The Systematic Dataset Review Process: from Linked Dataset Discoverability to their FAIRness

Sana Latif[1], Maria Angela Pellegrino[1,*]

[1]Università degli Studi di Salerno, via Giovanni Paolo II, 132, 84084 Fisciano (SA), Italy

Abstract

This work presents the `Systematic Dataset Review` process, a systematic, cyclic, and FAIR-aligned process for dataset discovery, curation, and quality assessment, inspired by Systematic Literature Review methodologies. The process consists of three phases: (i) a (multivocal) literature review to identify and document references about datasets; (ii) a Linked Dataset Discoverability phase to curate, filter, and publish datasets as domain-specific Linked Open Data sub-clouds; and (iii) a Quality Assessment phase that enables FAIRness evaluations through automated tools. The framework is adaptable across different use cases: monitoring the quality of existing sub-clouds, enriching current sub-clouds with new datasets, or constructing entirely new thematic sub-clouds. By ensuring transparency, reproducibility, and ongoing quality monitoring, this approach supports the creation of accessible and sustainable dataset ecosystems.

Keywords

Process, FAIRness, Dataset discoverability, Literature Review, Quality assessment, Reproducibility

1. Introduction

A systematic literature review (SLR) is a structured and methodical approach to identifying, evaluating, and synthesizing all relevant research on a particular topic or research question, with the aim of minimizing bias and ensuring reproducibility. SLRs follow a well-defined protocol to ensure comprehensiveness, traceability, and replicability of the process. This typically involves defining explicit inclusion and exclusion criteria, conducting a systematic search across multiple databases, applying quality assessment techniques, and organizing the findings through qualitative or quantitative synthesis.

Over the past two decades, numerous guidelines have been proposed to support researchers in conducting SLRs effectively, ranging from guidelines for conducting literature mappings [1, 2] to scoping reviews [3, 4], for snowballing [5] and a wide range of different checklist for rigorously reviewing literature [6, 7, 8, 9, 10, 11]. Prominent among these are the guidelines introduced by Kitchenham and Charters [12], which formalize the planning, conducting, and reporting stages of an SLR.

In response to the growing complexity and interdisciplinarity of research, the notion of multivocal literature reviews (MLRs) has also emerged. MLRs extend traditional SLRs by incorporating not only peer-reviewed academic literature but also the so called gray literature sources such as industry reports, blog posts, white papers, and technical documentation [13]. This is especially valuable in applied fields like computer science and engineering, where practice-oriented insights may precede academic publication or remain outside conventional scholarly venues. Incorporating multivocal sources helps capture a more holistic view of the topic, reflecting both theoretical advances and practical implementations.

While SLR guidelines are originally designed for synthesizing evidence from scholarly publications, this poster proposes to adapt them to dataset discovery and assessment. It is not a straightforward process as this adaptation poses several challenges that necessitate methodological adaptation. Unlike academic papers, datasets are often published in non-traditional sources like data portals, GitHub, or Zenodo, lacking standardized metadata, persistent identifiers, or peer review validation. This requires rethinking the search strategy, often shifting from bibliographic queries to web crawling or

ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan

*Corresponding author.

slatif@unisa.it (S. Latif); mapellegrino@unisa.it (M. A. Pellegrino)

0000-0001-8927-5833 (M. A. Pellegrino)

