A Systematic Literature Review on the Representation of Texts as Linguistic Linked Open Data

Michela Bandini

Cnr-Istituto di Linguistica Computazionale "A. Zampolli"
michela.bandini@ilc.cnr.it

Valeria Quochi

Cnr-Istituto di Linguistica Computazionale "A. Zampolli"
valeria.quochi@ilc.cnr.it

Abstract

Despite the growing interest in publishing linguistic data as Linked Open Data (LOD), the representation of ancient language corpora within the Semantic Web remains challenging. While LOD principles have been successfully applied to linguistic resources such as dictionaries, lexicons, and terminologies, their use for textual corpora — particularly those related to ancient languages — is still limited. Through a systematic literature review, we investigate how textual data has been represented as Linguistic Linked Open Data (LLOD), evaluating the potential and limitations of existing approaches and methodologies for enhancing data integration and interoperability in the Digital Humanities. This systematic literature review follows a rigorous methodology encompassing literature identification, screening for inclusion, and quality assessment. By classifying and analysing relevant studies, we provide a comprehensive overview of current practices and offer insights into their benefits and limitations.

Keywords: Linguistic Linked Open Data, Semantic Web, Systematic literature review, Ancient texts, DigitAnt, Ancient languages

Nonostante il crescente interesse per la pubblicazione di dati linguistici come Linked Open Data (LOD), la rappresentazione di corpora di lingue antiche all'interno del web semantico rimane una sfida. Mentre i principi LOD sono stati applicati con successo a risorse linguistiche come dizionari, lessici e terminologie, il loro utilizzo per i corpora testuali — in particolare quelli relativi alle lingue antiche — solleva problemi di modellazione e granularità dei dati. Attraverso una revisione sistematica della letteratura, analizziamo come i dati testuali possano essere rappresentati come Linguistic Linked Open Data (LLOD), valutando il potenziale e i limiti degli approcci e delle metodologie esistenti per migliorare l'integrazione e l'interoperabilità dei dati nelle Digital Humanities. Questa revisione sistematica della letteratura segue una metodologia rigorosa che comprende l'identificazione della letteratura, lo screening per l'inclusione e la valutazione della qualità. Classificando e analizzando studi rilevanti, forniamo una panoramica completa delle pratiche attuali e offriamo spunti di riflessione sui vantaggi e le sfide della pubblicazione di corpora antichi come LLOD.

Parole chiave: Dati aperti collegati, Rassegna sistematica della letteratura, Web Semantico, DigItAnt, Testi antichi, Lingue antiche

1. Introduction

The Semantic Web, with its focus on data representation and publication —particularly in the form of Linked Open Data (LOD) —is receiving growing attention across various fields, including computational linguistics and digital humanities (cf. [21], [15]). This attention is not limited to the representation of cataloguing metadata and conceptual-semantic knowledge, e.g., vocabularies, taxonomies and ontologies, but also extends to other textual data including primary and secondary data sources —such as artefacts containing texts and images, possibly with annotations.

The Semantic Web and Linked Open Data are indeed said to provide a significant advantage of enabling the creation of ecosystems of linked data, which enhance the availability and interoperability of data resources, allowing them to be federated and distributed across the Internet [23]. This, in turn, increases their usefulness both for research and development and for culture sharing and preservation. However, a prerequisite for the uptake and success of these technologies is the availability of high-quality linguistic resources represented and published as Linked Open Data.

Over the past decade, the linguistic community has primarily focused on defining and spreading Linked Data representational models and formats for lexical-semantic and conceptual resources (i.e., dictionaries, lexica, thesauri, terminologies, controlled vocabularies, ontologies), and significant efforts have gone into creating and sharing such resources, for instance, [9], [24], [6]. In the digital humanities, Linked Open Data has gained prominence for its role in standardising and representing different kinds of (semantic) data including descriptive, historical, and contextual metadata. This aligns with recent practices in digital archaeology, archives, and libraries, where LOD facilitates the contextualization and dissemination of tangible and intangible cultural heritage, enabling broader accessibility, as well as cross-disciplinary and cross-border collaboration ([37], [7]). In contrast, fewer initiatives have focused on modelling the representation of texts, corpora, and digital scholarly editions within the Semantic Web, despite their formal eligibility for inclusion in the LOD cloud [12]. Most works in this area have focused on creating datasets that represent, annotate, tag, and enrich humanities-relevant entities and knowledge at a holistic level —as knowledge objects— without addressing the semantic representation of their (textual) content at a finer-grained level. Scholarly editions of textual materials (such as manuscripts, inscriptions, papyri, or printed books) and linguistically annotated corpora are typically represented and published using other data models and formats, notably TEI-based XML documents for the former and CoNLL(-like) tabular formats for the latter. Recently, however, there has been a growing interest in representing texts as Linked Open Data, particularly within historical digital humanities projects, with the *LiLa Knowledge Base* serving as a notable example [22].

We undertook this review recognizing a lack of prior studies offering a similarly detailed focus on the representation of textual data as LOD. The few comparable efforts, such as [21] in particular, offer useful surveys of vocabularies and models within the Linguistic LOD framework, but tend to focus on lexical resources, metadata structures, or technological platforms, devoting limited attention to the representation of corpora and textual data.

Our aim, by contrast, is to map emerging practices and tools for the fine-grained semantic representation of textual data — specifically, primary text documents relevant to historical or philological research — as Linked Open Data.

Therefore, although initially motivated by the needs of a specific research project archaic languages and cultures ([49] and [50]), the scope of this review extends beyond that context. It seeks to offer a broader overview of available models and formats for publishing and interlinking textual data in the LOD ecosystem. While standards such as OntoLex-Lemon are relatively mature for lexical resources, the semantic modelling of full-text corpora — especially those documenting ancient or under-resourced languages — remains largely underexplored. Through this review, we thus aim to uncover current practices that can be taken as future best practices for how texts can effectively be represented and made interoperable as Linked Open Data.

The paper is structured as follows: the *'Reviewing Methodology'* section outlines the protocol adopted for conducting the systematic literature review. It proceeds with *'Literature Identification Criteria'*, where the research questions, inclusion/exclusion criteria, and search strategy are defined. The next section, *'Screening for Inclusion'*, details the step-by-step process used to identify and retain relevant works. This is followed by *'Literature Analysis and Classification*', which presents the analysis of the selected studies, focusing on two dimensions: the granularity of data representation and the models and formats used for representing textual data as Linked Open Data. The section '*Discussion'* reflects on the main findings, while '*Limitations*' acknowledges the scope and shortcomings of the study. The final section, '*Conclusions'*, summarizes the key insights, highlights emerging practices, and outlines possible directions for future research and updates to the review.

2. Reviewing Methodology

This work follows a qualitative systematic literature review approach to provide state-of-the-art information on the main interesting works that aim to represent or convert and publish text corpora or documents following the LOD principles and make them available as Linked Data on the Semantic Web.

A systematic literature review is a critical and in-depth synthesis of existing research, carried out by systematically searching for relevant available studies in a structured, non-biased and concise manner [5]. This process follows a rigorous and transparent methodological approach to ensure replicability and reliability of data and results. Systematic reviews can be considered an essential element of academic research and a fundamental cornerstone in different disciplines, including digital humanities. As stated by [35], literature reviews can take two main forms: a) a review that serves as a background for an empirical study, frequently used to provide theoretical context, justify methodological decisions, or identify gaps in the literature that the study intends to fill; and b) a stand-alone review, which attempts to make sense of a broader body of existing literature through a structured interpretation and analysis. While background reviews are typically more limited in scope and directly related to the specific objectives of an empirical study, stand-alone reviews aim to comprehensively summarise knowledge within a given field, often setting the stage for new theoretical insights or further investigation. Generally speaking, their main objective is to explore the scope of existing knowledge, identify possible gaps still to be investigated and\or answer targeted research questions [38]. In addition to these purposes, a systematic review allows the validity and quality of existing studies to be evaluated by applying objective, pre-defined criteria. The present work can be classified as a descriptive stand-alone review, which examines the state of the literature on a specific research topic providing a detailed account of its status at the time of the review (cf. [25] and [35] on the definition of "stand-alone" and "Describe" type reviews)[1].

