PRIVACY ONTOLOGY FOR HEALTH DATA SHARING IN RESEARCH

DSAP: DATA SHARING AGREEMENT PRIVACY ONTOLOGY

BY

MINGYUAN LI, HBSc

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER AND SOFTWARE

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Master of Science

Master of Science (2018) McMaster University

(computer and software) Hamilton, Ontario, Canada

TITLE:	DSAP: Data Sharing Agreement Privacy Ontology
AUTHOR:	Mingyuan Li HBSc (Human Biology), University of Toronto, Toronto, Canada
SUPERVISOR:	Dr. Reza Samavi
NUMBER OF PAGES:	x, 85

Abstract

Medical researchers utilize data sharing agreements (DSA) to communicate privacy policies that govern the treatment of data in their collaboration. Expression of privacy policies in DSAs have been achieved through the use of natural and policy languages. However, ambiguity in natural language and rigidness in policy languages make them unsuitable for use in collaborative medical research. Our goal is to develop an unambiguous and flexible form of expression of privacy policies for collaborative medical research. In this thesis, we developed a DSA Privacy Ontology to express privacy policies in medical research. Our ontology was designed with hierarchy structure, lightweight in expressivity, closed world assumption in interpretation, and the reuse of other ontologies. The design allows our ontology to be flexible and extensible. Being flexible allows our ontology to express different types of privacy policies. Being extensible allows our ontology to be mapped to other linkable ontologies without the need to change our existing ontology. We demonstrate that our ontology is capable of supporting the DSA in a collaborative research data sharing scenario through providing the appropriate vocabulary and structure to log privacy events in a linked data based audit log. Furthermore, through querying the audit log, we can answer privacy competency questions relevant to medical researchers.

I dedicate this thesis to my parents, for their endless and unconditional love.

Acknowledgements

I would like to thank my supervisor Dr. Reza Samavi for your mentorship, support, and guidance throughout my Master's degree. I am grateful for McMaster for providing me with the opportunity to carry out research in the ehealth field.

I would also like to thank my supervisory committee, Dr. Fei Chiang and Dr. Norm Archer, for your time and support in my thesis research. I appreciate your expertise and suggestions in my research.

I would like to express my sincere gratitute to Andrzej Parkitny, Drew van Camp, and Octavian Ghiugan for providing me with a rewarding, educational, and memorable MSc ehealth internship experience at eHealth Ontario.

I owe deep gratitute to Andrew Sutton and Omar Boursalie for your input in my research and also for your help in getting me started with LaTeX.

My Master's degree would not be complete without my "meme groups," who made the past two years much more bearable and enjoyable. I can always count on you for fresh memes.

I would like to acknowledge Freepik, Vectors Market, Smashicons, and Eucalyp of `flaticon.com` for providing icons used in Figure 3.1.

Last but certainly not least, I would like to thank my parents for all your support in this long journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Data sharing is the cornerstone of modern medical research. Many benefits to data sharing motivate its ubiquitous practice among medical researchers. For example, data sharing allows medical researchers to reproduce and verify research, make publicly funded research available to the public, allow research data to go under peer scrutiny, enable the analysis of health of different populations across geographical barriers, avoid duplications in research, and data reuse [15, 24, 34].

Though benefits to data sharing are straightforward, its challenges are far more complex. The challenges are more apparent in the context of medical research where privacy-sensitive datasets are shared [55]. Patients value medical research advancements that promise dramatic improvements to the way we understand and treat debilitating illnesses, promoted by medical data sharing. However, patients also value personal privacy, something that could be undermined by medical data sharing due to possible inappropriate disclosure of personal information involved in sharing. These conflicting demands raise complex challenges to open data sharing among medical researchers, which make data sharing agreements (DSAs) necessary to constrain the

degree of data sharing among medical researchers to protect the privacy of their research subjects.

A DSA is a written document composed of static text expressing the constraints that researchers must follow when sharing their research data containing personal information [58]. The composition of the expressed constraints, or privacy policies, is guided by the goal to foster data sharing but also to protect research subjects' personal information and identity. Typically, legal constraints and ethical guidelines in the researcher's jurisdiction and institution is also referenced when composing the DSA [32]. Once completed, the DSA must then be actively enforced in the scientific workflow [39]. Enforcement of DSAs may be achieved through manual auditing of data access and data sharing logs, or through automated mechanisms embedded in the software used in the scientific workflow [63].

Medical researchers are busy scientists preoccupied with running their experiments and analyzing data. The researchers have little time to manage the intricate details contained in dozens of pages of free text in DSAs that constrain their freedom of data sharing, storage, and analysis. Moreover, because the privacy policies in DSAs are expressed in natural language, there also exist ambiguity. Therefore, the semantics of the privacy policy may be misinterpreted by different parties involved in the collaborative medical research [50]. For example, in the following DSA excerpt:

> "Upon termination of this Agreement, receiving party shall promptly destroy all documents, files and other *tangible materials* representing disclosing party's Information."

It is unclear what "other tangible materials" refer to. Macmillan Dictionary defines "tangible" as "something that you can touch" [6]. When information is stored

in the cloud, or as soft copies, it could not be "touch[ed]". Importantly, the electronic information is also material that represents "disclosing party's Information". Some would rightfully say the author of this policy meant to include electronic information formats as well. Still others might disagree. The semantics in the above policy is not properly communicated due to ambiguity that inevitably exist in the natural language [50]. When ambiguity does exist, the semantics of the privacy policy is open to interpretation by each individual party involved in the data sharing. Such openness may permit parties to use ambiguity to each party's advantage or the openness may act as a deterrence to sharing and processing of data because the party might be afraid to violate a privacy policy. An unambiguous means to express DSAs is necessary to foster a medical research community that enables transparency in data sharing and to facilitate mutual understanding of each party's obligations to the sharing and processing of personal information.

The objective of this research is to develop a semantic model to unambiguously express DSA privacy policies for collaborative medical research. An ontology is a semantic model consisting of explicitly defined concepts and relations within a domain [28]. In this thesis, we built Data Sharing Agreement Privacy (DSAP) Ontology, an ontology of privacy constraints in the collaborative medical research domain by adapting the ontology design methodology described by [29]. To build our ontology, we utilized concepts and relations derived from privacy policies in collaborative medical research DSAs.

In assembling concepts and relations together to form the DSA Privacy Ontology, a number of design decisions were made. Specifically, our ontology was designed with a hierarchical structure, with lightweight expressiveness, with closed world assumption

in interpretation, and with reuse of other ontologies. These design decisions make our ontology flexible to be used by different medical disciplines, extensible to be mapped to other ontologies, unambiguous in expressing semantics to enable mutual understanding of policy terms, and easy to maintain. Furthermore, we demonstrated that it is possible to utilize our ontology to create a Linked Data Log for Transparency, Accountability, and Privacy (L2TAP) audit log [54] to answer privacy competency questions relevant to the medical researcher.

Work to enable unambiguous expression of DSA policies is not new. Others have proposed policy languages [12, 9, 20, 10] to unambiguously express DSA policy terms. Their focus however, is expressing obligation clauses and intellectual property rights for data sharing in general, and not privacy policies for collaborative medical research. With a different focus in mind, their highly structured forms of representing DSAs lack the flexibility and extensibility required to meet the needs of researchers across medical disciplines. Our DSA Privacy Ontology designed in this thesis fulfills this gap.

Thesis Contributions

In this thesis, we make the following contributions:

1. We extensively investigated the privacy aspects of DSAs in health research data sharing (manuscript in progress).

2. We developed an ontology to unambiguously express all categories of privacy policies in DSAs for collaborative medical research.

3. We demonstrated that our ontology can be combined with other existing ontologies to log privacy events in existing audit logging frameworks [54].

4. We demonstrated that through the use of SPARQL, we can query the audit log and inform the medical researcher of their privacy-related obligations as specified by the DSA.

Thesis Organization

This thesis begins with an introduction in Chapter 1 to provide the reader with some background information regarding our thesis work. Next, we review related work in privacy challenges of DSAs in Chapter 2. Then, we discuss the design of our ontology and the reasoning behind some design choices in Chapter 3. Following the design, we evaluate our ontology in Chapter 4. Finally, we conclude and discuss some future avenues for research in Chapter 5.

Chapter 2

Related Work

In this chapter, we discuss the related work to examine different ways in which DSA privacy policies have been expressed. In Section 2.1, we provide an overview of our review methodology. The first step is to understand what privacy constraints are to be expressed in collaborative medical research. In Section 2.2, we review DSAs and data sharing guidelines to identify the types of privacy constraints required in collaborative medical research. Next, we review policy languages that attempt to express DSA privacy constraints (Section 2.3). Finally, we review the DSA document itself to put our work in context of the dynamic lifecycle of the DSA document (Section 2.4).

2.1 Review Methodology

We have reviewed medical research data sharing privacy literature from both the computer science and medical discipline in order to gain a comprehensive overview of privacy challenges in DSAs (Figure 2.1). Investigating in both literature allowed us to

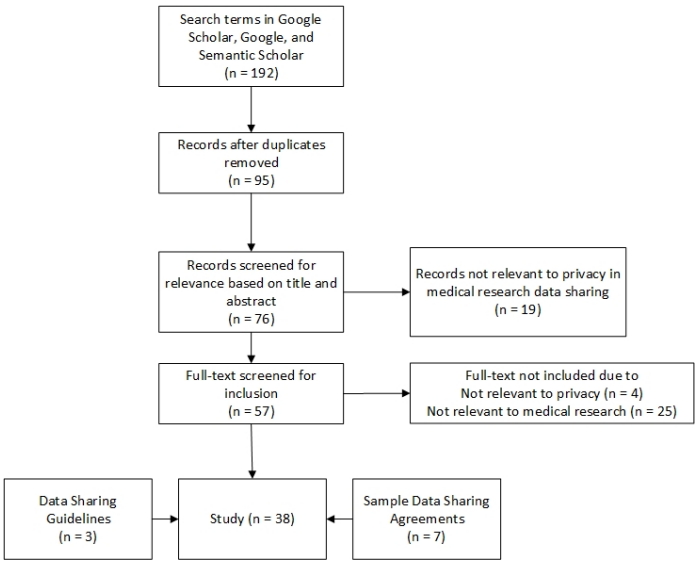

Figure 2.1: DSA Review Method

have an interdisciplinary view on privacy with respect to medical research data sharing. Overall, our search encompassed ACM Digital Library, ISI Web of Knowledge, IEEE Xplore, ScienceDirect, Springer Link, PubMed, Medline, and Scopus.

After duplicate papers were removed, the title and abstracts were scanned to examine for relevance to the review. Criteria for inclusion covered papers that were relevant to determining privacy constraints or expressing privacy constraints. Once the non-relevant records were removed from the survey, a more thorough review of the full-text paper was performed to examine the paper's relevance to our survey. The relevant papers after this stage were included in our survey. A total of 48 records were excluded and 28 full-text articles were extracted for inclusion.

In addition to literature searches in academic publication journals and databases, government data sharing guidelines from research-intensive countries and regions such as US, Canada, and Europe were also examined. Finally, sample data sharing agreements from medical researchers were obtained and included in our review. A total

of 10 medical data sharing guidelines and policies from United States, United Kingdom, and Canada as well as sample data sharing agreements from United States and United Kingdom were included in our review.

2.2 Types of Privacy Constraints in Collaborative Medical Research DSAs

In this section, we review the privacy constraints that are required to be expressed in medical research DSAs. Our review is organized based on the types of privacy constraints expressed: Authorization (Section 2.2.1), Obligation (Section 2.2.2), Administrative (Section 2.2.3), Data Retention (Section 2.2.4), Purpose Definition (Section 2.2.5), and Data Transfer Channel Definition (Section 2.2.6).

2.2.1 Authorization

To protect the privacy of research subjects, medical researchers must specify a limited number of trusted users to receive privileges to access, analyze, delete, or share stored medical data. The trusted users are identified either individually, by belonging to a certain medical research group, or by their affiliations with a certain organization [32]. A DSA is thus required to specify authorization policies to uniquely identify an individual or a group of individuals with specific credentials to gain authorized privileges as authorized users. In so doing, the DSA should answer the "who" part of: Who will have access to the data? Who can share the data? The authorization aspect of DSAs can include temporal constraints such that authorization to the data is only granted during a specified period of validity. The period of validity can be a

fixed date range or relative to an event. In the case of a pre-defined fixed date range, a patient can grant researcher access to personal health data between dates A and B. The case of a relative date can be illustrated with an example of organ donation. In the donation of an organ, a person can consent researchers access to their medical information only after they have deceased [4].

2.2.2 Obligation

On one hand, medical researchers are granted with access privileges to work with personal health data, on the other, medical researchers are also required to fulfill social responsibilities to ensure the privacy of research subjects [32]. For that reason, a DSA contains an obligation aspect to define the behavior or actions that authorized individuals must comply with to refrain researchers from potential unethical handling of personal health data. Obligations can be positive or negative, and can contain a combination of temporal, protection, or sharing constraints [58]. In the form of a positive obligation, a researcher is required to comply to a certain action. For example, medical researchers can be obliged to delete the data after 5 days of its receipt. Prohibitions are negative obligations where a medical researcher is restricted from performing an action. For example, a researcher may be refrained from further sharing received personal health data to minimize the risk of privacy breaches.

2.2.3 Administrative

Administrative aspects of a DSA in medical research contain information relating to the DSA itself [63]. Administrative privacy policies include definition of terms used in the DSA itself, parties involved, and mechanisms to update the DSA. Administrative

information also typically alludes to the environmental context of the DSA [63]. For example, the privacy policy can refer to legal constraints that each medical researcher is to follow based on the medical researcher's jurisdiction.

2.2.4 Data Retention

Privacy breaches in medical research can occur through unauthorized access to stored personal health data [64]. To address unauthorized access issue, the DSA outlines clear instructions on methods of data retention to minimize the impact of personal health data leakage resulting from security vulnerabilities. In medical research however, specifying data retention is particularly difficult as any form of privacy-preserving data manipulation may result in a loss of data value to medical research due to possible loss of data patterns, relationships, or trends [36]. The DSA may highlight specific data transformation requirements to stored data. Data transformation requirements can obligate a non-reversible hash for postal codes to deidentify individuals or reversible transformations such as data encryption [35]. Another useful method to data protection can be achieved through data obfuscation [44]. Alternatively, the data retention requirement can be expressed by referring the reader to external constraints such as legal guidelines. In Ontario, medical researchers are required to follow the strong encryption guideline mandated by the Information and Privacy Commissioner of Ontario [2].

The data retention aspect of DSAs can also have a geographical and temporal constraint [32]. In protecting confidentiality of the data, the DSA may specify where the data must be stored and for how long it can be kept. Temporally, allowing future access to data risks re-identification of individuals due to the possibility that new

data may be available in the future that can help with reidentification - even though the current data is deidentified [38].

2.2.5 Purpose Definition

Unethical use of personal health data risk the potential breach of privacy of research subjects [61]. Therefore, a DSA defines the purposeful use of personal health data. In genetics research for example, defining the purpose of personal health data use is paramount. Genetics data can be used for both good and bad purposes. For good, genetics data can be used to discover genetic predisposition to disease [62], or predict medication responses in personalized medicine research [60]. For evil, genetics data can be used to uniquely identify individuals [36], biological relatives [51], or for commercial reasons [52] without the knowledge or consent of the genetic data contributor. Thus, if identifiable genetics information is leaked, the misuse of information can have important repercussions on data contributor's quality of life such as refusal from a health insurance plan due to being identified to be genetically predisposed to disease. In protecting the privacy of research subjects, the DSA in collaborative medical research characterizes acceptable uses of personal medical data to prevent breaches of privacy [32].

