

ADA: Automatic Data Annotation for Data Ecosystems

Natalie Gdanitz¹, Sabine Janzen¹, Hannah Stein^{1,2}, Amin Harig¹ and Wolfgang Maaß^{1,2}

¹German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany

²Saarland University, Saarbrücken, Germany

Abstract

Data ecosystems have emerged as versatile platforms for managing and analyzing data from diverse sources, facilitating integration, collaboration and governance across organizations and systems. Annotated data are crucial for efficient and effective large-scale data ecosystems. However, there is a lack of full-fledged automatic annotation approaches for data ecosystems, with manual annotation by experts being the current requirement. Addressing specific annotation requirements of data ecosystems, we introduce ADA, an approach for automatic data annotation. ADA applies a semantic representation model called Data Product Description Object (DPDO) in JSON-LD and combines state-of-the-art models for metadata embeddings within an annotation pipeline. The approach extends technical metadata by essential concepts for data ecosystems, such as data provenance, quality, and accessibility. The effectiveness of ADA was evaluated using competency questions and data sets from diverse domains within the GAIA-X data ecosystem.

Keywords

Data ecosystems, Automatic data annotation, Metadata, Ontology

1. Introduction


Data ecosystems consist of centralized or decentralized platforms for managing and analyzing data from various sources, e.g., structured data, text, or images [1]. They are designed to facilitate data integration, sharing, collaboration, and governance across different systems, applications, and organizations, e.g., International Data Spaces¹, GAIA-X², Manufacturing-X³. Data ecosystems intend to help organizations to overcome data silos, enhance data-driven decision-making, and foster collaboration among data stakeholders [1]. Annotated data represent a fundamental requirement for efficient and effective large-scale data ecosystems [2, 3], referring to embedded metadata about structure, content, quality, and meaning of data. Automatically annotated data create a basis for high-quality curated data and allow data consumers to understand and interpret data without relying on external documentation or knowledge. Thus, seamless integration, sharing, data governance and trust, exploration, and large-scale data analysis within data ecosystems is enabled [4, 5]. While there exist conceptual ideas of generic cross-domain data

ISWC 2023 Posters and Demos: 22nd International Semantic Web Conference, November 6–10, 2023, Athens, Greece

✉ Natalie.Gdanitz@dfki.de (N. Gdanitz); Sabine.Janzen@dfki.de (S. Janzen); Hannah.Stein@dfki.de (H. Stein); Amin.Harig@dfki.de (A. Harig); Wolfgang.Maass@dfki.de (W. Maaß)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹<https://internationaldataspaces.org>

²<https://gaia-x.eu/>

³<https://www.plattform-i40.de/IP/Navigation/EN/Manufacturing-X/Manufacturing-X.html>

annotations for data ecosystems [6, 7, 8], there is a lack of full-fledged automatic annotation approaches for data ecosystems. So far, manual annotation by experts is required [9] (e.g., labeling training data, assigning data to developed concepts). Tackling the specific annotation requirements of data ecosystems with respect to data provenance, quality and context, accessibility, availability, and contractual information of open domain data is beyond the scope of existing research on automatic data annotation [2], e.g., [10, 11, 12]. In this work, we introduce ADA – an approach for automatic data annotation for data ecosystems. ADA works with a semantic representation model called Data Product Description Object (DPDO) operationalized in JSON-LD and combines multiple state-of-the-art models for metadata embeddings within an annotation pipeline (e.g., ontology development and knowledge graph population[13], metadata harvesting and extraction [14]). ADA builds up on existing ontological standards (e.g., Data Catalogue Vocabulary⁴) and extends technical metadata by essential concepts for data ecosystems, e.g., data provenance, quality, and accessibility. ADA supports open domain structured data sets, i.e., tabular data (CSV format). The approach was exemplified within an annotator service for automatic data annotation in data ecosystems. We were able to evaluate ADA by means of competency questions as well as data sets of diverse domains listed by the GAIA-X data ecosystem² extracted from Kaggle⁵, e.g., agriculture, construction, energy, geoinformation, or culture.

2. Automatic Data Annotation for Data Ecosystems (ADA)

In order to satisfy the requirements of data ecosystems, we first developed the DPDO which serves as semantic foundation for the annotation process. Our automatic annotation pipeline consists of 3 components: *Analyzer*, *controller*, and *provision engine* (see. figure 1).