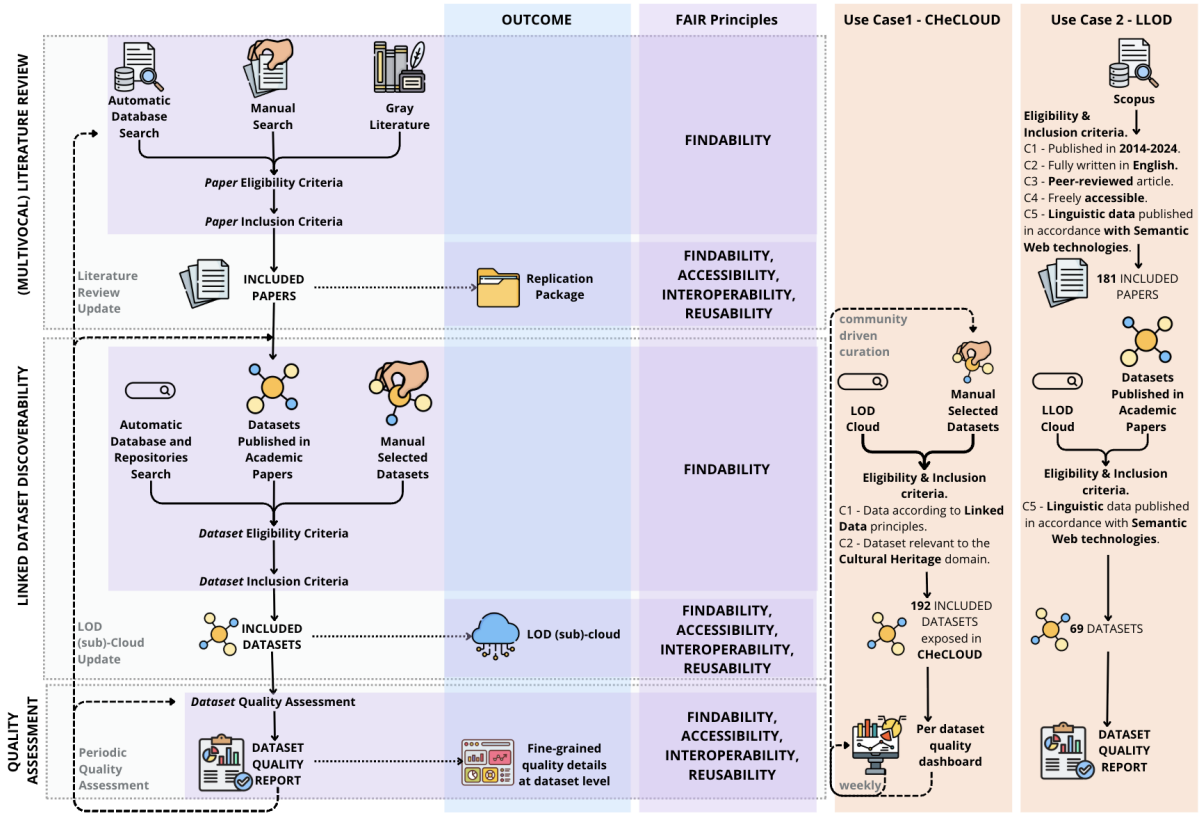

Figure 1: Systematic Dataset Review process. Overview of a three-phase process integrating multivocal literature review, linked dataset discovery, and quality assessment, aligned with FAIR principles to support transparent, reusable, and high-quality dataset curation. The high-level process is instantiated in two use cases.

API-based retrieval. The unit of analysis also differs: while SLRs assess research rigour and validity, dataset assessment must consider technical and semantic properties such as licensing, provenance, format (e.g., RDF), and accessibility, often operationalized through frameworks like FAIR [14] or Linked Data quality dimensions [15]. Moreover, datasets are dynamic—they can be updated, versioned, or deleted—so temporal aspects and version tracking become essential in defining inclusion criteria and in maintaining reproducibility. Quality assessment in this context must go beyond narrative synthesis, incorporating metrics-based evaluations (e.g., completeness, availability, interlinking) using tools like FAIR-Checker [16] or KGHeartBeat [17]. Finally, synthesis methods may rely less on narrative summaries and more on dashboard visualizations, semantic mappings, or clustering of metadata. These differences highlight the need for a tailored methodology that retains the systematic rigour of SLRs while embracing the unique demands of dataset discovery and evaluation in the Semantic Web.

This poster introduces the `Systematic Dataset Review` (SDR) process. To the best of our knowledge, this is the first methodology that proposes guidelines to systematic identify datasets. While the process is detailed in Section 2, Section 3 illustrates its practical relevance through concrete use cases. The article concludes with final thoughts, limitations, and future directions.

2. `Systematic Dataset Review` Process

This section outlines the SDR process, a rigorous and *Systematic* three-phase process for conducting *Dataset Reviews*, aimed at improving dataset discoverability and quality assessment. The approach draws on established SLR guidelines and is aligned with the FAIR principles—Findability, Accessibility, Interoperability, and Reusability. This proposal builds upon a comprehensive analysis of established SLR guidelines, concrete experience in curating domain-specific collections of datasets (as detailed in Section 3), and consolidated expertise in evaluating Linked Data quality through FAIR-aligned

frameworks and tools. The overall process is illustrated in Figure 1 which graphically reports phases, each phase output, and the alignment with FAIR principles.

