Towards a semantic repository of data mining and machine learning datasets

Ana Kostovska
Jožef Stefan IPS &
Jožef Stefan Institute
Ljubljana, Slovenia
ana.kostovska@ijs.si

Sašo Džeroski
Jožef Stefan Institute &
Jožef Stefan IPS
Ljubljana, Slovenija
saso.dzeroski@ijs.si

Panče Panov
Jožef Stefan Institute & Jožef
Stefan IPS
Ljubljana, Slovenia
pance.panov@ijs.si

ABSTRACT

With the exponential growth of data in all areas of our lives, there is an increasing need of developing new approaches for effective data management. Namely, in the field of Data Mining (DM) and Knowledge Discovery in Databases (KDD), scientists often invest a lot of time and resources for collecting data that has already been acquired. In that context, by publishing open and FAIR (Findable, Accessible, Interoperable, Reusable) data, researchers could reuse data that was previously collected, preprocessed and stored. Motivated by this, we conducted extensive review on current approaches, data repositories and semantic technologies used for annotation, storage and querying of datasets for the domain of machine learning (ML) and data mining. Finally, we identify the limitations of the existing repositories of datasets and propose a design of a semantic data repository that adheres to FAIR principles for data management and stewardship.

1. INTRODUCTION

One of the main use of data is in the process of knowledge discovery, where scientist employ ML and DM methods and try to solve various real-life problems from diverse fields, from systems biology and medicine, to ecology and environmental sciences. In order to obtain their objectives, they need high-quality data. The quality of the data is crucial to a DM project's success. Ultimately, no level of algorithmic sophistication can make up for low-quality data. On the other hand, progress in science is best achieved by reproducing, reusing and improving someone else's work. Unfortunately, datasets are not easily obtained, and even if they are, they come with limited reusability and interoprability.

A key-aspect in advancing research is making data open and **FAIR**. FAIR are four principles that have been recently introduced to support and promote good data management and stewardship [17]. Data must be easily findable (**Findability**) by both humans and machines. This means data should be semantically annotated with rich metadata and all the resources must be uniquely identified. The metadata should always be accessible (**Accessibility**) by standardized communication protocols such as HTTP(S) or FTP, even when the data itself is not. Data and metadata from different data sources can be automatically combined (**Interoperabiliy**). To do so, the benefits of formal vocabularies and ontologies should be exploited. Data and metadata is released with provenance details and data usage licence, so that humans and machines know whether data can be replicated and reused or not (**Reusabiliy**).

The benefits of publishing FAIR data are manifold. It speeds up the process of knowledge discovery and reduces the consumption of resources. When the FAIR-compliant data at hand does not contain all the information needed it can be easily integrated with data from external sources and boost the overall KDD performance [12].

Semantic data annotation, being very powerful technique, is massively used in some domains, i.e. medicine, but it is sill in the early phases in the domain of data mining and machine learning. To the best of our knowledge, there are no semantic data repositories that adhere to the FAIR principles. We recognize the ultimate benefits of having one and we are going in depths of the research covering semantic data annotation, ontology usage, storing and querying of data.

2. BACKGROUND AND RELATED WORK

The Semantic Web (Web 3.0) is an extension of the World Wide Web in which information is given semantic meaning, enabling machines to process that information. The aim of the Semantic Web initiative is to enhance web resources with highly structured metadata, known as semantic annotations. When one resource is semantically annotated, it becomes a source of information that is easy to interpret, combine and reuse by the computers [13]. In order to achieve this, the Semantic Web uses the concept of Linked Data. Linked data is build upon standard web technologies [7] including HTTP, RDF, RDFS, URIs, Ontologies, etc.

For uniquely identifying resources across the whole Linked Data, each resource is given a **Unified Resource Identifier (URI).** The resources are then enriched with terms from controlled vocabularies, taxonomies, thesauruses, and ontologies. The standard metadata model used for logical organization of data is called **Resource Description Framework (RDF).** Its basic unit of information is the triplet compiled from a subject, a predicate, and an object. These three components define the concepts and relations, the building blocks of an ontology.

