STEDI - A Support Tool for Ethical Data Integration

Kavithvajen Kamaraj

A Dissertation

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science (Intelligent Systems)

Supervisor: Declan O'Sullivan

September 2020

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Kavithvajen Kamaraj

September 7, 2020

Permission to Lend and/or Copy

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

Kavithvajen Kamaraj

September 7, 2020

Acknowledgments

I would first like to thank my supervisor *Declan O'Sullivan* for his guidance and motivation throughout this dissertation. His thought-provoking questions, observations, and suggestions were crucial to how this dissertation shaped up.

I am grateful to my supportive parents, who have always gone to great lengths to ensure I get the best of everything. They have always motivated me and encouraged me to pursue my interests. I would also like to thank my best friends - *Mathilde, Toshita, Malavika, and Rahul*, for their moral support during the last year. It really helped me get through tough times. I'm eternally grateful for your friendship, support, and understanding.

Finally, I would like to thank Trinity College Dublin for giving me the wonderful experience of doing this M.Sc course. I truly learned a lot from my time here.

Kavithvajen Kamaraj

University of Dublin, Trinity College
September 2020

STEDI - A Support Tool for Ethical Data Integration

Kavithvajen Kamaraj, Master of Science in Computer Science

University of Dublin, Trinity College, 2020

Supervisor: Declan O'Sullivan

After the integration of a few seemingly virtuous datasets, there exists a possibility that the fused dataset might lead to unethical insights and practices. However, the problem is magnified as human data-integrators cannot look at datasets and determine if it might cause an ethics-related issue. This dissertation aims to solve the aforementioned problem by developing a prototype support-tool called STEDI (Support Tool for Ethical Data Integration). Data integrators can use STEDI as a decision support tool to identify ethical issues present in datasets and to also predict the materialisation of ethical problems that could arise from integrating these datasets. The current version of STEDI focuses on working with linked-data datasets. When two or more datasets are loaded into it, STEDI will start by analysing the vocabularies and predicates present in the datasets to understand the kind of data that is present in it. State-of-the-art Natural Language Processing techniques are deployed to understand the data that the predicate entails. The results of the analysis are stored in a specially-designed ethics ontology that can answer a limited amount of questions regarding the ethical views of the datasets. To accurately identify data-integration related ethics issues, the data in the ethics ontology is retrieved to identify four highly-specific ethical issues. Finally, an ethics report is generated containing the results of the ethical analysis. Through evaluation, it is concluded that the proposed approach is viable and can be leveraged in the future to identify ethics-related problems in linked-data datasets.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

It is a known fact that all significant for-profit organisations collect and mine data. Analytics is performed on this raw data to convert them into actionable knowledge, and to discover new patterns and relationships that are not apparent to human beings [6]. Hence, all modern organisations have moved away from a traditional instinct-based decision-making process and have since adopted a more data-driven decision-making process [7]. Though these methods have proven to be very profitable for these organisations, sometimes there are unforeseen ethical consequences to applying these learning algorithms on large datasets. The problem is often more deep-rooted than it seems on the surface; it begins with how the engineers behind such technological advancements are not ethicists by profession. Their primary focus is usually on advancing science, and thinking about ethical behaviour often tends to slow them down [8]. Although the engineers are not entirely at fault in this instance, the ignorance of ethics in technology has led to a wide range of issues (such as user-privacy leakage [9] and biased Machine Learning models [10]), which have impacted the end-users. Data integration is a standard operation performed while working with multiple datasets [7]; it is used to extract additional information that would not be available otherwise. However, in some cases, after the integration is complete, the datasets might reveal more information than initially intended. Here are some examples of ethical issues arising due to the integration of datasets:

- In October 2006, the movie rental service Netflix had announced a public competition called the "Netflix Prize". As part of the contest, the company released a massive database containing movie ratings from 480,189 users. The challenge required the contestants to develop a movie recommendation algorithm that could beat Netflix's internal algorithm. Netflix had taken some necessary precautions to anonymise the dataset by removing all customer identifying details [11]. However, in 2007, two researchers from the University of Texas at Austin cross-correlated data between the anonymous Netflix dataset and a similar open dataset from IMDb (Internet Movie Database). Not only were they able to correctly identify the real identities of the users in the Netflix dataset, but they were also able to glean information about the users like their sexual and political preferences based on their ratings for movies revolving around these topics [9]. Hence, even from seemingly unimportant data like movie ratings, the user's privacy can be at risk by integrating it with another appropriate dataset.

- In 2015, Latanya Sweeney purchased a patient-level health dataset from The State of Washington for $50. This dataset contained a lot of information about patients and the reason for hospitalisation, but it did not have the patient's name or address. Sweeney was able to cross-reference these records with another publicly available dataset - old newspaper articles. These news articles often used the term "hospitalised" and then proceeded to mention the name of patients and the reason for which they were hospitalised. Thus, she was able to correctly identify the actual names of the patients in the dataset for 43 per cent of the news articles she found [12].

- In 2016, Larson et al. analysed a tool called COMPAS, which was widely used by judges, probation and parole officers to estimate the likelihood of a convicted criminal reoffending. The results of their study revealed that the algorithm behind COMPAS was racially-biased. Black defendants were usually predicted to be at a higher risk of reoffending than they actually did, whereas the opposite was true for white defendants [13]. The underlying learning algorithm was not built with ethics in mind, and a racially-biased recommendation algorithm was the result.

Companies often look into integrating publicly available open-datasets with their

internal datasets to provide better value for their customers and often justify these integrations by claiming that they are open datasets. The individuals working on the data integration are then put in a spot to decide what open data is ethically acceptable to be integrated and what is considered as a potential ethical risk. More often than not, considering ethics might seem impractical, and the potential benefits of data integration might overweigh the potential ethical issues that might arise due to the integration [14]. However, ethical companies should not unmindfully combine datasets; they must accept responsibility and carefully examine the moral consequences of their integration [6].

Previous work by Ashish Lochan [5] delved in the realm of identifying ethical issues in datasets but it focussed only on GDPR (General Data Protection Regulation) issues, and the proposed G-DIEP tool could only process one dataset. There is currently no existing gold-standard method or application to automatically check for such ethical data-integration issues. Hence, it would be useful to have an automated support tool for data integrators that could take multiple datasets as inputs, analyse them for ethical issues and output a simple ethics report. This ethics report can then aid the data integrator to decide whether or not to integrate the datasets.

1.2 Research Question

To what extent can a knowledge-driven system accurately predict and report ethics-related issues that could arise after the integration of two or more linked-data datasets?

1.3 Research Objectives

The following objectives need to be achieved to answer the research question:

1. Determine an approach to identify the type of data stored in a linked-data dataset.

2. Establish a method to analyse multiple datasets and identify ethical concerns in them.

3. Establish a method to predict ethical issues that might arise due to data integration.

4. Develop a prototype tool that can analyse multiple datasets, predict and generate a report about the materialisation of an ethical issue if the datasets were to be integrated.

5. Evaluate the performance and viability of the approach.

1.4 Contribution

The main contribution of this dissertation is a prototype tool called STEDI, which is an acronym for Support Tool for Ethical Data Integration. Given a few linked-data datasets, STEDI first analyses the vocabularies used in the datasets and checks for the presence of any potentially risky vocabularies. Next, it uses Natural Language Processing (NLP) techniques on the predicates of all the triples in the dataset to identify the kind of data stored in it. The results of the analysis are then pushed onto a specifically developed ethics ontology - it is designed to answer a limited set of ethics-related questions about the datasets. The ethics ontology is then queried to identify ethical issues either in the datasets itself or to predict the ethical problems that might arise after the integration of the datasets. Finally, based on the results of the queries, STEDI will generate an easy-to-read ethics report for the benefit of the human data-integrator.

1.5 Structure Of The Dissertation

The rest of the dissertation is organised as follows:

Chapter 2 - *Background:* Reviews the relevant background and state-of-the-art techniques for analysing datasets with an ethical perspective

Chapter 3 - *Design:* Discusses the identified requirements for building the prototype tool, the functional components of the approach and the high level operation of the tool.

Chapter 4 - *Implementation:* Discusses the implementation details, the security considerations, the challenges faced and how they were addressed.

Chapter 5 - *Evaluation:* Explains the performance tests and the user evaluations that were carried out to test the viability of this tool.

Chapter 6 - *Conclusion:* Summaries the work done and concludes the dissertation.

Chapter 2

Background

This chapter reports on the various kinds of ethical concerns that are generally present in technology, and in particular those that arise due to data integration. It also illustrates the popularity of linked data in general and how some ontologies help model ethics-related data. Also, specific research work is explored in detail as its objectives were close to this dissertation's goals and thus helped lay the foundation for this work. Finally, some commercially available state-of-the-art tools in this area are explored.

2.1 Ethical Concerns Of Data

Technology has been growing at a rapid pace in the past few decades, and businesses worldwide have been doing everything they can to capitalise on this growth as it always makes them more efficient with their resources. With the advent of data science (a field in which scientific methods and algorithms are used to extract useful knowledge from raw data [15]), businesses can gain more from technology than they ever have. The only requirement to capitalise on the power of data science is data; more data usually results in better outcomes. Hence, we are currently in the age of "*data deluge*" [16] where every product and service operated by a for-profit company collects an immense amount of data in hopes that this data will ultimately result in more profits for the company. The problem with collecting and processing such large amounts of information is that it might not always be in the best interests of the users of the services or products sold by a company.

An example would be the analysis of a person's medical and social data, which can be used to give access to personalised medicines, care and predictive measures. However, it can also lead to increased health insurance rates, making them too expensive (maybe even unaffordable) for those at risk [17]. Hence, it is of utmost importance that organisations understand the ethical ramifications of their data management practices and act accordingly.

2.2 Google's Advertising Services

Datta et al. identified that Google's ad services were not as ethical as they claimed to be [18]. By developing a tool called AdFisher, the authors found that immediately after they browsed websites regarding substance abuse, they received advertisements related to rehab, but the *"Ads Settings"* page did not reflect this change [18]. Therefore, this demonstrates that despite Google portraying themselves as being transparent, their advertisement recommendation algorithm was actually arcane, and their *"Ads Settings"* page did not always accurately represent the change in data. Furthermore, the authors were also able to identify that the ad system was discriminating against users based on their genders because switching the gender from male to female significantly reduced the number of high-paying jobs the user was shown [18].

2.3 Businesses, Profits And Ethics

In the paper titled "Is Ethical Behavior Good for Business?", author Miller declares that it is suitable for businesses to indulge in behaviour that lies in the intersection of both ethical and profitable behaviour (as shown in Figure 1) [2]. However, the problem arises when organisations assume that data analytics depends on mathematical algorithms and therefore, cannot be biased [17]. They often overlook the fact that all inherent biases and discriminations present in the data get amplified during the data analysis [17]. Hence businesses need to operate and develop technologies with ethics in mind.

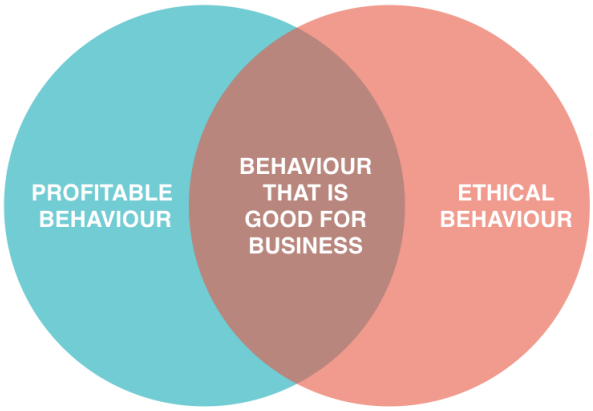

Figure 2.1: Venn Diagram showing what is good for business [2]

2.4 Privacy Issues

A common issue that stems from organisations collecting too much data is privacy-leakage. The data subject's privacy must be of utmost importance but is often neglected [19]. When publishing statistics and other open data, organisations tend to redact privacy-sensitive data by using simple anonymisation techniques. However, as multiple studies have shown [12, 9, 20, 21, 22], it is extremely hard to anonymise data properly. By integrating the anonymised dataset with another appropriate dataset, the data can be de-anonymised in most cases, thereby allowing to identify the users uniquely and expose their sensitive details.

2.5 Ethical Issues In Data Integration

An essential step in the information extraction process is the integration of two or more datasets to create a unique dataset, which will, in turn, be used by complex statistical, machine learning or data mining models [23]. Data integration usually involves three main steps, and they are all prone to different ethical issues [23]:

1. ***Schema matching:*** It is the alignment of the schemas (if available) of the datasets. The groups treated to be fair and diverse in the source can become under- or over-represented after data-integration.

2. ***Entity resolution:*** It is the identification of the same entities stored in different

datasets. Integrating datasets that generally protect the identities of the data subjects might generate a new dataset that violates the privacy of the users.

3. ***Data fusion:*** It is the process of merging two or more datasets to construct a new integrated dataset. Violation of user's privacy may happen if appropriate care is not taken while anonymising the datasets [12, 9, 20, 21, 22].

2.6 Requirements For Responsible Data Management

Based on Abiteboul and Stoyanovich's work [24, 25], the following are the requirements for any tool or algorithm involved in data management:

- Be fair, i.e., without any unintentional bias or discrimination
- Be transparent
- Be accountable
- Be aware of the data subject's rights and preferences
- Ensure diversity
- Responsible by design
- Privacy by design

2.7 Linked Data

In 2006, Tim Berners-Lee, the inventor of the World Wide Web (WWW) and the director of the World Wide Web Consortium (W3C), coined the term *"Linked Data"* to indicate the standards and practices suggested by the W3C to publish interconnected data on the web [26]. The interconnected web of data is called the semantic web, and it is an extension of the WWW. The semantic web is essentially meant to publish structured information that is linked with each other, making the stored data more useful through semantic queries [26]. The standout feature of linked data lies in its

ability to reference pieces of data that were published by different users on the web. It is built on standard web technologies and can share the structured data in such a way that both humans and computers can consume it [27]. Linked Data has the capability to reuse data in ways that were unforeseen at the data source, and this is precisely what exponentiates the value provided by the semantic web. Linked data is the core of what the semantic web stands for, and it is essentially the integration and automated reasoning of a web of data [28]. Thus, a core concept of linked-data is data integration, and it is required for data to be useful.

2.7.1 Linked Data Principles

The Linked Data principles, as stated by Tim Berners-Lee, are [29]:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs, so that they can discover more things.

2.7.2 Vocabularies & Namespaces

Linked-data datasets often use pre-existing vocabularies because that allows for proper expansion of the semantic web. Vocabularies are well-documented domain-specific ontologies that can be used to model other data. Every vocabulary has its own namespace, which is an identity space on the internet that the vocabulary's maintainers own and control.

2.7.3 Triple

A triple is the fundamental unit of expression in linked data. It is called a triple because any data can be represented with a three-section relationship - subject, predicate and object. The subject is the primary entity that the data is about, the predicate is the type of data, and the object is the data itself.

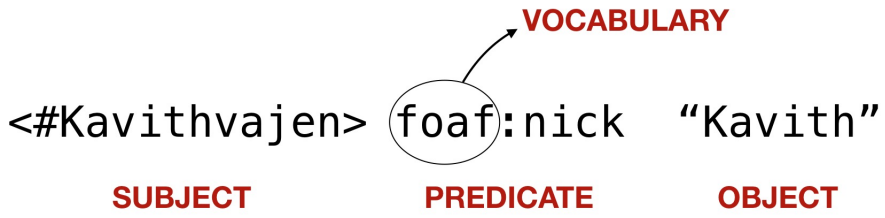

Figure 2.2: Example of a triple

In the example depicted in figure 2.2, <#Kavithvajen> is a URI and it is the **subject**. `foaf` stands for *"Friend Of A Friend"* which is a vocabulary used to describe people. `nick` is the term depicting a nickname in the FOAF ontology[1], and `foaf:nick` as a whole is known as the **predicate**. Finally, the string `"Kavith"` is the **object**. Therefore, the example above conveys the information *"Kavithvajen's nickname is Kavith"*.

2.7.4 Linked Data Applications

Since linked data almost always provides more value to the data when compared to traditional database management systems, there exists a lot of real-life uses for it, and some of them are:

- Linked open government data has the potential to be very useful for applications that require public administration knowledge [30, 31, 32].
- Cultural heritage institutions are developing ontologies, vocabularies and applications that can take advantage of the semantic web [33].
- Biomedical data is converted into actionable knowledge with the use of a knowledge graph that is driven by linked data principles [34].

[1] `http://xmlns.com/foaf/spec/`

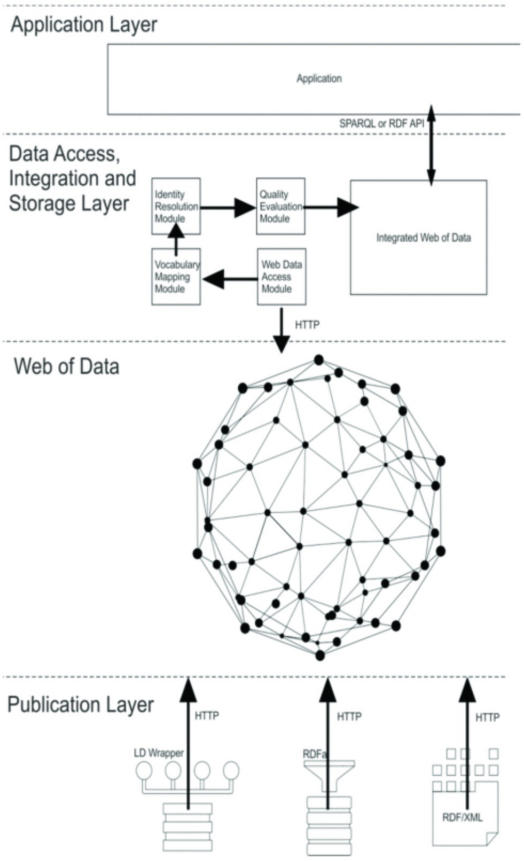

Figure 2.3: A typical linked data application architecture [3]

2.8 Ethics In Linked Data

2.8.1 Data Privacy Vocabulary (DPV)

The W3C Data Privacy Vocabulary and Controls Community Group (DPVCG) has created a semantic web ontology to represent personal data [4]. Before DPV, there were no clear vocabularies to describe information about the processing of personal data. The core concepts of DPV (as shown in figure 2.4) are the data controller, recipient, data subject, technical and organisational measures, legal basis, purpose, processing, personal data category [4]. Some of the concepts defined in DPV are reused in the ethics ontology that is designed as part of this dissertation.

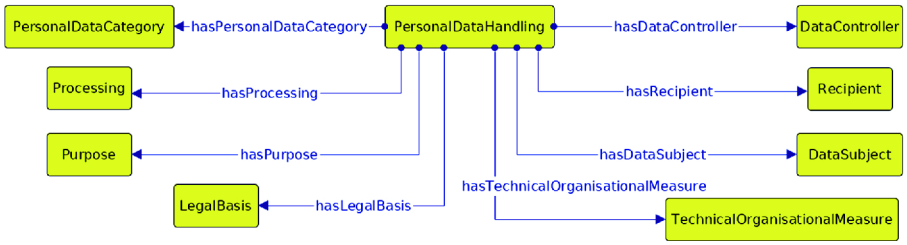

Figure 2.4: Core concepts of DPV [4]

2.8.2 GDPR & Linked Data

Quite a bit of work has been done in this area of linked data and GDPR(General Data Protection Regulation). GDPR is a regulation in the European Union (EU) and European Economic Area (EEA) that pertains to the processing and movement of personal data. The main aim of GDPR is to give the individuals located in the EU or EEA full control over the data they generate, and prohibit multinational companies from transferring data generated in the EU or EEA outside these regions [35, 36].

GDPR text extensions (GDPRtEXT) is a linked open dataset that exposes the GDPR as a linked data resource [37]. GDPRtEXT enables other linked-data resources to reference the information that addresses specific articles in the GDPR. This allows businesses to build applications that can automate the retrieving and generation of GDPR-related details - such as compliance towards specific obligations, privacy policy generation or management of certain business processes.

Since consent is a vital part of processing personal data under the GDPR, Pandit et al. propose a semantic web ontology called GConsent to model user consent and its related information [38]. GDPR imposes restrictions on the validity of the consent given by the data subject and gives them the right to withdraw their consent at any time. GConsent can model all of the information related to the data subject's consent, including the current validity status.

