An Ontology on the imminent POPI Act

Yahlieel Jafta
2858132@myuwc.ac.za
University of the Western Cape
Cape Town, Western Cape

ABSTRACT

The increase of connectivity due to technology advances is shifting the attention of legislators in various jurisdictions to the protection of personal information and data. The focus of this paper is the protection of privacy information, specifically the POPIA within South Africa. The Act is yet to be implemented and role players within the POPIA still have time to be compliant and understand how it affects them. This paper follows the engineering process of an ontology to establish a small knowledge base for the regulations in the POPIA and how it affects these role players. It provides a background on semantic technologies and outlines the scope and methodology for the development of an ontology.

KEYWORDS

Legal ontology, data protection, information privacy, Protection of Personal Data Act. POPIA, GDPR, compliance, data subject, responsible party

1 INTRODUCTION

We are living in an era in which connectivity is progressing rapidly. During the past decade this increase in connectivity shifted the focus of legislators in various jurisdictions to the protection of personal information and data. The motivation for this focus is a matter of concern, due to the impact this have on the privacy rights of individuals.

Embedded in section 14 of the Constitution of the Republic of South Africa, is the right to privacy by each South African citizen. This includes individuals and businesses. The right to privacy includes a right to protection against the unlawful collection, retention, dissemination and use of personal information [9]. While businesses have a legitimate reason to acquire personal data as information assets to achieve their business goals, they are required to comply with any regulatory requirements [4]. Nevertheless, this information is susceptible to abuse. With the rise of social media, businesses have new means to gain traction in the information space. A prime example of this was brought to the fore on the international stage by the recent Facebook–Cambridge Analytica data scandal [34] [7]. In the wake of this, the European Union (EU) was in the process of reforming their current data protection law, the General Data Protection Regulation (GDPR). This reform was developed since 2010 and came into effect on 25 May 2018. Therefore, legislators are in the process of enacting laws to secure proper information handling procedures.

The objective of privacy and data protection domains of law is to protect the personal information of the individuals in a given jurisdiction. Businesses operational in the South Africa are facing the enactment of the Protection of Personal Information Act 4 of 2013 (POPIA) which has been signed into law by the President on 19 November 2013 [32]. The POPIA is yet to be implemented, and the regulations to the Act only recently published on 14 December 2018. Once implementation is confirmed there will be a 1 year grace period for entities to be compliant, before enforcement. The Act aims to align the regulation of personal information in South Africa with international standards, which will put pressure on businesses operating within South Africa.

The POPIA seeks to protect the right of privacy that applies to individuals and juristic entities (referred to as data subjects) by the establishment of strict guidelines on how to obtain and process information [32]. These guidelines affects organizations (referred to as the responsible party [32]) and data subjects in various ways and is therefore important that these role players are well informed with regards to the implications. It is with this in mind that a knowledge base, through a semantic representation, for the legislation of the POPIA will be valuable for assisting with the education of both the data subjects and organizations on the POPIA.

In this paper I provide the analysis, design and implementation of a basic ontology on the data protection domain in the context of the POPIA. The paper is structured as follows. Section 2 provides a background on semantic technologies, their importance and influences in the scope of knowledge bases and information formalization. Section 3 describes the related work concerning domain legal ontologies within the context of data protection and privacy. Section 4 will outline the initial functional and non-functional requirements. The methodology for developing the ontology will be discussed in section 5. Finally, I provide an initial summary of the paper and future development.

2 BACKGROUND

2.1 Semantic Technologies

One of the biggest problems we face today is an overload of information. This is evident in various domains as the availability of large scale information is more abundant than ever before. Businesses which conduct their activities in a globalized world have become dependent on information systems in order to survive [22]. This is amplified by the huge size of the World Wide Web (WWW) and the resources it provides. In the WWW context, search engines have come a long way in handling vast amounts of information and making it available on the Web. However, with the continuous growth of the Web, search engines will have a hard time maintaining the quality of retrieval results [6]. An argument is made for Semantics being the single most important factor for advancing the Web to its next phase.

Semantic technology encodes meanings separately from data and content files, and separately from application code. It deals with large data sets, and links them together via self-describing interrelations allowing it to be processed by machines. Semantic technologies are one of the initiatives to address the challenge that large scale information, known as Big Data, presents. The advances in hardware along with Big Data provides enormous opportunities for individuals, businesses and society, but we haven't seen similar advances in software development capability. This could potentially give rise to a second Software Crisis [15]. Semantic technology seeks to address this potential crisis in the Big Data space. It is considered to be the best framework to deal with the diverse and massive scale of resources on the Web [30]. Even though the Web is a major force in the scope of Big Data it is not the only push factor. Push factors includes the processing of large data in enterprises and various domains, where making good decisions efficiently is essential to allowing organizations to manage and make better sense of their data.