With this systematic literature review, we set out to exhaustively search for all the available works and associated models or formats used by the digital humanities community for the representation of textual data as LOD and expect to gather information on existing approaches, models, and format, and insights on the most common or suitable approach(es), especially in the domain of historical digital humanities.

In detail, our methodology follows a structured process to identify, filter and analyse studies, adopting predefined criteria for classifying works, as we will elaborate on below. The results of the review are based on a synthesis of the different methodologies and models for representing corpora as LOD that allows the classification of studies into subgroups based on specific models (e.g., NIF, CoNLL-RDF, POWLA) and the level of granularity of the data representation itself (e.g., document level, sentence level, word level). Furthermore, the review includes the statistics of the selection process (e.g. the number of articles excluded at the various stages) integrating quantitative indications that enhance transparency and replicability.

Our approach and its steps are inspired by the framework and protocols developed by [25] and [35], which propose concise steps to develop and implement a systematic review. One of the key points of both approaches, which are very similar and almost complementary, is not only the essential focus on reproducibility and transparency of the research but also the flexibility and adaptability of the process and methodology: the protocol (and the steps themselves) can be adapted to different topics and purposes, serving as inspiration for creating *ad hoc* systematic review protocols.

This systematic review consists thus of three main phases; each divided into sub-steps. The entire process, along with the methodology adopted for each phase, will be explained in greater detail in the following sections.

2.1 Reviewing Protocol

The defined review protocol can be briefly summarized as follows:

1. **Literature identification Criteria**: definition of the terms and questions of the review, including search, selection and classification criteria;

2. **Screening for Inclusion**: search within selected digital libraries for articles describing relevant works;

[1] This systematic review was conducted over a period of 6 months starting from March 2023; consequently, it takes into account publications that appeared before this timeframe. The study and its various phases were carried out mostly in parallel by the two authors.

3. **Literature Analysis**: classification and description of the studies and projects selected according to the specific criteria.

In the first phase, **Literature Identification Criteria**, the search questions are defined, specific keywords are selected, criteria are established, and the online databases or information sources in which the search will be conducted are identified.

The second phase, **Screening for Inclusion**, concerns searching and selecting relevant studies. This phase is a stepwise sequence of screening for inclusion, in which each study is examined to see whether it meets the inclusion and exclusion criteria defined at the beginning of the review. This screening process involves reading first the titles, then abstracts, introductions, and conclusions, followed by a complete scanning of the articles to ensure that only the relevant studies are included in the analysis process. Figure 1 provides an overview of the workflow, or concept map, which visually explains the steps we followed in the process of selecting the works to be reviewed. This concept map outlines the detailed decision-making process we have adopted.

The workflow thus begins with searching articles using keywords or seed authors in the selected information sources (as detailed further below). Titles are screened first to determine relevance; if a title is considered relevant, the article is saved in a dedicated *Zotero* library.[2] Next, the abstract, introduction, and conclusion are analysed to check whether the topic aligns with the scope of this review. Articles deemed relevant proceed to a deeper screening, consisting of a full skimming of the entire article, focusing on the topics most relevant to our research (e.g., checking if it is related to textual data). Only after this step, a full-article reading is performed for those articles that pass the screening. The final step involves thorough analysis and classification of the selected works. Articles that fail any of these steps are skipped at the corresponding stage.

[2] The Text4LOD Zotero library is publicly available here https://www.zotero.org/groups/2552746/itant_project/collections/PQAUQ6YS. A frozen export of the library, reflecting the exact dataset described in this survey, has been deposited on Zenodo to support transparency and reproducibility: https://zenodo.org/doi/10.5281/zenodo.10978178 . The bibliographic entries within the library are categorized into "relevant" and "not-relevant" works and all tags used for classification and analysis have been preserved.

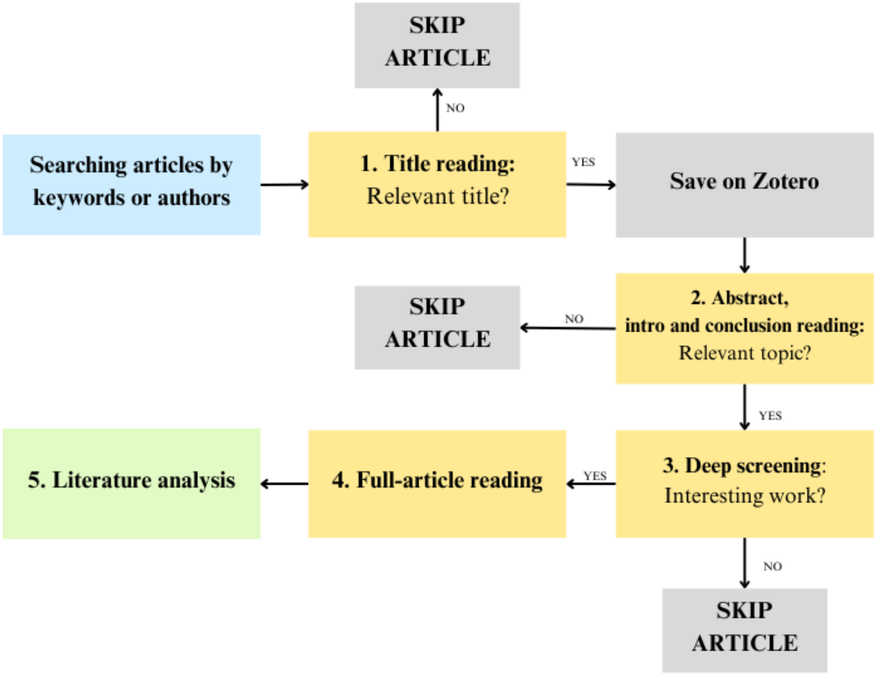

SKIP ARTICLE
NO
Searching articles by keywords or authors
1. Title reading: Relevant title?
YES
Save on Zotero
SKIP ARTICLE
NO
2. Abstract, intro and conclusion reading: Relevant topic?
YES
5. Literature analysis
4. Full-article reading
YES
3. Deep screening: Interesting work?
NO
SKIP ARTICLE

Figure 1: Reviewing workflow

Finally, the **Literature Analysis** phase focuses on analysing the selected articles and their relevance to classify them in terms of granularity and formats or methodologies applied for the Semantic Web representation.

3. Literature Identification Criteria (Phase 1)

3.1 Defining The Research Questions and Goals

The main goal of this review is to identify and analyse works and projects in which texts are represented as LOD. We, therefore, develop this review to understand:

- what are the most relevant projects and initiatives that have already attempted to transform – or represent – text documents for publication on the Semantic Web;
- what are the models and formats already in use, esp. within the digital humanities, for representing texts as LOD;
- what extent, in terms of granularity, reaches the representation and for what purpose (or research sub-community).

In accordance with our main research questions, we define the following overall criteria for the inclusion or exclusion of papers in the final analysis, to guide the review consistently throughout its phases. When in doubt, papers are passed on to the next phase of the reviewing process for further assessment. In general, a paper is excluded if it:

- Addresses only general or theoretical aspects (e.g., XML, RDF);
- Focuses on LLOD models unrelated to text (e.g., OntoLex-Lemon, SKOS);
- Covers marginally relevant topics (e.g., digitisation);
- Targets derivative data formats (e.g., lexicons, TSV/CSV);
- Mentions texts but primarily discusses platforms, implementations, or ontologies;
- Does not include full and somehow accessible textual resources in its data model;
- Does not deal with data relevant to the humanities.

Conversely, a paper is included if it either:

- focuses on the representation, modelling, or publication of texts as LOD, with particular attention to historical and ancient texts;
- describes a case study or project about the publication of texts on the Semantic Web;
- deals with a model or formats for representing textual content as LOD;
- Involves data of interest for the digital humanities, in terms of content, context, or application.

Relevance assessments are occasionally subject to interpretative variability, and, in such cases, decisions are made collaboratively through discussion and comparison among the researchers.

Given the high number of studies and articles available regarding the publication of data following LOD principles, defining a systematic and well-defined protocol helps us focus exclusively on strictly relevant studies.

3.2 Defining Criteria: Sources, Keywords/Seed Terms and Authors

As for sources of information, we focus on conference papers, journal articles, extended abstracts, dissertations, specific case studies, and book chapters.

