Data Privacy Issues in Big Biomedical Data

Maria-Esther Vidal[2] [0000−0003−1160−8727], Mayra Russo[0000−0001−7080−6331], and Philipp Rohde

TIB Leibniz Information Center for Science and Technology & L3S, Germany
{maria.vidal, mayra.botero, philipp.rohde}@tib.eu

Abstract. The amount of available Big data has grown drastically in the last decade, and a faster growth rate is expected in the coming years. Specifically, various biomedical domain methods, e.g., liquid biopsies, medical images, and genome sequencing, produce large volumes of data from where new biomarkers, or biological characteristics and medical signs, can uncover the incidence of a disease. Clinicians are faced with several challenges when analyzing biomedical data sources during diagnosis and treatment prescriptions. Biomedical data are presented in countless formats such as medical records, images, or genome data, that have to then be combined for optimal therapy decisions. Lastly, different regulation for enforcing data protection and privacy may hinder free access to biomedical data. This chapter addresses challenges present during the management of big biomedical data and presents a data-driven framework that resorts to ontologies to describe the main characteristics of data sources whose access is regulated by different data access regulations. The Privacy Ontology is defined as a formalism for representing the various entities that play a relevant role in the collection, anonymisation, integration, processing, and distribution of big biomedical data. As proof of concept, we illustrate the expressiveness of the proposed approach in the context of the European Union funded project iASiS, which aims at transforming big data into actionable knowledge to pave the way for personalised medicine and individualised treatments.

Keywords: Ontologies; Data-driven technologies; Biomedical data; Privacy-aware query processing

1 Introduction

Data is recognized as a fundamental asset for fostering industrial competitiveness, a country's economy, and assuring citizens with a high quality of life. Industrial digitalization, and the use of information technologies in public and health sectors, provide evidence of the pivotal role that data plays in our lives. Data provides the basis for supporting the understanding of a population's current characteristics; they represent building blocks for prediction and discovery. However, albeit recognized as critical infrastructures, data-driven systems have not been globally adopted yet. The interpretation and transparency of the outcomes of all the decisions made during data management are crucial to accounting for data quality, bias, and traceability. Therefore, relevant Big data developments at a large scale demand the ability to transform Big data into knowledge. More importantly, they must ensure that all these transformations are done following ethical and legal guidelines that guarantee the protection of sensitive personal data such as financial transactions, medical procedures, biometric identifiers, etc.

To address Big data challenges, effective and efficient data-centric applications must consider various privacy and access control regulations as well as enforce privacy constraints once data consumers access data. Existing works suggest the specification of access control ontologies data [4,16] and their enforcement on centralized or federated data repositories (e.g., [3,9]). Moreover, the European Union guidelines for Trustworthy AI [1] put forth detailed best practices to enhance trustworthiness. Nevertheless, trustable data-driven frameworks demand the fulfillment of requirements at both a legal and ethical level. Thus, special attention needs to be paid to deploy privacy-aware data-driven frameworks that ensure trustworthiness and reliability during the whole pipeline of data management. In this chapter, we present specific challenges required for the deployment of trustable data-driven systems to ethically foster industrial competitiveness and economies, and assure citizens of a high-quality of life.

Semantic data integration techniques towards trustable data-driven systems are illustrated in this chapter in the context of the European Union Horizon 2020 funded project iASiS. In iASiS, the proposed data-driven frameworks are being utilized to integrate big biomedical data, e.g., drugs, genes, mutations, side effects, with aggregated clinical records, medical images, and geneomic data. As a result, a knowledge graph is created. These data-driven frameworks are utilized to create holistic profiles of persons who suffer from two complex diseases: lung cancer and Alzheimer's. These profiles are part of knowledge graphs that integrate Electronic Health Records with scientific literature, and genomic and pharmacology data. These integrated bases of knowledge provide medical science with entirely new analytical methods that empower the understanding of individual conditions that may influence the risk of suffering a disease or its progression. More importantly, they facilitate the prescription of individualized treatments that may increase the chances of survival and quality of life. The benefits of the techniques discussed are not limited to medicine. They can be applied in various areas where ethical guidelines are required for quality assurance and the responsible and fair handling of data. They represent essential building blocks for enhancing trustworthiness of data-driven technologies.

The contributions of the work are summarized as follows:

- An analysis of big biomedical data characteristics and ethical requirements needed to be fulfilled in trustable data-driven frameworks.
- Characterization of data integration conflicts whose resolution require the satisfaction of legal and ethical requirements.
- A case study illustrating the application of the data-driven frameworks developed for the iASiS project[1],a European Union Horizon 2020 project funded to increase the understanding of two complex diseases: lung cancer and

[1] https://project-iasis.eu/

Alzheimer's and to pave the way for precision medicine with the use of data-driven technologies.

The remainder of the chapter is structured as follows: section 2 presents the basic terminology of big data and the most dominant dimensions that characterized big biomedical data. section 3 describes the problems of privacy in the context of Big data and the requirements to be satisfied whenever data protection and privacy regulations need to be fulfilled during biomedical data management. section 4 describes a privacy-aware framework to describe, exchange, and process big biomedical data. section 5 illustrates the application of the proposed privacy-aware framework in iASiS. Lastly, conclusions and future directions are outlined in section 6.