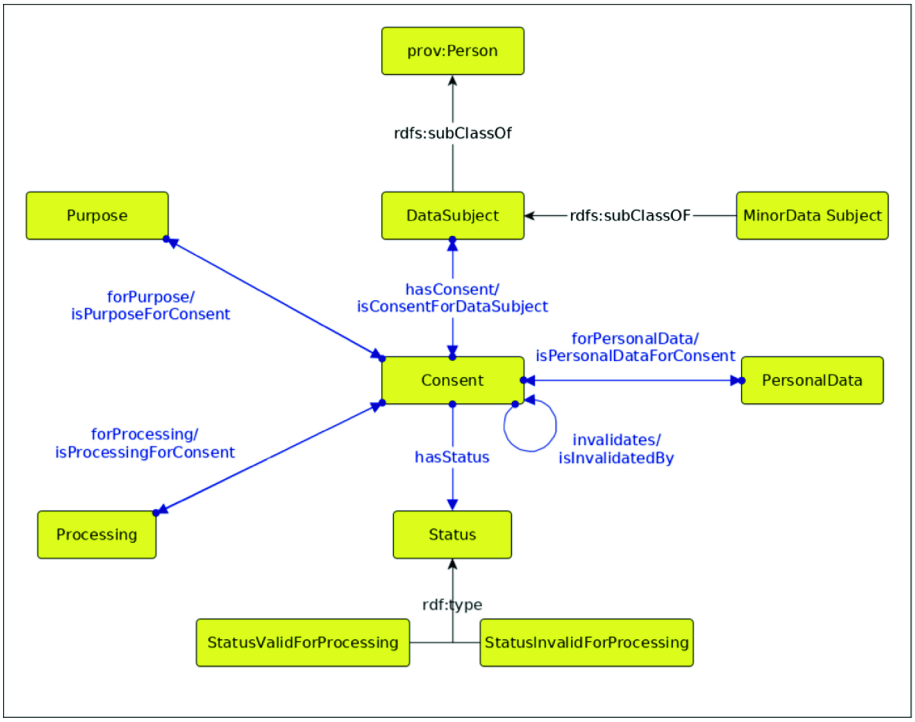

Figure 2.5: Core concepts of DPV [4]

2.9 State Of The Art

2.9.1 G-DIEP (GDPR Data Integration & Ethical Perspective)

The research question in Lochan's work is very similar to the research question stated in this dissertation [5]. The only differentiator is that Lochan's work revolves around the identification of GDPR-related issues before the integration of linked-data datasets. In contrast, this dissertation does not focus only on GDPR; it instead takes into account a broader scope of ethical concerns.

Lochan proposes a decision-making tool called GDPR Data Integration & Ethical Perspective (G-DIEP); it incorporates an ethics-based decision-making phase into the data integration process. First, the G-DIEP tool identifies the vocabularies present in the linked-data datasets using the data from Linked Open Vocabularies (LOV) [2]

[2] `https://lov.linkeddata.es/dataset/lov/api`

[39], and this would give it some insights as to what kind of data was contained in the dataset. Next, the information about the type of data and the vocabularies used is pushed onto an ethics ontology; the ethics ontology's sole purpose was to store the GDPR-related aspects of the dataset. The ethics ontology was specifically designed to answer ethics-related questions that have been alluded from Horizon 2020 - EU Ethics Questions, and the Trinity College Dublin ethics questions [5]. Finally, the ethics ontology is queried using the chosen ethics questions, and a report is generated.

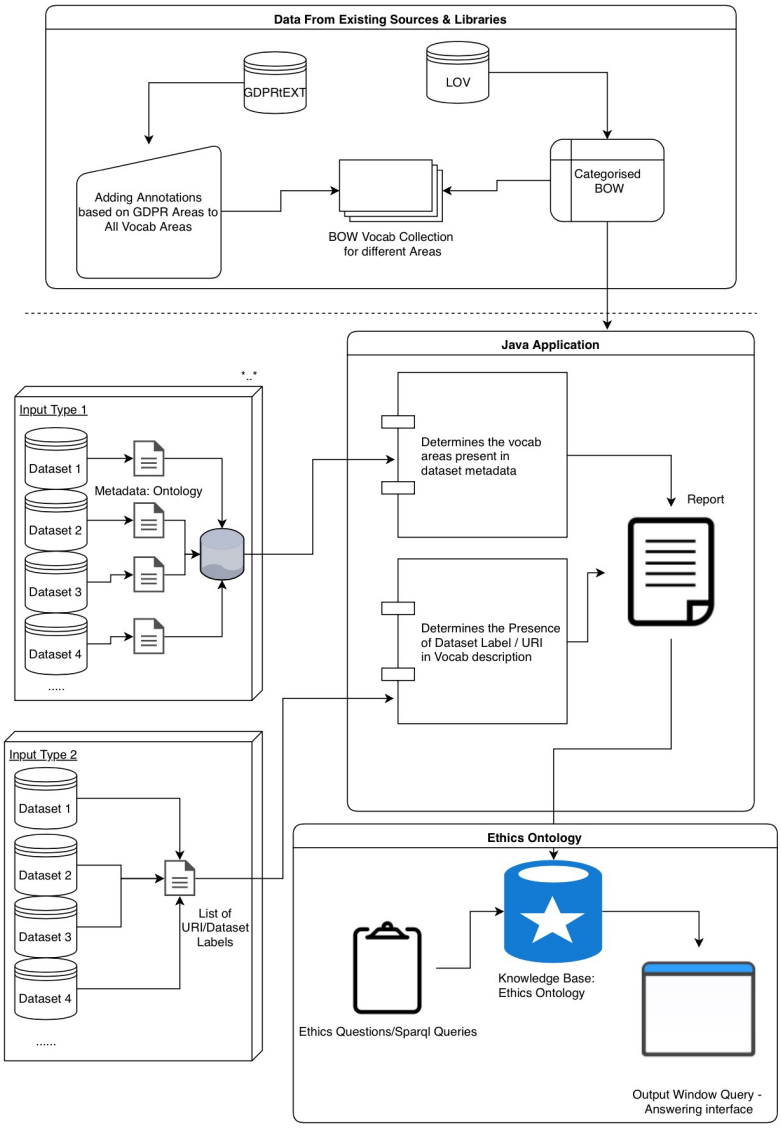

Figure 2.6: Detailed Architecture Diagram of G-DIEP [5]

A significant limitation of G-DIEP is that it depends on the metadata and vocabulary of the datasets to accurately identify GDPR-related issues. Hence, G-DIEP will not be able to retrieve the correct information if: (i) The metadata is missing, incomplete or inaccurate. (ii) The vocabulary is not open and well-known.

2.9.2 Other Significant State-Of-The-Art Approaches

- **SPECIAL (Scalable Policy-aware Linked Data Architecture For Privacy, Transparency and Compliance):** This project uses semantic web technologies to provide solutions for compliance-related tasks. It is a European H2020 project and is focused on working with privacy-preserving big data technologies. SPECIAL aims to allow the data subjects to share more data while guaranteeing the protection of the data they share [40].

- **BigID:** This commercial automation tool simplifies the compliance of the following EU GDPR requirements - data minimisation, data subject rights, consent management, data residency, and breach notification [41]. The tool uses machine learning techniques to understand personal data and its context, thereby accurately determining the changes in personally identifiable information.

- **DAPRECO (DAta Protection REgulation COmpliance):** This project deals with the creation of a knowledge-base modelling the legal informatics of GDPR compliance [42]. The knowledge-base can host a lot of legal interpretations that occur in the law-related domain where the various laws are perceived differently by different subjects (lawyers, judges and regulators). It is therefore regarded as an innovative tool for authorities in the GDPR compliance assessment [42].

- **SAS for Personal Data Protection:** This commercial tool provides a one-stop solution for organisations to comply with data protection regulations like GDPR and CCPA (California Consumer Privacy Act). It helps in every step of the process - identifying, governing and protecting personal data [43].

- **BPR4GDPR (Business Process Re-engineering and functional toolkit for GDPR compliance):** It is an EU Horizon 2020 project that aims to provide

a set of tools under the concept of Compliance-as-a-Service (CaaS) which can benefit any kind of organisation that is required to comply to GDPR [44]. It provides a comprehensive framework that is scalable and can support end-to-end GDPR-compliance-related processes [45].

- **CREDS (Cyber Research Ethics Decision Support):** The project provides a decision support methodology, conceptual framework and an interactive tool that facilitates cyber-based research (network and system security) by accurately identifying, rationalising and managing ethical issues that might arise due to the research [46].

2.10 Gaps & Opportunities For Further Work

A quick glance at the state-of-the-art approaches is enough to note that all of them focus only on GDPR. Though GDPR is a well-formed regulation, it concentrates exclusively on data protection and privacy, and there are a plethora of other ethical issues that it does not consider. Ethics is not a set of rules or regulations; it is instead a more philosophical way of thinking that refers to the principles concerning the distinction between right and wrong or good and bad behaviour. A tool that can identify a broad range of ethical issues in data can help steer technical innovations in a way that benefits all of humanity. Since data integration is a crucial step that is involved in data management [23], the proposed tool would benefit the most if it were to operate at this layer. Predicting ethical issues before they materialise, and providing a clear report of what data points are causing the issues will be beneficial for a data integrator. By using this tool, the informed data integrator will now be able to decide if it is worth merging these datasets as is or if any further precautions need to be taken.

Chapter 3

Design

In this chapter, the overall design of STEDI will be discussed, and the various functional components of the system will be introduced. Every design choice that was made during the development of this tool will be explored, and a high-level operational workflow of the tool will be portrayed.

3.1 Requirements

The proposed tool needs to be able to:

1. Allow loading of one or more linked-data datasets
2. Identify ethical concerns in datasets
3. Predict the materialisation of complex ethical issues that might arise after the datasets have been integrated
4. Store the results of the ethics analysis in a specially-designed ethics ontology
5. Query the ethics ontology and generate an ethics report for the consumption of the data integrator

3.1.1 UML Use Case

Since STEDI is marketed as an automated decision-support tool, the user only has to upload the datasets and fill a small questionnaire regarding the datasets. STEDI will

Figure 3.1: UML use case diagram

automatically identify all the ethical concerns in the datasets and generate a detailed ethics report for the consumption of the user. In the background, STEDI stores the identified ethical issues in a specially-designed ethics ontology which will be queried while generating the report.

3.2 Design Overview

For STEDI to satisfy the requirements stated in Section 3.1, an appropriate design choice would involve five main components - User Interface, Loading Service, Risk Identification Service, Ontology Management Service, Reporting Service. Figure 3.2 is the sequence diagram that showcases the working of STEDI.

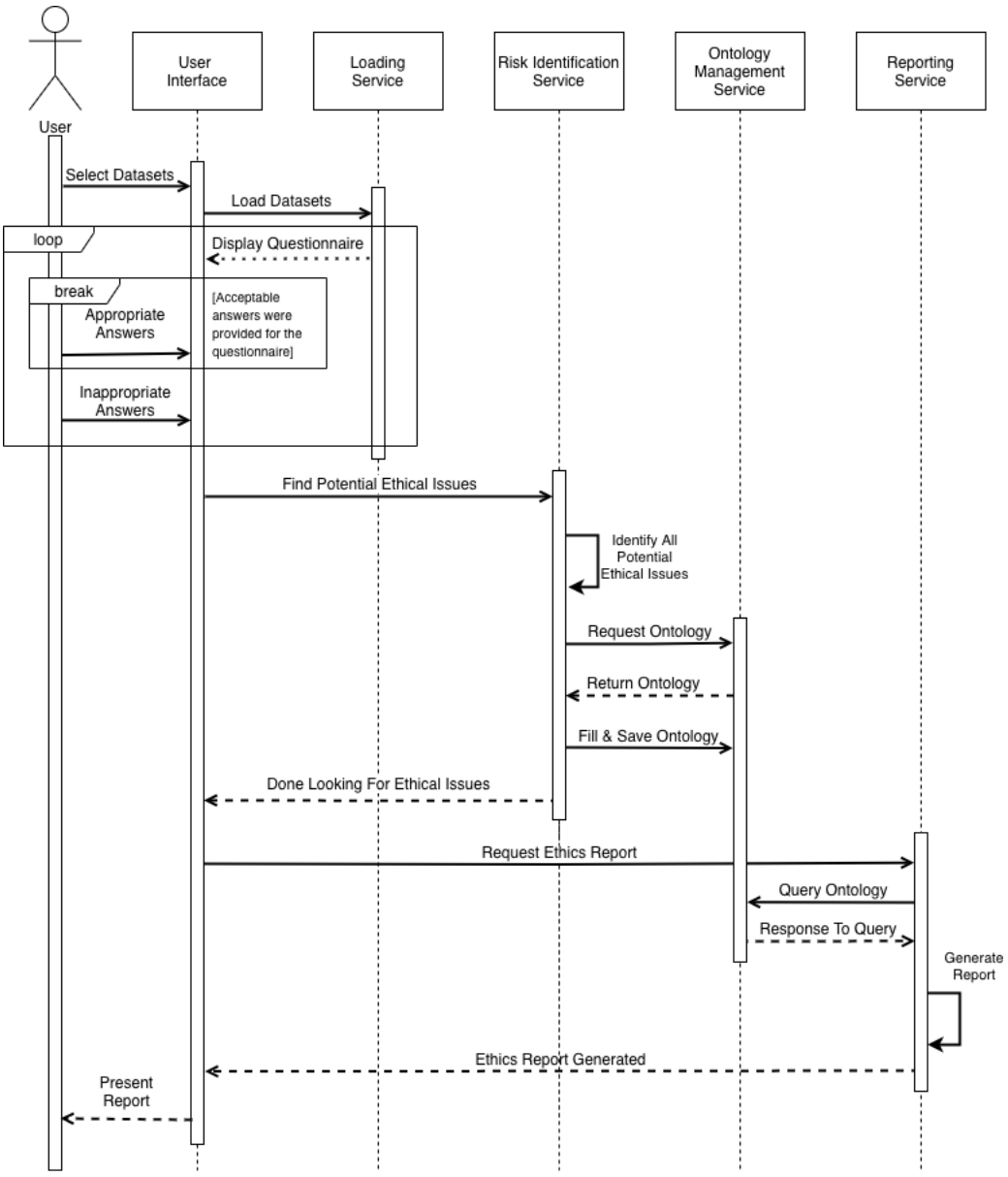

Figure 3.2: Sequence diagram depicting the working of STEDI

3.3 User Interface

The User Interface (UI) is the primary interacting medium for the data integrator. This component will also be the controller for the whole tool; that is, it will be responsible for liaising with the other components. Also, the look and feel of the tool do not fit in the scope of this dissertation and are not considered as a primary objective in the design of the UI. The final requirements of the UI are:

1. It must liaise with all the other components and ensure successful operation of the tool

2. It should provide appropriate feedback to the user and keep them informed about the ongoing process

3. It must be minimalistic, responsive and easy to use.

3.4 Loading Service

The loading service deals with correctly loading the datasets and priming them for analysis in the coming stages. A downside with G-DIEP (discussed in Section 2.9.1) [5] is that the tool depended on the metadata of the datasets. Hence, to overcome that issue, STEDI implements a different strategy where the data integrator is asked to fill in a small questionnaire regarding every dataset. The questions are:

1. Enter the name of the data controller that the data subject originally agreed to share their data with.

2. If any files are attached to the dataset, then enter some keyword(s) describing the file. Otherwise, leave the entry blank.

3. Are the data subjects, individuals or groups?

The answers to these questions are stored in the specially-designed ethics ontology that helps STEDI understand the kind of data stored in the datasets. These answers are crucial in identifying particular ethical concerns in the datasets.

3.5 Risk Identification Service

The risk identification service is the most critical component in this system; it deals with identifying ethical concerns in the datasets and predicting the actualisation of ethical issues that might arise after merging the datasets. Instead of relying only on the metadata of the dataset like G-DIEP (discussed in Section 2.9.1) [5], state-of-the-art Natural Language Processing (NLP) techniques are also used to identify the ethical issues in the datasets.

3.5.1 Vocabulary-Based

The vocabularies used in the linked-data datasets will usually present a rough idea as to what kind of data is present in the dataset. STEDI uses this technique to gain some knowledge about the data stored in the datasets. This technique was inspired by the work done in G-DIEP by Lochan (discussed in Section 2.9.1) [5]. Linked Open Vocabularies [1] [39] play a significant role in this stage as their vast and detailed database of linked-data vocabularies is very useful to get an understanding of what each vocabulary normally represents. When vocabularies dealing with sensitive data are identified in a dataset, the ethics ontology is updated accordingly. Table 3.1 lists some common ethically-sensitive vocabularies used in linked-data datasets.

Name	Prefix	Namespace URI	Describes
Friend-of-a-Friend	foaf:	`http://xmlns.com/foaf/spec/`	People
WGS84	geo:	`http://www.w3.org/2003/01/geo/`	Geo positioning
An ontology for vCards	vcard:	`http://www.w3.org/2006/vcard/ns`	vCards (virtual contact files)
BIO: A vocabulary for biographical information	bio:	`http://purl.org/vocab/bio/0.1/`	Biographical Information

Table 3.1: Some vocabularies that model ethically-sensitive data

[1] `https://lov.linkeddata.es/dataset/lov/api`

3.5.2 Predicate-Based

As was explained in Section 2.7.3, the predicate denotes what the data is about. Hence, a significant design choice was made to identify a way to extract the predicate and thus gain an understanding of what the data was about. STEDI would move through every triple in the dataset and check to see if any of the predicates showed indications of being part of a triple storing potentially unethical data. The advantage of using the predicate to identify ethics issues is that it does not depend on the creator of the dataset to have reused vocabularies or to have added proper metadata.

Since all entities in linked data are URIs, predicates must also have a typical URI syntax that follows a hierarchical component structure. The standard URI syntax structure separates the components using the slash ("/"), question mark ("?"), and number sign ("#") [47]. The term in the URI describing the predicate must be the last component in the URI. Hence, regular expressions[2] can be used to pattern-match and find the predicate term in any URI. Once the predicate term(s) have been retrieved from the URI, comparing the semantic similarity between the predicate term(s) and a list of available issue-causing terms will help understand the predicate efficiently. The results of the semantic similarity check are then pushed onto the ethics ontology.

Example: As depicted in figure 3.3, consider a situation where a custom ontology is used to model the information: Online user *"youngSinatra85"* has full name *"Sir Robert Bryson Hall II"*. Let the predicate be `"lu:hasFullName"`, then the expansion of this predicate would include the whole URI : "`http://liveuser.org/ontology/lu/hasFullName`". Now, the last component, which is "hasFullName", is retrieved using regular expressions. The string "hasFullName" in itself does not mean anything and would make sense only if the camelCase[3] was split, so another regular-expressions-based pattern-matching system is set up to make sense of such terms. This would render three separate terms *"has"*, *"Full"* and *"Name"* which is closer to regular English syntax, and can be dealt with Natural Language Processing (NLP) techniques to identify that some online user's full name is being exposed. Thus, even if the ontology is not standard, the system will be able to identify such ethical concerns with

[2] `https://en.wikipedia.org/wiki/Regular_expression`
[3] `https://en.wikipedia.org/wiki/Camel_case`

confidence.

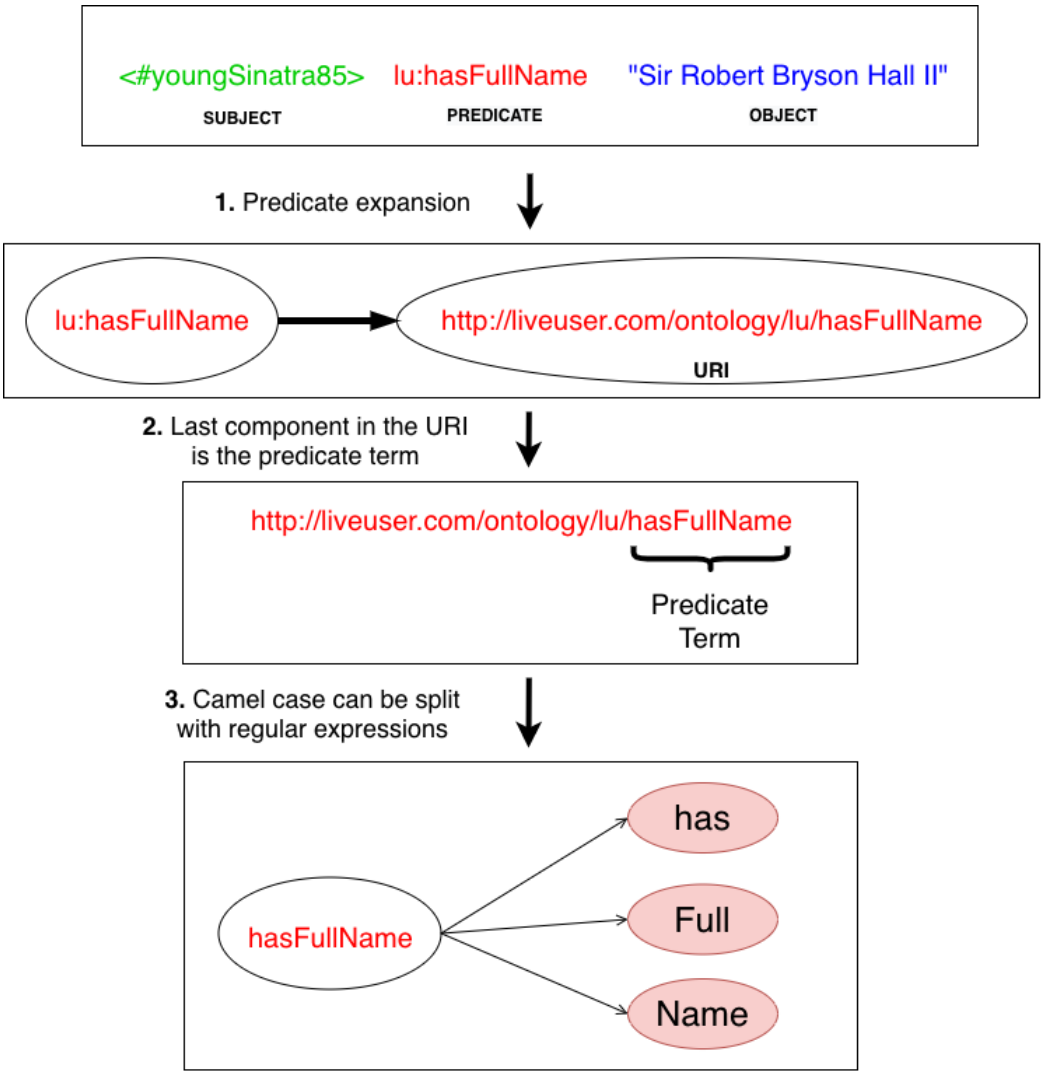

Figure 3.3: An example showcasing the technique used to understand predicates

3.5.3 Data Integration Risks

After the previous step, STEDI will have a deeper understanding of the kind of data present in the datasets, and also of the ethical concerns that are present in the said dataset. The ethics ontology is queried to identify specific patterns in the type of data the dataset possesses, and this will help to predict if an ethical issue will arise after

data integration. A simple pattern to check for would be the presence of ethnic data and crime rates; there would be critical ethical issues in an algorithm linking those two pieces of information together and making predictions on future crimes based on a person's ethnicity. Thus, if a similar pattern is noticed, details regarding the issue and the issue-causing area are sent to the reporting service. A total of four such complex, data-integration-related ethics issues can be identified using STEDI; the details of which will be presented in the next chapter.