2.2 Ontology

In Computer Science, Ontology is a formal representation of the knowledge by a set of concepts within a domain and the relationships between those concepts [25]. It is the process of formalizing knowledge and expressing the concepts and their relations in a given domain. An ontology defines a common vocabulary for researchers who need to share information in a domain. It includes machine–interpretable models of the concepts in a domain [12]. As a result, ontologies introduces a sharable and reusable knowledge base, enabling extension of knowledge of a given domain. Ontology is a W3C standard and is part of the Semantic Web stack, thus making it one of the building blocks of Semantic Technologies.

Ontologies are a big factor in promoting automated processes to access information. It is expected to provide structured vocabularies that analyze the relationships between different terms, allowing machines and humans to interpret their meaning unambiguously. For example, an appropriate pizza ontology might include the information that Mozzarella and Parmesan are variants of cheese, that cheese is not a kind of meat or seafood, and that a vegetarian pizza is one whose toppings do not include any meat or seafood. This information allows the term "pizza topped with only Mozzarella and Parmesan" to be unambiguously defined as a specialization of the term "vegetarian pizza" [19]. The formal representation of the information in an ontology allows for better data management. This is achieved by the common understanding of information an ontology provides and the explicit assumptions made by the various concepts in a domain.

Ontologies are used in various domains as a form of knowledge representation about the world or some subset of it [25]. Domains includes Artificial Intelligence, the Semantic Web, Biomedical Informatics and even in the legal domain. The use of ontologies in the legal domain assists with the organization of legal documents and providing support for legal reasoning [4]. Even though the focus of this paper is on the legal domain we have seen a number of successful ontologies developed in the health sector domains. This includes the Gene Ontology [8] and the Protein ontology [31].

2.2.1 Gene Ontology (GO). Applying to the domain of biology. GO is a bioinformatics resource for describing the roles genes play in the life of an organism, covering a variety of species from humans to bacteria and viruses [27]. The GO project seeks to provide a set of structured vocabularies for specific biological domains that can be used to describe gene products in any organism. The objective of GO was to develop cross-species biological vocabularies that are consumed by multiple databases to annotate genes and gene products in a consistent way [8].

2.2.2 Protein ontology (PO). The motivation for the development of PO was to efficiently represent the protein annotation framework and integrating the existing data representations into a standardized protein data specification for the bioinformatics community. The objective of PO was to correlate information about multiprotein machines with data in major protein databases to better understand sequence, structure and function of protein machines [31].

2.3 Owl

The Web Ontology Language (OWL) [26] is a Semantic Web language designed to represent and capture rich and complex knowledge about things, and their relations between them. OWL is a computational logic-based language [29]. The knowledge expressed by OWL enables it to be reasoned about. This is achieved by means of software applications that verifies consistency of the domain knowledge within an ontology or make implicit knowledge explicit [29]. OWL ontologies promotes reuse and modularity, as it can be published in the WWW and may refer to or be referred from other OWL ontologies. Knowledge concepts captured from data, in a given domain, is reasoned about in a rich hierarchical structure of concepts and their inter-relationships [31]. These relationships helps with the matching of concepts even if the data sources describing these concepts is not 100% uniform.

OWL is not restricted only to the Web, and has been applied successfully for knowledge modelling in many application areas, such as [8] and the [31]. Modelling information in OWL provides two practical benefits. As a descriptive language, it can be used to express domain knowledge in a formal way, and as a logical language, it can be used to infer conclusions from this knowledge. The latter point is what distinguishes OWL from other modelling languages such as UML [23].

Since 2009 there is a second version of OWL. This is due to challenges faced in the initial version, OWL 1. The challenges presented relates to the efficiency and scalability of the reasoning process. Reasoning was typically a multi-exponential process, which did not always yield results [23]. To address these issues, OWL 2 [3] was released. OWL 2, a W3C standard, introduced three profiles: OWL EL, OWL RL, and OWL QL. Also called sublanguages. These sublanguages of OWL restricts the available modelling features in order to simplify reasoning. This has yielded great results in improving the performance and scalability which has made the OWL 2 profiles very attractive for ontology engineers [23].

OWL EL is used in large biomedical ontologies, with applications of this in [8] and [31] [23]. OWL RL is the preferred approach for reasoning with Web data. OWL QL provides database applications with an ontological data access layer. This is demonstrated in [20].

