Dataset Management Powered by Semantic Web Technologies

Björn Andersson[1], Patrik Kompuš[2], Sachiko Lim[1], and Michaela Skans[1]

[1] Dun & Bradstreet Sweden AB, Solna, Sweden
[2] Prague University of Economics and Business, Prague, Czech Republic

1 Introduction

When developing data supply chains, more and more companies strive for *compliance by design*. This process aims to ensure that computer software meets business compliance rules and policies. Some examples of compliance questions that require an answer are *On what legal grounds can we store and process the data?*, *Does the data need to be encrypted at rest?*, *How long can we keep the data?* and *Who shall be able to access the data?*.

To automate these decisions in the data supply chain, the software needs to retrieve required policies and rules to be able to take appropriate actions. Also, to be able to request for policies, the software needs to be aware of the data asset that it is currently processing; i.e., the data needs an asset identifier.

Dun & Bradstreet is a leading global provider of business decisioning data and analytics for almost 200 years. In our Nordic data supply chain, we use Semantic Web technologies and knowledge graphs to govern which datasets are allowed to be processed. By assigning each dataset a unique identifier at the earliest stage in the data supply chain, any subsequent software decision point can retrieve the associated policies and take appropriate decisions. The information in the knowledge graph is governed and maintained by data owners and can be updated (when required) without changing any of the software in the data supply chain.

2 Related work

Among previous work that utilized an approach based on Semantic Web technologies to ensure data compliance, Debruyne et al. proposed an ontology, an extension of the provenance ontology called PROVO, to represent collected informed consent and its changes over time [2]. Castro et al. introduced an autonomous data governance system building on semantic techniques and ontology-driven reasoning based on defined rules [1]. Palmirani et al. proposed a Privacy Ontology (PrOnto) representing the main legal norms of data protection under GDPR [3]. These proposed frameworks have not yet been fully implemented in production, however. Real-world validation with live datasets would enhance the practical applicability of Semantic Web technologies. Our paper introduces an approach that has been actively employed in large-scale production since 2017.

3 Dataset management framework

In our data supply chain, a dataset is defined as a collection of information elements that abide under the same judicial policies and share the same data controller/processor and data owner.

3.1 Dataset ontology model

The center of our dataset ontology model is the `Dataset` concept, which serves as the `rdfs:domain` for our properties. Several `owl:Class`es are defined to represent finite lists of possible values where each value is defined as an `owl:NamedIndividual`. We use qualified cardinality restrictions (QCRs) to define the shape of a `Dataset` instance. Fig. 1 shows an example of our model.

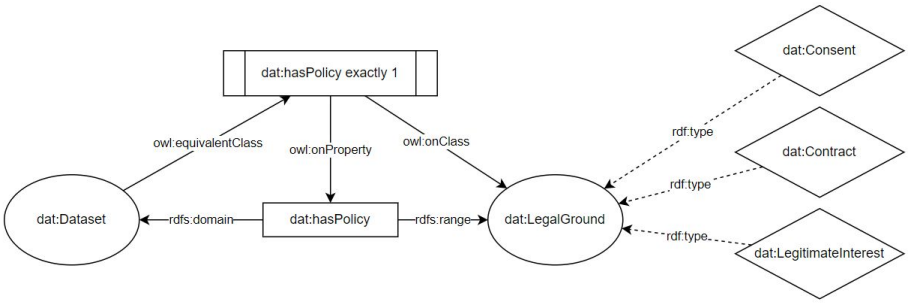

Fig. 1. An example of how some legal grounds are modelled in relation to a dataset

The properties in our dataset ontology model cover various metadata and data compliance policies divided into four categories; *General Information*, *Legal Framework*, *Information Security* and *Data Usage*. Fig. 2 shows some of the properties in each category.

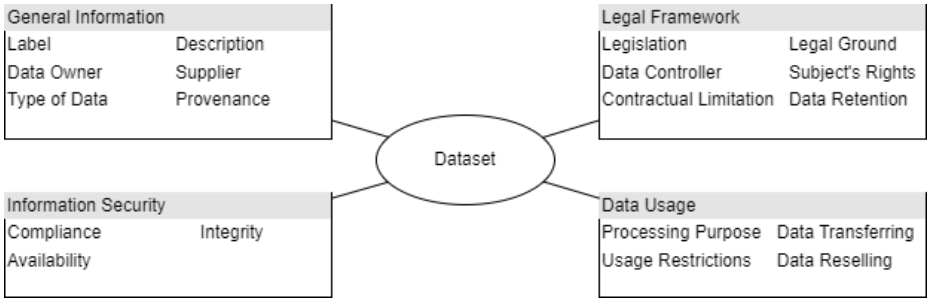

Fig. 2. Examples of properties within the dataset ontology model

3.2 System design and implemented use cases

The dataset management framework is implemented using Semantic Web technologies. The most important services, storage systems, clients and actors are depicted in Fig. 3 where the numbered use cases are further explained below.

