OPPO: An Ontology for Describing Fine-Grained Data Practices in Privacy Policies of Online Social Networks

Sanonda Datta Gupta*, Torsten Hahmann

School of Computing and Information Science, University of Maine, ME, USA

Abstract

Privacy policies outline the data practices of Online Social Networks (OSN) to comply with privacy regulations such as the EU-GDPR and CCPA. Several ontologies for modeling privacy regulations, policies, and compliance have emerged in recent years. However, they are limited in various ways: (1) they specifically model what is required of privacy policies according to one specific privacy regulations – GDPR; (2) they provide taxonomies of concepts but are not sufficiently axiomatized to afford automated reasoning with them; and (3) they do not model data practices of privacy policies in sufficient detail to allow assessing the transparency of policies. This paper presents an OWL **O**ntology for **P**rivacy **P**olicies of **O**SNs – OPPO – that aims to fill these gaps by formalizing detailed data practices from OSNS' privacy policies. OPPO is grounded in BFO, IAO, OMRSE, and OBI, and its design is guided by the use case of representing and reasoning over the content of OSNs' privacy policies and evaluating policies' transparency in a greater detailed.

Keywords

Privacy Policy, Web Ontology Language (OWL), Conceptual Modeling, Data Practices

1. Introduction

The widespread use of Online Social Network (OSN) raises many privacy concerns, such as storing users' personally identifiable information (PII) longer than required or without following proper security mechanisms. To mitigate such concerns, various privacy regulations (eg., GDPR [1] and CCPA[2]) have been introduced, which require organizations to be transparent about their data practices (collection, processing, or storage of users' data). An essential aspect to comply with these regulations are the privacy policies that describe the data practices of OSNs. However, privacy policies are usually long and complex and often do not explain the data practices in sufficient detail [3, 4]. For instance, a policy may mention that it follows best practices in security without actually specifying what mechanisms it uses and to which types of data (personal vs. non-personal) they apply. Such generic or overly vague descriptions are indicators of an OSN's *lack of transparency* in informing the users of its *actual* data practices. In this paper, we present the **O**ntology for **P**rivacy **P**olicies of **O**SNs (OPPO) that is designed to encode the data practices from the privacy policies of OSNs (and other companies) in as much detail as possible. The ontology is intended to let the OSN, users, and others (e.g. privacy

Ontology Showcase and Demonstrations Track, 9th Joint Ontology Workshops (JOWO 2023), co-located with FOIS 2023, 19-20 July, 2023, Sherbrooke, Québec, Canada.

*Corresponding author.

sanonda.gupta@maine.edu (S. D. Gupta); torsten.hahmann@maine.edu (T. Hahmann)

advocacy groups) query the formal ontological representations of OSNs' data practices for specific details or the general level of detail they provide.

Prior research on formal representations of online privacy has focused on extracting rights and obligations to check compliance with regulations [5, 6, 7] and developed privacy related vocabularies and taxonomies [8, 9, 10, 11, 12]. However, most of these works miss more detailed yet important concepts and relations, such as response time for data rectification or erasure requests, retention duration (limited or unlimited) for different data types (personal vs. non-personal), or security mechanisms that are applied to different storage entities (device vs. data center). These details are not omitted just for a lack of methods that are able to extract such details but, more fundamentally, the ontologies (e.g., Pronto [8] and PrivOnto [10]) lack important concepts for modeling data practices at such levels of details in the first place.

Towards the goal of filling this gap, we develop OPPO as an upper-level ontological model for the domain of privacy policies and data practices. It is grounded in top-level ontological distinctions and capable of describing data practices at more granular levels. As a proof-of-concept of OPPO's ability to describe fine-grained data practices, we focus on introducing classes and properties for capturing data retention, storage, and security practices in detail, including aspects such as how users can request data rectification or what types of security mechanisms are applied to various data types. OPPO is developed following the methodology METHONTOLOGY [13] and is guided by a set of 45 competency questions, manual analysis of the GDPR (Art. 5, 13–17, and 32), CCPA (Art. 1798.100, 1798.125, 1798.135, 1798.140, 1798.81.5), and the privacy policies of ten major OSNs[1]. OPPO is grounded in the Basic Formal Ontology (BFO) [24] as well as the Information Artifacts Ontology (IAO) [25], the Ontology for Biomedical Investigations (OBI) [26], and the Ontology for Medically Related Social Entities (OMRSE) [27] as widely-used specializations of BFO. OPPO further reuses parts of the Data Privacy Vocabulary (DPV) [9] and the OWL Time ontology [28]. The ontology is shared via GitHub at https://github.com/SanondaDattaGupta/OPPO-Ontology; it currently contains 60 classes and relations and is formalized in OWL2 [29][2]. To demonstrate its use and to verify its consistency, we populate the ontology with data from the privacy policy of Telegram as an example. Its logical consistency has been verified using the OWL Reasoner Hermit [30]. It has been validated on a subset of the guiding competency questions, expressing and executing them as SPARQL queries over the Telegram data.

[1]WhatsApp [14], Signal [15], Telegram [16], Twitter [17], Tiktok [18], WeChat [19], SnapChat [20], Reddit [21], Pinterest [22], and Instagram [23]

[2]`OPPO_Ontology_Full_Import.ttl` fully imports the ontologies BFO, IAO, OBI, OMRSE, DPV and OWL-Time from which concepts are reused or extended. Because only a small fraction of these concepts are relevant to OPPO, we also provide a version of OPPO that imports only small subsets of IAO, OBI, OMRSE and DPV that are limited to the classes (and their superclasses) and relations that are reused or refined. This version is called `OPPO_Ontology_Minimal_Import.ttl` and gives a better impression of the concepts introduced by OPPO when loaded into an ontology editor like Protégé.

