Semantic Data Link: Bridging Domain-Specific Needs with Universal and Interoperable Semantic Models

Maximilian Stäbler*[1], Paul Moosmann*[2], Patrick Dittmer[3], DanDan Wang[4], Frank Köster[1] and Christoph Lange[2,5]

[1]German Aerospace Center (DLR) Institute for AI Safety & Security, Ulm, Germany
[2]Fraunhofer Institute for Applied Information Technology (FIT), Sankt Augustin, Germany
[3]Behörde für Verkehr und Mobilitätswende (BVM), Hamburg, Germany
[4]T-Systems International GmbH, Bonn, Germany
[5]RWTH Aachen University, Aachen, Germany

Abstract

The emergence of data-driven systems necessitates enhanced interoperability across diverse data ecosystems. Traditional approaches to semantic interoperability have been hindered by the complexity and specificity of ontologies, demanding significant expertise and resources for their development and maintenance. This paper introduces the Semantic Data Link (SDL) framework, a novel approach that aims to democratize data description and enhance semantic interoperability. SDL offers a domain and ontology-independent methodology, focusing on a multi-layered architecture that emphasizes decentralized semantics and categorizes data into definitional, structural, and contextual aspects. Developed as part of the Gaia-X 4 Future Mobility initiative, SDL is particularly pertinent to the mobility sector, where real-time data exchange and interoperability are crucial. This framework promises to bridge the gap between varying levels of expertise in semantic technologies and accelerate the development of semantically interoperable applications and services. We provide an in-depth discussion on the conceptual framework, design rationale, and implementation of SDL. The paper concludes with insights into the practical implications of SDL and prospective directions for future research in the quest for a seamless, interoperable data landscape.

Keywords

Semantic Interoperability, Data Ecosystems, Dataspaces, Domain Agnostic Framework

1. Introduction and Motivation

Efficient data exchange and interoperability are crucial in various ecosystems [1], especially in the mobility sector [2], where they face significant challenges due to real-time data sharing demands across domains [3]. This situation exacerbates urban issues like congestion and pollution, as interoperability deficits limit the development of smart, connected urban mobility services [2]. Data heterogeneity, marked by incompatible message formats, complicates seamless data interoperability [4]. Although traditional ontology approaches have aimed at

The Second International Workshop on Semantics in Dataspaces, co-located with the Extended Semantic Web Conference, May 26 – 27, 2024, Hersonissos, Greece

maximilian.staebler@dlr.de (M. Stäbler*); paul.moosmann@fit.fraunhofer.de (P. Moosmann*)
0000-0003-1311-3568 (M. Stäbler*); 0009-0005-2114-8578 (P. Moosmann*); 0000-0001-9879-3827 (C. Lange)

[*] These authors contributed equally to this work

alignment [5, 4, 6], merging [7, 8], and matching [9, 10, 5] to address these issues, they adhere to the “80/20” principle, where automated solutions handle most discrepancies but still leave complex cases needing manual refinement [11]. The advancement of semantic interoperability is constrained by the complexity and specialized knowledge required for ontology development and deployment [11, 12]. This complexity necessitates that stakeholders possess advanced skills in semantic technology and ontology engineering, limiting the prevalence of semantically interoperable solutions.

However, the legislative landscape is evolving to address some of these barriers. A noteworthy advancement in this direction is the adoption of the European Data Governance Act (DGA), which represents a important step towards enhancing trust and expanding data availability across Europe. Under the European Data Act (Data Act), mobility service providers are mandated to make data exportable, understandable, and reusable for other stakeholders in the value chain, including end-users and manufacturers. This legislative move underscores the importance of interoperability and data sharing, potentially easing some of the complexity and expertise barriers associated with semantic technologies. In a communication to the European Parliament, the European Commission has called for the development of a common European Mobility Data Space. This communication also calls for interoperability (interlinking) between different dataspaces, which are defined as collaborative digital architectures enabling secure and sovereign data exchange among diverse stakeholders. In parallel, initiatives such as the European Open Science Cloud EOSC and the adoption of the FAIR (Findable, Accessible, Interoperable, Reusable) Data Principles further reinforce the importance of developing robust dataspaces. These challenges highlight the critical need for solutions that bridge semantic gaps between interconnected dataspaces, moving from inefficiency to semantic interoperability [3].