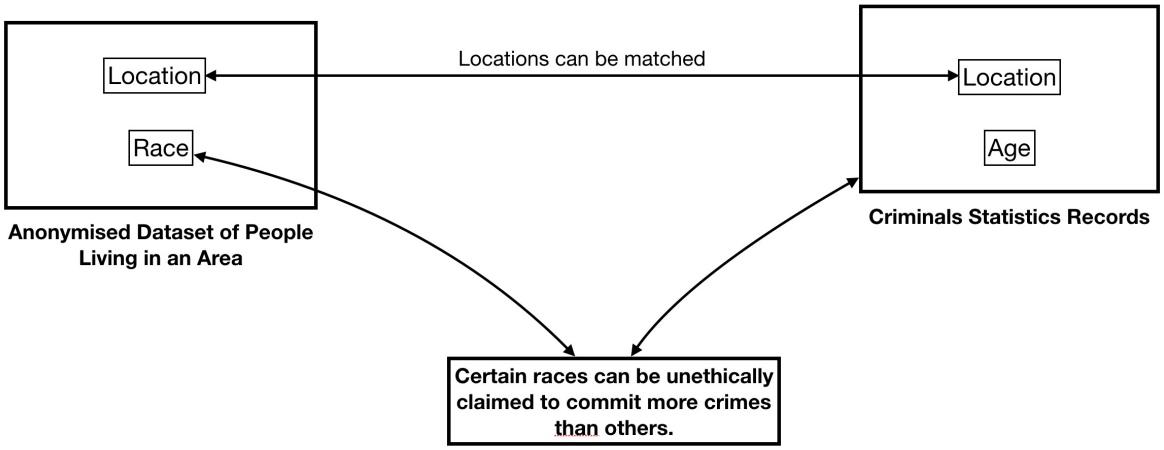

Figure 3.4: An example integration-related ethics issue that STEDI can handle

3.6 Ontology Management Service

The ontology management service is involved in importing and managing the specially-designed ethics ontology. When STEDI is launched, the ethics ontology is loaded into the memory. Following that, whenever new ethical concerns are detected in the datasets, the ontology management service makes sure the issues are correctly represented in the ethics ontology. Finally, when the reporting service is generating the ethics report, the ontology management service returns the problems detected in the relevant dataset.

3.7 Reporting Service

The reporting service only deals with generating the ethics report. Its functions are:

- Asking the user for the correct directory in which to save the report
- Liaising with the ontology management service and querying the ethics ontology
- Generating a detailed text report of all the issues and the entities that caused the issues

3.8 Tool Operation

This section explains in detail how the entire tool was designed to operate:

1. The user will have to start STEDI up and then select the input datasets.
2. A questionnaire will appear for every dataset that is loaded into the system.
3. The user will have to accurately answer the questionnaires to ensure the proper functioning of the system. If the system detects that there are no answers, or if the answers are inappropriate, it will display a relevant error message to let the user know that they need to modify their responses.
4. If the questionnaire was filled-in successfully, then STEDI starts to process the datasets. It does the following in a loop until all the datasets have been processed:
    - Retrieves the vocabularies used in the dataset and identifies the various domains that can be modelled with those vocabularies.
    - Iterates through every triple in the dataset and expands every predicate to obtain its URI. Using regular expressions[4] to match specific patterns in the URI, STEDI will be able to get the predicate's name.
    - STEDI will be able to gain some knowledge about what the predicates mean by using word vectors and semantic similarities. Using this knowledge, STEDI will be able to classify if the predicate represents a potentially unsafe datapoint or not.

[4] `https://en.wikipedia.org/wiki/Camel_case`

- The knowledge gained about the dataset from the previous step is pushed into the ethics ontology for future reference and reporting purposes.

5. After all the datasets have been processed, STEDI moves on to check the datasets for complex ethical issues that might arise due to data integration. It has its own knowledge of a limited set of complex ethical issues that occur due to data integration. Hence, it uses the complex ethical issues knowledge in conjunction with the knowledge it gains from querying the ethics ontology to check if any complex integration-related issues might arise.

6. Based on the results from the previous step and by querying the ethics ontology, a text report detailing all of the ethical issues that were identified or predicted will be generated in a file location that the user chooses.

Chapter 4

Implementation

This chapter introduces the intricate implementation details of STEDI. An essential part of STEDI is the specially-developed ethics ontology; thus, the details as to how the ontology was developed and the competency questions it aims to answer will be detailed in this chapter. The programming infrastructure, mainly developed with Python 3 and other external libraries, will be detailed in the coming sections. Lastly, the security considerations and the implementation challenges will be discussed towards the end of this chapter.

4.1 Competency Questions

A crucial step in developing an ontology is listing out the competency questions that the ontology should be able to answer [48, 49]; this gives the designer a clear idea of the scope of the ontology and helps them stay on track while developing the ontology [50, 49]. Some of the competency questions were inspired by the concepts modelled in the Data Privacy Vocabulary (DPV) [4], whereas others were inspired by unethical practices that were identified and reported in the past few years.

4.1.1 Individual Datasets

The following ethical concerns were targeted for individual datasets:

1. **Presence of Personally Identifiable Information (PII)**

Personally Identifiable Information (PII) is any data that can uniquely identify an individual [51]. Some examples of PII are name, biometrics, passport number, home address, and telephone number [52]. Datasets should never contain information that can uniquely identify an individual because that might lead to user-privacy leakage (as explained in Section 2.4).

2. **Neglecting the principle of data minimisation**

   The data collected from a data subject must be just enough to perform the required data analysis task [53, 54]. Organisations tend to collect more data than necessary to cover every edge case, but research shows that it unnecessarily increases the damage done in case of a security threat [53]. According to the GDPR [36, 35], the collected data must be adequate, relevant and limited to what is necessary for the purpose for which it was collected.

3. **Presence of ethically-sensitive data**

   Presence of sensitive data is not an ethical issue. However, if these data points are linked with other appropriate information about an individual, they can reveal a lot of unnecessary details about the individual. The following is a list of sensitive areas that were inspired by the concepts modelled in the Data Privacy Vocabulary (DPV) [4]: age, behaviour, criminal activity, ethnicity, health, income, loan records, location, non-disclosure agreements, physical characteristics, political opinions, religion, and user tracking data.

4. **Presence of attached documents that might contain PII**

   Documents like résumés, scanned medical prescriptions and photographs are commonly attached to datasets. These documents contain PII and other sensitive details, and thus precaution must be taken before such datasets are integrated.

5. **The data subject is a child**

   Regulations like Children's Online Privacy Protection Act (COPPA) [55] make sure that children's data is managed very safely. It is well-known that a child's data is considered to be more sensitive than an adult's data. Hence, it is justified to look for children-related data in datasets before integrating them.

6. **Proper consent was not given to the data controller**

   Consent is the basis of ethical data handling. No data controller can use any data without the data subject's explicit consent. Consent is also a significant part of the GDPR, and as discussed in Section 2.8.2, there exists an ontology called GConsent [38] to model the data subject's consent and other related information. Since GDPR has a well-structured policy for working with consent, STEDI also follows the same guidelines to make sure appropriate consent is given to the data controller.

4.1.2 Data Integration Scenarios

This dissertation aims to show with the help of a prototype tool that the materialisation of ethical issues can be predicted before datasets are integrated. However, it is almost impossible at this time to be able to predict all kinds of ethical issues. Hence, this tool focuses only on the following scenarios wherein data integration will cause some ethical issues:

- **Scenario - 1: Unethical linking of criminal activities and ethnicity data**

  As discussed in Section 1.1, the COMPAS recidivism algorithm [10] was biased towards black people as it usually predicted them to be at a higher risk of re-offending than they actually were. This scenario was included to combat such unethical linking of criminal activities to the data subject's ethnic background.

- **Scenario - 2: Unfair use of a person's online behaviour to decide if financial institutions should grant them a loan.**

  Companies like CASHe and Lenddo track everything a user does online before granting them a loan [56, 57]. For youngsters with low credit scores, this seems like an excellent way to get a loan, but it is not an ethical way to conduct business for a financial institution. The potential borrowers are meticulously profiled by going through their social media posts, online purchases, mobile phone usage, and allegedly even how quickly they respond to texts [56, 57, 58].

- **Scenario - 3: Unethically using a person's social media activities and online behaviour to increase medical insurance rates**

The Health Service Executive (HSE)[1] provides public health and social care services for the people living in the Republic of Ireland [59]. In March of 2019, it was reported that the HSE's website was filled with third-party ad trackers that were leaking sensitive information about the users of the website [60]. This included information on people looking for advice on unplanned pregnancy and methods to tackle HIV. Such sensitive details about the users of a public health website should never be accessible to anyone.

If the HSE has embedded ad trackers, it is highly possible that medical insurance providers also have such trackers embedded on their websites. A hypothetical example would be a medical insurance provider gaining access to a user's online behaviour and social media activities; in that case, the company can unethically regulate the prices depending on the user's medical conditions and other online behaviour. It can also study the user's social media posts and gain an understanding of the user's regular activities *(e.g., adventure sports)*. With that knowledge, they can make wrongful assumptions about the safety of the user and the people they are connected to, which in turn could lead to hiked insurance rates for the user and their connections.

- **Scenario - 4: Integrating datasets to create a tailored reality for users, especially with regards to their political opinions.**

  Eli Parser originally coined the term "filter bubble" which is another name for a tailored online reality where everything is customised to the user's preferences. Technology has advanced so much that it can create a fully-customised universe of information for every user; though it sounds pleasant, it can actually alter the basis of how humans typically encounter new ideas and information [61]. Users unconsciously start believing that everyone thinks like them; which can in turn render them oblivious to others' thoughts and perspectives.

  A well-known example of a harmful filter bubble is the one that occurred during the US presidential elections of 2016 [62, 63]. Most people who were sure that he would lose the elections to Hillary Clinton were oblivious of the popularity

[1] `https://www.hse.ie/eng/`

Trump had gained in the United States of America. This was due to the news sources providing them with the information they wanted to hear - that Trump would lose.

4.1.3 Final List Of Competency Questions

1. Are the principles of data minimisation followed in the dataset?

2. Does the dataset contain any Personally Identifiable Information (PII)?

3. Does the dataset collect data that is associated with the following areas - *age, behaviour, criminal activity, ethnicity, health, income, loan records, location, non-disclosure agreements, physical characteristics, political opinions, religion, and user tracking data.*

4. Does the dataset have any attached document(s) that may contain PII?

5. Does the dataset have any children-related data?

6. Did the data subject give explicit consent to the data controller for handling their data?

7. Can the datasets be integrated in a way that allows for unethical assumptions to arise about certain ethnicities being more inclined to criminal behaviour?

8. Can the datasets be combined in a way that will enable financial institutions to unfairly decide who gets a loan based on their online behaviour and interests?

9. Can the datasets be integrated in a way that allows insurance companies to unscrupulously hike medical insurance rates based on one's social media activities or online behaviour?

10. Can the datasets be combined in a way that allows for the creation of a tailored reality, especially with regards to political opinions?

4.2 Ethics Ontology

An ethics ontology was created to specifically answer the ten competency questions that were introduced in the previous section. The ontology development was guided by the paper titled *"Ontology Development 101: A Guide to Creating Your First Ontology"* by Natalya F. Noy and Deborah L. McGuinness [49]. The steps followed were:

1. **Determining the domain and scope of the ontology:** it immediately appeared that the ontology would be the backbone of STEDI. Its primary function was to store details regarding the ethical views of the datasets. This stored information would then be used to predict data integration issues and to generate an ethics report about the datasets. Defining the competency questions at an early stage in the design process helped ensure that the ontology could hold enough information to actually answer those questions.

2. **Considering existing ontologies:** other existing ontologies were studied, both for their ontologies and also the concepts they presented. Data Privacy Vocabulary (DPV) [4] and GConsent [38] displayed values that aligned closely with the values of this dissertation. The modelling of user consent was heavily influenced by the design strategies made for GConsent. Some of the object properties even reused concepts from the GConsent ontology.

3. **Listing all important terms:** the exact concepts and the relationships between them were identified to clearly express the knowledge required to answer the competency questions. It was an iterative process, and the ontology was changed multiple times throughout the research period to better fit the information it had to represent.

4. **Defining the classes and their hierarchy:** a "top-down development process" [49] was followed to define the classes and their hierarchy. The base classes identified were: Consent, Data Controller, Data Subject and Status. Figure 4.1 visualises the relationships between all the base classes.

   The Consent class is reused from the GConsent ontology [38] to model the user's consent. The Consent class does not include all the concepts that GConsent models, instead only a few of them are used. However, a lot more of GConsent's

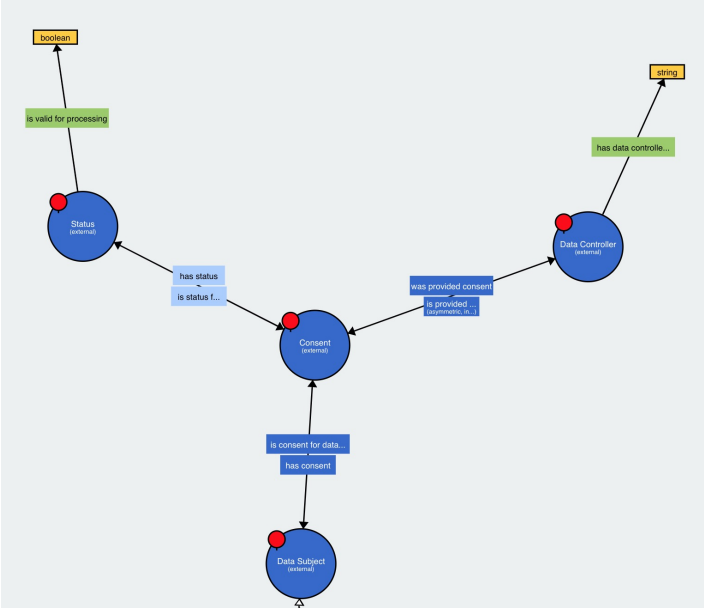

Figure 4.1: Graphical visualisation of all the base classes

concepts can be added in the future to enable a more detailed consent-modelling strategy. The Consent class has relations with all the other three base classes: Data Controller, Data Subject and Status.

The Data Controller class is reused from the DPV ontology [4], and it represents the current organisation/team/person that is handling the dataset. In this ontology, the Data Controller class only has one data property, and that is its name.

The Data Subject class is reused from the DPV ontology [4], and it is used to model the primary data subject of the dataset. It has two subclasses: Group and Individual. Since the subclasses will be able to model the information regarding the data subjects better, it does not have any natural properties to itself.

The Status class is reused from the GConsent ontology [38], and it is used to model the current status of the consent. The status can either be "Valid for processing" or "Not valid for processing". Hence, this class has been modified to have only one boolean data property: "`isValidForProcessing`", which can only hold the values "`True`" or "`False`".

5. **Defining the properties:** The internal structure of the concepts are modelled with the use of properties. All of the relevant properties and their types (data-property or object-property) were defined clearly. The properties were based on the competency questions defined earlier in Section 4.1.3.

6. **Defining the facets of the properties:** The properties have many features (also known as facets) such as cardinality, value type, domain and range. These facets determine how many different values they can store, what the data type of the value will be, the domain of the property, and the range of the property. All the data properties, except "`hasDataControllerName`", are boolean data properties. The "`hasDataControllerName`" stores a string value of the data controller's name.

   The boolean data properties can only store "`True`" or "`False`", and the default value for every data property is "`False`". Therefore, only if an issue is identified, the relevant value will change to "`True`".

7. **Defining the instances:** Each dataset will have an instance for itself in the ethics ontology. This way, multiple datasets can be analysed, and their information will be stored safely in the same ethics ontology. The instances will be named after the dataset's name. Once the datasets have been processed, the ethics ontology will contain an instance for every dataset. The instances will have a boolean value (`True` or `False`) for every issue-related property, and a string (containing the name of the data controller who was initially given consent to process the data) for the "`hasDataControllerName`" property. See Appendix A for a detailed report of all the properties in the ethics ontology.

4.2.1 Ontology Development Infrastructure

- The ethics ontology was built as an OWL2 ontology using the popular, open-source tool Protégé[2] [64].

- The in-built HermiT[3] reasoner in Protégé was used to identify logical inconsistencies in the ethics ontology.

[2] `https://protege.stanford.edu/`
[3] `http://www.hermit-reasoner.com/`

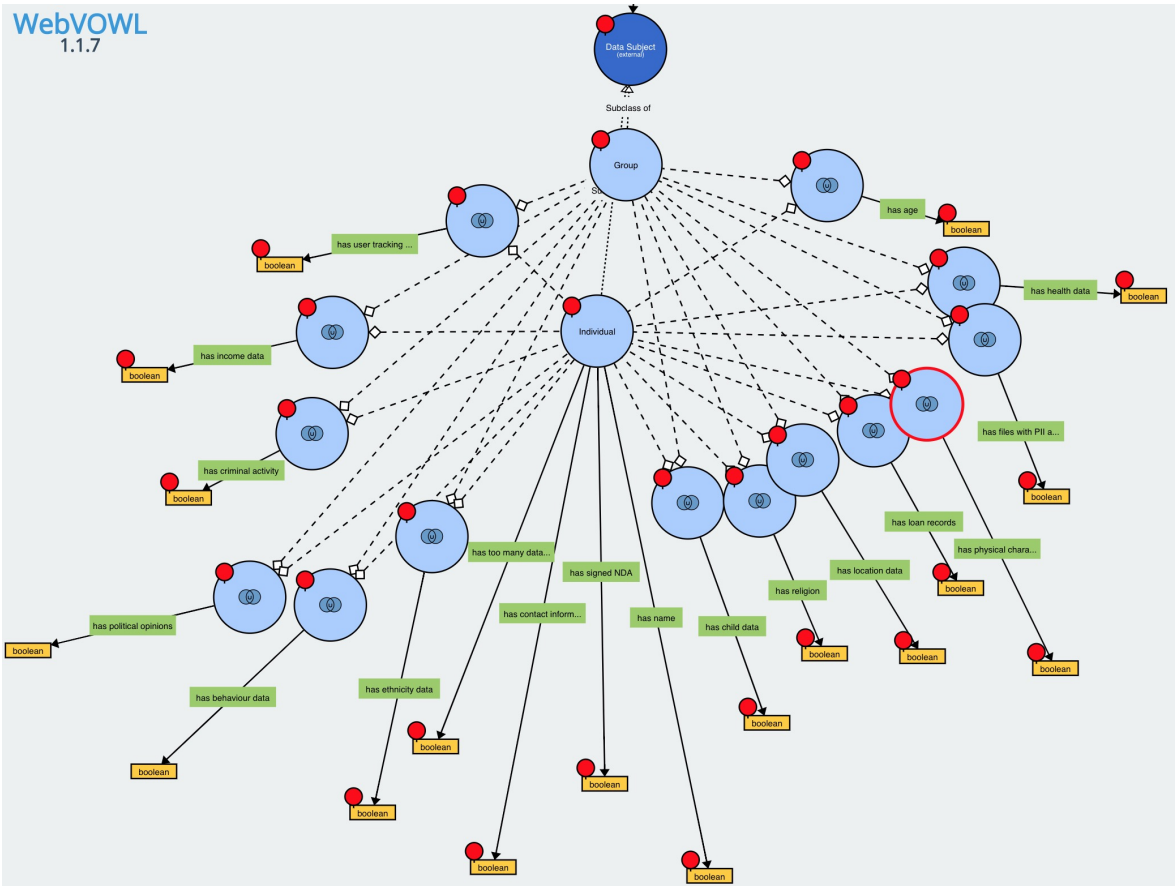

Figure 4.2: Graphical visualisation of all the data properties

- The ethics ontology was also checked for common pitfalls and bad design patterns using OOPS! (OntOlogy Pitfall Scanner!)[4].
- Sufficient metadata is provided for the ethics-ontology in the form of detailed comments, labels and other annotations.

