An Integrated Knowledge Graph to Automate GDPR and PCI DSS Compliance

Lavanya Elluri, Ankur Nagar, Karuna Pande Joshi
Information Systems Department
University of Maryland Baltimore County
Baltimore, MD, USA, 21250
Email: {lelluri1, anku2, karuna.joshi}@umbc.edu

***Abstract*—Big data analytics related to consumer behavior, market analysis, opinions, and recommendation often deal with end user's derived and inferred data, along with the observed data. To ensure consumer data protection, rules defined by the European Union's General Data Protection Regulation (EU GDPR) must be adhered to by every organization using Personally Identifiable Information (PII) data for Big Data analysis. Similarly, Payment Card Industry Data Security Standard (PCI DSS) has policy guidelines specifically for organizations handling consumer's payment card data. Both data regulation policies are currently available only in textual format and require significant manual effort to ensure their compliance. We have developed an integrated, semantically rich Knowledge Graph (or Ontology) to represent the rules mandated by both PCI DSS and EU GDPR. In the Ontology, we have also identified the obligations defined in these regulations and related them with corresponding Cloud Security Alliance (CSA) controls. We have validated this Knowledge Graph against the data policies of major vendors that deal with Big Data. This Knowledge Graph that is available in the public domain can be used by Big Data practitioners to automate data protection compliance in their organization.**

Keywords: Data Protection; Ontology; General Data Protection Regulation; Organizations.

I. INTRODUCTION

Companies are analyzing large consumer datasets to determine behavior patterns related to market trends, fraud detection or for forecasting customer loyalty. Along with observed data this analysis also uses derived or inferred data and includes consumer's Personally Identifiable Information (PII) data. Moreover, rapid adoption of Cloud computing for big data analytics has also resulted in a large volume of PII data being managed and transferred across the Internet. Security and Privacy of observed or derived PII managed by vendors is of key concern to consumers.

As a result, regulatory bodies throughout the globe are releasing new data protection laws, like European Union's General Data Protection Regulation (EU GDPR) [17] and Payment Card Industry Data Security Standard (PCI DSS) [4], etc. that must be adhered to by Big Data Providers and Consumers. This spurt in data protection regulations has resulted in overwhelming legal compliance challenges of Big Data, and businesses often fixate on a single tree or branch in the forest of laws, regulations, standards, and seldom step back to gain an overall view of the compliance forest. [20]

GDPR specifies rules and policies for organizations using any EU customer data for their analytics [18]. On the other hand, any organization utilizing cardholder's data or handling transaction related to credit/debit card must follow PCI DSS guidelines. The main difference between the two is that GDPR is less specific than PCI DSS since they differ in the type of data being regulated [8].

The PCI DSS regulation deals with payment card data and cardholder information, such as debit/credit card numbers, Primary Account Numbers (PAN), and Sensitive Authentication Data (SAD) such as Card Verification Value (CVV) and magnetic stripe data, from all the major card schemes [4]. The GDPR has a broader scope and covers any PII data related to EU residents connected to their private, professional or public life. It includes personal name, home address, photo, email address, bank details, medical records, social media posts, computer's IP address. It is noteworthy that a data breach that violates PCI DSS compliance also violates the GDPR [9] [10]. On the other hand, a breach that violates GDPR compliance does not necessarily violate the PCI DSS regulation. Both GDPR and PCI DSS in the UK are regulated by the Information Commissioner's Office (ICO) [17] which investigates every data breach, be it a PII or cardholder's data.

Data protection regulations are currently available only in textual format and so require significant human time and effort to ensure compliance and thereby prevent data breaches. We envision that an integrated, semantically rich, machine processable knowledge graph (or ontology) that captures the various data compliance regulations, as they apply to Big Data on the Cloud, will significantly help in automating an organization's data compliance processes. In addition to saving organizational resources dedicated to compliance adherence, it will also help in proactively identifying data breaches. Another advantage of building this integrated knowledge graph is that potential contradictory policies in the organization can be easily identified and rectified as needed.
As a first step towards this vision of a holistic data compliance knowledge graph, we have created a semantically rich policy-based knowledge representation of the PCI DSS and GDPR regulations [17] with corresponding CSA controls [18]. We have validated this Knowledge Graph against the

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

data policies of five major vendors that deal with Big Data. This Knowledge Graph that is available in the public domain can be used by Big Data practitioners to automate data protection compliance in their organization significantly.