The *first phase*, titled **(Multivocal) Literature Review**, follows conventional SLR practices and focuses on the identification and selection of relevant scholarly contributions. Literature is retrieved via automated database searches, manual searches, and exploration of gray literature, and then filtered based on eligibility and inclusion criteria. The resulting set of selected papers forms the knowledge base for further investigation. To enhance transparency and reproducibility, the outcome of this phase should be published as a **replication package**, which encapsulates the review protocol, selection rationale, and data extracted from included studies. By eating-our-own-food, the replication package might be structured via ontologies, as SLRONT [18]. While the SLR process inherently supports findability here interpreted as discoverability of relevant resources, the replication package significantly broadens the impact by supporting all FAIR principles: it ensures the accessibility of review materials, interoperability with related studies via shared standards and formats, and reusability for follow-up analysis or meta-reviews. Moreover, when replication packages are published in searchable repositories and assigned a persistent identifier—such as those provided by Zenodo—findability is also ensured. Besides enabling transparency, replication packages enable reproducibility. It is important to note that at this stage, the process is format-agnostic and focuses primarily on scholarly literature rather than datasets.

The *second phase*, called **Linked Dataset Discoverability**, adapts SLR principles to dataset identification. This involves discovering datasets mentioned in academic publications or indexed in online repositories, using both automated and manual approaches. By applying dataset-specific eligibility and inclusion criteria, a curated list of relevant datasets is compiled. These datasets can then be published as part of a thematic **LOD (Linked Open Data) sub-cloud**, modelled after well-known examples like the LOD Cloud[1]. This phase further reinforces findability of linked datasets and the LOD sub-cloud as an output facilitates the implementation of the other FAIR principles by encouraging standard access points (e.g., the LOD Cloud as a standard mechanism to visualize inter-linked datasets), proper licensing, and reuse-oriented metadata. Indexing linked datasets in a unified search engine enhances findability; adopting standardized publication mechanisms, such as linked data clouds, supports accessibility; following linked data principles facilitates interoperability through interlinking; and defining appropriate licensing terms enables reusability. Unlike the first phase, this stage is data format dependent, requiring adjustments based on the data structures (e.g., RDF, JSON-LD) and the specific characteristics of the community or domain. Repository selection, filtering criteria, and publication strategies must be tailored accordingly to ensure compatibility and engagement.

The *third phase*, **Quality Assessment**, evaluates the sub-cloud datasets using periodic, metric-based evaluations. These assessments are grounded in quality dimensions commonly addressed in SLRs (e.g., accuracy, completeness, timeliness) and can be mapped to the FAIR principles. The outcome is a **Dataset Quality Report** offering both aggregated and fine-grained metrics. This phase supports long-term monitoring of dataset quality, ensuring that the curated datasets remain aligned with FAIR principles over time. Tools as FAIR-Checker [16], SPARQLES [19], and KGHeartBeat [17] can support this stage, enabling automated FAIRness scoring and deeper quality diagnostics. This aligns with recent efforts to transit from traditional linked data quality assessments to FAIR-oriented evaluations [20].

Importantly, the entire framework is designed to be iterative and extensible. The cyclical nature of the process accommodates regular updates to dataset quality scores, the inclusion of newly published datasets, and revisions to the literature review itself. This dynamic workflow ensures that the knowledge ecosystem remains current, reflective of emerging standards, and responsive to the evolving landscape of data publication and reuse. To alleviate the manual effort involved in the eligibility and inclusion phases, recent literature has highlighted the potential of automating these steps through machine learning and language model-based approaches [21, 22, 23, 24, 25, 26, 20]. These methods aim to streamline the review process while preserving human oversight through a human-in-the-loop approach, where experts retain final control to ensure relevance and contextual accuracy. This hybrid model balances automation with the nuanced judgment essential to scholarly curation.

[1]LOD Cloud: https://lod-cloud.net

3. Use Cases

This process can be instantiated across a variety of scenarios to support dataset discoverability, enrichment, and quality monitoring within the LOD ecosystem. Three main use cases can be distinguished:

- **Quality Assessment of an Existing Sub-Cloud.** In cases where a sub-cloud already exists, the process can begin from the third phase. The curated list of datasets serves as a stable foundation, and the third phase (Quality Assessment) can be directly integrated to enable ongoing, periodic FAIRness evaluations. This supports long-term monitoring and maintenance of the sub-cloud's quality, ensuring its continued alignment with community standards. As a driving example in this direction, we report the **Linguistic LOD use case** [27]. Besides relying on the Linguistic LOD Cloud [28] as a main source, the extraction of linguistic LOD use case followed a systematic literature review approach inspired by Kitchenham's [12] guidelines and structured in planning, conducting, and reporting phases. In the planning stage, the objectives and research questions were defined, focusing on publishing trends of linguistic datasets modeled with Semantic Web technologies, the search engines most often employed, and evidence of dataset reuse, with the scope restricted to peer-reviewed works published in English between 2014 and 2024. The conducting phase relied on Scopus as the sole source, chosen for its broad coverage, where a tailored search string was applied, explicitly including terms such as "`knowledge graph`", "`linked data`", and "`linguistic`", while excluding acronyms and "`ontology`" to privilege materialized datasets over conceptual models. This search retrieved $1,788$ records, which were screened by two independent reviewers following PRISMA guidelines: first by title and abstract, then by full text, with disagreements resolved through weekly discussions and, when necessary, collaborative arbitration. The multi-stage screening resulted in 181 primary studies, of which 92 made use of LLOD resources and 89 defined 69 distinct linguistic linked datasets. In the reporting phase, these datasets were analyzed in relation to the FAIR principles—findability, accessibility, interoperability, and reusability—with those already indexed in the LLOD Cloud monitored automatically on a weekly basis via KGHeartBeat [17], while the others were manually checked through their accompanying publications and dataset websites, ensuring a consistent and comprehensive assessment of linguistic linked data practices.
- **Enrichment of an Existing Sub-Cloud.** The process also supports the expansion of existing sub-clouds. For example, the Life Sciences LOD Cloud [29] can be enriched with additional biomedical datasets identified through automated tools or manual review. After extending the dataset list (phase two), the newly enriched sub-cloud can undergo quality assessment (phase three). This ensures the sub-cloud evolves with the growing needs and contributions of the community while maintaining FAIR compliance.
- **Creation of a New Thematic Sub-Cloud.** The full three-phase workflow is particularly useful for constructing a novel sub-cloud—such as the Cultural Heritage LOD Cloud (CHeCLOUD[2]). Based on the process described in [20], the construction of the cloud began with the LD discoverability phase, using the December 2024 LOD Cloud snapshot as the primary source. From an initial pool of $1,658$ datasets, two independent annotators manually examined dataset titles, descriptions, and keywords to identify datasets both compliant with LD principles (C1) and relevant to the Cultural Heritage domain (C2). Conflicts (105 cases, approximately 6%) were resolved through discussion, with a third reviewer acting as arbiter. In addition, 49 further datasets were manually collected from complementary sources, including GitHub and Zenodo searches, expert recommendations, and relevant literature. A final inclusion screening was then performed, involving detailed reviews of metadata, dataset landing pages, and available data. This step generated additional discussions on 19 cases (around 10% of the eligible pool), ultimately leading to the inclusion of 192 datasets in CHeCLOUD. Each dataset in the index is accompanied by both coarse-grained and fine-grained quality assessment reports generated automatically via KGHeartBeat [17], as demonstrated in Tuozzo et al. [30]. While coarse-grain visualization

[2]CHeCLOUD: http://isislab.it:12280/CHe-cloud

enables datasets comparison, the fine-grain one provides punctual access to individual dataset quality scores over time. These scores are complemented by an automatically verbalized quality summary, currently produced through Gemini 2.5 Pro, though the framework is designed to support interchangeable LLMs. A preliminary evaluation confirms the accuracy and usefulness of these generated summaries, while a more structured comparative study across alternative LLMs remains a necessary next step. While the discoverability and metadata curation phases still require human oversight, the quality assessment pipeline is fully automated and executed on a weekly basis, ensuring the continuous monitoring of dataset FAIRness within the Cultural Heritage domain.

4. Conclusion, Limitations, and Future Direction

The SDR process represents a high-level proposal for enabling reproducible, high-quality, and format-agnostic dataset discovery and quality assessment, while remaining adaptable to the specific needs of individual research communities. Although a universal, one-size-fits-all toolkit is difficult to achieve, the provision of clear guidelines can ensure transparency, comparability, and reproducibility across domains. To illustrate how the process operates in practice, this poster presents two domain-specific pilots: CHeCLOUD, targeting Cultural Heritage, and the Linguistic LOD sub-cloud. Both pilots have demonstrated the feasibility of enriching existing dataset indexes while providing periodic, fully automated FAIRness assessments. Their partial outcomes, such as curated dataset lists, quality trends, and FAIR scores, are openly shared through public repositories, thereby strengthening transparency, accountability, and open science practices.