4.3 Programming Infrastructure

4.3.1 Python 3

Version: 3.8.5

[4] `http://oops.linkeddata.es/`

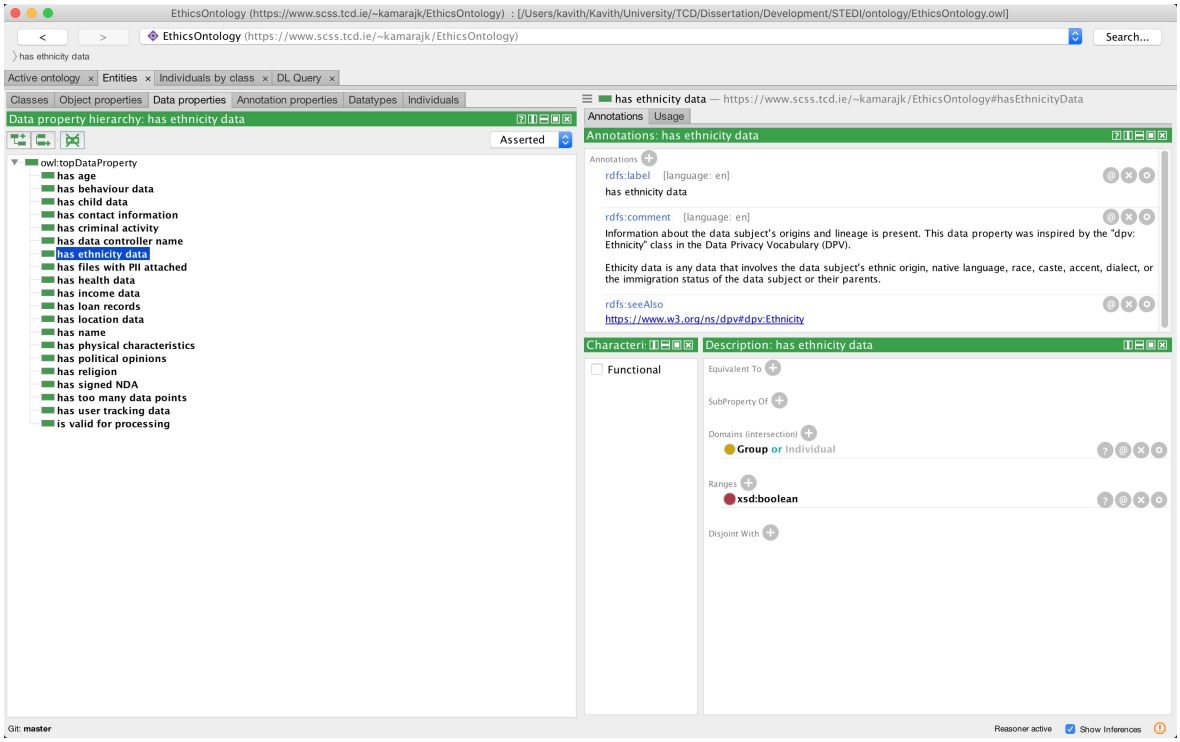

Figure 4.3: Screenshot of the Protégé ontology editor

The programming language used throughout the development of STEDI was Python 3. This programming language was chosen for its simplicity and its broad applicability.

4.3.2 RDFLib

Version: 5.0.0

RDFLib is a popular Python package for working with RDF-based datasets[5]. It includes built-in parsers, serialisers, querying interfaces and many other utility functions. STEDI uses the RDFLib package to work with the input linked-data datasets and the ethics ontology. RDFLib makes it very easy to parse linked-data datasets and query them.

[5] `https://rdflib.readthedocs.io/en/stable/`

4.3.3 LOV API

The Linked Open Vocabularies (LOV) Application Programming Interface (API)[6], is an open-source online repository that contains a lot of information regarding most standard linked-data vocabularies [39]. The LOV API is used by STEDI to understand what information is being presented by the dataset.

4.3.4 spaCy

Version: 2.3.2

spaCy is a performance-oriented Natural Language Processing (NLP) tool[7]. The spaCy Python package is used by STEDI to analyse predicates and understand the type of data being stored. The semantic similarity function is used together with a spaCy English-language model to gain an understanding of what the predicate terms mean.

4.3.5 tkinter

Python's standard GUI toolkit is called tkinter, and it comes packaged with the official Python distribution[8]. The front-end for STEDI was entirely built with tkinter due to its simplicity and ease of use. The biggest downside of tkinter is that the programs developed with tkinter tend to look old-fashioned. However, as mentioned in Section 3.3, the look and feel of the tool is outside the scope of this dissertation. Thus, tkinter was the perfect tool to use for this project.

4.4 Front-End

Section 3.2 introduced the five components of STEDI - the first two components being User Interface and Loading Service. Though they perform two different functions, they operate on the same layer, that is the front-end. Hence, both the User Interface and the Loading Service components are built into the front-end of the tool. As depicted

[6] `https://lov.linkeddata.es/dataset/lov/api`
[7] `https://spacy.io/`
[8] `https://docs.python.org/3/library/tkinter.html`

in figure 4.4, STEDI's front-end has satisfied all the requirements of the User Interface and Loading Service components (as discussed in Section 3.3 and Section 3.4):

- It successfully loaded the datasets.
- It displayed the questionnaire.
- It liaised with all of the other components to successfully process the datasets.
- It provided appropriate feedback to the user by mentioning the chosen datasets, displaying a progress bar, giving a short description of the current process and finally, using a message box to notify the user that the process has finished running.
- Finally, the UI is very responsive and requires minimal action on the user's part. It is designed in such a way that it is easy to use even for first-time users.

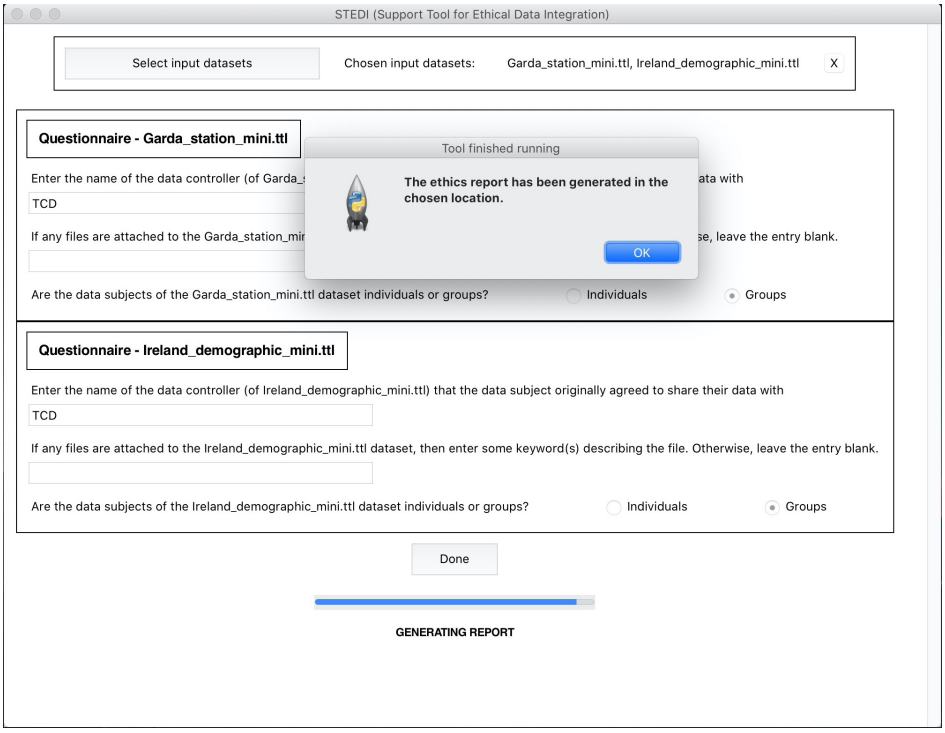

Figure 4.4: Screenshot of STEDI's front-end

4.5 Back-end

The back-end of this tool consists of the other three components - Risk Identification Service, Ontology Management Service and Reporting Service. These services work in tandem to identify a wide range of ethical issues in the datasets and report them to the user. The execution strategy to identify these issues is to use a phased approach. Figure 4.5 showcases the architecture diagram for this phased approach.

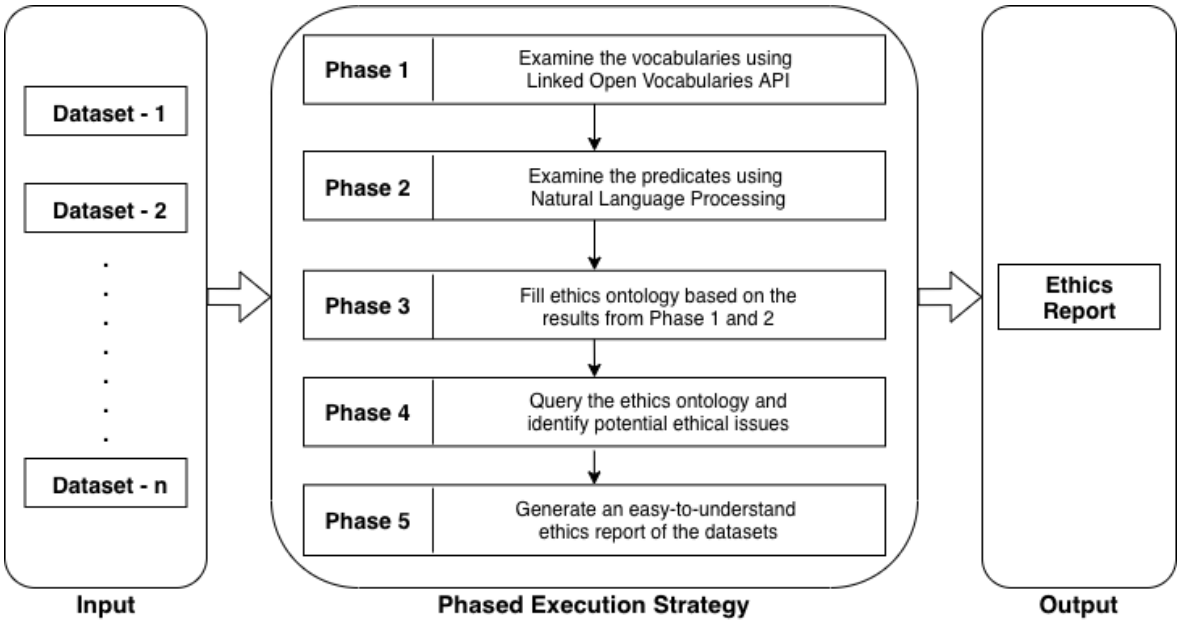

Figure 4.5: Architecture diagram for STEDI's back-end

4.5.1 Phase 1: Examine Vocabularies

The first phase was inspired by the approach taken by G-DIEP [5] (explained in Section 2.9.1). The vocabularies of the dataset are retrieved and examined using the Linked Open Vocabularies API[9] [39]. An HTTP GET request is made to the API containing the details regarding the vocabularies present in the dataset, and the response is in JSON. In this response, there is an attribute called "tags" which contains information regarding the areas that the vocabulary focuses on modelling. Therefore, retrieving

[9] `https://lov.linkeddata.es/dataset/lov/api`

the value of the tags attribute will reveal a lot of data about the dataset. Figure 4.6 showcases the tags attribute as seen in a JSON response about the "geo" vocabulary[10].

```
"tags": [
  "Geography"
],
```

Figure 4.6: The tags attribute for the "geo" vocabulary

After retrieving the tags, STEDI compares them with a built-in list of tags to identify if the vocabulary's domain is ethically-sensitive or not. Ethically-sensitive domains do not always represent a risk, which simply means that the data might be sensitive, and hence a lot of care needs to be taken while using the dataset. Table 4.1 indicates the list of sensitive tags and the matching data-property it sets to "`True`" in the ethics ontology.

Sensitive tags	**Data-property triggered**
Geography	`hasLocationData`
Society	`hasEthnicityData`
Health	`hasHealthData`
Biology	`hasHealthData`
Government	`hasPoliticalOpinions`

Table 4.1: STEDI's list of sensitive tags and their matching data-properties

4.5.2 Phase 2: Examine Predicates

The second phase is focussed on extracting the predicate and analysing it for ethical issues. As explained in Section 3.5.2, the system starts off by iterating through every triple in the dataset and retrieves all the predicates. When the predicate is expanded into its full URI, the last component of the URI will represent the predicate term. A regular-expression-based pattern-matcher is used to extract the last component from the URI. It is common practice to have predicate names in camelCase; thus, the next step would be to use another regular-expression-based pattern-matcher to split the camelCase and include spaces between the terms. Once this is done, the result will

[10] `https://www.w3.org/2003/01/geo/`

be a string of predicate tokens that can be utilised by NLP techniques to understand what the terms mean.

```
predicate_parts = p.split("/")
predicate = predicate_parts[-1]
predicate = re.sub("[^A-Za-z0-9 ]+", " ", predicate)
# To split camelCase
predicate = re.sub(r"([A-Z])", r" \1", predicate)
predicate = predicate.lower()
```

Figure 4.7: The regular expression code written to extract the predicate terms

To comprehend the predicates, there exists an extensive list of ethically-sensitive words that are modelled as a Bag-of-Words (BoW). The BoW is stored as part of a massive dictionary inside STEDI. The semantic similarity function in spaCy is used to check if any of the tokens in the predicate string are related to any of the words in the BoW. If two terms seem to have a high similarity, then the appropriate data property in the ethics ontology is set to "`True`". If there is not enough similarity between the terms, then the data property is left in its default value which is "`False`". Table 4.2 contains the exact BoW that indicates the presence of an ethical concern, and the data-property it modifies in the ethics ontology.

Appendix B.1 contains a code snippet that shows how the predicate processor method was implemented in Python; it first uses regular expressions to retrieve the predicate tokens and then uses spaCy's similarity function to identify semantically similar predicates. Based on the results of the similarity check, the applicable data property's value is set to "`True`".

4.5.3 Phase 3: Fill Ethics Ontology

After the ethical concerns in the dataset have been identified by Phase 1 and Phase 2, the results need to be stored in the ethics ontology. The results are not pushed into the ethics ontology as and when an ethical concern is detected. Instead, the results of the analysis are temporarily stored in a Python dictionary until the dataset has gone through the first two phases. Once the dataset has been thoroughly processed, the third phase begins by pushing the results of the analysis (that is, the values stored in

Data property	Bag-of-Words
hasAge	"age", "birthday", "dob"
hasBehaviourData	"behaviour", "interest", "myers", "opinion", "personality"
hasChildData	"baby", "child", "juvenile", "kid", "minor", "teenager", "young"
hasContactInformation	"account", "contact", "email", "phone", "skype"
hasCriminalActivity	"criminal", "felony", "jail", "prison"
hasEthnicityData	"accent", "community", "dialect", "immigrant", "language", "race", "religion"
hasHealthData	"DNA", "consult", "doctor", "health", "medical"
hasIncomeData	"earning", "income", "salary"
hasLoanRecords	"debt", "loan"
hasLocationData	"address", "area", "city", "county", "lives", "location", "postal", "resident", "station", "zip-code", "x", "y"
hasName	"name"
hasPhysicalCharacteristics	"body", "colour", "disability", "gender", "hair", "height", "piercings", "size", "skin", "tattoos", "weight"
hasPoliticalOpinions	"politics"
hasReligion	"divinity", "faith", "religion", "worship"
hasSignedNDA	"NDA", "non-disclosure", "nondisclosure", "confidential", "secrecy", "agreement"
hasUserTrackingData	"advertisement", "browser", "history", "search", "tracking"

Table 4.2: Ethics ontology's data properties and the matching BoW

the dictionary) into the ethics ontology.

The "saving" of values in the ethics ontology is performed in the following manner (see Appendix B.2 for the Python implementation):

1. Create a new individual for the ethics ontology with the same name as that of the dataset being processed.

2. Identify if the data subject is an individual or a group, and update the ethics ontology accordingly.

3. With the individual as the subject, predicates are created for every ethical issue,

and the object becomes a boolean value stating if the ethical concern is present (`True`) or not present (`False`).

4. Finally, information regarding which vocabulary or predicate triggered the issue is also stored in the ethics ontology.

4.5.4 Phase 4: Identify Integration Issues

Specific data-patterns identifying the four scenarios from Section 4.1.2 are stored in STEDI as existing knowledge. Once a similar pattern is noticed in the input datasets, the scenario is marked as present in the datasets. The information regarding the scenarios is also temporarily stored in Python dictionaries (as shown in Figure 4.8). The code snippets of the method that is used for identifying the integration issues are presented in Appendix B.3.

```
self.scenario_1_issues = {
    "hasCriminalActivity":  False,
    "hasEthnicityData": False,
    "hasLocationData": False,
    "hasReligion": False
}

self.scenario_2_issues = {
    "hasBehaviourData": False,
    "hasLoanRecords": False,
    "hasUserTrackingData": False
}
```

Figure 4.8: Code snippet showing how the scenario's data-patterns are stored

4.5.5 Phase 5: Generate report

The final phase is focused on reporting the detected ethical issues. The generated ethics report is a text file that clearly states all the ethical concerns in the datasets and the predicted ethical issues that might arise from data integration. STEDI adds the following information in the report:

- The name of the data controller with whom the data subjects initially consented to share their data.
- If the dataset is valid for processing or not; this is identified by comparing the name of the approved data controller with the name of the current organisation handling this dataset. If they are the same, then the dataset is marked as valid for processing.
- The data subject's type (Individual or Group).
- The identified ethical concerns and the predicate or vocabulary that caused the issue.
- A detailed explanation of the data-integration issues that were identified.

4.6 Application Flow

In this section, a typical workflow of STEDI from start to end is illustrated. See Appendix - C.1 for screenshots.

1. As soon as the tool is launched, a small dialogue box pops up asking for the name of the current organisation that is using this tool. The application will only proceed to the next step if the organisation name is entered.
2. Clicking on "*Select input datasets*" makes a file dialogue appear, through which the user can select the input datasets.
3. Once the datasets are selected, the dataset-questionnaire is displayed in this instance. After answering the questionnaire correctly, the "*Done*" button needs to be pressed for the dataset analysis to start.
4. STEDI starts to analyse the datasets by going through all the phases that were discussed in Section 4.5.
5. It provides ample feedback to the user by displaying a progress bar and a status widget. The status widget briefly describes the current process running in the background.

6. A message box appears on the screen to notify the user that the datasets have been processed.

7. Once the "*Ok*" button is clicked, a file dialogue opens to allow the user to choose a location in which to save the ethics report. Once a directory is selected, STEDI proceeds to save the ethics report in that location.

8. Finally, after saving the ethics report at the chosen location, another message box pops up to let the user know that the ethics report has been generated.

4.7 Security Considerations

4.7.1 Incorrect Risk Identification

STEDI cannot afford to be negligent when it comes to identifying ethical issues because it operates in a critical environment; a wrong prediction might cause tremendous harm [10, 12, 14]. Therefore, the tool has been designed in such a way that it would permit a lot more false-positives than false-negatives. The barrier set for a vocabulary or predicate to classify as an ethical concern is shallow. Furthermore, the predicate tokens do not have to be an exact match or a synonym of the in-built BoW; Instead, they simply have to be somewhat related to one of the terms in the BoW to be classified as an ethical concern. This ensures that even if the tool made a mistake, this mistake would be a precautionary false-positive issue rather than a missed ethical issue.

4.7.2 Programming Infrastructure

STEDI uses some third-party Python libraries, and these libraries must not have any significant security issues. Hence, only well-reputed, standard third-party dependencies were used to develop STEDI. It uses only the latest, fully-patched versions of these dependencies.