2. Use Case

OPPO is designed as a tool for OSN users, privacy researchers, policy makers, regulatory bodies, and organizations to help analyze how OSNs' privacy practices relate to different privacy regulations, identify inconsistencies between practices, or better judge their transparency – just to name a few examples. Generally, OPPO intends to serve as a tool to formulate and answer questions about the data practices described in privacy policies and, thus, achieve greater transparency while dealing with often long and complex policies. For the first released version of OPPO described here, we focus on capturing the practices that describe storage and retention of data, including the security mechanisms an OSN may employ.

To guide OPPO's development, we have defined 45 competency questions (CQ) [31][3]. The competency questions help define what terms should be included in the ontology and, later on, they serve as questions to measure whether the ontology can express and answer the questions. These questions include simple queries such as *Where does the OSN store my information?* to more complex questions such as *Which of my personal information will still be available after I delete my account?* Note that our competency questions are not limited to questions about data storage and retention practices but also cover other data practices such as collection or notification mechanisms as well. However, in this paper, we focus on the 27 CQs that pertain to data storage, retention, and security practices. In the following, we present three of those competency questions to illustrate their coverage and the kind of answers we would expect.

1. *Where does the social media site/company store my message?*
   An OSN may store users' personal information in several places which differ both in the type of storage (on a device vs. a data center) and the storage location (Europe vs. United States). Hence, this CQ enables us to capture the location and physical devices where an OSN may store users' personal information.
2. *What contents may be stored for a maximum of 12 months?*
   Both GDPR and CCPA require organizations to explicitly mention the retention period of the collected personal information. While analyzing the ten privacy policies, we found that some privacy policies (such as Telegram), specify different duration description (such as maximum of X months or as long as they need) for different information types. Thus, this CQ can help us capture and query such details about retention practices and also identify where OSNs lack specificity in such details.
3. *What types of security mechanism are applied to my photos and private chats?*
   Privacy regulations require organizations to describe how they ensure the security of the retained information. While most organizations *briefly* describe the security mechanisms, some organizations (such as Signal, WhatsApp, and Telegram) do explicitly mention different security mechanisms (hashing mechanism vs. encryption mechanism) that they apply to different information types (public chat vs. media). Moreover, certain information types are more sensitive (such as bio-metric or health information), and thus may pose higher privacy risks, if not stored following sufficiently secure mechanisms. Hence, this CQ exemplifies the kind of competency questions that help evaluate how transparent and concrete OSNs are about their security and data practices.

[3]The complete set of competency questions are available from the GitHub repository.

3. Related Work

In recent years, several vocabularies, ontologies, and conceptual models for modeling privacy policies, regulations, and compliance have been introduced [32, 33, 34, 11, 35] but are primarily concerned with modeling rights and compliance issues with respect to regulations (e.g., BPR4GDPR [11] and SPECIAL [35]) and permissions and prohibitions (e.g., CDMM [12], ODRL [36]). Most closely related to our work are PrivOnto [10], PrOnto [8], and DPV [9], which are vocabularies and conceptual models that also cover – to some extent – data practices. However, they do not capture more granular aspects of data practices such as the duration for which data is stored, the location it is stored in, the specific security protocols that are employed, or the specific process to request to rectify user data. While outlining such detailed practices is not explicitly required by most existing privacy regulations, it can help organizations be *more transparent*. In the following, we briefly describe the differences in scope, representation, and other limitations of PrivOnto, PrOnto and DPV that distinguish them from OPPO.

The purpose of PrivOnto [10] is quite different from OPPO: it is designed to annotate paragraphs from policies with concepts from their vocabulary as a kind of semantic tagging but PrivOnto does not allow encoding the content of the policies in a formal representation that it can be automatically reasoned with later on. Moreover, the development of PrivOnto predates the release of GDPR and other privacy regulations; hence the concepts do not tie well to data practices mentioned in regulations.

PrOnto [8] overlaps with OPPO in aspects such as the modeling of different privacy-related data types but lacks a full coverage of types of personal data, such as financial, identity or activity data that we add. A more fundamental difference between PrOnto and OPPO is that Pronto is encoded in the LegalRuleML [37] language, a defeasible logic, which restricts the ability to automatically reason with PrOnto[4] or to integrate it with other non-legal ontologies.

DPV [9] provides a comprehensive set of privacy-related terms, including for different data types, purposes, legal entities, and data processing operations (e.g. transmit or store) and encodes it in OWL2. However, it still is more of a taxonomy that does not describe or restrict how these concepts are related to one another axiomatically. Moreover, DPV does not tie the concepts to top-level concepts and does not use ontological analysis tools to structure its taxonomies. For example, while DPV loosely defines a group of *legal roles*, data subjects or regulation authorities are not treated as such. Likewise, multiple distinctions between different types of personal data are made, but these distinctions are not integrated into a single coherent taxonomy. However, we reuse some of DPV's terms where appropriate. But OPPO still introduces an additional 60 concepts such as specific security mechanisms (encryption and hashing), different data types (personal and non-personal), duration descriptions (definite and indefinite) and location descriptions (storage type and spatial location). OPPO further introduces concepts to model specific practices, such as for describing how an OSN deals with requests to rectify or delete data, and the associated properties (e.g., request and response type and response delay). These allow capturing data practices of OSNs more fully and in greater detail, thus enabling querying and evaluating the transparency and level of detail across OSNs.

[4]Reasoning with OWL2 ontologies is supported by a wide range of off-the-shelf reasoners but we are not aware of any tools that can reason directly with LegalRuleML. Instead, specifications in LegalRuleML first need to be converted to another format to enable reasoning as described in [38].

