of the search terms and core concepts, the core concept being a URI which they would be contributing to. Users were asked to spend no more than ten minutes on each term so that all terms would receive equal attention and data collected would not be skewed somewhat should a user run out of time before completing all tasks.

6.2.3 Debriefing

Upon completion of their interlinking tasks participants were required to complete a questionnaire which is discussed later in section 6.6.

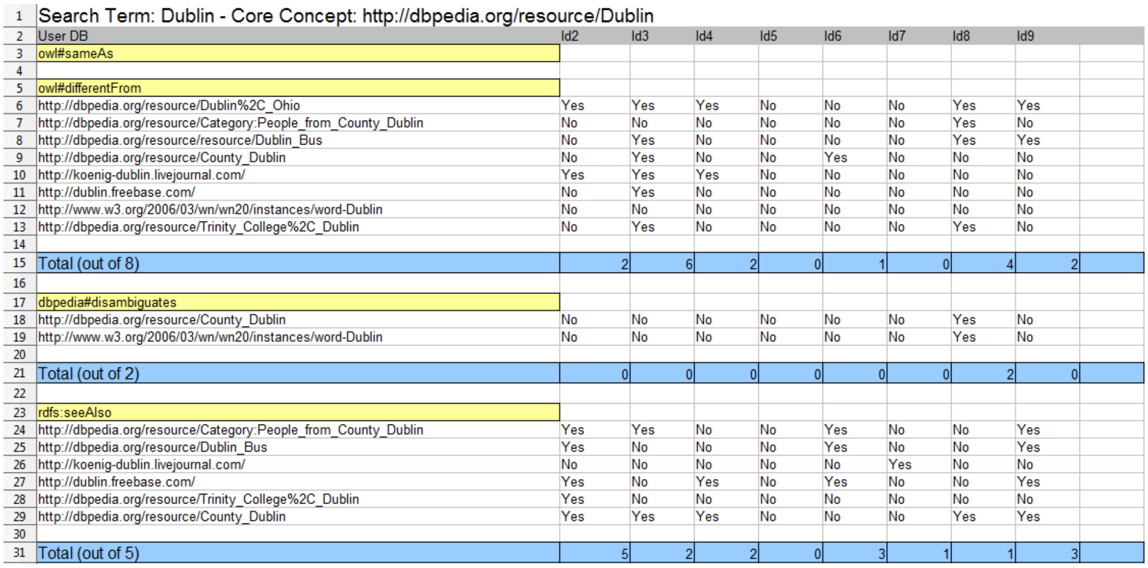
6.3 Gold Standard

A Gold Standard was drawn up by a Linked Data expert in relation to the three search terms and Core Concepts mentioned. The Linked Data expert is a Postdoc Research Fellow from the Knowledge and Data Engineering Group (KDEG), at Trinity College, Dublin. User contributions were compared against this Gold Standard in order to determine if those contributions were deemed to be quality interlinks. A critique of using a Gold Standard to measure quality interlinks however, is that semantic meaning or relatedness of URIs is generally open to human interpretation to some extent.

Should a Gold Standard be drawn up by two different Linked Data experts, it is unlikely that the two Gold Standards would be identical. This means that results may vary depending on the Gold Standard used. Producing a Gold Standard is also difficult, time consuming and “results remain sometimes controversial among domain experts” [41]. The Gold Standard consisted of interlinks shown in 6.3:

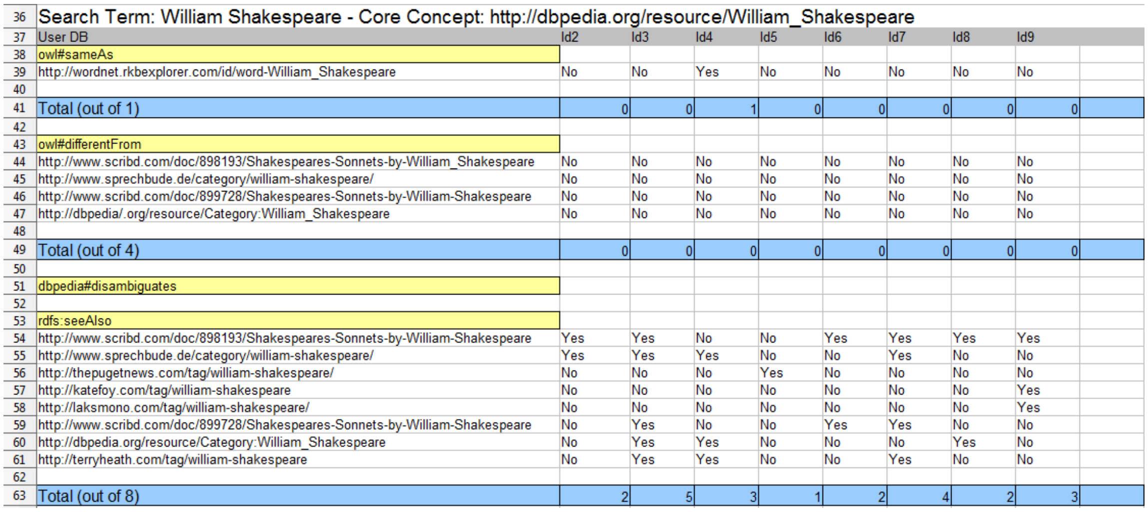
Table 6.1: Gold Standard Interlink Count by Predicate

Predicate	**Count**
owl#sameAs	4
owl#differentFrom	18
rdfs:seeAlso	19
dbpedia#disambiguates	2
Total	**44**



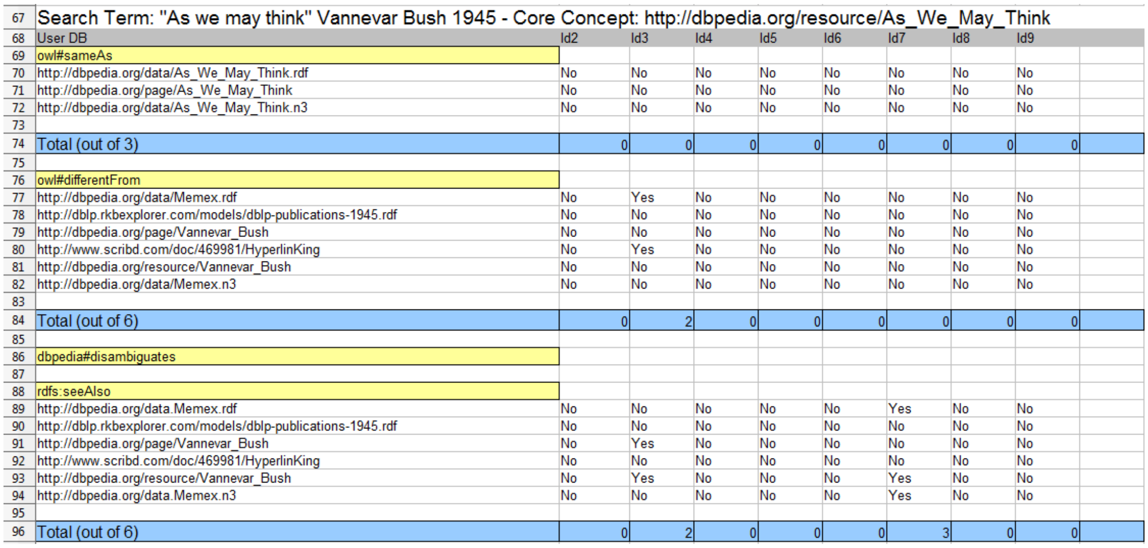
1	Search Term: Dublin - Core Concept: http://dbpedia.org/resource/Dublin								
2	User DB	Id2	Id3	Id4	Id5	Id6	Id7	Id8	Id9
3	owl#sameAs								
4									
5	owl#differentFrom								
6	http://dbpedia.org/resource/Dublin%2C_Ohio	Yes	Yes	Yes	No	No	No	Yes	Yes
7	http://dbpedia.org/resource/Category:People_from_County_Dublin	No	No	No	No	No	No	Yes	No
8	http://dbpedia.org/resource/resource/Dublin_Bus	No	Yes	No	No	No	No	Yes	Yes
9	http://dbpedia.org/resource/County_Dublin	No	Yes	No	No	Yes	No	No	No
10	http://koenig-dublin.livejournal.com/	Yes	Yes	Yes	No	No	No	No	No
11	http://dublin.freebase.com/	No	Yes	No	No	No	No	No	No
12	http://www.w3.org/2006/03/wn/wn20/instances/word-Dublin	No	No	No	No	No	No	No	No
13	http://dbpedia.org/resource/Trinity_College%2C_Dublin	No	Yes	No	No	No	No	Yes	No
14									
15	Total (out of 8)	2	6	2	0	1	0	4	2
16									
17	dbpedia#disambiguates								
18	http://dbpedia.org/resource/County_Dublin	No	No	No	No	No	No	Yes	No
19	http://www.w3.org/2006/03/wn/wn20/instances/word-Dublin	No	No	No	No	No	No	Yes	No
20									
21	Total (out of 2)	0	0	0	0	0	0	2	0
22									
23	rdfs:seeAlso								
24	http://dbpedia.org/resource/Category:People_from_County_Dublin	Yes	Yes	No	No	Yes	No	No	Yes
25	http://dbpedia.org/resource/Dublin_Bus	Yes	No	No	No	Yes	No	No	Yes
26	http://koenig-dublin.livejournal.com/	No	No	No	No	No	Yes	No	No
27	http://dublin.freebase.com/	Yes	No	Yes	No	Yes	No	No	Yes
28	http://dbpedia.org/resource/Trinity_College%2C_Dublin	Yes	No	No	No	No	No	No	No
29	http://dbpedia.org/resource/County_Dublin	Yes	Yes	Yes	No	No	No	Yes	Yes
30									
31	Total (out of 5)	5	2	2	0	3	1	1	3

Figure 6.1: Comparing Gold Standard Matches and User Derived Matches for Search Term ‘Dublin’



36	Search Term: William Shakespeare - Core Concept: http://dbpedia.org/resource/William_Shakespeare								
37	User DB	Id2	Id3	Id4	Id5	Id6	Id7	Id8	Id9
38	owl#sameAs								
39	http://wordnet.rkbexplorer.com/id/word-William_Shakespeare	No	No	Yes	No	No	No	No	No
40									
41	Total (out of 1)	0	0	1	0	0	0	0	0
42									
43	owl#differentFrom								
44	http://www.scribd.com/doc/898193/Shakespeares-Sonnets-by-William_Shakespeare	No	No	No	No	No	No	No	No
45	http://www.sprechbude.de/category/william-shakespeare/	No	No	No	No	No	No	No	No
46	http://www.scribd.com/doc/899728/Shakespeares-Sonnets-by-William-Shakespeare	No	No	No	No	No	No	No	No
47	http://dbpedia/.org/resource/Category:William_Shakespeare	No	No	No	No	No	No	No	No
48									
49	Total (out of 4)	0	0	0	0	0	0	0	0
50									
51	dbpedia#disambiguates								
52									
53	rdfs:seeAlso								
54	http://www.scribd.com/doc/898193/Shakespeares-Sonnets-by-William-Shakespeare	Yes	Yes	No	No	Yes	Yes	Yes	Yes
55	http://www.sprechbude.de/category/william-shakespeare/	Yes	Yes	Yes	No	No	Yes	No	No
56	http://thepugetnews.com/tag/william-shakespeare/	No	No	No	Yes	No	No	No	No
57	http://katefoy.com/tag/william-shakespeare	No	No	No	No	No	No	No	Yes
58	http://laksmono.com/tag/william-shakespeare/	No	No	No	No	No	No	No	Yes
59	http://www.scribd.com/doc/899728/Shakespeares-Sonnets-by-William-Shakespeare	No	Yes	No	No	Yes	Yes	No	No
60	http://dbpedia.org/resource/Category:William_Shakespeare	No	Yes	Yes	No	No	No	Yes	No
61	http://terryheath.com/tag/william-shakespeare	No	Yes	Yes	No	No	Yes	No	No
62									
63	Total (out of 8)	2	5	3	1	2	4	2	3

Figure 6.2: Comparing Gold Standard Matches and User Derived Matches for Search Term ‘William Shakespeare’



67	Search Term: "As we may think" Vannevar Bush 1945 - Core Concept: http://dbpedia.org/resource/As_We_May_Think								
68	User DB	Id2	Id3	Id4	Id5	Id6	Id7	Id8	Id9
69	owl#sameAs								
70	http://dbpedia.org/data/As_We_May_Think.rdf	No	No	No	No	No	No	No	No
71	http://dbpedia.org/page/As_We_May_Think	No	No	No	No	No	No	No	No
72	http://dbpedia.org/data/As_We_May_Think.n3	No	No	No	No	No	No	No	No
73									
74	Total (out of 3)	0	0	0	0	0	0	0	0
75									
76	owl#differentFrom								
77	http://dbpedia.org/data/Memex.rdf	No	Yes	No	No	No	No	No	No
78	http://dblp.rkbexplorer.com/models/dblp-publications-1945.rdf	No	No	No	No	No	No	No	No
79	http://dbpedia.org/page/Vannevar_Bush	No	No	No	No	No	No	No	No
80	http://www.scribd.com/doc/469981/HyperlinKing	No	Yes	No	No	No	No	No	No
81	http://dbpedia.org/resource/Vannevar_Bush	No	No	No	No	No	No	No	No
82	http://dbpedia.org/data/Memex.n3	No	No	No	No	No	No	No	No
83									
84	Total (out of 6)	0	2	0	0	0	0	0	0
85									
86	dbpedia#disambiguates								
87									
88	rdfs:seeAlso								
89	http://dbpedia.org/data.Memex.rdf	No	No	No	No	No	Yes	No	No
90	http://dblp.rkbexplorer.com/models/dblp-publications-1945.rdf	No	No	No	No	No	No	No	No
91	http://dbpedia.org/page/Vannevar_Bush	No	Yes	No	No	No	No	No	No
92	http://www.scribd.com/doc/469981/HyperlinKing	No	No	No	No	No	No	No	No
93	http://dbpedia.org/resource/Vannevar_Bush	No	Yes	No	No	No	Yes	No	No
94	http://dbpedia.org/data.Memex.n3	No	No	No	No	No	Yes	No	No
95									
96	Total (out of 6)	0	2	0	0	0	3	0	0

Figure 6.3: Comparing Gold Standard Matches and User Derived Matches for Search Term: “‘As We May Think” Vannevar Bush 1945’

6.4 Quantitative Evaluation

Information retrieval performance measures such as Precision, Recall and F-Measure are commonly used and widely accepted in the evaluation of ontology matching algorithms [26].

With regard to figure 6.4, in this research:

A represents all cases where a user deleted or failed to create an interlink which was present in the Gold Standard.

B represents all cases where a user contributed an interlink which was present in the Gold Standard.

C represents all cases where a user created an interlink but the interlink was not present in the Gold Standard.

D represents all cases where a user contributed to wrong graph.

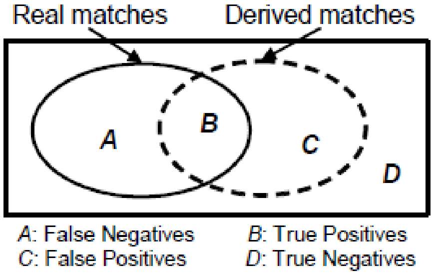


Figure 6.4: Comparing Gold Standard Matches and User Derived Matches [23]

6.4.1 Precision

Precision is used to test for exactness. It measures the ratio of correctly found correspondences (true positives) over the total number of returned correspondences (true positives and false positives).

Precision is calculated using the formula 6.1 [23]

$$\text{Precision} = \frac{|B|}{|B| + |C|} \tag{6.1}$$

The average Precision of participants who took part in the evaluation was 48.68 %. As you can see in figure 6.5, the user with ID 6 was well above average with regard to Precision achieving 80 % in the evaluation. This was calculated by substituting the following values into formula 6.1

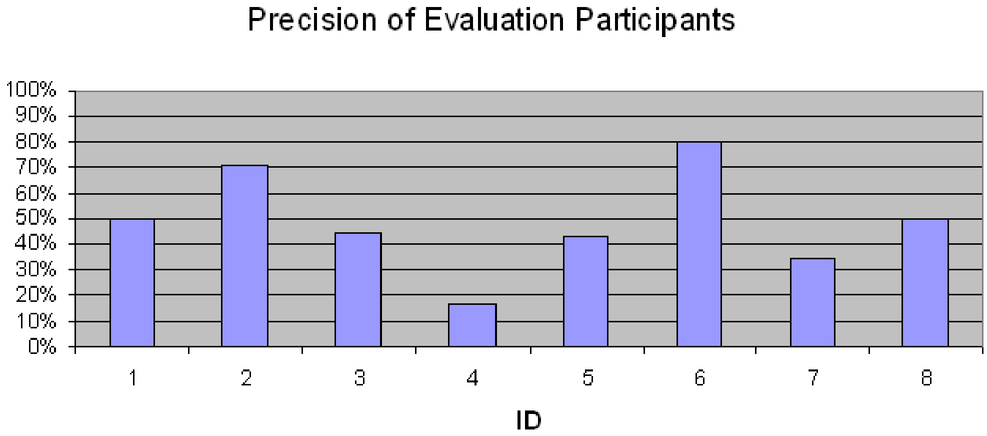


Figure 6.5: Precision

$$\text{Precision}_{\text{User ID 6}} = \frac{|8|}{|8| + |2|} = \frac{|8|}{|10|} = 80\,\%$$

This means that 80 % of the time, when user ID 6 contributed an interlink, that interlink was present in the Gold Standard.

∴ User ID 6 was "exact" 80 % of the time that User ID 6 contributed.

6.4.2 Recall

Recall is used to measure completeness. It measures the ratio of correctly found correspondences (true positives) over the total number of expected correspondences (false negatives and true positives).

Recall is calculated using the formula 6.2 [23]

$$\text{Recall} = \frac{|B|}{|A| + |B|} \tag{6.2}$$

The average Recall of participants who took part in the evaluation was 18.75 %. In figure 6.6, we can see that User ID 2 had the highest recall which was calculated by substituting User ID 2's values into formula 6.2 as shown in 6.4.2:

$$\text{Recall}_{\text{User ID 2}} = \frac{|17|}{|27| + |17|} = \frac{|17|}{|44|} = 39\,\%$$

This means that User ID 2 found 39 % of the total number of interlinks which were on the Gold Standard list.

∴ User ID 2 attained 39 % "completeness".