4.7.3 Human Error

Sufficient precaution has been taken while developing the tool to account for common human errors. The tool has gone through rigorous tests of irregular human behaviour

to make sure it does not fail when a user does something out of the ordinary. All corner-case errors are caught and handled efficiently. The following precautions have been taken to handle human errors (See Appendix-C.2 for screenshots):

- The user will not be allowed to circumvent the first dialogue box. They need to enter the name of their current organisation to proceed with using the tool.
- If the user wants to clear their current selection of datasets, they are free to do so by clicking on the "X" button.
- The name of the data controller needs to be given in the questionnaire. Otherwise, a message box asking the user to input the name will pop up.
- The type of data subject (Individual or Group) needs to be selected. Otherwise, a message box will pop up asking the user to select one.
- Once the tool is done processing the dataset, if the user clicks on the "cancel" button in the file dialogue, the tool will loop back and ask the user to choose a location to save the ethics report.

Apart from the above anticipated human errors, it is of utmost importance that the data integrator is honest and knowledgeable in the areas of privacy and ethics. It is easy for a data integrator to run STEDI, identify the areas triggering an ethics issue, and modify them so that STEDI's checks are cleared. Or, they could also just ignore the identified issues and proceed to integrate the datasets. The only solution to this problem is that organisations need to be aware of such issues and take the precautionary steps to hire well-trained and ethical employees.

4.8 Implementation Challenges & How They Were Addressed

4.8.1 Tool Platform

Deciding whether the tool had to be a desktop application or a web application was one of the biggest challenges while developing this tool. The original idea was to

make the tool a desktop application as it reduced the cyberattack surface. However, towards the end of the development period, the web-application route seemed to be more favourable owing to faster processing and up-to-date risk identification services. Since not all computers are created equal, the processing times tend to vary a lot between machine to machine, and this is primarily due to the NLP technique used in identifying ethical issues in predicates. Hence, a server-side analysis of datasets might provide faster results and more importantly, a more consistent processing time. This dissertation implemented STEDI as a desktop application due to a lack of time, but future work could include making STEDI a web application.

4.8.2 Data Minimisation

The principle of data minimisation states that only the required amount of data must be collected [53, 54]. However, one question it does not answer is "*How much data is too much?*". This is purposefully left out as a single number cannot be applied for all data collection requirements. Hence, it is hard to identify data minimisation violations. This problem was temporarily addressed by setting 10 as the number. This number was chosen arbitrarily as it was better to flag datasets that had more than ten data-points about a single individual than not considering this issue at all. It would be a false-positive in most cases, but it is best left for the data integrator to decide if it is an actual ethical concern.

4.8.3 Comprehending Data

Identifying the kind of data present in datasets was a significant issue during the early stages of this dissertation. Though it seems like a simple matter to humans, it is tough to make a software program understand what a certain kind of data looks like. Similar issues are being dealt within the area of text analytics, and NLP techniques seem to be very successful in solving such problems [65, 66]. Hence, NLP was considered as a viable option to make sense of the predicates found in linked-data datasets.

Chapter 5

Evaluation

This chapter describes the input datasets that were used during the testing and evaluation of this tool, the overall workflow of the application, and the ethics report that is generated. In addition, this chapter describes the approach taken to test the viability of STEDI through user-evaluation. For reproducibility purposes, all the input datasets[1] used in this evaluation and STEDI[2] itself are available online for download.

5.1 Input Datasets

Two groups of input datasets were used to test STEDI's abilities. One of the groups have two open datasets with real-life data, and the other group contains two synthetic datasets that were specifically designed to trigger certain ethical issues.

5.1.1 Real-Life Datasets

- **Dataset - 1: Ireland's demographic data**

  This dataset contains demographic data regarding the people who lived in Ireland in 2011 [67]. It contains a wide range of information such as place of birth, nationality, ethnicity, religion and languages spoken. Data is present for 18,488 small areas present in the Republic of Ireland. Figure 5.1 shows some of the datapoints present in this dataset.

[1] `https://drive.google.com/drive/folders/1kk-KyGw3I-BDJh8nNQ2BZytm51jezOsN`
[2] `https://github.com/Kavithvajen/STEDI`

Figure 5.1: A few datapoints in the Demographic dataset

- **Dataset - 2: Crimes recorded at Garda stations**

  This dataset contains information regarding the various crimes that were committed throughout the Republic of Ireland from 2003 to 2016 [68]. However, only data pertaining to the year 2011 was considered as it could be directly integrated with the Irish demographic data that was discussed previously. The Garda[3] is the national police service of Ireland, and there are multiple Garda stations present throughout the country. This dataset contains information regarding the location of the Garda station and the number of crimes that were reported in every crime-category over 13 years. Some crime-categories did not exist until 2010, hence why the data up until then is erratic - but from 2010 to 2016 there is a standard set of crime-categories. This was not an issue as the interested period was only the year 2011, and the data was consistent throughout this period. Figure 5.2 shows some of the datapoints present in the dataset.

The problem with these two datasets is that they are presented only in the CSV (Comma-separated values) format, and they have to be uplifted into linked-data. The datasets have been published under licenses that allow the use and modification of the data for non-profit, personal, research and educational purposes [67, 68]. Hence, the datasets were extracted and uplifted to RDF. The uplifting of the dataset was done using Juma - a jigsaw puzzle metaphor-based representation of linked-data mappings[4]

[3] `https://www.garda.ie/en/`

[4] `http://juma.adaptcentre.ie/juma-editor/login`

attempts_or_threats_to_murder_assaults_harassments_and_related_offences
burglary_and_related_offences
controlled_drug_offences
damage_to_property_and_to_the_environment
dangerous_or_negligent_acts
divisions
fraud_deception_and_related_offences
kidnapping_and_related_offences
offences_against_government_justice_procedures_and_organisation_of_crime

Figure 5.2: A few datapoints in the Garda dataset

[69]. Regarding the Garda dataset, only the statistics from 2011 were extracted to ensure that the datasets are accessible.

Both these datasets use a new vocabulary that is unknown to LOV [39] and hence proves to be excellent testing datasets. Phase 1 of the tool will not be able to identify the ethical concerns in either of these datasets, hence it is dependent only on Phase 2 to identify the concerns using predicate analysis. Also, these datasets were specifically chosen as they do have the potential to cause a complex data-integration-based ethical issue, specifically Scenario - 1 (as described in Section 4.1.2) where crime statistics are merged with ethnicity-related data. The locations in both the datasets can be matched as they are all places in the Republic of Ireland, and unfair assumptions can be made regarding the future criminal possibilities of certain ethnicities.

After some initial tests, it was noted that the ethical analysis took too long to be feasible. This was later identified as a cause of these two datasets being too big for personal computers to process. Hence, mini-versions of the datasets were created for testing and evaluating purposes. The mini-versions represented the same data and had all the predicates, but the number of individuals was reduced.

5.1.2 Synthetic Datasets

Apart from the real datasets, two synthetic datasets were also created to see if STEDI was able to identify other integration-issues correctly. The datasets were specially designed to trigger Scenario - 2 (as described in Section 4.1.2), where a user's online behaviour is used by financial institutions to decide if the user can be granted a loan. The fabricated datasets are:

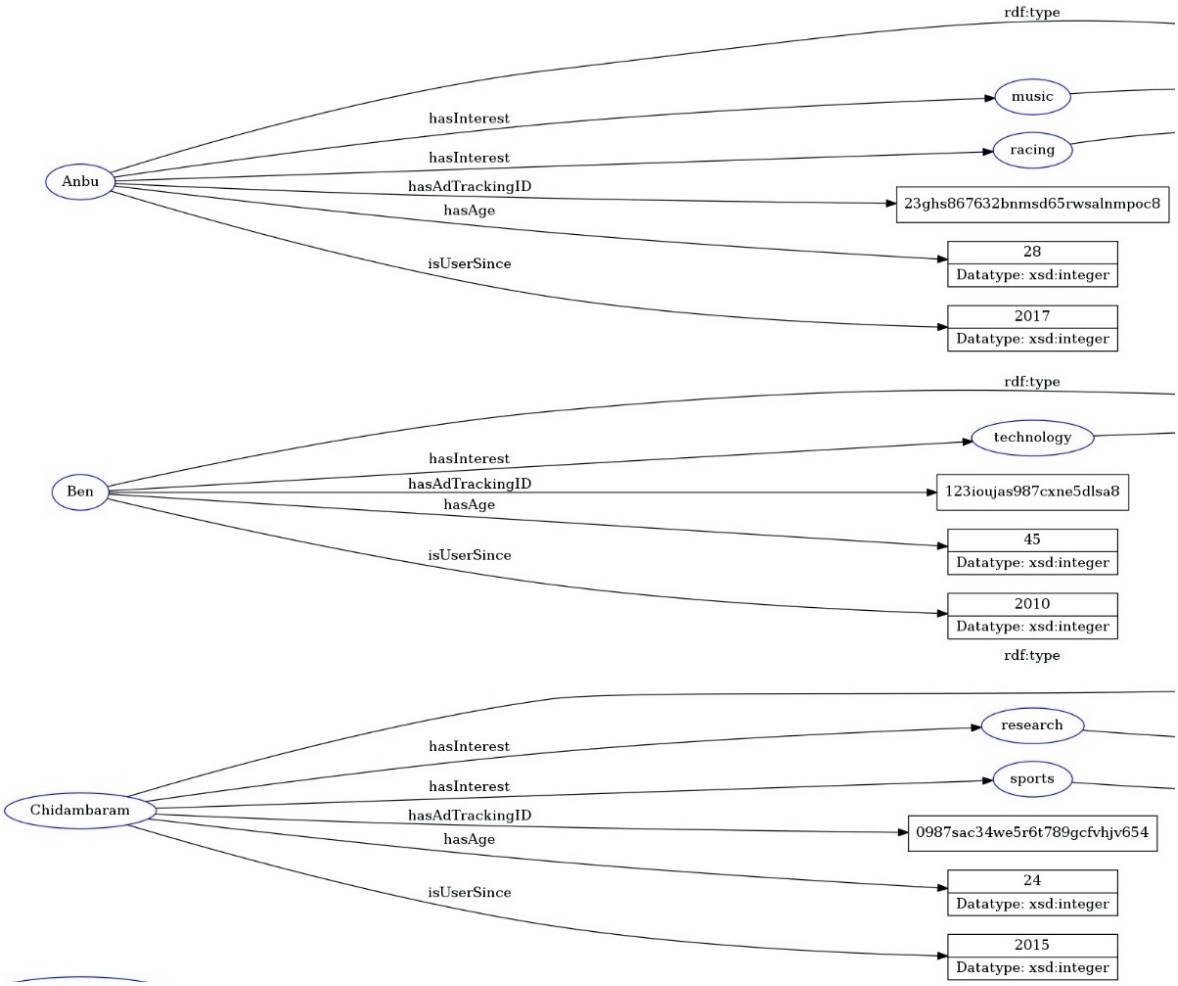

Figure 5.3: Graph representation of the borrower's dataset

- **Dataset - 1: User's online profile**

  This linked-data dataset was created to store information regarding a user's online profile and behaviour. The data stored in this dataset consisted of three individual data subjects, their unique ad tracking ID, their age, date of joining the platform, and their interests (obtained from online browsing behaviour). Figure 5.3 illustrates a part of the graph-representation of this dataset.

- **Dataset - 2: Financial institution data**

  This linked-data dataset was created to store information about the loan status and history of customers belonging to a financial institution. The same indi-

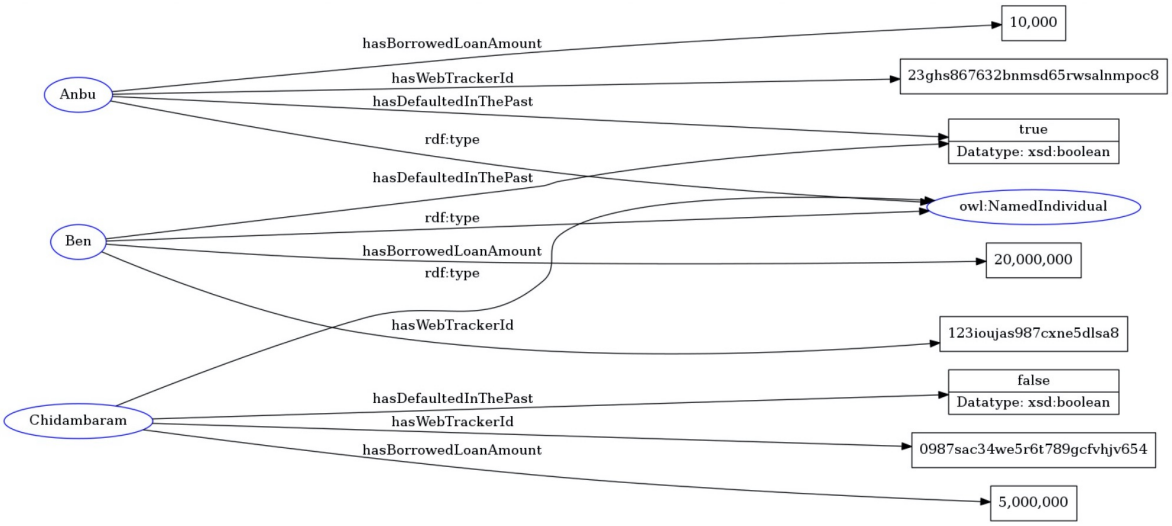

Figure 5.4: Graph representation of the financial institution dataset

viduals from the previous user-dataset are present in this dataset. However, in this dataset, their ad tracking ID, their borrowed loan amount and information on their loan repayment history is stored. The two datasets can be integrated entirely using the matching ad tracking ID. Figure 5.4 shows a part of the graph-representation of this dataset.

5.2 Performance Of STEDI

STEDI was locally tested with each of the two groups of input datasets mentioned in Section 5.1. The ethics reports generated by both the runs are deeply analysed in this section. A point to note is that the integration-scenario results are not considered for this evaluation because they are entirely dependent on the results of Phase 1 & 2 (as discussed in Section 4.5).

5.2.1 Execution - 1

During the first execution, the real-life datasets (discussed in Section 5.1.1) have been used. Since the datasets are publicly-available open-datasets, the answer to the data controller question in the questionnaire can be the same as the current organisation's name. Table 5.1 compares the results of the execution with the ground truth that was

manually identified in the datasets. The full ethics report is presented in Appendix D.1.

Ethical Concern Areas	Demographic Dataset		Garda Dataset	
	Observed	Ground Truth	Observed	Ground Truth
Age	FALSE	FALSE	FALSE	FALSE
Behaviour	FALSE	FALSE	FALSE	FALSE
Child	TRUE	FALSE	FALSE	FALSE
Contact	FALSE	FALSE	FALSE	FALSE
Criminal	FALSE	FALSE	TRUE	TRUE
Ethnicity	TRUE	TRUE	TRUE	FALSE
Files with PII	FALSE	FALSE	FALSE	FALSE
Health	FALSE	FALSE	FALSE	FALSE
Income	FALSE	FALSE	FALSE	FALSE
Loan	FALSE	FALSE	FALSE	FALSE
Location	TRUE	TRUE	TRUE	TRUE
Name	FALSE	FALSE	FALSE	FALSE
Physical Characteristics	FALSE	FALSE	FALSE	FALSE
Politics	TRUE	FALSE	FALSE	FALSE
Religion	TRUE	TRUE	FALSE	FALSE
Non-Disclosure Agreement	FALSE	FALSE	FALSE	FALSE
No data minimisation	FALSE	FALSE	FALSE	FALSE
User Tracking	FALSE	FALSE	FALSE	FALSE
Valid for processing	TRUE	TRUE	TRUE	TRUE

Table 5.1: Execution - 1: Observed values vs. Ground Truth

5.2.2 Execution - 2

During the second execution, the synthetic datasets (discussed in Section 5.1.2) have been used. Since the datasets are synthetic and in order to check if STEDI identifies improper consent, data controller names are purposefully changed. Table 5.2 compares the results of the execution with the ground truth that was manually identified in the datasets. The full ethics report is presented in Appendix D.2.

Ethical Concern Areas	Financial Institution Dataset		Borrower's Dataset	
	Observed	Ground Truth	Observed	Ground Truth
Age	FALSE	FALSE	**TRUE**	**TRUE**
Behaviour	**TRUE**	FALSE	**TRUE**	**TRUE**
Child	FALSE	FALSE	FALSE	FALSE
Contact	FALSE	FALSE	FALSE	FALSE
Criminal	FALSE	FALSE	FALSE	FALSE
Ethnicity	FALSE	FALSE	FALSE	FALSE
Files with PII	FALSE	FALSE	FALSE	FALSE
Health	FALSE	FALSE	FALSE	FALSE
Income	FALSE	FALSE	FALSE	FALSE
Loan	**TRUE**	**TRUE**	**TRUE**	FALSE
Location	FALSE	FALSE	FALSE	FALSE
Name	FALSE	FALSE	FALSE	FALSE
Physical Characteristics	FALSE	FALSE	FALSE	FALSE
Politics	FALSE	FALSE	FALSE	FALSE
Religion	FALSE	FALSE	FALSE	FALSE
Non-Disclosure Agreement	FALSE	FALSE	FALSE	FALSE
No data minimisation	FALSE	FALSE	FALSE	FALSE
User Tracking	**TRUE**	**TRUE**	**TRUE**	**TRUE**
Valid for processing	FALSE	FALSE	FALSE	FALSE

Table 5.2: Execution - 2: Observed values vs. Ground Truth

5.2.3 Metrics

The chosen metrics for this analysis are Accuracy, True Positive Rate, True Negative Rate, Precision and F-Score.

- **Accuracy:** It gives a rough idea of the number of correct predictions STEDI has made.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **True Positive Rate (TPR):** As discussed in Section 4.7.1, the dissertation aims not to have any false-negative predictions. Hence, this metric provides the best understanding of how well the tool performs with respect to not allowing an ethical concern to go unnoticed.

$$TruePositiveRate = \frac{TP}{TP + FN}$$

- **Precision:** Due to the high penalties given to false positives, it gives a good understanding of the ethical issues that are being falsely classified as present in the dataset. Precision is not the primary metric used to evaluate STEDI. However, a very low precision value will classify every datapoint as an ethical concern.

$$Precision = \frac{TP}{TP + FP}$$

- **F-Score:** The F-Score takes both TPR and Precision into account to provide a good understanding of how the model performs in general. False-negatives and false-positives are both weighted higher than the true-negatives and true-positives.

$$F - Score = \frac{2 * TP}{2 * TP + FP + FN}$$

5.2.4 Performance Results

A confusion matrix is built with the total values observed in Section 5.2.1 and Section 5.2.2. Once the confusion matrix is built, the metrics are calculated using their appropriate formulas.

	Observed as an issue	Not observed as an issue
Actually an issue	12	0
Actually not an issue	5	59

Table 5.3: The confusion matrix

METRIC	VALUE
Accuracy	0.93
True Positive Rate	1
Precision	0.7
F-Score	0.83

Table 5.4: The calculated metrics

As seen in the metrics, it is clear that no existing ethics issue went unnoticed (as explained in Section 4.7.1). However, there were some discrepancies with issues that did not exist in the dataset but were flagged by STEDI. Looking deeper into this issue reveals that certain predicates like "birth place" are identified as both a location and a child. This issue is caused by spaCy's semantic similarity function relating "birth" to a child. This cannot be regarded as a mistake because the two words "birth" and "child" are definitely related tokens; just not in this case. Hence further work in the NLP part of this dissertation is required to be able to identify these discrepancies accurately. As was expected, the weakest metric is precision. STEDI is built in a way that it will allow more false-positives than false-negatives.

The dataset analysis process was timed during both the executions. During the first execution, 216 triples were analysed in 220 seconds, and during the second execution,

50 triples were analysed in 66 seconds. That is an average processing speed of 0.87 triples per second, or in other words; it takes approximately 1.165 seconds to process a single triple. These speeds were achieved on a Mid-2015 15" MacBook Pro with a 2.2GHz Quad-Core Processor and 16GB of RAM.

5.3 User Evaluation

This section describes the user evaluation done to evaluate the usability of STEDI.

5.3.1 PSSUQ

The PSSUQ (Post-Study System Usability Questionnaire) was used to measure the users' perceived satisfaction from using the tool. It is a 16-item standardised questionnaire that helps the developers of a system (hardware or software) understand how satisfied their users were by using their system. Each question has seven possible answers ranging from "Strongly Agree" to "Strongly Disagree", out of which the user must select one. A global score will be obtained in the end with three subscales - System Usefulness, Information Quality, and Interface Quality. See Appendix D.3 for the full list of questions that are a part of PSSUQ.

5.3.2 PSSUQ Metrics

The metrics that were considered for this study were all the values that could be extracted from PSSUQ, and they are [1]:

- **Overall:** the standard metric that indicates the overall usability of the tool. It is calculated by taking the average of all responses received.
- **System Usefulness (SysUse):** this metric depicts the usefulness and ease-of-use of the system. It is calculated by averaging the responses for questions 1 to 6.
- **Information Quality (InfoQual):** this metric shows how easy it is for a user to use only the tool and the documentation provided with the tool to recover

from a mistake they made. It is calculated by averaging the responses given for questions 7 to 12.

- **Interface Quality (IntQual):** this metric indicates how good the User Interface of the system is. The metric is calculated by taking the average of the responses received for questions 13 to 15.