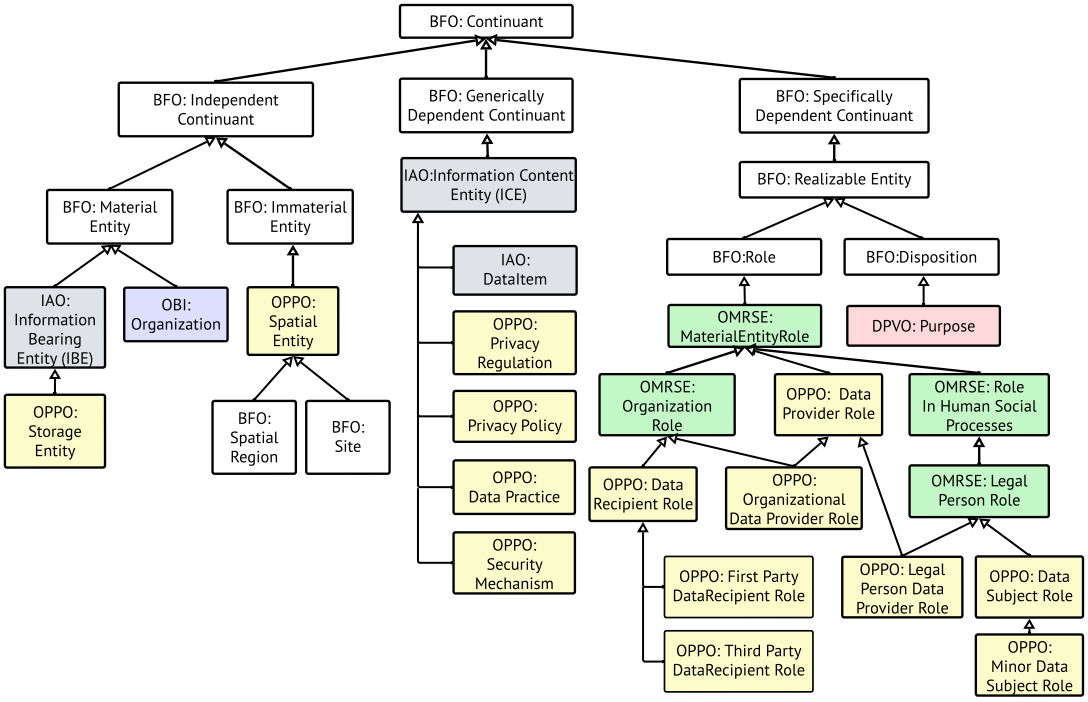

Figure 1: OPPO's high-level concepts and how they extend concepts from BFO, IAO, OBI, and OMRSE. Concepts in yellow are introduced by OPPO, while all other concepts are being reused.

4. The Conceptual Model of OPPO

OPPO builds and reuses a number of existing ontologies. We directly reuse the classification of purposes from DPV [9] and time-related concepts (e.g. to capture storage duration) from OWL-Time [39] as discussed in more detail later on. In addition, we use top- and mid-level ontologies to ground OPPO concepts. Because privacy policies and data practices are primarily about information, we rely on the Information Artifact Ontology (IAO)[5] [25] as the main source for upper-level reference concepts while OMRSE [27] and OBI [26] provide additional high-level role classifications of interest to OPPO. All of IAO, OMRSE, and OBI build on the top-level ontology BFO [24]. Thus, to maximize compatibility we also reuse BFO as top-level ontology. Figure 1 demonstrates how OPPO's most generic concepts relate to BFO and the other ontologies. In the remainder of this section we will provide overview of OPPO's key concepts and their relationship to top- and mid-level ontologies.

Information content entities: Privacy policies describe the data practices an OSN employs to collect, store and process users' data in compliance with privacy regulations. We have identified the four underlined concepts as central to this endeavor, which all are modeled as subclasses of `IAO:InformationContentEntity`: (i) a `PrivacyRegulation` such as GDPR or

[5]Available from https://github.com/information-artifact-ontology/IAO

CCPA that regulate the storage, collection, and processing of user's personal information; (ii) a `IAO:DataItem` is a piece of personal information directly or indirectly associated with a user (e.g., name or age); (iii) a `DataPractice` describes a way that an OSN processes user data (e.g., how long or where it retains data); and (iv) a `PrivacyPolicy` describes the entirety of the data practices of an OSN or, more generally, any kind of organization. Note that by modeling both `PrivacyPolicy` and `DataPractice` as subclasses of `InformationContentEntity` we make the intentional choice to treat them as the content of a privacy policy (`PrivacyPolicy`) or a portion of that content that describes a specific practice (`DataPractice`). They are distinct from the text itself and from what the OSN actually does in practice with the data. Data practices may mention specific techniques or tools employed to protect user data, which are modeled as `SecurityMechanisms`[6].

Roles: We have identified three distinct roles relevant to privacy policies and the data practice described therein, which align well with roles distinguished in the OMRSE ontology. (i) `DataSubjectRole` is a role played by a person whom the collected data is about (e.g. a user's date of birth or a message they posted). It specializes `OMRSE:LegalPersonRole`. (ii) `DataRecipientRole` is a role played by an organization (*OBI:Organization* in OBI's terms) that receives information either directly from a person or from a third party. It specializes `OMRSE:OrganizationRole`. (iii) `DataProviderRole` is a role played by either a person or an organization that shares data with others. To distinguish whether the data is shared by a person or an organization, we introduce the subclasses `LegalPersonDataProviderRole` and `OrganizationalDataProviderRole` that also specialize `OMRSE:OrganizationRole` and `OMRSE:LegalPersonRole`, respectively. Note also that users and organizations can play multiple roles for a particular piece of data, for example, a user can share data about themselves, in which case the user acts in both a `DataSubjectRole` and a `DataProviderRole`. Likewise, one organization can act in a `DataRecipientRole` when it receives some data from a third party and in a `DataProviderRole` when sharing data with the same or other third parties.