In section I, we described the motivation for this work, and in section II we discuss the background and related work in this area. In section III, we describe our methodology of building the knowledge graph and detail the ontology we have developed using OWL. In this section, we also discuss the text mining and NLP approaches we took to extract and populate policy documents of various cloud-based service providers as instances of our knowledge graph and present the results of our validation in section IV. We end with conclusions and future work.

II. RELATED WORK

A. Semantic Web

The Semantic Web deals primarily with data instead of documents. It allows data to be annotated with machine understandable meta-data, permitting the automation of their retrieval and their usage in incorrect contexts. Semantic Web technologies include languages such as Resource Description Framework (RDF) [21] and Web Ontology Language (OWL) [22] for defining ontologies and describing meta-data using these ontologies as well as tools for reasoning over these descriptions. These technologies can be used to provide the common semantics of privacy information and policies enabling all agents who understand basic Semantic Web technologies to communicate and use each other's data and Services effectively.

In our prior works, we developed a new integrated methodology for the lifecycle of IT services delivered on the cloud and demonstrate how it can be used to represent and reason about services and service requirements, and so automate service acquisition and consumption from the cloud [3]. We have also developed ontologies to represent legal documents pertaining to cloud data like Service Level Agreements [5] and Data Privacy policies [6]. We are now extended this work to build an integrated Data Compliance Knowledge Graph.

B. Key components of GDPR

As part of our previous work [2], we have identified the key classes of a knowledge graph to represent the GDPR rules. We have referenced the GDPR regulation available at [17] [18] for this. Key classes for this component are as follows: *'Consumers and Providers', 'Fines and Enforcement', 'Breach & Notification', 'Data Protection Officer', 'Data Subject'.*

C. Key components of PCI DSS

In our previous work, we have developed a simple ontology for the PCI DSS regulation based on the 12 requirements defined by the PCI DSS council [1][4]. The goal of the PCI DSS is to protect cardholder data wherever the card data is processed, stored or transmitted [1][4]. In general, if an organization deals in card transactions, then it must follow the key policies listed in the sections below. These policies are part of the latest PCI DSS Version 3.2 released in 2016 [1][4]. Key classes for this component are defined as follows: '*Build and maintain a Secure Network', 'Protect Cardholder Data', 'Maintain a Vulnerability Management Program', 'Implement Strong Access Control Measures', 'Regularly Monitor and Test Networks', 'Maintain an Information Security Policy'.*

III. METHODOLOGY

In this section, we describe our methodology to build and validate our integrated Bigdata compliance ontology. We aim to present a rich policy-based knowledge representation of the PCI DSS and GDPR regulations with the corresponding CSA controls. We created this Ontology using Protégé [5] which has reasoner like HermiT etc. The methodology has three phases for processing the repository and checklist of GDPR & PCI DSS respectively. Figure 1 is the representation of our architecture flow.

The three phases of our methodology are:

- **Preprocessing stage:** For both the regulations we extracted relevant chapters and key terms and then mapped them with corresponding CSA controls. A detailed explanation can be found in section A.
- **Knowledge Graph/Ontology Development:** We have developed a comprehensive Data Compliance ontology that integrates the knowledge representation for both GDPR and PCI DSS rules. Detailed information can be found in section B. For creating the knowledge graph; we utilized the Protégé tool [5].
- **Validation:** We validated the knowledge graph that is built using five publicly available organization policies dealing in PII and cardholder's data. Section C has detailed information related to this.

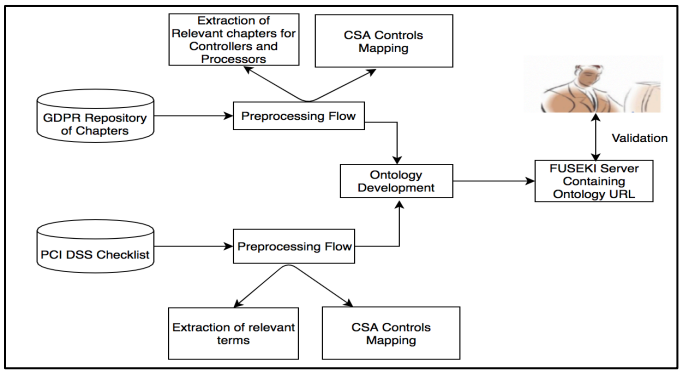

Figure 1: Architecture Flow

A. Preprocessing stage

In the first stage of our system, we extracted the repository & checklist of GDPR [17] and PCI DSS [4] respectively. In our previous work [1], we were able to extract certain key terms from the 12 PCI DSS documents and build knowledge graph accordingly. Similarly, to map