2.2.6 Data Transfer Channel Definition

In planning data sharing, medical researchers identify the channel that they wish to use for data transfer [19]. Such channels should be secure and reliable such that it minimizes the possibility of data leakage resulting in privacy breaches. The DSA formally enforces preplanned avenues of data transfer through defining the accepted

channel of data transfer. Data can be shared via a third party such as an existing database, via journal websites, or other open access repositories. Other means of data sharing can be accomplished through direct transfer from sender to receiver, or sharing within a closed loop of community [19].

2.3 Policy Languages

In medical research, it is of paramount importance to express DSAs clearly and unambiguously to ensure mutual understanding of its terms and constraints between all parties involved. Expression of DSAs among medical researchers have been achieved through natural languages [27]. Though natural language is flexible to express many different constraints, natural language is also highly ambiguous as different people can interpret the same statement in different ways due to the semantic plurality of natural languages. To enable unambiguous expression of the semantics conveyed in DSAs, policy languages have been adapted [9, 12, 40, 33]. However, the policy language's structured representations of privacy policies capture semantics at the cost of flexibility.

The Event-B specification language has been adapted by Arenas and colleagues [9] to express DSA constraints. They represent temporal, protection, and sharing constraints in obligation clauses with linear temporal logic (LTL), a linear sequence of events defined by time. Implementation support for this model is available through the Rodin platform[1], which utilizes the ProB animator and model checker to analyze the DSA for conflicts and also to validate a researcher's actions with obligations

[1] `www.event-b.org/platform.html`

outlined in the DSA. Though interpretation of the model can be achieved unambiguously, pertinent clauses of DSA such as purpose definition could not be expressed with Event-B. In contrast to LTL modelling of DSA, Swarup and colleagues [58] modelled obligation clauses in DSA using distributed temporal logic predicates over data resources, data stores and data flows, which allows reasoning about properties of several components of the system as well as both past and future events. Swarup and colleagues' DSA framework is also able to express penalty if an obligation is not complied with. Electronic implementation of this framework is underway [58]. Just like Event-B, Swarup and colleagues' framework place a focus on obligation and cannot express purpose definition in DSAs.

Another more flexible language, controlled natural language (CNL4DSA) [40] has also been proposed to model DSA. The main purpose of CNL4DSA is to assist in DSA authoring as it is designed with syntax that is more intuitive and readable to researchers [40]. An implementation of CNL4DSA is developed as a web application authoring tool [42]. Following DSA authoring, CNL4DSA is translatable to more formal policy languages such as POLicy-based Proccess Algebra (POLPA) [11] allowing DSA analysis of conflicts with existing implementation frameworks like Maude [18]. While the CNL4DSA has been demonstrated to be useful in ensuring the security in sharing electronic health information [41], CNL4DSA is insufficient in capturing certain privacy requirements in medical research such as expressing domain-specific sharing constraints in the purpose definition. Thus, prior to adopting CNL4DSA, a medical researcher needs to create a list of their domain-specific words for use in authoring a DSA in their specialty to extend CNL4DSA.

Support for the most flexible language, natural language, in DSA expression also

exist. Brodie and colleagues [16] proposed SPARCLE Policy Workbench as a means to parse natural languages into a machine-readable XML expression of policy elements. The biggest advantage of SPARCLE Policy Workbench is that the system can readily parse policies written in natural languages, the popular language that DSAs have traditionally been written in. While the accuracy of SPARCLE Policy Workbench in parsing healthcare privacy constraints is high ($\geq$ 91%), but the accuracy is not perfect, so the user is cautioned that manual work is required to verify the accuracy of parsed policies in using SPARCLE [16].

SecPAL is another policy language proposed to express DSA policies [12]. In addition to permission and obligation clauses, SecPAL4DSA can also incorporate penalties and risks into the expression. SecPAL4DSA's biggest limitation is that it does not allow for the expression of prohibitive clauses due to the nature of the SecPAL policy language [12].

Through this review, we learned that in the development of policy languages, researchers are generally required to identify specific information from their domain for the policy language to reason over [9, 12, 33]. As a result, the policy languages are bound to the domain from which they were created and not flexible to express other constraints across domains [33]. The lack of flexibility make policy languages unable to adapt to specific data sharing scenarios across various medical research domains. Moreover, the machine-oriented syntax of policy languages force policy languages to be highly structured, difficult for a human to read, and difficult to extend. While policy languages can be extensible [33], it is not obvious to a medical researcher how extension of a policy language can be achieved due to poor human readability. A

flexible, extensible, and human-readable way to express privacy constraints in collaborative medical research is therefore needed.

2.4 DSA Lifecycle

In this section, we review the DSA as a dynamic document that undergoes various stages throughout its lifespan (Figure 2.2). Privacy constraints in medical research data sharing are relevant in each stage of the DSA lifecycle.

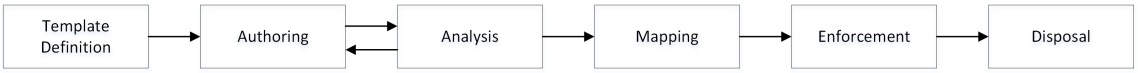

Figure 2.2: The DSA Lifecycle displaying template definition, authoring, analysis, mapping, enforcement, and disposal stage. (Figure adapted from: [39])

2.4.1 Template Definition

DSA templates [27, 7] are typically created by host institutions [32] to define a broad set of constraints in data sharing in general so that the medical researcher can then customize additional constraints to suit their specific field of academic research. For example, a cancer researcher who works with both genetic and clinical data may start from a DSA template for medical data sharing. With this template, the cancer researcher can then add more specific data sharing constraints to suit the needs of genetic data sharing and the needs of clinical data sharing separately. With DSA templates, researchers can be certain not to miss crucial constraints required in DSAs.

2.4.2 Authoring

In the authoring stage, the DSA document itself is written. Since the DSA is the vehicle for communicating and documenting privacy policies among all research collaborators, clear and unambiguous expression of the privacy policies in this stage is then critical to ensure shared understanding of the same policies. While accepted templates are used as a basis for authoring DSAs, researchers are required to customize the template policies for specific data sharing scenarios to suit unique needs. The constant customization of DSAs places consistent demand to DSA authors as the authors constantly struggle to close loopholes in the interpretation of their customized DSA policies.

DSAs have traditionally been authored in natural languages. For a medical researcher, authoring DSAs in natural language is a straightforward task using their familiar word processing software. However, DSAs authored in natural language are prone to misinterpretation - something that a privacy-sensitive medical research domain could not afford. On the other hand, electronic DSA authoring methods supporting policy languages also exist [56]. Although an initial learning curve might be required of the medical researcher to author the DSA using electronic tools, its benefits are enormous. DSAs authored using electronic tools can express privacy policies unambiguously. In cases where DSA tool support exist, the machine-understandable policies can also be automatically enforced [56].

2.4.3 Analysis

In the analysis stage, rules defined in the DSA are checked to ensure that they do not conflict one another. For example, if researcher A is obligated to delete the patient

information after 5 days, then it follows that researcher A must first be authorized to delete the patient information data after 5 days. If researcher A is not authorized to do so in the first place, then this results in a conflict in the DSA. If one or more conflict is discovered, then the DSA returns to the authoring stage to correct all conflicts. The authoring and analysis stages iterate until all conflicts in the DSA are resolved.

2.4.4 Mapping

Electronic enforcement of DSAs require privacy policies to be expressed in machine-understandable ways. Mapping refers to translating the natural languages in DSA into machine-understandable and enforceable security policies. Mapping can be achieved through defined function that maps DSA text to formal, enforceable policy languages [40]. Additionally, mapping capabilities can be directly built into the DSA authoring tool. A built-in mapper can leverage standard web ontologies (ie. SNOMED CT for healthcare) to define domain-specific vocabularies for mapping input to enforceable policy rules [39].

2.4.5 Enforcement

Enforcement refers to formally enacting and enforcing the policies described by the DSA. Enforcement of DSA includes identifying and holding individuals accountable for breaches of the terms and constraints specified in the DSA. Given that the DSA support each step of the data sharing process, then it follows that enforcement procedures must also occur in each step. Privacy constraints in the first step of collaborative

medical research involving patient consent is out of scope for this related work review. For more information regarding patient consent and its enforcement, please refer to [53].

The enforcement of data retention requires auditing metadata pertaining to the creation and storage of data. For example, the data collection date can be used to calculate the age of data in the enforcement of temporal constraints that describe the length of data retention

Enforcement of compliance in delivery, receiving, using, and reusing in data sharing requires correctly identifying trusted users defined by the DSA through authentication mechanisms. The privileges and obligations of the trusted user defined by the DSA must be compared with their activities. Therefore, logging and auditing researcher activity can be used to enforce DSA compliance [38].

In addition to researcher activity, scrutiny of data activity is also required to enforce DSA terms. Data activity, captured as provenance, describes the history of data, what has happened to the data, and who contributed to each version of the data throughout the medical research workflow. There are two levels of workflow provenance: system-level provenance and application-level provenance [14]. System-level provenance refers to the context at which the workflow is executed. System-level provenance relevant to enforcement include geographic location of workflow execution, user identity, and time of data access. Application-level provenance refers to provenance generated by application logic. Information on specific data analysis procedures, and outcomes of data analysis are application-level provenance relevant to the enforcement of DSAs. In enforcing the DSA, a provenance system is required to collect and store both system-level and application-level provenance. Additionally,

another mechanism is required to compare the provenance with DSA privacy policies to identify possible inconsistencies.

2.4.6 Disposal

Finally, a DSA can be disposed of when all contracting parties no longer wish to use this DSA or the DSA has expired. Responsibilities, however, remain on researchers even after the disposal of the DSA. Data retention constraints, for example, institute the researchers to manage the data in a certain way after the ending of the collaboration and the agreement. The DSA can ask the researcher to destroy the data or preserve the data in anonymized formats. In some cases, the researcher may be obligated to share deidentified data through open access channels for publicly-funded research.

2.5 Summary

In this chapter, we reviewed the literature that identified different privacy constraints expressed in DSAs and how these privacy constraints have been expressed. Privacy policies expressed in medical research data sharing were organized into six categories. The large variety of privacy policies communicated in medical research DSAs demand the form of expression for these policies to be flexible. Within the DSA lifecycle, the enforcement of DSA privacy policies require the medical researcher to understand their own obligations. As a result, it is important for the medical researcher to express privacy policies unambiguously during the authoring phase of DSA lifecycle.

The most popular form of DSA privacy policy expression is natural language,

owing to its easy human readability. However, ambiguity in natural language create problems to researchers during DSA enforcement when they become confused about their own obligations. While unambiguous expression of DSA terms have been achieved through policy languages, their lack of flexibility, extensibility, and human readability make policy languages less useful in collaborative medical research. Importantly, policy languages are also unable to express all aspects of privacy constraints in medical research DSAs. In this thesis, we developed an ontology to express privacy-related policy terms in collaborative medical research DSAs with the intention of addressing the flexibility, extensibility, and human readability issues.

Chapter 3

DSAP Ontology Design Method

In this chapter, we adapt the ontology design methodology outlined by [29] to develop our DSAP ontology. We begin by examining some competency questions that our ontology is required to answer in Section 3.1. Next, we explain the ontology design in Section 3.2. Then, we devise concepts and relations for our ontology in Section 3.3. Finally, we discuss our DSAP ontology in Section 3.4.

3.1 Competency Questions

Competency questions are questions that our ontology is required to answer [29]. Different stakeholders in the collaborative medical research environment will ask different competency questions. Since our ontology was designed specifically for medical researchers, therefore, our competency questions were devised based on the point of view of the medical researcher. We first describe a motivating scenario where our ontology can be used (Section 3.1.1). Next, we devise competency questions for our scenario that capture the needs of a medical researcher (Section 3.1.2).

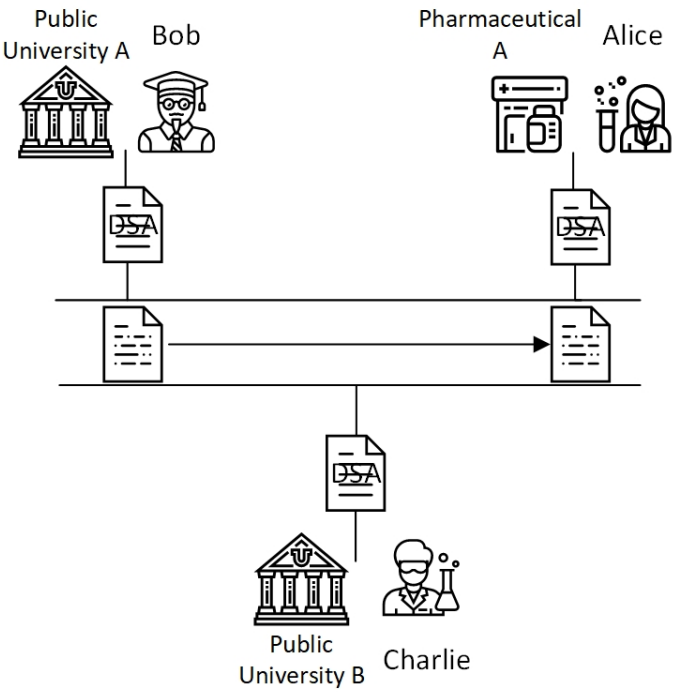

Figure 3.1: Motivating Scenario

3.1.1 Motivating Scenario

The real-world basis for the application of our ontology is illustrated with a motivating scenario [29]. The motivating scenario sets the requirements of our ontology through identifying informal competency questions that our ontology is required to answer. In Figure 3.1, Alice, Bob, and Charlie are all researchers. Alice is from Pharmaceutical A, a privately held pharmaceutical, while Bob and Charlie are both from publicly funded research universities - Public University A and Public University B respectively. Bob currently holds a genetics dataset which both Alice and Charlie would like to request for access. The researchers are bound by a DSA, which contain a privacy policy that limits the genetics dataset to be only shared with publicly funded research. In this complex data sharing environment, Alice, Bob, and Charlie

are helplessly confused about their obligations set by the DSA and thus unable to determine exactly what they are permitted to and not permitted to do. The medical researchers' confusion of responsibilities is important because mishandling of patient data is serious and can lead to loss of credentials of the researcher, even if the act is done with good intentions [45].

3.1.2 Informal Competency Questions

Since our ontology was designed for medical researcher, therefore, the informal competency questions were devised from the perspective of the medical researcher. The researchers' lack of mutual understanding and confusion about DSA's specified proper handling of patient data leads to informal competency questions that our ontology needs to answer. The questions are informal because they are not yet expressed by vocabulary in our ontology [29]. The informal competency questions were broken down into two main categories: prohibitive and prescriptive.

Prohibitive Questions

Prohibitive questions asks the ontology whether or not a researcher or another actor in the research pipeline is permitted to perform a certain task. It demands a simple yes or no answer from our ontology. Examples of prohibitive questions include:

1. Is a specific researcher allowed to access the dataset?
2. Is a specific researcher allowed to share the dataset with another researcher? Or with a third party?
3. Is a specific researcher allowed to disclose the data?

Prescriptive Questions

Prescriptive questions asks the ontology to identify a certain piece of information from the DSA. These questions demand a clear answer from information directly conveyed by the DSA. Examples of prescriptive questions include:

1. What are my obligations if I access this dataset?

2. What are my obligations after I receive this dataset?

In summary, both prohibitive and prescriptive informal competency questions are closed questions. They both demand a specific, and direct, answer from our ontology. The short, focused answers leave the researcher without ambiguity, thus improving understanding of the DSA.