We apply regular and advanced online searches on the *ACL Anthology, DBPL, Google Scholar, IEEE Xplore*, and *Semantic Scholar*. These digital libraries, archives and tools for scholarly literature are chosen for their extensive coverage, the relevance of their topics to digital humanities, and their high volume of publications, ensuring both inclusiveness and reliability in our review process.

We use 3 different seed keywords combined with additional terms as reported in Table 1. For the choice of keywords, we identify 2 main groups of search terms to ensure inclusiveness and relevance to our research questions. Terms in the first group are seed keywords (second column), which consist of general and foundational terms related to core topics such as "Linked Open Data," "LOD", or "XML/RDF". These keywords are selected to align with our research focus on Linked Open Data and its fundamental aspects. The second group includes terms addressing more specific topics directly linked with our project and humanities in general, such as "ancient languages," "corpora," or "historical text", reflecting the specific needs of our investigation.

For example, we combine these terms in queries like "Linked Open Data" OR "LOD" AND "corpora" OR "ancient languages" OR "historical texts" OR "edition."

Linked Open Data		LOD		XML/RDF		
ancient languages	linked open data	ancient languages	LOD	from	XML	RDF
corpora	linked open data	corpora	LOD	transitioning	XML	RDF
edition	linked open data	edition	LOD	convert	XML	RDF
historical text	linked open data	historical text	LOD	corpora		RDF
transform	linked open data	transform	LOD	conversion		RDF

Table 1: Seed Keywords combinations

This strategy allows us to combine broad foundational concepts with more detailed terms, ensuring that the retrieved documents are comprehensive and aligned with the core objectives of this study. We broaden the search with a multi-term approach to capture resources closely related to the focus of our project.

Other extra keywords targeting known relevant formats/models and authors, such as "NIF", "POWLA"[3], or "Chiarcos", as reported in Table 2, are additionally used to ensure that key contributors in the field are included. These keywords are specifically chosen to address established standards and formats relevant to our research questions.

The seed authors selected are closely tied to LOD works concerning corpora and are among the most frequently cited contributors in this domain. The rationale here is to uncover both new works authored by these individuals and to retrieve similar articles which may offer relevant insight for our survey.

Extra Keywords	Seed Authors
POWLA	C. Chiarcos
NIF in corpora	L. Romary
NLP interchange format	F. Mambrini
	M. C. Passarotti

Table 2: Extra Keywords and Seed Authors

For filtering the search results to a manageable and reasonable number[4], additional criteria are defined. These include:

[3] Respectively, the POWLA Model and the NLP Interchange Format (NIF), better exemplified in the following paragraphs.

[4] To give a rough sense of scale, considering queries that already include the publication date filter, the number of results varies significantly across platforms: *IEEE Xplore* and *DBLP* return on average fewer than 50 results per query; *ACL Anthology* generally stays below 5000 results per query;

- filtering by publication date: we focus exclusively on works published from 2000 onward as the early 2000s mark the formal introduction of the Semantic Web as theorized by [3];
- filtering by language: we narrow the scope to the articles written in English or in Italian;
- sorting and reordering results by relevance, where possible (by applying the functionalities of the respective sources of scientific literature consulted);
- in cases where large volumes of results are generated, concentrating on the first 15 pages of search outputs.

Additionally, citations included in highly relevant studies are examined to discover other related works and ensure that the review process is as comprehensive as possible.

4. Screening for Inclusion (Phase 2)

4.1 Title Reading

For each search on the selected digital libraries and literature service, we carefully review all the resulting titles (and snippet, where available) and ignore all those papers that clearly are not relevant. For example, while conducting an advanced search on *IEEE Xplore* using the seed keyword "Linked Open Data" AND "ancient languages' or "corpora" filtering by year and including only work published from 2000 onward, we exclude clearly unrelated works, such as "An AI Based Automatic Translator for Ancient Hieroglyphic Language" [34], which deals with an AI-based automatic translation designed to recognize and translate Egyptian hieroglyphs into English, or "Online Writing Data Representation: A Graph Theory Approach" [48], a theoretical paper on graph-theory applied to textual data. We also exclude theoretical, introductory or didactic chapters and papers on the Semantic Web, Linked Open Data, or RDF models, like the book "Open Data—Linked Data—Linked Open Data—Linguistic Linked Open Data (LLOD): A General Introduction" [14]. Whenever in doubt, we save the paper for the next screening stage.

A total number of 219 articles are passed to the next stage and saved in the dedicated *Zotero* library.

4.2 Abstract, Introduction, and Conclusion Reading

We now proceed with screening the abstract, introduction, and conclusion of the items saved in the *Zotero* library to decide which ones are truly relevant to our research (recall the workflow in Figure 1 above). At this stage, we can exclude works that do not deal with the semantic web and linked data or with textual resources, and include only works that describe some model or approach to represent or convert datasets of or about (possibly annotated) texts in compliance with the L(O)D principles, i.e., papers that even generically mention the representation of textual

Google Scholar averages around 16700; while *Semantic Scholar* consistently yields the highest number of results, averaging approximately 180000 per query.

data for the Semantic Web. At the end of this step, we retain 136 relevant papers and exclude 83 papers that mostly fall into one of the following three categories:

A. Illustrate general and theoretical topics about (e.g., description of formats such as XML or RDF);

B. Provide a LLOD-compliant model or activity clearly not related to texts (e.g., OntoLex Lemon, SKOS);

C. Provide not interesting works or not relevant topics (e.g., describing the process used in the digitisation, or topics not directly related to LOD).

To give examples of our choices and illustrate the practical application of our pre-selected criteria, we present below some examples of excluded and included papers. For instance, an article excluded at this stage is "The Semantic Web: the Roles of XML and RDF" [16]. This paper is a perfect example of the type of article that falls into category A. In fact, while addressing topics related to LOD and data representation, the article focuses exclusively on theoretical concepts. Specifically, this work explores the use of XML and RDF for semantic interoperability on the Web. Although the article highlights RDF's advantages for semantic LOD representation, its focus remains on general and theoretical concepts, such as the comparative analysis between XML and RDF and the role of ontologies, without directly addressing concrete models or approaches for text representation. For this reason, this type of work is considered irrelevant to the present study[6]. However, since a paper's exact focus is not always clearly identifiable by reviewing only the abstract, introduction, and conclusions, certain theoretical papers were advanced to the subsequent screening stage. This is the case especially of works that include mention to corpora, or state-of-art works about the POWLA model [8] or Ligt [11], an RDF vocabulary which allows the representation of linguistic entities such as words or morphemes, their alignment to grammatical annotations and even their ability to supports the concepts of text hierarchy and segmentation. From a different yet similar perspective, other works are excluded for their focus on specific LOD representation models related to different language aspects that are not strictly related to corpora or textual data. These fall under category B of the exclusion criteria described above. For instance, "Historical Lexicography of Old French and Linked Open Data" [36] explores the transformation of the *Dictionnaire étymologique de l'ancien français*, into LOD exploiting the OntoLex-Lemon model, which deals with the modelling and representation of a lexical resource and does not align with our goal to review models and approaches for the representation of textual data as LOD. Yet again, some apparently similar papers are retained for further screening when they appear to discuss projects or platforms that integrate various types of linguistic data, including text corpora. For instance, the *LiLa: Linking Latin* ERC project [27] and all the articles related to or concerning it are considered relevant at this stage and passed on to the next step.

To provide an example of an article deemed irrelevant under category C of our exclusion criteria above, "Archives, Linked Data and the Digital Humanities: Increasing Access to Digitised and Born-Digital Archives via the Semantic Web" [40] discusses how LOD can improve access to digitised (born-digital) archives, focusing on frameworks and methodologies rather than specific models for representing textual data following LOD principles. For similar reasons, many other papers such as "Access to cultural heritage data: a challenge for the Digital Humanities" [44], or "ARTchives: a linked open data native catalogue of art historians' archives" [43] are excluded, as their main topics are not directly relevant to our research questions and do not explicitly mention textual data representation

4.3 Deep Screening

Finally, the 136 remaining papers are skimmed through entirely to determine whether they are relevant for the analysis phase. At this stage, we also pay attention to specific keywords present inside the texts such as "RDF", "text", "corpora" or "corpus", "sentences", "books", "manuscripts". As a result, we can exclude another 59 articles, which fall into 2 further categories:

D. Do not target text corpora directly, but rather corpus-derived data, represented in the form of lexicons, terminologies, or other structured data formats, such as CSV or TSV files.