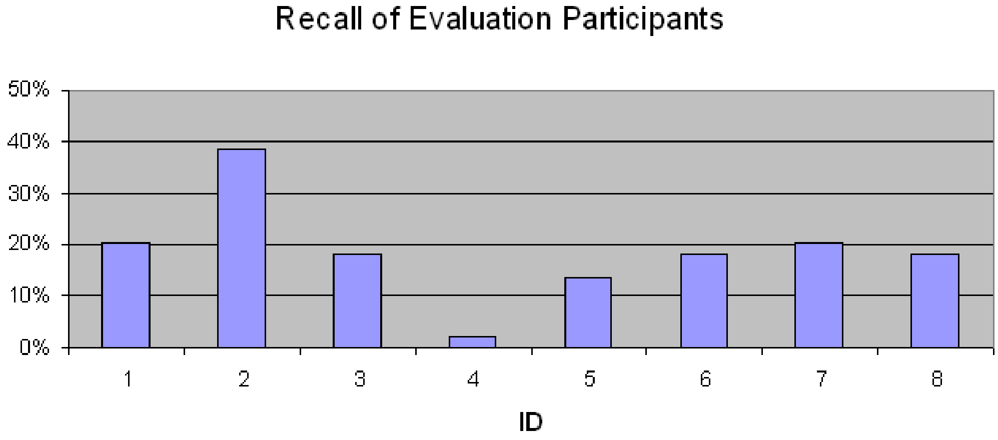


Figure 6.6: Recall

6.4.3 F-Measure

According to Do et al.[23], Precision and Recall alone cannot be used to assess quality. Recall can be maximised at the expense of a poor precision and likewise a high precision can be achieved at the expense of a poor recall. F-Measure is the weighted harmonic mean of Precision and Recall. In the evaluation of this research equal importance was placed on Precision and Recall. F-Measure is calculated using the formula: 6.3 [23]

$$\text{F-Measure}(\alpha)^1 = \frac{\text{Precision} * \text{Recall}}{(1-\alpha) * \text{Precision} + \alpha * \text{Recall}} \tag{6.3}$$

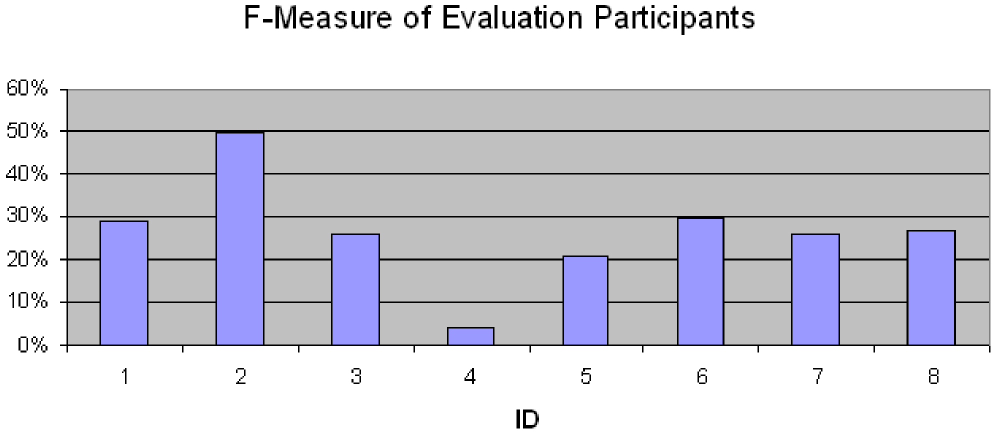


Figure 6.7: F-Measure

---

[1] α set to 0.5 for equal importance

Looking at figure 6.7, we can see that user ID 4 has the lowest F-Measure. The average F-Measure achieved by participants was 26.44 %, however, user ID 4 achieved a significantly lower than average F-Measure — 4 %. This makes sense as this user only achieved 17 % Precision and 2 % Recall. Substituting these values into the formula given in 6.3,

$$\begin{aligned}\text{F-Measure}_{\text{User ID 4}}(0.5) &= \frac{0.17 * 0.02}{(1 - 0.5) * .17 + 0.5 * 0.02}\\ &= \frac{0.0034}{(0.5) * .17 + 0.01}\\ &= \frac{0.0034}{0.09}\\ &= 0.0377\end{aligned}$$

∴ The harmonic mean of Precision, that is, “exactness” and Recall, that is, “completeness” for User ID 4 is 4 %.

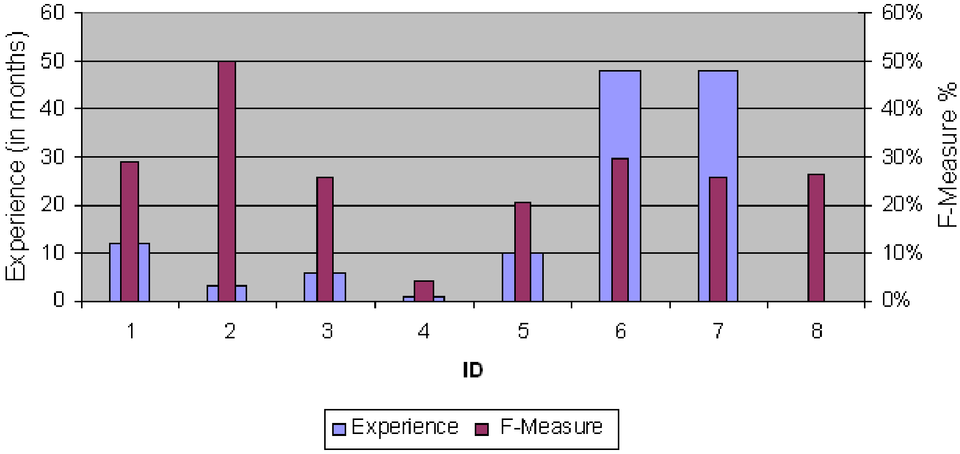


Figure 6.8: Participant F-Measure Relative to Participant Experience

6.4.4 Discussion

User ID 6 had 4 years experience of Linked Data and the highest Precision: 80 % (Recall was 18 %, F-Measure was 30 %).

User ID 2 had only 3 months Linked Data experience but had the highest Recall: 39 % and highest F-Measure — 50 % (Precision was 71 %). This participant is known to work as a software quality assurance tester. This could be the reason this user had the highest Recall which measures completeness. A software tester is generally considered to be conscientious and methodical in the carrying out of tasks.

User ID 5 rated themselves a 5 on a scale (1–6) with regard to familiarity with Linked Data. This participant had the lowest Precision (17 %), Recall (2 %) and F-Measure (4 %)

of all participants in the evaluation.
User ID 8 was completely unfamiliar with Linked Data and had no experience. Despite this, Precision and F-Measure was slightly above average, 50 % and 27 % respectively, compared to the average, which was 49 % and 26 % respectively. User ID 6 and User ID 7 had the most experience of all users — 4 years. User ID 8 had a higher F-Measure than User ID 7. User ID 6 also only did marginally better with an F-Measure of 30 %.
User ID 8 is from a non-computer science background and had no experience or familiarity with regard to Linked Data upon commencement of the evaluation. The participant was given the same briefing information, overview of Linked Data, demonstrator tutorial as all participants from a computer science background. This could suggest either of the following:

1. There is no correlation between participant experience and participant Precision, Recall and F-Measure.
2. The possibility of users with a higher level of experience and familiarity was achieving better results than others was reduced owing to the time limitation/not enough cognitive support for advanced interlinking.

With the question in relation to correlation between participant experience and participant Precision, Recall and F-Measure arising from the fact that a user with no experience achieved a similar F-Measure as two users with four years experience, we felt it important to carry out a statistical analysis to determine if a correlation did in fact exist. This analysis is discussed in section 6.4.5.

6.4.5 Correlation

Correlation is a measure of the relation between two or more variables [1]. It is calculated using the formula 6.4.

$$\text{Correlation} = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}} \tag{6.4}$$

The correlation coefficient of Participant Experience and Participant F-Measure is 0.076 which would indicate that there is an extremely weak positive correlation. As one can see in figure 6.9, the trendline in the scatter diagram is almost horizontal, indicating that as experience increases, F-measure increases, but only marginally.

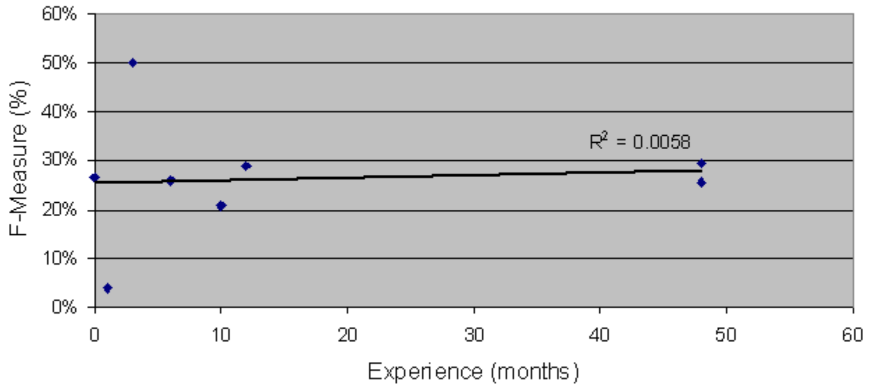


Figure 6.9: Correlation between Participant Experience and Participant F-Measure

6.4.6 User Contribution Count

Each participants "contribution count" was recorded for two purposes. Firstly, the contribution count was displayed to the participant with the aim of motivating them to contribute more. Secondly, the contribution count was used to record the class of interlinking performed by the participant:

- New
- Edit
- Remove

Table 6.2: Interlinks by Class of Contribution

User	**new**	**edit**	**remove**
ID 2	18	1	0
ID 3	20	2	2
ID 4	21	1	1
ID 5	1	17	2
ID 6	19	0	0
ID 7	19	0	3
ID 8	27	1	0
ID 9	16	1	0
Totals	**141**	**23**	**8**

This information was stored in the database. The idea being that this information may indicate which class of interlinking task required more cognitive support. Figure 6.10 and table 6.4.6 show the breakdown of User Contributed Interlinks by task.

As one can see, most of the participant contributions – 82 % were "new" interlinks which is an extremely high proportion of new interlinks.

For the core concepts specified in 6.2.2 participants could potentially edit/remove 23 existing interlinks in total.

The tutorial (see Appendix) suggested participants compare each core concept to 9 other URIs, creating new interlinks where appropriate, which may potentially lead to 27 new interlinks.

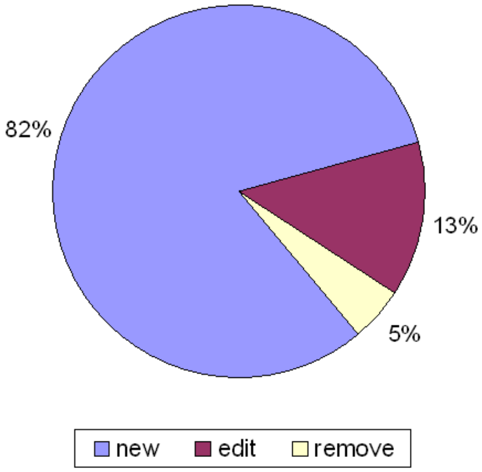


Figure 6.10: Breakdown of User Contributed Interlinks by Task Type

6.5 Discussion

The high proportion of new interlinks could be due to one or more of the following:

1. Users started with the "new" task and ran out of time.
2. Users were not confident enough to question the accuracy of existing interlinks.
3. Existing interlinks were already accurate.
4. Hyperlinks on edit/remove panes did not lead to more information which meant a decision could not be reached.
5. There was more cognitive support provided for creating new interlinks than edit/remove interlink.

6.6 Qualitative Evaluation

This section contains a qualitative evaluation of the experiences reported by users whilst they were using the Improv-A-link demonstrator to contribute to interlink quality.

6.6.1 System Usability Scale (SUS)

The System Usability Scale (SUS) shown in table 6.3, was used as part of the Qualitative Evaluation process. SUS is a simple ten item questionnaire used to gather feedback in relation to user interfaces [16].

Table 6.3: System Usability Scale Results

#	Statement	Mean Value	Mean Description
1.	I think that I would like to use this website frequently	3.00	Indifferent
2.	I found this website unnecessarily complex	2.44	Mildly disagree
3.	I thought this website was easy to use	3.67	Indifferent
4.	I think that I would need assistance to use this website	2.22	Mildly disagree
5.	I found the various functions in this website were well integrated	3.78	Indifferent
6.	I thought there was too much inconsistency in this website	1.44	Strongly disagree
7.	I would imagine that most people would learn to use this website very quickly	3.44	Indifferent
8.	I found this website very cumbersome/awkward to use	2.33	Mildly disagree
9.	I felt very confident using this website	4.22	Mildly agree
10.	I needed to learn a lot of things before I could get going with this website	2.11	Mildly disagree

6.6.2 User Evaluation Questionnaire Feedback

With regard to participants familiarity with Linked Data/Ontology Matching upon commencement of the evaluation:

Participants claimed to have levels of Linked Data experience ranging from none up to four years (see figure 6.11).

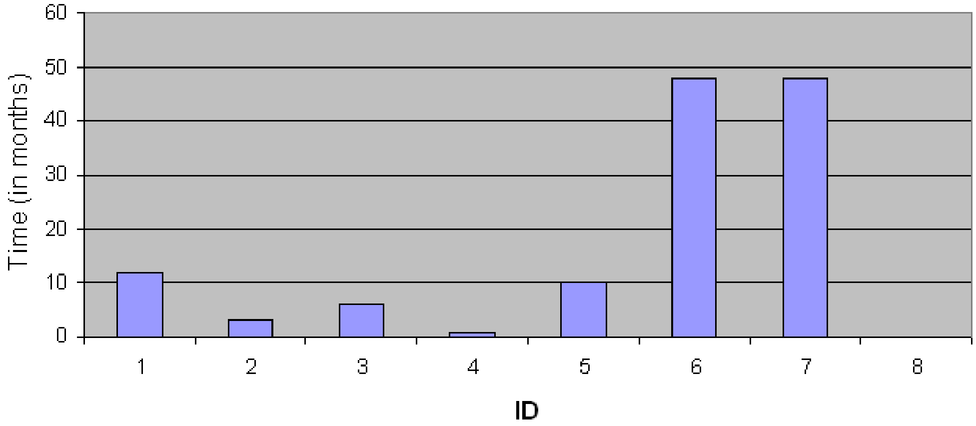


Figure 6.11: Participant Experience of Linked Data/Ontology Matching

On a scale of 1–6, the majority of participants rated themselves four or above with regard to Linked Data familiarity (see figure 6.12).

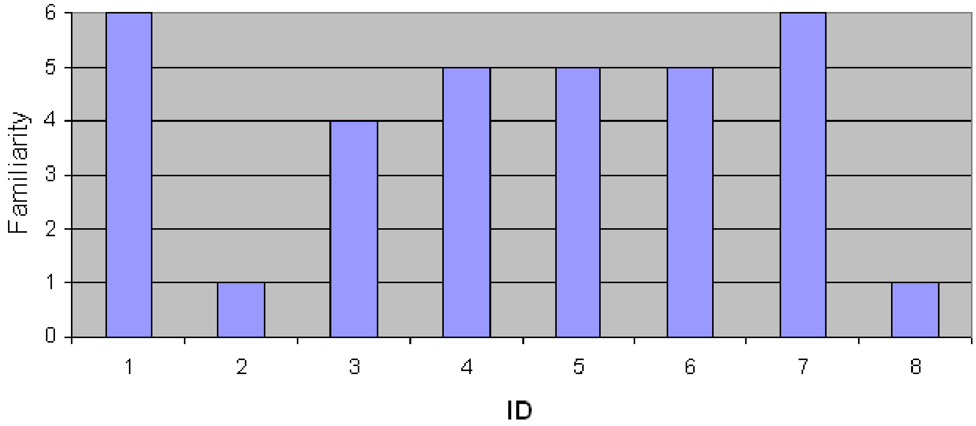


Figure 6.12: Participant Familiarity with Linked Data/Ontology Matching

The user annotation study conducted in [24] concludes that ‘making use of the annotations was only as good as the quality of the annotations themselves’ which would suggest that the tool plays an important part in the quality of interlink.

6.7 Discussion

Overall, the results of the System Usability Scale suggest that no significant problems were experienced by evaluation participants using the demonstrator.

The scale indicated the following:

With regard to **ease of use**, the mean description for statement 3. "I thought this system was easy to use", the mean result was "indifferent", that is, most users neither strongly agreed nor strongly disagreed that the system was easy to use, which would indicate no significant problems were experienced by evaluation participants using the demonstrator.

With regard to **time needed to become familiar** with using the system, the mean description for statement 7. " I imagine that most people would learn to use this system quickly", the mean result again was "indifferent" no significant problems were experienced by evaluation participants using the demonstrator.

With regard to **prior knowledge required** to get going, the mean description was "mildly disagree", which would indicate no user, including those with little Linked Data knowledge felt the need to have prior knowledge in order to be able to use the system.

User feedback gathered from open ended questions in the questionnaire highlighted three main issues with the demonstrator which led to some frustration:

- The necessity to switch back and forth, clicking on "Show More Information" with the the "Core Concept" URI selected to ensure the new or edited triple was added to the correct graph "eats up time".
- Very long URIs were "difficult to read" in "edit" and "remove" tab tables.
- Hyperlinks which did not lead to more information (page which has been removed or link is to an rdf or n3 file).

Elimination of the above issues would inevitably lead to improved cognitive support which may lead to higher F-Measures for future participants.

Participants also suggested the following improvements to the demonstrator:

- "a select subject function" that was different to the "show more information" button may solve the need to "re-expand" the graph prior to performing interlinking tasks.
- The ability to follow hyperlinks whilst on the new tab as opposed to having to open the explore tab.
- More predicates in the comboBox.
- An integrated browser to preview current subject/object.

Launching of the browser was problematic in many cases according to questionnaire feedback. The browser was supposed to provide users with cognitive support for decision making. If the user was unsure what a resource was they could click on a hyperlink to get more information. The main problem was with regard to URIs which led to raw RDF files, jrdf files, n3 files or .css files. Clicking on these hyperlinks, a user would be presented with a Windows File Download dialog asking “Do you want to save this file, or find a program online to open it?”. This caused confusion for some users and might explain the low number of matches with the Gold Standard recorded for search term — “As we may think” Vannevar Bush 1945, as this was the term which most of these files were noticed. Special handling to render the file may help users with Linked Data experience in these cases, but may provide little support for those with no experience as they may remain as confused as before when presented with an RDF or n3 file.

One user asked in relation to difficulty he had selecting a box to tick for SUS statement 1. “I think that I would like to use this website frequently”, “Like why would I use it?, I don’t get it, why would I want to use it again”. This seems a very fair question and highlights the fact that users may not see any benefit to contributing interlinks, at least no immediate benefit to contributing interlinks.

6.8 Summary

In this chapter the extent to which users could improve the quality of Linked Data interlinks and the level of cognitive support offered by the Improv-A-link demonstrator was evaluated. Quantifiable improvements made by the users to interlink quality were measured using Precision, Recall and F-Measure. Qualitative methods determined how supported participants were whilst contributing to interlink quality using the System Usability Scale (SUS) and questionnaire. The correlation between between Participant Experience and Participant F-Measure was also investigated.

Chapter 7

Conclusion

7.1 Introduction

This chapter discusses the main conclusions of this dissertation and suggests ideas for future work

7.2 Conclusions

The user evaluation indicates that there is an extremely weak positive correlation between F-Measure of participants and participant experience. This would suggest that Improv-A-link provided enough cognitive support to enable participants with little or no Linked Data/Ontology experience to achieve an F-Measure similar to experienced participants. If participants with no experience can be supported to to achieve an F-Measure similar to experienced participants, one might conclude that anyone can contribute to the interlinking of data sets on the Linked Data cloud.

However, more cognitive support for participants who had Linked Data experience prior to commencing the evaluation may have led to higher F-Measures for those participants. Two participants with Linked Data experience suggested in the questionnaire that more predicates in the combo box would make the system better. There was a text box where the user could enter their own predicate URI but in reality, even those with experience are unlikely to remember full predicate URIs and the inclusion of a search function for looking up predicates 'LIKE' X predicate may have supported these participants more. The two participants highlighted their need for more cognitive support. One might assume that with the inclusion of more predicates, and provided those additional predicates had been considered in the Gold Standard, that these two users would have attained higher F-Measures. That said, there is nothing to say that additional predicates would not have assisted non-experienced users to attain higher F-Measures also.

Results may have been more varied without the time constraint per term. In any user study, it is difficult to determine what results may have been influenced by the way the study was designed.