3.2 Ontology Design

Building the ontology involve piecing together devised concepts and relations. There are many ways this can be achieved, and there is no one correct way to build an ontology [21]. As a result, we first consider some design requirements of our ontology (Section 3.2.1). Next, we describe the design decisions that were made to satisfy our design requirements (Section 3.2.2).

3.2.1 Design Requirements

Our ontology was designed specifically for collaborative medical research use. In health data sharing, stakeholders include data contributors, privacy auditors, and medical researchers. Data contributors can be patients or research volunteers. Their

main interest is the preservation of privacy of their contributed personal data throughout the medical research process. Privacy auditors are concerned with the compliance of medical researchers with privacy policies governing the use of personal data. Meanwhile, medical researchers need to communicate privacy policies regarding the treatment of data with their collaborators. Additionally, the researchers are also concerned with their own compliance with privacy policies specified by their collaborators. To achieve compliance, the researchers first need to understand their own obligations and privileges with the health data. Our ontology was designed for use by medical researchers. Therefore, we derive the following design requirements from the point of view of medical researchers.

Unambiguous. For medical researchers, the protection of data contributor's privacy is communicated through privacy policies in DSAs. Importantly, researchers need to understand their own obligations with the proper handling of data as specified in DSA. The mutual understanding of privacy policies between researchers can be achieved through unambiguous expression of privacy policies.

Flexible. The medical researcher should be able to express all privacy policies relevant to their collaboration using our ontology. One challenge to achieving the expression of all privacy policies is that different medical research disciplines require different privacy policies to be expressed. For instance, with respect to data retention policies, public health researchers may only work with aggregate and anonymized data. Therefore, public health research DSAs may simply require the researcher to dispose the data after a year. On the other hand, genetics researchers work with more personal data. As a result, genetics research DSAs may specify more strict policies to

include an encryption requirement in addition to disposal requirement in their data retention privacy policy. Therefore, our ontology will need to be flexible to be able to express all different privacy policies.

Extensible. Our ontology should be extensible to work with other linkable ontologies. With an extensible ontology, the medical researcher will have the freedom to extend our ontology with other ontologies with more or less expressive power. When extending our ontology with ontologies with more expressive power, the medical researcher will be able to use more vocabulary to express privacy policies to more details.

Human Readable. Our ontology needs to be easy for the researcher to read and understand. With human readability, the medical researcher can author DSA privacy policies directly with our ontology. Therefore, a parser would not be required to convert the privacy policies into machine-understandable format for enforcement. Ontology is machine-readable.

3.2.2 Design Decisions

In this section, we outline the design decisions made to achieve the design requirements specified in the previous section. First, a hierarchy design was utilized. Second, the expressiveness of the ontology was designed to be lightweight. Third, a closed world assumption was applied to our entire ontology. Finally, other ontologies were used by mapping mutually overlapping concepts.

Hierarchy

Our ontology was organized into a hierarchical structure. One key advantage to having this structure is inheritance. The child of any parent inherits the definitions and relations of the parent. As a result, the semantics already defined in the parent does not need to be redefined in the child. The property of inheritance greatly simplifies the construction of our ontology.

Easy human readability is another advantage for having a hierarchical design. One can think of a parent of a child as a category that the child falls in. When reading a hierarchy, the medical researcher can intuitively organize the concepts and relations based on parent-child relationships. The simple and self-explanatory organization of concepts in a hierarchy greatly simplifies the complexities of DSAs and enables the user to quickly get started with using the ontology without technical knowledge.

The hierarchy structure also promotes extensibility of the ontology, important for medical researchers of different domains to add domain-specific vocabularies to the ontology. Concepts within the hierarchy can be extended without the need to disrupt the rest of the ontology. We illustrate the extensibility of a hierarchical design with the `Funding` hierarchy in Section 3.3.2.

Lightweight

Another design decision we made was to design a lightweight ontology over fully formal ontology. Lightweight ontologies expressed using OWL Lite [43], RDFS [1], and RDF [3], with limited expressiveness has multiple advantages over highly formal ontology design using first order logic. The most important property of lightweight ontology is its decidability. Lightweight ontologies are also easy to understand by a

non-technical person such as a medical researcher and are typically used either for description or classification purposes [26]. On the other hand, fully formal ontologies, while highly expressive, are composed of detailed relationships between concepts [49]. Fully formal ontologies are more difficult to build and to understand due to the added complexities introduced by detailed relationships between concepts. Fully formal ontologies are typically built only when absolutely necessary in software engineering where the software requires semantic interoperability or automated information extraction from text [49]. Similarities also exist between the two design choices. Both lightweight and fully formal ontologies enable unambiguous expression of concepts and relations [49].

Another distinct advantage of lightweight ontology is its extensibility and flexibility. The simplicity of a lightweight ontology allows the ontology to be easily extended with other linkable ontologies. Similarly, the plain concepts defined in a lightweight ontology enables flexibility for the medical researcher to define their own vocabulary to express different privacy policies. The extensible and flexible property of our lightweight ontology is illustrated with the `Purpose` hierarchy in Section 3.3.2.

Closed World Assumption

To enhance decidability of our ontology, our ontology was limited to only closed world assumption (CWA). CWA states that what is not true in the ontology are to be considered as false and only what is explicitly expressed in the ontology is considered true [31]. The opposite of CWA is open-world assumption (OWA). In OWA, a statement may be true even if the statement is not explicitly stated as true [31]. The uncertainty in OWA may cause our ontology to become undecidable,

a situation we want to avoid. As a design decision, the interpretation of our ontology was therefore limited to CWA.

Reuse Other Ontologies

Finally, we chose to reuse other ontologies as part of our design. Ontology reuse has many benefits and is encouraged in the ontology design literature [37]. Advantages of reusing existing ontologies include saving the labour involved in creating ontology from scratch, bettering the quality of ontology, and reducing ontology maintenance overhead [37]. We thus reviewed and reused other ontologies where appropriate. The reuse of other ontologies is discussed in more detail in Section 3.3.2.

3.3 Concepts and Relations

In this section, we devise concepts and relations to provide the vocabulary and structure for our ontology. The first step is to acquire domain knowledge (Section 3.3.1). In Section 3.3.2, we describe the ontology concepts and relations.

3.3.1 Acquire Domain Knowledge

Acquiring domain knowledge refers to gaining knowledge in the domain of medical research data sharing and understanding privacy constraints expressed in DSAs used by collaborative medical researchers. To achieve that, a literature survey of data sharing agreements and data sharing guidelines related to collaborative medical research was conducted. A total of 10 medical research data sharing guidelines and 9 medical research data sharing agreements were obtained from Google search and from

medical research collaboration. The obtained data sharing guidelines and agreements included from both Europe and North America and represent a range of medical research topics for example cancer research and genetics research.

There are two factors to consider to ensure the collected agreements and guidelines are representative of the privacy requirements in collaborative medical research across the world - 1) geographic diversity and 2) research topic diversity. First, since various jurisdictions have their own privacy regulations to the handling of personal information, therefore, the collected agreements and guidelines must encompass jurisdictional requirements where collaborative medical research are carried out. Since Europe and North America together represent a majority of biomedical research productivity across the world [46], therefore, the agreements and guidelines obtained are representative of collaborative medical research geographically. Second, though our review included multiple medical research topics, it does not guarantee that all of collaborative medical research is represented. However, data sharing guidelines and agreements in human genetics research was thought to be representative of expression of privacy constraints in the collaborative medical research field. We think this is true for two reasons. First, genetics data hold good scientific value when shared, driving motivation to share genetic data [17]. Second, genetics data is some of the most intimate and personal types of data shared in collaborative medical research [59], so privacy is more of a concern in the sharing of genetics data. For these reasons, genetics research data sharing guidelines and data sharing agreements are an ideal source to review in understanding privacy constraints required in collaborative medical research data sharing.

The process of devising concepts and relations as well as building the ontology is

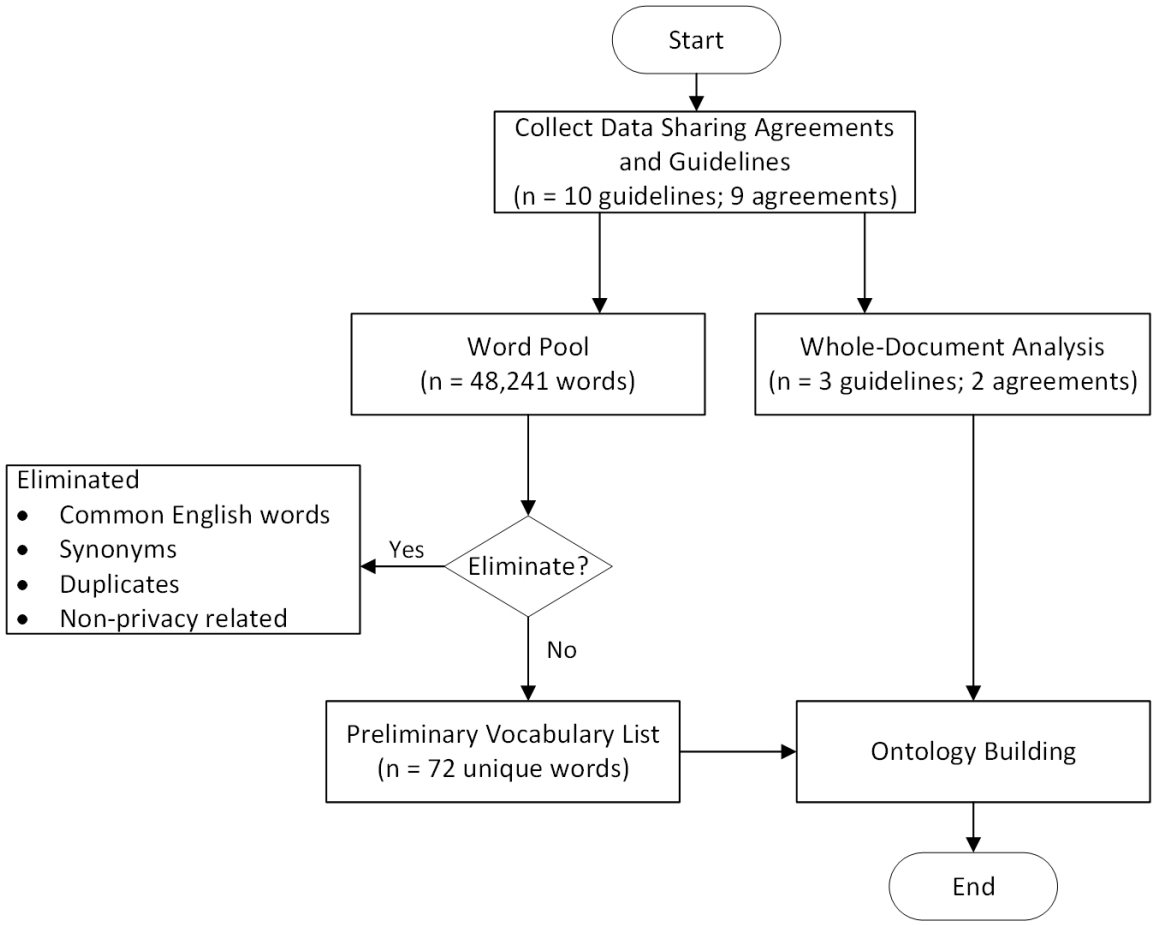

Figure 3.2: Methodology in Building the Ontology

described in Figure 3.2. First, a word pool was created by parsing the data sharing guidelines and data sharing agreements through a word counter [5]. Within the 10 guidelines and 9 agreements collected, a total of 48,241 words were parsed. To eliminate common words in the English language and filler words for sentences irrelevant to the medical research domain such as "the", "a", "on", words that contained 3 letters or less were excluded in our word count. To further narrow down and focus on words that are important to identifying common privacy constraints in the guidelines and agreements, only words that appeared at least four times in the same document

were included. Next, after synonyms and duplicates were removed, 72 unique words were identified and were used to create a preliminary vocabulary list from which our ontology was constructed (Table A.1).

A total of 5 data sharing agreements and guidelines were selected for detailed whole-document analysis. Selection process ensured that this subset of agreements and guidelines represent the pool of 19 obtained. The selected documents analyzed include cancer research and genetics research from both Canada and Europe. In detailed analysis, each policy was categorized as privacy related and non-privacy related. The privacy related clauses were analyzed for required concepts and relations to be expressed. Non-privacy related policy clauses excluded in our analysis were related to other aspects of collaborative medical research such as intellectual property rights and authorships.

3.3.2 Devise Concepts and Relations

Table 3.1 shows the main concepts and relations organized by type of privacy constraint expressed (Section 2.2) derived from our detailed whole-document analysis. Authorization, Obligation, and Administrative privacy constraints require an `Agent`, or a person, to be identified. This person assumes some sort of `Role` (Table 3.2), which defines the function of the individual within an `Organization`, within the data sharing pipeline, or within the `Agreement`. The assignment of `Role` to an `Agent` is captured with the `hasRole` relation.

The assignment of `Role` is important in DSAs because `Action` required, prohibited, or authorized to be performed by the individuals in the collaborative medical

Table 3.1: Main Concepts and Relations

Privacy Constraint Type	**Concepts**	**Relations**
Authorization (Subsection 2.2.1)	Agent Role Action Date	hasRole authorizedAction hasDate
Obligation (Subsection 2.2.2)	Agent Role Action Date Dataset Database	hasRole requiredAction authorizedAction prohibitedAction requiredStorage hasDate
Administrative (Subsection 2.2.3)	Agent Role Action Organization Jurisdiction Subject Consent ResearchStudy Funding	requiredAction authorizedAction prohibitedAction partOfJurisdiction requiredJurisdictionCompliance consentsTo requiredConsentCompliance associatedWithJurisdiction hasResearchStudy hasFunding collaboratesWith
Data Retention (Subsection 2.2.4)	Data Dataset Database Date Agreement	requiredStorage hasDate
Purpose Definition (Subsection 2.2.5)	Role Purpose Organization ResearchStudy	eligiblePurpose hasResearchStudy aboutData
Data Transfer Channel Definition (Subsection 2.2.6)	Role Action Data	requiredAction aboutData
Other	Policy Penalty PolicyCondition	hasPenalty hasPolicyCondition

research environment are based on their `Role` assignment. The semantics in the relations between `Role` and `Action` are captured with `requiredAction`, `prohibitedAction`, and `authorizedAction`. First, `requiredAction` denotes that the `Action` must be performed by the corresponding `Role`. Second, `prohibitedAction` refers to that the `Action` is not permitted to be performed by the corresponding `Role`. Finally, the `authorizedAction` relation indicates that the `Action` is permitted to be performed by a specific `Role`, but the `Action` is not required to be performed.

In addition, the `Action` and `Agreement` concepts in Authorization, Obligation, and Data Retention constraints all have a temporal component. This time representation is captured with the `Date` concept. The temporal semantics of `Action` and `Agreement` are in turn captured with the `hasDate` relation with the `Date` concept.

The original personal health information collected from Research `Subject`, or the patients, are captured as the `Data` concept. The aggregate of `Data` forms the `Dataset`. Certain Obligation and Data Retention constraints specify how the `Dataset` should be stored upon receipt. The point of storage is captured as the `Database` concept. The relation `requiredStorage` captures the requirement to store `Dataset` in a particular way or form in the `Database`.

Access to, or sharing of this personal health information usually requires a defined `Purpose`. The anticipated goal or outcome for an `Action` is captured through the `Purpose` concept. The `eligiblePurpose` relation captures the semantics of specific allowed `Purpose` for the `Action` to be performed. If the eligible `Purpose` refers to an `Action` on a `Data` or a `Dataset`, the `aboutData` relation can be used to capture this.