the PCI DSS policy with CSA control we looked at the key terms which were extracted from the policies and mapped all the 12 requirements to CSA controls based on keyword comparison. In the preprocessing stage, we extracted chapters 3 and 4 of the GDPR regulation which is for Consumers and Providers.

During the process, we observed the alignment of some of the rules of PCI DSS and GDPR. Both the data protection rules mandate that the organization should secure personal data. If an organization is PCI DSS compliant, then it is on track for achieving GDPR compliance as well. There are commonalities in both the data protection rules. Some of which we were able to relate include:

- Both Data protection rules focus primarily on building the secure infrastructure environment
- Both Data protection rules focus on securing personal data
- Both regulate access to personal data
- Both policies require auditing of security provisions
- Both impose hefty fines in case of breach

A breach in PCI DSS can also be regarded as a breach in GDPR. However, it is not necessary that if an organization is PCI DSS compliant, then it is also GDPR compliant. PCI DSS deals with a very small set of data- cardholder's data which consists of debit/credit card numbers, Primary Account Numbers (PAN), and Sensitive Authentication Data (SAD) such as CVVs and magnetic stripe data, from all the major card schemes [1][4].

On the other hand, GDPR has a broader scope in terms of Big Data analysis usage because it covers any PII. This PII can include any EU customer's personal details such as Name, Address, Phone numbers, Medical records location, Race, gender, birth date, Criminal convictions, etc. Figure 2 shows the mapping of the scope of GDPR and PCI DSS. After identifying similarities and differences between the regulations, we mapped GDPR rules to the CSA controls.

In this stage, we also determined the permissions and obligations for both data protection rules. The process to determine that is detailed below.

1) Permission & Obligations

Modal logic is a broad term used to cover various other forms of logic such as temporal logic and deontic logic[19]. Deontic logic labels statements containing permissions and obligations, and temporal logic defines time-based requirements. Deontic logic further consists of four types of modalities:

1. **Permissions** / Rights: Permissions are expressions or rules that describe the rights or authorizations for an entity.
2. **Obligations**: Obligations expressions are the compulsory actions that an entity must accomplish.
3. **Dispensations**: Dispensations that describe optional expressions and describe non-mandatory conditions.
4. **Prohibitions**: Prohibitions are the expressions that specify the actions which are prohibited.

To classify the data protection policies as Permissions and Obligations, we extracted certain modal keywords like 'will', 'should', 'can', 'could', 'shall', 'must' etc. These modal verbs helped us in determining whether the sentence is classified as permission or an obligation. These permissions and obligations determine how the polices in GDPR and PCI DSS affect consumer, provider and end user. To extract the modal verbs, we did a frequency count of these verbs in the GDPR policy for controllers & processors and in PCI DSS checklist. Table 1 list the frequent occurrences of verbs in both the documents.

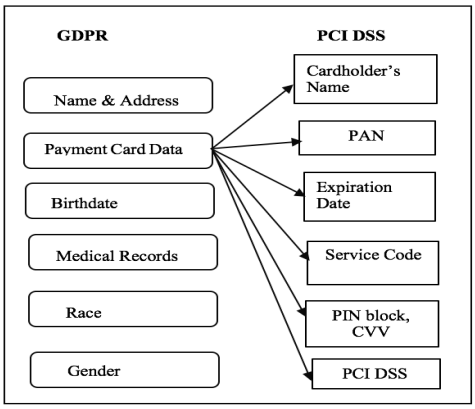

Figure 2: Mapping the scope of GDPR & PCI DSS

Modal Verbs	Occurrence
will	3
shall	119
can	1
could	1
may	36

Modal Verbs	Occurrence
must	11
Can	6
could	1
may	7
should	7

Table 1: Modal Frequency for GDPR& PCI DSS respectively

In our paper, we have used permissions & obligations to categorize sentences into any one of them. Sentences that have verbs like *'may', 'can', 'could' 'will'* were categorized as Permissions and sentences having verbs like 'shall', 'must' 'should' were categorized as Obligations. Below mentioned are some examples of our context:

Permissions (PCI DSS):

"Requirement 7: Restrict access to cardholder data by business need to know. To ensure critical data can only be accessed by authorized personnel, systems and processes must be in place to limit access based on need to know and according to job responsibilities" [4].