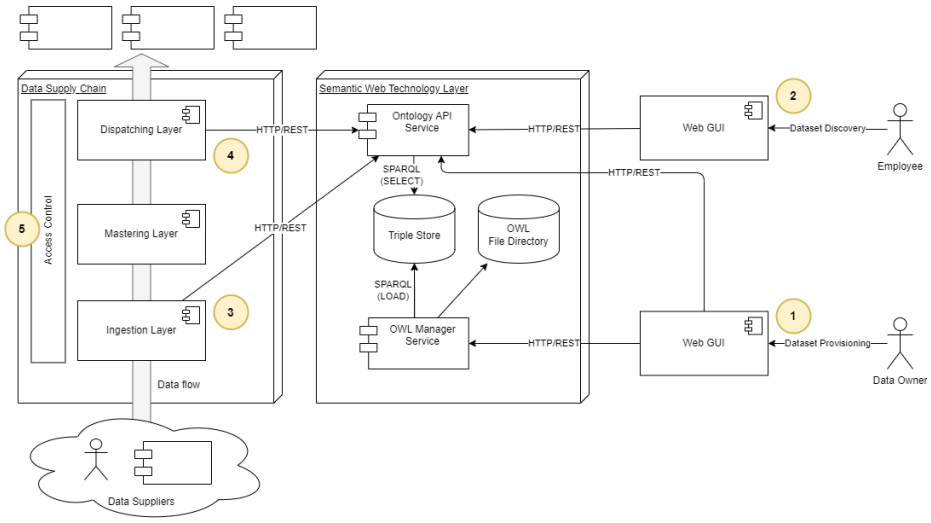

Fig. 3. High level system design of the dataset management framework

1. *Dataset Provisioning* Data owners use a `Web GUI` to create and update their datasets. The GUI renders a form based on the implemented QCRs. The form output is transferred to an `OWL Manager` service, which updates the appropriate OWL-file and finally loads the new version of the file into the `Triple Store`. Each instantiated dataset is assigned a globally unique IRI further used as the identifier across the data supply chain.
2. *Dataset Discovery* Employees may access a `Web GUI` where they can search for datasets and apply filters based on certain criteria selections, e.g., *"List all datasets in Finland that contain contact data"*. An `Ontology API` service executes a set of pre-defined SPARQL-queries towards the `Triple Store` and returns the content to the `Web GUI`. Also, the dataset IRI itself redirects any employee directly to the description of the dataset in the `Web GUI`.
3. *Dataset Ingestion* During the data ingestion process, every data content that shall be ingested is tagged with a dataset IRI. The `Ingestion Layer` verifies that the IRI exists, is active and that there is a provided legal ground for the dataset. If these checks are true, the data is ingested and further processed by the data supply chain; otherwise the data is rejected. The dataset IRI is now tagged to the ingested file and to all data messages originating from this file.

4. *Dataset Dispatching* Every data message is intercepted and inspected by the `Dispatching Layer`. The dataset IRI is retrieved from the message header and the `Ontology API` service is used to query the `Triple Store` for dispatching-related policies. Based on the returned policies, the `Dispatching Layer` distributes the data message to the appropriate and compliant destinations.
5. *Dataset Authorization* Access to operational tools and physical data within our data supply chain is authorized using attribute-based access control (ABAC) where the dataset IRI is one of the possible policy attributes.

4 Conclusion and future work

This dataset management framework has been in use since 2017, governing approximately 80 datasets. On a weekly basis over 25 million messages are processed in our Nordic data supply chain, all explicitly tagged as belonging to one of these datasets. Data owners, who previously had to manage their datasets using Excel or Word documents, have a modern user interface where they themselves can easily create new datasets or update existing ones if necessary, without needing any help from other resources such as developers. The user interface has over 100 active users.

By implementing this framework, we:

1. have a dataset registry that is centralized, unified, digitalized and accessible for humans and machines
2. ensure legal basis for data storage and processing
3. ensure information security by granting access to data per user/dataset to provide privacy by default
4. provide data compliance by design through automated data delivery decisions

To further standardize our solution, as future work we would like to explore using Shapes Constraint Language (SHACL) to steer the web forms in the user interface. Another future point of investigation could potentially be a dynamic data lineage solution.

References

1. Castro, A., Villagra, V.A., García, P., Rivera, D., Toledo, D.: An ontological-based model to data governance for big data. IEEE Access **9**, 109943–109959 (2021)
2. Debruyne, C., Pandit, H.J., Lewis, D., O'Sullivan, D.: "just-in-time" generation of datasets by considering structured representations of given consent for gdpr compliance. Knowledge and Information Systems **62**(9), 3615–3640 (2020)
3. Palmirani, M., Martoni, M., Rossi, A., Bartolini, C., Robaldo, L.: Pronto: Privacy ontology for legal compliance. In: Proc. 18th Eur. Conf. Digital Government (ECDG). pp. 142–151 (2018)