Users spent 10 minutes per search term. Most users seemed to lose interest with regard to interlinking toward the end of the evaluation. If User Contributed Interlinking is to make a difference to the quality of Linked Data interlinks, cognitive support for interlinking should be considered along with approaches that focus on user retention, such as Game Based Interlinking.

Improving the quality of Linked Data interlinks is a very slow process. Users took 30 minutes to improve three interlinks with an average F-Measure of 26 %. Estimates suggest that that there are 150 million Linked Data interlinks. One might question the practicality of adding a human 'in the loop' with the intention of improving the quality of interlinks.

7.3 Future Work

This section suggests potential related research areas which could be investigated further.

7.3.1 Advanced User Evaluation

It would be interesting to make small changes to the code to improve cognitive support, having taken into account user feedback and conduct another study. It would also be interesting to see if adding more cognitive support for advanced users or allowing users unlimited time to complete tasks would alter user evaluation results.

7.3.2 Gold Standard

It might also be interesting to see how a Gold Standard impacts on these type of studies. If a Gold Standard was drawn up by two different Linked Data experts, it would be interesting in my opinion to compare a users contributions to the two Gold Standards evaluating the difference between Precision, Recall and F-Measure in both cases.

7.3.3 Multiple Users

A multi-user demonstrator may be an interesting area to research. Allowing many users to contribute to one central cache may be interesting as users may have different ideas of how a specific subject relates to a specific object. Potential for lots of overwriting of contributions which could be recorded and evaluated.

7.3.4 Publishing User Improved Interlinks Back to the LOD Cloud

Publishing of the user improved interlinks back to the LOD Cloud would make a very small contribution to the quality of Linked Data interlinks. In the case of the participant who could not see the point of using Improv-A-Link again, maybe the publishing of that participants interlinks back to the cloud might offer some encouragement to use the demonstrator again. That said, publishing of interlinks is unlikely to be instantaneous in the near future. Interlinks will most likely have to be verified by the publishers of the Linked Data data sets that those interlinks refer to.

7.3.5 Maintaining Interlink Quality Over Time

There exists a problem in that Linked Data on the LOD Cloud is constantly evolving. The question arises as to whether the interlinks contributed by a user are still accurate when a Linked Data set is updated. In the case of this research, the NG4J Quad Store cache needs to be kept up to date to be kept relevant, however, it is complicated to re-cache only the data from Sindice pertaining to a URI. The ng4j_quads table is highly normalized and some kind of 'flag' may be required to be stored with each quad to identify interlinks which were contributed by a user. Comparisons would need to be made to prevent duplicates being stored upon updating plus all existing User Contributed Interlinks stored in the cache may need to be re-checked by a user to ensure that they are still accurate following each update of the cached data from Sindice.

7.3.6 Interlinking with Incentives

Much research has highlighted the need for human involvement in interlinking processes. The interlinking process, regardless of the level of cognitive support afforded by a tool, can be extremely repetitive and tedious. A 'games with a purpose approach' may increase the likelihood that users will want to get involved and stay involved over a longer time period [17].

Bibliography

[1] Basic statistics. Online. URL `http://www.statsoft.com/textbook/basic-statistics/#Correlationsa`. Last Accessed: 6 September 2010.

[2] Jena2 database interface. Online. URL `http://jena.sourceforge.net/DB/index.html`. Last accessed: 2 August 2010.

[3] Named graphs/ semantic web interest group. Online, 2008. URL `http://www.w3.org/2004/03/trix/`. Last accessed: 10 August 2010.

[4] Camicatzee. Online, 2009. URL `http://sw.joanneum.at/CaMiCatzee/`. Last accessed: 1 April 2010.

[5] Ng4j - named graphs api for jena. Online, 2010. URL `http://www4.wiwiss.fu-berlin.de/bizer/ng4j/`. Last accessed: 25 August 2010.

[6] Fabian Abel. The benefit of additional semantics in folksonomy systems. In *PIKM '08: Proceeding of the 2nd PhD workshop on Information and knowledge management*, pages 49–56, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-257-3. doi: http://doi.acm.org/10.1145/1458550.1458560.

[7] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC+ASWC 2007)*, pages 722–735, November 2008. doi: 10.1007/978-3-540-76298-0_52.

[8] Tim Berners-Lee. Linked data-design issues, June 2009. URL `http://www.w3.org/DesignIssues/LinkedData.html`. Last accessed: 25 March 2010.

[9] Tim Berners-Lee. The next web of open, linked data. `http://www.youtube.com/watch?v=OM6XIICm_qo`, February 2009. TED Talks.

[10] Chris Bizer. The lod cloud, July 2009. URL `http://www4.wiwiss.fu-berlin.de/bizer/pub/lod-datasets_2009-07-14_colored.png`. Last accessed: 3 April 2010.

[11] Chris Bizer, Richard Cyganiak, and E. Rowland Watkins. Ng4j named graphs api for jena. In *ESWC*, 2005.

[12] Christian Bizer. The emerging web of linked data. *Intelligent Systems, IEEE*, 24(5): 87–92, Sept.-Oct. 2009. ISSN 1541-1672. doi: 10.1109/MIS.2009.102.

[13] Christian Bizer, Tom Heath, and Tim Berners-Lee. The story so far. In T. Heath, M. Hepp, and C. Bizer, editors, *Linked Data*, 2009.

[14] Christian Bizer, Tom Heath, Tim Berners-Lee, and Michael Hausenblas. 3rd international workshop on linked data on the web (ldow2010). Online, April 2010. Intro Slides.

[15] John G. Breslin, Simon. Scerri, Stefan. Decker, Hak Lae. Kim, and Hong-Gee. Kim. The state of the art in tag ontologies: A semantic model for tagging and folksonomies. In *Proc. Intl Conf. on Dublin Core and Metadata Applications*, 2008.

[16] John Brooke. Sus - a quick and dirty usability scale. URL `http://www.usabilitynet.org/trump/documents/Suschapt.doc`. Last accessed: 6 April 2010.

[17] Tobias Bürger and Michael Hausenblas. Interlinking multimedia - principles and requirements. In *Proceedings of the First International Workshop on Interacting with Multimedia Content on the Social Semantic Web, co-located with SAMT 2008*, 2008.

[18] Jeremy J. Carroll, Christian Bizer, Pat Hayes, and Patrick Stickler. Named graphs, provenance and trust. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 613–622, New York, NY, USA, 2005. ACM. ISBN 1-59593-046-9. doi: http://doi.acm.org/10.1145/1060745.1060835.

[19] Semantic Web Company. The web must be a web, not a series of isolated data islands, April 2008. URL `http://www.semantic-web.at/1.36.resource.235.x22-the-web-must-be-a-web-not-a-series-of-isolated-data-islands-x22.htm`. Last accessed: 1 April 2010.

[20] Colm Conroy, Rob Brennan, Declan O'Sullivan, and David Lewis. User evaluation study of a tagging approach to semantic mapping. In *6th Annual European Semantic Web Conference (ESWC2009)*, pages 623–637, June 2009. URL `http://data.semanticweb.org/conference/eswc/2009/paper/142`.

[21] M.P. Consens. Managing linked data on the web: The linkedmdb showcase. In *Latin American Web Conference, 2008. LA-WEB '08.*, pages 1–2, Oct. 2008. doi: 10.1109/LA-WEB.2008.30.

[22] Isabel F. Cruz, Flavio Palandri Antonelli, and Cosmin Stroe. Efficient selection of mappings and automatic quality-driven combination of matching methods. In *In Proc. 4th International Workshop on Ontology Matching (OM-2009), collocated with ISWC-2009*, 2009.

[23] Hong Hai Do, Sergey Melnik, and Erhard Rahm. Comparison of schema matching evaluations. In Akmal B. Chaudhri, Mario Jeckle, Erhard Rahm, and Rainer Unland, editors, *Web, Web-Services, and Database Systems*, volume 2593 of *Lecture Notes in Computer Science*, pages 221–237. Springer, 2002. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.11.4792&rep=rep1&type=pdf`.

[24] Laura Downey. Designing annotation mechanisms with users in mind: A paper prototyping case study from the scientific environment for ecological knowledge (seek) project. In *International Semantic Web User Interaction Workshop*, 2006.

[25] Marc Ehrig and Steffen Staab. Qom - quick ontology mapping. In *The Semantic Web ISWC 2004*, pages 683–697. Springer, 2004.

[26] Jérôme Euzenat. Semantic precision and recall for ontology alignment evaluation. In *IJCAI*, pages 348–353, 2007.

[27] Jérôme Euzenat and Pavel Shvaiko. Ten challenges for ontology mapping. Technical Report DISI-08-042, University of Trento, August 2008. Also: in Proceedings of The 7th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE), 2008.

[28] Sean M. Falconer. Cognitive support for human-guided mapping systems. Technical report, 2007.

[29] Sean M. Falconer. *Cognitive support for semi-automatic ontology mapping.* PhD thesis, University of Victoria, 2009.

[30] Sean M. Falconer and Margaret-Anne Storey. A cognitive support framework for ontology mapping. In *The Semantic Web*, volume 4825/2008, pages 114–127, 2008. doi: 10.1007/978-3-540-76298-0_9.

[31] Wolfgang Halb, Yves Raimond, and Michael Hausenblas. Building linked data for both humans and machines. In *Linked Data on the Web (LDOW2008)*, 2008. URL `http://data.semanticweb.org/workshop/LDOW/2008/paper/10`.

[32] Harry Halpin. Is there anything worth finding on the semantic web? In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 1065–1066, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-487-4. doi: http://doi.acm.org/10.1145/1526709.1526858.

[33] Oktie Hassanzadeh, Anastasios Kementsietsidis, Lipyeow Lim, Renée J. Miller, and Min Wang. A framework for semantic link discovery over relational data. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1027–1036, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. doi: http://doi.acm.org/10.1145/1645953.1646084.

[34] Michael Hausenblas. Exploiting linked data to build web applications. In *IEEE Internet Computing*, volume 13, pages 68–73, Piscataway, NJ, USA, 2009. IEEE Educational Activities Department. doi: http://dx.doi.org/10.1109/MIC.2009.79.

[35] Michael Hausenblas and Wolfgang Halb. Interlinking of resources with semantics. In *5th European Semantic Web Conference (ESWC2008)*, 2008.

[36] Michael Hausenblas and Wolfgang Halb. Interlinking multimedia data. In *Linking Open Data Triplification Challenge at the International Conference on Semantic Systems (I-Semantics08)*, 2008.

[37] Michael Hausenblas, Wolfgang Halb, and Yves Raimond. Scripting user contributed interlinking. In *4th Workshop on Scripting for the Semantic Web (SFSW2008)*, 2008. URL `http://data.semanticweb.org/workshop/scripting/2008/paper/6`.

[38] Michael Hausenblas, Tobias Brger, Raphal Troncy, and Yves Raimond. Interlinking multimedia: How to apply linked data principles to multimedia fragments. In *Proceedings of Linked Data on the Web (LDOW2009), co-located with the 18th International World Wide Web Conference (WWW2009), Madrid, Spain, April 20th 2009 (forthcoming)*, 2009.

[39] Tom Heath. Linked data - connect distributed data across the web. URL `http://linkeddata.org/faq`. Last accessed: 6 Dec 2009.

[40] HP. com.hp.hpl.jena.rdf.model ; interface model. Online, 2009. URL `http://jena.sourceforge.net/javadoc/com/hp/hpl/jena/rdf/model/Model.html`. Last accessed: August 8 2010.

[41] Antoine Isaac, Lourens van der Meij, Stefan Schlobach, and Shenghui Wang. An empirical study of instance-based ontology matching. In *ISWC/ASWC*, pages 253–266, 2007.

[42] Eyal Oren and Giovanni Tummarello. A lookup index for semantic web resources. In *SFSW*, 2007.

[43] Sean B. Palmer. The semantic web: An introduction. Online, 2009. URL `http://infomesh.net/2001/swintro/`. Last accessed: 9 September 2010.

[44] Y. Raimond, C. Sutton, and M. Sandler. Interlinking music-related data on the web. In *IEEE Multimedia*, volume 16, pages 52–63, April-June 2009. doi: 10.1109/MMUL.2009.29.

[45] Yves Raimond, Christopher Sutton, and Mark Sandler. Automatic interlinking of music datasets on the semantic web. In *Linked Data on the Web workshop (LDOW)*, volume 16, pages 52–63, 2008.

[46] Feng Shi, Juanzi Li, Jie Tang, Guotong Xie, and Hanyu Li. Actively learning ontology matching via user interaction. In *ISWC 2009: Proceedings of the 8th International Semantic Web Conference, Chantilly, VA, USA*, volume 5823, pages 585–600. Springer, 2009.

[47] The HP Semantic Web Team. Hp labs semantic web research. Online, 2009. URL `http://www.hpl.hp.com/semweb/`. Last accessed: 15 August 2010.

[48] H Thomas, D O'Sullivan, and R Brennan. Evaluation of ontology mapping representations. In *AISB 2009 Convention*, 2009.

[49] Giovanni Tummarello, Eyal Oren, and Renaud Delbru. Sindice.com: Weaving the open linked data. In *6th International and 2nd Asian Semantic Web Conference (ISWC2007+ASWC2007)*, pages 547–560, November 2007. URL `http://data.semanticweb.org/conference/iswc-aswc/2007/tracks/research/papers/547`.

[50] Foaf WIKI. Using based near. URL `http://wiki.foaf-project.org/w/UsingBasedNear`. Last accessed: 1 April 2010.

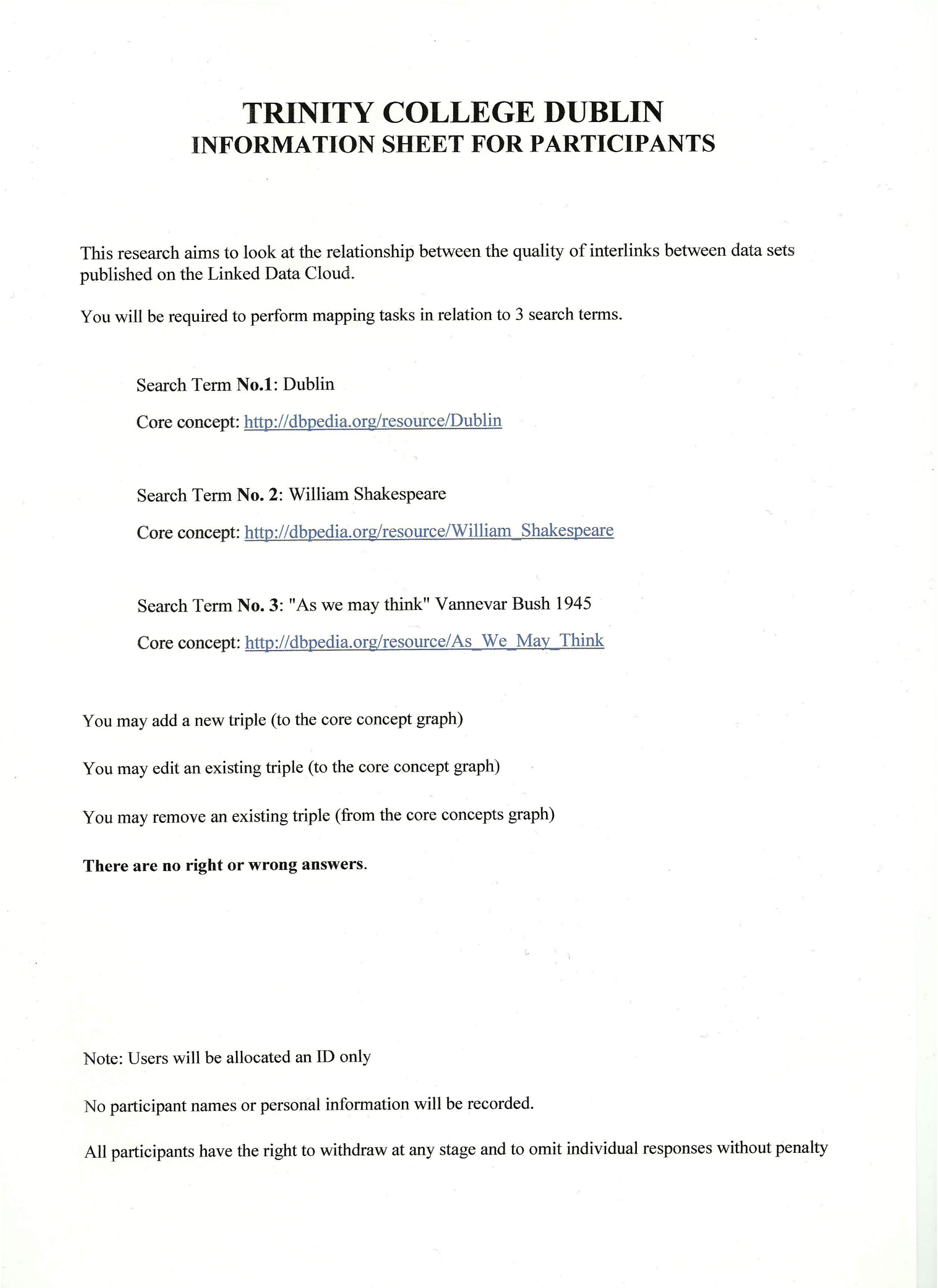
[51] Zhaohui Wu, Huajun Chen, Yuxing Mao, and Guozhou Zheng. *Web Technologies Research and Development - APWeb 2005*. Springer Berlin / Heidelberg, 2005. ISBN 978-3-540-25207-8. doi: 10.1007/b106936.

[52] Chunying Zhou, Chengli Xu, Huajun Chen, and Kingsley Idehen. Browser-based semantic mapping tool for linked data in semantic web. In *Linked Data on the Web (LDOW2008)*, 2008. URL `http://data.semanticweb.org/workshop/LDOW/2008/paper/15`.

Appendix

8.1 Participant Briefing Materials

8.1.1 Information Sheet



TRINITY COLLEGE DUBLIN
INFORMATION SHEET FOR PARTICIPANTS

This research aims to look at the relationship between the quality of interlinks between data sets published on the Linked Data Cloud.

You will be required to perform mapping tasks in relation to 3 search terms.

Search Term **No.1**: Dublin

Core concept: http://dbpedia.org/resource/Dublin

Search Term **No. 2**: William Shakespeare

Core concept: http://dbpedia.org/resource/William_Shakespeare

Search Term **No. 3**: "As we may think" Vannevar Bush 1945

Core concept: http://dbpedia.org/resource/As_We_May_Think

You may add a new triple (to the core concept graph)

You may edit an existing triple (to the core concept graph)

You may remove an existing triple (from the core concepts graph)

There are no right or wrong answers.