2 Big Biomedical Data

2.1 The Big Data Model

Big data is an artefact of individual and collective intelligence generated and collected using technological environments, where virtually every real-world entity can be captured digitally and stored in data sources [7]. The complexity of Big data is characterised by its dimensions, usually known as the V's of Big Data [14]. Dominant dimensions of Big data include:

- **Volume** refers to the ability to ingest and store very large datasets; they may consist of terabytes, petabytes, zettabytes of data, or even more. This colossal increase of large-scale data sets brings new challenges for the tasks of integrating, managing, and analysis.
- **Velocity** denotes the speed of data generation and delivery, being challenges the ingestion of high rate of data inflow with heterogeneous and evolving structures.
- **Variety** indicates multiple data formats and models. Enormous volumes of data is not consistent nor does it follow a specific template or format; it is captured in diverse forms and diverse sources e.g., genome and healthcare data coming from data-intensive experiments. These different forms clearly indicate that heterogeneity is a natural property of Big data and it is a big challenge to integrate, manage, and analyse such data sources.
- **Veracity** refers in part to the biases, ambiguities, and noise in data, as well as it is about understanding the data, as there are integral discrepancies in almost all the data collected. Thus, the necessity to deal with inaccurate and ambiguous data is another facet of Big data that needs to be tackled to ensure the appropriate management and mining of unreliable data.
- **Value** concerns data quality, including aspects like trustworthiness, authenticity, provenance, accountability, and availability of the data. The challenge is extracting knowledge/value from vast amounts of structured and unstructured data without loss of their meaning and properties.

2.2 Big Biomedical Data

According to the Big Data model, Biomedical data can be characterized as follows:

- **Volume**: biomedical data sources and particularly, genomics, make available large volumes of data. Public websites from scientific organizations like the European Genome-Phenome Archive (EGA) [2], EMBL-EBI [3], and the Centre for Genomic Regulation (CRG) [4] make available volumes of biomedical research data that include thousands of data sources. Furthermore and despite its already large volume and its demanding storage requirements, genomic data is growing at an unprecedented rate, with a projected exponential growth that will reach more than one Zettabytes per year by 2025 [15].
- **Variety**: biomedical data is ingested and harvested using a great variety of devices and protocols, e.g., liquid biopsies, medical images, or genome sequencing tests. Moreover, clinical notes encode relevant knowledge about the conditions and treatments of a patient; irregularities in patient's visits generates heterogeneity in the granularity of the entries of collections of clinical records. Additionally, treatments, interventions, and outcomes are diverse, and there are no standard schemas or protocols for reporting them in clinical notes.
- **Velocity**: clinical data is composed of data generated from different devices and the results of medical tests regularly done to the patients. Additionally, patients' vital signs can be registered in real-time, as the evolution of a tumor, and the reaction to a particular treatment.
- **Veracity**: patients' characteristics in a given instance of time may be affected in many cases for uncertainty generated by missing observations, errors in the interpretations of the conditions of a patient, and incorrect values due to inaccuracy of existing interventions and procedures.
- **Value**: the role of biomedical data in the improvement of healthcare has been shown in diverse scenarios and significant contributions have been achieved by conducting Big data-driven studies over clinical and genomic data with the aim of supporting precision medicine [12]. Just to mention some: Big data analytics over Electronic Health Records (EHRs) of nearly 3 million people and trillions of pieces of medical data has allowed for identifying associations between the use of proton-pump inhibitors and the likelihood of incurring a heart attack [13]. Also, using Big data-driven analysis, a study was conducted with over more than 7.700 brain images from 1.171 people in various stages of Alzheimer's progression; outcomes of the study revealed that intra-brain vascular dysregulation, i.e. a change in the brain blood flow, is an early pathological event during the development of the disease [17]. Lastly, Big data mining techniques implemented in DMET-Miner [2] are able to link

[2] https://www.ebi.ac.uk/ega/home

[3] https://www.ebi.ac.uk/

[4] https://www.crg.eu/

allelic variants in more than one probe with the conditions of patients, and suggest genes related to drug absorption, metabolism, and excretion.

3 Big Data Privacy

Big data usually consists of various data types from a wide variety of disparate sources. Heterogeneities vary from different methods to fetch data, variability of data formats, and inequalities of the technologies used in data generation. Additionally, data sets can be so large and complex that it becomes extremely challenging to capture and store them using traditional tools. More importantly, data sets can comprise data whose access is regulated by diverse policies. Research challenges include ingesting flows of massive volumes of data - open and protected- both at rest or in motion, presented in many formats and with potential quality issues. Furthermore, the meaning or interpretation of the data may change over time and become inconsistent and incomplete with periodic peaks. All these characteristics demand novel data-driven technologies able to scale up all these data complexity problems in the most effective and efficient way. In this section, problems of big data privacy are discussed as well as the requirements to be satisfied whenever big biomedical data is processed and analyzed.

3.1 Data Privacy and Controlled Access

As clearly stated by the EU General Data Protection Regulation (GDPR) law [5], the definition of privacy relies on the individual's right of preserving control access to personal data, e.g. health information, political and religious preferences, and genomics. Big data brings the opportunity to develop management and analytics methods that allow for unraveling unknown patterns as well as for the generation of actionable insights. Exploiting the benefits of Big data demands the implementation of techniques like record linkage and distributed data analysis; a major challenge in record linkage is the identification of duplicate entities and to integrate attributes present in the duplicate entities. In the biomedical domain, particularly, identifying entities that refer to the same real-world entity across one or more data sources is extremely important. Nevertheless, record linkage can risk confidentiality of personal data, and a major challenge is the identification of duplicate entities while protecting sensitive private information and preserving linkage quality. Therefore, preserving privacy during Big data management and analytics demands to ensure anonymity and confidentiality, i.e., hiding individuals' identities and avoiding information be shared with a second party without the identity being publicly revealed. In order to identify potential risks of privacy violations, confidential attributes should be identified in the data sources that will be integrated during Big data management and analytics. Privacy characteristics to be identified include: **Licensed**, **Anonymised**, **Authoritised**, and **Attributed**.

[5] `https://www.eugdpr.org/article-summaries.html`

Licensed requirements to access, manage, and distribute a data source following a particular license or agreement. In the case of open data, the European Commission clearly states the different type of licenses or regulations that can guide publications and access [6]. Depending on the nature of the data sources, the requirements of licensing differ. For instance, some sources are openly accessibly while others are subject to access or management constraints. An example of licensing is the *Creative Commons* licenses which enable data owners to maintain their copyright while granting others access or control over their data sets [7]. Creative commons licenses are defined in terms of three layers: **i) Legal Code:** provides the legal aspects of the license in a language understood by lawyers. **ii) Human readable:** presents the textual description of the license in a format that humans can read and understand. **iii) Machine readable:** makes available a fine-grained description of the principal characteristics of a license in a format manageable by computers. *Creative Commons* are exemplar licenses that regulate the exchange and use of creative work. We describe them in the context of data exchange, processing, and distribution. We refer the reader to the website of *Creative Commons* for an extensive description of all the licenses [7].