A specific example of this trend towards increasing data transparency and interoperability at a more localized level is seen in the City of Hamburg, Germany. Due to the Hamburg Transparency Act (Hamburger Transparenzgesetz), public sector datasets in Hamburg must be made directly accessible and published as open data, while ensuring the protection of personal data. Data is published via the Hamburger Transparenzportal using the European metadata standard DCAT-AP. However, this standard primarily concerns how data is cataloged, but not its format, which is chosen individually. Consequently, the need for semantic interoperability is not just a broad European challenge but indeed also exists at a localized level, as demonstrated by the Hamburg initiative.

This paper presents the Semantic Data Link (SDL) framework as an innovative solution to establish semantic interoperability between previously incompatible systems, data sources and applications. With an emphasis on ease of use and domain agnosticism, SDL addresses a wide range of domains and applications. In particular, it is designed to be accessible to domain experts without prior knowledge of semantic technologies such as Resource Description Framework (RDF) or Shapes Constraint Language (SHACL), enabling them to create meaningful and universally applicable descriptions. Developed as part of the Gaia-X 4 Future Mobility (GX4FM) project family – a Gaia-X initiative – SDL embodies the vision of a more connected, efficient and innovative mobility future. It enables the free and meaningful flow of data across borders and sectors. The project family involves more than 80 partners from industry, research and the public sector, and each of them requires continuous ontology updates to keep pace with evolving domain knowledge and practices. Updates often lag behind due to the dynamic and resource-intensive nature of these revisions. In particular, manual intervention is normally needed, which slows down the interoperability process and introduces the risk of inconsistencies. With these challenges in mind, we developed SDL to facilitate semantic interoperability, the effectiveness of which will be validated in partnership with the City of Hamburg to enhance future mobility applications. The collaboration with industry partners and municipalities demonstrates the industry-driven approach of this project and highlights its potential for widespread adoption and impact.

The paper is structured as follows: We begin with a background section that sets the context for our study, highlighting the current state of semantic interoperability. We then present the SDL, detailing its conceptual underpinnings and describing its implementation process. The paper concludes with a discussion of our findings, implications for practice, and directions for future research in this evolving field.

2. Background

In this section, we will give an overview of the background of our work. Specifically, we will cover the topics of Semantic Web technologies, corresponding tool support, and give a short introduction to the dataspace initiative Gaia-X. This work emerged in the context of building a dataspace based on Gaia-X, which is directly connected to the topic of Semantic Web technologies and tools since their use is central to the implementation of federated dataspaces in Gaia-X. A general overview of the role of semantics in dataspaces is given by Theissen-Lipp et al. [13].

Semantic Web Technologies and Tool Support. There have been several reviews of the current status of Semantic Web technologies in recent years, such as the works of Hitzler [14] or Patel and Jain [15]. These works identify the W3C standards RDF, RDF Schema, OWL, and SPARQL as core technologies. In our work, we extend this list with SHACL, which builds upon RDF and is used for validating RDF graphs against a set of conditions. SHACL plays a vital role not only in the functionality of the SDL but also in underpinning semantic technologies within dataspace initiatives, such as Gaia-X or IDS. Based on these core technologies, further vocabularies were defined, which today, due to their widespread adoption, can also be seen as part of the core of the Semantic Web. Examples include the SKOS and DCAT vocabularies [15, 16]. These are also used in the context of the SDL and Gaia-X. While the development of further vocabularies leads to a stronger (and more standardized) core, Hogan [16] compiled various criticisms regarding the Semantic Web, with one being that the standards are complex and difficult to understand. To tackle this problem, various tools have been developed to aid users of Semantic Web technologies. The survey of Khamparia and Pandey gives a good overview of existing Semantic Web reasoners and tools [7]. Some prominent examples include the Protégé ontology editor, the ELK reasoner or the Linked Open Vocabularies (LOV) database [14, 17]. Even though the Semantic Web is being criticized for its lack of usable systems and tools [16], a variety of isolated tools exist, that can be built upon or integrated into the SDL. That way, the SDL decreases the impact of lacking Semantic Web expertise by its users, without being redundant. E.g., we use LOV to reuse existing vocabularies and foster interoperability. We also reduce the complexity of creating OWL and SHACL schemas, by using the LinkML model to generate them. LinkML is a general purpose modeling language. While LinkML is designed to work in harmony with semantic RDF-based frameworks, it uses the human-readable data serialization language YAML, making it more approachable for non RDF experts [18].