A defining feature of the SDR process is its community-driven nature: domain experts and dataset curators are actively invited to refine inclusion criteria, validate the relevance of candidate datasets, and update metadata through collaborative platforms such as GitHub and Zenodo. Replication packages and interactive dashboards further support external validation, creating open feedback loops that allow the process to evolve in tandem with community needs. To enable customization of the visualizations offered in CHeCLOUD, we publicly released the code for generating thematic LOD sub-clouds and rendering dataset-related quality scores (https://github.com/isislab-unisa/Systematic-Dataset-Review).

Nonetheless, several challenges remain. Persistent issues include heterogeneous and incomplete metadata, the absence of persistent identifiers, and the need to tailor discovery and inclusion strategies across diverse data formats [31, 32]. The proposed use cases heavily rely on the LOD Cloud, while its obsolescence and quality issues are well known in the literature [31, 32]. Moreover, the exponential growth in both the volume and heterogeneity of available data underscores the urgency of robust automated mechanisms for dataset discovery. While human oversight will remain indispensable for ensuring contextual accuracy and final quality, sustainable and scalable dataset discoverability can only be achieved by placing automation at the backbone of the process, complemented by community expertise to ensure trust and relevance. These challenges highlight key avenues for future improvement, including the development of more advanced automation pipelines, metadata enrichment techniques, and reliable linking mechanisms to guarantee long-term usability and alignment with FAIR principles.

Acknowledgment

We thank Gabriele Tuozzo for publicly releasing the sub-cloud generation code and for valuable discussions that informed this work.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT in order to: Grammar and spelling check.

References

[1] K. Petersen, S. Vakkalanka, L. Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering: An update, Information and software technology 64 (2015) 1–18. doi:10.1016/j.infsof.2015.03.007.

[2] B. Barn, S. Barat, T. Clark, Conducting systematic literature reviews and systematic mapping studies, in: Proceedings of the 10th innovations in software engineering conference, 2017, pp. 212–213. doi:10.1145/3021460.3021489.

[3] A. C. Tricco, E. Lillie, W. Zarin, K. K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M. D. Peters, T. Horsley, L. Weeks, et al., PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation, Annals of internal medicine 169 (2018) 467–473. doi:10.7326/M18-0850.

[4] H. Arksey, L. O'malley, Scoping studies: towards a methodological framework, International journal of social research methodology 8 (2005) 19–32. doi:10.1080/1364557032000119616.

[5] C. Wohlin, Guidelines for snowballing in systematic literature studies and a replication in software engineering, in: Proceedings of the 18th international conference on evaluation and assessment in software engineering, 2014, pp. 1–10. doi:10.1145/2601248.2601268.

[6] D. Budgen, P. Brereton, Performing systematic literature reviews in software engineering, in: Proceedings of the 28th international conference on Software engineering, 2006, pp. 1051–1052. doi:10.1145/1134285.1134500.

[7] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, et al., The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, British Medical Journal (BMJ) 372 (2021). doi:10.1136/bmj.n71.

[8] J. F. Wolfswinkel, E. Furtmueller, C. P. Wilderom, Using grounded theory as a method for rigorously reviewing literature, European journal of information systems 22 (2013) 45–55. doi:10.1057/ejis.2011.51.

[9] Y. Xiao, M. Watson, Guidance on conducting a systematic literature review, Journal of planning education and research 39 (2019) 93–112. doi:10.1177/0739456X17723971.

[10] A. Booth, M. Martyn-St James, M. Clowes, A. Sutton, Systematic approaches to a successful literature review, SAGE Publications Ltd, 2021. URL: https://digital.casalini.it/9781529759648.

[11] J. v. Brocke, A. Simons, B. Niehaves, B. Niehaves, K. Riemer, R. Plattfaut, A. Cleven, Reconstructing the giant: On the importance of rigour in documenting the literature search process, European Conference on Information Systems (2009). URL: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1145&context=ecis2009.

[12] S. Keele, et al., Guidelines for performing systematic literature reviews in software engineering, Technical Report, Technical report, ver. 2.3 ebse, 2007. URL: https://shorturl.at/Z19ct.

[13] V. Garousi, M. Felderer, M. V. Mäntylä, Guidelines for including grey literature and conducting multivocal literature reviews in software engineering, Information and software technology 106 (2019) 101–121. doi:10.1016/j.infsof.2018.09.006.

[14] A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, M. Courtot, M. Crosas, M. Dumontier, C. T. Evelo, et al., FAIR principles: interpretations and implementation considerations, 2020. doi:10.1162/dint_r_00024.

[15] T. Berners-Lee, 5-star deployment scheme, 2012. URL: https://5stardata.info/, [Online] Last access 2025, July.

