Towards ShowVoc: dataset publication and browsing

Armando Stellato[1,2,*,†], Manuel Fiorelli[1,2,†], Tiziano Lorenzetti[2,†] and Andrea Turbati[2,†]

[1]Tor Vergata University of Rome, Italy
[2]Lore Star s.r.l., Rome, Italy

Abstract

ShowVoc is a web-based, multilingual, platform for publication and consumption of datasets complying with Semantic Web standards. Born in the context of the ISA[2] European programme for the development of digital solutions for interoperable cross-border and cross-sector public services, ShowVoc aims at providing a one-stop shop for maximizing the diffusion of semantic and lexical resources as Linked Open Data. To this end, ShowVoc combines traditional data provisioning following LOD policies with global activities (e.g. global search, navigation of dataset relationships/alignments, translation API benefiting from multilingual datasets and linksets). A rich dataset browsing interface provides dedicated support for diverse data models: OWL ontologies, SKOS/SKOS-XL thesauri, OntoLex-Lemon lexicons and generic RDF datasets and linkage possibilities (EDOAL, XKOS). A metadata registry completes the offer combining different metadata vocabularies into an advanced catalog that can be inspected through a convenient user interface and LOD best practices. Finally, ShowVoc is an ideal companion to VocBench, a popular collaborative editing environment for Semantic Web resources, complementing it for realizing an entire workflow embracing all stages of a dataset life, from realization and maintenance, to release and publication.

Keywords

Semantic Web, Linked Open Data, Dataset Catalogs, Metadata repositories, Data consumption

1. Introduction

The Semantic Web [1], which is being built according to Linked Data [2] best practices, is based on the decentralized publication of disparate but interlinked datasets that together form a huge global graph. Although resolvable URIs and query-through-discovery are the defining access mechanism for a machine-accessible Web and the focus is on linking records, there is still a need, especially for humans, for a coarse-grained perspective made of browsing, querying and visualization capabilities over the published resources.

Discovery by link traversal - a la “follow your nose” - is closely related to people surfing the Web in search of information. If we take this analogy seriously, then we should consider people's reliance on search engines as an entry point to the Web. Although semantic web search engines are not as common as they could be, there has been a proliferation of dataset catalogs, both in specific domains and across the web, which play a similar role.

In this paper, we present ShowVoc, a platform for dataset publication and exploitation, which addresses both needs: it allows for the publication of datasets with resolvable URIs and a more sophisticated browsing experience than simple subject pages while offering a fully-fledged data portal for linked datasets. ShowVoc can be seen as a companion to VocBench 3 [3], a platform for dataset development and maintenance, inheriting many of its features, such as its advanced multi-model support. However, while

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy
*Corresponding author.
†These authors contributed equally.
stellato@uniroma2.it (A. Stellato); manuel.fiorelli@uniroma2.it (M. Fiorelli); tiziano.lorenzetti@lorestar.it (T. Lorenzetti); andrea.turbati@lorestar.it (A. Turbati);
0000-0001-5374-2807 (A. Stellato); 0000-0001-7079-8941 (M. Fiorelli); 0000-0001-5676-8877 (T. Lorenzetti); 0000-0002-6214-4099 (A. Turbati)
© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3762/577.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

most of the operations in VocBench 3 deal with individual datasets, ShowVoc adds a number of cross-dataset operations that rely on managing multiple datasets. These include global search, translation and alignment management, which are based on the idea that multiple datasets contribute to a sort of giant virtual reference for terminology and translation.

ShowVoc is open source and made available under the BSD-3-Clause license. The project official web site is https://showvoc.uniroma2.it/. Source code and deployment artifacts are hosted on Bitbucket at https://bitbucket.org/art-uniroma2/showvoc.

The paper is structured as follows. Section 2 discusses related work. Section 3 briefly describes the architecture of the ShowVoc. Section 4 delves into the main features of ShowVoc. Then, Section 5 argues for its impact. Finally. Section 6 draws conclusions.

2. Related work

We should probably start our discussion of related efforts on data portals with CKAN[2], which had established itself as a de facto standard, particularly in the public sector, with its rich API and support for federation of catalogs. Within the scientific community, Zenodo[3] (based on the open-source software Invenio[4]) has established itself as the go-to solution for ensuring data persistence, similar to what arXiv[5] has achieved for preprint publication. Regarding the impact of archiving, the Open Archive Initiative [4] (OAI) is certainly of interest, especially for its metadata harvesting protocol (OAI-PMH).