- **Attribution generic (CC BY):** This is the most generic and less restrictive license. It allows data consumers to distribute, remix, adapt, and build upon data sets, even for commercial purposes, as long as the credit to the data provider and owner is mentioned. CC BY is recommended for maximizing dissemination and reuse of a data set.
- **Attribution-ShareAlike (CC BY-SA):** This license allows data consumers to distribute, remix, adapt, and build upon data sets, even for commercial purposes. The constraint is that credit to the data provider and owner should be mentioned, and the outcome of the process of the original data set be licensed under identical terms. CC BY-SA is recommended for maximizing dissemination and reusing a data set that has already included parts of a CC BY-SA data set.
- **Attribution-NoDerivs (CC BY-ND):** This license allows data consumers to reuse the data set for any purpose, including commercially. Nevertheless, the changes to a data set cannot be distributed in any form, and the credit must also be attributed to the data owner.
- **Attribution-NonCommercial-NoDerivs (CC BY-NC-ND):** This license is the most restrictive of CC licenses. It only enables data consumers to access the data set and share it with others as long as the credit to the data provider and the owner is stated. Data consumers cannot change the data sets in any way or commercialised them.

Anonymised values of the attributes are processed and stored in a way that the real identity of the donor of the piece of data cannot be detected. Anonymisation processes must ensure that sensitive data can be distributed while privacy is preserved. Nonetheless, in some instances removing information that may serve

[6] https://www.europeandataportal.eu/sites/default/files/d2.1.2_training_module_2.5_data_and_metadata_licensing_en_edp.pdf

[7] https://creativecommons.org/licenses/

as an identifier of a data subject is not possible. In those cases, a process of pseudonymisation is a more feasible solution. A pseudonymised data set does not remove all the attributes that identify a piece of data or entity, but does reduce the connection of a pseudonymised entity with its original identity.
Authorised data usage consent is granted to data consumers through the signature of an access agreement or a license. These legal documents should establish permissions for performing operations like reading, aggregating, or analysing the data. The distribution of the results of data processing should also be specified in the data access authorization agreement.
Attributed similarly to the authorisation of the usage and processing, the attribution and recognition of the credit to the data providers and owners should be established legally. This document should be signed by the data providers, owners, and consumers, and state the copyright permissions and acknowledgements of sources and provenance of the data.

3.2 Privacy and Data Access Control Requirements and Operations

In addition to access regulation characteristics for data sets, operations performed on these should also meet a set of requirements. According to Zeng et al. [18], the following requirements need to be satisfied for ensuring access control policies during data management and exploration in a federation of users (parties) and data sources.

- **Per-party authorized view (R1)** the data view is different among the parties.
- **Authorization autonomy (R2)** each party should have full and autonomous control over the authorization definition.
- **Fine-granularity access control over records (R3)** access control can be applied to specific data fragments and not only the whole data source.

Requirements **R1** and **R2** conform to data sovereignty while **R3** states control at different levels. Additionally, several operations can be performed over data provided by one or more parties. The execution of these operations should be regulated by access control policies to enforce them. These operations can be executed at different levels, e.g. over attributes, records, or data sets, and as stated by **R3**, access policies can also regulate the execution of these operations at different granularities.

- **Read (R)**: a record or attribute is collected from a data source.
- **Write (W)**: a record or attribute is (re-)written in a data source.
- **Merge (M)**: an operation can be performed over a record or the value of an attribute to merge, integrate, or join with records or values from another data source.
- **Storage (S)**: a record or attribute collected from a data source is stored in a different data source.
- **Distribution (D)**: a different party distributes a record or attribute provided by a party.

The scientific community has defined several approaches to address the problem of describing access control policies, as well as the problem of enforcing them; exemplar approaches are described. Kirrane et al. [10,11] surveyed various access control models, policy representations, and standards for the Resource Description Framework (RDF). The described models include:

- **Mandatory Access Control (MAC)** MAC uses a central authority to control access to the resource.
- **Discretionary Access Control (DAC)** In contrast to the centralized approach of MAC, DAC allows resource owners to control access to their resources.
- **Role Based Access Control (RBAC)** RBAC restricts access to resources to groups of users, which allows for a more dynamic and inclusive permission allocation.
- **Attribute Based Access Control (ABAC)** ABAC is designed for distributed systems, where the subject may not be known to the system. In ABAC, access to resources is limited based on the attributes of the subject and/or the resource.
- **Context-Based Access Control (CBAC)** CBAC uses properties of the users, resources, and the environment to control access to resources.
  Kirrane et al. also present the following types of access policy representations:
- **Ontology-based approach** Access control policies in this approach are represented using the vocabulary and entity relations defined in the ontology. Using the deductive capabilities of ontologies, it is possible to infer new policies based on the existing ones.
- **Rule-based approach** This approach supports access control policies that can only be evaluated at runtime due to the dynamic nature of access constraints.
- **Hybrid approach** This approach combines both the ontology-based and rule-based approaches to benefit from their capabilities.

The privacy-aware techniques presented in this document describe data sources based on these operators. They provide the foundations for enforcing data privacy and data access regulations during data access or query processing. Furthermore, this approach implements hybrid privacy-aware techniques and follows the **DAC** and **RBAC** models where data providers restrict access to their resources to certain parties.

3.3 Requirements in Big Biomedical Data

The richness, variety, and potential data quality issues of biomedical data demand the satisfaction of further requirements with the aim of increasing trustworthiness and reliability. The main requirements to be fulfilled include:

- **Data transparency (RBDD$_1$):** Data set characteristics and decisions made during processing and analysis must be documented. These descriptions should be accessible and processed for humans and machines. They should be done at a fine-grained granularity to facilitate tracking down the whole data-driven pipeline and causality relationships between the input and output of each of the pipeline steps.