While DSA specify constraints to the access or sharing of personal health information that need to be met, legal requirements governing the handling of health

information also need to be satisfied. Legal obligations come from the legal system in rule in the area which the researcher operates. This is captured as the `Jurisdiction` concept. The `partOfJurisdiction` relation identifies the `Jurisdiction` associated with the `Oriqanization` that an `Agent`, or a researcher, is part of.

Finally, certain policies within the DSA may have a condition under which the `Policy` comes into effect. We capture this with the `PolicyCondition` concept and the `hasPolicyCondition` relation. In addition to `PolicyCondition`, there are also policies that contain clauses to represent the punishment for behavior inconsistent with the `Policy`. We capture this semantic with the `Penalty` concept and provide the `Policy` the `Penalty` with the `hasPenalty` relation.

Below, we describe `Role`, `Action`, `Date`, `Purpose`, and `Funding` concepts along with their relations in more detail to capture enough semantics to enable meaningful expression of DSA privacy policies with our ontology.

Role

Table 3.2: Various roles specified by the DSA

Type of Role	Description	Example
`PartyRole`	Role as a party assumed by a person or group of people pertaining to the DSA.	Third Party
`DataSharingRole`	Role assumed by a person or group of people with respect to the data sharing pipeline.	Data Sender
`OrganizationRole`	Role assumed by a person or group of people within an organization.	Medical Researcher

To enable the expression of various types of roles specified in the DSA, the `Role`

concept was extended. There are three major types of `Role`s expressed in DSAs (Table 3.2): `PartyRole`, `DataSharingRole`, and `OrganizationRole`.

`PartyRole` captures the participation of person or group of people involved in the DSA. It identifies the relationship of each party with respect to each other in the medical research collaboration. Examples of `PartRole` include `AgreeingParty` to capture the person or group of people directly involved in the medical research collaboration, or `ThirdParty` to capture the person or group of people indirectly or not principally involved in the medical research collaboration as defined by the DSA. There can also be representatives of each type of party that can act on behalf of the represented party. This is captured as the `Representative` concept. The type of party being represented is captured through the `representativeOfAgreeingParty` and the `representativeOfThirdParty` relation.

`DataSharingRole` captures the intended or assigned function of a person or group of people in the collaborative research data sharing pipeline (Figure 3.1). Examples of `DataSharingRole` are `DataSender` to capture the person involved in the sending of data, `DataRequestor` to capture the person making the data request, `DataReceiver` to capture the person receiving the data, and `DataSubject` to capture the person donating the data.

`OrganizationRole` captures the intended or assigned function of a person or group of people in an `Organization`. They can be a physician who sees patients captured as the `Clinician` concept, or a researcher who only conducts research captured as the `Researcher` concept, or an auditor who make privacy audits for that organization captured as the `Auditor` concept.

The `Role`s were organized into a hierarchy. In Figure 3.3, `PartyRole` is a child of

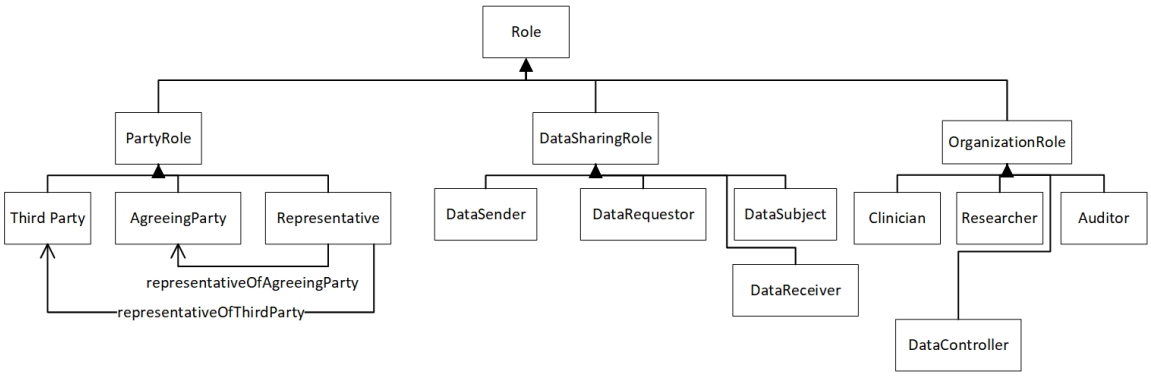

Figure 3.3: DSA Privacy Ontology Role Hierarchy

`Role`. From the `Role` concept, we know that `Role` is an intended or assigned function of a person or group of people. By extension, we therefore also know that `PartyRole` is some sort of function assumed by an `Agent`, so it does not need to be repetitively defined. Similarly, `DataSharingRole` and `OrganizationRole` (Figure 3.3) also inherit the definition of `Role`. The same inheritance property is true for the `Action` hierarchy (Figure 3.4).

Action

The `Action` concept was extended to capture the various types of actions specified in the DSA. Three major types of actions were found in the DSA (Table 3.3), `DataAction`, `ResultsAction`, and `AgreementAction`.

The `DataAction` concept captures the action that could be performed on a data or dataset. For example, data or dataset could be shared, this action is captured through the `ShareAction` concept. Other actions that can be performed on data or dataset as specified by the DSA include derive results (`DeriveResultsAction`), protect (`ProtectAction`), receive (`ReceiveAction`) , store (`StoreAction`), copy (`CopyAction`),

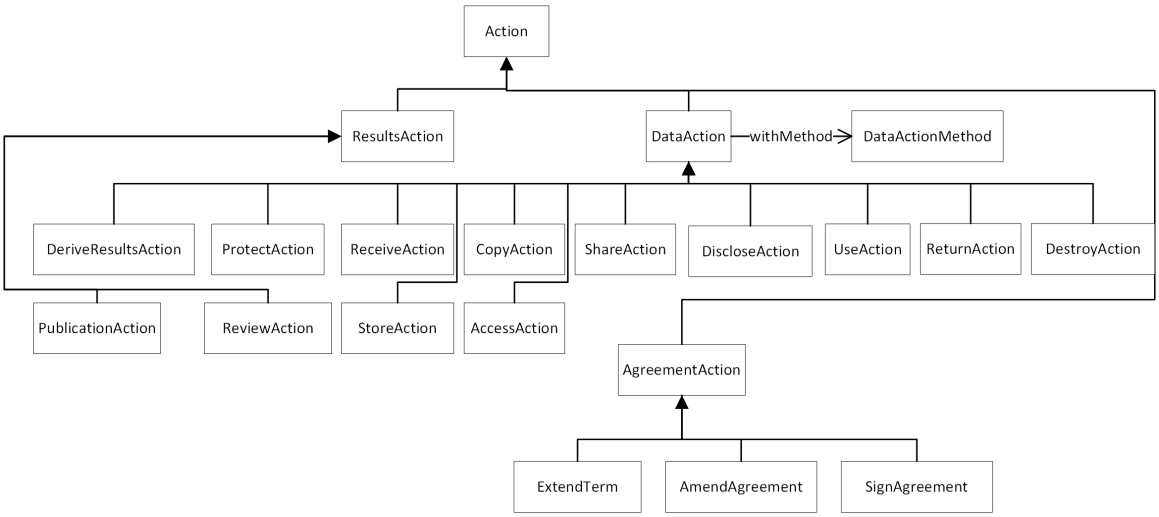

Figure 3.4: DSA Privacy Ontology Action Hierarchy

access (`AccessAction`), disclose (`DiscloseAction`), use (`UseAction`), return (`ReturnAction`), and destroy (`DestroyAction`).

The `ResultsAction` concept captures action that could be performed on the `Result`, which is an outcome of an analysis derived from data or dataset. The `derivedResult` relation can be used to specify the `Data` or `Dataset` from which the `Result` was derived from. Examples of actions that could be performed on a `Result` include publish (`PublicationAction`) and review (`ReviewAction`).

The `AgreementAction` concept was used to capture action related to the DSA itself. Action that could be performed on the DSA include extend the term (`ExtendTerm`), amend (`AmendAgreement`), and sign (`SignAgreement`).

Date

The temporal aspects of `Action` and `Agreement` concepts require the capturing of different types of time-related constraints in the DSA. These time-related constraints come in the form of key dates.

Table 3.3: Various actions specified by the DSA

Type of Action	Description	Example
`DataAction`	Action that can be performed on a data or dataset.	Receive Data
`ResultsAction`	Action that can be performed on results derived from data or dataset.	Publish Results
`AgreementAction`	Action that can be performed on the DSA itself.	Amend Agreement

The `Date` concept was therefore extended to capture the variety of dates specified in DSAs. Dates expressed include the date at which data was collected from the research subject (`DataCollectionDate`), the date of data release (`DataReleaseDate`), the date at which the data should be destroyed (`DataDestroyDate`), the date at which the agreement comes into effect (`EffectiveDate`), the corresponding date at which the agreement ends (`EndDate`), the date at which the agreement was signed (`SignatureDate`), the date at which the research study is to be completed (`ResearchStudyEndDate`), and finally the earliest permitted date for the publication of the results (`PublicationDate`).

Purpose

As mentioned before, the `Purpose` concept captures the goal or the initial intention for data access or sharing. While such definition permit virtually unlimited possibilities for `Purpose`, we found that only two types of purposes would be required to be captured in order to unambiguously express the semantics of medical research DSA policies - `ResearchPurpose` and `CommercialPurpose`. The `ResearchPurpose` concept captures the initial intent of the data access or sharing purely for academic

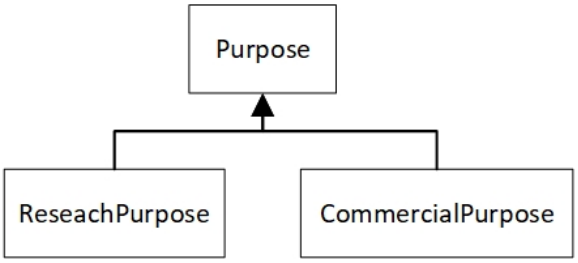

Figure 3.5: DSA Privacy Ontology Purpose Hierarchy

or research reaons. On the other hand, The `CommercialPurpose` concept captures the goal for data access or sharing as related to business activities to produce goods or services for selling or making a profit.

In our case, the flexible and extensible nature of lightweight ontology can be illustrated by the `Purpose` hierarchy (Figure 3.5). While the purpose of medical research project can be many, ranging from drug discovery to public health reporting, we have intentionally kept our `Purpose` hierarchy lightweight. The parent `Purpose` concept only consists of two child, `ResearchPurpose` and `CommercialPurpose`, with no additional breakdown of the two (Figure 3.5). Being general and lightweight, medical researchers would be able to use our ontology as is while also having the flexibility to extend the ontology with their own, more specific, `Purpose` concepts if they wish. Flexibility is necessary especially when the purpose to data access or sharing may sometimes be unclear even to medical researchers. The very goal of collaborative data sharing in medical research is to discover the unknown within the dataset. Thus, to a medical researcher, it is very difficult to tell upfront, before accessing the data, what the "purpose" to data access is. They may access the data for the purpose of finding causation to a disease, but through data analysis, they may discover other insights within the data. In fact, accidental discoveries are not uncommon in the

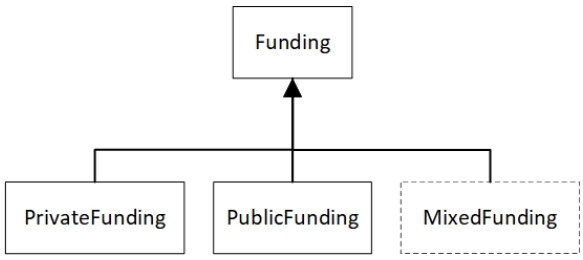

Figure 3.6: DSA Privacy Ontology Funding Hierarchy

medical field - Penicillin, Warfarin, and Nitrous Oxide are examples of medical discoveries where the use of material or data was not intended from its original purpose. Thus, our lightweight ontology supports the type of flexibility inherent in the process of discoveries in medical research where `Purpose` may be difficult to define before data access.

Funding

The source of financial support, captured as the `Funding` concept, to carry out the research is also relevant to privacy in data sharing. Two main streams of funding were identified through our detailed analysis of DSAs and data sharing guidelines - `PrivateFunding` and `PublicFunding`. The `PrivateFunding` concept captures funding of research solely from individual persons or companies and not from the government. On the other hand, the `PublicFunding` concept captures research funding solely from the government. Policies within DSAs and data sharing guidelines permit certain actions on datasets based on the type of funding used to support the research study. The `permittedFunding` relation is used to capture the semantics of the associated `Funding` that is permitted on an `Action` concept.

The extensible nature of the hierarchy design can be seen from the `Funding` hierarchy (Figure 3.6). In our analysis of DSA policies, only `PrivateFunding` and `PublicFunding` concepts were required to be expressed. The general `Funding` concept was not required but was added to be parent of `PrivateFunding` and `PublicFunding` specifically enable a hierarchy design to our ontology. The addition of `Funding` concept as the parent enables extensibility of our ontology. For instance, some medical research are funded through public-private partnerships [48], which is neither solely privately funded nor solely publicly funded, so we require a new `MixedFunding` concept in our ontology to properly express a mixed funding model (shown in dotted box in Figure 3.6). To add `MixedFunding` concept to our ontology, we can easily extend the `Funding` concept by having `MixedFunding` as a child to thus inherit the semantics of `Funding` without disrupting the rest of our ontology.

Other Ontology Reuse

We reviewed other existing ontologies for concepts and relations for reuse. We found that existing ontologies also capture some of the concepts related to privacy in DSAs. Specifically, provenance (PROV) ontology (`https://www.w3.org/TR/prov-o/`), friend of a friend (FOAF) ontology (`http://xmlns.com/foaf/spec/`), and Timeline ontology (`http://motools.sourceforge.net/timeline/timeline.html`) were good candidates for reuse. We mapped overlapping concepts between our ontology and the existing ontologies to take advantage of the semantics already expressed in these existing ontologies.

PROV Ontology. The PROV Ontology captures the concepts and relations to represent the provenance information of data sharing and processing [13]. While this

information is not useful to express policy statements in DSAs, they can be used during the enforcement of the DSA as data sharing happens (Figure 2.2). Mapping of DSA Privacy ontology onto the PROV ontology therefore enables interoperability between the static DSA document and the dynamic aspect of DSA during enforcement (Table 3.4).

Table 3.4: Mapping of overlapping concepts between DSA Privacy Ontology and PROV Ontology

DSA Privacy Ontology	**PROV Ontology**
Dataset, Data, Result	Entity
Agent	Agent
Organization	Organization
Action	Activity

FOAF Ontology. The FOAF Ontology captures the semantics in the linking of people [22]. In terms of collaborative medical research, it is useful to express the semantics in the relationship between research collaborators. Overlapping concepts of DSA Privacy Ontology and the FOAF Ontology are presented in Table 3.5. The properties `member`, `knows`, `currentProject` defined by FOAF are particularly useful as they capture the semantics in the relationship between `Researchers` with their `Organization`, with each other, and also between the `Researchers` with the `Research Projects` that the `Researchers` are undertaking.

Timeline Ontology. The Timeline Ontology captures the semantics in date and the continuousness of time [47]. Many of the policy clauses in the DSAs that we have reviewed contain a temporal component. The timeline ontology is then useful

Table 3.5: Mapping of overlapping concepts between DSA Privacy Ontology and FOAF Ontology

DSA Privacy Ontology	**FOAF Ontology**
Agent	Agent
ResearchStudy	Project
Organization	Organization

to capture the temporal aspect of DSA privacy policies. By expressing the `Date` hierarchy of our ontology as `xsd:date`, we can enable extension of our ontology to use the properties defined in the Timeline Ontology. The Timeline Ontology properties useful in DSA privacy are `before`, `after`, and `atDate`. These properties all have a Range of `xsd:date` with an arbitrary Domain. By reusing the Timeline Ontology, we can enable the semantics of time to be expressed without creating new concepts and relations.