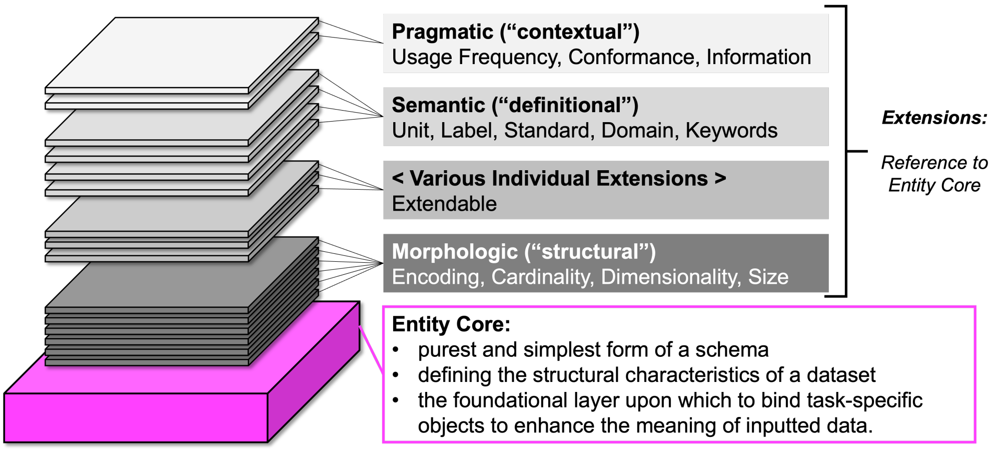
Gaia-X. Gaia-X, a European initiative for secure data sharing, employs federated services with Semantic Web technologies to ensure data trustworthiness. Participants and services within Gaia-X must provide credentials as specified by the W3C Verifiable Credential Data Model. The content of the credentials is detailed in the Gaia-X Trust Framework and their corresponding OWL and SHACL schemas are retrievable from the Gaia-X Registry. Also, implementations using the LinkML schema already exist and can be found in the repository of the Gaia-X Working Group Service Characteristics. Since SDL was developed to streamline the creation of credentials and associated OWL and SHACL schemas for usage within and beyond the Gaia-X context, we decided to build SDL on top of LinkML to support already existing resources.

3. Semantic Data Link

Recognizing the vital necessity of enhancing semantic interoperability within data ecosystems, the SDL framework emerges as a transformative solution. SDL proposes a domain agnostic approach to synchronize data models, data services and representation formats. This methodology paves the way towards the democratization of the development of meaningful data descriptions for individuals lacking comprehensive proficiency in semantic technologies. This section elaborates SDL's conceptual foundation with key functionalities, the rationale behind its innovative design, and the implementation, which aims to provide a comprehensive solution for data heterogeneity and interoperability.

3.1. Conceptual Framework

SDL enables a uniform description framework for individual data records without mandating particular standards. This flexibility allows for the integration of existing semantic descriptions from diverse domains, including but not limited to datasets, services, digital twins, and other relevant fields, facilitating compatibility with industry standards such as OPC UA. Its intermediate layer bridges disparate data descriptions without necessitating alterations to the source systems, thereby providing advantages to a broad range of stakeholders. *Publishers* do not need to adopt new formats for compatibility, which lowers the entry barrier and therefore enhances data availability. *Consumers* benefit from standardized descriptions and improved data comparability and usability from various sources. Also, enhanced data discoverability allows *all users* to precisely locate necessary data for applications and services. Figure 1 illustrates the stacked approach of the SDL and emphasizes decentralized semantics by categorizing data into definitional (*semantic*), structural (*morphologic*), and contextual (*pragmatic*) aspects for complex entities like services or datasets. This categorization aids in aggregating digital representations with rich contextual understanding. At its core, SDL features an *Entity Core* encapsulating essential dataset or service attributes (e.g., provider ID, name) and assigns a unique identifier to


Pragmatic ("contextual")
Usage Frequency, Conformance, Information
Semantic ("definitional")
Unit, Label, Standard, Domain, Keywords
< Various Individual Extensions >
Extendable
Morphologic ("structural")
Encoding, Cardinality, Dimensionality, Size
Extensions:
Reference to Entity Core

Entity Core:

- purest and simplest form of a schema
- defining the structural characteristics of a dataset
- the foundational layer upon which to bind task-specific objects to enhance the meaning of inputted data.