- **Integrity constraint validation ($\mathbf{RBDD_2}$):** The domain restriction descriptions should be specified in formal languages that enable human and machine understanding and processing. Verification should be tractable to allow for practical implementations of the approach.
- **Data certification ($\mathbf{RBDD_3}$):** Attribute values should be certifiable at every data-driven pipeline step. If needed, the results of external processes of certification, e.g., preferred validation sources, should be indicated.
- **Description data quality assessment ($\mathbf{RBDD_4}$):** The results of the processes of data quality assessments and curation should be documented in a human and machine understandable fashion.
- **Documentation and explanation of data bias ($\mathbf{RBDD_5}$):** In this chapter, we refer to bias as the *inclination or prejudice* in how a group of entities is represented or processed in a data-driven framework. The documentation of bias and diversity of the groups or cohorts present in a data set should be comprehensible to humans and machines.

Our proposed approach resorts to ontologies to describe data provenance and operations to be executed during query processing. It also enables the documentation of the entity or organization that authorised data access and processing. These ontology-based descriptions represent building blocks for the traceability of data access and query processing in a data-driven pipeline.

4 Enforcing Privacy and Data Access Control

In this section, we describe the methodology followed to analyse data sources and document their main properties. Furthermore, an ontology named Privacy Ontology (PO) is defined in terms of its main classes and properties. Lastly, we show how the descriptions of the data sources using the metadata encoded in PO empower the data-driven framework to ensure data access regulations during query processing.

4.1 Ontologies as a Formalism for Knowledge Representation

Knowledge engineering is a field that covers scientific, technical, and social paradigms required to model, create, maintain, and deploy knowledge-based infrastructures. Knowledge modeling facilitates the description and formal representation of the entities and their properties in a universe of discourse using a data model. Data models range from expressive formalisms like ontologies to less expressive models like the relational model. Knowledge representation provides the basis for the definition of the main properties of a real-world entity. They enable the specification of syntactical and semantic characteristics of an entity, facilitating, thus, successful communication. An *ontology* refers to *the specification of a conceptualization.* They allow for the representation of explicit and implicit properties of an entity; the specification of the main characteristics of these properties, i.e. in an ontology data and metadata can be represented.

Ontologies are specified using knowledge representation models, making the expressiveness of the ontology dependant on the expressive power of representation model. For example, the Resource Description Framework (RDF) enables the description of entities in terms of classes and properties; while subsumption relations between classes and properties can be modelled with the RDF Schema (RDFs). Main features of these two formalisms can be summarized as follows:

Resource Description Framework is a model developed by the W3C consortium to describe the metadata of resources. RDF follows the natural intuition of representing factual statements as expressions of the form subject-predicate-object, named RDF triples. Subjects correspond to entities uniquely identified in the web by the means of **Uniform Resource Identifier (URI)**[8]. A predicate defines a relation between subject and object; predicates are also identified with URIs. Lastly, objects can be represented by URIs or literals. they are in the form of URIs while objects can be of any form.

RDF Schema is an extension of the basic RDF that allows for the definition of classes, properties, hierarchies of classes and properties. On the other hand, more expressive formalisms like the Ontology Web Language (OWL) make available a larger number of operators which enable the representation not only of classes, properties, and subsumption relations, but also class and property constraints, negative statements, general equivalence relations, and restrictions of cardinality.

4.2 Ontologies and Data Privacy

Several ontology-based approaches have been proposed to model data privacy regulations. Kamateri et al. [8] present the Linked Medical Data Access Control (LiMDAC) framework with the aim of enabling access control over medical data aggregated by the multi-dimensional data cubes. LiMDAC exploits data cubes' metadata to restrict access to cubes and access policies can be defined over specific datasets and access spaces to which a number of users belong. Grando et al. [6] propose a hybrid approach where an ontology and a set of access control rules allow for reasoning about access permissions. As a proof of concept, Grando et al. apply this formalism to the biomedical domain and define rules that take the form of a consent statement signed by a patient and led by a researcher. Furthermore, each consent is associated with consent rules that regulate the processing to be executed against an entity representing a consent. Finally, Zeng et al. [18] devise a query evaluation scheme that supports access control in a federated database system where different collaborative parties are sharing and exchanging data described using the relational model. The framework that we describe in this document, also implements an ontology-based approach to describe data access policies and a set of rules for reasoning about the privacy and access control policies to apply when these sources are accessed. However, in contrast to the above ontology-based approaches, this formalism provides the

[8] For simplicity, we assume that the RDF graphs are free of blank nodes.

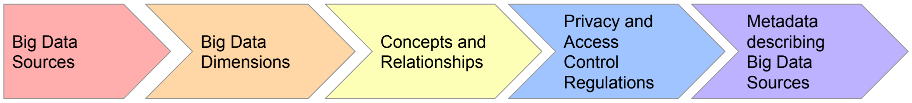
Big Data Sources
Big Data Dimensions
Concepts and Relationships
Privacy and Access Control Regulations
Metadata describing Big Data Sources

Fig. 1: **Data Source Description Methodology**. Steps of the methodology to create the metadata that describes data sources in terms of big data and privacy issues characteristics. Structured questionnaires are utilized to capture the description of the sources from their providers or owners.

basis to be used as part of a query engine in order to ensure that data privacy regulations are respected during query processing. In this way, each of the operations required to execute a query, e.g., Read (R) or Merge (M), respect the access policies imposed by their data providers. The definition of the privacy-aware formalism and the evaluation of the impact and benefits of enforcing privacy-aware query processing over knowledge graphs was originally reported in Endris et al. [5]. subsection 4.4 presents the ontology that models the main concepts needed for data protection and regulated access, and subsection 4.5 outlines the main characteristics of privacy-aware framework that implements these privacy aware techniques.