None of these solutions are specifically tailored to semantic web datasets, beyond the ability to store dumps as files. For this reason, we now consider catalogs of semantic web datasets. LOV [5] is a catalog of Linked Data Vocabularies, while LOD Cloud[6] hosts, in addition to the eponymous figure, a catalog of the datasets that actually drove the creation of the former. There are also domain-specific catalogs, most notably BioPortal [6] for ontologies related to the biomedical domain. Today, the OntoPortal Alliance [7] has taken over BioPortal's original source code, which is being adopted by portals across various domains, such as agrifood [8] and biodiversity and ecology [9]. Within the field of solid Earth science, we mention a European initiative [10] using metadata and semantic technologies for integration and access of data from diverse sources.

Alignment management, which is addressed by many Semantic Web catalogs, including LOV and OntoPortal, can also be a use case in its own right. For example, the Alignment API [11] ships with a server that can handle an ontology network, with the ability to compute, retrieve, combine, and otherwise manipulate alignments between ontologies. In a related vein, the ELEXIS [12] project aims at linking (legacy) language resources via linked data, and has developed a standard REST API for accessing a catalog of dictionaries. Both of these applications are covered by ShowVoc, as we will see later.

We conclude the section on related work by discussing the publication of linked data. Pubby[7] implements resolvable URIs by querying a SPARQL endpoint. This software is now discontinued, but newer alternatives such as LodView[8] and Loddy[9] have emerged. The triple store Virtuoso [13] has even integrated this feature without the need for third-party software. Subject pages were even took as a paradigm for data editing, in systems such as TemaTres [14] or OntoWiki [15].

Subject pages are not always the best choice for browsing through your data. For example, SKOSMOS[10] became a popular choice for publishing a collection of SKOS thesauri with more sophisticated browsing capabilities, including search and indexing. For ontologies, the need for more organized documentation became apparent. This can be automatically generated from the ontology definitions themselves using tools such as LODE [16] or, more recently, WIDOCO [17]. This feature has also been developed within VocBench 3 using its custom reporting facility. In fact, both browsing tools and documentation pages can be used to resolve URIs and both use cases are supported by ShowVoc.

3. Architecture

ShowVoc has been designed as a single-page application (SPA), with a frontend running inside a web browser that communicates with a back-end server through a REST-like API.

The frontend is developed in TypeScript using the Angular framework and can be delivered to users by any web server or CDN (Content Delivery Network).

The backend server is based on Semantic Turkey [18], the same RDF services platform that powers VocBench 3. The platform, based on an opinionated

[2] https://ckan.org/
[3] https://zenodo.org/
[4] https://inveniosoftware.org/
[5] https://arxiv.org/
[6] https://lod-cloud.net/
[7] https://github.com/cygri/pubby
[8] https://github.com/LodLive/LodView
[9] https://bitbucket.org/art-uniroma2/loddy
[10] https://skosmos.org/

combination of the Spring Boot[11] and PF4J[12] frameworks, supports the development and publication of services related to RDF data. Prebuilt services address multiple models and various concerns such as history, validation, and import/export. PF4J makes it easy to deploy new services and extend the capabilities of existing ones by providing implementations of the extension points on which they depend. For example, the *export service* defines extension points that can be used to provide both the conversion logic to a particular serialization format (i.e., reformatting exporter) and the ability to deploy data to particular targets (i.e., deployer). Semantic Turkey ships with implementations of these extension points for common use cases, but (as mentioned) new ones can be added to the system.

Semantic Turkey relies on the RDF4J framework to process RDF data and interact with triple stores (i.e., RDF database management systems), both in-process within Semantic Turkey or managed as separate processes. The latter option is the preferred method, allowing the use of enterprise-grade triple stores such as Ontotext 's GraphDB[13].

VocBench 3 and ShowVoc can share the same backend server, with a common set of projects that can be conveniently made accessible through ShowVoc. However, the common, recommended scenario is to have separate backend servers (and, most important, different storage solutions with different expected workloads) for ShowVoc and VocBench 3 so that managers of projects developed within VocBench can submit datasets to the ShowVoc instance for publication [19].