In the context of computer science, **ontology** is "an explicit formal specifications of the concepts and relations among them that can exist in a given domain" [3]. As computational artifacts, they provide the basis for sharing meaning both at machine and human level. When creating an ontology, there are multiple languages to choose from. **RDF Shema (RDFS)** is ontology language with small expressive power. It provides mechanisms for creating simple taxonomies of

concepts and relations. Another commonly used ontology language is the **Web Ontology Language (OWL).** OWL supports creation of all ontology components: concepts, instances, properties (or relations). Finally, **SPARQL**[1] is standard, semantic query language used for querying fast-growing private or public collections of structured data on the Web or data stored in RDF format.

There are diiferent technologies for storing data and metadata. The most broadly used are **relational databases**, digital databases based on the relational model of data organized in tables, forming entity-relational model. Another approach that became popular with the appearance of Big Data are **NoSQL** databases [5], which are flexible databases that do not use relational model. **Triplestores** are specific type of NoSQL databases, that store triples instead of relational data. Triplestores use URIs and can be queried over trillions of records, which makes them very applicable.

Data in an information system can reside in different heterogeneous data sources, both internal and external to the organization. In this setting, the relevant data from the diverse sources should be integrated. Accessing disparate data sources has been a difficult challenge for data analysts to achieve in modern information systems, and an active research area. **OBDA** [1, 11] is much longed-for method that addresses this problem. It is a new paradigm, based on a three-level architecture constituted of the ontology, the data sources, and the mappings between the two (see **Figure 1**). With this approach, OBDA provides data structure description, as well as semantic description of the concepts in the domain of interest and roles between them.

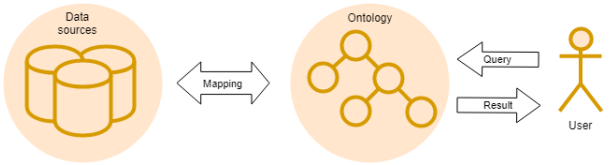

Figure 1. The OBDA architecture

In the context of semantic ML data repository, we group ontologies in three categories, i.e., ontologies for describing machine learning and data mining, ontologies for provenance information, and domain ontologies. **OntoDM** ontology describes the domain of data mining. It is composed of three sub-ontologies: OntoDT [10] - generic ontology for representation of knowledge about datatypes; OntoDM-core [8] - ontology of core data mining entities (e.g., data, DM task, generalizations, DM algorithms, implementations of algorithms, DM software); OntoDM-KDD [9] - ontology for representing the knowledge discovery process following CRISP-DM process model. **The Data Mining OPtimization Ontology (DMOP)** [6] has been designed to support automation at various choice points of the DM process, i.e., choosing algorithms, models, parameters. **The PROV Ontology (PROV-O)**[2] **and Dublin Core vocabulary** [16] facilitate the discovery of electronic resources by providing a base for describing provenance information about resources.

[1] https://www.w3.org/TR/rdf-sparql-query/
[2] https://www.w3.org/TR/prov-o/

There are numerous repositories of ML datasets available online. The UCI repository[3] is the most popular repository of ML datasets. Each dataset is annotated with several descriptors such as dataset characteristics, attribute characteristics, associated task, number of instances, number of attributes, missing values, area, etc. Similarly, Kaggle Datasets[4], Knowledge Extraction based on Evolutionary Learning (KEEL), and Penn Machine Learning Benchmarks (PMLB)[5] are well-known dataset repository that provide users with data querying based on the descriptors attached to the datasets. OpenML[6] is an open source platform designed with the purpose of easing the collaboration of researchers within the machine learning community [14]. Researchers can share datasets, workflows and experiments in such a way that they can be found and reused by others. When the data format of the datasets is supported by the platform, the datasets are annotated with measurable characteristics [15]. These annotations are saved as textual descriptors and are used for searching through the repository.