[16] A. Gaignard, T. Rosnet, F. De Lamotte, V. Lefort, M.-D. Devignes, FAIR-Checker: supporting digital resource findability and reuse with knowledge graphs and semantic web standards, Journal of Biomedical Semantics 14 (2023) 7. doi:10.1186/s13326-023-00289-5.

[17] M. A. Pellegrino, A. Rula, G. Tuozzo, KGHeartBeat: An open source tool for periodically evaluating the quality of knowledge graphs, in: International Semantic Web Conference, Springer, 2024, pp. 40–58. doi:10.1007/978-3-031-77847-6_3.

[18] Y. Sun, Y. Yang, H. Zhang, W. Zhang, Q. Wang, Towards evidence-based ontology for supporting systematic literature review, in: 16th International Conference on Evaluation Assessment in Software Engineering (EASE), 2012, pp. 171–175. doi:10.1049/ic.2012.0022.

[19] P.-Y. Vandenbussche, J. Umbrich, L. Matteis, A. Hogan, C. Buil-Aranda, SPARQLES: Monitoring public SPARQL endpoints, Semantic web 8 (2017) 1049–1065. doi:10.3233/SW-170254.

[20] A. Lieto, M. A. Pellegrino, G. Tuozzo, The FAIRness of CHeCLOUD, the cultural heritege linked open data cloud, Semantic web (2025). URL: https://www.semantic-web-journal.net/content/fairness-checloud-cultural-heritage-linked-open-data-cloud, [Under review].

[21] K. R. Felizardo, J. C. Carver, Automating systematic literature review, Contemporary empirical methods in software engineering (2020) 327–355. doi:10.1007/s00607-023-01181-x.

[22] K. R. Felizardo, M. S. Lima, A. Deizepe, T. U. Conte, I. Steinmacher, ChatGPT application in systematic literature reviews in software engineering: an evaluation of its accuracy to support the selection activity, in: Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2024, pp. 25–36. doi:10.1145/3674805.3686666.

[23] J. de la Torre-López, A. Ramírez, J. R. Romero, Artificial intelligence to automate the systematic review of scientific literature, Computing 105 (2023) 2171–2194. doi:10.1007/s00607-023-01181-x.

[24] F. Osborne, H. Muccini, P. Lago, E. Motta, Reducing the effort for systematic reviews in software engineering, Data Science 2 (2019) 311–340. doi:10.3233/DS-190019.

[25] F. Octaviano, K. R. Felizardo, S. C. Fabbri, B. M. Napoleão, F. Petrillo, S. Hallé, SCAS-AI: a strategy to semi-automate the initial selection task in systematic literature reviews, in: 2022 48th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), IEEE, 2022, pp. 483–490. doi:10.1109/SEAA56994.2022.00080.

[26] S. Götz, Supporting systematic literature reviews in computer science: The systematic literature review toolkit, in: Proceedings of the 21st ACM/IEEE International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings, 2018, pp. 22–26. doi:10.1145/3270112.3270117.

[27] P. Esposito, M. A. Pellegrino, G. Tuozzo, FAIRness of the Linguistic Linked Open Data Cloud: an Empirical Investigation, Special Issue on Data quality dimensions in Data FAIRification design and processes in the Journal of Data and Information Quality (JDIQ) (2025).

[28] C. Chiarcos, S. Hellmann, S. Nordhoff, Linking Linguistic Resources: Examples from the Open Linguistics Working Group, 2012, pp. 201–216. doi:10.1007/978-3-642-28249-2_19.

[29] A. Hasnain, S. Sana E Zainab, M. Kamdar, Q. Mehmood, C. Warren, Jr, Q. Fatimah, H. Deus, M. Mehdi, S. Decker, A roadmap for navigating the life sciences linked open data cloud, 2014. doi:10.1007/978-3-319-15615-6_8.

[30] G. Tuozzo, M. A. Pellegrino, A. Lieto, CHeCLOUD—the cultural heritage linked open data cloud, in: International Semantic Web Conference ISWC - Poster&Demo, 2025.

[31] J. Debattista, C. Lange, S. Auer, D. Cortis, Evaluating the quality of the LOD cloud: An empirical investigation, Semantic Web 9 (2018) 859–901. doi:10.3233/SW-180306.

[32] T. Gabriele, Navigating the LOD Subclouds: Assessing Linked Open Data Quality by Domain, in: Companion Proceedings of the Web Conference, Association for Computing Machinery, New York, NY, USA, 2025. doi:10.1145/3701716.3717569.