4. Features

We will introduce here the main features of ShowVoc and discuss their relevance to the system's use cases (see **Figure 1** for an overview of its UI).

Contributions. A ShowVoc dataset portal can optionally allow contributions from visitors. These can request the addition of a new dataset, possibly after conversion from a non-RDF format, and the creation of a development environment within an associated VocBench instance.

Content negotiation. ShowVoc's use cases extend beyond cataloging third-party datasets, as it also addresses the needs of original dataset publishers. These can set up ShowVoc as an advanced browser for their datasets; however, Linked Data rules require that entity identifiers be resolvable via HTTP, which is perhaps the defining characteristic of the Linked Data paradigm. ShowVoc supports this as well, with some endpoints that can be queried by a reverse proxy associated with the domain to implement content negotiation and generate different variants, including machine-readable serializations and a human-friendly page.

Multi-model support. ShowVoc inherits from Semantic Turkey the ability to manage arbitrary RDF datasets, coupled with convenient facilities for OWL ontologies and other less formal Knowledge Organization Systems (KOS) modeled in SKOS, as well as OntoLex/Lemon lexicons. In addition, ShowVoc is aware of various lexicalization models for grounding data in natural language, including RDFS, SKOS(-XL), and OntoLex-Lemon.

At the user interface level, this flexibility is first visible in ShowVoc's resource view, which can display the description of any resource, divided into sections that roughly correspond to different properties. As such, the resource view can display any type of resource, but it can be specialized and made efficient for specific modeling vocabularies through a combination of customized templates (defining the prominent sections for different resource types), specialized sections (e.g., the one grouping class axioms), as well as dedicated support for specific mechanisms (e.g., proper rendering of class axioms in Manchester syntax). ShowVoc works seamlessly with different lexicalization models, which are taken into account when selecting the "labels" for displaying a resource (instead of its IRI or qname) or when populating the "lexicalizations" section of the resource view (which abstracts over the specific lexicalization model).

ShowVoc also provides a number of views to browse the content of the dataset, depending on its nature, such as a class tree, instance list, property tree, concept tree, etc.

Seamless navigation between local and remote datasets. A user who encounters a reference to a resource outside the dataset being browsed can easily jump to it. If the resource belongs to another dataset in the same ShowVoc installation, the user interface automatically switches to that dataset and focuses on the target resource. External resources can also be displayed in a modal dialog populated with information retrieved by deferecentiation or a SPARQL endpoint, if known to the system.

[11] https://spring.io/projects/spring-boot
[12] https://pf4j.org/
[13] https://www.ontotext.com/products/graphdb/