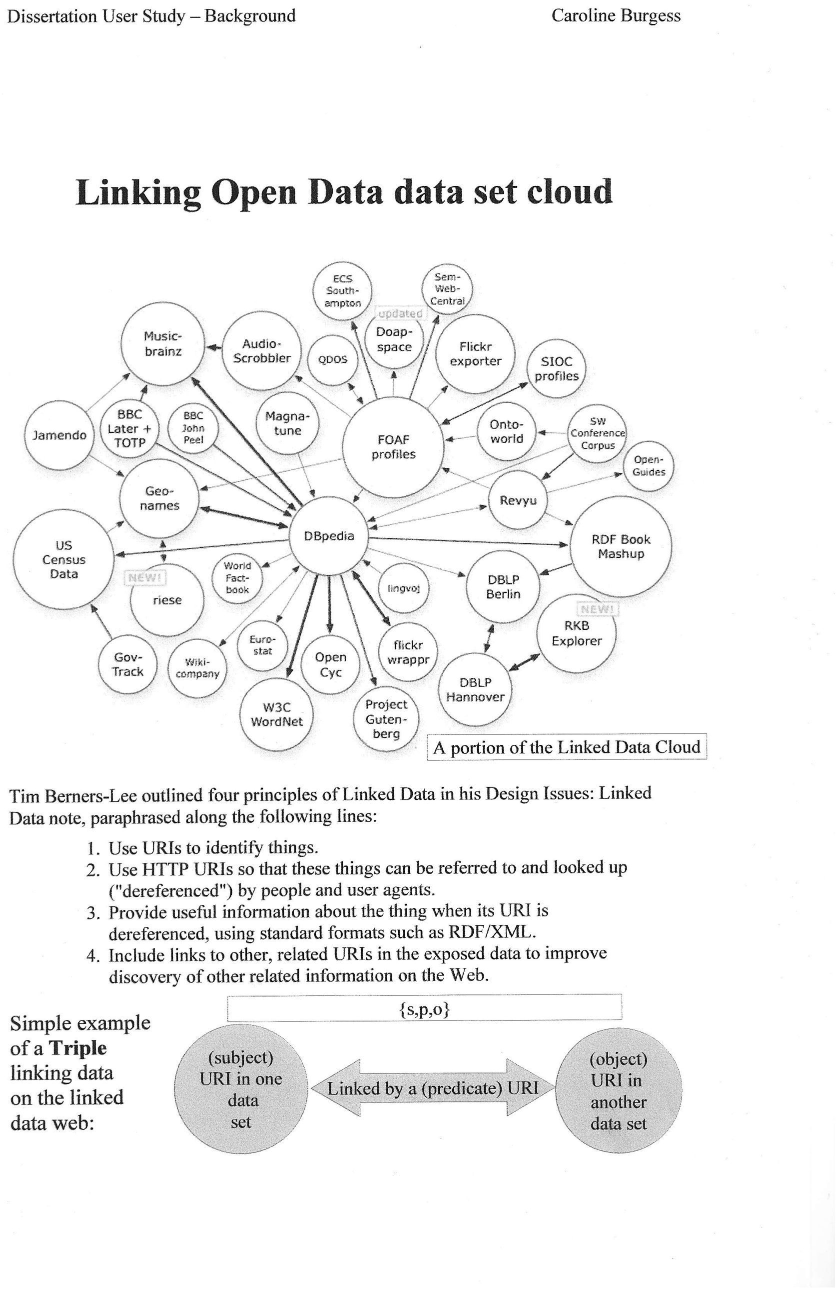
Note: Users will be allocated an ID only

No participant names or personal information will be recorded.

All participants have the right to withdraw at any stage and to omit individual responses without penalty

Figure 8.1: Participant Information Sheet

8.1.2 Linked Data Overview



Dissertation User Study – Background   Caroline Burgess

Linking Open Data data set cloud

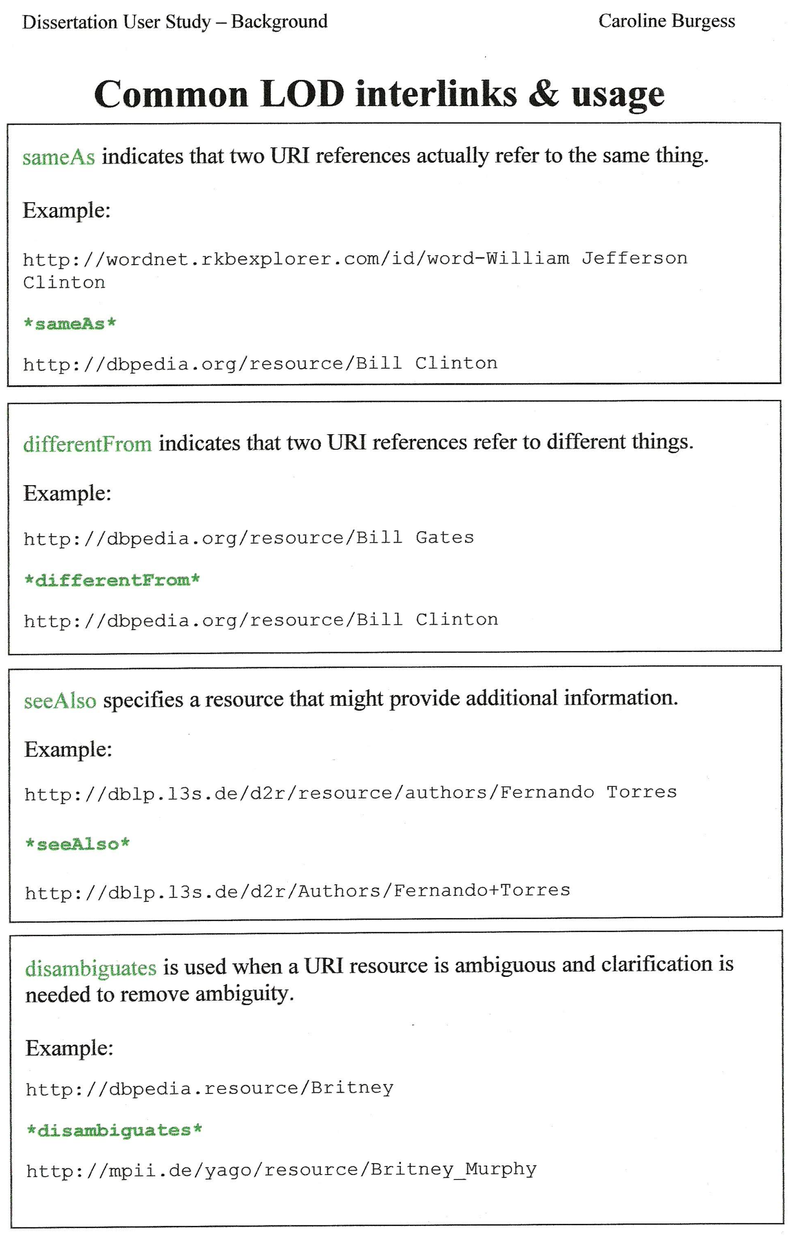
A portion of the Linked Data Cloud

Tim Berners-Lee outlined four principles of Linked Data in his Design Issues: Linked Data note, paraphrased along the following lines:

1. Use URIs to identify things.
2. Use HTTP URIs so that these things can be referred to and looked up ("dereferenced") by people and user agents.
3. Provide useful information about the thing when its URI is dereferenced, using standard formats such as RDF/XML.
4. Include links to other, related URIs in the exposed data to improve discovery of other related information on the Web.

Simple example of a **Triple** linking data on the linked data web:

Figure 8.2: Linked Data Overview *a*



Dissertation User Study – Background Caroline Burgess

Common LOD interlinks & usage

sameAs indicates that two URI references actually refer to the same thing.

Example:

`http://wordnet.rkbexplorer.com/id/word-William Jefferson Clinton`

`*sameAs*`

`http://dbpedia.org/resource/Bill Clinton`

differentFrom indicates that two URI references refer to different things.

Example:

`http://dbpedia.org/resource/Bill Gates`

`*differentFrom*`

`http://dbpedia.org/resource/Bill Clinton`

seeAlso specifies a resource that might provide additional information.

Example:

`http://dblp.l3s.de/d2r/resource/authors/Fernando Torres`

`*seeAlso*`

`http://dblp.l3s.de/d2r/Authors/Fernando+Torres`

disambiguates is used when a URI resource is ambiguous and clarification is needed to remove ambiguity.

Example:

`http://dbpedia.resource/Britney`

`*disambiguates*`

`http://mpii.de/yago/resource/Britney_Murphy`

Figure 8.3: Linked Data Overview *b*

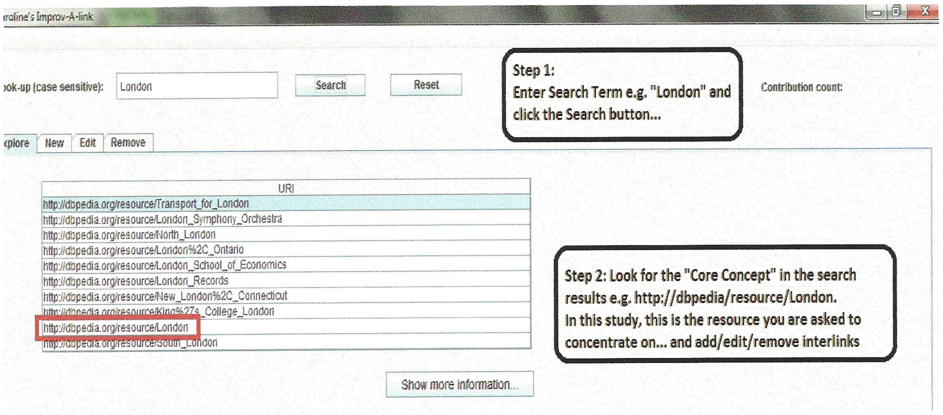
8.1.3 App Tutorial

App Tutorial

Search Term: London

Core Concept: http://dbpedia/resource/London

Fig. 1

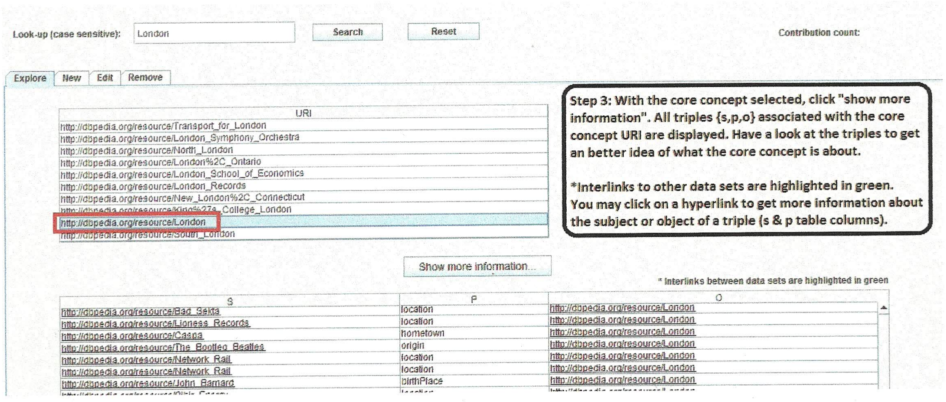


URI
http://dbpedia.org/resource/Transport_for_London
http://dbpedia.org/resource/London_Symphony_Orchestra
http://dbpedia.org/resource/North_London
http://dbpedia.org/resource/London%2C_Ontario
http://dbpedia.org/resource/London_School_of_Economics
http://dbpedia.org/resource/London_Records
http://dbpedia.org/resource/New_London%2C_Connecticut
http://dbpedia.org/resource/King%27s_College_London
http://dbpedia.org/resource/London
http://dbpedia.org/resource/South_London

Step 1:
Enter Search Term e.g. "London" and click the Search button...

Step 2: Look for the "Core Concept" in the search results e.g. http://dbpedia/resource/London. In this study, this is the resource you are asked to concentrate on... and add/edit/remove interlinks

Fig. 2



URI
http://dbpedia.org/resource/Transport_for_London
http://dbpedia.org/resource/London_Symphony_Orchestra
http://dbpedia.org/resource/North_London
http://dbpedia.org/resource/London%2C_Ontario
http://dbpedia.org/resource/London_School_of_Economics
http://dbpedia.org/resource/London_Records
http://dbpedia.org/resource/New_London%2C_Connecticut
http://dbpedia.org/resource/King%27s_College_London
http://dbpedia.org/resource/London
http://dbpedia.org/resource/South_London

Step 3: With the core concept selected, click "show more information". All triples {s,p,o} associated with the core concept URI are displayed. Have a look at the triples to get an better idea of what the core concept is about.

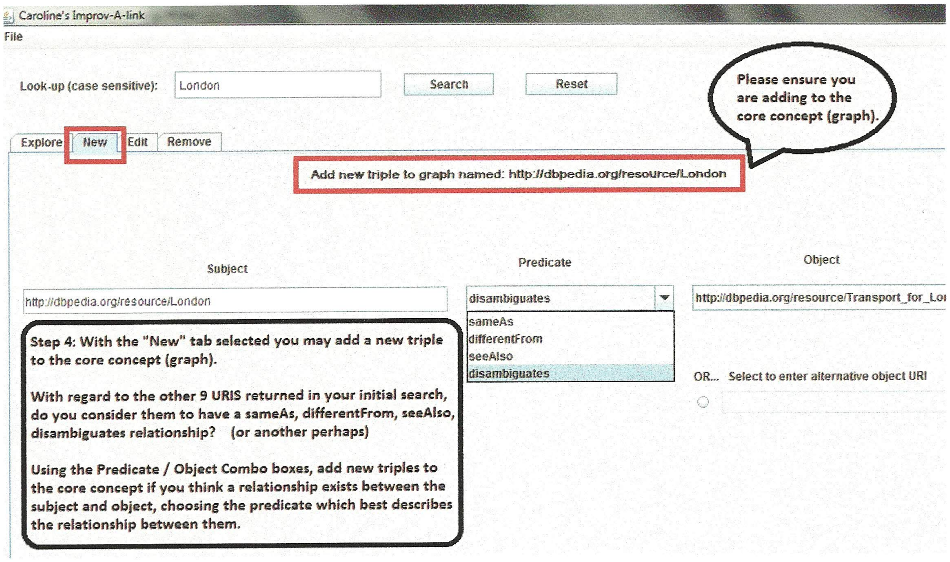
*Interlinks to other data sets are highlighted in green. You may click on a hyperlink to get more information about the subject or object of a triple (s & p table columns).

* Interlinks between data sets are highlighted in green

S	P	O
http://dbpedia.org/resource/Bad_Sekta	location	http://dbpedia.org/resource/London
http://dbpedia.org/resource/Lioness_Records	location	http://dbpedia.org/resource/London
http://dbpedia.org/resource/Caspa	hometown	http://dbpedia.org/resource/London
http://dbpedia.org/resource/The_Bootleg_Beatles	origin	http://dbpedia.org/resource/London
http://dbpedia.org/resource/Network_Rail	location	http://dbpedia.org/resource/London
http://dbpedia.org/resource/Network_Rail	location	http://dbpedia.org/resource/London
http://dbpedia.org/resource/John_Barnard	birthPlace	http://dbpedia.org/resource/London

Figure 8.4: App Tutorial *a*

Fig. 3



Look-up (case sensitive): London

Please ensure you are adding to the core concept (graph).

Add new triple to graph named: http://dbpedia.org/resource/London

Subject: http://dbpedia.org/resource/London

Predicate: disambiguates

Object: http://dbpedia.org/resource/Transport_for_Lo

Step 4: With the "New" tab selected you may add a new triple to the core concept (graph).

With regard to the other 9 URIS returned in your initial search, do you consider them to have a sameAs, differentFrom, seeAlso, disambiguates relationship? (or another perhaps)

Using the Predicate / Object Combo boxes, add new triples to the core concept if you think a relationship exists between the subject and object, choosing the predicate which best describes the relationship between them.

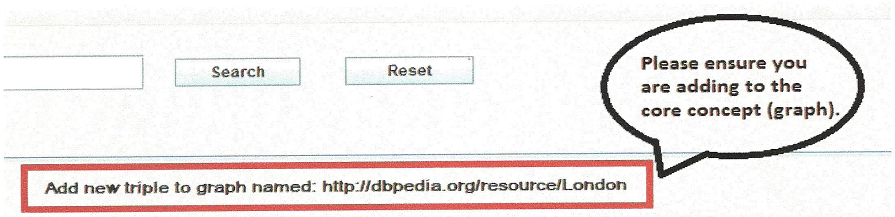
<u>Note</u>: To assist with decision making you will need to revert back to "explore" to determine what the other 9 URI's refer to. Simply select the URI -> "Show more information" as in Fig. 2

Determine what the URI is about...then...

Important: If you have been exploring other URIs, you must reselect the core concept -> "show more information" in order to add the new triple to the core concept.

You can see what graph is selected at the top of the "new", "edit" and "remove" panes as shown below:

Fig. 4

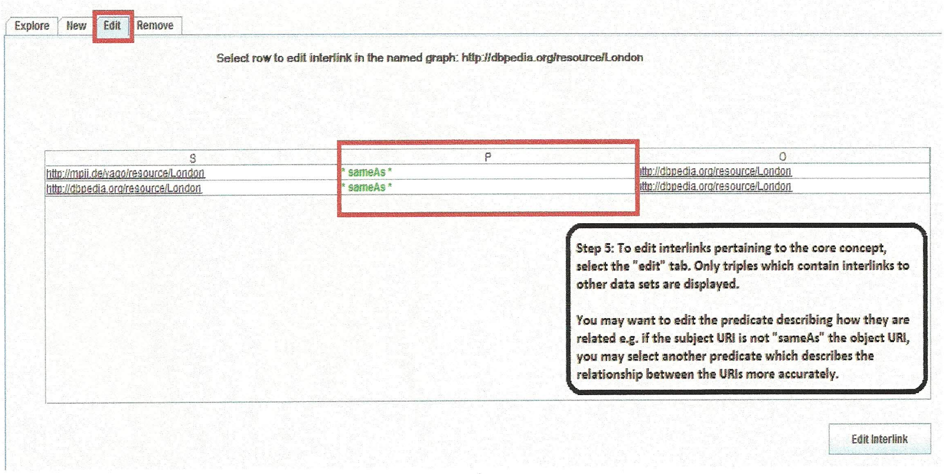


Please ensure you are adding to the core concept (graph).

Add new triple to graph named: http://dbpedia.org/resource/London

Figure 8.5: App Tutorial *b*

Fig. 5



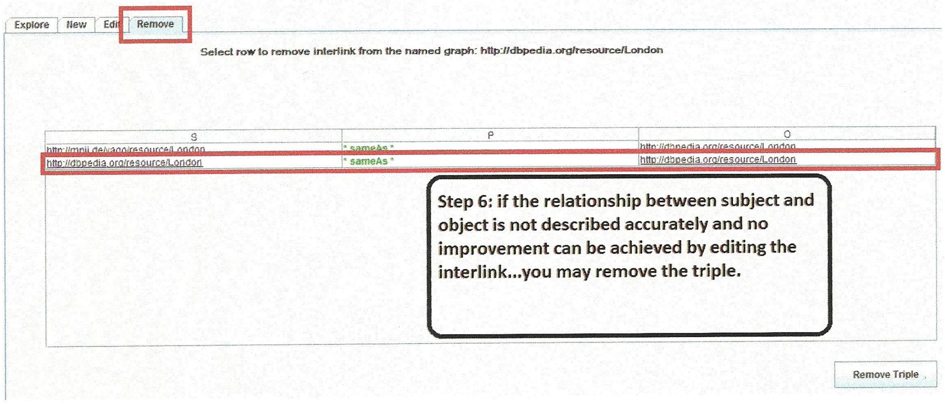
Select row to edit interlink in the named graph: http://dbpedia.org/resource/London

S	P	O
http://mpii.de/yago/resource/London	* sameAs *	http://dbpedia.org/resource/London
http://dbpedia.org/resource/London	* sameAs *	http://dbpedia.org/resource/London

Step 5: To edit interlinks pertaining to the core concept, select the "edit" tab. Only triples which contain interlinks to other data sets are displayed.

You may want to edit the predicate describing how they are related e.g. if the subject URI is not "sameAs" the object URI, you may select another predicate which describes the relationship between the URIs more accurately.

Fig. 6



Select row to remove interlink from the named graph: http://dbpedia.org/resource/London

S	P	O
http://mpii.de/yago/resource/London	* sameAs *	http://dbpedia.org/resource/London
http://dbpedia.org/resource/London	* sameAs *	http://dbpedia.org/resource/London

Step 6: if the relationship between subject and object is not described accurately and no improvement can be achieved by editing the interlink...you may remove the triple.