The usage of OWL is applied in various successful ontologies, thus demonstrating the benefits it provides.

3 RELATED WORK

The POPIA is very closely aligned with the GDPR with one of the differences being that the POPIA applies the term data subject to natural and juristic persons, whereas the GDPR only applies this to natural persons. However, the GDPR is more extensive than the POPIA. There have been a few approaches related to expressing the GDPR as an ontology. An ontology for the GDPR was developed for the data protection requirements. It modeled the data protection requirements in the context of the GDPR reform and presented an approach for integrating it into a workflow to express these requirements within a business process by means of the ontology [5]. The goal of the ontology was to provide support for data controllers in achieving compliance with the GDPR legislation. This was done to create an ontologoical representation of the duties of data controllers and the corresponding rights of data subjects.

The GDPRov project is an ontology concerned with the management of compliance by means of identifying provenance information related to consent and personal data required for compliance documentation. It is an OWL 2 linked open data ontology that represent provenance of consent and data lifecycle workflows for the GDPR.

The proposal of this paper addresses the use of an ontology to model a set of concepts expressed within the POPIA and how it affects the data subjects and responsible parties once the Act is implemented. These concepts and the extent to which this knowledge will be used is still being explored with the goal of outlining POPIA compliance for various industries. The implementation of the Act impacts various industries differently, one such example being the digital marketing industry. Current practices such as direct marketing are regarded as a cost effective option for driving sales in various organizations and this option will largely be removed once the POPIA is implemented [1].

This section is still a work in progress as I'm reviewing additional literature that will help with the scope of my ontology and provide a framework for competency questions that the ontology can answer.

4 REQUIREMENTS

4.1 Functional Requirements

The requirements for this project will be defined by a set of competency questions the ontology should answer. These questions will serve as the litmus test in the evaluation phase of development process and will help define the scope of the ontology [12]. An initial set of questions is listed below.

How will the Act affect Direct Marketing?

What does the impact of the Act on Direct Marketing have on data subjects?

What are some of the exclusions of the Act?

The set of questions is not finalized yet and is subject to change. These questions will be revised once the scope of the ontology is determined.

4.2 Non-functional Requirements

The development tool that will be used for this project is Protégé [28]. Protégé is a free open-source ontology editor that is being developed at the Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine. It is supported by a large community of academic, government, and corporate users, who use it to build knowledge-based solutions for various domains [16]. One of the main strengths of protégé is its user interface, and the flexible manner in which it can be extended to provide additional functionality in the form of plug-ins [24]. One of the design goals of protégé is to be compatible and adaptable with other systems for knowledge representation and knowledge extraction [24]. This compatibility will be a great feature for a POPIA ontology as the information privacy domain is large, enabling possible integration with other knowledge bases within the same domain.

The ontology language that will be used for this project is OWL and it is available as a plugin for protégé. The OWL plugin allow users to exploit the features of OWL within the protégé editor and allow them to make use of intelligent guidance to find mistakes similar to a debugger in a programming environment. The plugin also provides an open testing framework in which code similar to JUnit test cases can be executed [21]. The features and benefits outlined for the combination of protégé and the OWL plugin aligns with the non-functional requirements that needs to be satisfied. Since the ontology is based on an Act, it will always be susceptible to change, due to change or amendments in regulations. Based on this, an initial list of non-functional requirements that will need to be satisfied are Adaptability, Reusability, Configurability, Testability, Maintainability and Quality. Considering these requirements, the protégé editor provides a suitable environment for the development of an ontology to meet these requirements.

5 METHODOLOGY

This project will be implemented through 4 phases: Analysis, Design, Implementation and Evaluation. Analysis: This involves literature study and review. Defining the requirements for the project within the scope of the POPIA. Design: Includes the modelling of various concepts within the POPIA. Implementation: The implementation forms part of the design phase, as the modelling is done within the protégé editor. This allows for quick feedback in the design phase. Evaluation will be performed at the latter stage of the project to test conformance of ontology to requirements.

A series of methodologies for developing ontologies have been reported in literature since the early 1990's demonstrating that there is no single set of processes for ontology engineering. Therefore the methodology that will be implemented needs to be relatively applicable, to the domain of information privacy and fulfill the necessary requirements for the duration of the engineering life cycle. Methodologies includes: The Cyc methodology, methodology of Uschold and King, methodology of Grūninger and Fox, METHONTOLOGY, The KACTUS approach and the SENSUS-based methodology [13].

The Cyc methodology is a product of the development of the Cyc Knowledge Base (Cyc KB) [11]. Each phase includes the development of knowledge representation, outlining abstract concepts and representing the rest of the knowledge using these concepts [13].