Obligations (PCI DSS):

"Requirement 10.7 Retain audit trail history for at least one year; at least three months of history must be immediately available for analysis" [4].

Permissions (GDPR):

"A group of undertakings may appoint a single data protection officer provided that a data protection officer is easily accessible from each establishment" [17].

Obligations (GDPR):

"The controller shall implement appropriate technical and organizational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed" [17].

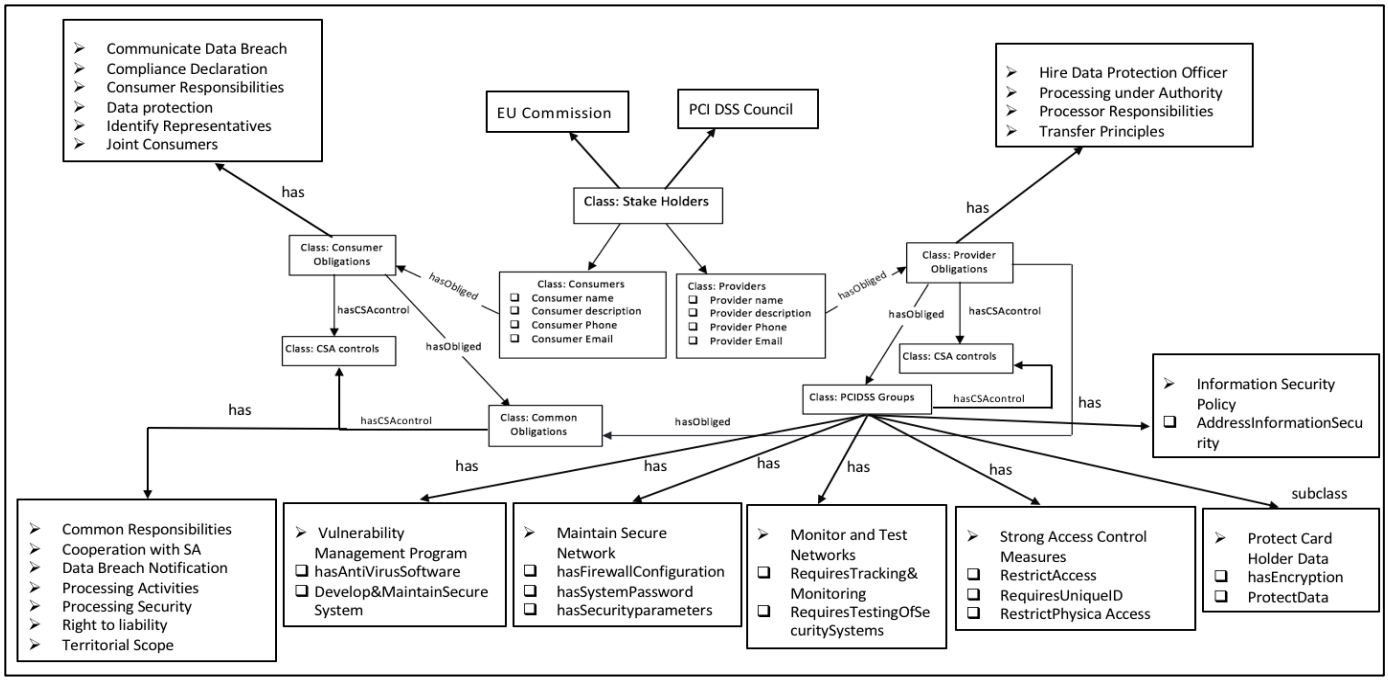

Figure 3: Ontology for GDPR, PCI DSS vs. CSA Control

2) Key terms Extraction

After identifying permissions and obligations for the data protection rules, we wanted to look for key terms both in GDPR and PCI DSS. As mentioned above, in previous work of PCI DSS [1], we extracted key terms that were important in context when an organization falls under PCI DSS compliance. For GDPR regulation, we applied the similar approach to extract the relevant key terms from the repository of controller & processor. As defined in EUGDPR [7] repository there are several key terms which should be taken into consideration when an organization is falling under GDPR compliance. We have used Python to develop the code to extract [11] the key terms from the large corpus of GDPR. In our code, we made a list of stop words which were not needed and were irrelevant to our context. Also, we did make sure that certain words like will', 'should', 'can', 'could', 'shall', 'must' were not part of stop words list since these words contribute towards defining permission & obligations expressions. This approach helped us in segregating any irrelevant terms. We made use of regular expressions which helped us in identifying the key terms which have already been shared by EU GDPR. Table 2 below shows the list of PCI DSS and GDPR key terms frequencies [1] [17].

Key terms	Frequency
maintain	10
control	13
establish	5
access	43
unauthorized	6
ensure	10

Key terms	Frequency
data subject	375
processor	528
controller	1008
profiling	46
data breach	37
personal data	1148
consent	144
notification	28

Table 2: Key terms of PCI DSS & GDPR respectively

B. Ontology Development

We used Protégé software to build the integrated Big Data compliance knowledge graph which combines CSA controls, PCI DSS, and GDPR. Figure 3 illustrates the high-level combined view of all the classes. Due to page limitations, we have restricted the description to first level classes. In our previous work, we have developed the semantically rich ontology to capture obligations of only GDPR and the associated CSA controls [2] and PCI DSS [1]. We had manually identified the key terms and extracted the obligations of Consumer, Provider and common obligations. We have now developed tools to automate the process of extracting the key terms of GDPR, PCI DSS and associating it with corresponding CSA controls from the legal texts. The main classes of our knowledge graph are:

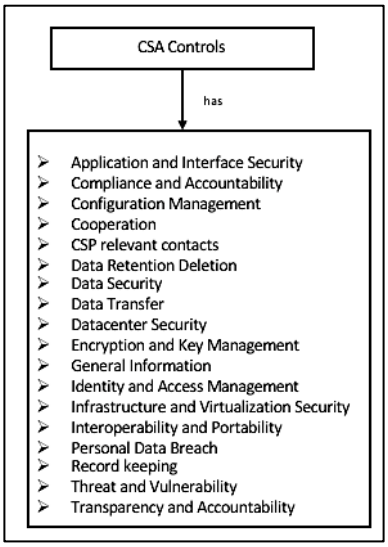

Figure 4: CSA controls subclasses

- The **Stakeholder** class is the main class that represents the key organizations that are affected by the regulations. This class has four main subclasses. These are the Big Data consumers, providers, EU Commission (regulates via GDPR) and the PCI DSS Council (regulates PCI DSS). Stake Holders class includes *hasObliged* property associated with all the Obligation classes, and *hasCSAcontrol* property has the domain as the Obligation classes and range as CSA control classes.
- The **Consumer** class represents the data users and includes properties of end users.
- **Consumer Obligations**: The consumers have obligations that they have to adhere to for GDPR. The main subclasses of the consumer obligations are Compliance declaration, responsibilities, communication of data breach, data protection, representative and other joint consumers.
- **CSA Control**: This class represents the security controls recommended by the Cloud Security Alliance [18]. It has 18 subclasses that define the various categories of Cloud security. In [2] we have presented tables that include the specific consumer, provider and common obligations common between GDPR and CSA controls. In this paper, we have related PCI DSS groups as well to CSA controls as shown in Table 3. CSA controls subclasses are in Figure 4.
- **Provider**: Provider class represents the data providers and includes properties of providing organization.
- **Provider Obligations**: The providers have a separate set of obligations that they have to adhere to for GDPR. The provider is also obligated to adhere to the PCI DSS requirements that are divided into six main classes in our Knowledge Graph.
- **Common Obligations:** Obligations in GDPR that are the responsibilities of both Consumer and Provider are represented in this class. Its main subclasses are responsibilities, cooperation, breach notification, processing activities, processing security, liability, and scope.
- The **PCI-DSS group** class consists of six main class which incorporate the 12 PCI-DSS requirements. The classes are Control Measures, Data Protection, Secure Network, Secure Policy, Monitor and Test Network, and Management Program. Each class is disjoint from other classes which means that an individual (or object) cannot be an instance of more than one of these six classes. Each of the main class has sub-classes with their properties.