5.3.3 Participants

A group of 10 participants who had at least an undergraduate degree in Computer Science were considered for this study. All of them had at least elementary knowledge of the semantic web and how linked-data worked. The participants were initially sent a recruitment message for volunteering, and once they agreed they were sent a link to a webpage containing further instructions on how to proceed with the study.

5.3.4 Tasks

The primary tasks for the participants were:

1. Download STEDI from an online repository [5] and install it on a local machine.
2. Use STEDI until comfortable using it. The two groups of input datasets that were explained in Section 5.1 were also provided. The participants were free to use STEDI with other linked-data datasets that they might have had access to.
3. Fill the PSSUQ form that was hosted on Google Forms [6].

5.3.5 Evaluation Results

Table 5.5 depicts the results derived from the PSSUQ. Lewis and James [1] have done extensive research on the PSSUQ and have concluded on a mean-value for all of the PSSUQ metrics. Systems that have scored lower than the mean-value have performed well, and if the scores are higher than the mean-value, then there is room for improvement on the usability of the system. Table 5.6 shows the average score that was

[5] `https://github.com/Kavithvajen/STEDI`
[6] `https://forms.gle/TuYjfBpFhWiRBMdm9`

achieved by STEDI in each of the PSSUQ metrics and the corresponding mean-value for that metric, as stated by Lewis and James [1].

QUESTIONS	PARTICIPANTS										
	P-1	P-2	P-3	P-4	P-5	P-6	P-7	P-8	P-9	P-10	Average response for question
Q-1	1	1	3	2	4	2	3	5	1	2	2.4
Q-2	1	1	3	1	4	2	4	7	1	3	2.7
Q-3	3	1	2	2	4	3	2	5	1	2	2.5
Q-4	1	1	3	2	4	1	3	6	1	3	2.5
Q-5	1	1	2	6	3	3	3	6	1	1	2.7
Q-6	4	1	2	2	4	2	2	4	2	2	2.5
Q-7	1	1	1	2	4	4	3	7	2	4	2.9
Q-8	1	1	4	2	4	4	4	6	1	3	3
Q-9	1	1	1	1	4	2	3	6	1	2	2.2
Q-10	1	1	2	2	4	2	4	6	1	3	2.6
Q-11	1	1	2	2	4	3	3	6	1	3	2.6
Q-12	1	1	1	1	4	3	4	6	1	3	2.5
Q-13	4	1	2	2	4	2	4	4	4	2	2.9
Q-14	5	1	3	3	5	2	4	4	3	2	3.2
Q-15	2	1	1	2	5	3	4	5	2	3	2.8
Q-16	3	1	2	2	4	2	4	5	2	2	2.7
Overall score of participant	1.9375	1	2.125	2.125	4.0625	2.5	3.375	5.5	1.5625	2.5	2.66875

Table 5.5: The responses received from the participants

PSSUQ	Average Score Achieved	PSSUQ Means
SysUse	2.55	2.8
InfoQual	2.634	3.02
IntQual	2.9	2.49
Overall	2.66875	2.82

Table 5.6: The average score achieved and the mean value given[1]

The mean value of the "overall" metric as stated by Lewis and James [1] is 2.82, but the participants P-5, P-7 and P-8 have values exceeding that. Hence, it can be said that these three participants are not satisfied with the usability of the tool.

Figure 5.5 shows a box plot that depicts the outliers of the metrics. There are two outliers noted in the box plot; one of them is for the SysUse-metric, and the other is for the Overall-metric. Both these outliers were identified as part of participant P-8's responses.

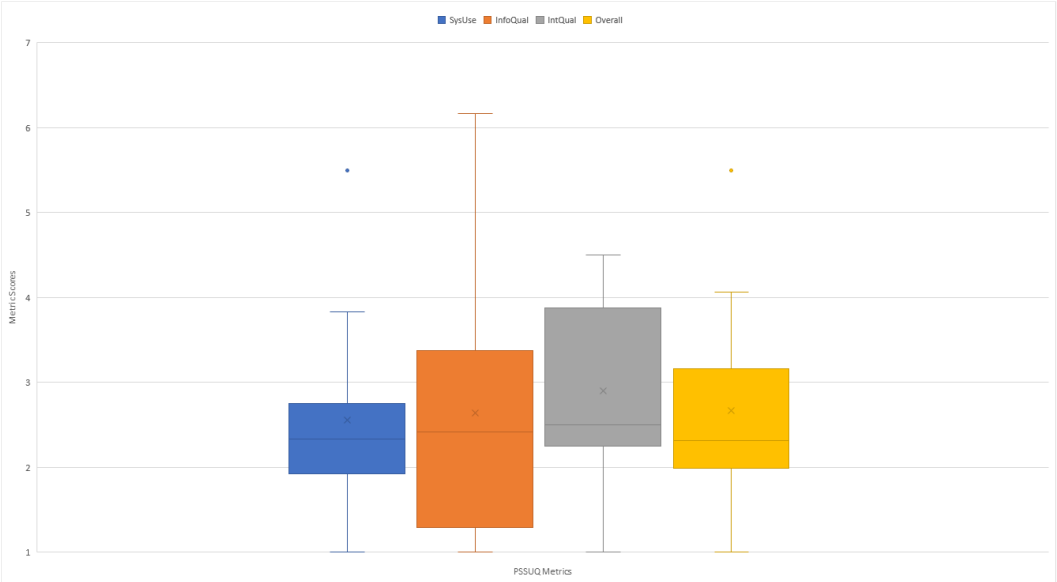

Figure 5.5: The responses received from the participants

It indicates that all participants, except P-8, thought that the overall usability of the tool was all right. P-8 seems to have had a bad experience using STEDI, as their scores are considerably outside the accepted range (seen in table 5.5). P-5 and P-7 were not entirely satisfied either, as their overall scores were higher than the mean-value stated by Lewis and James [1]. Another fascinating insight is that Participant P-2 seems to have strongly agreed with the tool's overall usability as all of the values submitted by them are "1", which casts some doubt over the legitimacy of their answer.

It is perceived that though the tool is usable and 7 out of 10 people were satisfied with it, it needs some work with regards to the User Interface (UI). It could include a more pleasant UI, and some additional functionalities could be provided. However, as STEDI is the first prototype tool in this domain, it is acceptable to have some room for improvement. Future work on STEDI should look into improving the PSSUQ scores achieved now.

Chapter 6

Conclusion

This chapter concludes this dissertation by stating briefly how the research objectives were achieved, and consequently, the research question was answered. The limitations of this approach are explained, and finally, some ideas are suggested for future work in this domain.

6.1 Revisiting The Research Question & Objectives

The research question (as stated in Section 1.2): To what extent can a knowledge-driven system accurately predict and report ethics-related issues that could arise after the integration of two or more linked-data datasets? The following research objectives were set (see Section 1.3) and have been achieved to answer the research question:

1. **Determine an approach to identify the type of data stored in a linked-data dataset.**

   The identification of the type of data takes place in two separate phases, as described in Section 4.5.1 and Section 4.5.2. The first phase focuses on retrieving the vocabularies used in the datasets and identifying the domain it usually models. This is accomplished by using the Linked Open Vocabularies API [1]. The second phase aims to analyse the predicates in the dataset using regular expressions and NLP techniques. The exact predicate terms are extracted from the URI

[1] `https://lov.linkeddata.es/dataset/lov/api`

using regular-expression-based pattern-matching. Then, an inbuilt Bag-of-Words containing ethically-sensitive terms is used in conjunction with spaCy's semantic similarity function to identify the type of data that is stored in the linked-data dataset.

2. **Establish a method to analyse multiple datasets and identify ethical concerns in them.**

   An ethics ontology was proposed as part of this dissertation to act as a knowledge-base that can store information regarding the ethical views of a dataset. When the ethical analysis of a dataset is completed, the results of the analysis will be pushed into the ethics ontology (as described in Section 4.5.3). Subsequently, as new datasets are being analysed, the knowledge-base (that is the ethics ontology) grows in size. The ethics ontology is queried during the generation of the ethics report. Though the ethics ontology does not identify the ethical concerns by itself, it plays a significant role in the approach taken in this dissertation to identify ethical issues across multiple datasets.

3. **Establish a method to predict ethical issues that might arise due to data integration.**

   The ethics ontology is queried to understand the ethical perspectives held by each dataset that is about to be integrated. Once that knowledge is available, it is compared with existing knowledge about data-integration-related ethical issues (as discussed in Section 4.5.4). If a similar pattern is identified in the datasets, then it is predicted that an ethical issue will arise if the datasets are integrated.

4. **Develop a prototype tool that can analyse multiple datasets, predict and generate a report about the materialisation of an ethical issue if the datasets were to be integrated.**

   A decision-support tool called STEDI (Support Tool for Ethical Data Integration) is proposed in this dissertation. It uses the techniques discussed above to accurately identify ethical concerns in the datasets and predict the possibility of an ethical issue arising if the datasets were to be integrated. Based on the analysis, it generates an ethics report that will be helpful for the data integrator.

5. **Evaluate the performance and viability of the approach.**

   As described in Section 5.2, the performance of STEDI was measured using various metrics. From the performance testing, it was concluded that STEDI was robust enough to not let any ethical issues go unnoticed. However, it did falsely claim a few non-issue predicates as ethical issues, and this was because of the NLP model falsely categorising some terms as ethically-sensitive. The viability of this approach was tested by requesting human participants to use the tool and answer the PSSUQ. The results of the PSSUQ showed that the tool is usable for now and that most people would be satisfied with such a tool - but the long-term viability of the tool is guaranteed only if more work is done with regards to the User Interface.

By achieving all of the research objectives, it is possible to say that a knowledge-driven system can very accurately predict and report ethics-related issues that could arise after the integration of two or more linked-data datasets. The approach used in this dissertation has been tested to be viable. If more work was done on this approach, it could become a standardised tool for predicting ethical issues in linked-data datasets.

6.2 Limitations Of STEDI

1. Ethically-sensitive vocabularies will not be correctly identified in Phase 1 (see Section 4.5.1) if the LOV API fails or if that database gets corrupted. Though STEDI will still be able to identify ethical issues using Phase 2, it is still a limitation of the tool.

2. A significant issue with Phase 2 (see Section 4.5.2) is that semantic similarity function does not always correctly categorise the issues; a few predicates are sometimes falsely identified as ethical issues. Though it does not affect the main aim of the tool, it does bring in unnecessary problems that could lead to decreased reputability for STEDI.

3. Only four complex data-integration-related ethical issues were considered during the development of STEDI. Hence, if STEDI encountered some other ethical-issue scenario, the issue will go unnoticed.

4. As noted in Section 5.2.4, STEDI took 220 seconds to analyse 216 triples. These 216 triples were from the mini-versions of the datasets. For comparison, the full-versions of the datasets have over 1.6 million triples in total. So a significant limitation of this tool is how slow it is on personal computers.

5. As seen in Section 5.3.5, the interface is one of the biggest concerns for STEDI.

6.3 Future Work

The limitations of STEDI have been listed in the previous section. Future work in this domain can revolve around fixing those limitations. Some of the suggested areas to work on in the future are:

1. A local repository of the vocabularies listed in LOV could be created. The local repository can be kept updated by fetching the database from LOV at regular intervals.

2. Cutting-edge NLP techniques need to be studied to combat the issue of the semantic similarity method producing some false-positives. The field of text analytics might provide some more insights into how this problem can be handled.

3. It is tough to expand the number of integration-related ethical-issue scenarios that this approach can handle. The list can be manually increased to handle all the known ethical issues for now, and when a new issue is discovered, the list can be updated. Machine Learning techniques can be deployed to learn from previous ethical issues and predict issues in the future.

4. The focus of this dissertation was not on the engineering of the tool; hence STEDI is a single-threaded application. Converting it into a multi-threaded application will make it considerably faster on personal computers. As explained in Section 4.8.1, it is clear that a personal computer is not the right platform for STEDI. Instead, STEDI would thrive if it were a web application with sufficient processing power on its end. Then, users could connect to the web application, upload their datasets and receive an ethics report. The processing power of the

user's computer would be removed from the equation, thereby allowing for faster processing of datasets.

5. The UI of the tool can be considerably more appealing. Usability and the interface play a significant role in user satisfaction and therefore, must be improved in the future.

6. The final report can be made interactive and explorable. It would be advantageous if the users could explore the issues in the datasets in a more visually engaging and informative way.

7. The ethics ontology models consent by taking some components from GConsent. However, a different approach can be explored to better model concepts for specifically for STEDI.

8. Different weights can be given to different kinds of ethical issues that have been detected. This could allow the user to understand the seriousness of the ethical issues a lot better.

6.4 Final Remarks

Pushing the boundaries of what is possible through technology is considered to be far more interesting than looking at the ethical ramifications of these advancements. This needs to change as modern learning algorithms, and data analytics often cause real damage to humanity. Most modern technologies will require the integration of two or more datasets to gain some additional knowledge about the problem being solved. This dissertation proposes a prototype tool called STEDI that can predict the materialisation of ethical issues before they are integrated. STEDI was built as a decision-support tool for data integrators to use before they integrate the datasets. STEDI uses NLP techniques to identify ethical concerns in individual datasets and uses that knowledge to predict the materialisation of ethical issues due to data integration. The results of the ethical analysis are presented in an ethics report for the consumption of the data integrator.

Bibliography

[1] J. Lewis, "Psychometric evaluation of the pssuq using data from five years of usability studies," *Int. J. Hum. Comput. Interaction*, vol. 14, pp. 463–488, 09 2002.

[2] K. W. Miller, "Is ethical behavior good for business?," *IT Professional*, vol. 14, no. 1, pp. 10–11, 2012.

[3] R. Beniwal, V. Gupta, M. Rawat, and R. Aggarwal, "Data mining with linked data: Past, present, and future," in *2018 Second International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1031–1035, Feb 2018.

[4] H. J. Pandit, A. Polleres, B. Bos, R. Brennan, B. Bruegger, F. J. Ekaputra, J. D. Fernández, R. G. Hamed, E. Kiesling, M. Lizar, E. Schlehahn, S. Steyskal, and R. Wenning, "Creating a vocabulary for data privacy," in *On the Move to Meaningful Internet Systems: OTM 2019 Conferences* (H. Panetto, C. Debruyne, M. Hepp, D. Lewis, C. A. Ardagna, and R. Meersman, eds.), (Cham), pp. 714–730, Springer International Publishing, 2019.

[5] Ashish Lochan, *Ethical Data Integration and General Data Protection Regulation*. M.Sc. Dissertation, University of Dublin, Trinity College, Aug. 2018.

[6] D. Linstedt, "Data Warehousing Ethical Concerns: Security, Access and Control." `"https://tdan.com/data-warehousing-ethical-concerns-security-access-and-control/5186"`, Apr. 2004. [Online; accessed 19-August-2020].

[7] P. M. Schwartz, "Data protection law and the ethical use of analytics, privacy and

security law report, 10 pvlr 70," 2011. [Available at: http://works.bepress.com/paul_schwartz/110/; accessed 05-September-2020].

[8] J. Delcker, "Europe's AI ethics chief: No rules yet, please." "https://www.politico.eu/article/pekka-ala-pietila-artificial-intelligence-europe-shouldnt-rush-to-regulate-ai-says-top-ethics-adviser/", Oct. 2018. [Online; accessed 20-August-2020].

[9] A. Narayanan and V. Shmatikov, "How to break anonymity of the netflix prize dataset," *ArXiv*, vol. abs/cs/0610105, 2006.

[10] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias." "https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing", May 2016. [Online; accessed 2-August-2020].

[11] Wikipedia contributors, "Netflix prize — Wikipedia, the free encyclopedia." "https://en.wikipedia.org/w/index.php?title=Netflix_Prize&oldid=970689854", 2020. [Online; accessed 19-August-2020].

[12] L. Sweeney, "Only You, Your Doctor, and Many Others May Know," *Technology Science*, Sept. 2015.

[13] Jeff Larson, Surya Mattu , Lauren Kirchner, and Julia Angwin, "How We Analyzed the COMPAS Recidivism Algorithm." "https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm", May 2016. [Online; accessed 2-August-2020].

[14] K. M. Boyd, "Ethnicity and the ethics of data linkage," *BMC Public Health*, vol. 7, no. 1, p. 318, 2007.

[15] Wikipedia contributors, "Data science — Wikipedia, the free encyclopedia." "https://en.wikipedia.org/w/index.php?title=Data_science&oldid=974700734", 2020. [Online; accessed 26-August-2020].

[16] Wikipedia contributors, "Information explosion — Wikipedia, the free encyclopedia." "https://en.wikipedia.org/w/index.php?title=Information_explosion&oldid=969034653", 2020. [Online; accessed 26-August-2020].

[17] Julia Stoyanovich, "Data, Responsibly Introduction to Data Mining." "https://dataresponsibly.github.io/documents/lecture.pdf". [Online; accessed 22-August-2020].

[18] A. Datta, M. C. Tschantz, and A. Datta, "Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination," *CoRR*, vol. abs/1408.6491, 2014.

[19] P. Fule and J. F. Roddick, "Detecting privacy and ethical sensitivity in data mining results," in *Proceedings of the 27th Australasian Conference on Computer Science - Volume 26*, ACSC '04, (AUS), p. 159–166, Australian Computer Society, Inc., 2004.

[20] Serge Abiteboul, "Issues in Ethical Data Management." "https://hal.inria.fr/hal-01621687/file/17.Ethics.Data.Namur.pdf", Oct. 2017. [Online; accessed 19-August-2020].

[21] Mr Data Insight, "Text mining on personal data in the age of gdpr." "https://medium.com/analytics-vidhya/text-mining-on-personal-data-in-the-age-of-gdpr-f8e7a2c138c6". [Online; accessed 22-August-2020].

[22] M. Douriez, H. Doraiswamy, J. Freire, and C. T. Silva, "Anonymizing nyc taxi data: Does it matter?," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 140–148, 2016.

[23] D. Firmani, L. Tanca, and R. Torlone, "Ethical dimensions for data quality," *J. Data and Information Quality*, vol. 12, Dec. 2019.

[24] Julia Stoyanovich and Serge Abiteboul, "Data, responsibly." "https://wp.sigmod.org/?p=1900", Nov. 2015. [Online; accessed 22-August-2020].

[25] Julia Stoyanovich , "Towards responsible data science." "https://dataresponsibly.github.io/documents/SIGMOD18_Julia.pdf". [Online; accessed 22-August-2020].

[26] Wikipedia contributors, "Linked data — Wikipedia, the free encyclopedia." "https://en.wikipedia.org/w/index.php?title=Linked_data&oldid=974464741", 2020. [Online; accessed 29-August-2020].

[27] Wikipedia contributors, "Semantic web — Wikipedia, the free encyclopedia." "https://en.wikipedia.org/w/index.php?title=Semantic_Web&oldid=973815071", 2020. [Online; accessed 29-August-2020].

[28] W3C, "LINKED DATA." "https://www.w3.org/standards/semanticweb/data". [Online; accessed 25-August-2020].

[29] Tim Berners-Lee, "Linked Data." "https://www.w3.org/DesignIssues/LinkedData.html", July 2006. [Online; accessed 25-August-2020].

[30] R. Fleiner, "Linking of open government data," *2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pp. 1–5, 2018.

[31] N. Shadbolt and K. O'Hara, "Linked data in government," *IEEE Internet Computing*, vol. 17, no. 4, pp. 72–77, 2013.

[32] C. Debruyne, E. Clinton, L. McNerney, A. Nautiyal, and D. O'Sullivan, "Serving ireland's geospatial information as linked data," in *International Semantic Web Conference*, 2016.

[33] N. Freire and S. d. Valk, "Automated interpretability of linked data ontologies: : an evaluation within the cultural heritage domain," in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 3072–3079, Dec 2019.

[34] M. Vidal, S. Jozashoori, and A. Sakor, "Semantic data integration techniques for transforming big biomedical data into actionable knowledge," in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 563–566, 2019.

[35] "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)." "https://eur-lex.europa.eu/eli/reg/2016/679/2016-05-04". [Online; accessed 26-August-2020].

[36] Wikipedia contributors, "General data protection regulation — Wikipedia, the free encyclopedia." "https://en.wikipedia.org/w/index.php?title=General_Data_Protection_Regulation&oldid=974946401", 2020. [Online; accessed 29-August-2020].

[37] H. J. Pandit, K. Fatema, D. O'Sullivan, and D. Lewis, "Gdprtext - gdpr as a linked data resource," in *The Semantic Web* (A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, eds.), (Cham), pp. 481–495, Springer International Publishing, 2018.

[38] H. J. Pandit, C. Debruyne, D. O'Sullivan, and D. Lewis, "Gconsent - a consent ontology based on the gdpr," in *The Semantic Web* (P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A. J. Gray, V. Lopez, A. Haller, and K. Hammar, eds.), (Cham), pp. 270–282, Springer International Publishing, 2019.