4.3 Methodology to describe biomedical data sources in terms of data privacy and data access control regulations

This section presents the main steps of a methodology proposed to capture the main characteristics of the data sources of a big data project. Figure 1 depicts the main steps of this methodology. First, data sources are described by the data owners or providers according to the main more dominant dimensions of the big data model. A questionnaire is used to collect these descriptions; it is composed of five parts. In total, the questionnaire comprises 30 questions; Table 1 summarises the main parts of the questionnaire.

- **Overview:** general description of the data set is provided.
- **Big data Vs:** the data set is described in terms of the dimensions big data models volume, velocity, variety, veracity, and value.
- **Data provider:** captures the protocols followed to access the data, who is the data owner and administrator, and permission status.
- **Data set detailed features:** outlines the main characteristics of the data in the source. These features include: **a)** data formats; **b)** language; **c)** assumptions and standards followed during data collection and harvesting; **d)** ontologies and vocabularies used to described the data; **e)** accessibility, permissions, and anonymization, and **f)** data collection frequency.
- **Use cases:** presents the use cases where the described data set can be utilized and the coverage of the data set.

Once the data sources are described in terms of the main big data characteristics, dominant privacy and data access properties are identified. These

Table 1: **Big Data Questionnaire**. The dominant characteristics of the data sources are captured in a questionnaire composed five sessions.

Questionnaire Section	Question Description
Overview	Data source title Data source acronym Data set general description Temporal coverage Status/ Maintenance
Big Data Vs	Volume- Data size Velocity- Frequency of the observations (Longitudinal data) Variety- Various formats Veracity - Type of quality problems Value - Key Performance Indicators
Provider	Data provider Data provider URI Protocol Used to Access Data Experimental Strategy Data Owner Data administrator Permission status
Data Detailed Description	Data format Data language Data collection assumptions Standards Ontologies and Vocabularies Accessibility, Permissions, Anonymisation Data collection frequency Data schema documentation Raw data sample
Use Case	Application scenario Possible scenario coverage

descriptions are collected with the help of the data owners and providers using another questionnaire. The questionnaire for data privacy and access regulation description is composed of two parts: the first part describes the privacy and data access regulations of the data source, while the second part indicates the field-specific permissions and regulations of a data source. The questionnaire is composed of 26 questions.

- **Controlled Access:** a data access agreement must be approved and signed to retrieve the data set.
- **License based Access:** the terms and use of the data set are regulated by a specific license.
- **Open Access:** no restrictions are imposed on access to, or use of the data set.
- **Regulated Operators:** operations restricted according to the policies established by the data providers of the data sources with license-based access.

Operations like Read (R), Merge (M), and Distribution (D) cannot be usually executed over controlled data sources.

Table 2: **Data Privacy Questionnaire**. Data privacy and data access characteristics are captured in a questionnaire of two levels of representation granularity: data source- and attribute-based description.

Data Source Level	
Questionnaire Section	**Question Description**
Overview	Data source title Data type (raw or analysis-derived) License Type (e.g., CC BY, CC BY-NC) License URL on data set website
Detailed Access Information	Data must be anonymised Partners can read data Partners can analyse data End-users can read the data Partners can read analysis results produced from the data Partners can export analysis results produced from the data End-users can read analysis results produced from the data End-users can export analysis results produced from the data
Attribution	Requirements
Field Level	
Questionnaire Section	**Question Description**
Overview	Data set name Field name
Detailed Access Information (per field)	Data of this field must be anonymised Partners can read this field Partners can analyse this field End-users can read this field Partners can read analysis results produced from this field Partners can export analysis results produced from this field End-users can read analysis results produced from this field End-users can export analysis results produced from this field
Attribution	Requirements

Once the description of the data sources are captured by using the questionnaires depicted in Table 1 and Table 2, metadata describing the results of this analysis is expressed using the Privacy Ontology (PO). Main concepts of PO are described in the next section.

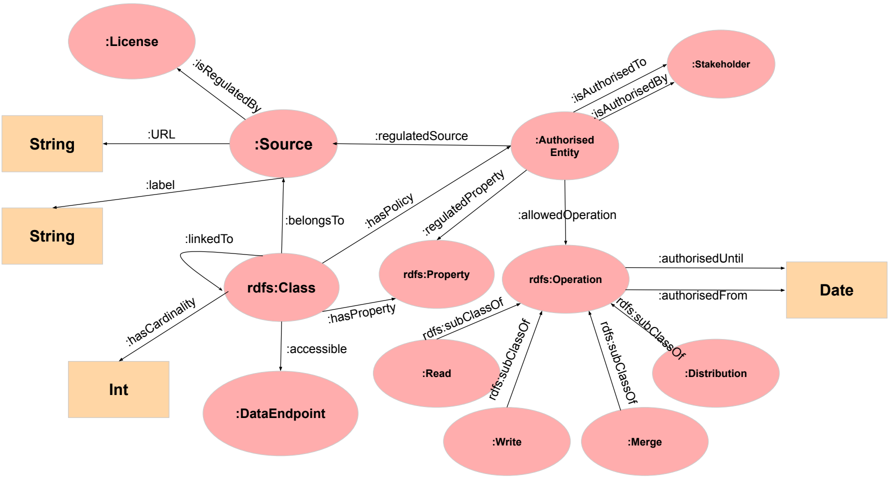

Fig. 2: **A Portion of the Privacy Ontology**.

4.4 An Ontology-based Approach for Describing Privacy and Access Control

We propose the Privacy Ontology (PO) to describe data sources in terms of privacy regulations and data provenance. Moreover, the proposed ontology describes the operations Read (R), Merge (M), Storage (S), and Distribution (D) that can be executed over the entities of a particular knowledge graph. Figure 2 depicts the main classes and properties of the Privacy Ontology (PO).
Source This class of the Privacy Ontology represents the data sources that compose a data-driven system. Data sources are described in terms of the following properties:

- **:isRegulatedBy** indicates the type of license that regulates the access of the data source.
- **:URL** represents the URL where the data set can be accessed or where is described.
- **:label** denotes the title of the data source.