3.4 DSAP Ontology

Adapting the methodology described in [29] and after applying our own ontology design decisions in organizing the concepts and relations, we arrived at our DSA Privacy Ontology (Figure 3.7). Four concepts were extended to form hierarchy structures to sufficiently express privacy-related medical research DSA policies. Specifically, the `Role`, `Action`, `Purpose`, and `Funding` concepts were extended and their hierarchy are shown in Figures 3.3, 3.4, 3.5, and 3.6 respectively. In combining the concepts together, new relations were required to capture the semantics in the link between the concepts. For instance, when the `Role` hierarchy is combined with the `Action` hierarchy, the relations `prohibitedAction`, `requiredAction`, and `authorizedAction`

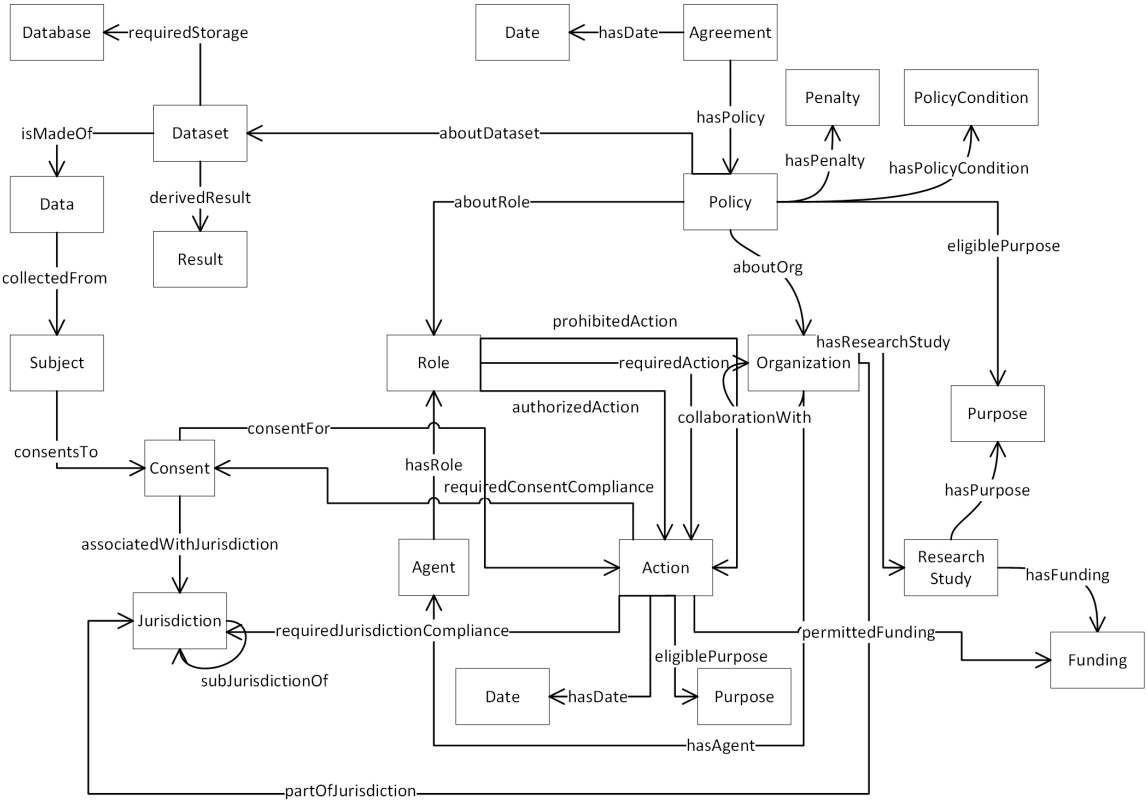

Figure 3.7: DSA Privacy Ontology

were added to capture the prohibited, required, and authorized `Action`s of `Role` respectively. When the `Purpose` hierarchy was incorported to our DSAP ontology, we added the relation `eligiblePurpose` from `Action` concept to capture the permitted `Purpose` of `Action`. Similarly, when the `Funding` hierarchy was combined into our DSAP ontology, a new relation, `permittedFunding`, was added to capture the permitted `Funding` that is associated with a specific `Action`.

Validity Constraints

The assembly of previously defined concepts and relations into an ontology also reveal new insights into conflicts that need to be addressed. For example, the ontology currently permits a role to hold an authorized action as well as a same prohibited action. In the medical research data sharing environment, while a role can be authorized to perform a certain action, naturally the same role cannot be prohibited to perform the same action. To address such conflicts, limitations need to be imposed onto the concepts and relations in the ontology to reflect reality.

Table 3.6: Sample DSA Privacy Ontology Validity Constraints

Constraint	**Example Affected Concepts and Relations**	**Description**
Cardinality	Agent hasRole Role	The minimum and/or maximum of instances of each concepts that can be related per DSA.
Contradiction	authorizedAction requiredAction prohibitedAction	The concept and relation instances that cannot be defined together in the same DSA.

Two types of constraints were added to the DSA Privacy Ontology (Table 3.6): cardinality and contradiction. The cardinality constraint limits the number of relations that instances of each concept can take. In our ontology, each `Agent` can only assume at most one of each `PartyRole`, `DataSharingRole`, and `OrganizationRole` (Figure 3.3) through the `hasRole` relation. For example, an `Agent` of the `AgreeingParty` can also take on a `DataSender` role, but not a `ThirdParty` role simultaneously. That is because both `AgreeingParty` and `ThirdParty` are `PartyRole` while

`DataSender` is a `DataSharingRole`. The other type of constraint, contradiction constraint, is imposed on the ontology to ensure the concepts and relations make logical sense. As previously mentioned, it does not make sense for a role to be authorized and prohibited to perform the same action. As a result, `authorizedAction` and `prohibitedAction` are disjoint, and cannot be used together. Similarly, `requiredAction` and `prohibitedAction` are also disjoint.

3.5 Summary

The DSA Privacy Ontology (Figure 3.7) was developed through adapting the methodology described by [29]. Design decisions made in the design of our ontology were consistent with the requirements to support collaborative medical research. The hierarchy design, limited expressivity, closed world assumption interpretation, and reuse of other ontologies enable our ontology to be flexible, extensible, and human-readable. These qualities make our ontology more suitable than the highly structured policy languages to express privacy-related DSA policy terms in collaborative medical research. Therefore, our ontology fulfills the gap in literature. In the next chapter, we will evaluate the ontology built in this chapter.

Chapter 4

Ontology Design Evaluation

In this chapter, we evaluate our ontology in three ways. First, we show that our ontology is capable of expressing DSA privacy policies unambiguously by translating sample DSA excerpts in natural language to RDF format (Section 4.1). Next, we evaluate the ability for our ontology to support collaborative medical research by demonstrating the use of our ontology in a data sharing scenario (Section 4.2). Finally, we evaluate the ability of our ontology to answer privacy competency questions relevant to the medical researcher (Section 4.3).

4.1 Translating DSA Privacy Policy to RDF

The Resource Description Framework (RDF) is a standard way to serialize information derived from an ontology in the form of RDF triples [3]. Each RDF triple is composed of 3 elements, a subject, a predicate, and an object, written in the triple form as subject-predicate-object. The subject and object are concepts in our ontology while the predicate denote the relation between the subject and the object concepts.

Since each element of the triple are explicitly defined in the ontology, they form an unambiguous expression. Thus, to evaluate our ontology's ability to express DSA privacy policies unambiguously, we translate all different types of DSA privacy policies (Section 2.2) from DSAs written in natural language to RDF triples serialized in Turtle syntax [23]. We use the prefix `dsap` to denote the namespace for our DSA Privacy Ontology. Any terms without a prefix indicate an example.

Table 4.1: Expressing Authorization Privacy Policy in RDF Triples

DSA Excerpt	**RDF Triples**
Only researchers are permitted to access personal information.	1 `:univResearcher a dsap:Researcher.` 2 `:accessData a dsap:AccessAction.` 3 `:univResearcher dsap:authorizedAction :accessData.`

We start with translating authorization policies. Authorization policies are straightforward to express in RDF triples. The translation of authorization policies simply requires the relation `authorizedAction` as the predicate to indicate the relationship between a `Role` and an `Action` concept. The example in Table 4.1 line 3 shows the authorization of a `Researcher` role (line 1) to perform an access action (line 2). With closed world assumption, the word "only" in the DSA excerpt (Table 4.1) is included in the interpretation of the RDF triple statements. It is therefore not necessary to define that all other roles are not authorized to perform the same action. Similarly, in the likely event that the dataset shared by the medical researchers contain personal information, it is not necessary to define "personal information". However, if that is not the case, the DSA author may elect to specify that the access action refers to a particular dataset containing personal information using the triple `dsap:AccessAction dsap:aboutDataset dsap:Dataset`.

Table 4.2: Expressing Obligation Privacy Policy in RDF Triples

DSA Excerpt	**RDF Triples**
Any Information obtained shall not be copied or shared with other researchers who are not a party to this Agreement.	1 :univResearcher a dsap:Role; 2 dsap:prohibitedAction :exShareAction. 3 :exShareAction a dsap:ShareAction; 4 dsap:aboutDataset :exDataset; 5 dsap:targetRole :exThirdParty. 6 :exDataset a dsap:Dataset. 7 :exThirdParty a dsap:ThirdPartyRole.

A single obligation privacy policy may require multiple triples to express. In natural language, multiple smaller sentences can be shortened when combined to a single sentence. While shortening technique make sentences a pleasure to read, the shortening of sentences adds a layer of complexity to the sentence therefore making its interpretation more ambiguous. In generating RDF triples, combined natural language sentences must first be disassembled to its components. The obligation DSA excerpt shown in Table 4.2 can be broken down into 3 short and clear statements. 1) Everyone in the agreement is prohibited from sharing. The prohibition of the action is captured using the relation `dsap:prohibitedAction` (lines 1-2); 2) Sharing refers to sharing of our dataset. The `dsap:aboutDataset` relation captures the target dataset of the `dsap:ShareAction` concept (lines 3-4); 3) Sharing refers to sharing with any third parties. The target role of the `dsap:ShareAction` can be captured by the `dsap:targetRole` relation (lines 3, 5). When these 3 statements are interpreted together, they unambiguously express the same semantics as the DSA excerpt shown without the loss of any meaning (Table 4.2).

In translating natural language DSA privacy policies to RDF triples, open-ended sentences must be given fixed restrictions. For the administrative DSA privacy policy

Table 4.3: Expressing Administrative Privacy Policy in RDF Triples

DSA Excerpt	**RDF Triples**
Personal data shall be processed in accordance with the rights of data subjects under the law.	1 :dataAction a dsap:DataAction; 2 dsap:requiredJurisdictionCompliance :HIPAA. 3 :HIPAA a dsap:Jurisdiction.

excerpt shown in Table 4.3, the word "processed" is open-ended and must be defined. In this case, to be on the safe side, any performed action on data qualify as "processed". As such, the `DataAction` concept is used to capture the semantics of "processed" (line 1). This include any action that are subclasses of the `DataAction` concept shown in Figure 3.4. Moreover, the legal constraint expressed in the administrative policy (Table 4.3) is broad and open-ended, without specifying which law the researcher must comply to. In writing RDF triples, the open-ended expression of "the law" is given a fixed restriction by the explicit definition of a specific jurisdiction using the `Jurisdiction` concept (lines 2-3). The `Jurisdiction` is attributed to `DataAction` through the `dsap:requiredJurisdictionCompliance` relation (line 2).

Table 4.4: Expressing Data Retention Privacy Policy in RDF Triples

DSA Excerpt	**RDF Triples**
Upon termination of this Agreement, receiving party shall promptly destroy all documents, files and other tangible materials representing disclosing party's Information.	1 :receiver a dsap:DataReceiver; 2 dsap:requiredDestroyAction :deleteData. 3 :deleteData a dsap:DestroyAction; 4 dsap:aboutDataset :healthData; 5 tl:atDate "2018-08-01"^^xsd:date.

Data retention privacy policies is important for its temporal component. In translating the data retention excerpt shown in Table 4.4 to RDF triples, the policy is broken down into 3 components. 1) An obligation clause stating that the receiver

of the data must destroy data (lines 1-2); 2) An informational clause indicating the specific dataset to be destroyed (line 3-4); and 3) A temporal clause stating the date at which dataset must be destroyed (lines 3, 5). With the help of `tl:atDate` from the Timeline Ontology, the semantics of the temporal component captured as `xsd:date` can be incorporated onto `DestroyAction` seamlessly (line 5). On the other hand, the `dsap:Dataset` concept refers to a single distinct dataset, which does not capture semantics of dataset that "[represent] disclosing party's Information". The RDF triple eliminates ambiguity by forcing the author to clearly name specific datasets instead of vaguely describing a type of dataset (line 4).

Table 4.5: Expressing Purpose Definition Privacy Policy in RDF Triples

DSA Excerpt	**RDF Triples**
Personal data shall be obtained only for research purposes, and shall not be further processed in any manner incompatible with that purpose or those purposes.	1 :researchPurpose a dsap:ResearchPurpose. 2 :healthData a dsap:Dataset. 3 4 :receiveAction a dsap:ReceiveAction; 5 dsap:eligibleDataActionPurpose :researchPurpose; 6 dsap:aboutDataset :healthData. 7 :anyAction a dsap:DataAction; 8 dsap:eligibleDataActionPurpose :researchPurpose; 9 dsap:aboutDataset :healthData.

Expressing purpose related privacy policies is also quite simple. In our DSA Privacy Ontology, we define concepts to capture 2 types of purposes: `ResearchPurpose` and `CommercialPurpose` (Figure 3.5). While other data handling purpose may exist, privacy related policies we reviewed only specify these 2 types of purposes. An example of `ResearchPurpose` usage is shown in Table 4.5. The RDF triples capture the semantics of the DSA excerpt by breaking the statement down into 4 components: 1) Receiving the data requires a research purpose. The relationship between `ResearchPurpose` and `ReceiveAction` is captured with `dsap:eligibleDataActionPurpose`

(lines 4-5); 2) The data to receive refers to a specific health dataset. The `dsap:aboutDataset` relation is used to declare the dataset associated with the `ReceiveAction` (lines 4, 6); 3) Performing any action on the dataset requires a research purpose.; and 4) The dataset in statement 3 refers to a specific health dataset. The same concepts and relations in statements 1 and 2 (lines 1-6) are used to translate statements 3 and 4. In decomposing a convoluted sentence expressed in natural language, the RDF triples expresses the same semantics unambiguously (Table 4.5).

Table 4.6: Expressing Data Transfer Channel Privacy Policy in RDF Triples

DSA Excerpt	**RDF Triples**
Genome Canada recognizes publication as a vehicle for data release, and, at a minimum, expects data to be released and shared no later than the original publication date of the main findings from any datasets generated by that project.	1 :researcher a dsap:Researcher; 2 dsap:authorizedAction :publishResults; 3 dsap:requiredAction :discloseData. 4 :publishResults a dsap:PublicationAction; 5 dsap:aboutDataset :healthData. 6 :healthData a dsap:Dataset. 7 :discloseData a dsap:DiscloseAction; 8 dsap:aboutDataset :healthData; 9 tl:before "2018-08-01"^^xsd:date.