[39] Pierre-Yves Vandenbussche, Ghislain Atemezing, Maria Poveda, and Bernard Vatant, "Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web," *IOS Press*, vol. 8, pp. 437–452, Dec. 2016.

[40] S. Kirrane, J. D. Fernández, W. Dullaert, U. Milosevic, A. Polleres, P. A. Bonatti, R. Wenning, O. Drozd, and P. Raschke, "A scalable consent, transparency and compliance architecture," in *The Semantic Web: ESWC 2018 Satellite Events* (A. Gangemi, A. L. Gentile, A. G. Nuzzolese, S. Rudolph, M. Maleshkova, H. Paulheim, J. Z. Pan, and M. Alam, eds.), (Cham), pp. 131–136, Springer International Publishing, 2018.

[41] BigID, "BigID - Satisfy EU GDPR Data Protection Requirements With Automation.." "https://bigid.com/home/eu-gdpr/". [Online; accessed 27-August-2020].

[42] G. Lenzini, "DAta Protection REgulation COmpliance." "https://www.fnr.lu/projects/data-protection-regulation-compliance/", Aug. 2020. [Online; accessed 26-August-2020].

[43] SAS Institute Inc., "SAS® for Personal Data Protection." "https://www.sas.com/en_ie/solutions/personal-data-protection.html". [Online; accessed 27-August-2020].

[44] "BPR4GDPR - Deliverables." "https://www.bpr4gdpr.eu/results/deliverables/, author = Spiros Alexakis and Kalaboukas Konstantinos and Georgios V. Lioudakis and Marwan Hassani, note = "[Online; accessed 27-August-2020]".

[45] Spiros Alexakis, Kalaboukas Konstantinos, Georgios V. Lioudakis, and Marwan Hassani, "BPR4GDPR." "https://www.bpr4gdpr.eu/". [Online; accessed 27-August-2020].

[46] E. Kenneally and M. Fomenkov, "Cyber research ethics decision support (creds) tool," in *Proceedings of the 2015 ACM SIGCOMM Workshop on Ethics in Networked Systems Research*, NS Ethics '15, (New York, NY, USA), p. 21, Association for Computing Machinery, 2015.

[47] T. Berners-Lee, R. T. Fielding, and L. Masinter, "Uniform resource identifier (uri): Generic syntax," STD 66, RFC Editor, January 2005. "http://www.rfc-editor.org/rfc/rfc3986.txt".

[48] M. Grüninger and M. Fox, "Methodology for the Design and Evaluation of Ontologies," in *IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing, April 13, 1995*, 1995.

[49] N. Noy and D. Mcguinness, "Ontology development 101: A guide to creating your first ontology," *Knowledge Systems Laboratory*, vol. 32, 01 2001.

[50] D. Wisniewski, J. Potoniec, A. Lawrynowicz, and C. M. Keet, "Competency questions and SPARQL-OWL queries dataset and analysis," *CoRR*, vol. abs/1811.09529, 2018.

[51] Wikipedia contributors, "Personal data — Wikipedia, the free encyclopedia." "https://en.wikipedia.org/w/index.php?title=Personal_data&oldid=976273104", 2020. [Online; accessed 2-September-2020].

[52] "Guide to Identifying Personally Identifiable Information (PII)." "https://www.technology.pitt.edu/help-desk/how-to-documents/guide-identifying-personally-identifiable-information-pii", Feb. 2017. [Online; accessed 1-September-2020].

[53] Information Commissioner's Office, "Principle (c): Data minimisation." "https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/data-minimisation/". [Online; accessed 2-September-2020].

[54] Experian Information Solutions, Inc., "What is Data Minimisation?." "https://www.experian.co.uk/business/glossary/data-minimisation/". [Online; accessed 2-September-2020].

[55] "Children's Privacy." "https://www.ftc.gov/tips-advice/business-center/privacy-and-security/children%27s-privacy". [Online; accessed 3-September-2020].

[56] N. S. Vageesh, "Tracking online behaviour for granting loans." "https://www.thehindubusinessline.com/money-and-banking/tracking-online-behaviour-for-granting-loans/article8991329.ece", Jan. 2018. [Online; accessed 1-September-2020].

[57] R. Kumar, "Not CIBIL, this lender uses your social media behaviour for loan up to Rs 2 lakh!." "https://www.financialexpress.com/money/not-cibil-this-lender-uses-your-social-media-behaviour-for-loan-up-to-rs-2-lakh/1761934/", Nov. 2019. [Online; accessed 1-September-2020].

[58] Livemint, "How your social media behaviour can affect your loan." "https://www.livemint.com/Home-Page/rWieRBvuk8U3HdvuszdBXI/How-your-social-media-behaviour-can-affect-your-loan.html", Feb. 2018. [Online; accessed 1-September-2020].

[59] "Ireland's Health Services." "https://www.hse.ie/eng/". [Online; accessed 2-September-2020].

[60] J. Horgan-Jones, "Data from HSE website users 'leaked to commercial actors'." "https://www.irishtimes.com/business/technology/data-from-hse-website-users-leaked-to-commercial-actors-1.3829547". [Online; accessed 2-September-2020].

[61] "How Filter Bubbles Distort Reality: Everything You Need to Know." "https://fs.blog/2017/07/filter-bubbles/", July 2017. [Online; accessed 2-September-2020].

[62] D. Baer, "The 'Filter Bubble' Explains Why Trump Won and You Didn't See It Coming." "https://www.thecut.com/2016/11/how-facebook-and-the-filter-bubble-pushed-trump-to-victory.html", Nov. 2016. [Online; accessed 2-September-2020].

[63] M. M. El-Bermawy, "Your Filter Bubble is Destroying Democracy," *Wired*, Nov. 2016. [Online; accessed 2-September-2020].

[64] T. Tudorache, C. Nyulas, N. F. Noy, and M. A. Musen, "Webprotégé: A collaborative ontology editor and knowledge acquisition tool for the web," *Semant. Web*, vol. 4, p. 89–99, Jan. 2013.

[65] "Understanding Language Syntax and Structure: A Practitioner's Guide to NLP." "https://www.kdnuggets.com/understanding-language-syntax-and-structure-a-practitioners-guide-to-nlp.html/". [Online; accessed 5-September-2020].

[66] "10. Analyzing the Meaning of Sentences." "https://www.nltk.org/book/ch10.html". [Online; accessed 5-September-2020].

[67] "Migration, ethnicity and religion (T2) SA | All-Island Research Observatory." "https://airo.maynoothuniversity.ie/datastore/migration-ethnicity-and-religion-t2-sa". [Online; accessed 5-September-2020].

[68] "Crimes at Garda Stations Level 2010-2016 - data.gov.ie." "https://data.gov.ie/dataset/crimes-at-garda-stations-level-2010-2016". [Online; accessed 5-September-2020].

[69] A. Crotti Junior, C. Debruyne, and D. O'Sullivan, "Juma uplift: Using a block metaphor for representing uplift mappings," in *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pp. 211–218, 2018.

Appendix A

Property Name	Property Type	Domain	Range
has consent	Object property	Data Subject	Consent
has status	Object property	Consent	Status
is consent for data subject	Object property	Consent	Data Subject
is provided to controller	Object property	Consent	Data Controller
is status for consent	Object property	Status	Consent
was provided consent	Object property	Data Controller	Consent
has age	Data property	Group or Individual	xsd:boolean
has behaviour data	Data property	Group or Individual	xsd:boolean
has child data	Data property	Group or Individual	xsd:boolean
has contact information	Data property	Individual	xsd:boolean
has criminal activity	Data property	Group or Individual	xsd:boolean
has data controller name	Data property	Data Controller	xsd:string
has ethnicity data	Data property	Group or Individual	xsd:boolean
has files with PII attached	Data property	Group or Individual	xsd:boolean
has health data	Data property	Group or Individual	xsd:boolean
has income data	Data property	Group or Individual	xsd:boolean
has loan records	Data property	Group or Individual	xsd:boolean
has location data	Data property	Group or Individual	xsd:boolean
has name	Data property	Individual	xsd:boolean
has physical characteristics	Data property	Group or Individual	xsd:boolean
has political opinions	Data property	Group or Individual	xsd:boolean
has religion	Data property	Group or Individual	xsd:boolean
has signed NDA	Data property	Individual	xsd:boolean
has too many data points	Data property	Individual	xsd:boolean
has user tracking data	Data property	Group or Individual	xsd:boolean
is valid for processing	Data property	Status	xsd:boolean

Table 1: All the properties in the ethics ontology

Appendix B

B.1 Processing Predicates

```
def predicate_processor(self, word_lists_dict):
    for p in self.graph.predicates():
        if p not in _Dataset.common_schema_predicates:
            predicate_parts = p.split("/")
            predicate = predicate_parts[-1]
            predicate = re.sub("[^A-Za-z0-9 ]+", " ", predicate)
            # To split camelCase
            predicate = re.sub(r"([A-Z])", r" \1", predicate)
            predicate = predicate.lower()

            predicate_tokens = nlp(predicate)

            for token in predicate_tokens:
                if token.text not in _Dataset.common_words_to_ignore:
                    for word_list in word_lists_dict.values():
                        for issue in word_list[0]:
                            # 1st condition checks if it's the same thing, this eliminates the unnecessary use of NLP.
                            # 2nd condition avoids checking similarity for empty vectors first and then does the similarity check.
                            # "0.5" allowed a wider range of words to creep in as issues, so after trial and error I settled on "0.6".
                            if (str(token).lower() == issue.lower()) or (token.has_vector and token.similarity(nlp(issue)) > 0.6) :
                                data_property = [word_list[1]][0]
                                self.ethics_ontology_dictionary[data_property]["value"] = True
                                if str(p) not in self.ethics_ontology_dictionary[data_property]["trigger_items"]["PREDICATES"]:
                                    self.ethics_ontology_dictionary[data_property]["trigger_items"]["PREDICATES"].append(str(p))
                                    print(f"\nIssue detected: {data_property}")
```

Figure 1: Code snippet showing the predicate processor method

B.2 Filling Ethics Ontology

```python
def fill_ethics_ontology(self, ethics_ontology):
    # Cleaning up dataset_name so the individuals of the ethics ontology follow a consistent naming convention.
    dataset_name = os.path.splitext(self.dataset_name)[0]
    dataset_name = re.sub("[^A-Za-z0-9 ]+", " ", dataset_name)
    dataset_name = re.sub("[ ]+", " ", dataset_name)
    dataset_name = dataset_name.replace(" ", "_")

    # Creates a named individual with the name of the dataset
    ethics_ontology.add((EONS[dataset_name], RDF.type, OWL.NamedIndividual))

    # Identifying data-subject type and removing that information from the dictionary
    if self.ethics_ontology_dictionary["representsIndividuals"]["value"] == True:
        ethics_ontology.add((EONS[dataset_name], RDF.type, EONS.Individual))
        del self.ethics_ontology_dictionary["representsIndividuals"]
        del self.ethics_ontology_dictionary["representsGroups"]

    else:
        ethics_ontology.add((EONS[dataset_name], RDF.type, EONS.Group))
        del self.ethics_ontology_dictionary["representsIndividuals"]
        del self.ethics_ontology_dictionary["representsGroups"]

    for issue, issue_info in self.ethics_ontology_dictionary.items():
        ethics_ontology.add((EONS[dataset_name], EONS[issue], rdflib.term.Literal(issue_info["value"])))

        if "trigger_items" in self.ethics_ontology_dictionary[issue]:
            if self.ethics_ontology_dictionary[issue]["trigger_items"]["VOCABULARIES"] or self.ethics_ontology_dictionary[issue]["trigger_items"]["PREDICATES"]:
                ethics_ontology.add((EONS[dataset_name], EONS[f"{issue}TriggeredBy"], rdflib.term.Literal(issue_info["trigger_items"])))
```

Figure 2: Code snippet showing the method used to fill the ethics ontology

B.2 Identifying Data-Integration Issues

```
def quick_issue_checker(self, issue, dataset):
    # A method to just check for specific issues to see if they can be a linking point.
    for d, e_dict in self.ethics_dicts_dict.items():
        if d != dataset and e_dict[issue]["value"] == True:
            return True

    return False

def check_integration_issue_scenarios(self):
    for dataset, ethics_dict in self.ethics_dicts_dict.items():
        # Scenario-1 : Check for ethnicity-criminal associations being made.
        for issue in self.scenario_1_issues.keys():
            if ethics_dict[issue]["value"] == True:
                if issue == "hasLocationData": # A minimum of 2 datasets need have location data to provide linkage between the datasets.
                    self.scenario_1_issues[issue] = self.quick_issue_checker(issue, dataset)
                else:
                    self.scenario_1_issues[issue] = True

        # Scenario-2 : Check for behaviour-loan repayment associations being made.
        for issue in self.scenario_2_issues.keys():
            if ethics_dict[issue]["value"] == True:
                if issue == "hasUserTrackingData": # A minimum of 2 datasets need have tracking data to provide linkage between the datasets.
                    self.scenario_2_issues[issue] = self.quick_issue_checker(issue, dataset)
                else:
                    self.scenario_2_issues[issue] = True

        # Scenario-3 : Check for social media activity used to manipulate insurance rates.
        for issue in self.scenario_3_issues.keys():
            if ethics_dict[issue]["value"] == True:
                if issue == "hasBehaviourData":
                    self.scenario_3_issues[issue] = True
                else: # Tracking, name & location need to be common to provide some linkage between the datasets.
                    self.scenario_3_issues[issue] = self.quick_issue_checker(issue, dataset)

        # Scenario-4 : Check for tailored reality/ filter bubble issue caused by grouping of political opinions and other factors.
        for issue in self.scenario_4_issues.keys():
            if ethics_dict[issue]["value"] == True:
                if issue == "hasPoliticalOpinions":
                    self.scenario_4_issues[issue] = True
                else: # Age, behaviour, ethnicity, income, location, religion need to be common to provide some linkage between the datasets.
                    self.scenario_4_issues[issue] = self.quick_issue_checker(issue, dataset)
```

Figure 3: Code snippet showing how data integration issues are identified

```
self.scenario_1_issues = {
    "hasCriminalActivity":  False,
    "hasEthnicityData": False,
    "hasLocationData": False,
    "hasReligion": False
}

self.scenario_2_issues = {
    "hasBehaviourData": False,
    "hasLoanRecords": False,
    "hasUserTrackingData": False
}

self.scenario_3_issues = {
    "hasUserTrackingData": False,
    "hasName": False,
    "hasBehaviourData": False,
    "hasLocationData": False
}

self.scenario_4_issues = {
    "hasAge": False,
    "hasBehaviourData": False,
    "hasEthnicityData": False,
    "hasIncomeData": False,
    "hasLocationData": False,
    "hasPoliticalOpinions": False,
    "hasReligion": False
}
```