RDF Classes A class denotes a set of entities of the same type; it is characterised by the following properties:

- **:hasProperty** represents the properties of the class, e.g., **:hasBiosy** and **:hasTumorStage**. The type of the property p is **rdfs:Property**.
- **:hasPolicy** allows for representing the operations that can be performed over a property whose data is provided by a given data source; the object of the property **:hasPolicy** is related to an entity which is an instance of the class **:AuthorisedEntity**.
- **:hasCardinality** represents the cardinality of the entities in a class **C**.

- **:accessible** corresponds to the data endpoint or web service from where the entities of a class **C** can be accessed.
- **:linkedTo** represents the links between a class **C** and other classes.

Authorised Entity An entity of the class **:AuthorisedEntity** represents the relation between a property whose access is authorised by a data provider **s** to an access consumer **c** during a given time period. The following properties are part of the class **:AuthorisedEntity**:

- **:regulatedProperty** represents a property **p** whose access is regulated. The type of the property **p** is **rdfs:Property**.
- **:allowedOperation** represents an operation **o** that can be performed over the property **p**. The operation is of type **:Operation**.
- **:isAuthorisedBy** represents a stakeholder which provides the data for populating a property **p** and authorised an operation **o**. The stakeholder is type **:Stakeholder**.
- **:isAuthorisedTo** represents a stakeholder which consumes the data of property **p** and is authorised to execute an operation **o**. The stakeholder is type **:Stakeholder**.
- **:authorisedFrom** represents the starting date when the access is granted.
- **:authorisedUntil** represents the ending date when the access.

Operation A resource of the class **:Operation** represents an operation that can be executed over a property **p**. Operations can be **:Read**, **:Write**, **:Merge**, and **:Distribution**.

4.5 Preserving Privacy and Access Control Regulations in a Data-driven Pipeline

This section presents a data-driven framework that exploits the descriptions of the data sources captured by following the methodology presented in Figure 1, and modeled using the Privacy Ontology in Figure 2. The novelty of this component is: **1)** Different granularities of data privacy and access control regulations, e.g., at the level of entities or attributes; **2)** Formal representation of the data privacy and access control regulations as privacy rules using a formal language, and **3)** The data access plans that collect data by respecting data access policies. Privacy rules are modeled using the Privacy Ontology and represent data access and management required for executing a given query. They include: Read (R), Merge (M), Storage (S), and Distribution (D), enabling, thus, the description of privacy characteristics, e.g., authority for processing the data. Thus, the privacy-aware techniques meet the requirements of Zeng et al.[18] and respect the DAC and RBAC models, proposed by Kirrane et al. [10,11], where data providers restrict data access to certain parties.

Figure 3 depicts the main components of the privacy-aware component for query processing. This component receives a request expressed as a query using a formal query language, e.g., SPARQL [9], and identifies the data sources that need

[9] SPARQL is a formal language to express queries over RDF data sources `https://www.w3.org/TR/rdf-sparql-query/`

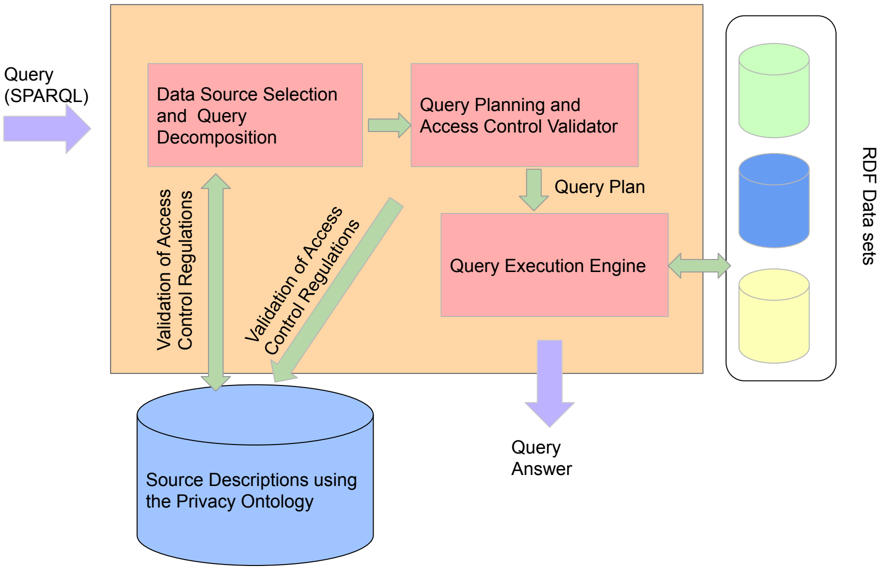
Query (SPARQL)
Data Source Selection and Query Decomposition
Query Planning and Access Control Validator
Query Plan
Query Execution Engine
Validation of Access Control Regulations
Validation of Access Control Regulations
RDF Data sets
Source Descriptions using the Privacy Ontology
Query Answer

Fig. 3: **A privacy-aware query engine**. The main components of a query engine able to exploit data access control regulations described by the Privacy Ontology. A query in a formal language, e.g., SPARQL, is received as input and the output is the answer of the query produced by a query plan that respects the data access regulations imposed by the data providers. The data sources required to answer the query are selected and a query plan for these sources is generated in a way that data access regulations are respected. A query engine executes the query plan and produces the query answers.

to be accessed to answer the query. The component **Data Source and Query Decomposition** exploits the descriptions of the data sources with the aim of identifying the relevant data sources and the parts of the original query that will be executed in each of these selected data sources. As a result, a decomposition of the original query into subqueries that will be executed in the selected data sources is generated. The component **Query Planning and Access Control Validator** receives as input a decomposition of the original query and making use of the information about the operators that can be executed by each of the partners, decides a plan on the data sources. Lastly, the query execution engine evaluates the privacy-aware query on the selected RDF data sources and produces the query answers. In case no data source that respects the data access regulations can be selected, the output of the query is empty.