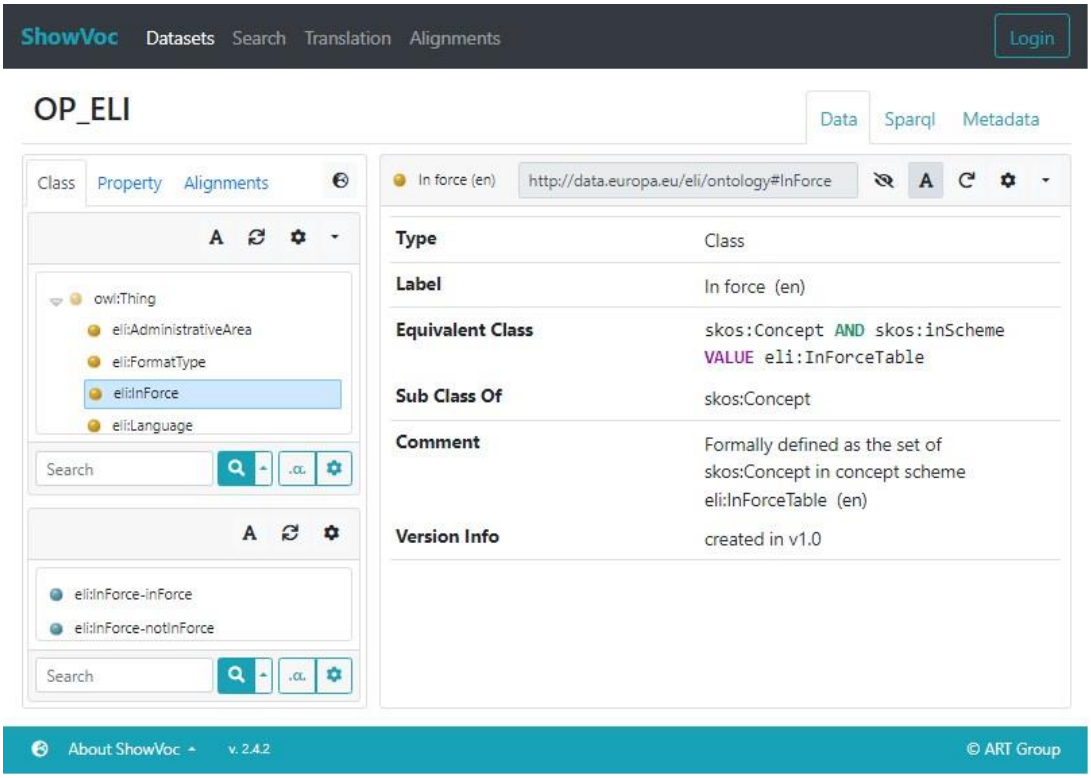

Figure 1: ShowVoc data view displaying the ELI (European Legislation Identifier) ontology.

SPARQL querying. ShowVoc provides a SPARQL editor with syntax highlighting and completion that can be used to query individual datasets. It also supports federated queries involving other hosted datasets or remote SPARQL endpoints. The former can be done more efficiently if the chosen triple store is GraphDB, which has specific optimizations for local federation.

Results can be downloaded in a variety of formats, and queries can be loaded from different scopes, e.g., system scope for general purpose queries or dataset-specific queries.

Dataset metadata. Proper metadata is considered critical for publishing a dataset according to the FAIR principles [20]. As such, ShowVoc manages a complete description for each dataset, including general metadata (e.g., title), customizable facets (e.g., category, organization), access metadata (e.g., SPARQL endpoint), structural metadata (e.g., URI space), and various metrics that provide insight into the richness of the available content both at the conceptual level (e.g., number of classes, concepts, etc.) and at the lexical level (i.e., regarding the degree of coverage of different natural languages). These metrics can be visualized as a table or as a chart of various types.

ShowVoc manages the dataset descriptions using the Metadata Registry (MDR) component of Semantic Turkey. This in turn manages the available datasets as a DCAT [21] catalog, using a metadata profile based on a combination and interpretation of existing metadata vocabularies (e.g. DCTERMS, FOAF, VoID [22], LIME [23]) together with a small ontology addressing concerns (mostly related to access) not covered by the former.

Distributions. ShowVoc maintains multiple distributions of each dataset. In the first place, these distributions can be used to provide a downloadable data dump of a (given version) of the dataset. However, they also include any additional files, including documentation.

Global Search. ShowVoc allows users to perform a full-text search across all hosted datasets. Matched entities, grouped by dataset, are displayed with their labels, IRIs, and `skos:note` specializations (which include definitions, examples, etc.).

Translation. Similar to global search, this feature allows users to search lookup a term in a given natural language inside the datasets in the catalog, searching for a translation in one or more natural languages.

Alignments. ShowVoc keeps track of the alignments contained in each dataset, providing both a per-dataset and a global view of these alignments.

The former consists of an expanding tree whose roots are the datasets directly aligned with the current datasets. These can be expanded to show the datasets to which they are aligned, and thus to which the current dataset is indirectly aligned. Each node is decorated by the number of links: when one of them is clicked, ShowVoc displays a paginated list of the correspondences (with some filtering features).

Global visualization of alignments is supported by a similar tree view as well as by graph visualization: nodes represent datasets and edges represent alignments. By clicking on a node or an edge, users can view metadata about a dataset or an alignment.

5. Impact

First released in September 2021, ShowVoc is younger than its editing companion VocBench 3, which has become a reference platform since its launch in September 2017. Despite ShowVoc's relatively short history, we can point to some notable adopters. The Food and Agriculture Organization (FAO) of the United Nations (UN) adopted ShowVoc for the Caliper portal[14], which publishes statistical classifications as linked data. The Italian branch of LifeWatch ERIC – the European Research Infrastructure Consortium for biodiversity and ecology – used ShowVoc in addition to OntoPortal as a data publication platform supporting resolvable URIs. Last but not least, the Publications Office (OP) of the European Union (EU), which managed the development of the system, has deployed an instance of ShowVoc[15] "to support interested teams and professionals working for the EU institutions and agencies". The Publications Office also integrated ShowVoc into the EU Vocabularies Portal[16] to provide an "advanced view" of the datasets content, complementing the portal's own capabilities.