Figure 8.6: App Tutorial *c*

8.2 Participant Trial Questionnaires

Note: **9** participants in total took part in the trial.

Data pertaining to the first participant, User ID 1, was not included in this research as severe network difficulties on the day lead to the trial being terminated early for this user. Comparison, therefore would not be on a like-for-like basis and it was decided to discard the data, as it may lead to skewed results.

For simplicity, in the main document text,

Appendix User ID 2 — became Participant 1

Appendix User ID 3 — became Participant 2

Appendix User ID 4 — became Participant 3

Appendix User ID 5 — became Participant 4

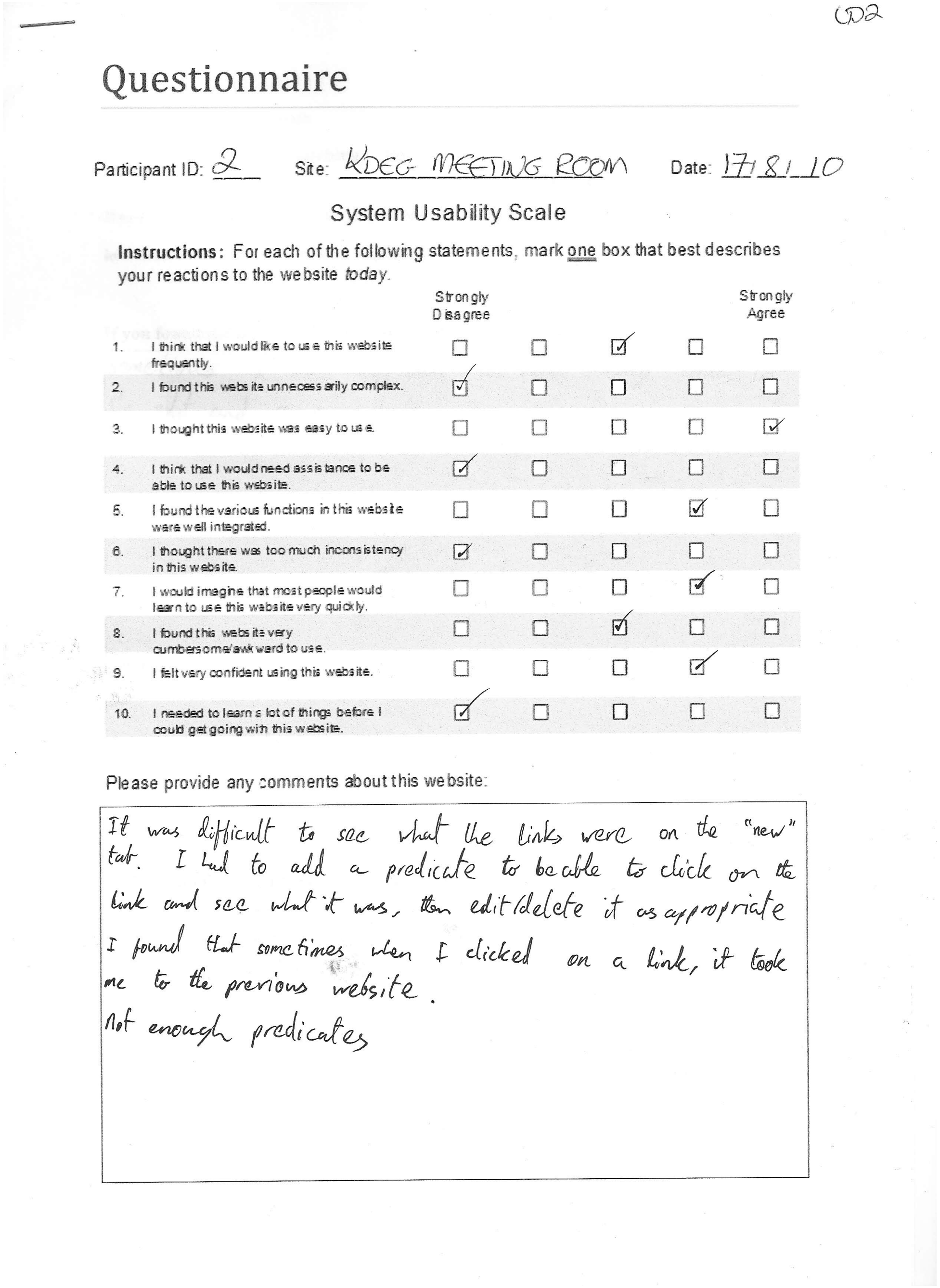
Appendix User ID 6 — became Participant 5

Appendix User ID 7 — became Participant 6

Appendix User ID 8 — became Participant 7

Appendix User ID 9 — became Participant 8

8.2.1 Participant ID — 2



CD2

Questionnaire

Participant ID: 2 Site: KDEG MEETING ROOM Date: 17/8/10

System Usability Scale

Instructions: For each of the following statements, mark one box that best describes your reactions to the website *today*.

		Strongly Disagree				Strongly Agree
1.	I think that I would like to use this website frequently.	☐	☐	☑	☐	☐
2.	I found this website unnecessarily complex.	☑	☐	☐	☐	☐
3.	I thought this website was easy to use.	☐	☐	☐	☐	☑
4.	I think that I would need assistance to be able to use this website.	☑	☐	☐	☐	☐
5.	I found the various functions in this website were well integrated.	☐	☐	☐	☑	☐
6.	I thought there was too much inconsistency in this website.	☑	☐	☐	☐	☐
7.	I would imagine that most people would learn to use this website very quickly.	☐	☐	☐	☑	☐
8.	I found this website very cumbersome/awkward to use.	☐	☐	☑	☐	☐
9.	I felt very confident using this website.	☐	☐	☐	☑	☐
10.	I needed to learn a lot of things before I could get going with this website.	☑	☐	☐	☐	☐

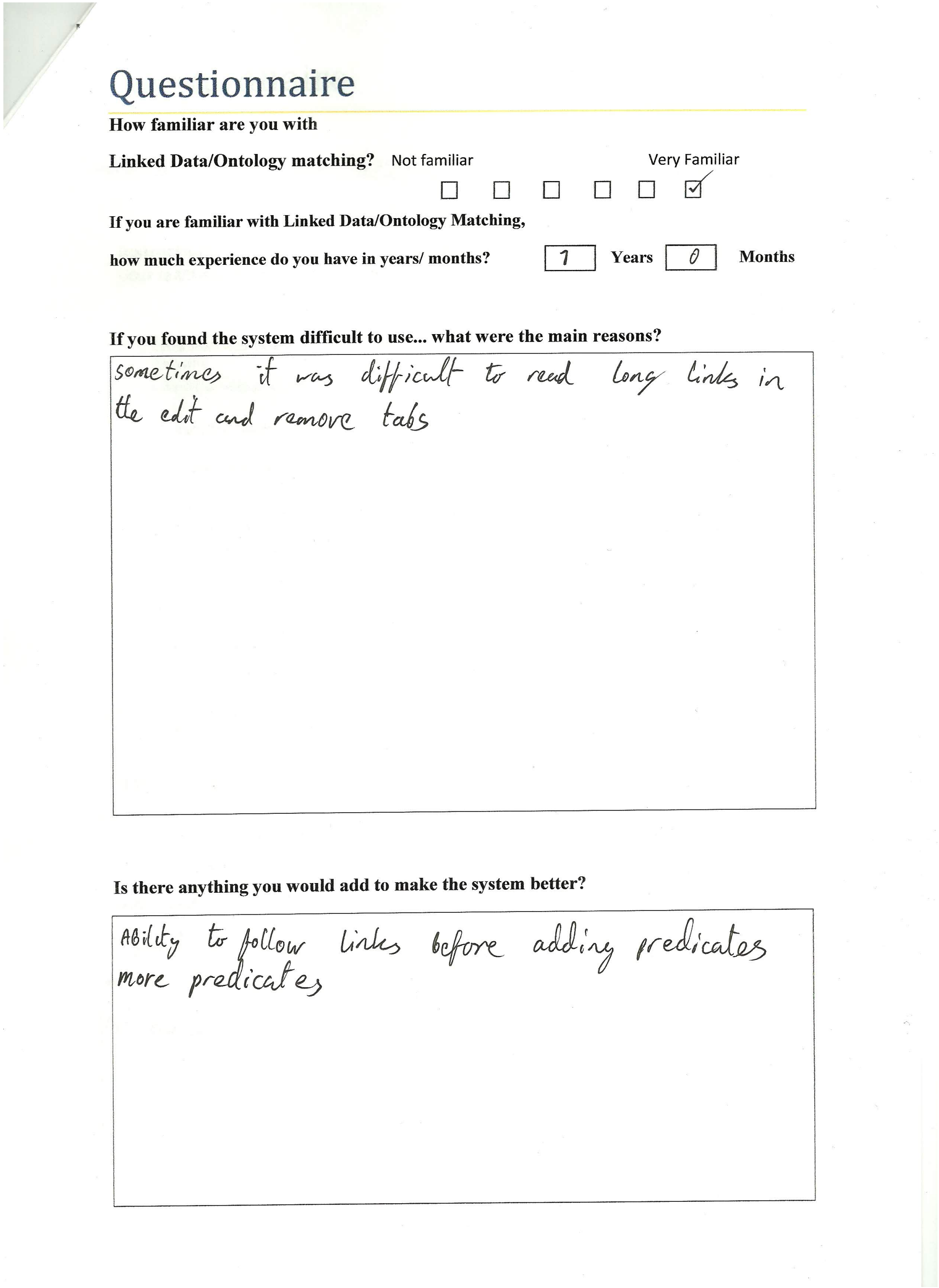
Please provide any comments about this website:

It was difficult to see what the links were on the "new" tab. I had to add a predicate to be able to click on the link and see what it was, then edit/delete it as appropriate

I found that sometimes when I clicked on a link, it took me to the previous website.

Not enough predicates

Figure 8.7: Participant ID — 2 *a*



Questionnaire

How familiar are you with

Linked Data/Ontology matching? Not familiar ☐ ☐ ☐ ☐ ☐ ☑ Very Familiar

If you are familiar with Linked Data/Ontology Matching,

how much experience do you have in years/ months? 1 **Years** 0 **Months**

If you found the system difficult to use... what were the main reasons?

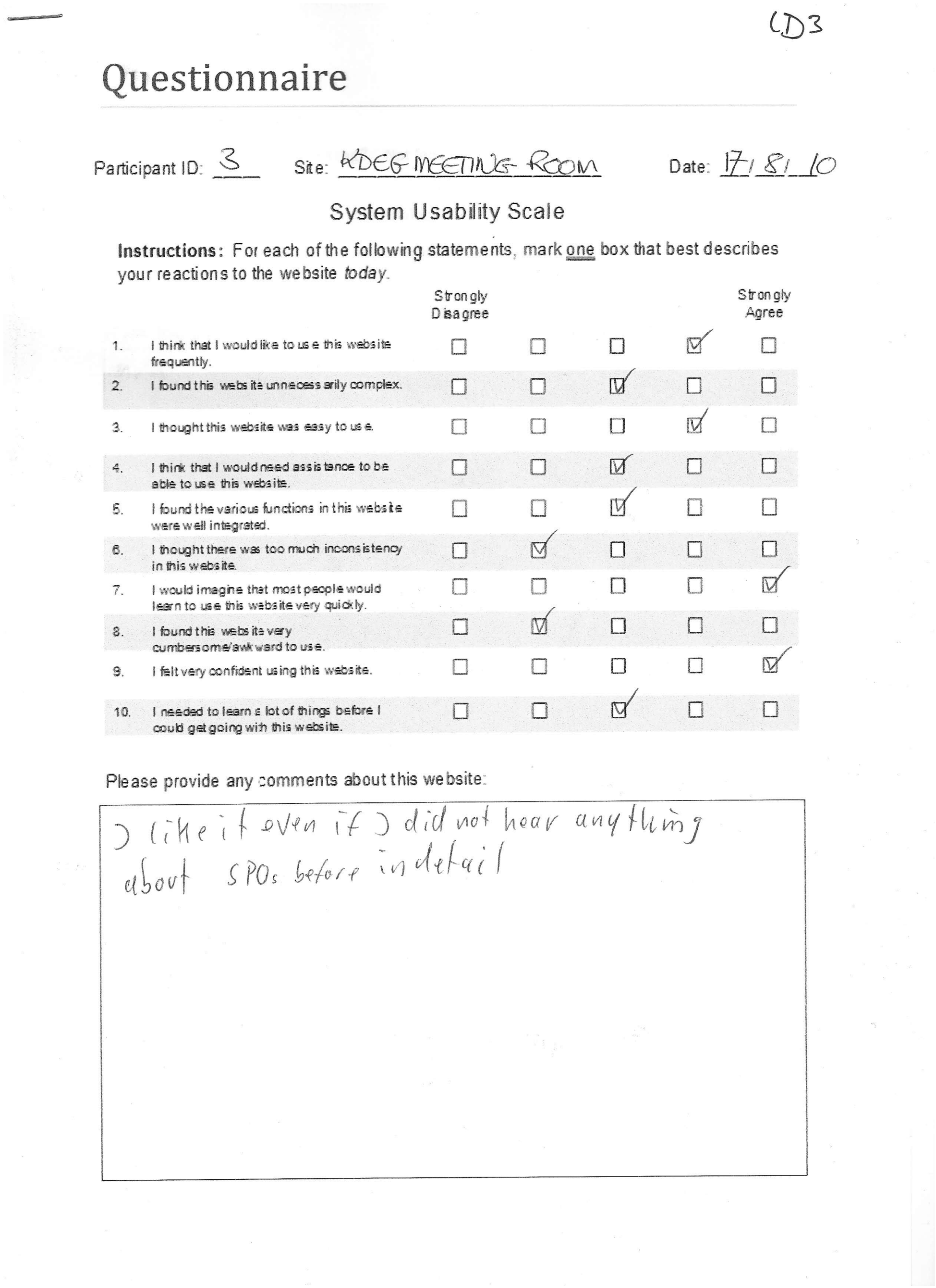
sometimes it was difficult to read long links in the edit and remove tabs

Is there anything you would add to make the system better?

Ability to follow links before adding predicates
more predicates

Figure 8.8: Participant ID — 2 *b*

8.2.2 Participant ID — 3



CD3

Questionnaire

Participant ID: 3 Site: KDEG MEETING ROOM Date: 17/8/10

System Usability Scale

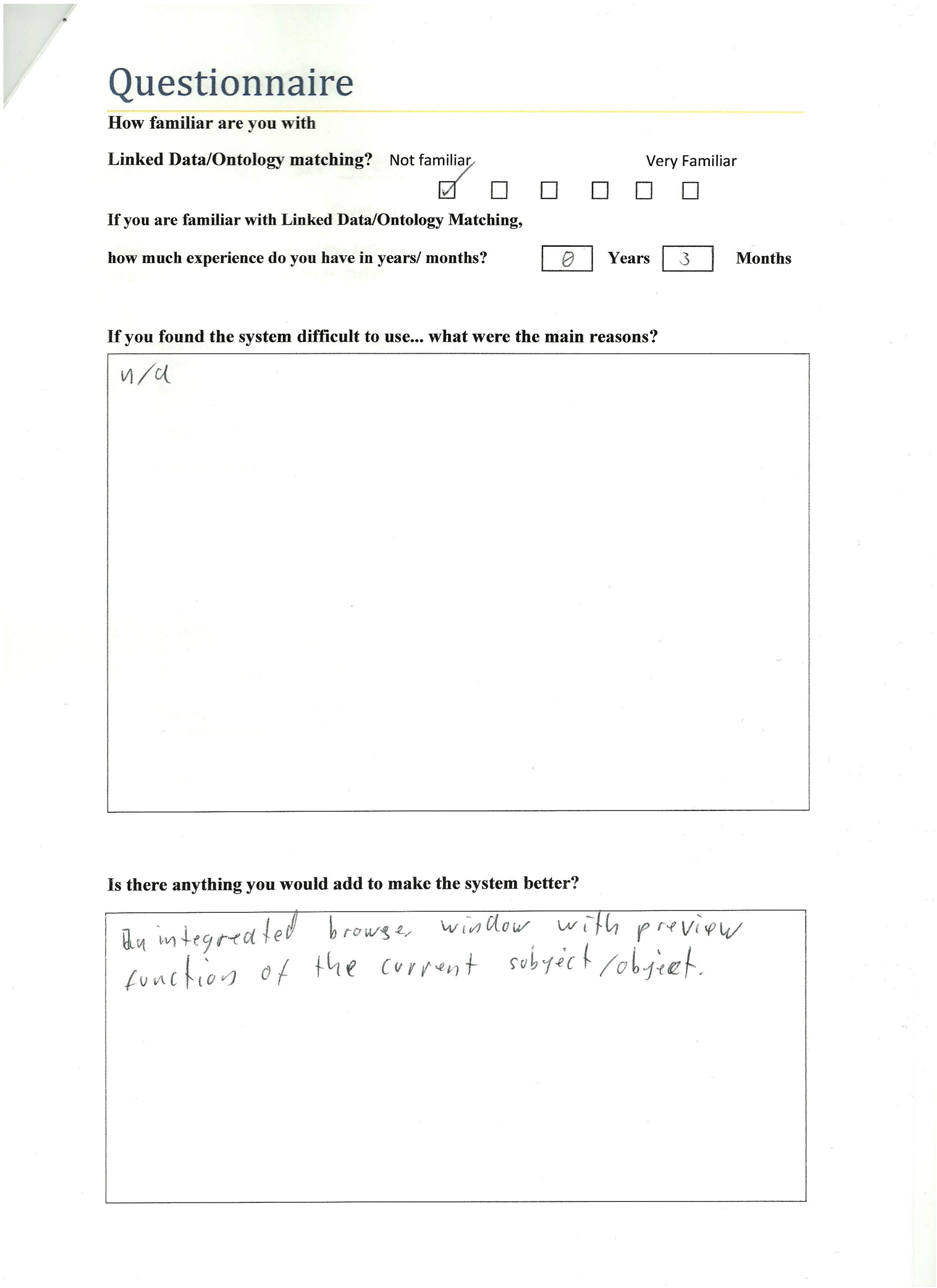
Instructions: For each of the following statements, mark one box that best describes your reactions to the website *today*.

		Strongly Disagree				Strongly Agree
1.	I think that I would like to use this website frequently.	☐	☐	☐	☑	☐
2.	I found this website unnecessarily complex.	☐	☐	☑	☐	☐
3.	I thought this website was easy to use.	☐	☐	☐	☑	☐
4.	I think that I would need assistance to be able to use this website.	☐	☐	☑	☐	☐
5.	I found the various functions in this website were well integrated.	☐	☐	☑	☐	☐
6.	I thought there was too much inconsistency in this website.	☐	☑	☐	☐	☐
7.	I would imagine that most people would learn to use this website very quickly.	☐	☐	☐	☐	☑
8.	I found this website very cumbersome/awkward to use.	☐	☑	☐	☐	☐
9.	I felt very confident using this website.	☐	☐	☐	☐	☑
10.	I needed to learn a lot of things before I could get going with this website.	☐	☐	☑	☐	☐

Please provide any comments about this website:

I like it even if I did not hear anything about SPOs before in detail

Figure 8.9: Participant ID — 3 *a*



Questionnaire

How familiar are you with

Linked Data/Ontology matching? Not familiar ☑ ☐ ☐ ☐ ☐ ☐ Very Familiar

If you are familiar with Linked Data/Ontology Matching,

how much experience do you have in years/ months? 0 **Years** 3 **Months**

If you found the system difficult to use... what were the main reasons?