In contrast to the above mentioned repositories, there are frameworks in other domains that offer advanced techniques for describing, storing and querying datasets. One cutting-edge framework in the domain of neuroscience is **Neuroscience Information Framework(NIF)** [4]. Its core objective is to create a semantic search engine that benefits from semantic indexes when querying distributed resources by keywords. **The Gene Ontology Annotation (GOA)**, is a database that provides high-quality annotations of genome data [2]. The annotations are based on GO, a vocabulary that defines concepts related to gene functions and relation among them. Large part of the annotations are generated electronically by converting existing knowledge from the data to GO terms. Electronic annotations are associated with high-level ontology terms. The process of generating more specific annotations can hardly be automated with the current technologies, therefore it is done manually.

3. CRITICAL ASSESSMENT

In this section, we conduct critical assessment of the current research based on the review presented in the previous section.

Semantic Web technologies. The whole stack of semantic technologies provide ways of making the content readable by machines. The metadata that describes the content can be used not only to disregard useless information, but also for merging results to provide a more constructed answer. A major drawback of this process of giving data a semantic meaning is that it is time consuming and requires great amount of resources, thus people sometimes feel unmotivated to do it. Another point to make is that semantic annotations cannot solve the ambiguities of the real world.

Technologies for storing data and metadata. The data in relational databases is stored in a very structured way, making them a good choose for applications that relay

[3] https://archive.ics.uci.edu/ml/
[4] https://www.kaggle.com/datasets
[5] https://github.com/EpistasisLab/penn-ml-benchmarks
[6] https://www.openml.org/

on heavy data analysis. Moreover the referential integrity guarantees that transactions are processed reliably. While relational databases are a suitable choice for some applications, they have difficulties dealing with large amounts of data. On the other hand, NoSQL databases were designed primarily for big data and can be run on cluster architectures. Non-relational databases store unstructured data, with no logical schema. They are flexible, but this comes with the price of potentially inconsistent data.

Describing data and metadata. OntoDM is an ontology that describes the domain of DM, ML and KDD with a great level of detail. Because it covers a wide area, some parts would be irrelevant for our application. DMOP is ontology built with the special use case of optimizing the DM process. Nevertheless, both of them can be used for describing ML and DM datasets. DC vocabulary and PROV-O define a wide range of provenance terms, therefore both of them can be employed in the provenance metadata generation.

Repositories of machine learning datasets. The UCI repository offers a wide range of datasets, but they are not available through a uniform format or API. Although it also provides data descriptors for searching the data, a major setback is that none of the descriptors is based on any vocabulary or ontology, which certainly limits interoperability. Kaggle Datasets, KEEL, PMLB also provide similar meta annotations, but they all lack semantic interpretability. Another shortcoming of the UCI repository, KEEL and PMLB is that they don't allow uploading new datasets. All datasets stored in the OpenML repository can be downloaded in CSV or ARFF format. The annotation are based on Exposé ontology, and they can be downloaded in JSON, XML or RDF format. A major weakness of this repository is that annotations are not stored, but they are calculated on-the-fly and can not be used for semantic inference.

Frameworks for describing, storing and querying domain datasets. The NIF framework is very progressive in terms of semantic annotation, storing, and querying. Its advantages come from providing domain experts with the ability to contribute to the ontology development, by adding new terms through the use of Interlex. It has a powerful search engine, and it follows the OBDA paradigm. Heterogeneous data is stored in its original format. The user defined, keyword query is mapped to ontological terms to find synonyms, and then translated to a query relevant to the individual data store. With respect to the genomics domain, GOA database is favourable because of its high-quality annotations. Curators put extreme efforts in generating manual annotations. To speed up the query execution it uses the Solr document store. Another superiority of GOA database is that it provides advanced filtering of the annotations, for downloading customized annotation sets. The deficiency of NIF and GOA database is that they are not able to query and access the annotations in RDF format, which is an emerging standard for representing semantic information

4. PROPOSAL FOR SEMANTIC REPOSITORY OF DM/ML DATASETS

In this section, we propose three possible architecture designs of the semantic data repository for the domain of ML and DM. The proposals are based on the critical review of the approaches and technologies. Each of the proposed architectures has positive and negative sides, so there will be trade-off when choosing one.