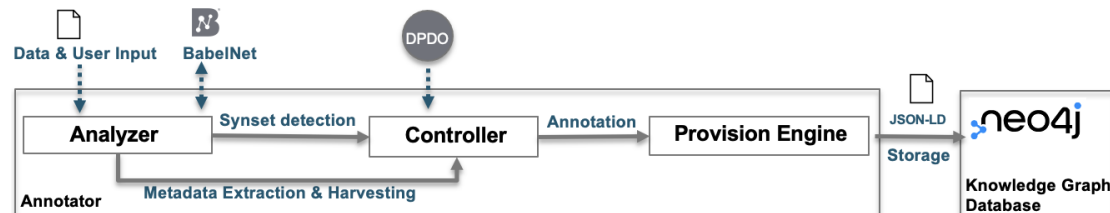


Figure 1: Sequence of automatic data annotation shown in a pipeline

Semantic representation model: Based on existing literature [15, 16, 17, 18, 19] and publicly available vocabularies (schema.org⁶, Data Quality Vocabulary⁷, Open Vocab⁸,

⁴<https://www.w3.org/TR/vocab-dcat-3/>

⁵<https://www.kaggle.com/datasets>

⁶<https://schema.org/>

⁷<https://www.w3.org/TR/vocab-dqv/>

⁸<https://vocab.org/open/>

Data Catalog Vocabulary⁹, DCMI Metadata Terms¹⁰, GAIA-X ontology^{11 12}), the semantic representation models information on data in data ecosystems within five facets¹³ (see figure 2). Potential data consumers require a *product description* in terms of context and metadata (e.g., topic, datatypes, data size). *Data quality* (e.g., referencing existing data quality standards such as ISO 8000¹⁴, metrics for the calculation of a quality score or accuracy) lays the foundation for the usability of data. Information on accessibility (i.e., access URL, technical support) and timeliness (i.e., last data modification, historical information on data versioning) enables the actual *usage* of the data. A data *business* transaction between ecosystem participants requires contractual information (i.e., contract description, price specifications, sanctions) and information about usage rights (e.g., license). Furthermore, *trust* between participants can be established by having information on the data provenance (i.e., name, locality, and contact of the data provider).

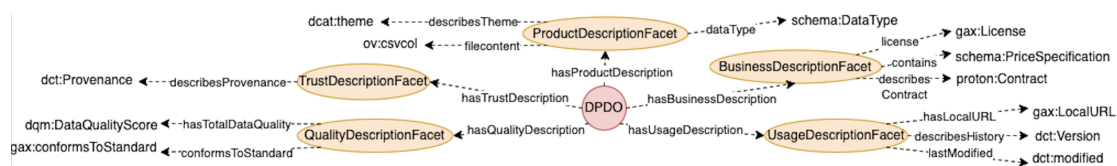


Figure 2: Simplified overview of the semantic representation model - Data Product Description Object

Analyzer component: Using the *GUI*, users can upload their data intended to be annotated. Within our demonstrator, they are additionally provided with sample JSON files that need to be uploaded alongside the data as additional input on the provider, contractual details, and data quality in order to fill the DPDO. In a real-life scenario, while participating in a data ecosystem, this information is expected to be filled once and then to be stored within respective platforms as described by existing concepts^{1 2}. The *analyzer* (see figure 1) then reads all files in order to extract the information required by the DPDO. Embedded metadata within the data file are harvested as described in [14] (i.e., title of the file, file size) or extracted directly from the file's content as in [14] (i.e., column names, data types). Column names and title of the file are used to generate a thematic allocation and to disambiguate the content of the data file by using them as lemmas to search for synsets and hypernyms in BabelNet¹⁵. Furthermore, the additional input JSON files are parsed and given alongside all other extracted information to the controller.

Controller component: The *controller* (see. figure 1) assigns all gathered information by the analyzer to a semantic description. The DPDO is operationalized in the form of JSON-LD, serving as a marker template that is filled rule-based with the results of the processed input files. The controller automatically maps extracted metadata, synsets and hypernyms to entities

⁹<https://www.w3.org/TR/vocab-dcat-3/>

¹⁰<https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>

¹¹<https://gaia-x.gitlab.io/gaia-x-community/gaia-x-self-descriptions/core/core.html>

¹²<https://gaia-x.gitlab.io/technical-committee/service-characteristics/widoco/participant/participant.html>

¹³A complete overview of modeled entities is given in our repository

¹⁴<https://www.iso.org/standard/81745.html>

¹⁵<https://babelnet.org/>

of the product description facet (see figure 2). Information on the user registry is mapped to the trust description facet and contractual information to the business and usage description facet. The additional user input is mapped onto remaining entities of all facets of the specified semantic scheme of the DPDO.

Provision engine: The *provision engine* (see. figure 1) transforms the generated JSON-LD file into a graph to be stored within the knowledge graph database Neo4j¹⁶ using a Cypher script. While we could have used a triplestore for storing our knowledge graph, one reason why we chose Neo4j is, that information on entities and relationships can be stored without creating extra nodes, leading to a more condensed representation of the data, which is particularly relevant in the case of large-scale data ecosystems. The resulting knowledge graph serves then as data catalogue or knowledge base within a data ecosystem^{1 2 3} that can be explored and queried.

Evaluation: We evaluated our approach using 16 competency questions (CQ) extracted from related work within the domain of semantic annotations and data ecosystems. For deriving these competency questions, we specifically focused on information to be required by stakeholders within data ecosystems. Examples would be '*Which context is the focus of the dataset?*', '*Which datatypes are used?*', '*What's the size of the dataset?*'¹⁷. Furthermore, we annotated publicly available data sets¹⁷ matching listed domains by the GAIA-X ecosystem, which we extracted from Kaggle (i.e., agriculture, energy, construction, finances, geo data, industry, culture, education, mobility, public sector, smart living). We investigated the resulting knowledge graph based on the defined CQ with the help of Cypher queries. We were able to answer all 16 CQs.