The same is true for Data Transfer Channel Definitions. The excerpt shown in Table 4.6 is a data transfer channel policy expressed as a compound sentence consisting of several sentences joined together. When broken down into its components, we can unambiguously express the same semantics in the form of RDF triples based on our ontology. 1) The researcher is authorized to publish (lines 1-2); 2) The publishing in statement 1 refers to the health dataset (lines 4-5); 3) The researcher must disclose dataset (lines 1, 3); 4) The dataset in statement 3 refers to the same dataset being published (lines 4-5, 7-8); and 5) The disclosure of the dataset in statement 3 must happen before the publication date in statement 1 (lines 7, 9). We hereby showed that it is also possible to express data transfer channel privacy policies unambiguously

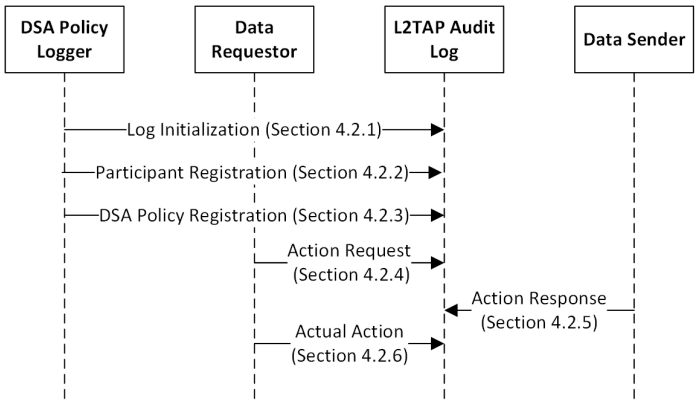

Figure 4.1: Sequence Diagram of Events in L2TAP-Based Audit Log (Adapted from [54])

through RDF triples based on our ontology (Table 4.6).

Through translating DSA policies written in natural language to RDF triples based on our DSA Privacy Ontology, we demonstrated that it is possible for our ontology to express all types of DSA privacy policies (Section 2.2). While some policies are straightforward to translate, others require more effort. The level of effort required is independent from the policy type but rather how the policies were expressed in natural language. Compound sentences and convoluted statement with open-ended interpretation require extra effort to decompose and apply fixed restrictions. As a result, a single sentence may require multiple RDF triples to fully capture its semantics (ie. Table 4.5). Meanwhile, the semantics of other simple clauses can easily be captured with a single RDF triple (ie. Table 4.1). In summary, our DSA Privacy Ontology is sufficient to express the semantics of DSA privacy policies unambiguously in collaborative medical research.

4.2 DSA in Action

In this section, we evaluate our ontology in the context of our motivating scenario described in Section 3.1.1. We demonstrate how our ontology is able to help support the researcher's to compliance to the DSA in this data sharing scenario through logging privacy events and DSA privacy policies in an L2TAP (Linked Data Log to Transparency, Accountability and Privacy) audit log [54].

L2TAP is a family of ontologies that is designed to support audit logging of privacy events using linked data [54]. In supporting our data sharing scenario, we use our DSA Privacy Ontology together with the L2TAP ontologies to populate the audit log in the sequence of privacy events shown in Figure 4.1. Each event: Log Initialization, Participant Registration, DSA Policy Registration, Action Request, Action Response, and Actual Action are described in more detail in the next subsections.

The main participants of the audit logging process are DSA Policy Logger, Data Requestor (Alice and Charlie), the L2TAP Audit Log, and the Data Sender (Bob) (Figure 4.1). For each event generated by the participants in our data sharing scenario (Figure 3.1), we show an example of the L2TAP log serialized in the Turtle syntax [23]. Namespace prefixes used in our log include: **l2tap**, **l2tapi**, **l2tapp**, and **scip** from the L2TAP Ontology [54]; **dsap** from our DSA Privacy Ontology; **tl** from the Timeline Ontology [47]; **foaf** from the Friend of a Friend Ontology [22]; **exlog**, **exAgreement**, **exOrg**, and **exStudy** are instance of log, agreement, organization, and study terms specific to our example; and finally, **rdfg** and **xsd** are RDF graph and XML schema definition terms respectively. A comprehensive list of the namespace prefixes used is shown in Listing 4.1.

```
1 @prefix l2tap:<http://purl.org/l2tap#>.
2 @prefix l2tapi:<http://purl.org/l2tapi#>.
```

```
3 @prefix l2tapp:<http://purl.org/l2tapp#>.
4 @prefix scip:<http://purl.org/scip#>.
5 @prefix dsap: <http://dsap.mingli.ca>.
6 @prefix tl:<http://purl.org/NET/c4dm/timeline.owl#>.
7 @prefix foaf:<http://xmlns.com/foaf/0.1/>.
8 @prefix exlog: <http://dsap-action-example.mingli.ca/log/>.
9 @prefix exAgreement: <http://dsap-action-example.mingli.ca/agreement/>.
10 @prefix exOrg: <http://dsap-action-example.mingli.ca/org/>.
11 @prefix exStudy: <http://dsap-action-example.mingli.ca/study/>.
12 @prefix rdfg:<http://www.w3.org/2004/03/trix/rdfg-1/>.
13 @prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
```

Listing 4.1: Namespace Prefixes Used for Our L2TAP Audit Log

4.2.1 Log Initialization

For each collaborative medical research data sharing scenario, an instance of an L2TAP log is generated. During the first step, the L2TAP log is initialized to define log characteristics for our instance. Each privacy event log contains a log header and a log body. The log header holds information about the event log itself. In our log initialization example (Listing 4.2), the log header is defined in lines 1-12. In our header, `exlog:logevent1` is an instance of `l2tap:LogInitializationEvent` (line 1) and a member of `exlog:log1` (line 2). `exlog:log1` is the URI for the group of logs that together captures all events in our data sharing scenario. Line 3 captures that the logger for our log initialization event is `exlog:dsapLogger`. Lines 4-12 captures the timestamps for the log, utilizing the Timeline Ontology and `xsd:dateTime`.

```
1 exlog:logevent1 a l2tap:LogInitializationEvent;
2   l2tap:memberOf exlog:log1;
3   l2tap:eventParticipant exlog:dsapLogger;
4   l2tap:receivingTimestamp exlog:logevent1-time1;
5   l2tap:publicationTimestamp exlog:logevent1-time2;
6   l2tap:eventData exlog:log-namedgraph1.
7 exlog:logevent1-time1 a tl:Instant;
8   tl:atDateTime "2018-07-01T12:00:00Z"^^xsd:dateTime;
9   tl:onTimeLine exlog:tlphysical.
10 exlog:logevent1-time2 a tl:Instant;
11   tl:atDateTime "2018-07-01T12:00:00Z"^^xsd:dateTime;
12   tl:onTimeLine exlog:tlphysical.
13 exlog:log-namedgraph1 a rdfg:Graph.
14 exlog:log-namedgraph1 {
15   exlog:log1 a l2tapi:Log;
```

```
16     l2tapi:hasLogger exlog:dsapLogger;
17     l2tapi:logTimeline exlog:tlphysical;
18     l2tapi:logDiscreteTimeline exlog:tldiscrete.
19   exlog:tldiscrete a tl:DiscreteTimeLine;
20     time:unitType time:unitDay.
21   exlog:tlphysical a tl:PhysicalTimeLine;
22     owl:sameAs tl:universaltimeline.
23   exlog:dsapLogger a foaf:Agent;
24     l2tapi:initializesLog exlog:log1.
25 }
```

Listing 4.2: Log Initialization

The log body is enclosed within a named graph (lines 14-25). Lines 15-16 indicate that our log `exlog:log1` is an instance of `l2tapi:Log` and that it has a logger `exlog:dsapLogger`. Lines 17-22 specifies that our log uses a discrete timeline and a physical timeline defined by `tl:DiscreteTimeLine` and `tl:PhysicalTimeLine` classes from the Timeline Ontology respectively. Finally, lines 23-24 denote that our logger, `exlog:dsapLogger`, is a `foaf:Agent` and has initialized our log (`exlog:log1`).

4.2.2 Participant Registration

Following log initialization, the participants and their characteristics are registered in our log through the Participant Registration event (Listing 4.3). Lines 1-12 contain the log header, consisting of similar information explained earlier in Log Initialization. Briefly, lines 1-2 asserts that this log event is a `l2tap:ParticipantRegistrationEvent`, and that it is a member of our `exlog:log1` group. Meanwhile, lines 4-12 contain timestamp information regarding this privacy event. In the log body (lines 15-57), Alice (lines 15-24), Bob (lines 25-34), Charlie (lines 35-44), their dataset of interest (line 45), and their DSA end date (lines 46-47) are registered. Specifically, the logger registers Alice denoting that Alice is an instance of `dsap:Agent` (line 15) and that Alice assumes the role of an agreeing party of the DSA (lines 16 and 18) as well as the role of a researcher within Pharmaceutical A (lines 16-17, 19-20) using the

`dsap:hasRole` and `dsap:partOfOrg` relation. Lines 21-24 asserts that Alice is working on a privately funded drug development project within Pharmaceutical A using the `dsap:PrivateFunding` concept from the `dsap:Funding` hierarchy (Figure 3.6). The participants Bob and Charlie are registered in the same format in lines 25-34 and 35-44 respectively according to our data sharing scenario described earlier and shown in Figure 3.1 The logger registers the genetics dataset (line 45) and the agreement end date (lines 46-47) using the `dsap:Dataset` and `dsap:EndDate` concepts.

```
1  exlog:logevent2 a l2tap:ParticipantRegistrationEvent;
2    l2tap:memberOf exlog:log1;
3    l2tap:eventParticipant exlog:dsapLogger;
4    l2tap:receivingTimestamp exlog:logevent2-time1;
5    l2tap:publicationTimestamp exlog:logevent2-time2;
6    l2tap:eventData exlog:log-namedgraph2.
7  exlog:logevent2-time1 a tl:Instant;
8    tl:atDateTime "2018-07-01T12:00:00Z"^^xsd:dateTime;
9    tl:OnTimeLine exlog:tlphysical.
10 exlog:logevent2-time2 a tl:Instant;
11   tl:atDateTime "2018-07-01T12:00:01Z"^^xsd:dateTime;
12   tl:OnTimeline exlog:tlphysical.
13 exlog:log-namedgraph2 a rdfg:Graph.
14 exlog:log-namedgraph2 {
15   exAgreement:alice a dsap:Agent;
16     dsap:hasRole exAgreement:agreeingParty1, exAgreement:pharmaceuticalResearcher;
17     dsap:partOfOrg exOrg:Pharmaceutical.
18   exAgreement:agreeingParty1 a dsap:AgreeingParty.
19   exAgreement:pharmaceuticalResearcher a dsap:Researcher.
20   exOrg:Pharmaceutical a dsap:Organization;
21     dsap:hasResearchStudy exStudy:drugDevelopment.
22   exStudy:drugDevelopment a dsap:ResearchStudy;
23     dsap:hasFunding exStudy:privateFunding.
24   exStudy:privateFunding a dsap:PrivateFunding.
25   exAgreement:bob a dsap:Agent;
26     dsap:hasRole exAgreement:agreeingParty2, exAgreement:univAResearcher;
27     dsap:partOfOrg exOrg:mcmaster.
28   exAgreement:agreeingParty2 a dsap:AgreeingParty.
29   exAgreement:univAResearcher a dsap:Researcher.
30   exOrg:mcmaster a dsap:Organization;
31     dsap:hasResearchStudy exStudy:geneticsStudy.
32   exStudy:geneticsStudy a dsap:ResearchStudy;
33     dsap:hasFunding exStudy:publicFunding.
34   exStudy:publicFunding a dsap:PublicFunding.
35   exAgreement:charlie a dsap:Agent;
36     dsap:hasRole exAgreement:agreeingParty3, exAgreement:univBResearcher;
37     dsap:partOfOrg exOrg:uoft.
38   exAgreement:agreeingParty3 a dsap:AgreeingParty.
39   exAgreement:univBResearcher a dsap:Researcher.
40   exOrg:uoft a dsap:Organization;
41     dsap:hasResearchStudy exStudy:publichealthStudy.
42   exStudy:publichealthStudy a dsap:ResearchStudy;
43     dsap:hasFunding exStudy:publicFunding.
44   exStudy:publicFunding a dsap:PublicFunding.
```

```
45   exAgreement:geneticsData a dsap:Dataset.
46   exAgreement:endDate a dsap:EndDate;
47     tl:atDate "2018-08-01"^^xsd:date.
48   exlog:dsapLogger l2tapp:registersAgent exAgreement:alice, exAgreement:bob.
49   exlog:participants-pharmaceuticalResearcher a l2tapp:Participant;
50     l2tapp:registeredAgent exAgreement:alice;
51     owl:sameAs exAgreement:alice.
52   exlog:participants-univAResearcher a l2tapp:Participant;
53     l2tapp:registeredAgent exAgreement:bob;
54     owl:sameAs exAgreement:bob.
55   exlog:participants-univBResearcher a l2tapp:Participant;
56     l2tapp:registeredAgent exAgreement:charlie;
57     owl:sameAs exAgreement:charlie.
58 }
```

Listing 4.3: Participant Registration

4.2.3 DSA Privacy Policy Registration

Next, the DSA privacy polices are registered in the L2TAP log (Listing 4.4). The header of this log event is shown in lines 1-12, and contain same information as previously described in log initialization and participant registration events. The registration of DSA privacy policies is categorized as a `l2tap:PrivacyEvent` (line 1). Each DSA privacy policy is captured by the `dsap:Policy` concept in the body of the log. For simplicity, our data sharing scenario will have 3 DSA privacy policies. The first registered policy (lines 15-18) asserts that researchers from Public University A are permitted to share genetics data using the `dsap:aboutRole` and `dsap:aboutDataset` relation. The permission to share data is captured using the `dsap:authorizedAction` relation to a `dsap:ShareAction` concept. The second and third registered policy (lines 19-22, 23-28 respectively) denote that researchers from Pharmaceutical A and Public University B are both authorized to receive data, also captured with `dsap:aboutRole` and `dsap:aboutDataset` relation. The third policy (lines 23-28) further mandate the researcher from Public University B to destroy

data (line 26-27) after the end date of the agreement (line 33-34). The obligation to destroy data is captured through the `dsap:requiredAction` relation to a `dsap:DestroyAction` concept. The same researcher from Public University B is also prohibited from further sharing the genetics data (lines 28-30). The prohibition of researcher to share data is captured using the `dsap:prohibitedAction` relation to a `dsap:ShareAction`. The log further asserts that the research project to receive genetics data for must only be publicly funded (lines 31-32, 35). The permission to exclusively use public funding is captured using the `dsap:permittedFunding` relation to a `dsap:PublicFunding` concept.

```
1  exlog:logevent3 a l2tap:PrivacyEvent;
2    l2tap:memberOf exlog:log1;
3    l2tap:eventParticipant exlog:dsapLogger;
4    l2tap:receivingTimestamp exlog:logevent3-time1;
5    l2tap:publicationTimestamp exlog:logevent3-time2;
6    l2tap:eventData exlog:log-namedgraph3.
7  exlog:logevent3-time1 a tl:Instant;
8    tl:atDateTime "2018-07-01T12:00:02Z"^^xsd:dateTime;
9    tl:OnTimeLine exlog:tlphysical.
10 exlog:logevent3-time2 a tl:Instant;
11   tl:atDateTime "2018-07-01T12:00:03Z"^^xsd:dateTime;
12   tl:OnTimeline exlog:tlphysical.
13 exlog:log-namedgraph3 a rdfg:Graph.
14 exlog:log-namedgraph3 {
15   exlog:policy1 a dsap:Policy;
16     dsap:aboutRole exAgreement:univAResearcher;
17     dsap:aboutDataset exAgreement:geneticsData.
18   exAgreement:univAResearcher dsap:authorizedAction exAgreement:shareData.
19   exlog:policy2 a dsap:Policy;
20     dsap:aboutRole exAgreement:pharmaceuticalResearcher;
21     dsap:aboutDataset exAgreement:geneticsData.
22   exAgreement:pharmaceuticalResearcher dsap:authorizedAction exAgreement:receiveData.
23   exlog:policy3 a dsap:Policy;
24     dsap:aboutRole exAgreement:univBResearcher;
25     dsap:aboutDataset exAgreement:geneticsData.
26   exAgreement:univBResearcher dsap:authorizedAction exAgreement:receiveData;
27     dsap:requiredAction exAgreement:destroyData;
28     dsap:prohibitedAction exAgreement:shareData.
29   exAgreement:shareData a dsap:ShareAction;
30     dsap:aboutDataset exAgreement:geneticsData.
31   exAgreement:receiveData a dsap:ReceiveAction;
32     dsap:permittedFunding exStudy:publicFunding.
33   exAgreement:destroyData a dsap:DestroyAction;
34     dsap:hasDate exAgreement:endDate.
35   exStudy:publicFunding a dsap:publicFunding.
36 }
```

Listing 4.4: DSA Privacy Policy Registration

After the initialization and the registration of participants and privacy policies of the L2TAP log, the researchers can now start to share their data. The events that occur during the sharing of data are captured in three steps: 1) request for sharing by a data requestor (an action request), 2) respond to the sharing request by a data sender (an action response), 3) actual receiving of the data by the data requestor (an actual action).