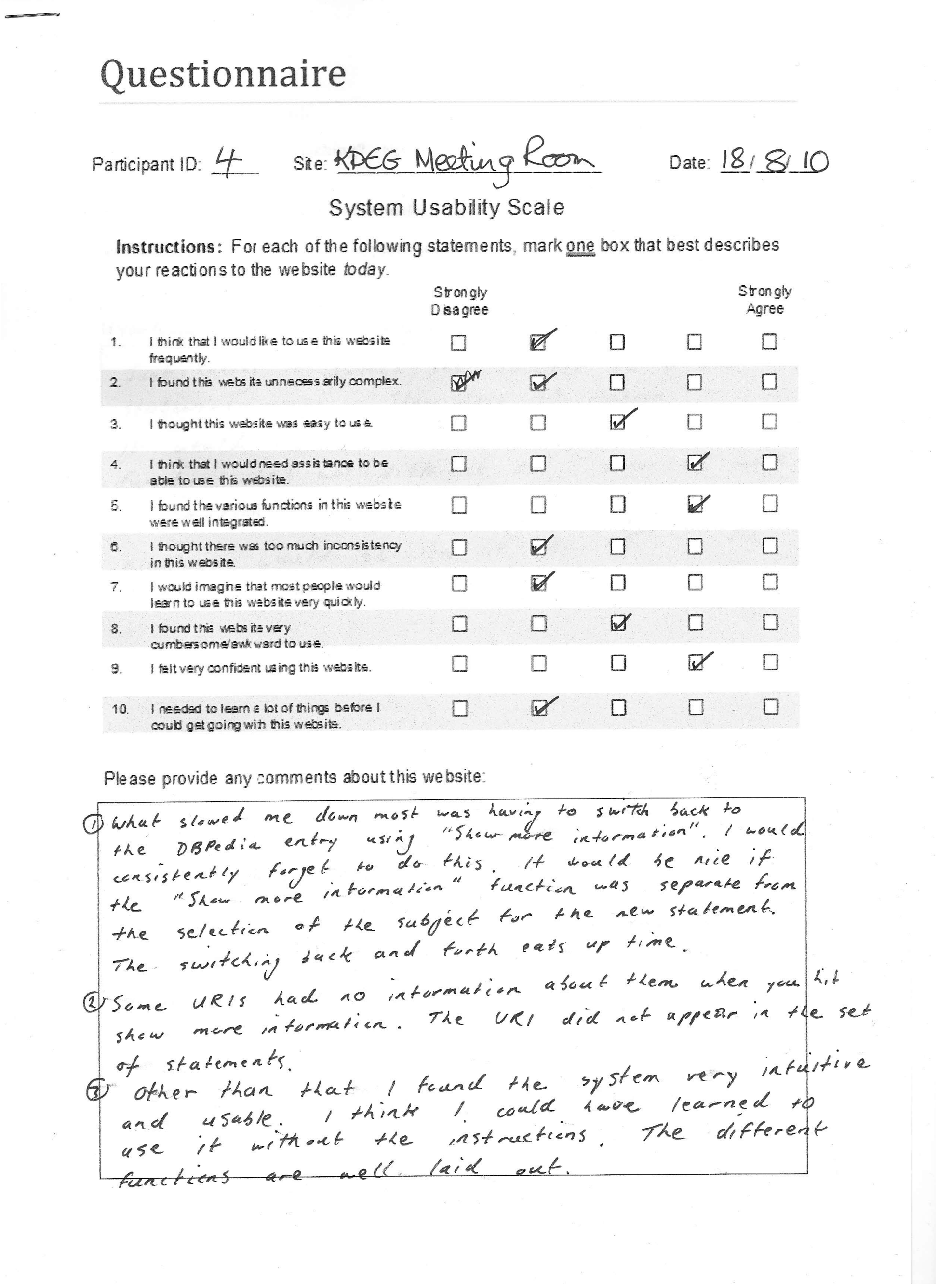
n/a

Is there anything you would add to make the system better?

An integrated browse window with preview function of the current subject/object.

Figure 8.10: Participant ID — 3 *b*

8.2.3 Participant ID — 4



Questionnaire

Participant ID: 4 Site: KDEG Meeting Room Date: 18/8/10

System Usability Scale

Instructions: For each of the following statements, mark <u>one</u> box that best describes your reactions to the website *today*.

		Strongly Disagree				Strongly Agree
1.	I think that I would like to use this website frequently.	☐	☑	☐	☐	☐
2.	I found this website unnecessarily complex.	☒	☑	☐	☐	☐
3.	I thought this website was easy to use.	☐	☐	☑	☐	☐
4.	I think that I would need assistance to be able to use this website.	☐	☐	☐	☑	☐
5.	I found the various functions in this website were well integrated.	☐	☐	☐	☑	☐
6.	I thought there was too much inconsistency in this website.	☐	☑	☐	☐	☐
7.	I would imagine that most people would learn to use this website very quickly.	☐	☑	☐	☐	☐
8.	I found this website very cumbersome/awkward to use.	☐	☐	☑	☐	☐
9.	I felt very confident using this website.	☐	☐	☐	☑	☐
10.	I needed to learn a lot of things before I could get going with this website.	☐	☑	☐	☐	☐

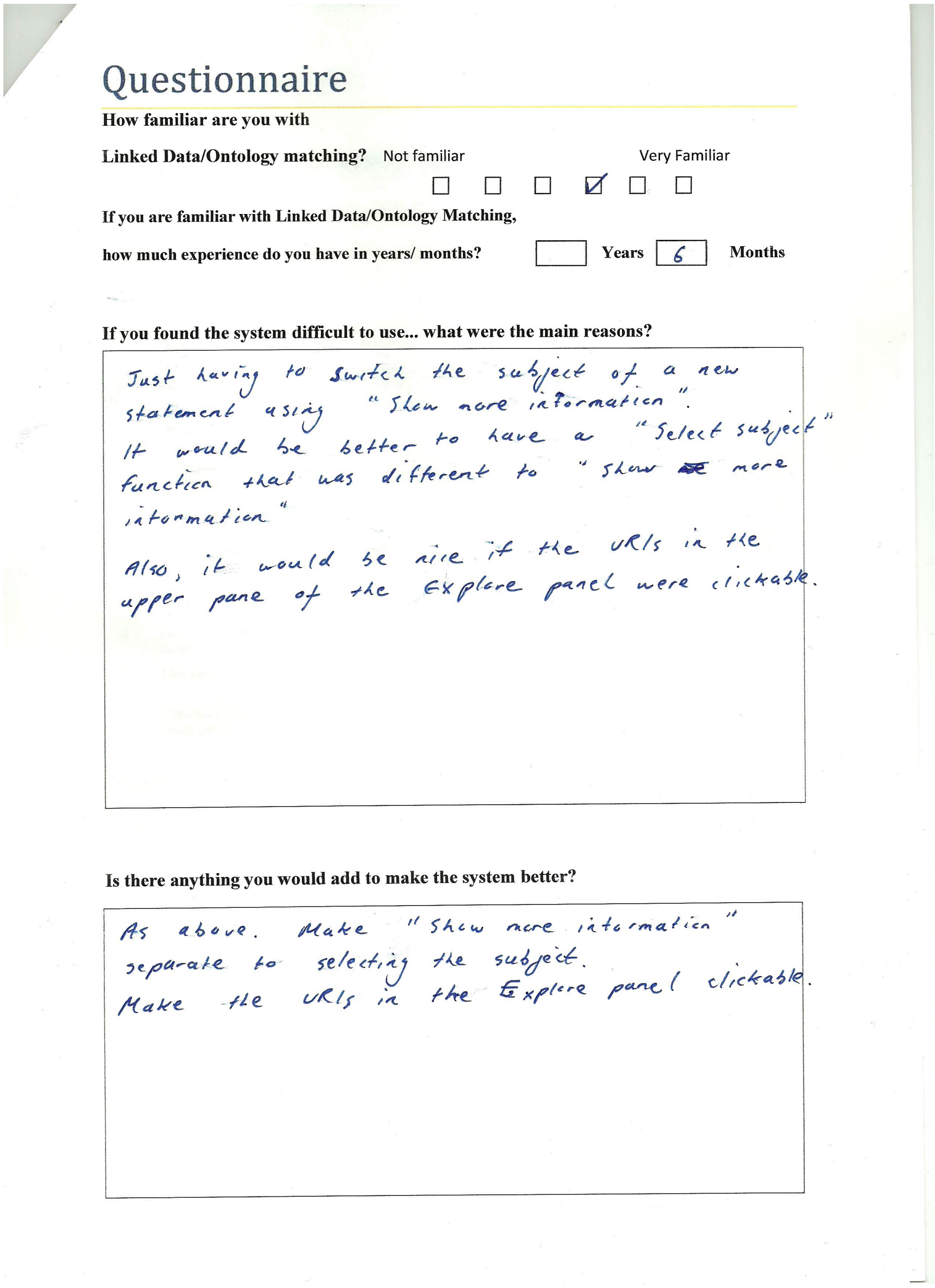
Please provide any comments about this website:

① What slowed me down most was having to switch back to the DBPedia entry using "Show more information". I would consistently forget to do this. It would be nice if the "Show more information" function was separate from the selection of the subject for the new statement. The switching back and forth eats up time.

② Some URIs had no information about them when you hit show more information. The URI did not appear in the set of statements.

③ Other than that I found the system very intuitive and usable. I think I could have learned to use it without the instructions. The different functions are well laid out.

Figure 8.11: Participant ID — 4 *a*



Questionnaire

How familiar are you with

Linked Data/Ontology matching? Not familiar — Very Familiar

☐ ☐ ☐ ☑ ☐ ☐

If you are familiar with Linked Data/Ontology Matching,

how much experience do you have in years/ months? ☐ **Years** 6 **Months**

If you found the system difficult to use... what were the main reasons?

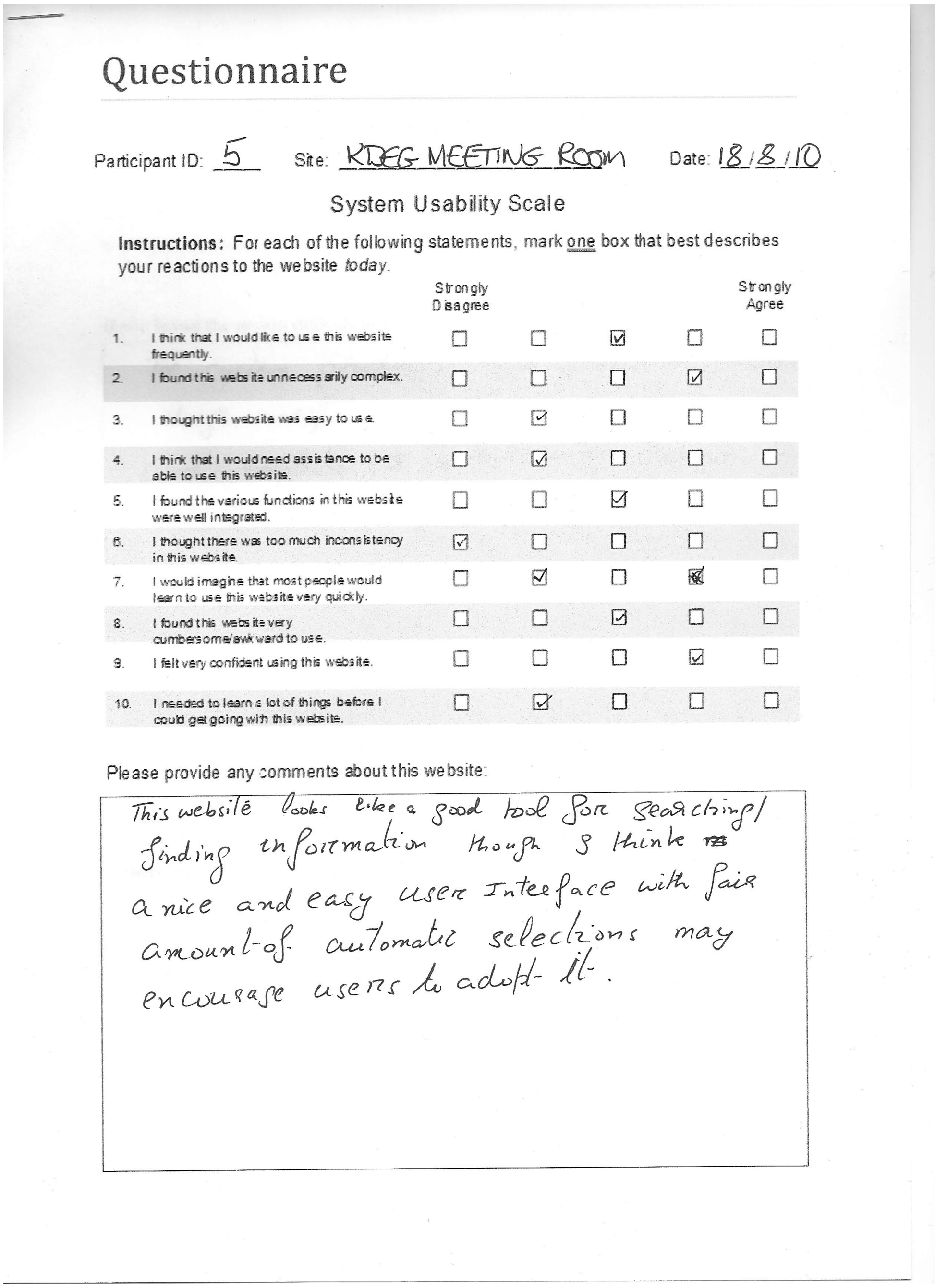
Just having to switch the subject of a new statement using "Show more information".
It would be better to have a "Select subject" function that was different to "Show more information"
Also, it would be nice if the URIs in the upper pane of the Explore panel were clickable.

Is there anything you would add to make the system better?

As above. Make "Show more information" separate to selecting the subject.
Make the URIs in the Explore panel clickable.

Figure 8.12: Participant ID — 4 *b*

8.2.4 Participant ID — 5



Questionnaire

Participant ID: 5 Site: KDEG MEETING ROOM Date: 18/8/10

System Usability Scale

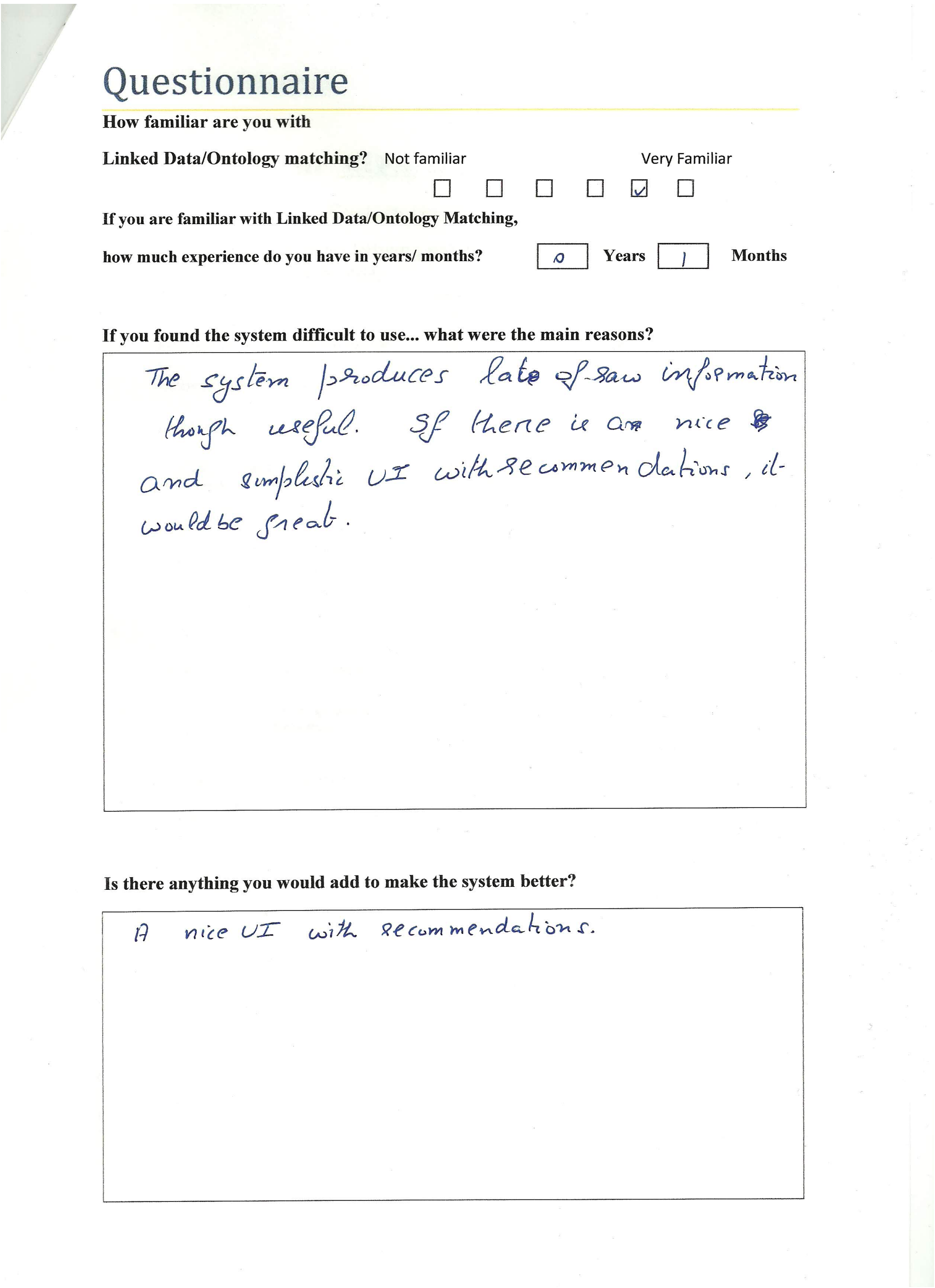
Instructions: For each of the following statements, mark one box that best describes your reactions to the website *today*.

		Strongly Disagree				Strongly Agree
1.	I think that I would like to use this website frequently.	☐	☐	☑	☐	☐
2.	I found this website unnecessarily complex.	☐	☐	☐	☑	☐
3.	I thought this website was easy to use.	☐	☑	☐	☐	☐
4.	I think that I would need assistance to be able to use this website.	☐	☑	☐	☐	☐
5.	I found the various functions in this website were well integrated.	☐	☐	☑	☐	☐
6.	I thought there was too much inconsistency in this website.	☑	☐	☐	☐	☐
7.	I would imagine that most people would learn to use this website very quickly.	☐	☑	☐	☒	☐
8.	I found this website very cumbersome/awkward to use.	☐	☐	☑	☐	☐
9.	I felt very confident using this website.	☐	☐	☐	☑	☐
10.	I needed to learn a lot of things before I could get going with this website.	☐	☑	☐	☐	☐

Please provide any comments about this website:

This website looks like a good tool for searching/
finding information though I think
a nice and easy user Inteeface with fair
amount of automatic selections may
encourage users to adopt it.

Figure 8.13: Participant ID — 5 *a*



Questionnaire

How familiar are you with

Linked Data/Ontology matching? Not familiar ☐ ☐ ☐ ☐ ☑ ☐ Very Familiar

If you are familiar with Linked Data/Ontology Matching,

how much experience do you have in years/ months? 0 **Years** 1 **Months**

If you found the system difficult to use... what were the main reasons?