The methodology of Uschold and King is based on the development of the Enterprise Ontology, an ontology for enterprise modelling processes at the Artificial Intelligence Applications Institute (AIAI) of Edinburgh [33].

The methodology of Grūninger and Fox is based on the development of the TOVE [17] project ontology within the domain of business processes and activities modelling [18]. It involves building a logical model of a knowledge source that is to be specified by means of an ontology. It is a logic-based formal methodology that transforms informal scenarios expressed in a natural language into a computable model expressed through logic [13].

The METHONTOLOGY method enables the construction of ontologies at the knowledge level and follows an iterative approach by means of evolving prototypes. The foundation of this methodology is in the main activities identified by the software development process [2].

The methodology framework being considered for building an ontology based on the POPIA is the METHONTOLOGY [14] method since the scope of the ontology won't be that large. This methodology have been implemented within the legal knowledge domain [10] [5] and is a mature methodology compared to the other methodologies mentioned above [13]. It allows for the construction of ontologies at the knowledge level, and is based on the main activities identified by the IEEE software development process [10]. Furthermore, it provides a framework for iterative development which is suitable for the development of this project.

6 CONCLUSION AND FUTURE WORK

This paper highlights the importance of private information regulation in a global connected world. It highlights the shift in focus of legislators to enact legislation to support the privacy rights of individuals and juristic entities. The focus of this paper is the South African POPIA and the impact it will have on various role players, such as the responsible party and data subjects. It outlines the development of an ontology to provide a knowledge base on various concepts within the Act that will promote transparency and education that can aid with the inception of this Act. The development of the ontology is still in progress. The full scope of what will be covered in the ontology is still to be established, such as the set of competency questions the ontology should provide answers on. An initial list of questions is given and will be used as a basis for the next development phase, design. Once the design phase is completed, the construction of the ontology will commence followed by an evaluation to assert satisfaction of requirements within the scope of the project.

REFERENCES

[1] [n. d.]. OPINION: Direct marketing - SMEs best sales driver. https://www.workspace.co.uk/community/homework/marketing/opinion-direct-marketing-smes-best-sales-drive [Online; accessed 24-March-2019].

[2] 1996. IEEE Standard for Developing Software Life Cycle Processes. *IEEE Std 1074-1995* (1996), i–. https://doi.org/10.1109/IEEESTD.1996.79663

[3] 2012. *OWL 2 Web Ontology Language Document Overview (Second Edition).* W3C Recommendation. W3C. http://www.w3.org/TR/2012/REC-owl2-overview-20121211/.

[4] Cesare Bartolini. 2015. Reconciling Data Protection Rights and Obligations : An Ontology of the Forthcoming EU Regulation.

[5] Cesare Bartolini, Robert Muthuri, and Cristiana Santos. 2017. Using Ontologies to Model Data Protection Requirements in Workflows. In *New Frontiers in Artificial Intelligence*, Mihoko Otake, Setsuya Kurahashi, Yuiko Ota, Ken Satoh, and Daisuke Bekki (Eds.). Springer International Publishing, 233–248.

[6] V. Richard Benjamins, Jesus Contreras, Oscar Corcho, and Asuncion Gomez-Perez. 2002. Six challenges for the Semantic Web. (04 2002).

[7] MacKenzie F. Common. 2018. Facebook and Cambridge Analytica: let this be the high-water mark for impunity. http://eprints.lse.ac.uk/88964/.

[8] T. G. O. Consortium. 2001. Creating the Gene Ontology Resource: Design and Implementation. *Genome Research* 11, 8 (Aug 2001), 1425âĂŞ1433. https://doi.org/10.1101/gr.180801

[9] Constitution of the Republic of South Africa. 1996. Section 14. http://www.justice.gov.za/legislation/constitution/SAConstitution-web-eng.pdf.

[10] Oscar Corcho, Mariano Fernández-López, Asunción Gómez-Pérez, and Angel López-Cima. 2005. *Building Legal Ontologies with METHONTOLOGY and WebODE.* Springer Berlin Heidelberg, Berlin, Heidelberg, 142–157. https://doi.org/10.1007/978-3-540-32253-5_9

[11] Cycorp. [n. d.]. Knowledge Base. http://www.cyc.com/kb/

[12] N F. Noy and Deborah Mcguinness. 2001. Ontology Development 101: A Guide to Creating Your First Ontology. *Knowledge Systems Laboratory* 32 (01 2001).