C. *Validation*

For the validation process, we referenced data policies of major cloud data providers that have access to their customer PII data. These included AWS [12], Facebook [13], Google [14], Microsoft [15] and WhatsApp [16].

PCI DSS Groups	CSA controls
Build and maintain a secure network	PY-04, MOS-01, STA-03, TVM-01, IVS-12, IVS-06, MOS-19
Protect Card Holder Data	AIS-03, AIS-04, DSI-02, DSI-03, DSI-05, EKM-03, EKM-02, MOS-11, AIS-02
Maintain a Vulnerability Management Program	TVM-01, TVM-02, MOS-01, TVM-03
Implement Strong Access Control Measures	DCS-02, DCS-07, DCS-08, DCS-09, EKM-04, IAM-06, IAM-12
Regularly Monitor and test networks	CCC-03, CCC-04, CCC-05, IAM-03
Maintain an Information Security Policy	DSI-04, DCS-06, IAM-04, MOS-17

Table 3: PCI DSS Groups vs. CSA controls

Table 4 lists the organization policies used for validation. We wanted to verify if key terms and obligations specified in these data policies can be populated as instances of our data compliance knowledge graph. After downloading the publicly available data policies, we applied them to the pre-processing tools that we have created. We used the privacy/terms of service policies to look for terms similar to the ones defined by GDPR and PCI DSS. This applied approach helped us in extracting the key terms from their Terms of service/Privacy policy. We did find similar key terms in the organizational policies along with the number of times that term has occurred. The graph in Figure 5 gives us a snapshot of key terms and its count for various organizations. With the help of these terms, each organization's policies were populated as instances of our knowledge graph. The data policies are now available as an RDF graph and are machine processable. It will now be possible to automate the compliance validation by using policy reasoning engines that can alert any potential compliance violation.

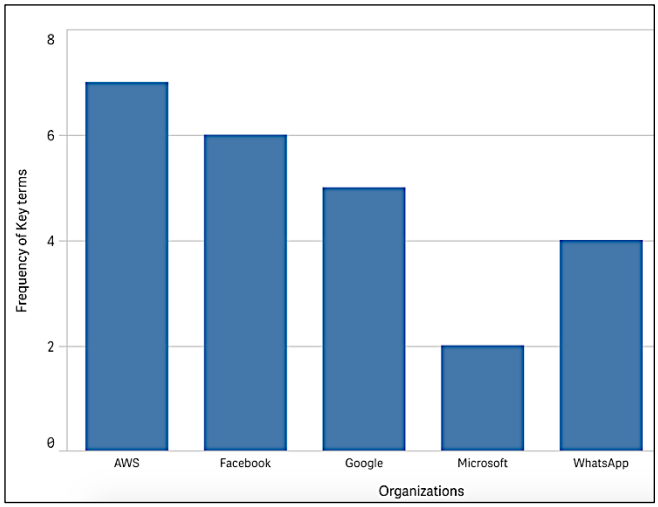

Figure 5: Validation Results

Organizations	Policies
AWS	https://d1.awsstatic.com/legal/aws-gdpr/AWS_GDPR_DPA.pdf
Facebook	https://www.facebook.com/business/gdpr
Google	https://privacy.google.com/businesses/compliance/#?modal_active=none
Microsoft	https://privacy.microsoft.com/en-us/updates
WhatsApp	https://www.whatsapp.com/legal/#privacy-policy

Table 4: List of policies for validations

IV. CONCLUSION & FUTURE WORK

Regulatory bodies throughout the globe are releasing new data protection laws to ensure data security and privacy. These data protection regulations are currently available only in textual format and so require significant human time and effort to ensure compliance. We envision that a semantically rich, machine processable knowledge graph (or ontology) that captures the various data compliance regulations, as they apply to Big Data on the Cloud, will significantly help in automating an organization's data compliance process. We have developed an integrated semantically rich, machine processable knowledge graph (or ontology) to represent knowledge embedded in the PCI DSS and GDPR regulations. We have also studied the CSA code of conduct controls and included associated GDPR articles with the CSA controls in our Ontology. We used Semantic Web technologies, Natural Language Processing (NLP) and text mining techniques to create this graph. In this paper, we describe this knowledge graph in detail along with the methodology we have used to build it. We have validated this Knowledge Graph against the data policies of five major vendors that deal with Big Data. Our knowledge graph will help Big Data practitioners to get a well-defined integrated view of the data regulations, and they can reference it as a compliance checklist. As part of our future work, we plan to build a reasoning component in our system that will automatically detect compliance violations.