The common part of the three designs is that DM and ML datasets will be annotated through a semantic annotation engine. The semantic query engine will receive SPARQL query as input, and it will bring back results in form of set of RDF triples. There will be SPARQL endpoint through which users can specify the query used as input in the semantic query engine. Another open possibility is to enable users to query data and metadata by simply writing keywords. Later, the system itself generates SPARQL query based on those keywords. The anotation schema used by the semantic annotation engine will be based on three different types of ontologies such as ontologies for DM and ML (e.g., OntoDT, OntoDM-core, Onto-KDD, DMOP), domain ontologies, and ontologies and schemes for describing provenance information (e.g., Dublin Core ontology, PROV-O). Part of the annotations will be generated automatically, e.g., annotations related to datatypes, while others will be semi-automatically because they require concept mapping, e.g., annotations based on domain ontologies.

We plan to build a web-based user interface that will enable users to search and query both datasets and metadata annotations. Users will be given a chance of uploading new datasets in CSV or ARFF fromat. Besides the dataset, users will be expected to specify some additional information about it such as data mining task they plan to execute on the data, domain, provenance information, descriptions of the attributes, etc. Since the whole process of semantic annotation can't be automatic, when new dataset is uploaded, it won't be immediately available on the site. First it must be curated, and only when the complete set of metadata annotations is generated, the metadata will be published online. The dataset itself will be released under clear data usage licence.

The three architectural designs differ in the way of storing the datasets. The metadata annotations will be RDF triplets and they will be stored in triplestore that optimizes physical storage. Next, we briefly explain the differences between storing the datasets and what are the effects on querying.

Proposal I. The simplest approach of storing a dataset would be to store it in RDF format in the same triplestore as the metadata. The datasets from their original format, will be converted to RDF triples. Having only one triplestore will ease querying, but it will require more storage capacity (see Figure 2).

Proposal II. The second option is to store the datasets in a relational database and the metadata in RDF triplestore. Datasets from CSV or ARFF format will be translated into a relational database. Here, querying becomes more complicated, for which we will need a federated query engine. A federated query engine allows simultaneous search on multiple data sources. A user makes a single query request, which is distributed across the management systems participating in the federation and translated to a query written in a language relevant to the individual system. We will have two data stores, one for the data itself and one for the metadata. For querying the two data stores, we will still use the same

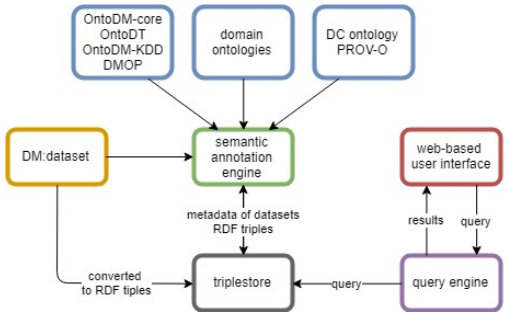

Figure 2. Architectural design I

RDF query language, SPARQL. In order to query the relational database with SPARQL, it will be mapped to virtual RDF graph (see Figure 3).

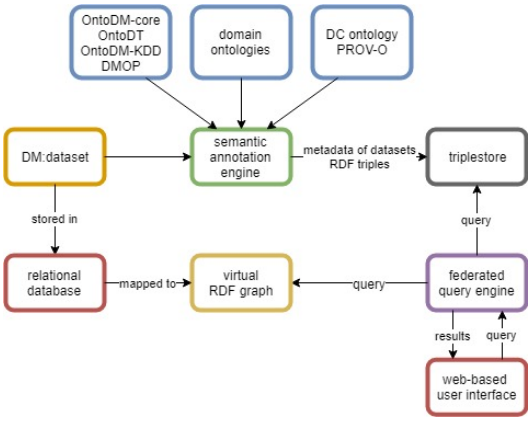

Figure 3. Architectural design II

Proposal III. Instead of mapping the relational database to virtual RDF graph, we can use the OBDA methodology and federated querying to use a combination of SQL queries and SPARQL queries. Metadata will be queried with SPARQL queries, but for the datasets, they will be mapped to SQL queries. The integrated results are brought back to the user (see Figure 4).