E. Although dealing with or mentioning texts, focus on technological aspects such as platforms, software tools, website implementations.

An example of exclusion, under category D above, is the paper "Adapting GermaNet for the Semantic Web" [39], which upon detailed screening is found not to involve direct representation of textual data, nor even natural language examples or glosses. Instead, the article describes transforming *GermaNet* into a Semantic Web-compliant format by adopting the OntoLex-Lemon model to ensure interoperability with other language resources and accessibility as Linked Open Data.

As an example of borderline article excluded at this stage, consider "The CoBiS Linked Open Data Project and Portal" [31]. The *CoBiS* project mostly deals with the mapping and harmonisation of bibliographic metadata coming from different sources belonging to a network of specialised libraries by applying LOD principles. The digitised original publications seem to be accessible and retrievable from the connected archives but not represented in their textual content in any way. As such it could be passed to the next stage; however, the paper focuses on the platform and on the procedures devised to map and harmonise the heterogeneous representations of bibliographic metadata, providing no details about the data model. Consequently, the article is considered not relevant under category E, as it is too little informative.

In contrast, we pass to the next stage many papers dealing with various texts or textual data. These range from unstructured texts, such as those discussed in "Extracting RDF Triples from Raw Text" [1] or in "LODifier: Generating Linked Data from Unstructured Text" [2], which examines both simple (e.g., short sentences) and complex ones (e.g., articulated documents), to the speech act as in "Towards a Linked Open Data Resource for Direct Speech Acts in Greek and Latin Epic" [18], which focuses on direct speech conversations in Greek and Latin epic poems.

4.4 Full-article Reading

Following the previous step, 77 articles remain for deep full-text reading aimed at further evaluating the quality and eligibility of the works, with some additional non-relevant works excluded during this phase. The interesting aspect of this process, and the strength of the methodology we apply for our systematic literature review, is that many articles considered relevant in previous stages can be excluded after a more detailed investigation. Specifically, 8 papers dealing with analysed texts in a Semantic Web context are eventually considered irrelevant after the "full-reading" phase. This is, for instance, the case of [1], mentioned earlier, which proposes an automated method for extracting RDF triples from unstructured texts (i.e., newspaper articles) by applying NLP techniques to syntactically analyse the texts and identify relationships between words, from which they subsequently extract RDF triples. However, unlike similar works that pass this phase, the text itself does not appear to be retained, and the resulting data — i.e., the extracted triples — maintain no link with the original information source. Consequently, we considered this and similar papers irrelevant to our survey.

Furthermore, of these remaining 69 relevant works, at full-text reading 12 papers reveal to be mostly theoretical or general in nature, not dealing with specific representations. It is the case, for instance, of surveys or state-of-the-art reviews on LOD representation models, such as [12] or [8], which focus primarily on vocabularies and frameworks for LOD representation of texts, described from a theoretical perspective. The remaining papers instead are predominantly case studies or project-based works that aim to represent texts or corpora as LOD, often by applying pre-existing models or community-shared vocabularies, and constitute the core of our analysis. It is worth noting that many of these papers are closely related, frequently describing different aspects of the same projects or initiatives. This overlap reduces the number of distinct relevant works or projects classified to less than 69. The assessment and analysis of these papers are conducted independently by the authors of this study, and disagreement is resolved through discussion.

To sum up, Figure 2 provides detailed statistics of the systematic review process, illustrating the progressive refinement of the literature selection. Initially, 219 papers with relevant titles are identified across the selected databases. Following a first screening phase based on the abstract, introduction, and conclusion, 83 papers are excluded for various reasons such as 46 are deemed irrelevant to the topic, 19 are theoretical in nature, and 18, although related to LOD formats, focus on other kinds of data, e.g. terminologies (SKOS), or dictionaries/lexica (OntoLex-Lemon). This process results in 136 relevant papers being retained for deeper evaluation. A following phase of in-depth reading leads to the exclusion of other 59 papers, including 36 that do not focus on corpora or textual formats and 23 that are still not sufficiently relevant to the research questions or criteria. Finally, 77 papers remain for full-article reading and further analysis during the quality and eligibility assessment phase. These remaining papers, 69, are ultimately classified based on criteria such as the granularity of data representation, as well as the models and formats applied (this will be discussed in detail further below).

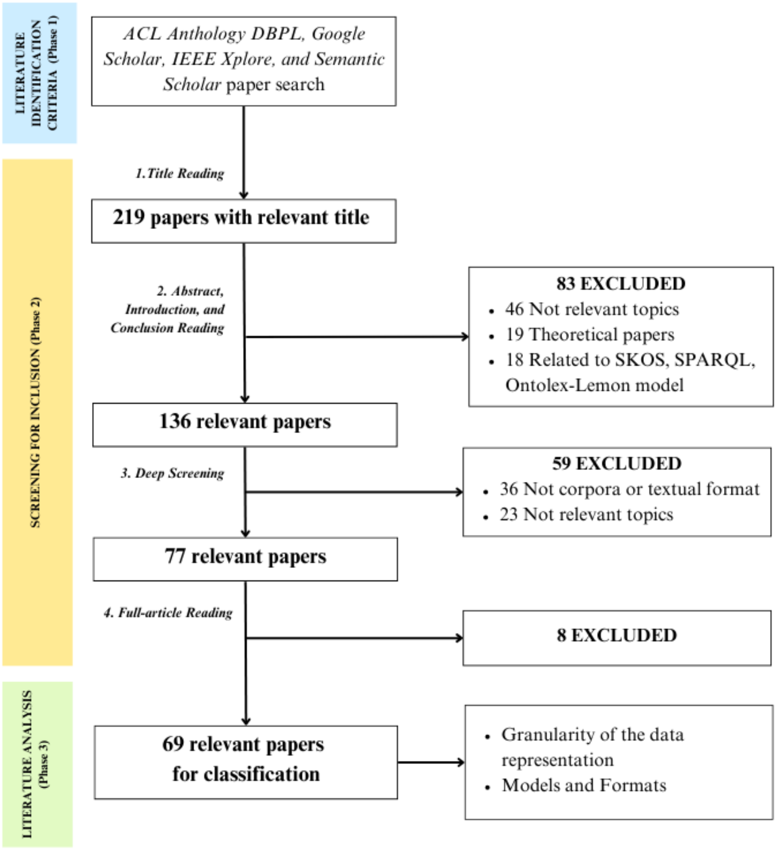

LITERATURE IDENTIFICATION CRITERIA (Phase 1)
ACL Anthology DBPL, Google Scholar, IEEE Xplore, and Semantic Scholar paper search
SCREENING FOR INCLUSION (Phase 2)
1.Title Reading
219 papers with relevant title
2. Abstract, Introduction, and Conclusion Reading
83 EXCLUDED
46 Not relevant topics
19 Theoretical papers
18 Related to SKOS, SPARQL, Ontolex-Lemon model
136 relevant papers
3. Deep Screening
59 EXCLUDED
36 Not corpora or textual format
23 Not relevant topics
77 relevant papers
4. Full-article Reading
8 EXCLUDED
LITERATURE ANALYSIS (Phase 3)
69 relevant papers for classification
Granularity of the data representation
Models and Formats

Figure 2: Statistics on articles and papers selected and excluded for the analysis

5. Literature Analysis and Classification (Phase 3)

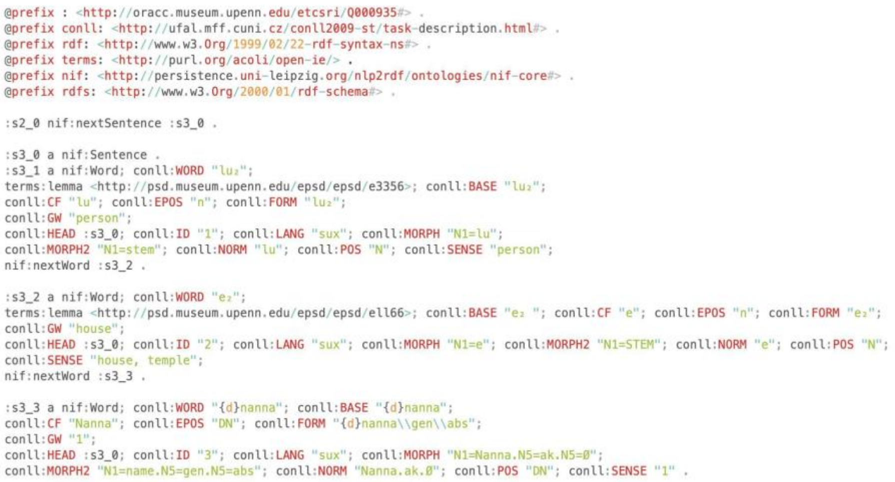
The 69 remaining papers are analysed and categorized based on two key dimensions: the level of granularity in data representation and the models and formats used for representing texts within the Semantic Web. Our primary goal is to elucidate prevalent practices and identify trends in making (annotated) texts available on the Semantic Web. The discussion below synthesizes the most significant findings from our analysis.