V. REFERENCES

[1] A. Nagar and K. P. Joshi, "A Semantically Rich Knowledge Representation of PCI DSS for Cloud Services", In Proceedings, 6th International IBM Cloud Academy Conference ICA CON 2018, Japan, May 2018

[2] L. Elluri and K. P. Joshi, "A Knowledge Representation of Cloud Data controls for EU GDPR Compliance", InProceedings, 11th IEEE International Conference on Cloud Computing (CLOUD), July 2018.

[3] Karuna Pande Joshi et al., "Automating Cloud Services Lifecycle through Semantic technologies", Article, IEEE Transactions on Service Computing, January 2014.

[4] Payment Card Industry (PCI) Data Security Standard, Version 3.2, https://www.pcisecuritystandards.org/document_library, April 2016

[5] Musen, M.A. The Protégé project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June 2015. DOI: 10.1145/2557001.25757003.

[6] Karuna P Joshi, Aditi Gupta, Sudip Mittal, Claudia Pearce, Anupam Joshi, and Tim Finin. Semantic Approach to Automating Management of Big Data Privacy Policies. In Proceedings, IEEE BigData, 2016.

[7] EU GDPR Portal. (2018). GDPR Glossary of Terms. [online] Available at: https://www.eugdpr.org/glossary-of-terms.html [Accessed 17 Aug. 2018].

[8] GDPR and PCI DSS: How They Differ, How They're Similar and... (2018, July 10). Retrieved from http://paymentsjournal.com/gdpr-and-pci-dss/

[9] Calver, N. (2018). How the PCI DSS can help you meet the requirements of the GDPR. [online] IT Governance Blog. Available at: https://www.itgovernance.co.uk/blog/how-the-pci-dss-can-help-you-meet-the-requirements-of-the-gdpr/

[10] Jones, A. and I.S. Partners, L. (2018). 4 Ways to Use PCI DSS to Achieve GDPR Compliance | I.S. Partners. [online] I.S. Partners. Available at: https://www.ispartnersllc.com/blog/4-ways-to-use-pci-dss-to-achieve-gdpr-compliance/

[11] PyPI.(2018). rake-nltk [online] Available at: https://pypi.org/project/rake-nltk/ [Accessed 17 Aug. 2018].

[12] Anon, (2018). [ebook] Available at: https://d1.awsstatic.com/legal/aws-gdpr/AWS_GDPR_DPA.pdf [Accessed 17 Aug. 2018].

[13] Facebook Business. (2018). General Data Protection Regulation. [online] Available at: https://www.facebook.com/business/gdpr [Accessed 17 Aug. 2018].

[14] Privacy.google.com. (2018). Compliance | How Google complies with data protection laws. [online] Available at: https://privacy.google.com/businesses/compliance/#?modal_active=none [Accessed 17 Aug. 2018].

[15] Privacy.microsoft.com. (2018). Change history for Microsoft Privacy Statement – Microsoft privacy. [online] Available at: https://privacy.microsoft.com/en-us/updates [Accessed 17 Aug. 2018].

[16] WhatsApp.com. (2018). WhatsApp Legal Info. [online] Available at: https://www.whatsapp.com/legal/#privacy-policy [Accessed 17 Aug. 2018]

[17] "General Data Protection Regulation (GDPR) – Final text neatly arranged." General Data Protection Regulation (GDPR), gdpr-info.eu/.

[18] Cloud Security Alliance Releases Code of Conduct for GDPR Compliance. (n.d.).from https://www.morganlewis.com/blogs/sourcingatmorganlewis/2017/11/cloud-security-alliance-releases-code-of-conduct-for-gdpr-compliance

[19] Modal Logic: http://plato.stanford.edu/entries/logic-modal/

[20] Michael R. Overly, Legal compliance challenges of Big Data: Seeing the forest for the trees, https://www.csoonline.com/article/2883796/big-data-security/legal-compliance-challenges-of-big-data-seeing-the-forest-for-the-trees.html, last retrieved 8/19/2018

[21] "Resource description framework (RDF)." [Online]. Available: http: //www.w3.org/RDF/

[22] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.