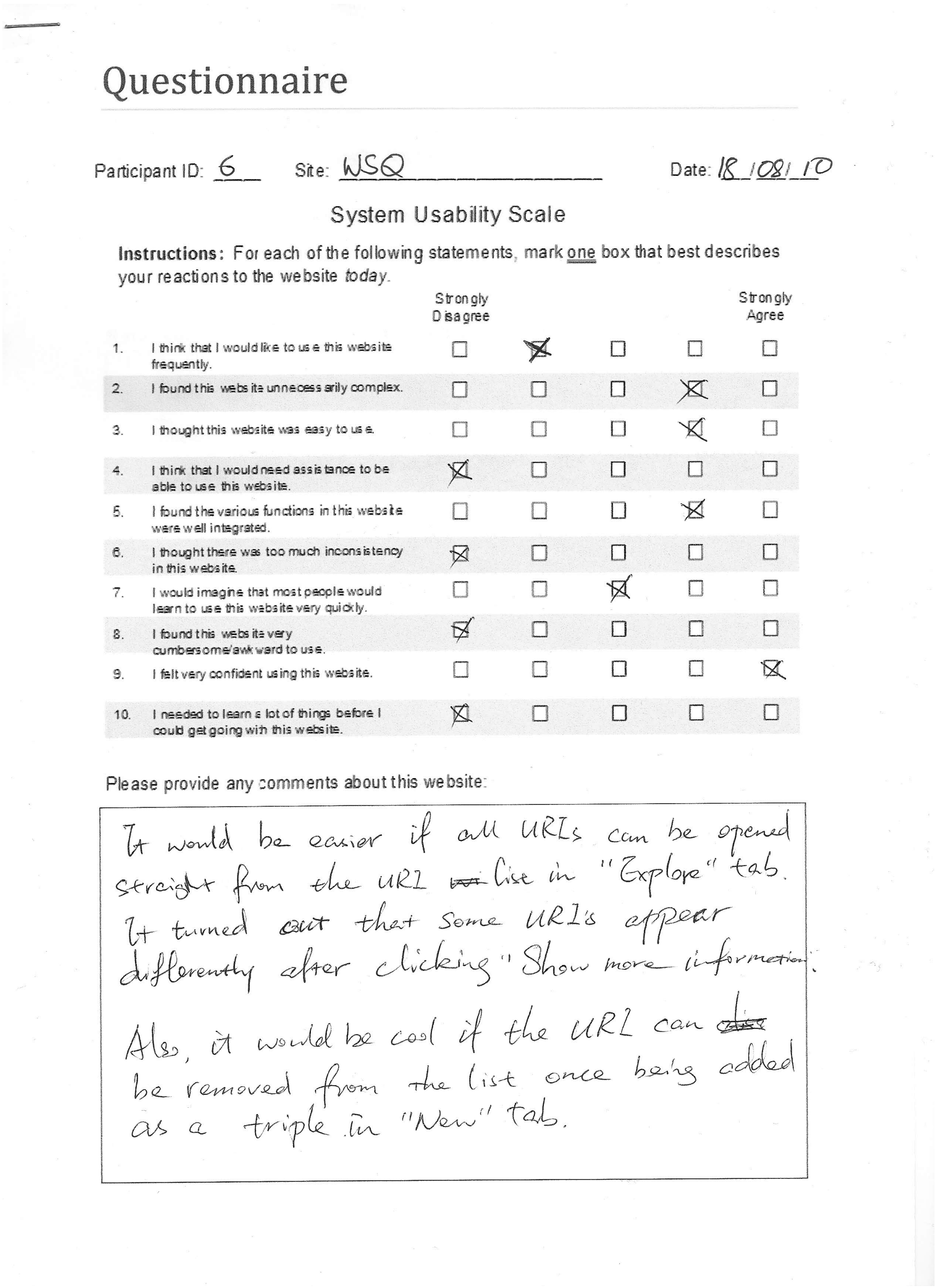
The system produces late of raw information though useful. If there is an nice and simplistic UI with recommendations, it would be great.

Is there anything you would add to make the system better?

A nice UI with recommendations.

Figure 8.14: Participant ID — 5 *b*

8.2.5 Participant ID — 6



Questionnaire

Participant ID: 6 Site: WSQ Date: 18/08/10

System Usability Scale

Instructions: For each of the following statements, mark one box that best describes your reactions to the website *today*.

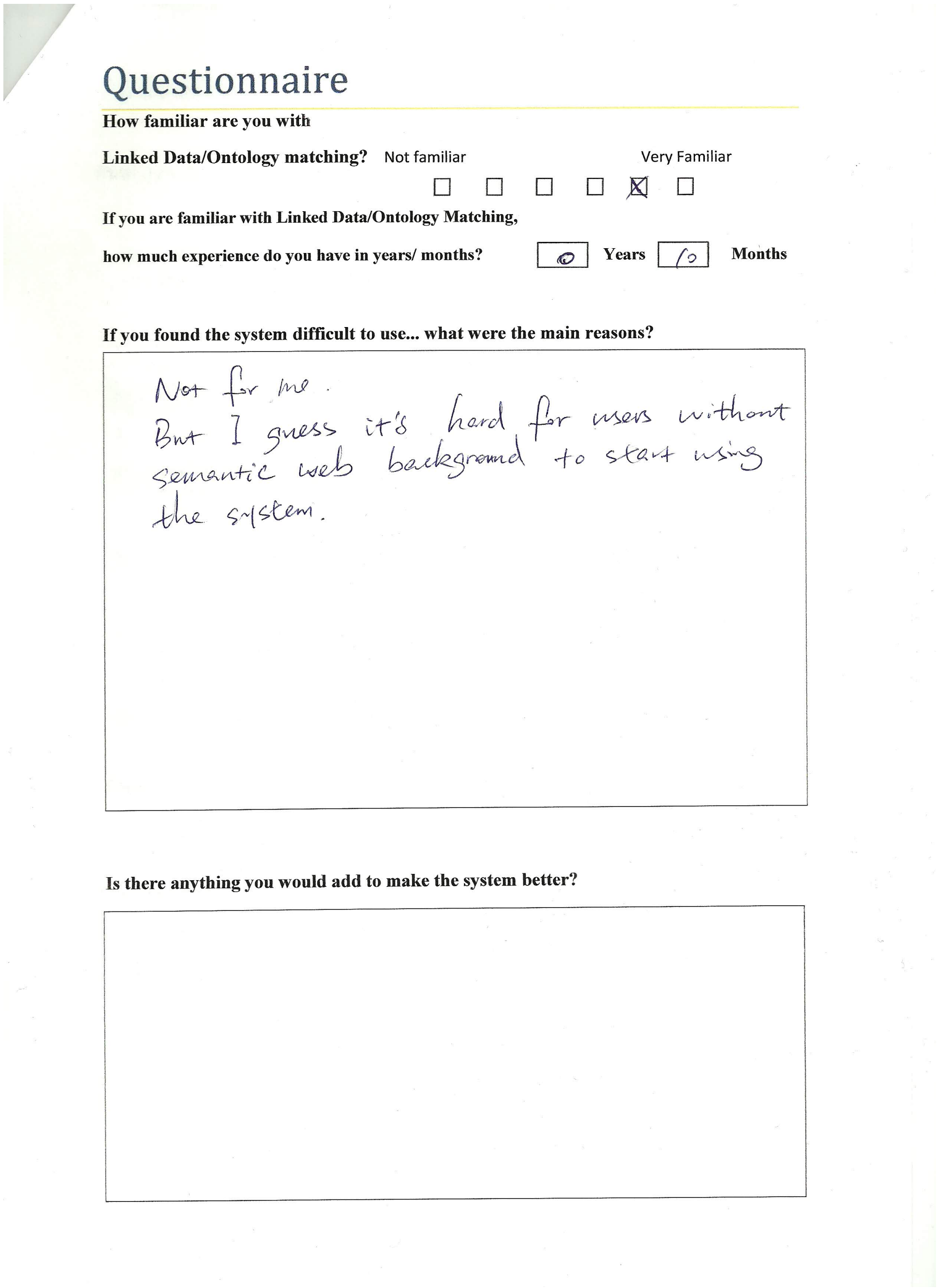
		Strongly Disagree				Strongly Agree
1.	I think that I would like to use this website frequently.	☐	☒	☐	☐	☐
2.	I found this website unnecessarily complex.	☐	☐	☐	☒	☐
3.	I thought this website was easy to use.	☐	☐	☐	☒	☐
4.	I think that I would need assistance to be able to use this website.	☒	☐	☐	☐	☐
5.	I found the various functions in this website were well integrated.	☐	☐	☐	☒	☐
6.	I thought there was too much inconsistency in this website.	☒	☐	☐	☐	☐
7.	I would imagine that most people would learn to use this website very quickly.	☐	☐	☒	☐	☐
8.	I found this website very cumbersome/awkward to use.	☒	☐	☐	☐	☐
9.	I felt very confident using this website.	☐	☐	☐	☐	☒
10.	I needed to learn a lot of things before I could get going with this website.	☒	☐	☐	☐	☐

Please provide any comments about this website:

It would be easier if all URIs can be opened streight from the URI list in "Explore" tab. It turned out that some URI's appear differently after clicking "Show more information".

Also, it would be cool if the URI can be removed from the list once being added as a triple in "New" tab.

Figure 8.15: Participant ID — 6 *a*



Questionnaire

How familiar are you with

Linked Data/Ontology matching? Not familiar Very Familiar

☐ ☐ ☐ ☐ ☒ ☐

If you are familiar with Linked Data/Ontology Matching,

how much experience do you have in years/ months? 0 **Years** 10 **Months**

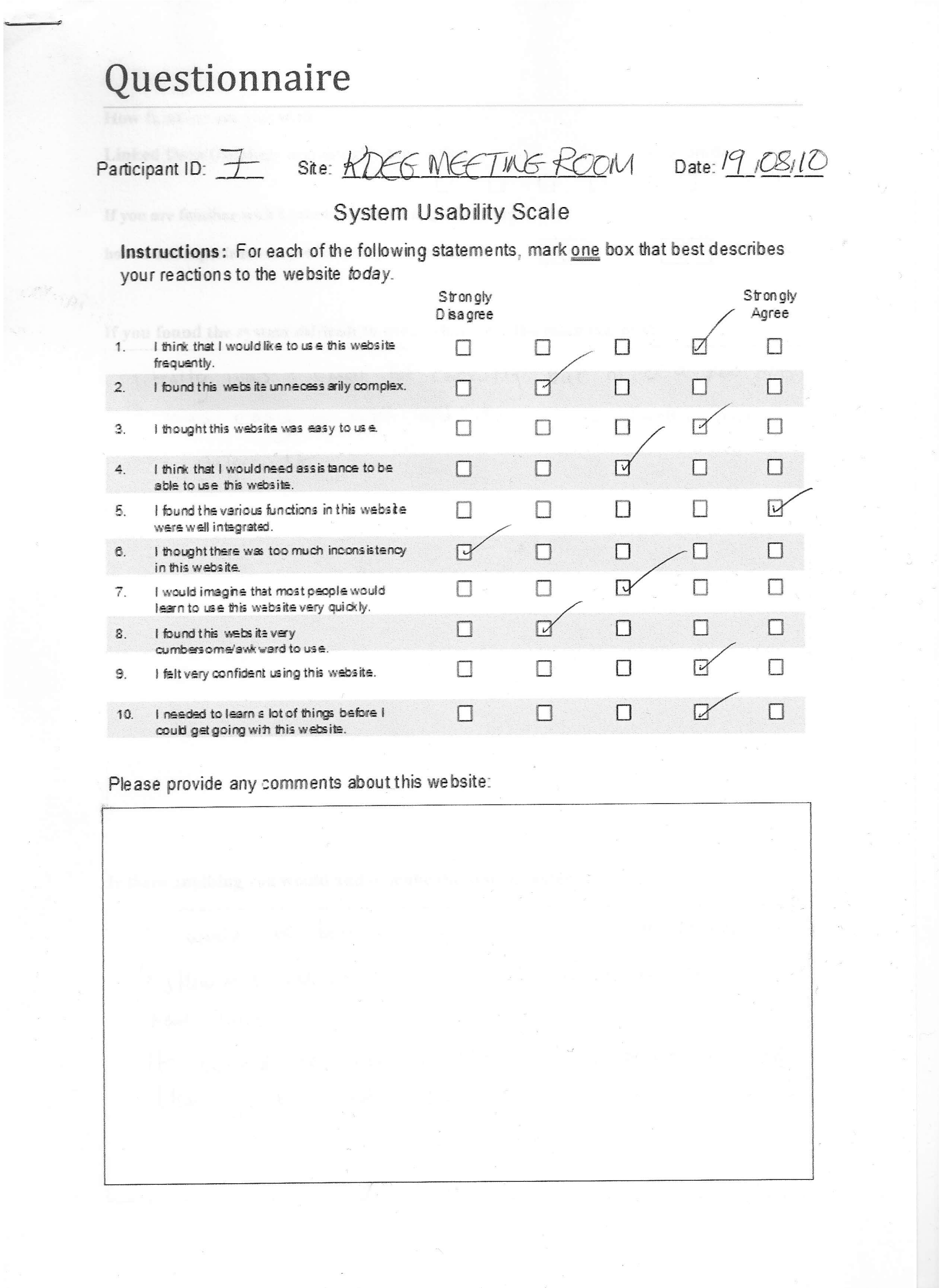
If you found the system difficult to use... what were the main reasons?

Not for me.
But I guess it's hard for users without
semantic web background to start using
the system.

Is there anything you would add to make the system better?

Figure 8.16: Participant ID — 6 *b*

8.2.6 Participant ID — 7



Questionnaire

Participant ID: 7 Site: KDEG MEETING ROOM Date: 19/08/10

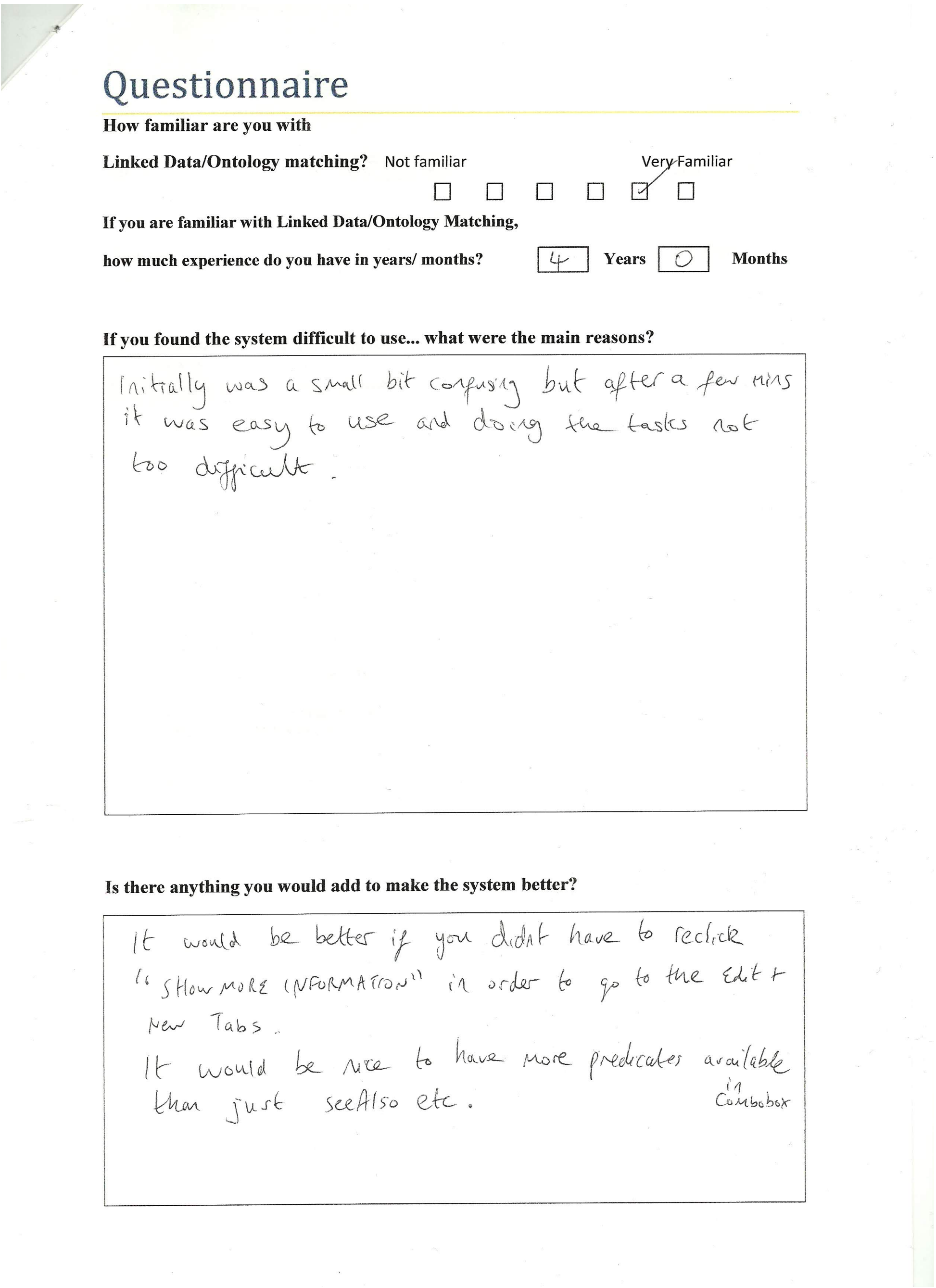
System Usability Scale

Instructions: For each of the following statements, mark one box that best describes your reactions to the website *today*.

		Strongly Disagree				Strongly Agree
1.	I think that I would like to use this website frequently.	☐	☐	☐	☑	☐
2.	I found this website unnecessarily complex.	☐	☑	☐	☐	☐
3.	I thought this website was easy to use.	☐	☐	☐	☑	☐
4.	I think that I would need assistance to be able to use this website.	☐	☐	☑	☐	☐
5.	I found the various functions in this website were well integrated.	☐	☐	☐	☐	☑
6.	I thought there was too much inconsistency in this website.	☑	☐	☐	☐	☐
7.	I would imagine that most people would learn to use this website very quickly.	☐	☐	☑	☐	☐
8.	I found this website very cumbersome/awkward to use.	☐	☑	☐	☐	☐
9.	I felt very confident using this website.	☐	☐	☐	☑	☐
10.	I needed to learn a lot of things before I could get going with this website.	☐	☐	☐	☑	☐

Please provide any comments about this website:

Figure 8.17: Participant ID — 7 *a*



Questionnaire

How familiar are you with

Linked Data/Ontology matching? Not familiar ☐ ☐ ☐ ☐ ☑ ☐ Very Familiar

If you are familiar with Linked Data/Ontology Matching,

how much experience do you have in years/ months? 4 **Years** 0 **Months**

If you found the system difficult to use... what were the main reasons?

Initially was a small bit confusing but after a few mins it was easy to use and doing the tasks not too difficult.