These roles can be further refined as shown in the right half of Figure 1. One additional refinement is motivated by regulations, such as GDPR and CCPA, imposing stricter conditions on handling data from minors, that is, users under a certain age (which may be defined differently by different regulations). Thus, we introduce `MinorDataSubjectRole` as a subclass of `DataSubjectRole`. Another distinction is between two kinds of `DataRecipientRoles`. An organization may act in a `FirstPartyDataRecipientRole` role when it receives data *directly* from a person in which case the person acts in both a `DataSubjectRole` and a `DataProviderRole`. An organization may also share the collected data with other organizations according to its own data practices. Any other organization that receives such data then acts in the `ThirdPartyDataRecipientRole` and is bound not only by its own privacy policies but also by the policy of the organization that it receives the data from. For example, if an organization that is a `FirstPartyDataRecipientRole` for some piece of data says that it shares data only for specific purposes with third parties, then these third parties are expected not to share that information with others for any other purposes either.

[6]For brevity, we omit the `OPPO:` namespace; concepts and properties without a namespace are implicitly assumed to be within that namespace.

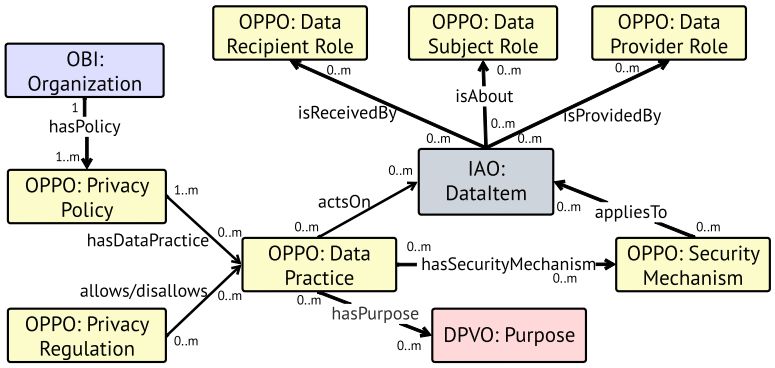

Figure 2: The core concepts in OPPO and the relations between them.

Purposes: Many privacy regulations, such as GDPR, require privacy policies to outline for what purpose an organization collects or processes data. For instance, GDPR Art. 5 [1] states that an organization may collect different information types for different purposes such as archiving statistical or research purpose. To model these different purposes we adopt the generic concept `DPVO:Purpose` from the DPV ontology [9], which is a subclass of *BFO:Disposition* in BFO. Thus, all specialized purposes from DPV can be reused as well.

Relations among OPPO's core concepts: Figure 2 shows the key relations among the core concepts of OPPO, forming its central Ontology Design Pattern (ODP). It consists of the following relations. A `PrivacyPolicy` of an OSN (modeled as an `OBI:Organization`) contains data practices ((`hasDataPractice`) relation to `DataPractice`) that describe how the OSN collects, stores, and processes users' data. These data practices may be explicitly allowed or disallowed (`allows` and `disallows`) by different `PrivacyRegulation`. `DataPractices` apply to (`actsOn`) specific kinds of data (`IAO:DataItem`), such as specific subclasses of data (e.g. `DemographicPersonalData` or `AnonymizedData` as elaborated in Figure 3) or data that is constrained in other ways, for example, what kind of legal person the data is about (a minor or not), what purpose it is collected for, or how it was received. More specifically, `IAO:DataItems` can be linked to specific data subjects, such as who the data is about (`isAbout` linking to a `DataSubjectRole`), data providers (`isProvidedBy` linking to a `DataProviderRole` describing who provides the data), and data recipients (`isReceivedBy` linking to a `DataRecipientRole` describing who receives the data). Because an OSN may describe different practices for storing or collecting data for distinct purposes (`DPVO:Purpose`), such as archiving or marketing purpose, a practice is related to purposes via the `hasPurpose` relation. By specifying both purposes and specific types, instances of the `DataPractice` class can capture that certain `DataItems` (e.g. personal data or statistical data) are stored or processed only for certain purposes. Similarly, an OSN may employ different security mechanism in different `DataPractice` and thus, implicitly, for different kinds of `DataItems`. For example, an organization may apply end-to-end encryption mechanism to biometric data while using a pseudonymization mechanism for personal technical data such as IP addresses. We capture this by relating `SecurityMechanisms` to `IAO:DataItem` using the `appliesTo` relation. In the next subsections, we will discuss further refinements of

IAO:DataItem, DataPractice, and SecurityMechanism.

4.1. Data Item Module

To describe the different kinds of data collected by an OSN, we reuse and refine IAO:DataItem [25] as shown in Figure 3. At the highest level, we distinguish data related to an individual (IndividualData), such as a name, age, or credit card information, from data that is an aggregated across multiple individuals (AggregatedData). IndividualData can be further distinguished based on whether it is anonymous, that is, data that cannot be used to personally identify any specific individual person (AnonymizedData). Data which may – directly or indirectly – reveal an individual's identity falls into the complementary class of PersonalData. It includes, for example, photos, fingerprints, posts, reviews, location, or credit card information. Based on the analysis of GDPR, CCPA, privacy policies, and prior works [40, 9], we have identified thirteen subclasses of PersonalData as shown in Figure 3. Their full definitions are provided in the ontology using the skos:definition relation. One noteworthy concept is that of PseudonymizedPersonalData, which is widely considered to be still PersonalData, though any personal identifiers have been replaced by a pseudo-identity. But it is different from AnonymizedData in that the data can be still ascribed to an individual by anyone who knows the mapping (or mapping algorithm) between pseudonyms and personal identifiers. We also include dpvo:InferredPersonalData from DPV as subclass of PersonalData. It includes any new data that is derived from existing data (e.g., demographic information from the browsing history) and which may, directly or indirectly, identify an individual.