5.1 Granularity of the Data Representation

From the perspective of granularity in data representation, many of the surveyed papers represent datasets at document level (that is as bibliographic items or cultural objects) without formalising the representation of the textual and linguistic data thereby contained. For example, “Mapping Manuscript Migrations on the Semantic Web: A Semantic Portal and Linked Open Data Service for Premodern Manuscript Research” [20] describes a search portal that merges catalogue records from various library and archival sources, including more than 200,000 manuscripts, which are described at document level with several kinds of metadata. In addition to standard bibliographic descriptors, the manuscript records are enriched with contextual, historical, geographic and provenance metadata. Moreover, unlike the *CoBis* paper [31] excluded at the previous stage, this work outlines a data model in which original digitised source documents are linked and made accessible through the respective preservation institutions, typically in PDF or image formats.

Other analyzed papers feature a "partial" representation of text contents as LOD; that is, only some predetermined extracted text parts are represented as RDF triples, such as named entities or events, which are then linked to some external KB/KG. For example, "Annotating Arabic Texts with Linked Data" [4] applies an NLP pipeline for extracting tokens from Arabic sentences and automatically maps them to DBPedia concepts with the goal of generating semantic triples as enrichments of the original text, with text documents and triple datasets remaining distinct. Thus, relevant words are "annotated" with DBPedia URIs, establishing a connection between the original text and the semantic information within the ontology. The outcome is an enriched and annotated text, linked to the DBPedia Semantic Web resource.

In a few other projects the representation is more granular and maintains the sentence structure, with the possibility to also represent consequentiality, which refers to the possibility to represent word and sentence order. The *Machine Translation and Automated Analysis of Cuneiform Languages project* (*MTAAC* project) of "Towards a Linked Open Data Edition of Sumerian Corpora" [13], for instance, employs the CoNLL [10] model to represent texts as LOD. In this context, the work is an example of the depth of the representation we are interested in, which includes sentence offsets, tokens, morphological information, as shown in Figure 3 below.

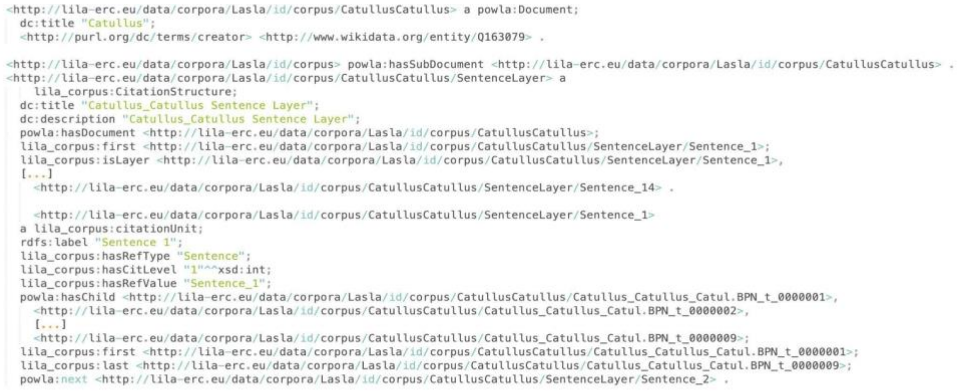


```
@prefix : <http://oracc.museum.upenn.edu/etcsri/Q000935#> .
@prefix conll: <http://ufal.mff.cuni.cz/conll2009-st/task-description.html#> .
@prefix rdf: <http://www.w3.Org/1999/02/22-rdf-syntax-ns#> .
@prefix terms: <http://purl.org/acoli/open-ie/> .
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .
@prefix rdfs: <http://www.w3.Org/2000/01/rdf-schema#> .

:s2_0 nif:nextSentence :s3_0 .

:s3_0 a nif:Sentence .
:s3_1 a nif:Word; conll:WORD "lu₂";
terms:lemma <http://psd.museum.upenn.edu/epsd/epsd/e3356>; conll:BASE "lu₂";
conll:CF "lu"; conll:EPOS "n"; conll:FORM "lu₂";
conll:GW "person";
conll:HEAD :s3_0; conll:ID "1"; conll:LANG "sux"; conll:MORPH "N1=lu";
conll:MORPH2 "N1=stem"; conll:NORM "lu"; conll:POS "N"; conll:SENSE "person";
nif:nextWord :s3_2 .

:s3_2 a nif:Word; conll:WORD "e₂";
terms:lemma <http://psd.museum.upenn.edu/epsd/epsd/e1166>; conll:BASE "e₂ "; conll:CF "e"; conll:EPOS "n"; conll:FORM "e₂";
conll:GW "house";
conll:HEAD :s3_0; conll:ID "2"; conll:LANG "sux"; conll:MORPH "N1=e"; conll:MORPH2 "N1=STEM"; conll:NORM "e"; conll:POS "N";
conll:SENSE "house, temple";
nif:nextWord :s3_3 .

:s3_3 a nif:Word; conll:WORD "{d}nanna"; conll:BASE "{d}nanna";
conll:CF "Nanna"; conll:EPOS "DN"; conll:FORM "{d}nanna\\gen\\abs";
conll:GW "1";
conll:HEAD :s3_0; conll:ID "3"; conll:LANG "sux"; conll:MORPH "N1=Nanna.N5=ak.N5=Ø";
conll:MORPH2 "N1=name.N5=gen.N5=abs"; conll:NORM "Nanna.ak.Ø"; conll:POS "DN"; conll:SENSE "1" .
```

Figure 3: Example of CoNLL-RDF representation of textual corpora representation in MTAAC project. This code-snippet is a simplified version of the example provided in [13:2224]

Within this LOD representation, the authors can represent the cuneiform corpus providing details about the beginning and end of sentences, their components and words' morphological information, and even the word order thanks to specific CoNLL-RDF attributes, derived from the NLP Interchange Format (NIF), which will be discussed better in the next section. To interpret the code provided in the figure above, the sentence is defined with `nif:Sentence`, word is defined as a `nif:Word`, followed by its `conll:WORD`, other annotations in alphabetical order of their properties are provided, concluding a `nif:next` statement pointing to the next word in the sentence. The relationship between words and sentences is established by `conll:HEAD` and `conll:WORD`. The attribute `nif:nextSentence` is used in case there are more sentences following the one represented.