Figure 1: SDL Layers: Scalable approach to semantic annotation, dividing data characteristics into three principal layers: Pragmatic, Semantic, and Morphologic, supported by additional extendable extensions for domain- and application specific details. This stratification facilitates modular interoperability and provides a robust framework for data description across varied applications and domains. Practitioners can decide individually how many layers they want to use, depending on the application. Each attribute of each extension represents one layer.

each entity, streamlining referencing and interactions in the data ecosystem. This approach draws upon the theoretical foundation laid out by the Overlay Capture Architecture (OCA). OCA is a framework designed to enable data harmonization and privacy compliant sharing across different governance frameworks. *Extensions* are critical in SDL because they add layers of metadata to the entity core, taking semantic, morphological, contextual and other individual and application-specific dimensions into consideration. The selection of attributes for the SDL framework was significantly influenced by a combination of the foundational principles from the OCA and extensive deliberations within the GX4FM project's Expert Group on semantics, ensuring alignment with both theoretical and practical requirements of data interoperability. This is an evolving, community-driven effort, open to incorporating additional attributes in future versions to better meet the emerging needs and insights of the diverse stakeholder community. This architecture supports modular interoperability and semantic integrity, is domain agnostic, and harmonizes data models and formats across boundaries. Its layered design enriches data and service descriptions, improving comparability and interoperability to efficiently meet the needs of diverse domains and applications.

3.2. Rationale Behind SDL's Design Choices

The SDL framework seeks to overcome the limitations of existing semantic interoperability frameworks, responding to industry calls for a solution that is more flexible and responsive than traditional, rigid systems, thereby offering a user-friendly alternative capable of evolving with technological and business needs. Compatibility with arbitrary ontologies creates a flexible framework for data description. This design choice directly responds to the industry's requirement for simplified semantic technologies that can accommodate the diverse backgrounds of

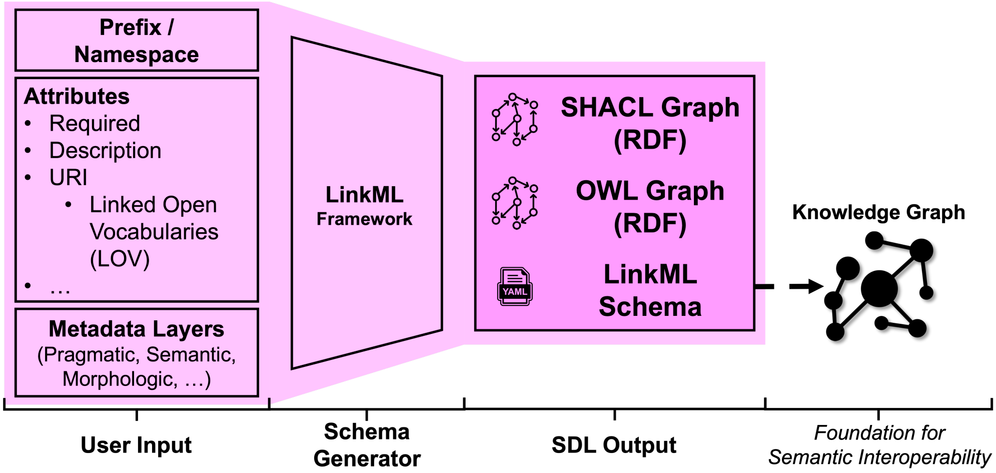


Figure 2: The frontend implementation of the SDL allows the user to input the namespace, prefix, class, and related attributes that are being described by the created schema. It supports the reuse of existing vocabularies by embedding LOV and guides the user by providing metadata layers to enhance the data description. This information is then translated using the LinkML Framework to generate a SHACL graph, an OWL graph and a LinkML Schema as output. In a future release the information stored in the LinkML schema will also be translated into a Knowledge Graph.

users, including domain experts, without any preliminary semantic knowledge. By decentralizing semantics, SDL broadens user participation, which leads to a more inclusive data ecosystem. The presented *extensions* and *entity core* concept enhances the data's contextual understanding, which is crucial for interoperability. This not only improves data description precision but also ensures the ecosystem's adaptability and scalability, meeting industry needs for evolving data landscapes.