6. Discussion and conclusion

In this work, we have introduced ShowVoc, a web-based multilingual platform for publishing and consulting OWL ontologies, SKOS(-XL) thesauri, Ontolex-lemon lexicons and generic RDF datasets. Its features and impact on the world of linked open datasets have been discussed. Future work includes further broadening the dedicated support for core modeling vocabularies (e.g. XKOS for statistical classifications), improving the publication workflow from VocBench to ShowVoc and further exploiting its linking metadata to broaden its possibilities as a authority-based translation system.

Acknowledgements

ShowVoc has been originally designed by Tor Vergata University of Rome and is now maintained and evolved by Lore Star srl in the context of the Digital Europe Programme, under management of the Publications Office of the EU in a provision contract with European Dynamics.

References

[1] T. Berners-Lee, J. A. Hendler, and O. Lassila, "The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities," *Scientific American*, vol. 284, no. 5, pp. 34-43, 2001, doi: 10.1038/scientificamerican0501-34 .

[2] T. Berners-Lee. (2006) Design Issues. [Online]. Available: https://www.w3.org/DesignIssues/LinkedData.html

[3] A. Stellato, M. Fiorelli, A. Turbati, T. Lorenzetti, W. van Gemert, D. Dechandon, C. Laaboudi-Spoiden, A. Gerencsér, A. Waniart, E. Costetchi, and J. Keizer, "VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons," *Semantic Web*, vol. 11, no. 5, pp. 855-881, Jan 2020, doi: 10.3233/SW-200370 .

[4] C. Lagoze and H. Van de Sompel, "The open archives initiative: building a low-barrier interoperability framework," in *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, Roanoke Virginia USA, June 24-28, 2001*, 2001, pp. 54-62, doi: 10.1145/379437.379449 .

[5] P.-Y. Vandenbussche, G. A. Atemezing, M. Poveda-Villalón, and B. Vatant, "Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web," *Semantic Web*, vol. 8, no. 3, pp. 437-452, December 2016, doi: 10.3233/SW-160213 .

[6] M. Salvadores, P. R. Alexander, M. A. Musen, and N. F. Noy, "BioPortal as a dataset of linked biomedical ontologies and terminologies in

[14] https://www.fao.org/statistics/caliper/en
[15] https://showvoc.op.europa.eu
[16] https://op.europa.eu/en/web/eu-vocabularies

RDF," *Semantic Web*, vol. 4, no. 3, pp. 277-284, 2013, doi: 10.3233/SW-2012-0086 .

[7] C. Jonquet, J. Graybeal, S. Bouazzouni, M. Dorf, N. Fiore, X. Kechagioglou, T. Redmond, I. Rosati, A. Skrenchuk, J. L. Vendetti, M. Musen, and m. o.t.O. Alliance, "Ontology Repositories and Semantic Artefact Catalogues with the OntoPortal Technology," in *The Semantic Web – ISWC 2023 (Lecture Notes in Computer Science)*, T. R. Payne et al., Eds.: Springer, Cham, 2023, vol. 14266, pp. 38-58, doi: 10.1007/978-3-031-47243-5_3 .

[8] C. Jonquet, A. Toulet, E. Arnaud, S. Aubin, E. Dzalé Yeumo, V. Emonet, J. Graybeal, M.-A. Laporte, M. A. Musen, V. Pesce, and P. Larmande, "AgroPortal: A vocabulary and ontology repository for agronomy," *Computers and Electronics in Agriculture*, vol. 144, pp. 126-143, 2018, doi: 10.1016/j.compag.2017.10.012 .