Lastly, the representation of linguistic corpora according to POWLA [8], discussed better in more detail below, generally also extends from sentence to morphology level and can include linking to external resources to provide richer morphological and linguistic information. As shown in Figure 4,**Error! Reference source not found.** in the *LASLA corpus,* part of the *LiLa: Linking Latin* ERC project [17],[5] the text has different layers of representation to encode different types of linguistic information: e.g., sentences through the `SentenceLayer`, and tokens using the `powla_hasChild` attribute. Specific properties are used to specify detailed information about the structure of the text, such as the ordering of the annotation units – e.g., sentences and word tokens — by means of the `lila_corpus:first` and `lila_corpus:next` properties. Also, `powla:hasDocument` is used to link the sentence layer with the main corpus, as shown in the extract taken from the "Catullus Catullus" corpus.



```
<http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus> a powla:Document;
  dc:title "Catullus";
  <http://purl.org/dc/terms/creator> <http://www.wikidata.org/entity/Q163079> .

<http://lila-erc.eu/data/corpora/Lasla/id/corpus> powla:hasSubDocument <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus> .
<http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/SentenceLayer> a
    lila_corpus:CitationStructure;
  dc:title "Catullus_Catullus Sentence Layer";
  dc:description "Catullus_Catullus Sentence Layer";
  powla:hasDocument <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus>;
  lila_corpus:first <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/SentenceLayer/Sentence_1>;
  lila_corpus:isLayer <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/SentenceLayer/Sentence_1>,
  [...]
    <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/SentenceLayer/Sentence_14> .

    <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/SentenceLayer/Sentence_1>
  a lila_corpus:citationUnit;
  rdfs:label "Sentence 1";
  lila_corpus:hasRefType "Sentence";
  lila_corpus:hasCitLevel "1"^^xsd:int;
  lila_corpus:hasRefValue "Sentence_1";
  powla:hasChild <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/Catullus_Catullus_Catul.BPN_t_0000001>,
    <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/Catullus_Catullus_Catul.BPN_t_0000002>,
    [...]
    <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/Catullus_Catullus_Catul.BPN_t_0000009>;
  lila_corpus:first <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/Catullus_Catullus_Catul.BPN_t_0000001>;
  lila_corpus:last <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/Catullus_Catullus_Catul.BPN_t_0000009>;
  powla:next <http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus/SentenceLayer/Sentence_2> .
```

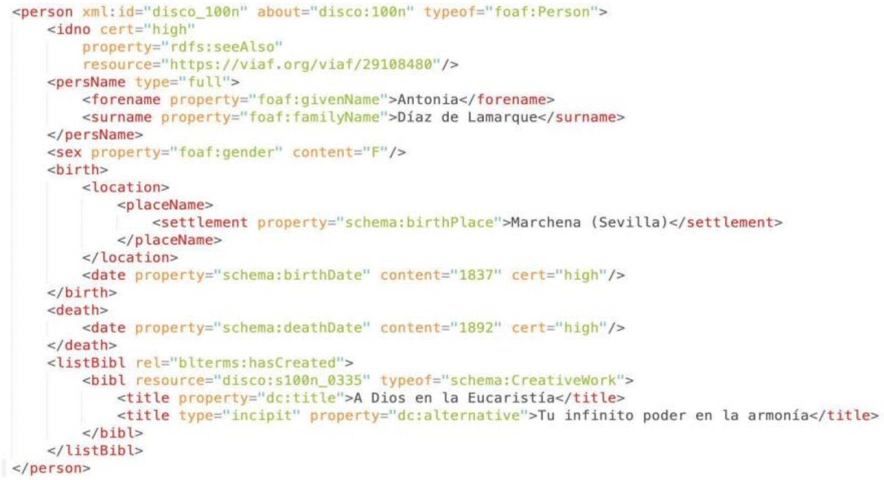
Figure 4: Example of POWLA representation of LASLA corpus in LiLa

5.2 Models and Formats

Looking at the literature in terms of the models and format more commonly adopted for representing texts for the Semantic Web, we observe that some surveyed works (about 5) rely on customizations of XML formats, allowing direct use of RDF within XML, mostly TEI-based, documents by exploiting RDFa. In such cases, RDF triples are directly encoded inline in the XML documents. Within the present review, 4 papers are RDFa-related. Just to provide an

[5] This code-snipped is extracted from the "Catullus Catullus" text of the LASLA corpus represented in POWLA (lines 14-26; 39-48; 55-58). See the project's github repository for the full text code: http://lila-erc.eu/data/corpora/Lasla/id/corpus/CatullusCatullus

example,[6] the *Diachronic Spanish Sonnet Corpus* (DISCO) described in "The Diachronic Spanish Sonnet Corpus: TEI and Linked Open Data Encoding, Data Distribution, and Metrical Findings" [30] makes use of TEI/XML for the digital edition of more than 4000 sonnets by Spanish and Latin American authors from the 15th to the 19th century and includes RDFa attributes to incorporate links to external metadata sources, such as VIAF and Wikidata for author biographical information (e.g. birthplace, date of birth and death, profession). Figure 5 below displays a simplification of the original XML representation used for the encoding of the DISCO[7] corpus. As we can see, an RDFa layer is encoded with different attributes: with the `@typeOf` attribute the domain of the properties is declared, with `@property` the predicates of the RDF triple are defined, `@about` is used to represent the subject, while its IRI is added with `@resource`.



```
<person xml:id="disco_100n" about="disco:100n" typeof="foaf:Person">
    <idno cert="high"
        property="rdfs:seeAlso"
        resource="https://viaf.org/viaf/29108480"/>
    <persName type="full">
        <forename property="foaf:givenName">Antonia</forename>
        <surname property="foaf:familyName">Díaz de Lamarque</surname>
    </persName>
    <sex property="foaf:gender" content="F"/>
    <birth>
        <location>
            <placeName>
                <settlement property="schema:birthPlace">Marchena (Sevilla)</settlement>
            </placeName>
        </location>
        <date property="schema:birthDate" content="1837" cert="high"/>
    </birth>
    <death>
        <date property="schema:deathDate" content="1892" cert="high"/>
    </death>
    <listBibl rel="blterms:hasCreated">
        <bibl resource="disco:s100n_0335" typeof="schema:CreativeWork">
            <title property="dc:title">A Dios en la Eucaristía</title>
            <title type="incipit" property="dc:alternative">Tu infinito poder en la armonía</title>
        </bibl>
    </listBibl>
</person>
```

Figure 5: Example of TEI/XML-RDFa representation of bibliographical information in DISCO project

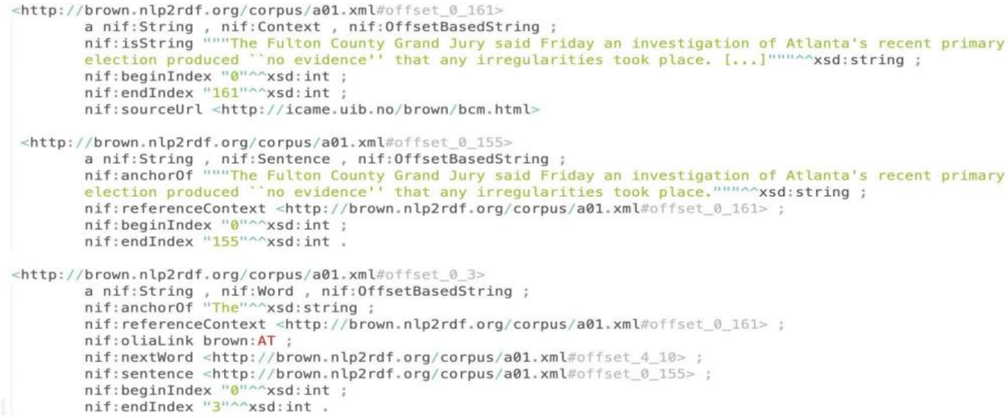
Other projects, specifically 7, explicitly rely on domain-specific RDF models and/or vocabularies, such as CoNLL-RDF, used to represent linguistically annotated natural languages and based on the CoNLL format, a tab-separated-values de-facto standard format typically used in NLP. The *MTAAC* project [13], mentioned and described earlier, is a good example of the application of this model to represent the rich morphologically annotated *Electronic Text Corpus of Sumerian Royal Inscriptions*. Once converted to CoNLL-RDF through their CoNLL2RDF tool, the texts are enriched with links to external resources depending on the type of information being linked. Each lemma of the texts, for instance, is associated with the *Electronic Pennsylvania Sumerian Dictionary* (*ePSD*) via a URI, exploiting the Ontolex–Lemon model; for example,

[6] Due to space constraints, this article reports only a few examples of our classification. For a comprehensive and detailed view, see the full dataset, which contains all the classified reviewed papers: https://doi.org/10.5281/zenodo.10978178

[7] This code-snipped is extracted from the DISCO project's GitHub public repository (READ-ME section). See https://github.com/pruizf/disco/tree/v2.1

referring to terms like “lu₂” for “person”, as shown in the snipped in Figure 3 above. Additionally, the metadata describing the objects are semantically described by recurring to the CIDOC-CRM ontology to model information including provenance, historical period, and the museum of preservation, whereas information about the supports on which the texts are engraved, such as bricks or tablets, are linked to external sources, such as the British Museum linked data.