3.3. Implementation of SDL

The implementation of SDL consists of a front-end component that allows user input and a back-end component that uses this input to generate an output consisting of structured data in YAML and RDF format. To collect user input, we provide a simple UI that allows the user to define the data they want to describe. Since SDL is designed to help non-Semantic Web experts create RDF data, the UI enables gathering all the necessary information without requiring the user to apply any Semantic Web technologies. Figure 2 shows an overview of the main components of our SDL implementation. The user input component consists of (1) two text fields that collect information about the namespace and prefix used, (2) a part where the user can add attributes necessary to describe the data, and (3) additional layers that can be used to extend the data by certain predefined attributes to improve the data description. While the currently implemented layers focus on the basic metadata description of datasets (see Figure 1), it is possible and intended to implement additional layers in future releases. Once the user input is complete, the schema generation component is used to create a YAML file that follows the LinkML model from the input. LinkML provides a framework for generating RDF schemas from the YAML

Table 1
Top 3 LOV results for attribute *longitude*. *score* indicates how well a *prefixName* matches the search attribute (*longitude*) [19].

Rank	prefixedName	uri	score
1	geo:long	http://www.w3.org/2003/01/geo/wgs84_pos#long	0.856
2	og:longitude	http://ogp.me/ns#longitude	0.556
3	geo:lat	http://www.w3.org/2003/01/geo/wgs84_pos#lat	0.507

file. We chose LinkML as the basis for SDL to seamlessly integrate future Gaia-X schemas already modeled in LinkML, leveraging existing work and its framework for straightforward implementation and user-defined attribute modeling through the pre-defined parameters of LinkML. These include (1) *required* and *multivalued* to model cardinality constraints, (2) a *description* of the attribute, (3) a regex *pattern* that defines constraints on the attribute value, and (4) a *URI* that can point to already defined attributes to facilitate reuse of existing vocabularies. We enhance vocabulary reuse and new attribute creation by automatically linking to the Linked Open Vocabularies (LOV) database. When users create an attribute and name it, a LOV search auto-executes, presenting the top ten results for selection. An example is shown in Table 1. Any of these attributes can be selected to be used if they meet the user's requirements. If none of the suggestions fit, a new attribute is created under the previously defined namespace. Finally, the LinkML framework generators are used to generate an OWL graph and a SHACL graph from the LinkML YAML created from the user input. While LinkML also provides generators to generate schemas in other formats from the YAML file, we limit the output to OWL and SHACL graphs, as these are the relevant schemas in the context of the Gaia-X dataspace initiative. This can be easily adapted for other use cases. A future implementation of the SDL will also use the LinkML YAML to transfer the stored information into a knowledge graph, which plays an important role in promoting semantic interoperability. The current implementation of the SDL, as described in this section, can be found in the form of a GitHub repository. This repository also contains instructions for installing the SDL locally using Docker.

4. Summary and Future Work

Conclusion. In summary, SDL represents a significant advancement towards achieving seamless semantic interoperability within data ecosystems. By abstracting the complexities of domain-specific ontologies and providing a user-friendly, multi-layered architecture, SDL democratizes data description and fosters an inclusive ecosystem of data exchange. The framework's potential was evidenced through its applicability in the mobility sector, with future enhancements poised to extend its utility across various domains. The development of SDL aims to improve interoperability between existing data infrastructures while lowering the complexity hurdles of semantic technologies.

Limitation. The SDL leverages the LinkML Framework for generating OWL and SHACL graphs, with its limitations bifurcating into: (1) LinkML's OWL and SHACL generators inadequately translating defined constraints within the schema to corresponding graphs, exemplified by the non-translated properties such as *any_of* or *equals_string_in*, and (2) the incapacity of the LinkML schema to represent certain semantic details expressible in OWL or SHACL. Addressing these limitations is imperative, involving the expansion of the LinkML schema to encompass broader semantic expressions, and enhancing the generators for full property translation. Continued refinement will explore extending the existing LinkML schema and evaluating alternative frameworks to ensure SDL's adaptability to future semantic interoperability requirements.

Future Work. Moving forward, we have identified several areas for future research. Firstly, we suggest focusing on the enhancement of SDL through the creation of a Knowledge Graph that enables advanced interoperability. This graph should be composed of the descriptions generated by the SDL. Secondly, there is a need to refine SDL's multi-layered architecture to broaden its adoption. Lastly, further evaluation in a real-world setting is essential to validate the effectiveness and applicability of these advancements. The ultimate goal of these proposed areas of research is to achieve scalable and resilient interoperability, democratize data usage across diverse ecosystems, and accomplish these without the necessity for specialized semantic expertise.