5 Applying a Data-driven Pipeline in the Context of Biomedical data

iASiS is a 36-month project supported by the European Commission under the program Horizon 2020 Research and Innovation Action. iASiS aims to transform clinical and big pharmacogenomics data into actionable knowledge to support personalized medicine in two life-threatening diseases: lung cancer and Alzheimer's. iASiS aims at integrating heterogeneous Big data sources into an integrated knowledge base or the iASiS knowledge graph. Data sources include clinical notes, medical images, genomics, medications, and scientific publications. To create the iASiS knowledge graph, iASiS offers a unified representation schema to represent knowledge encoded into the heterogeneous big data sources. Furthermore, to overcome heterogeneity conflicts across the heterogeneous sources, the iASiS infrastructure uses diverse data analytics methods. For example, Natural Language Processing and text-mining techniques are used to convert clinical notes into structured data. At the same time, state-of-the-art machine learning methods are utilized for image analysis. Moreover, the iASiS infrastructure relies on ontologies to semantically describe real-world entities, e.g., drugs, treatments, publications, genes, and mutations; these annotations provide the basis for semantic integration. The iASiS data sources are diverse in format and content, as well as in the data privacy policies and licenses that regulate access, integration, analysis, and distribution.

The methodology depicted in Figure 1 was followed to analyse more than 50 different biomedical data sources. For each of the data sources the questionnaires described in Table 1, and Table 2 were filled in. Data owners and providers– who were partners of the project– were responsible for describing the iASiS data sources based on the features captured in each of the questions. A repository with the descriptions of all the data sources was created. It is implemented as an instance of the Leibniz Data Manager [10]; big data characteristics, access policies, and main vocabularies utilized to describe the data, are included as part of the descriptions of the data sources. The repository also allows the user to verify the data sources' content without downloading them beforehand, and enables the visualization of data sources in different formats.

The iASiS data sources were harvested following different knowledge extraction methods. For example, clinical data was collected from the hospitals who were partners of the project. These clinical data sets include clinical notes, medical images, and hospital-derived genomic data; all these data sets were aggregated and annotated with biomedical vocabularies. The hospitals further performed anonymisation techniques to remove confidential attributes, e.g., national identifiers, name, addresses, and birth dates. The clinical partners respected national and European regulations for accessing and managing clinical data (e.g., the Spanish Law of Personal Data Access[11]). Internal agreements for accessing

[10] https://projects.tib.eu/datamanager/

[11] Laws 15/1999 and 41/2002 https://www.boe.es/buscar/act.php?id=BOE-A-2002-22188

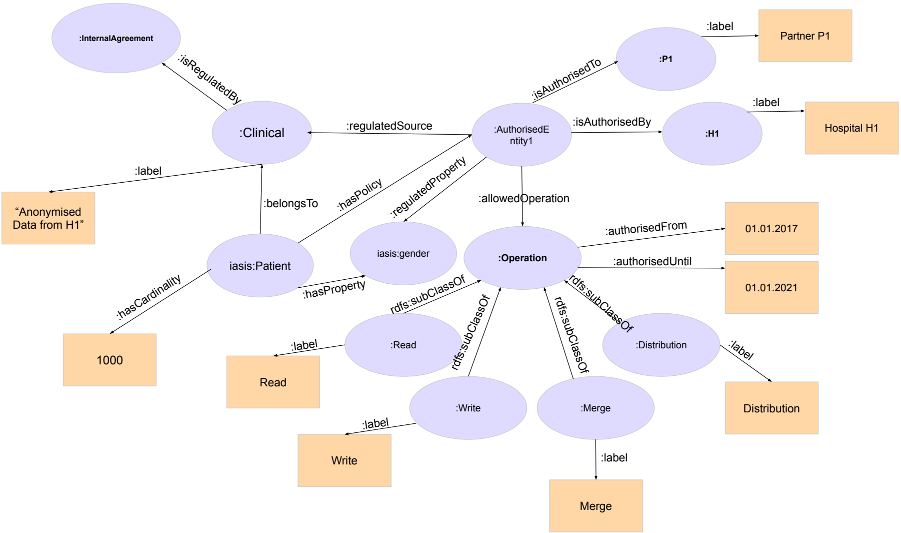

Fig. 4: **A portion of the Privacy Ontology**. The description of the access regulations of the property **iasis:age** of the class **iasis:Patient**.

anonymised clinical data were signed by the partners of the consortium responsible for processing, curating, and integrating the anonymised clinical data. Knowledge about anonymised electronic health records was extracted using EHR Text Analysis. The extracted knowledge was annotated with terms from the UMLS control vocabulary and made available via a REST API. The API access required the authentication of a *username* and *password* that were encoded in Base64 following the TLS protocol, i.e., the communication established to download the data was encrypted respecting a security protocol. Data downloaded using the API was stored in a server that was not accessible via the Internet, and only the partners who have accounts in the Intranet could access this data. The description of the semantified version of the downloaded clinical data stated that the hospitals only granted the permission to the partner who performed knowledge extraction to Read (R), Merge (M), Storage (S), and Distribution (D). On the other hand, the partner who executed the curation and integration of the clinical data was only granted with the operators Read (R), Merge (M), and Storage (S). The rest of the technical partners could only access aggregated properties and could Read (R) and Merge (M) this type of data. The Privacy Ontology (PO) was utilised to describe all these data access regulations.