As mentioned above, CoNLL-RDF was developed based on NIF [19], a stand-off representation model designed to integrate corpus data into the Semantic Web and specifically designed for use within NLP. NIF's key feature is claimed to be its string-based approach, in which every element that can be annotated is identified by a URI, which represents and identifies the element and serves as the subject in RDF triples, making it possible to express annotations on and relationships between strings, texts, and documents. The work “Linking Four Heterogeneous Language Resources as Linked Data” [32], for example, employ it to convert the *Manually Annotated Subcorpus (MASC) of the American National Corpus*, consisting of 500k words of written and transcribed spoken language, at a good granular level, where texts are represented down to the sentence level, with descriptions of their sequential relationships. An example is shown in Figure 6.[8] Similarly to CoNLL-RDF, information about sentence- and token order can be specified through specific attributes; NIF, however, does not provide support for encoding morphology, nor for representing the internal structure of words. This is perhaps the reason why this format does not appear very popular, as we found only 2 works that adopt it: [29] and [52].



```
<http://brown.nlp2rdf.org/corpus/a01.xml#offset_0_161>
        a nif:String , nif:Context , nif:OffsetBasedString ;
        nif:isString """The Fulton County Grand Jury said Friday an investigation of Atlanta's recent primary
        election produced ``no evidence'' that any irregularities took place. [...]"""^^xsd:string ;
        nif:beginIndex "0"^^xsd:int ;
        nif:endIndex "161"^^xsd:int ;
        nif:sourceUrl <http://icame.uib.no/brown/bcm.html>

 <http://brown.nlp2rdf.org/corpus/a01.xml#offset_0_155>
        a nif:String , nif:Sentence , nif:OffsetBasedString ;
        nif:anchorOf """The Fulton County Grand Jury said Friday an investigation of Atlanta's recent primary
        election produced ``no evidence'' that any irregularities took place."""^^xsd:string ;
        nif:referenceContext <http://brown.nlp2rdf.org/corpus/a01.xml#offset_0_161> ;
        nif:beginIndex "0"^^xsd:int ;
        nif:endIndex "155"^^xsd:int .

<http://brown.nlp2rdf.org/corpus/a01.xml#offset_0_3>
        a nif:String , nif:Word , nif:OffsetBasedString ;
        nif:anchorOf "The"^^xsd:string ;
        nif:referenceContext <http://brown.nlp2rdf.org/corpus/a01.xml#offset_0_161> ;
        nif:oliaLink brown:AT ;
        nif:nextWord <http://brown.nlp2rdf.org/corpus/a01.xml#offset_4_10> ;
        nif:sentence <http://brown.nlp2rdf.org/corpus/a01.xml#offset_0_155> ;
        nif:beginIndex "0"^^xsd:int ;
        nif:endIndex "3"^^xsd:int .
```

Figure 6: Example of NIF representation in the MASC corpus

Several papers in our review, instead, represent linguistic corpora of ancient languages according to POWLA, an OWL2/DL vocabulary for linguistic annotations based on the LAF ISO standard, made to support any text-oriented annotation [8], better explained in the previous section with an example from the *LASLA corpus*. It is exploited in many projects throughout the digital humanity community; we count at least 10 papers in our screening, 7 of which,

[8] The snippet is taken from https://bpmlod.github.io/report/nif-corpus/index.html

however, are about the *LiLa: Linking Latin* ERC project, which seeks to interlink and publish in a machine-actionable way different Latin language data resources, making connections among the linguistic tools developed for Latin research, combining text databases, digital archives, dictionaries, and natural language processing tools [26].

Other relevant projects do not adhere to any of the previously mentioned models and describe other, mostly custom or proprietary, models and formats for representing texts as LOD. Some of these projects provide detailed illustrations of the process of creating specific ontologies to represent corpora or textual data as LOD. For example, the *POSTDATA* (*Poetry Standardization and Linked Open Data*) ERC project is currently developing an ontology for the LOD representation of European poetry, offering a semantically enriched version customized for the domain [28]. This initiative highlights a targeted application of standardization and semantic enrichment specifically designed for poetry. Another example to mention is the *Orlando: Women's Writing in the British Isles from the Beginnings to the Present* project from "From XML to RDF in the Orlando Project" [33], a digital resource that includes over 1300 biocritical entries describing over 27000 individual people, represented as annotated text exploiting a detailed XML tag set In this case study, each entry (corresponding to an individual or a text) was converted from XML to RDF triples using a Python script. This work was considered relevant since this process demonstrates an adaptable approach to transforming structured XML data into LOD interoperable formats, which can be considered a practical way to align textual data to the Semantic Web.

6. Discussion

The analysis of the selected literature reveals several insights into the current state of text representation as Linked Open Data within the digital humanities. One of the clearest findings is the limited number of initiatives explicitly addressing this challenge and the relative immaturity of the field. Much of the scholarly focus remains directed toward the modelling, publication, interlinking of catalogues, archival records, and other GLAM-related metadata resources. Similarly, considerable attention is devoted to extracting structured knowledge from texts and representing it through established ontologies and vocabularies, such as CIDOC CRM and LRM. These areas, as the higher volume of related publications suggests, continue to dominate the Semantic Web applications in the humanities.

In contrast, projects that experiment with representing texts themselves as LOD are comparatively few. Even among these, many are concentrated within a small number of research initiatives or sub-communities — such as *LiLa* [17], *POSTDATA* [28], or *Sampo* model [51] indicating that while the technical feasibility is acknowledged, broader adoption remains limited. Where text modelling does occur, several projects resort to ad hoc or project-specific solutions, developed specifically for the use case at hand. It is the case, for example, of the *Orlando* project [33], which converts XML entries into RDF triples and represents a practical yet limited adaptation of existing data structures, or of the *POSTDATA* ERC project [28], aimed at developing an ontology for the semantic representation of European poetry. From the analysis, three models stand out as *de-facto* standard or community models: NIF and CoNLL-RDF, which align more closely with natural language processing practices of annotating texts linguistically, and the POWLA ontology, also capable of representing linguistic annotation and apparently preferred for classical and ancient languages. A third approach, adopted in a few projects, is encoding LOD information within XML documents, often in TEI/XML, by recurring to RDFa syntax. This last option is, however, less convincing as such data would not be directly exploitable in terms of actionability in the LLOD ecosystem.

Nevertheless, emerging convergence around certain models can be observed. POWLA appears particularly preferred in philological contexts, while NIF tends to be favoured in settings with stronger ties to NLP workflows, particularly for historical languages.

A predominant trend is the use of low- to medium-granularity representations. Many works treat texts primarily as bibliographic or cultural artifacts, employing RDF mainly to encode metadata such as authorship, chronology, and provenance. In other cases, the granularity is increased moderately, by encoding some content features of the texts at document level, i.e., assigning to each document descriptive metadata expressed according to some authoritative vocabulary or shared ontology. This is often the case for projects dealing with information extraction and knowledge graph creation for the semantic indexing of documents within digital libraries or archives. Typically, the extracted information concerns named entities (e.g., place names, person names, organisations). Access to the original text material, where present, is provided in the form of a PDF or image.

A smaller but noteworthy subset of studies address a finer-grained level of representation of textual corpora, including sentence-level representations to part-of-speech tagging, morphological analysis, and syntax. These works, mostly based on POWLA or NIF, demonstrate the encoding of detailed annotation layers that preserve sentence segmentation, tokenization, morphological features, and, in some instances, etymological or syntactic information. While these representations theoretically support expressive querying and integration with linguistic or conceptual knowledge bases, the literature reviewed does not report on practical applications of such capabilities — highlighting a gap between representational affordances and their actual exploitation. RDFa embedding within TEI/XML documents emerges as a hybrid strategy but raises concerns regarding its limited actionability, especially when the embedded triples are not extracted or published as dereferenceable resources within LOD infrastructures.