Acknowledgments

This work was supported by the German Federal Ministry for Economic Affairs and Climate Action and by the European Commission, whose funding has been crucial to our research efforts. We also thank the *Core-Working-Group Semantics* of the Gaia-X 4 Future Mobility project family for their significant contributions to the Semantic Data Link, enhancing our work on data interoperability and governance.

References

[1] R. Henßen, M. Schleipen, Interoperability between opc ua and automationml, Procedia CIRP (2014). doi:10.1016/j.procir.2014.10.042.

[2] S. Paiva, M. A. A. G. Tripathi, N. Feroz, G. Casalino, Enabling technologies for urban smart mobility: Recent trends, opportunities and challenges., Sensors (2021). doi:10.3390/s21062143.

[3] A. Kouroubali, D. G. Katehakis, The new european interoperability framework as a facilitator of digital transformation for citizen empowerment., Journal of Biomedical Informatics (2019). doi:10.1016/j.jbi.2019.103166.

[4] F. Ardjani, D. Bouchihaand, M. Malki, Ontology-alignment techniques: Survey and analysis, International Journal of Modern Education and Computer Science (2015). doi:10.5815/ijmecs.2015.11.08.

[5] M. A. Khoudja, M. Fareh, H. Bouarfa, Ontology matching using neural networks: Survey and analysis, International Conference on Independent Component Analysis and Signal Separation (2018). doi:10.1109/icass.2018.8652049.

[6] A. H. Nejhadi, B. Shadgar, A. Osareh, Ontology alignment using machine learning techniques, International Journal of Computer Science and Information Technology (2011). doi:10.5121/ijcsit.2011.3210.

[7] A. Khamparia, B. Pandey, Comprehensive analysis of semantic web reasoners and tools: a survey, Education and Information Technologies (2017). doi:10.1007/s10639-017-9574-5.

[8] M. Fahad, N. Moalla, A. Bouras, Detection and resolution of semantic inconsistency and redundancy in an automatic ontology merging system, Journal of Intelligence and Information Systems (2012). doi:10.1007/s10844-012-0202-y.

[9] X. Liu, Q. Tong, X. Liu, Z. Qin, Ontology matching: State of the art, future challenges, and thinking based on utilized information, IEEE Access (2021). doi:10.1109/access.2021.3057081.

[10] A. Bento, A. Zouaq, M. Gagnon, Ontology matching using convolutional neural networks., International Conference on Language Resources and Evaluation (2020). doi:null.

[11] Z. Boukhers, C. Lange, O. Beyan, Enhancing data space semantic interoperability through machine learning: a visionary perspective, The Web Conference (2023). doi:10.1145/3543873.3587658.

[12] M. Stäbler, T. M. Guggenberger, W. DanDan, R. Mrasek, F. Köster, C. Langdon Schlueter, Bridging Data Domains: Towards Semantic Interoperability in Heterogeneous Data Ecosystems and Data Spaces, [Manuscript submitted for publication] (2024).

[13] J. Theissen-Lipp, M. Kocher, C. Lange, S. Decker, A. Paulus, A. Pomp, E. Curry, Semantics in dataspaces: Origin and future directions, The Web Conference (2023). doi:10.1145/3543873.3587689.

[14] P. Hitzler, A review of the semantic web field, Communications of The ACM (2021). doi:10.1145/3397512.

[15] A. Patel, S. Jain, Present and future of semantic web technologies: a research statement, International Journal of Computers and Applications (2019). doi:10.1080/1206212x.2019.1570666.

[16] A. Hogan, The semantic web: Two decades on, Social Work (2020). doi:10.3233/sw-190387.

[17] F. Gandon, A survey of the first 20 years of research on semantic web and linked data, Ingénierie Des Systèmes D'information (2018). doi:10.3166/isi.23.3-4.11-38.

[18] S. Moxon, H. Solbrig, D. Unni, D. Jiao, R. Bruskiewich, J. Balhoff, G. Vaidya, W. D. Duncan, H. Hegde, M. Miller, M. H. Brush, N. Harris, M. Haendel, C. Mungall, The linked data modeling language (linkml): A general-purpose data modeling framework grounded in machine-readable semantics, International Conference on Biomedical Ontology (2021). doi:null.

[19] P.-Y. Vandenbussche, G. A. Atemezing, M. Poveda-Villalón, B. Vatant, Linked open vocabularies (lov): a gateway to reusable semantic vocabularies on the web, Sprachwissenschaft (2016). doi:10.3233/sw-160213.