[13] Mariano Fernández-López and Asunción Gómez-Pérez. 2002. Overview and Analysis of Methodologies for Building Ontologies. *Knowl. Eng. Rev.* 17, 2 (June 2002), 129–156. https://doi.org/10.1017/S0269888902000462

[14] M. Fernández-López, A. Gómez-Pérez, and N. Juristo. 1997. METHONTOLOGY: From Ontological Art Towards Ontological Engineering. In *Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series.* American Asociation for Artificial Intelligence. http://oa.upm.es/5484/ Ontology Engineering Group ? OEG.

[15] Brian Fitzgerald. 2012. Software Crisis 2.0. *IEEE Computer - COMPUTER* 45 (04 2012), 89–91. https://doi.org/10.1109/MC.2012.147

[16] Stanford Center for Biomedical Informatics Research. [n. d.]. A free, open-source ontology editor and framework for building intelligent systems. https://protege.stanford.edu/ [Online; accessed 24-March-2019].

[17] Mark S. Fox. 1992. The TOVE Project Towards a Common-Sense Model of the Enterprise. In *IEA/AIE.*

[18] Michael GrÃijninger and Mark S. Fox. 1995. Methodology for the Design and Evaluation of Ontologies.

[19] Ian Horrocks, Peter Patel-Schneider, and Frank Harmelen. 2003. From SHIQ and RDF to OWL: The Making of a Web Ontology Language. *SSRN Electronic Journal* (01 2003). https://doi.org/10.2139/ssrn.3199003

[20] Evgeny Kharlamov, Nina Solomakhina, Özgür Lütfü Özçep, Dmitriy Zheleznyakov, Thomas Hubauer, Steffen Lamparter, Mikhail Roshchin, Ahmet Soylu, and Stuart Watson. 2014. How Semantic Technologies Can Enhance Data Access at Siemens Energy. In *The Semantic Web – ISWC 2014*, Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble (Eds.). Springer International Publishing, 601–619.

[21] Holger Knublauch. 2004. Ontology-Driven Software Development in the Context of the Semantic Web: An Example Scenario with Protege/OWL.

[22] Sel ¸cuk Köylüoglu, Levent Duman, and Aykut Bedük. 2015. Information Systems in Globalization Process and Their Reflections in Education. *Procedia – Social and Behavioral Sciences* 191 (06 2015), 1349–1354. https://doi.org/10.1016/j.sbspro.2015.04.610

[23] Markus Krötzsch. 2012. *OWL 2 Profiles: An Introduction to Lightweight Ontology Languages.* Springer Berlin Heidelberg, Berlin, Heidelberg, 112–183. https://doi.org/10.1007/978-3-642-33158-9_4

[24] Patrick Lambrix, Manal Habbouche, and Marta Pérez. 2003. Evaluation of ontology development tools for bioinformatics. *Bioinformatics* 19 12 (2003), 1564–71.

[25] Diana Man. 2013. Ontologies in Computer Science. *Didactica Mathematica* 31 (2013), 43–46.

[26] Deborah McGuinness and Frank van Harmelen. 2004. *OWL Web Ontology Language Overview.* W3C Recommendation. W3C. http://www.w3.org/TR/2004/REC-owl-features-20040210/.

[27] Christopher J Mungall, Heiko Dietze, and David Osumi-Sutherland. 2014. Use of OWL within the Gene Ontology. In *OWLED.*

[28] Mark A. Musen. 2015. The protégé project: a look back and a look forward. *AI Matters* 1, 4 (2015), 4–12. https://doi.org/10.1145/2757001.2757003

[29] Bijan Parsia, Peter Patel-Schneider, Markus Krötzsch, Sebastian Rudolph, and Pascal Hitzler. 2012. *OWL 2 Web Ontology Language Primer (Second Edition).* Technical Report. W3C. http://www.w3.org/TR/2012/REC-owl2-primer-20121211/.

[30] Amit Sheth and Cartic Ramakrishnan. 2004. Semantic (Web) Technology In Action: Ontology Driven Information Systems For Search, Integration and Analysis. *IEEE Data Engineering Bulletin* 26 (01 2004).

[31] Amandeep Sidhu, Tharam S. Dillon, Elizabeth Chang, and Baldev S. Sidhu. 2005. Protein Ontology Development using OWL., Vol. 188.

[32] South African Government Gazette. 2013. Protection of Personal Information Act. http://www.justice.gov.za/legislation/acts/2013-004.pdf.

[33] Michael Uschold and Martin King Aiai. 1995. Towards a Methodology for Building Ontologies.

[34] Wikipedia. 2019. Facebook–Cambridge Analytica data scandal — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Facebook-Cambridge_Analytica_data_scandal&oldid=888391351. [Online; accessed 20-March-2019].