Figure 4: Code snippet showing how the data-integration scenarios are stored

Appendix C

C.1 Application Flow Screenshots

Figure 5: Application Flow Screenshot - 1

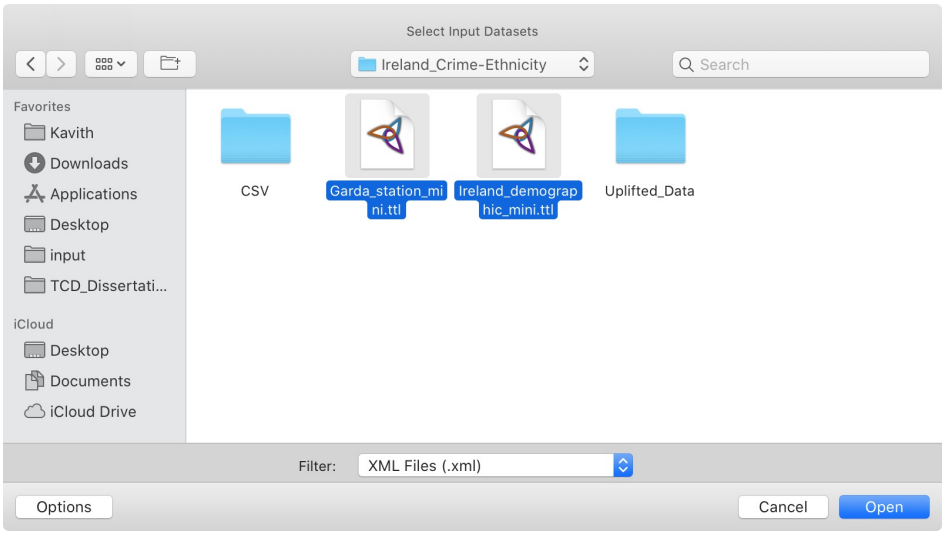

Figure 6: Application Flow Screenshot - 2

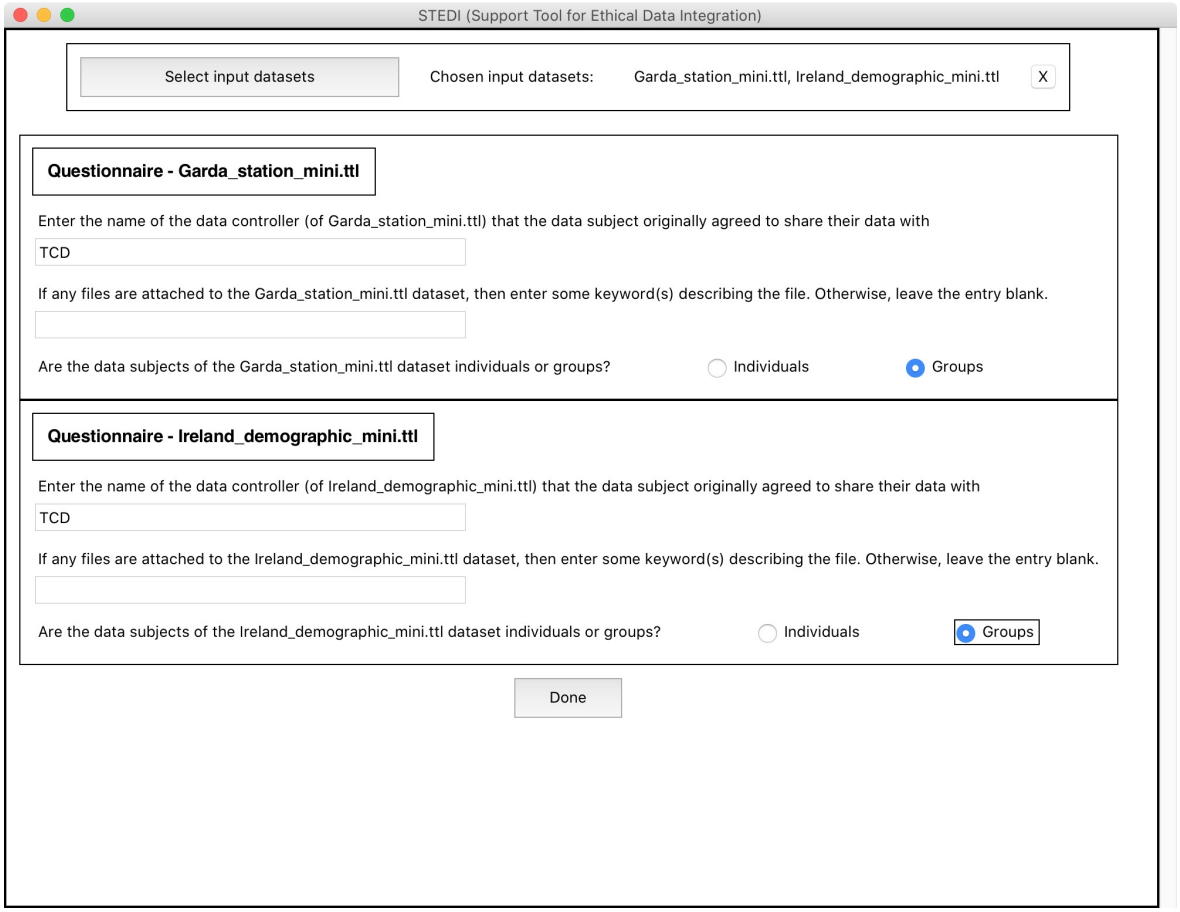

Figure 7: Application Flow Screenshot - 3

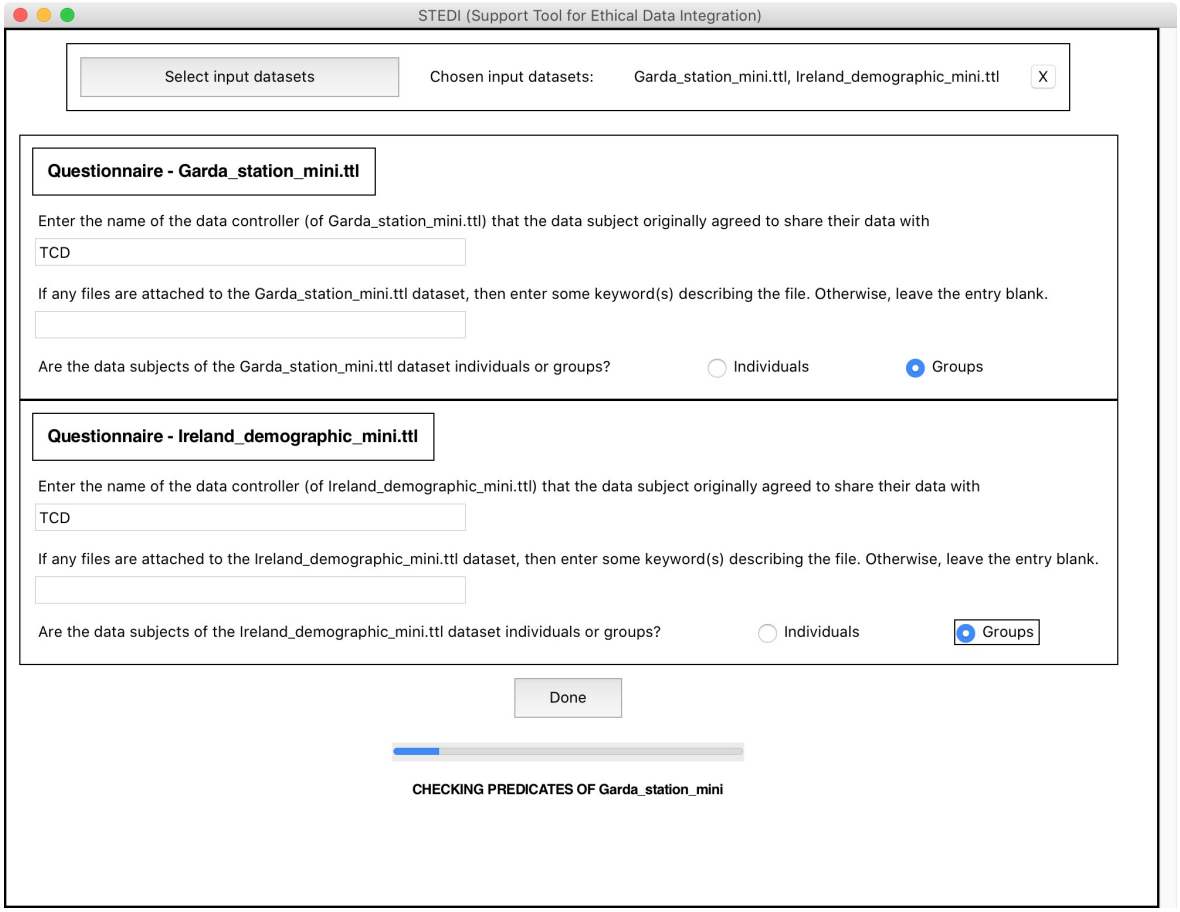

Figure 8: Application Flow Screenshot - 4

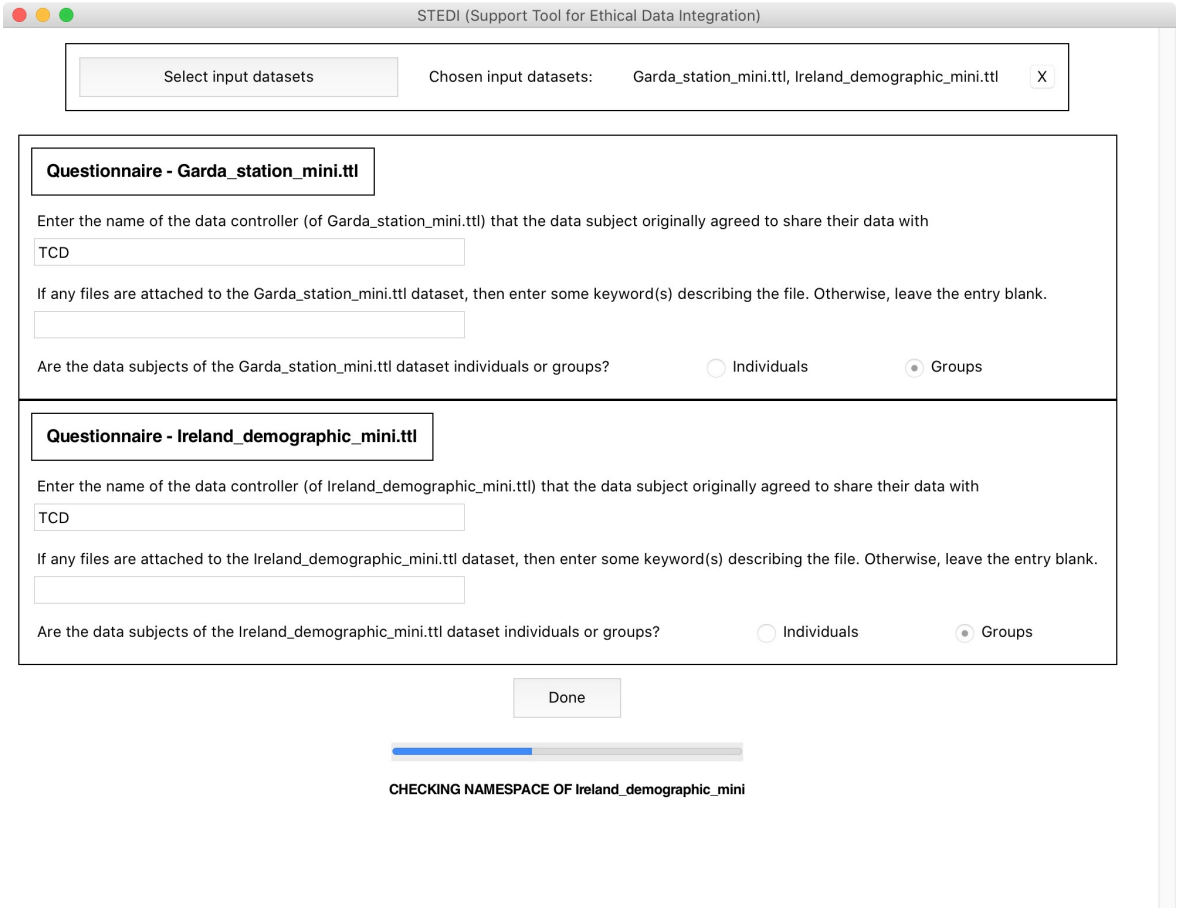

Figure 9: Application Flow Screenshot - 5

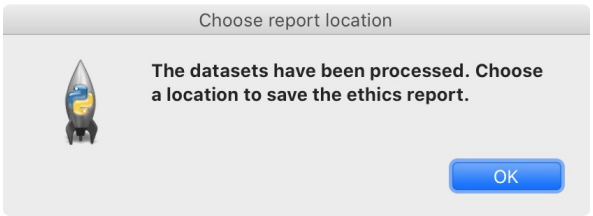

Figure 10: Application Flow Screenshot - 6

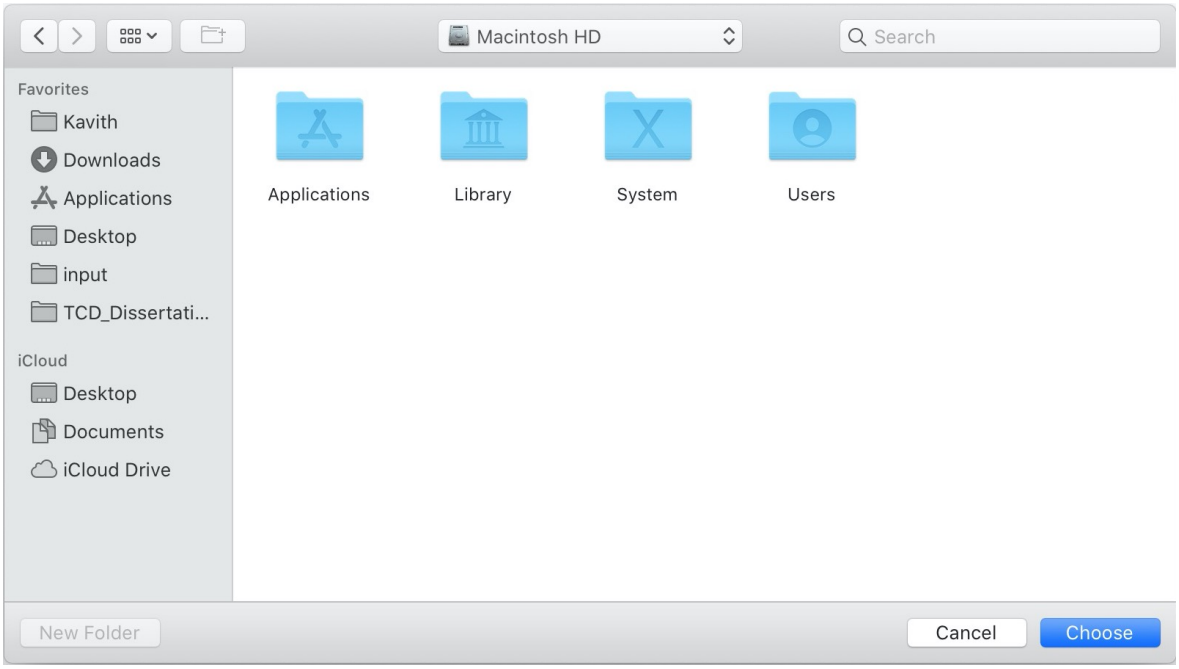

Figure 11: Application Flow Screenshot - 7

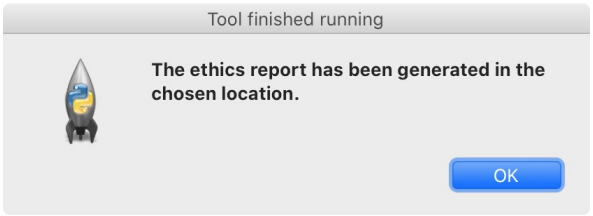

Figure 12: Application Flow Screenshot - 8

C.2 Error Messages In STEDI

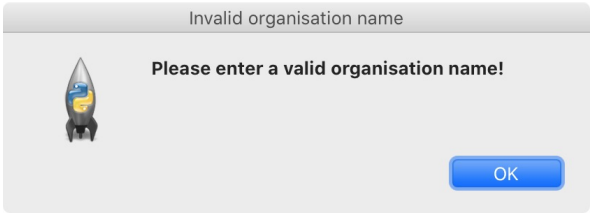

Figure 13: Error message - 1

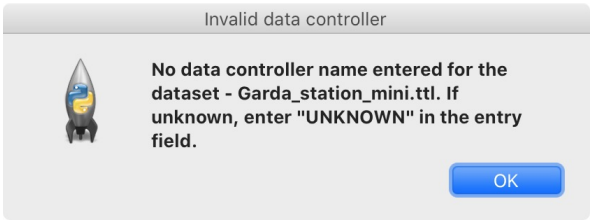

Figure 14: Error message - 2

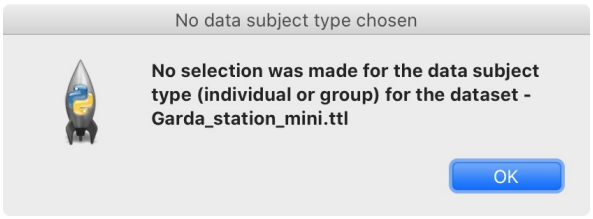

Figure 15: Error message - 3

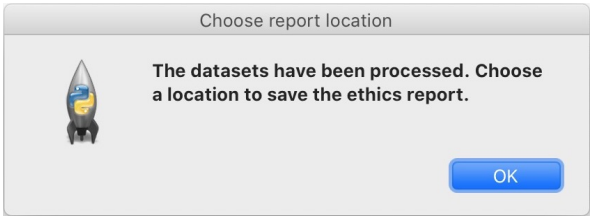

Figure 16: Error message - 4

Appendix D

D.1 Ethics Report Generated During Execution - 1

ETHICS REPORT FOR INDIVIDUAL DATASET - IRELAND_DEMOGRAPHIC_MINI

Data controller: tcd

Valid for processing: True

This dataset represents groups.

Issues present in the dataset:

1. HAS CHILD DATA

Predicates that triggered this issue:

* http://example.org/csv/perc_place_of_birth_lithuania_2011
* http://example.org/csv/place_of_birth_ireland_2011
* http://example.org/csv/perc_place_of_birth_poland_2011_
* http://example.org/csv/perc_place_of_birth_other_eu_28_2011
* http://example.org/csv/perc_place_of_birth_uk_2011
* http://example.org/csv/place_of_birth_rest_of_world_2011
* http://example.org/csv/perc_place_of_birth_ireland_2011
* http://example.org/csv/place_of_birth_total_2011
* http://example.org/csv/place_of_birth_uk_2011
* http://example.org/csv/perc_place_of_birth_rest_of_world_2011
* http://example.org/csv/place_of_birth_other_eu_28_2011
* http://example.org/csv/place_of_birth_poland_2011
* http://example.org/csv/place_of_birth_lithuania_2011

2. HAS ETHNICITY DATA

Predicates that triggered this issue:

* http://example.org/csv/perc_religion_not_stated_2011
* http://example.org/csv/perc_religion_no_religion_2011
* http://example.org/csv/foreign_languages_other_2011
* http://example.org/csv/foreign_languages_polish_2011
* http://example.org/csv/religion_catholic_2011
* http://example.org/csv/perc_religion_catholic_2011
* http://example.org/csv/religion_total_2011
* http://example.org/csv/perc_religion_other_stated_religion_2011
* http://example.org/csv/religion_not_stated_2011
* http://example.org/csv/religion_no_religion_2011
* http://example.org/csv/religion_other_stated_religion_2011
* http://example.org/csv/perc_foreign_languages_lithuanian_2011
* http://example.org/csv/perc_foreign_languages_other_2011
* http://example.org/csv/foreign_languages_french_2011
* http://example.org/csv/foreign_languages_lithuanian_2011
* http://example.org/csv/perc_foreign_languages_french_2011
* http://example.org/csv/perc_foreign_languages_polish_2011
* http://example.org/csv/foreign_languages_total_2011

3. HAS LOCATION DATA

Predicates that triggered this issue:

* http://example.org/csv/usual_residence_outside_ireland_2011
* http://example.org/csv/electoral_division_cso_code
* http://example.org/csv/perc_usual_residence_same_address_2011
* http://example.org/csv/perc_usual_residence_elsewhere_in_county_2011
* http://example.org/csv/county
* http://example.org/csv/perc_usual_residence_outside_ireland_2011
* http://example.org/csv/usual_residence_same_address_2011

* http://example.org/csv/planning_region
* http://example.org/csv/usual_residence_elsewhere_in_county_2011

4. HAS POLITICAL OPINIONS

Predicates that triggered this issue:
* http://example.org/csv/perc_religion_not_stated_2011
* http://example.org/csv/perc_religion_no_religion_2011
* http://example.org/csv/religion_catholic_2011
* http://example.org/csv/perc_religion_catholic_2011
* http://example.org/csv/religion_total_2011
* http://example.org/csv/perc_religion_other_stated_religion_2011
* http://example.org/csv/religion_not_stated_2011
* http://example.org/csv/religion_no_religion_2011
* http://example.org/csv/religion_other_stated_religion_2011

5. HAS RELIGION

Predicates that triggered this issue:
* http://example.org/csv/perc_religion_not_stated_2011
* http://example.org/csv/perc_religion_no_religion_2011
* http://example.org/csv/religion_catholic_2011
* http://example.org/csv/perc_religion_catholic_2011
* http://example.org/csv/religion_total_2011
* http://example.org/csv/perc_religion_other_stated_religion_2011
* http://example.org/csv/religion_not_stated_2011
* http://example.org/csv/religion_no_religion_2011
* http://example.org/csv/religion_other_stated_religion_2011

Data controller: tcd
Valid for processing: True
This dataset represents groups.

Issues present in the dataset:

1. HAS CRIMINAL ACTIVITY
Predicates that triggered this issue:
* http://example.org/csv/theft_and_related_offences
* http://example.org/csv/attempts_or_threats_to_murder_assaults_harassments_and_related_offenc
* http://example.org/csv/robbery_extortion_and_hijacking_offences
* http://example.org/csv/fraud_deception_and_related_offences
* http://example.org/csv/offences_against_government_justice_procedures_and_organisation_of_cri
* http://example.org/csv/burglary_and_related_offences

2. HAS ETHNICITY DATA
Predicates that triggered this issue:
* http://example.org/csv/public_order_and_other_social_code_offences

3. HAS LOCATION DATA
Predicates that triggered this issue:
* http://example.org/csv/station
* http://example.org/csv/y
* http://example.org/csv/x
* http://example.org/csv/public_order_and_other_social_code_offences

---

ETHICS REPORT FOR DATA INTEGRATION OF ALL DATASETS

+ SCENARIO-1 : Locations can be linked and certain races can be unethically claimed as more inclined to be criminals.

+ SCENARIO-1 : Location can be linked and certain ethnic groups can be unethically claimed as more inclined to be criminals.

+ SCENARIO-1 : Since locations can be linked and criminal data is involved, any datapoint from any of the datasets can be used to make ethically wrong assumptions.

+ SCENARIO-4 : Certain users can be unethically targeted with others' political opinions just because they belong to the same ethnic group.

+ SCENARIO-4 : Certain users can be unethically targeted with others' political opinions just because they reside in the same area.

D.2 Ethics Report Generated During Execution - 2

ETHICS REPORT FOR INDIVIDUAL DATASET - TRACKED_USER_DATA

Data controller: test2
Valid for processing: False
This dataset represents individuals.

Issues present in the dataset:

1. HAS AGE
Predicates that triggered this issue:
* http://www.semanticweb.org/kavith/ontologies/tracked-user-data#hasAge

2. HAS BEHAVIOUR DATA
Predicates that triggered this issue:
* http://www.semanticweb.org/kavith/ontologies/tracked-user-data#hasInterest

3. HAS LOAN RECORDS
Predicates that triggered this issue:
* http://www.semanticweb.org/kavith/ontologies/tracked-user-data#hasInterest

4. HAS USER TRACKING DATA
Predicates that triggered this issue:
* http://www.semanticweb.org/kavith/ontologies/tracked-user-data#hasInterest
* http://www.semanticweb.org/kavith/ontologies/tracked-user-data#isUserSince
* http://www.semanticweb.org/kavith/ontologies/tracked-user-data#hasAge
* http://www.semanticweb.org/kavith/ontologies/tracked-user-data#hasAdTrackingID

ETHICS REPORT FOR INDIVIDUAL DATASET - BANK_LOAN_DATA

Data controller: test1
Valid for processing: False
This dataset represents individuals.

Issues present in the dataset:

1. HAS BEHAVIOUR DATA

Predicates that triggered this issue:
* http://www.semanticweb.org/kavith/ontologies/bank-loan#hasBorrowedLoanAmount
* http://www.semanticweb.org/kavith/ontologies/bank-loan#hasWebTrackerId
* http://www.semanticweb.org/kavith/ontologies/bank-loan#hasDefaultedInThePast

2. HAS LOAN RECORDS

Predicates that triggered this issue:
* http://www.semanticweb.org/kavith/ontologies/bank-loan#hasBorrowedLoanAmount
* http://www.semanticweb.org/kavith/ontologies/bank-loan#hasWebTrackerId
* http://www.semanticweb.org/kavith/ontologies/bank-loan#hasDefaultedInThePast

3. HAS USER TRACKING DATA

Predicates that triggered this issue:
* http://www.semanticweb.org/kavith/ontologies/bank-loan#hasWebTrackerId

---

ETHICS REPORT FOR DATA INTEGRATION OF ALL DATASETS

+ SCENARIO-2 : By cross-site tracking a user, unethical assumptions can be made with regards to their loan repayment capabilities and their general interest/behaviour.

+ SCENARIO-2 : Cross-site tracking can be linked with the user's loan records to make any unethical assumption regarding the user.

+ SCENARIO-3 : Based on cross-site tracking data and the behavioural data of a user, unethical assumptions can be made about the user's activities thereby manipulating insurance rates.

+ SCENARIO-3 : Unethical assumption can also be made about the activities of the user's connections (friends, family, followers) on social media accounts.

+ SCENARIO-3 : Online tracking details of a user is very sensitive. It can be combined with any other data about the individual to gain extra information that the user did not consent to originally.

D.3 List of questions as part of the PSSUQ

1. Overall, I am satisfied with how easy it is to use this system.
2. It was simple to use this system.
3. I was able to complete the tasks and scenarios quickly using this system.
4. I felt comfortable using this system.
5. It was easy to learn to use this system.
6. I believe I could become productive quickly using this system.
7. The system gave error messages that clearly told me how to fix problems.
8. Whenever I made a mistake using the system, I could recover easily and quickly.

9. The information (such as online help, on-screen messages, and other documentation) provided with this system was clear.
10. It was easy to find the information I needed.
11. The information was effective in helping me complete the tasks and scenarios.
12. The organization of information on the system screens was clear.
13. The interface of this system was pleasant.
14. I liked using the interface of this system.
15. This system has all the functions and capabilities I expect it to have.
16. Overall, I am satisfied with this system.