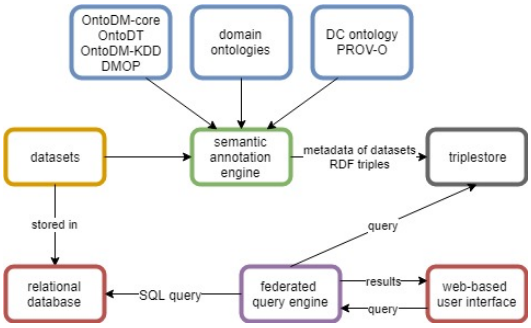

Figure 4. Architectural design III

5. CONCLUSION

We have conducted a literature overview of research being done in the field of semantic annotation, storage, and querying of ML and DM datasets. We also examined specific implementations of frameworks in the domain of neuroscience and genomics. Taking into consideration the critical assessment of the current state-of-the-art we will construct semantic data repository for ML and DM datasets. The semantic repository would be utilized for easy access of semantically rich annotated datasets and semantic inference. This, will improve the reproducibility and reusability in ML and DM research area. Moreover, annotating the datasets with domain ontologies will facilitate the process of understanding the analyzed data. As of now, we have three proposed architectural designs for the semantic data repository that differ in the way of storing the datasets. We will either store both data and metadata in a triplestore, or we will have multiple data stores which will require usage of tools and methods from the ontology based data access paradigm.

Acknowledgements

The authors would like to acknowledge the support of the Slovenian Research Agency through the projects J2-9230, N2-0056 and L2-7509 and the Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia through its scholarship program.

6. REFERENCES

[1] Mihaela A Bornea et al. Building an efficient rdf store over a relational database. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 121–132. ACM, 2013.

[2] Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic acids research*, 43(D1):D1049–D1056, 2014.

[3] Thomas R Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5-6):907–928, 1995.

[4] Amarnath Gupta et al. Federated access to heterogeneous information resources in the neuroscience information framework (nif). *Neuroinformatics*, 6(3):205–217, 2008.

[5] Jing Han et al. Survey on nosql database. In *Pervasive computing and applications (ICPCA), 2011 6th international conference on*, pages 363–366. IEEE, 2011.

[6] C Maria Keet et al. The data mining optimization ontology. *Web Semantics: Science, Services and Agents on the World Wide Web*, 32:43–53, 2015.

[7] Brian Matthews. Semantic web technologies. *E-learning*, 6(6):8, 2005.

[8] Panče et al. Panov. Ontology of core data mining entities. *Data Mining and Knowledge Discovery*, 28(5-6):1222–1265, 2014.

[9] Panče Panov et al. Ontodm-kdd: ontology for representing the knowledge discovery process. In *International Conference on Discovery Science*, pages 126–140. Springer, 2013.

[10] Panče Panov et al. Generic ontology of datatypes. *Information Sciences*, 329:900–920, 2016.

[11] Antonella Poggi et al. Linking data to ontologies. In *Journal on data semantics X*, pages 133–173. Springer, 2008.

[12] Petar Ristoski and Heiko Paulheim. Semantic web in data mining and knowledge discovery: A comprehensive survey. *Web semantics: science, services and agents on the World Wide Web*, 36:1–22, 2016.

[13] Gerd Stumme et al. Semantic web mining: State of the art and future directions. *Web semantics: Science, services and agents on the world wide web*, 4(2):124–143, 2006.

[14] Jan N Van Rijn et al. Openml: A collaborative science platform. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 645–649. Springer, 2013.

[15] Joaquin Vanschoren et al. Taking machine learning research online with openml. In *Proceedings of the 4th International Conference on Big Data, Streams and Heterogeneous Source Mining*, pages 1–4. JMLR. org, 2015.

[16] Stuart Weibel. The dublin core: a simple content description model for electronic resources. *Bulletin of the Association for Information Science and Technology*, 24(1):9–11, 1997.

[17] Mark D Wilkinson et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.