Finally, two distinct subclasses of AggregatedData are included in OPPO: (i) statistical aggregations of user data (StatisticalData), such as the number of views on a product page or the number of likes of a post, and (ii) artificial data produced to mimick real user data (dpvo:SyntheticData).

4.2. Data Practice Module

As mentioned in Section 1, the upper level of OPPO is currently only refined to the extent needed to model fine-grained data storage and retention practices (DataStoragePractice) and security practices (DataSecurityPractice). For each of them, we distinguish three subclasses as shown in Figure 4 and explained next.

Data Storage Practices: DataStoragePractices can specify restrictions on the duration of the storage, the location of the storage, and how to get stored data corrected or deleted. While all storage practices can give such details, those that do fall into a of three subclasses. A DataStorageDurationPractice must specify the duration of the stored data, which may be definite or indefinite. A DataStorageLocationPractice includes restrictions on where or how the data is stored. For instance, it may apply to data stored in specific geographic locations (e.g., EU-GDPR imposes restrictions on data practices while the data is stored outside of the EU) or sites of a specific company. We use the new concept SpatialEntity that generalizes both BFO:SpatialRegion and BFO:Site (as shown in Figure 1) as location to remain

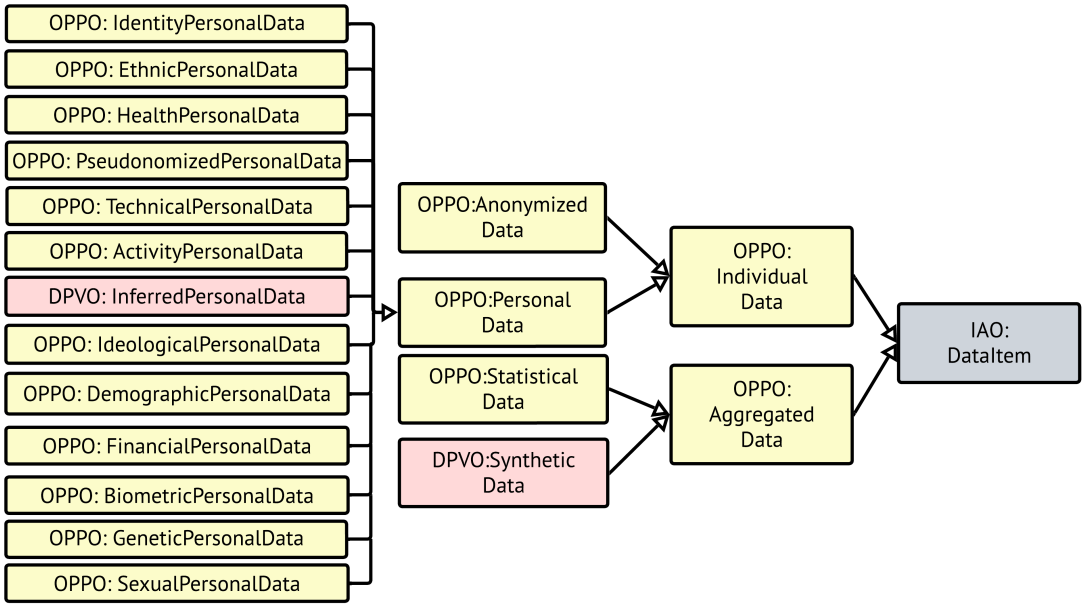

Figure 3: OPPO's taxonomy of subclasses of `IAO:DataItem`. The yellow concepts are introduced by OPPO; all other concepts are reused from DPVO and IAO.

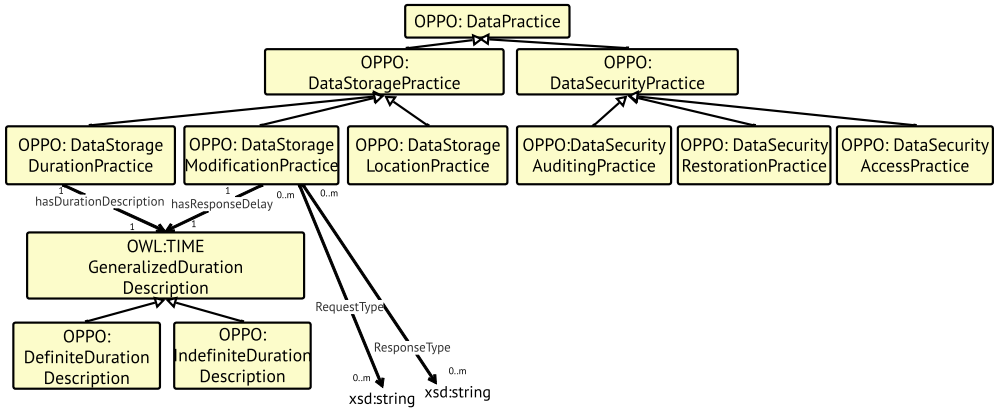

Figure 4: OPPO's taxonomy of `DataPractices` with indicative relations for the `DataStoragePractices`.

flexible and compatible with how different BFO-based ontologies may specify locations. Alternatively, a `DataStorageLocationPractice` may specify the kind of physical infrastructure (`StorageEntity`) the data is stored in, such as a data center or the user's device.