5.1 Example

In this section, the expressive power of the Privacy Ontology (PO) is illustrated with an example. Consider the class **iasis:Patient** with properties **iasis:gender**, **iasis:age**, **iasis:hasTumorStage**, and **iasis:ageRange**; it is populated with anonymised clinical data collected by a hospital $\mathbf{H}_1$. To illustrate how the Privacy Ontology (PO) can be used, let us assume that $\mathbf{H}_1$ allows a partner $\mathbf{P}_1$ to perform all the operators over the property **iasis:gender**, i.e., the operations Read (R), Merge (M), Storage (S), and Distribution (D) can be executed by $\mathbf{P}_1$. Figure 4 depicts a fragment of the Privacy Ontology (PO) that represents these data access regulations. As can be observed, PO allows for describing the class **iasis:Patient**, its properties and the authorization regulations that enables $\mathbf{P}_1$ to perform all the operations over the property **iasis:age**. More regulations can be imposed; for example, only aggregated properties of **iasis:Patient** can be read and merged by all the partners of the consortium, e.g., **iasis:ageRange**. The partner $\mathbf{P}_1$ can Read (R), Merge (M), Storage (S), and Distribution (D) over all the properties of the class, while another partner $\mathbf{H}_2$ can only Read (R), Merge (M) and Storage (S). Similarly, other regulations that control the access of open biomedical data based on diverse licenses, can be expressed using the Privacy Ontology. These rich source descriptions provide the basis for ensuring data access regulations during data access, integration, and query processing.

6 Conclusions and Future Work

This chapter presents a data-driven approach for documenting and managing data privacy and access control regulations. A methodology for analysing and describing the main characteristics of the data sources is presented. The Privacy Ontology is proposed as the formalism to document the data privacy regulations of big data sources. As a proof of concept, the proposed data-driven approach is illustrated with a simple example in the context of the EU H2020 project iASiS. The results of the analysis of the proposed data-driven methods suggest that accounting data privacy with ontologies provides expressive building blocks to support privacy-aware data management and query processing.

In the future, we plan to extend the Privacy Ontology to describe the process of data validation and curation. Furthermore, the integration of the Privacy Ontology with an ontology that documents data quality issues, integration conflicts, and bias in the data is part of our future agenda.

Acknowledgments This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions (grant agreement number 860630) for the project "NoBIAS - Artificial Intelligence without Bias", and under grant agreement number 727658 for the project "IASIS - Big Data for Precision Medicine"

Disclaimer This work reflects only the authors' views and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains.

References

1. Ethics guidelines for trustworthy ai. `https://ec.europa.eu/futurium/en/ai`, 2018. alliance- consultation.

2. G. Agapito, P. H. Guzzi, and M. Cannataro. Dmet-miner: Efficient discovery of association rules from pharmacogenomic data. *J. Biomed. Informatics*, 56:273–283, 2015.
3. M. Amini and R. Jalili. Multi-level authorisation model and framework for distributed semantic-aware environments. *IET Information Security*, 4(4):301–321, 2010.
4. L. Costabello, S. Villata, and F. Gandon. Context-aware access control for RDF graph stores. In L. D. Raedt, C. Bessière, D. Dubois, P. Doherty, P. Frasconi, F. Heintz, and P. J. F. Lucas, editors, *ECAI 2012 - 20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS-2012) System Demonstrations Track, Montpellier, France, August 27-31 , 2012*, volume 242 of *Frontiers in Artificial Intelligence and Applications*, pages 282–287. IOS Press, 2012.
5. K. M. Endris, Z. Almhithawi, I. Lytra, M. Vidal, and S. Auer. BOUNCER: privacy-aware query processing over federations of RDF datasets. In *Database and Expert Systems Applications - 29th International Conference, DEXA 2018, Regensburg, Germany, September 3-6, 2018, Proceedings, Part I*, pages 69–84, 2018.
6. M. A. Grando and R. Schwab. Building and evaluating an ontology-based tool for reasoning about consent permission. In *AMIA 2013, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 16-20, 2013*, 2013.
7. H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi. Big data and its technical challenges. *Commun. ACM*, 57(7):86–94, 2014.
8. E. Kamateri, E. Kalampokis, E. Tambouris, and K. A. Tarabanis. The linked medical data access control framework. *Journal of Biomedical Informatics*, 50:213–225, 2014.
9. Y. Khan, M. Saleem, M. Mehdi, A. Hogan, Q. Mehmood, D. Rebholz-Schuhmann, and R. Sahay. SAFE: SPARQL federation over RDF data cubes with access control. *J. Biomedical Semantics*, 8(1):5:1–5:22, 2017.
10. S. Kirrane, A. Abdelrahman, A. Mileo, and S. Decker. Secure manipulation of linked data. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, pages 248–263, 2013.
11. S. Kirrane, S. Villata, and M. d'Aquin. Privacy, security and policies: A review of problems and solutions with semantic web technologies. *Semantic Web*, 9(2):153–161, 2018.
12. T. J. Schmidlen, L. Wawak, R. Kasper, J. F. García-España, M. F. Christman, and E. S. Gordon. Personalized genomic results: Analysis of informational needs. *Journal of Genetic Counseling*, 23(4), 2014.
13. N. H. Shah, P. LePendu, A. Bauer-Mehren, Y. T. Ghebremariam, S. V. Iyer, J. Marcus, J. P. C. Kevin T. Nead, and N. J. Leeper. Proton pump inhibitor usage and the risk of myocardial infarction in the general population. *Plos One*, 10(7), 2015.
14. U. M. M. K. Sivarajah, Z. Irani, and V. Weerakkody. Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, 70:263–286, 2017.
15. Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, M. C. S. Ravishankar Iyer, S. Sinha, and G. E. Robinson. Big data: Astronomical or genomical? *Plos One*, 13(7), 2015.
16. J. Unbehauen, M. Frommhold, and M. Martin. Enforcing scalable authorization on SPARQL queries. In *Joint Proceedings of the Posters and Demos Track of*

the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016., 2016.

17. I.-M. Y, S. RC, and e. a. Toussaint PJ. Early role of vascular dysregulation on late-onset alzheimer's disease based on multifactorial data-driven analysis. *Nature Communications*, 7(11934), 2016.
18. Q. Zeng, M. Zhao, P. Liu, P. Yadav, S. B. Calo, and J. Lobo. Enforcement of autonomous authorizations in collaborative distributed query evaluation. *IEEE Trans. Knowl. Data Eng.*, 27(4):979–992, 2015.