3. Conclusion

In this paper, we proposed ADA, an approach for automatic data annotation in data ecosystems. ADA leverages the semantic representation model DPDO (JSON-LD) and integrates state-of-the-art models for metadata embeddings, enabling seamless integration, sharing, governance, exploration, and large-scale data analysis within data ecosystems. ADA supports open domain structured data sets, i.e., tabular data in CSV format, and can be used by annotating experts and non-experts. By extending technical metadata with essential concepts such as data provenance, quality, and accessibility, ADA addresses the specific annotation requirements of data ecosystems and their stakeholders. The evaluation of ADA through competency questions and diverse tabular data sets demonstrates its effectiveness in supporting open domain structured data sets within the GAIA-X data ecosystem and beyond¹⁸.

¹⁶<https://neo4j.com/>

¹⁷All competency questions, references of related work, used data sets, executed queries, and query results are listed within our repository.

¹⁸Demonstration is given within a screencast <https://youtu.be/ra6IxTy4NUk>; Code of the service and evaluation results can be found within our GitHub repository <https://github.com/InformationServiceSystems/pairs-project/tree/main/Modules/ADA>.

4. Acknowledgement

This work was partially funded by the German Federal Ministry of Economics and Climate Protection (BMWK) under the contracts 01MK21008D and 01MK20015A.

References

- [1] F. Tocco, L. Lafaye, Data platform solutions, *Designing Data Spaces* (2022) 383.
- [2] M. Fassnacht, C. Benz, D. Heinz, J. Leimstoll, et al., Barriers to data sharing among private sector organizations, *Proc. of the 56th HICSS* (2023).
- [3] C. Mertens, J. Alonso, O. Lázaro, C. Palansuriya, et al., A framework for big data sovereignty: The european industrial data space (eids), in: *Data Spaces: Design, Deployment and Future Directions*, Springer International Publishing Cham, 2022, pp. 201–226.
- [4] W. Maass, Contract-based data-driven decision making in federated data ecosystems, *Proc. of the 55th HICSS* (2022).
- [5] M. Jarke, B. Otto, S. Ram, Data sovereignty and data space ecosystems, *Bus Inf Syst* 61 (2019) 549–550.
- [6] G. Solmaz, F. Cirillo, J. Fürst, T. Jacobs, et al., Enabling data spaces: Existing developments and challenges, in: *Proc. of the International Workshop on Data Economy*, 2022, pp. 42–48.
- [7] GAIA-X, GAIA-X Core Ontology, <https://gaia-x.gitlab.io/gaia-x-community/gaia-x-self-descriptions/core/core.html>, Accessed: 2023-07-10, 2022.
- [8] D. R. Firdausy, P. de Alencar Silva, M. van Sinderen, M. E. Iacob, Semantic discovery and selection of data connectors in international data spaces, *Proc. of I-ESA 1613* (2022) 0073.
- [9] S. Sharma, S. Jain, Comprehensive study of semantic annotation: Variant and praxis, *Int J Comput Intell Appl (ACI 2021)* 2823 (2021) 102–116.
- [10] P. Nguyen, I. Yamada, N. Kertkeidkachorn, R. Ichise, H. Takeda, Mtab4wikidata at semtab 2020: Tabular data annotation with wikidata., *SemTab@ ISWC 2775* (2020) 86–95.
- [11] V. Janev, M. E. Vidal, K. Endris, D. Pujic, Managing knowledge in energy data spaces, in: *Companion Proc. of the Web Conf. 2021*, 2021, pp. 7–15.
- [12] H. Drees, D. O. Kubitzka, J. Lipp, S. Pretzsch, et al., Mobility data space—first implementation and business opportunities, in: *Proc. of the 27th ITS World Congress*, 2021, pp. 11–15.
- [13] N. Abdelmageed, B. König-Ries, Meta2kg: transforming metadata to knowledge graphs, in: *Proc. of the 17th OM*, volume 3324, 2022, pp. 226–228.
- [14] S. Patankar, M. Phadke, S. Devane, Wiki sense bag creation using multilingual word sense disambiguation, *IAES Int* 11 (2022) 319.
- [15] S. Janzen, W. Maass, Smart product description object (spdo), in: *Poster Proc. of the 5th FOIS*, Citeseer, 2008.
- [16] M. Abramovici, Smart products, *CIRP Encyclopedia of Prod Eng* 59 (2014) 1–5.
- [17] A. Oberweis, V. Pankratius, W. Stucky, Product lines for digital information products, *Inf Syst* 32 (2007) 909–939.
- [18] K. L. Hui, P. Y. Chau, Classifying digital products, *Commun ACM* 45 (2002) 73–79.
- [19] S. Neumaier, J. Umbrich, A. Polleres, Automated quality assessment of metadata across open data portals, *ACM J Data Inf Qual* 8 (2016) 1–29.