`DataStorageModificationPractice` is a subclass of `DataStoragePractice` that specifically captures practices that allow the modification (i.e., correction or deletion) of stored data. It provides relations to specify how users can request to rectify inaccurate data (a data property `RequestType`), how the OSN may respond to such requests (a data property `ResponseType`), or

how fast long the OSN may take to process such data rectification or erasure request (using `hasResponseDelay`). To capture different data retention practices as well as response delays, we reuse the `TIME:GeneralizedDurationDescription` concept and refine it by introducing `DefiniteDurationDescription` and `IndefiniteDurationDescription` as our own specializations. We refer to the ontology for details.

Data Security Practices: `DataSecurityPractices` are organizational data practices that are followed to maintain security to the collected/stored data. OPPO distinguishes three classes of security practices (`DataSecurityPractice`). A `DataSecurityAuditingPractice` is a practice that inspects whether and how the organization maintains proper safeguard mechanisms while collecting, storing, or processing personal data. A `DataSecurityRestorationPractice` is a practice that discusses how data will be recovered if the data has been lost, stolen, or compromised in other ways. A `DataSecurityAccessPractice` is a practice that limits access to the data, thus preventing unauthorized access.

4.3. Security Mechanism Module

Both GDPR and CCPA require organizations to apply suitable *techniques or tools* (referred to as `SecurityMechanism` in OPPO) to ensure the security of the collected data. The regulations themselves distinguish two types of security mechanisms: (i) `PseudonymizationMechanisms` that replace personal identifiers with a pseudo identity; and (ii) `EncryptionMechanisms` that make personal data unintelligible without the necessary keys for decryption. Our analysis of the ten OSN privacy policies, however, identified additional security mechanisms that are employed by these social networks. For instance, Signal's privacy policy [15] states that it applies cryptographic hashing mechanisms to collected data before transmitting it to their server. As another example, Telegram's policy [16] explicitly states that it employs (if the user enables it) two-factor authentication mechanisms to limit unauthorized access to their data. As a result, OPPO also distinguishes `HashingMechanism` and `AuthenticationMechanism` as two additional subclasses of security mechanisms, with further subclasses for `AuthenticationMechanism`. The hierarchy of subclasses of `SecurityMechanism` is shown in Figure 5.

5. Formalization and Evaluation

The ontology is encoded using the Web Ontology Language (OWL2) as computer-interpretable format. The axiomatization currently contains 60 new classes, 17 object properties, and 271 logical axioms[7] The syntactic correctness of the ontology has been verified with a simple RDFS validator[8]. We further checked for common pitfalls in the ontology such as missing domain and range restrictions using the OntOlogy Pitfall Scanner (OOPS!) [41].

Verification: We used the HermiT [30] OWL2 reasoner that is provided with Protégé to check the ontology for logical consistency. In the current version, no inconsistencies are found nor

[7]The numbers are based on a core version of the ontology that only replicates some upper level BFO, IAO and OMRSE concepts, but those have been excluded in the concept and property counts.
[8]http://rdfvalidator.mybluemix.net/

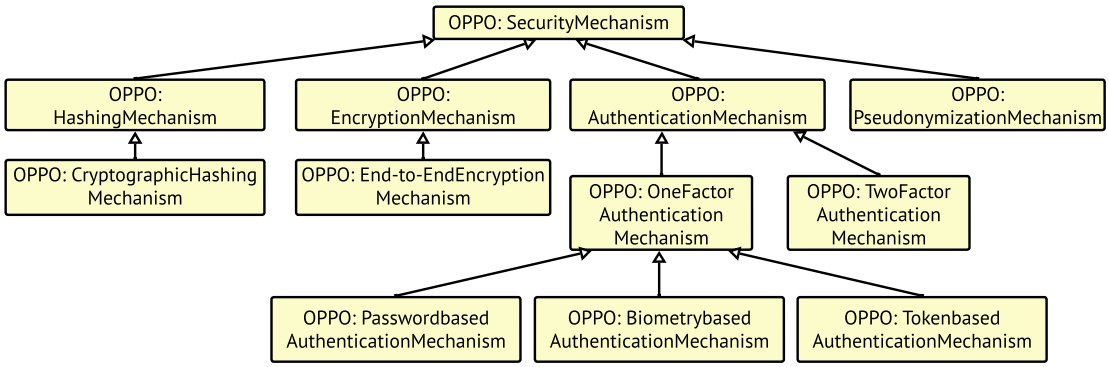

Figure 5: OPPO's taxonomy of `SecurityMechanism` and its subclasses.

are any classes inferred as being equivalent in any problematic ways, such as being equivalent to `owl:Thing`. Additionally, we manually created instances of the classes and properties from Telegram's real privacy policy[9]. When re-checked the consistency of this dataset together with the ontology using the HermiT reasoner.

Validation: To test the expressivity of OPPO, we were able to express 15 out of a total of 27 competency questions as SPARQL queries[10]. The remaining questions will guide the refinement and further development of the ontology. We loaded the ontology and the sample dataset into a GraphDB instance [42] that supports RDF and OWL reasoning. We selected the OWL-RL (Optimized) ruleset provided by GraphDB that implements OWL2 reasoning with the limitations described by the OWL2 RL profile. We executed each SPARQL query and analyzed the results to ensure that they match what we expect for our sample data. This validated that the ontology is sufficiently expressive to adequately encode and answer these competency questions. For instance, Figure 6 shows one example of a CQ that focuses on capturing the *specific data types* that are being stored by Telegram for a *maximum of 12 months*. The output indicates that Telegram stores four types of data for a maximum of 12 months. A more in-depth evaluation of the CQs on a larger dataset will be completed in the near future.