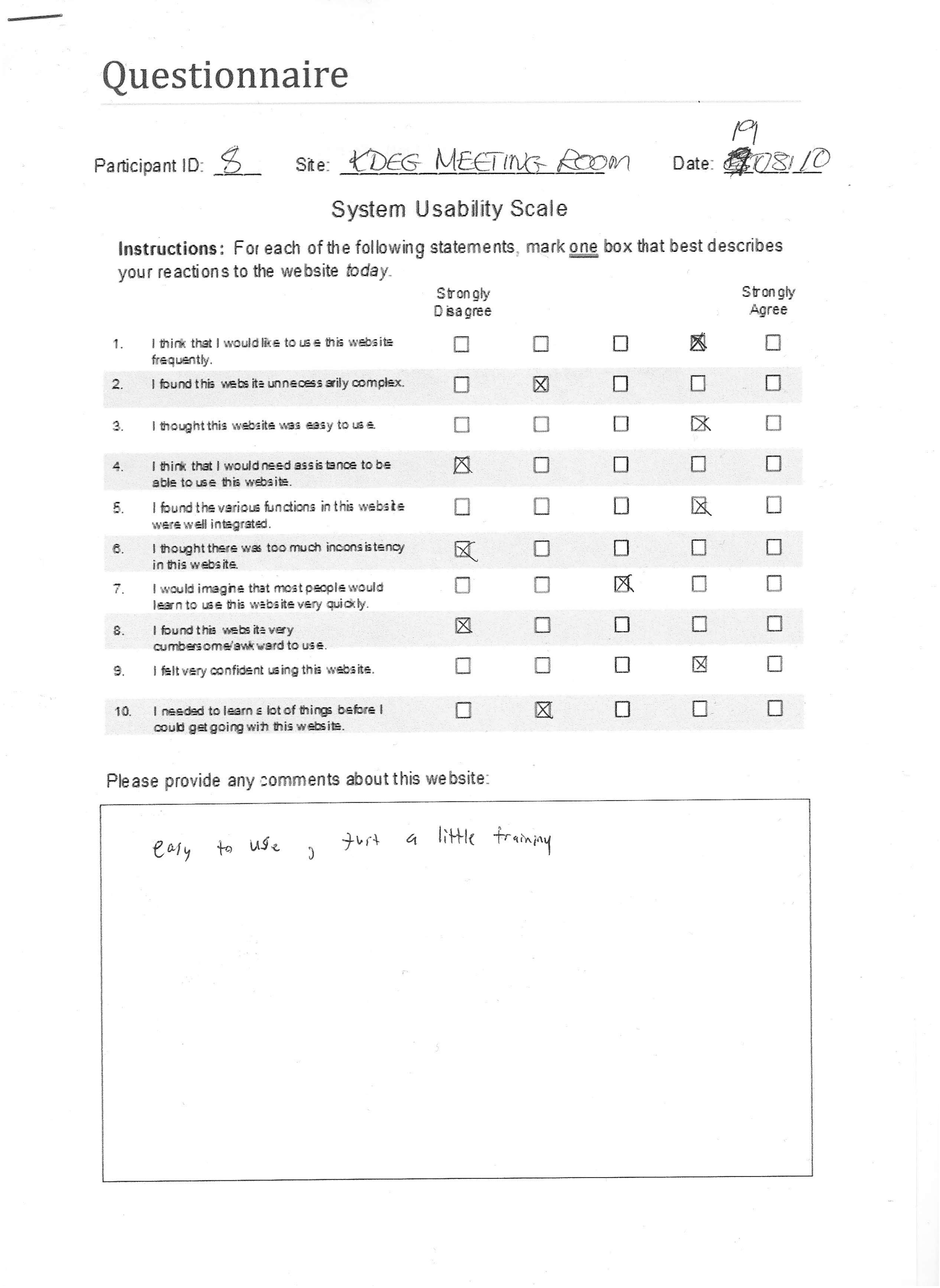
Is there anything you would add to make the system better?

It would be better if you didnt have to reclick "SHOW MORE INFORMATION" in order to go to the Edit + New Tabs.

It would be nice to have more predicates available in Combobox than just seeAlso etc.

Figure 8.18: Participant ID — 7 *b*

8.2.7 Participant ID — 8



Questionnaire

Participant ID: 8 Site: KDEG MEETING ROOM Date: ~~[illegible]~~ 19/08/10

System Usability Scale

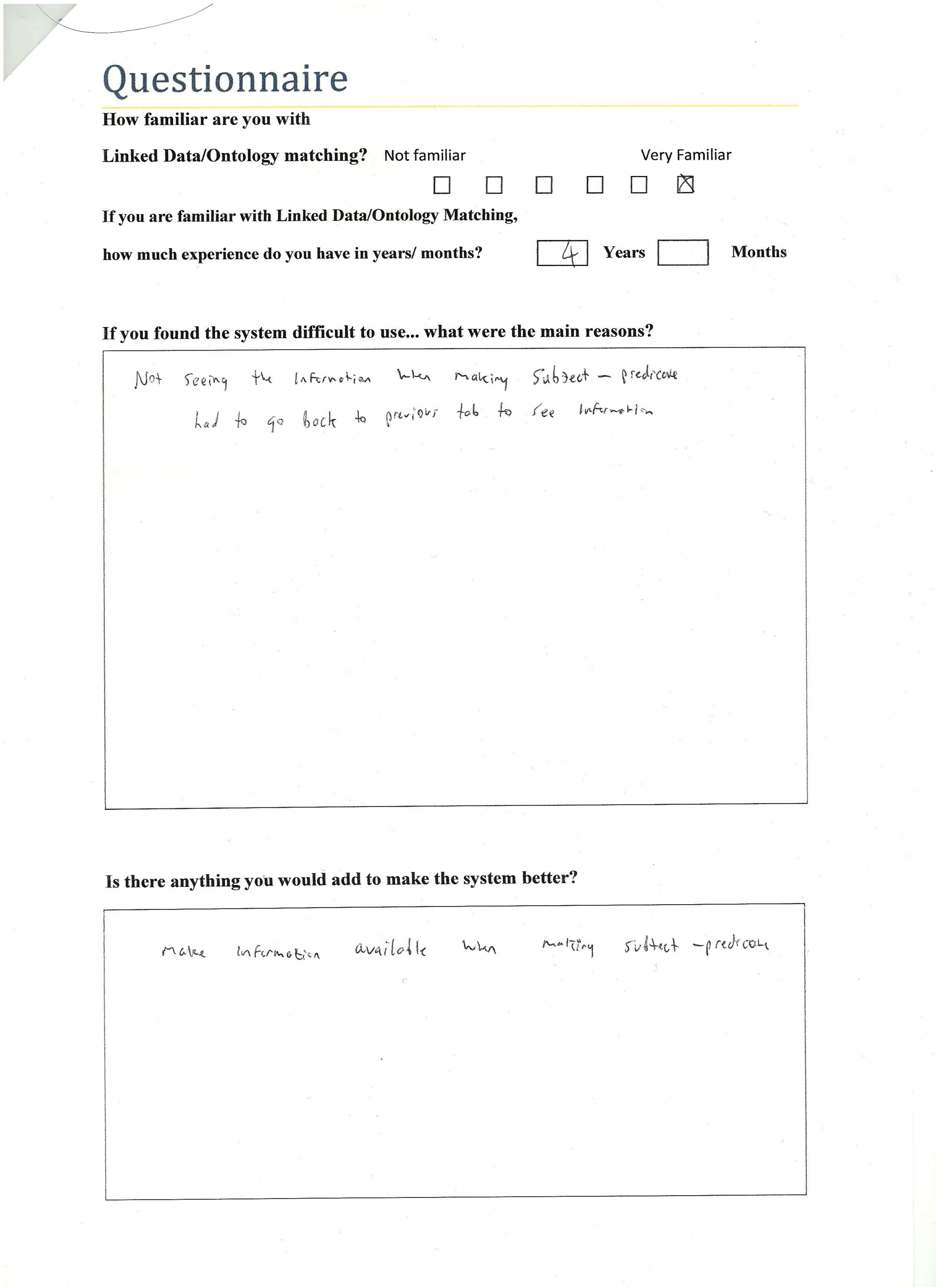
Instructions: For each of the following statements, mark one box that best describes your reactions to the website *today*.

		Strongly Disagree				Strongly Agree
1.	I think that I would like to use this website frequently.	☐	☐	☐	☒	☐
2.	I found this website unnecessarily complex.	☐	☒	☐	☐	☐
3.	I thought this website was easy to use.	☐	☐	☐	☒	☐
4.	I think that I would need assistance to be able to use this website.	☒	☐	☐	☐	☐
5.	I found the various functions in this website were well integrated.	☐	☐	☐	☒	☐
6.	I thought there was too much inconsistency in this website.	☒	☐	☐	☐	☐
7.	I would imagine that most people would learn to use this website very quickly.	☐	☐	☒	☐	☐
8.	I found this website very cumbersome/awkward to use.	☒	☐	☐	☐	☐
9.	I felt very confident using this website.	☐	☐	☐	☒	☐
10.	I needed to learn a lot of things before I could get going with this website.	☐	☒	☐	☐	☐

Please provide any comments about this website:

easy to use, just a little training

Figure 8.19: Participant ID — 8 *a*



Questionnaire

How familiar are you with

Linked Data/Ontology matching? Not familiar ☐ ☐ ☐ ☐ ☐ ☒ Very Familiar

If you are familiar with Linked Data/Ontology Matching,

how much experience do you have in years/ months? 4 **Years** ☐ **Months**

If you found the system difficult to use... what were the main reasons?

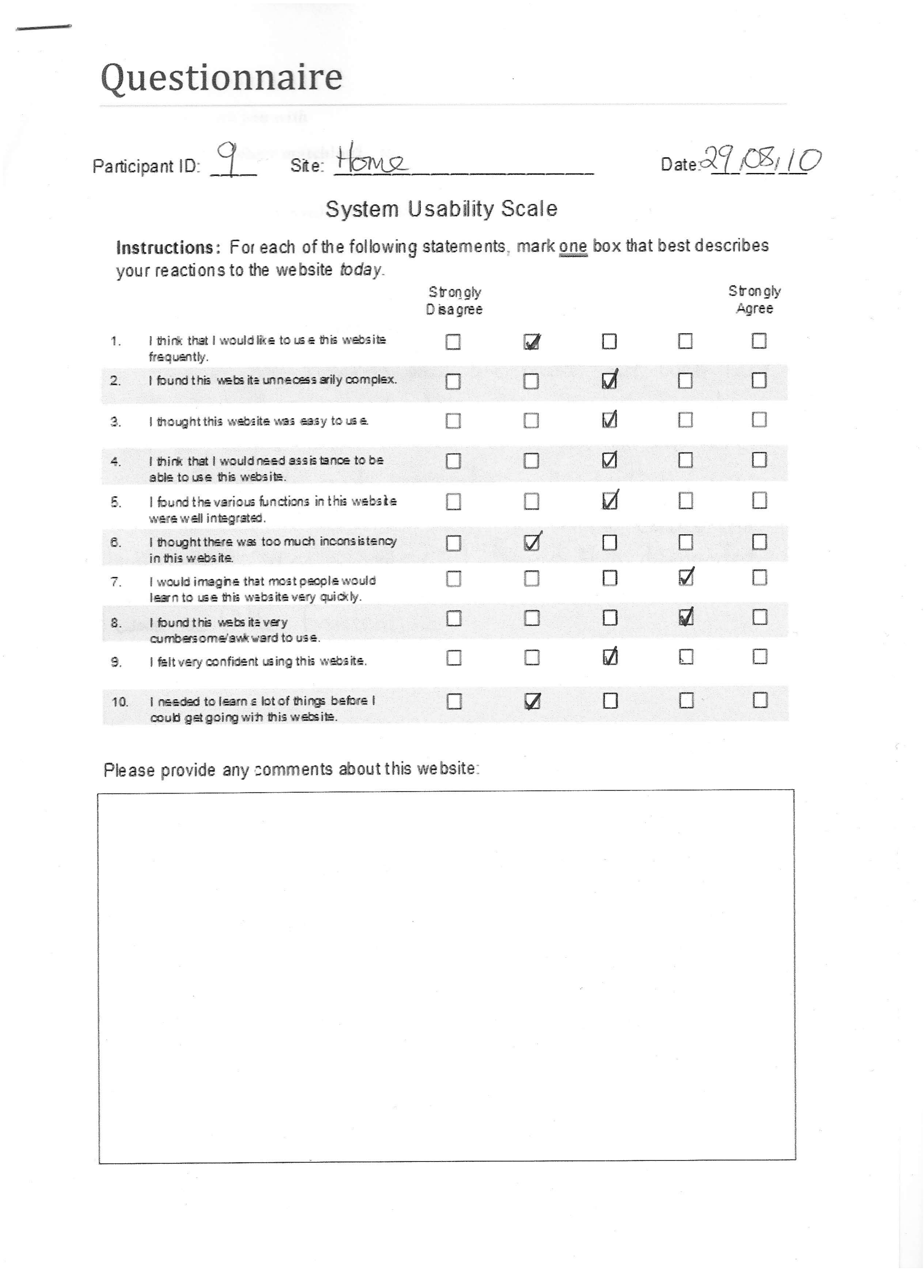
Not seeing the Information when making Subject – predicate
had to go back to previous tab to see Information

Is there anything you would add to make the system better?

make Information available when making subject –predicate

Figure 8.20: Participant ID — 8 *b*

8.2.8 Participant ID — 9



Questionnaire

Participant ID: 9 Site: Home Date: 29/08/10

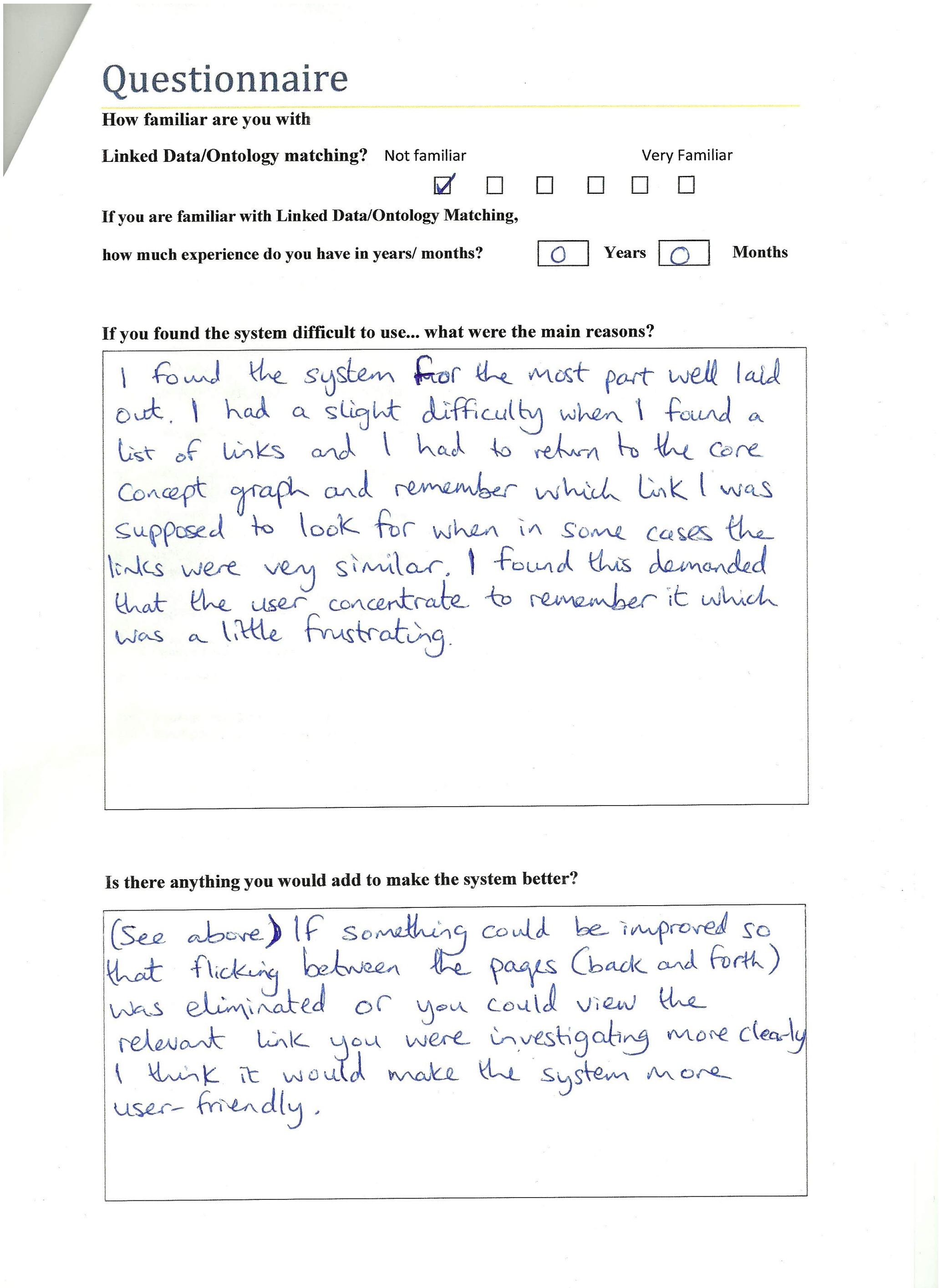
System Usability Scale

Instructions: For each of the following statements, mark one box that best describes your reactions to the website *today*.

		Strongly Disagree				Strongly Agree
1.	I think that I would like to use this website frequently.	☐	☑	☐	☐	☐
2.	I found this website unnecessarily complex.	☐	☐	☑	☐	☐
3.	I thought this website was easy to use.	☐	☐	☑	☐	☐
4.	I think that I would need assistance to be able to use this website.	☐	☐	☑	☐	☐
5.	I found the various functions in this website were well integrated.	☐	☐	☑	☐	☐
6.	I thought there was too much inconsistency in this website.	☐	☑	☐	☐	☐
7.	I would imagine that most people would learn to use this website very quickly.	☐	☐	☐	☑	☐
8.	I found this website very cumbersome/awkward to use.	☐	☐	☐	☑	☐
9.	I felt very confident using this website.	☐	☐	☑	☐	☐
10.	I needed to learn a lot of things before I could get going with this website.	☐	☑	☐	☐	☐

Please provide any comments about this website:

Figure 8.21: Participant ID — 9 *a*



Questionnaire

How familiar are you with

Linked Data/Ontology matching? Not familiar Very Familiar

☑ ☐ ☐ ☐ ☐ ☐

If you are familiar with Linked Data/Ontology Matching,

how much experience do you have in years/ months? 0 **Years** 0 **Months**

If you found the system difficult to use... what were the main reasons?

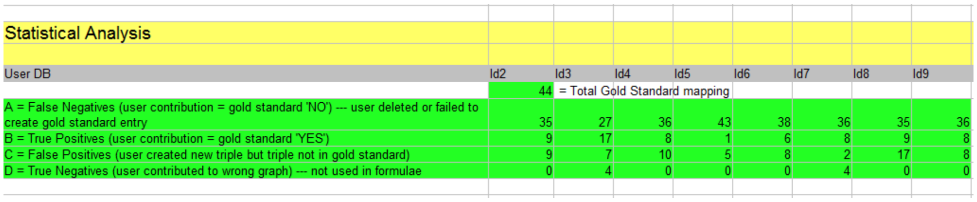
I found the system for the most part well laid out. I had a slight difficulty when I found a list of links and I had to return to the core concept graph and remember which link I was supposed to look for when in some cases the links were very similar. I found this demanded that the user concentrate to remember it which was a little frustrating.

Is there anything you would add to make the system better?

(See above) If something could be improved so that flicking between the pages (back and forth) was eliminated or you could view the relevant link you were investigating more clearly I think it would make the system more user-friendly.

Figure 8.22: Participant ID — 9 *b*

8.2.9 Comparing Gold Standard Matches and User Derived Matches - Overall Results



Statistical Analysis

User DB	Id2	Id3	Id4	Id5	Id6	Id7	Id8	Id9
	44	= Total Gold Standard mapping						
A = False Negatives (user contribution = gold standard 'NO') --- user deleted or failed to create gold standard entry	35	27	36	43	38	36	35	36
B = True Positives (user contribution = gold standard 'YES')	9	17	8	1	6	8	9	8
C = False Positives (user created new triple but triple not in gold standard)	9	7	10	5	8	2	17	8
D = True Negatives (user contributed to wrong graph) --- not used in formulae	0	4	0	0	0	4	0	0

Figure 8.23: Overall Results