4.2.4 Action Request

The first of three events that occur during the data sharing process involve an actor requesting to perform an action on the data. In our scenario, a data requestor (Alice and Charlie) requests to receive data from a data sender (Bob). Listing 4.5 is a log showing Charlie requesting data from Bob encoded with the Simple Contextual Integrity Privacy (SCIP) module from L2TAP [54]. Similar to other logs, lines 1-12 constitute the header of the log. The action request event is recorded as a `l2tap:PrivacyEvent` (line 1). In the body of the log, the request intent is captured with `scip:AccessRequest`. The role of data requestor (Charlie) and data sender (Bob) are captured in lines 16-17 using `scip:dataRequestor` and `scip:dataSender` relations respectively. Line 18 uses the `scip:requestedDataItem` to assert that the requested data item is the genetics data set. Finally, the requested action in our scenario is captured through the `scip:requestedPrivilege` relation to the `dsap:ReceiveAction` concept (line 19). The actors can request to perform any action described by concepts in the `dsap:DataAction` hierarchy (Figure 3.4).

```
1 exlog:logevent4 a l2tap:PrivacyEvent;
2   l2tap:memberOf exlog:log1;
3   l2tap:eventParticipant exlog:participants-pharmaceuticalResearcher;
4   l2tap:receivingTimestamp exlog:logevent4-time1;
5   l2tap:publicationTimestamp exlog:logevent4-time2;
6   l2tap:eventData exlog:log-namedgraph4.
```

```
exlog:logevent4-time1 a tl:Instant;
  tl:atDateTime "2018-07-01T12:00:02Z"^^xsd:dateTime;
  tl:OnTimeLine exlog:tlphysical.
exlog:logevent4-time2 a tl:Instant;
  tl:atDateTime "2018-07-01T12:00:03Z"^^xsd:dateTime;
  tl:OnTimeline exlog:tlphysical.
exlog:log-namedgraph4 a rdfg:Graph.
exlog:log-namedgraph4 {
  exlog:accessRequest1 a scip:AccessRequest;
    scip:dataRequestor exlog:participants-univBResearcher;
    scip:dataSender exlog:participants-univAResearcher;
    scip:requestedDataItem exAgreement:geneticsData;
    scip:requestedPrivilege exAgreement:receiveData.
}
```

Listing 4.5: Action Request Registration

4.2.5 Action Response

In response to action request by the data requestor, the data sender processes the request to return an action response. This response is also recorded in our L2TAP log. In our scenario, Charlie requests for the genetics data set from Bob. After Bob receives the request, he queries the existing L2TAP log to make an access decision. The decision is positive if Bob is authorized to share the data and if Charlie is authorized to receive the data. The SPARQL query used in this scenario is shown in Listing 4.6.

```
1  ASK
2  WHERE{
3    exAgreement:bob dsap:hasRole ?bobRole.
4    ?bobRole dsap:authorizedAction ?bobAuthorizedAction.
5    ?bobAuthorizedAction a dsap:ShareAction.
6    ?bobAuthorizedAction dsap:aboutDataset exAgreement:geneticsData.
7    exAgreement:charlie dsap:hasRole ?charlieRole.
8    ?charlieRole dsap:authorizedAction ?charlieAuthorizedAction.
9    ?charlieAuthorizedAction a dsap:ReceiveAction.
10   exAgreement:charlie dsap:partOfOrg ?org.
11   ?org dsap:hasResearchStudy ?study.
12   ?study dsap:hasFunding ?funding.
13   ?funding a dsap:PublicFunding.
14 }
```

Listing 4.6: SPARQL to Generate Access Decision

The SPARQL `ASK` statement is used to generate the access decision (line 1). The statement tests for the existence of RDF triple patterns registered within the L2TAP

log returning `TRUE` or `FALSE`. A `TRUE` result grants the access while a `FALSE` result denies the access. Lines 3-4 checks for actions that Bob is authorized to perform per DSA. Line 5 checks if any of the authorized action is a `ShareAction`. Should Bob be authorized to share data, line 6 checks if Bob is authorized to share the genetics dataset that Charlie requested. Correspondingly, lines 7-9 checks if Charlie is permitted to receive data. Finally, a check is added to ensure that Charlie is working on a publicly funded research project (lines 10-13) within his organization because our DSA specifies that only publicly funded research is permitted to receive data.

The execution of Listing 4.6 returns `TRUE`, granting the sharing of the genetics dataset from Bob to Charlie. The access decision is logged in the L2TAP shown in Listing 4.7. The header of the log is presented in lines 1-12. The body of the log uses the SCIP module of the L2TAP ontology to encode an "access granted" decision. Lines 15-16 assert that this access response log is in response to Charlie's data request. The access decision is registered in line 17 of the log as an `xsd:boolean` data type.

```
1 exlog:logevent6 a l2tap:PrivacyEvent;
2   l2tap:memberOf exlog:log1;
3   l2tap:eventParticipant exlog:participants-univAResearcher;
4   l2tap:receivingTimestamp exlog:logevent6-time1;
5   l2tap:publicationTimestamp exlog:logevent6-time2;
6   l2tap:eventData exlog:log-namedgraph3.
7 exlog:logevent6-time1 a tl:Instant;
8   tl:atDateTime "2018-07-01T12:00:02Z"^^xsd:dateTime;
9   tl:OnTimeLine exlog:tlphysical.
10 exlog:logevent6-time2 a tl:Instant;
11   tl:atDateTime "2018-07-01T12:00:03Z"^^xsd:dateTime;
12   tl:OnTimeline exlog:tlphysical.
13 exlog:log-namedgraph6 a rdfg:Graph.
14 exlog:log-namedgraph6 {
15   exlog:accessResponse1 a scip:AccessResponse;
16     scip:responseTo exlog:accessRequest1;
17     scip:accessDecision "true"^^xsd:boolean.
18 }
```

Listing 4.7: SPARQL Query to Generate Access Decision

4.2.6 Actual Action

Finally, the actual sharing of data from Bob to Charlie is registered in the log using the SCIP module of L2TAP ontology (Listing 4.8). Similar to other logs, lines 1-12 make up the header of the log. In the body of the log, line 15 records the occurence of the sharing of data captured using the `scip:ActualAccess` concept. The associated data sharing request is captured with the `scip:accessFor` relation on line 16. Lines 17-20 registers the date at which the data sharing occured using the Timeline Ontology.

```
1 exlog:logevent7 a l2tap:PrivacyEvent;
2   l2tap:memberOf exlog:log1;
3   l2tap:eventParticipant exlog:participants-univBResearcher;
4   l2tap:receivingTimestamp exlog:logevent7-time1;
5   l2tap:publicationTimestamp exlog:logevent7-time2;
6   l2tap:eventData exlog:log-namedgraph3.
7 exlog:logevent7-time1 a tl:Instant;
8   tl:atDateTime "2018-07-01T12:00:02Z"^^xsd:dateTime;
9   tl:OnTimeLine exlog:tlphysical.
10 exlog:logevent7-time2 a tl:Instant;
11   tl:atDateTime "2018-07-01T12:00:03Z"^^xsd:dateTime;
12   tl:OnTimeline exlog:tlphysical.
13 exlog:log-namedgraph7 a rdfg:Graph.
14 exlog:log-namedgraph7 {
15   exlog:accessActual1 a scip:ActualAccess;
16     scip:accessFor exlog:accessRequest1;
17     scip:accessOccurredIn exlog:actualAccessTime.
18   exlog:actualAccessTime a tl:Instant;
19     tl:atDate "2018-07-02"^^xsd:date;
20     tl:onTimeLine exlog:tlphysical.
21 }
```

Listing 4.8: Actual Action

In this section, we discussed how utilizing our ontology together with the L2TAP ontology and Timeline ontology can support the logging of DSA privacy policies and collaborative medical research data sharing privacy events in a linked data based L2TAP audit log with our sample scenario (Figure 3.1). In the next section, we show how this log can be used to answer competency questions asked by researchers.

4.3 Competency Questions

In the last part of our ontology evaluation, we demonstrate how using SPARQL queries on the L2TAP log in our scenario (Section 4.2) can answer the competency questions devised in Section 3.1. Two types of questions are of interest - prohibitive questions and prescriptive questions.

The prohibitive questions asks for the existence of permission of a researcher to perform a certain action. It therefore demands a `YES/NO` response. Answering prohibitive competency questions is analogous to generating access decision described in Section 4.2.5. In fact, in generating the access decision, the SPARQL query used to check for data sharing permission (Listing 4.6) is an example of answer to the question:

Is a particular researcher allowed to share data with another researcher?

In our scenario, Listing 4.6 answers the question: *Is Bob allowed to share the genetics data with Charlie?* The SPARQL query checks for the permission for both Bob to send the data (lines 3-6) and for Charlie to receive the data (lines 7-13). To check for permission of Bob to perform a different action, one can simply replace `dsap:ShareAction` on line 5 with any action concept defined within the `Action` hierarchy (Figure 3.4). For example, to answer the question: *Is Bob allowed to make a copy of the genetics dataset?*, one can replace `dsap:ShareAction` on line 5 with `dsap:CopyAction`. Since no other person is involved in copying data, only checks for Bob (lines 3-6) need to be included in the `WHERE` clause shown in Listing 4.6. In summary, through traversing the L2TAP log and identifying the RDF triple pattern coded

in our SPARQL `ASK` statement, the SPARQL query engine returns a `TRUE/FALSE` response. A `TRUE` response indicate at least one matched pattern and signfy a `YES` answer; Similarly, a `FALSE` response mean a `NO` answer to our prohibitive competency question.

Prescriptive questions on the other hand requests for a piece of information from the DSA. It demands an answer that represents this piece of requested information expressed unambiguously. We utilize SPARQL to query the L2TAP log to find and return the requested information in RDF format. In our scenario, Charlie has the obligation to destroy the data at the agreement end date. We use the SPARQL query shown in Listing 4.9 to answer Charlie's question:

What are my obligations after receipt of the genetics dataset?

A `SELECT` statement is used with the `DISTINCT` modifier to pull the requested information from the L2TAP log while avoiding duplicates (line 1). `?requiredAction` and `?prohibitedAction` are queried to return positive obligations and negative obligations respectively (line 1). Lines 3-6 looks for the matching access request for the genetics dataset by Charlie (line 4). Lines 7-10 attempts to find the actual access of the genetics dataset in the L2TAP log based on the access request and stores the date of actual access in `?actualAccessDate` (line 10). The `?actualAccessDate` is later used in the `OPTIONAL` clause to `FILTER` obligations that are required after the access of genetics dataset (line 16) from all of Charlie's obligations (lines 11-12).

```
1 SELECT DISTINCT ?requiredAction ?requiredActionDate ?actualAccessDate ?prohibitedAction
2 WHERE{
3   ?accessReq a scip:AccessRequest.
4   ?accessReq scip:dataRequestor exlog:participants-univBResearcher.
5   ?accessReq scip:requestedDataItem exAgreement:geneticsData.
6   ?accessReq scip:requestedPrivilege exAgreement:receiveData.
7   ?actualAccess a scip:ActualAccess.
8   ?actualAccess scip:accessFor ?accessReq.
9   ?actualAccess scip:accessOccurredIn ?accessDate.
```

```
10   ?accessDate tl:atDate ?actualAccessDate.
11   exAgreement:univBResearcher dsap:requiredAction ?requiredAction.
12   exAgreement:univBResearcher dsap:prohibitedAction ?prohibitedAction.
13   OPTIONAL{
14     ?requiredAction dsap:hasDate ?requiredActionDateType.
15     ?requiredActionDateType tl:atDate ?requiredActionDate.
16     FILTER (?requiredActionDate > ?actualAccessDate).
17   }.
18 }
```

Listing 4.9: SPARQL Query to Answer a Prescriptive Question

One key aspect to answering competency questions is to provide the medical researcher with an unambiguous and truthful response. We achieve unambiguity through returning direct SPARQL query results in RDF format to the researcher. Since the semantics of the RDF triples are explicitly defined in ontologies, their interpretation will be unambiguous. The responses stay as truthful as the L2TAP log since the SPARQL only searches within the L2TAP log for answer. If the L2TAP log had been unappropriately editted, our answers to competency questions too will be falsified. However, recent advancements in using blockchain to create tamper-proof L2TAP log can be used to address this problem [57]. Using SPARQL to query the L2TAP log, we demonstrate that our answers to competency questions are unambiguous and can be truthful if used with a tamper-proof L2TAP log.

4.4 Summary

In this chapter, our ontology was evaluated in three ways: 1) Our ontology's ability to express DSA privacy policies was demonstrated through translating DSA excerpts expressed in natural language to RDF format based on our DSAP ontology; 2) Our ontology's ability to support collaborative medical research was demonstrated through using our ontology to generate an L2TAP audit log in a hypothetical data sharing

scenario; 3) Our ontology's ability to answer competency questions for the medical researcher was demonstrated through querying the L2TAP audit log with SPARQL.

DSA privacy policies in natural language can be translated to RDF based on our DSAP Ontology. During the translation process, any complex and ambiguous sentences that are open-ended must first be decomposed to simpler, direct sentences. The simple sentences, when translated to RDF, unambiguously expresses privacy semantics conveyed by the DSA. By utilizing additional concepts and relations from L2TAP Ontology and Timeline Ontology, these RDF can support a collaborative medical research scenario during the dynamic aspect of the DSA lifecycle. Enforcement of DSA privacy policies was achieved through recording privacy events in RDF organized in an L2TAP log format [54]. The L2TAP log can be queried with SPARQL to answer privacy competency questions, such as "*what are my obligations?*", to the medical researcher. In summary, we demonstrated our ontology's ability to express DSA privacy policies unambiguously, to support DSA in a collaborative data sharing scenario, and to answer privacy competency questions for a medical researcher.

Chapter 5

Conclusion and Future Work

DSAs specify privacy constraints that limit medical researcher's freedom to share, process, or publish the data to protect data contributor's privacy. We reviewed ways in which privacy constraints are expressed and found gaps. Specifically, expression using natural language are too ambiguous while expression using policy languages are too structured to be flexible to meet the needs of different medical research fields. In this thesis, we proposed an ontology to address this gap to express privacy constraints unambiguously while also allowing flexibility.