6. Conclusion and Future Work

In this paper, we presented OOPO as an extensible ontology for the privacy domain that is designed to model and formally encode detailed data practices as described in OSNs' privacy policies. As a proof-of-concept, the ontology provides a core pattern that connects privacy policies and their contained data practices to data items that are described by their data types and roles. The pattern identifies refined data practices based on the kinds of constrains they impose (e.g. the duration, location, or type of practice) in order to allow formally representing and reasoning over data practices of different privacy policies of online social networks and

[9]The dataset can be found in our GitHub repository.
[10]The SPARQL queries are provided in our GitHub Repository.

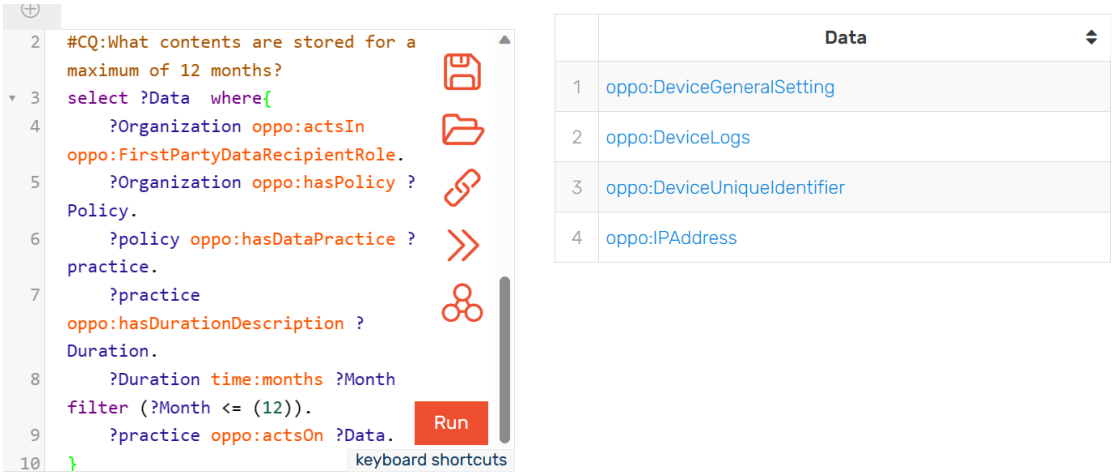

Figure 6: Encoding of a sample competency question as a SPARQL Query and the answer when executed within GraphDB over the ontology and test dataset.

similar companies. The ontology leverages and connects to a number of existing ontologies based on the Basic Formal Ontology (BFO). OPPO is encoded in OWL2 and provided as an open source resource to the community. We evaluated the ontology's logical consistency with some small but real dataset and using a set of competency questions. The competency questions were encoded in SPARQL, with the answers validating that the ontology can indeed express and answer queries about the details of data practices from privacy policies.

In the future, we plan to improve OPPO by (i) extending the ontology with other practices, such as those related to data collection and sharing; by (ii) constructing larger datasets from multiple privacy policies in order to evaluate and compare their degrees of transparencies while also further evaluating the ontology; and by (iii) expanding the ontology to also model different privacy regulation in order to pose and answer questions about compliance and non-compliance between individual privacy policies and different privacy regulations, which is becoming increasingly important as more regulations emerge.

Acknowledgments We thanks the anonymous reviewers for the helpful suggestions to improve the final version of the paper.

References

[1] European Union, The EU GDPR - Article 14, 09-29-2019. URL: https://eugdpr.org.
[2] State of California, California Consumers' Privacy Act, 01-01-2020. URL: https://oag.ca.gov/privacy/ccpa.
[3] F. H. Cate, The limits of notice and choice, IEEE Security & Privacy 8 (2010) 59–62.
[4] A. M. McDonald, L. F. Cranor, The cost of reading privacy policies, Isjlp 4 (2008) 543.
[5] T. D. Breaux, A. I. Anton, M. W. Vail, Towards regulatory compliance: Extracting rights