[9] X. Kechagioglou, L. Vaira, P. Tomassino, N. Fiore, A. Basset, and I. Rosati, "EcoPortal: An Environment for FAIR Semantic Resources in the Ecological Domain," in *Proceedings of the Joint Ontology Workshops 2021. Episode VII: The Bolzano Summer of Knowledge. co-located with the 12th International Conference on Formal Ontology in Information Systems, and the 12th International Conference on Biomedical Ontologies*, vol. 2969, 2021. [Online]. Available: https://ceur-ws.org/Vol-2969/paper6-s4biodiv.pdf

[10] D. Bailo, R. Paciello, J. Michalek, M. Cocco, C. Freda, K. Jeffery, and K. Atakan, "The EPOS multi-disciplinary Data Portal for integrated access to solid Earth science datasets," *Scientific Data*, vol. 10, 2023, doi: 10.1038/s41597-023-02697-9 .

[11] J. David, J. Euzenat, F. Scharffe, and C. Trojahn dos Santos, "The Alignment API 4.0," *Semantic Web Journal*, vol. 2, no. 1, pp. 3-10, 2011.

[12] J. P. McCrae, C. Tiberius, A. F. Khan, I. Kernerman, T. Declerck, S. Krek, M. Monachini, and S. Ahmadi, "The ELEXIS interface for interoperable lexical resources," in *Proceedings of the sixth biennial conference on electronic lexicography (eLex). eLex 2019*, 2019.

[13] O. Erling, "Virtuoso, a Hybrid RDBMS/Graph Column Store," *Data Engineering Bulletin*, vol. 35, no. 1, pp. 3-8, 2012.

[14] A. Gonzales-Aguilar, M. Ramírez-Posada, and D. Ferreyra, "TemaTres: software para gestionar tesauros," *El profesional de la información*, vol. 21, no. 3, pp. 319-325, 2012, doi: 10.3145/epi.2012.may.14 .

[15] P. Frischmuth, M. Martin, S. Tramp, T. Riechert, and S. Auer, "OntoWiki – An authoring, publication and visualization interface for the Data Web," *Semantic Web*, vol. 6, no. 3, pp. 215-240, 2015, doi: 10.3233/SW-140145 .

[16] S. Peroni, D. Shotton, and F. Vitali, "Making Ontology Documentation with LODE," in *Proceedings of the I-SEMANTICS 2012 Posters & Demonstrations Track, Graz, Austria, September 5-7, 2012*, 2012, pp. 63-67. [Online]. Available: https://ceur-ws.org/Vol-932/paper12.pdf

[17] D. Garijo, "WIDOCO: A Wizard for Documenting Ontologies," in *The Semantic Web – ISWC 2017. ISWC 2017 (Lecture Notes in Computer Science)*, vol. 10588, 2017, pp. 94-102, doi: https://doi.org/10.1007/978-3-319-68204-4_9 .

[18] M. T. Pazienza, N. Scarpato, A. Stellato, and A. Turbati, "Semantic Turkey: A Browser-Integrated Environment for Knowledge Acquisition and Management," *Semantic Web Journal*, vol. 3, no. 3, pp. 279-292, 2012, doi: 10.3233/SW-2011-0033 .

[19] M. Fiorelli, A. Stellato, I. Rosati, and N. Fiore, "Process-Level Integration for Linked Open Data Development Workflows: A Case Study," in *Metadata and Semantic Research (Communications in Computer and Information Science)*, E. Garoufallou and A. Vlachidis, Eds.: Springer, Cham, 2023, vol. 1789, pp. 148-159, doi: 10.1007/978-3-031-39141-5_13 .

[20] M. D. Wilkinson, et al., "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 160018, 2016, doi: 10.1038/sdata.2016.18 .

[21] R. Albertoni, D. Browning, S. Cox, A. N. Gonzalez-Beltran, A. Perego, and P. Winstanley, "The W3C Data Catalog Vocabulary, Version 2: Rationale, Design Principles, and Uptake," *Data Intelligence*, pp. 1-37, December 2023, doi: 10.1162/dint_a_00241 .

[22] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. (2011, March) World Wide Web Consortium (W3C). [Online]. Available: http://www.w3.org/TR/void/

[23] M. Fiorelli, A. Stellato, J. P. Mccrae, P. Cimiano, and M. T. Pazienza, "LIME: the Metadata Module for OntoLex," in *The Semantic Web. Latest Advances and New Domains (Lecture Notes in Computer Science)*, F. Gandon et al., Eds.: Springer International Publishing, 2015, vol. 9088, pp. 321-336, doi: 10.1007/978-3-319-18818-8_20 .