Our ontology was designed 1) with hierarchical structure to take advantage of inheritance, enabling easy human readability and allowing for flexibility; 2) to be lightweight, enabling flexibility; 3) with closed world assumption, allowing for better decidability; and 4) with the reuse of other ontologies, benefiting from semantics expressed in existing ontologies. The flexible and unambiguous expression of privacy policies with our ontology makes our ontology better suited for medical research use. Our ontology addresses the gaps left by the amibuity of natural language and the rigidness of structured policy languages. In evaluating our ontology, we demonstrate

that our ontology is capable of supporting a collaborative medical research scenario through logging privacy events in an L2TAP audit log [54].

Future Work

Our ontology has a number of limitations. First, although the DSAs and guidelines we reviewed already have geographic and research topic diversity, other DSAs and guidelines can also exist that may use different terminologies to express the same privacy policy or express different privacy policies not captured in our ontology. For different terminologies, the `OWL:sameAs` [43] property can be used to capture equivalent classes. For different privacy constraints, a more extensive review of DSAs and guidelines can be performed.

Second, even when the semantics of DSA privacy policies were captured in our ontology, there may also exist legal terms in Privacy Acts we may not have captured. The legal concepts are equally important to the governance of the treatment of privacy-sensitive data. To capture legal concepts, a review of privacy legal terms across different jurisdictions can be conducted.

Third, while using ontology and linked data to express privacy policies enable human readability, however our ontology is not readily usable by a human. To make our ontology usable to the medical researcher, an interface shell needs to be developed. The shell can be designed similar to other medical ontology viewers such as the Systematized Nomenclature of Medicine Clinical Terms Ontology Browser [8].

Finally, the use of our ontology in enforcement of DSA privacy policies is based on auditing and accountability. The retrospective review of privacy logs allows the researcher to first be non-compliant to the policies. To strictly prohibit non-compliance,

software tools can be developed to automate compliance checking during the medical research data sharing process.

Bibliography

[1] (2004). RDF Vocabulary Description Language 1.0: RDF Schema (RDFS). https://www.w3.org/2001/sw/wiki/RDFS. Last Accessed: July 1, 2018.

[2] (2010). Health-Care Requirement for Strong Encryption. https://www.ipc.on.ca/wp-content/uploads/Resources/fact-16-e.pdf. Last Accessed: July 5, 2018.

[3] (2014). Resource Description Framework (RDF). https://www.w3.org/RDF/. Last Accessed: February 10, 2018.

[4] (2016). Organ and tissue donor registration. https://www.ontario.ca/page/organ-and-tissue-donor-registration. Last Accessed: June 14, 2018.

[5] (2018). Free Online Word Counter. http://countwordsfree.com/. Last Accessed: May 12, 2018.

[6] (2018). Macmillan Dictionary. https://www.macmillandictionary.com/. Last Accessed: May 1, 2018.

[7] (2018). Sample Data-Sharing and Usage Agreement. https://www.cdc.gov/cancer/ncccp/doc/sampledatasharingusageagreement.doc. Last Accessed: April 22, 2018.

[8] (2018). SNOMED International SNOMED CT Browser. http://browser.ihtsdotools.org/? Last Accessed: September 25, 2018.

[9] Arenas, A. E., Aziz, B., Bicarregui, J., and Wilson, M. D. (2010). An event-B approach to data sharing agreements. In *International Conference on Integrated Formal Methods*, pages 28–42. Springer.

[10] Ashley, P., Hada, S., Karjoth, G., Powers, C., and Schunter, M. (2003). Enterprise privacy authorization language (EPAL). *IBM Research*.

[11] Aziz, B., Arenas, A., Martinelli, F., Matteucci, I., and Mori, P. (2008). Controlling usage in business process workflows through fine-grained security policies. In *International Conference on Trust, Privacy and Security in Digital Business*, pages 100–117. Springer.

[12] Aziz, B., Arenas, A., and Wilson, M. (2011). SecPAL4DSA: a policy language for specifying data sharing agreements. In *FTRA International Conference on Secure and Trust Computing, Data Management, and Application*, pages 29–36. Springer.

[13] Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., and Zhao, J. (2013). PROV-O: The PROV Ontology. https://www.w3.org/TR/prov-o/. Last Accessed: May 22, 2018.

[14] Belloum, A., Inda, M. A., Vasunin, D., Korkhov, V., Zhao, Z., Rauwerda, H., Breit, T. M., Bubak, M., and Hertzberger, L. O. (2011). Collaborative e-science experiments and scientific workflows. *IEEE Internet Computing*, **15**(4), 39–47.

[15] Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, **63**(6), 1059–1078.

[16] Brodie, C. A., Karat, C.-M., and Karat, J. (2006). An empirical study of natural language parsing of privacy policy rules using the SPARCLE policy workbench. In *Proceedings of the second symposium on Usable privacy and security*, pages 8–19. ACM.

[17] Campbell, E. G. and Bendavid, E. (2002). Data-sharing and data-withholding in genetics and the life sciences: Results of a national survey of technology transfer officers. *J. Health Care L. & Pol'y*, **6**, 241.

[18] Clavel, M., Durán, F., Eker, S., Lincoln, P., Martí-Oliet, N., Meseguer, J., and Talcott, C. (2007). *All about maude-a high-performance logical framework: how to specify, program and verify systems in rewriting logic*. Springer-Verlag.

[19] Collis, A., McAllister, D., and Ball, M. (2011). BBSRC Data Sharing Policy. *Nature*.

[20] Cranor, L., Langheinrich, M., Marchiori, M., Presler-Marshall, M., and Reagle, J. (2002). The Platform for Privacy Preferences 1.0 (P3P1.0) Specification. https://www.w3.org/TR/P3P/. Last Accessed: August 5, 2018.

[21] Cristani, M. and Cuel, R. (2005). A survey on ontology creation methodologies. *International Journal on Semantic Web and Information Systems (IJSWIS)*, **1**(2), 49–69.

[22] Dan Brickley, L. M. (2014). FOAF Vocabulary Specification 0.99. http://xmlns.com/foaf/spec/. Last Accessed: June 2, 2018.

[23] David Beckett, T. B.-L. (2011). Turtle - Terse RDF Triple Language. https://www.w3.org/TeamSubmission/turtle/. Last Accessed: June 22, 2018.

[24] Fecher, B., Friesike, S., and Hebing, M. (2015). What drives academic data sharing? *PLoS One*, **10**(2), e0118053.

[25] for Disease Control, C., Prevention, *et al.* (2003). HIPAA privacy rule and public health. Guidance from CDC and the US Department of Health and Human Services. *MMWR: Morbidity and mortality weekly report*, **52**(Suppl. 1), 1–17.

[26] Giunchiglia, F. and Zaihrayeu, I. (2009). Lightweight ontologies. In *Encyclopedia of Database Systems*, pages 1613–1619. Springer.

[27] GlaxoSmithKline, L. (2015). DATA SHARING AGREEMENT. https://study329.org/wp-content/uploads/2015/08/DATA-SHARING-AGREEMENT.pdf. Last Accessed: June 16, 2018.

[28] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, **5**(2), 199–220.

[29] Grüninger, M. and Fox, M. S. (1995). Methodology for the design and evaluation of ontologies. *Workshop on Basic Ontological Issues in Knowledge Sharing*.

[30] Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., and Dean, M. (2004). SWRL: A Semantic Web Rule Language Combining OWL and RuleML. https://www.w3.org/Submission/SWRL/. Last Accessed: May 2, 2018.

[31] Hustadt, U. *et al.* (1994). Do we need the closed world assumption in knowledge representation? In *KRDB*.

[32] Jarquin, P. B. (2012). Data Sharing: Creating Agreements. http://www.ucdenver.edu/academics/colleges/PublicHealth/research/centers/R-MPRC/resources/Documents/ToolsLast Accessed: April 15, 2018.

[33] Kagal, L., Finin, T., and Joshi, A. (2003). A policy language for a pervasive computing environment. In *Policies for Distributed Systems and Networks, 2003. Proceedings. POLICY 2003. IEEE 4th International Workshop on*, pages 63–74. IEEE.

[34] Kervin, K., Cook, R. B., and Michener, W. K. (2014). The Backstage Work of Data Sharing. In *Proceedings of the 18th International Conference on Supporting Group Work*, pages 152–156. ACM.

[35] Krishna, R., Kelleher, K., and Stahlberg, E. (2007). Patient confidentiality in the research use of clinical medical databases. *American journal of public health*, **97**(4), 654–658.

[36] Lin, Z., Owen, A. B., and Altman, R. B. (2004). Genomic research and human subject privacy. *Science*, **305**(5681), 183–183.

[37] Lonsdale, D., Embley, D. W., Ding, Y., Xu, L., and Hepp, M. (2010). Reusing ontologies and language components for ontology generation. *Data & Knowledge Engineering*, **69**(4), 318–330.

[38] Malin, B., Karp, D., and Scheuermann, R. H. (2010). Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *Journal of Investigative Medicine*, **58**(1), 11–18.

[39] Manea, M. and Petrocchi, M. (2015). Engineering the Lifecycle of Data Sharing Agreements. *ERCIM News*.

[40] Matteucci, I., Petrocchi, M., and Sbodio, M. L. (2010). CNL4DSA: a controlled

natural language for data sharing agreements. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 616–620. ACM.

[41] Matteucci, I., Mori, P., Petrocchi, M., and Wiegand, L. (2011). Controlled data sharing in E-health. *2011 1st Workshop on Socio-Technical Aspects in Security and Trust (STAST)*, pages 17–23.

[42] Matteucci, I., Petrocchi, M., Sbodio, M. L., and Wiegand, L. (2012). A design phase for data sharing agreements. In *Data Privacy Management and Autonomous Spontaneus Security*, pages 25–41. Springer.

[43] McGuinness, D. L. and van Harmelen, F. (2004). OWL Web Ontology Language. https://www.w3.org/TR/owl-features/. Last Accessed: August 2, 2018.

[44] Murphy, S. N., Weber, G., Mendis, M., Gainer, V., Chueh, H. C., Churchill, S., and Kohane, I. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, **17**(2), 124–130.

[45] O'herrin, J. K., Fost, N., and Kudsk, K. A. (2004). Health Insurance Portability Accountability Act (HIPAA) regulations: effect on medical record research. *Annals of surgery*, **239**(6), 772.

[46] Rahman, M. and Fukui, T. (2003). Biomedical research productivity: factors across the countries. *International journal of technology assessment in health care*, **19**(1), 249–252.

[47] Raimond, Y. and Abdallah, S. (2007). The Timeline Ontology.

http://motools.sourceforge.net/timeline/timeline.html. Last Accessed: May 25, 2018.

[48] Reich, M. R. (2002). Public-private partnerships for public health. *Public-private partnerships for public health*, pages 1–18.

[49] Reimer, U. (2011). Lightweight Ontologies (LWO) versus Full-Fledged Ontologies. http://www.ai-one.com/2011/05/31/lightweight-ontologies-lwo-versus-full-fledged-ontologies/. Last Accessed: August 1, 2018.

[50] Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, **11**, 95–130.

[51] Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature*, **405**(6788), 847–856.

[52] Rothenberg, K. H. (1995). Genetic information and health insurance: state legislative approaches. *The Journal of Law, Medicine & Ethics*, **23**(4), 312–319.

[53] Ruan, C. (2008). UML specification of e-consent requirements in a health care system. In *International Symposium on Computer Science and its Applications*, pages 275–280. IEEE.

[54] Samavi, R. and Consens, M. P. (2014). Publishing L2TAP Logs to Facilitate Transparency and Accountability. In *LDOW*.

[55] Savage, N. (2016). Privacy: The myth of anonymity. *Nature*, **537**(7619), S70–S72.

[56] Sbodio, M. L. (2014). Data sharing agreements. https://www.google.com/patents/US20140089212.

[57] Sutton, A. and Samavi, R. (2017). Blockchain enabled privacy audit logs. In *International Semantic Web Conference*, pages 645–660. Springer.

[58] Swarup, V., Seligman, L., and Rosenthal, A. (2006). A data sharing agreement framework. In *International Conference on Information Systems Security*, pages 22–36. Springer.

[59] Tabor, H. K., Berkman, B. E., Hull, S. C., and Bamshad, M. J. (2011). Genomics really gets personal: how exome and whole genome sequencing challenge the ethical framework of human genetics research. *American Journal of Medical Genetics Part A*, **155**(12), 2916–2924.

[60] Tantisira, K. G. and Drazen, J. M. (2009). Genetics and pharmacogenetics of the leukotriene pathway. *Journal of Allergy and Clinical Immunology*, **124**(3), 422–427.

[61] Trinidad, S. B., Fullerton, S. M., Bares, J. M., Jarvik, G. P., Larson, E. B., and Burke, W. (2010). Genomic research and wide data sharing: views of prospective participants. *Genetics in Medicine*, **12**(8), 486–495.

[62] Turnbull, C. and Hodgson, S. (2005). Genetic predisposition to cancer. *Clinical medicine*, **5**(5), 491–498.

[63] Wilson, M., Crompton, S., Matthews, B., and Orlov, A. (2011). Enforcing scientific data sharing agreements. In *E-Science (e-Science), 2011 IEEE 7th International Conference on eScience*, pages 271–278. IEEE.

[64] Win, K. T., Susilo, W., and Mu, Y. (2006). Personal health record systems and their security protection. *Journal of medical systems*, **30**(4), 309–315.

Appendix A

Preliminary Vocabulary List

Table A.1 shows the preliminary vocabulary list generated from data sharing agreements and data sharing guidelines according to methodology described in Figure 3.2. The words in the "Term" column were used as a starting point to build our ontology. Synonym(s) of the terms that appeared in data sharing agreements and data sharing guidelines are shown in the "Alias" column. Alias words are not used. The last column, "Concept/Relation" denote whether the term is a (c)oncept or a (r)elation.

Table A.1: Preliminary Vocabulary List

Term	**Alias**	**Concept/Relation**
data	information	c
dataset		c
database	repository, archive, storage	c
share		r
research	study, project	c
investigator	researcher, professor, doctor	c

Table A.1: Preliminary Vocabulary List Continued

Term	**Alias**	**Concept/Relation**
participant	subject, originator	c
requestor		c
plan		c
property		c
funding		c
from		r
use		r
restriction	obligation	c
ensure	necessary, must, require	r
policy		c
have		r
when		r
confidentiality		r
organization	community, agency, university, institute	c
department		c
public		r
made		r
preservation		r
publication		c
period		c

Table A.1: Preliminary Vocabulary List Continued

Term	**Alias**	**Concept/Relation**
grant		r
prior		r
party	third party, other party	c
agreement		c
collaboration		r
disclose		r
authorized		r
under		r
receive		r
date	term, current	c
representative		c
jurisdiction	Canada, legislation, regulation, statutory, province, federal HIPAA	c
operate		r
access		r
analysis		c
procedure	mechanism, approach, method, process, protocol	c
resource		c
will		r

Table A.1: Preliminary Vocabulary List Continued

Term	**Alias**	**Concept/Relation**
release		r
relevant		r
user		c
purpose	objective	c
document	agreement	c
consent		c
inform		r
identifiable		r
fund		c
public		r
trans-border	geographic	r
communication		c
transfer		r
collection		c
source		c
delete		r
associated		r
partner		c
staff		c
identify		r
commission		c

Table A.1: Preliminary Vocabulary List Continued

Term	**Alias**	**Concept/Relation**
relating		r
mutually		r
anonymised		r
responsible		r
behalf		r