and obligations to align requirements with regulations, Technical Report, North Carolina State University, Dept. of Computer Science, 2006.
[6] N. Kiyavitskaya, N. Zeni, T. D. Breaux, A. I. Antón, J. R. Cordy, L. Mich, J. Mylopoulos, Automating the extraction of rights and obligations for regulatory compliance, in: International Conference on Conceptual Modeling, Springer, 2008, pp. 154–168.
[7] T. Yue, L. C. Briand, Y. Labiche, A systematic review of transformation approaches between user requirements and analysis models, RE 16 (2011).
[8] M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, L. Robaldo, Pronto: privacy ontology for legal reasoning, in: International Conference on Electronic Government and the Information Systems Perspective, Springer, 2018, pp. 139–152.
[9] H. J. Pandit, A. Polleres, B. Bos, R. Brennan, B. Bruegger, F. J. Ekaputra, J. D. Fernández, R. G. Hamed, E. Kiesling, M. Lizar, et al., Creating a vocabulary for data privacy, in: Proceedings of the OTM 2019 Confederated Intern. Conf., Springer, 2019, pp. 714–730.
[10] A. Oltramari, D. Piraviperumal, F. Schaub, S. Wilson, S. Cherivirala, T. B. Norton, N. C. Russell, P. Story, J. Reidenberg, N. Sadeh, PrivOnto: A semantic framework for the analysis of privacy policies, Semantic Web 9 (2018) 185–203.
[11] G. Lioudakis, E. Papagiannakopoulou, M. Koukovini, N. Dellas, K. Kalaboukas, L. Bracciale, E. Raso, G. Bianchi, P. Loreti, P. Barracano, et al., GDPR compliance made easier: the BPR4GDPR project, ARIS2 - Advanced Research on Information Systems Security 1 (2021) 5–23.
[12] K. Fatema, E. Hadziselimovic, H. J. Pandit, C. Debruyne, D. Lewis, D. O'Sullivan, Compliance through informed consent: Semantic based consent permission and data management model., PrivOn@ ISWC 1951 (2017) 1–16.
[13] M. Fernández-López, A. Gómez-Pérez, N. Juristo, Methontology: from ontological art towards ontological engineering, in: Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series, AAAI, 1997.
[14] WhatsApp, Whats app privacy policy, 12-19-2022. URL: https://www.whatsapp.com/legal/privacy-policy/?lang=en.
[15] Signal, Signal privacy policy, 12-19-2022. URL: https://signal.org/legal/.
[16] Telegram, Telegram privacy policy, 12-19-2022. URL: https://tinyurl.com/39crcpat.
[17] Twitter, Inc., Twitter privacy policy, 12-19-2022. URL: https://twitter.com/en/privacy.
[18] Tiktok, TikTok privacy policy, 2023-21-03. URL: https://www.tiktok.com/legal/page/us/privacy-policy/en.
[19] WeChat, WeChat privacy policy, 2022-09-09. URL: https://www.wechat.com/en/privacy_policy.html.
[20] SnapChat, SnapChat privacy policy, 2022-29-07. URL: https://values.snap.com/privacy/privacy-policy.
[21] Reddit, Reddit privacy policy, 2022-15-12. URL: https://www.reddit.com/policies/privacy-policy.
[22] Pinterest, Pinterest privacy policy, 2022-16-12. URL: https://policy.pinterest.com/en/privacy-policy.
[23] Instagram, Instagram privacy policy, 2023-01-01. URL: https://privacycenter.instagram.com/policy.
[24] R. Arp, B. Smith, A. D. Spear, Building ontologies with basic formal ontology, MIT Press,

2015.
[25] B. Smith, T. Malyuta, R. Rudnicki, W. Mandrick, D. Salmen, P. Morosoff, D. K. Duff, J. Schoening, K. Parent, IAO-Intel: an ontology of information artifacts in the intelligence domain (2013).
[26] A. Bandrowski, R. Brinkman, M. Brochhausen, M. H. Brush, B. Bug, M. C. Chibucos, K. Clancy, M. Courtot, D. Derom, M. Dumontier, et al., The ontology for biomedical investigations, PloS one 11 (2016) e0154556.
[27] A. Hicks, J. Hanna, D. Welch, M. Brochhausen, W. R. Hogan, The ontology of medically related social entities: recent developments, Journal of Biomedical Semantics 7 (2016) 1–4.
[28] S. Cox, C. Little, Time Ontology in OWL (W3C Candidate Recommendation Draft), 2022. URL: http://www.w3.org/TR/owl-time.
[29] B. Motik, P. Patel-Schneider, B. Parsia, OWL 2 Web Ontology Language. Structural Specification and Functional-Style Syntax (Second Edition), 2012. URL: http://www.w3.org/TR/owl2-syntax/.
[30] B. Glimm, I. Horrocks, B. Motik, G. Stoilos, Z. Wang, HermiT: an OWL 2 reasoner, Journal of Automated Reasoning 53 (2014) 245–269.
[31] M. Grüninger, M. S. Fox, The role of competency questions in enterprise engineering, in: IFIP WG5.7 Workshop on Benchmarking – Theory and Practice, Trondheim, Norway, Springer, 1994, pp. 22–31.
[32] Y. Jafta, L. Leenen, K. F. P. Chan, An ontology for the south african protection of personal information act, in: Proceedings of the 19th European Conf. on Cyber Warfare and Security, Virtual Conf. University of Chester, UK, 2020, pp. 158–166.
[33] B. Esteves, V. Rodríguez-Doncel, Analysis of ontologies and policy languages to represent information flows in GDPR, Semantic Web (2022) 1–35.
[34] A. Kurteva, T. R. Chhetri, H. J. Pandit, A. Fensel, Consent through the lens of semantics: State of the art survey and best practices, Semantic Web (2021 (Preprint)).
[35] S. Kirrane, J. D. Fernández, W. Dullaert, U. Milosevic, A. Polleres, P. A. Bonatti, R. Wenning, O. Drozd, P. Raschke, A scalable consent, transparency and compliance architecture, in: European Semantic Web Conference, Springer, 2018, pp. 131–136.
[36] R. Ianella, Open digital rights language (ODRL), Open Content Licensing: Cultivating the Creative Commons (2007).
[37] T. Athan, G. Governatori, M. Palmirani, A. Paschke, A. Wyner, LegalRuleML: Design principles and foundations, in: Reasoning Web International Summer School, Springer, 2015, pp. 151–188.
[38] H.-P. Lam, M. Hashmi, Enabling reasoning with LegalRuleML, Theory and Practice of Logic Programming 19 (2019) 1–26.
[39] Time Ontology in OWL, 2006. URL: http://www.w3.org/TR/owl-time.
[40] S. D. Gupta, A. Nygaard, S. Kaplan, V. Jain, S. Ghanavati, PHIN: a privacy protected heterogeneous IoT network, in: Intern. Conf. on Research Challenges in Information Science, Springer, 2021, pp. 124–141.
[41] M. Poveda-Villalón, A. Gómez-Pérez, M. C. Suárez-Figueroa, OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation, International Journal on Semantic Web and Information Systems (IJSWIS) 10 (2014) 7–34.
[42] OntoText, GraphDB 10.2, 2023. URL: https://www.ontotext.com/products/graphdb/.