Ultimately, the review exposes a persistent gap between the theoretical potential of LOD technologies and their practical adoption for text representation. Despite the availability of tools and models, many linguistically annotated corpora are not published as LOD. This may stem not only from technical challenges but also from disciplinary and epistemic factors — particularly in fields where the added value of LOD publication has yet to be convincingly demonstrated.

7. Limitations

This review provides a solid and comprehensive account of the state of the art in the representation of historical and ancient textual data as Linked Open Data, following a rigorous and transparent methodology. Nevertheless, some limitations should be acknowledged to contextualize the scope of the study.

First, the temporal coverage is restricted to works published between 2000 and 2023, as the systematic search and data collection were carried out in 2023. While recent publications from 2024 and 2025 may offer additional insights (e.g. [47] and [46]), the specific and still emerging nature of the field suggests that the core findings and trends identified remain valid. Second, although the search was conducted across widely recognized academic search engines and digital libraries indexing a vast amount of scientific literature, it is possible that some relevant works were not retrieved. This applies particularly literature in the digital humanities published in journals or proceedings that are nor indexed or included in the sources we consulted, or to datasets that are not, or are only minimally, documented in scholarly publications. As a results, interesting papers, such as [53] or [54], have not been taken into consideration in this study[9]. Finally, while the screening and classification of works were conducted through a carefully designed, multi-phase process aimed at ensuring consistency and transparency, the selection of borderline cases inevitably involved a degree of subjective judgment, and unintentional human errors may have occurred. However, the openness of the *Zotero* library used for references — and the frozen dataset deposited on Zenodo — supports reproducibility and enables the community to revisit and refine the dataset.

8. Conclusions

This systematic review has surveyed existing approaches to the representation of (historical and ancient) texts as Linked Open Data, identifying trends, common practices, and areas where further development is needed. The findings reveal that, while there is growing interest in applying LOD principles to textual resources, the field is still emerging, and applications remain limited in number. The analysis identified a variety of strategies, ranging from metadata-centric descriptions to more sophisticated models that capture the internal linguistic structure of texts. Among these, POWLA and NIF emerge as more commonly used models, the former more common in philological contexts and the latter in NLP-oriented projects.

Despite the asserted potential of the Semantic Web for both NLP and DH, the reasons limiting its uptake for linguistic text representation remain unclear. We can only speculate that perhaps the verbosity of RDF formats and the associated costs of storage and publication may present tangible challenges, particularly to scholars in the humanities. Similarly, the limited availability of technical expertise, combined with a lack of interest in data sharing, may discourage humanists from undertaking such endeavours. Interestingly, however, even in the fields of computational linguistics and NLP — where technical expertise in such methods is more widespread — the publication of texts as Linked Open Data remains marginal, suggesting that these downsides, though real, may not be the primary limiting factors. The practical benefits of representing entire textual datasets as LLOD have yet to be convincingly demonstrated. Specifically, it remains to be shown that integrating LOD text corpora with other resources in the cloud (lexica, ontologies, concept systems) enables the discovery of new knowledge or facilitates answering open research questions, thanks to the power of reasoning over federated data. Indeed, one of the asserted main values of LOD approaches lies in enabling the discovery of new insights through reasoning and federated querying mechanisms over richly interlinked data. Projects such as *LiLa* hint at

[9] We thank one of the our reviewers for highlighting this limitation.

the potential of these technologies, demonstrating how the creation of linked data ecosystems can integrate diverse types of resources – lexical, conceptual, and textual – thereby facilitating complex queries available to both machines and humans alike.

Finally, we have acknowledged some limitations of our present work, which focuses exclusively on scholarly literature and may therefore have missed important and interesting datasets. Looking ahead, we intend to expand this review to include not only publications but also an analysis of the datasets themselves. Many datasets, in fact, may not be described in publications, or papers may lack the level of detail necessary for our analysis. However, the datasets and related documentation might be available in data repositories like *Zenodo*, in discipline-specific or institutional repositories such as those provided by the CLARIN Centres. Adopting this broader approach could offer a more comprehensive view of how textual data is represented and utilised in contemporary research.

Acknowledgements

This work was supported by the Cnr - Institute for Computational Linguistics and the PRIN 2017 project Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models (no. 2017XJLE8J), funded by the Italian Ministry of University and Research.

The authors acknowledge the use of generative AI for assistance in correcting typographical and grammatical errors and in enhancing the fluency of the language.

We thank the reviewers for their careful reading and constructive feedback, which helped us to clarify several aspects and improve the overall quality of the paper. Any remaining shortcomings or errors are entirely our responsibility.

Author contribution:

Both authors contributed equally to the overall study. According to the Contributor Role Taxonomy (CRediT), M. Bandini was responsible for the original drafting of the manuscript (Writing – original draft), the curation and organization of the data (Data curation), and the specification of the methodology (Methodology). V. Quochi supervised the entire project (Supervision) and was responsible for the critical review and editing of the manuscript (Writing – review & editing). Both authors equally contributed to the conceptual development of the research (Conceptualization), the verification and analysis of results (Validation), and the core research activities (Investigation). Both authors have read and agreed to the published version of the manuscript.

References

[1] Akter, Yeasmin Ara, and Md. Ataur Rahman. 2019. 'Extracting RDF Triples from Raw Text'. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, 1–4. https://doi.org/10.1109/ICASERT.2019.8934694.

[2] Augenstein, Isabelle, Sebastian Padó, and Sebastian Rudolph. 2012. 'LODifier: Generating Linked Data from Unstructured Text'. In: *The Semantic Web: Research*

and Applications, edited by Elena Simperl, Philipp Cimiano, Axel Polleres, Oscar Corcho, and Valentina Presutti, 7295:210–24. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30284-8_21.

[3] Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. 'The Semantic Web'. *Scientific American* 284 (5): 34–43. http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.

[4] Bouziane, Abdelghani, Djelloul Bouchiha, and Noureddine Doumi. 2020. 'Annotating Arabic Texts with Linked Data'. In *2020 4th International Symposium on Informatics and Its Application*s (ISIA), 1–5. https://doi.org/10.1109/ISIA51297.2020.9416543.

[5] Bruce, Julie, and Jill Mollison. 2004. 'Reviewing the Literature: Adopting a Systematic Approach'. *Journal of Family Planning and Reproductive Health Care* 30 (1): 13–16. https://doi.org/10.1783/147118904322701901.

[6] Buono, Maria Pia di, Philipp Cimiano, Mohammad Fazleh Elahi, and Frank Grimm. 2020. 'Terme-à-LLOD: Simplifying the Conversion and Hosting of Terminological Resources as Linked Data'. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, edited by Maxim Ionov, John P. McCrae, Christian Chiarcos, Thierry Declerck, Julia Bosque-Gil, and Jorge Gracia, 28–35. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.ldl-1.5/.

[7] Cayless, Hugh A. 2019. 'Sustaining Linked Ancient World Data'. In *Ancient Greek and Latin in the Digital Revolution*, edited by Monica Berti, 35–50. Berlin, Boston: De Gruyter Saur. https://doi.org/doi:10.1515/9783110599572-004.

[8] Chiarcos, Christian. 2012. 'POWLA: Modeling Linguistic Corpora in OWL/DL'. In *The Semantic Web: Research and Applications*, edited by Elena Simperl, Philipp Cimiano, Axel Polleres, Oscar Corcho, and Valentina Presutti, 7295:225–39. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30284-8_22.

[9] Chiarcos, Christian, Christian Fäth, and Maxim Ionov. 2020. 'The ACoLi Dictionary Graph'. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 3281–90. ELRA.

[10] Chiarcos, Christian, and Luis Glaser. 2020. 'A Tree Extension for CoNLL-RDF'. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 7161–69. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.885.

[11] Chiarcos, Christian and Maxim Ionov. 2019. 'Ligt: An LLOD-Native Vocabulary for Representing Interlinear Glossed Text as RDF'. *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Open Access Series in Informatics (OASIcs), Volume 70, pp. 3:1-3:15, Schloss Dagstuhl – Leibniz-Zentrum für Informatik https://doi.org/10.4230/OASICS.LDK.2019.3

[12] Chiarcos, Christian, Bettina Klimek, Christian Fäth, Thierry Declerck, and John Philip McCrae. 2020. 'On the Linguistic Linked Open Data Infrastructure'. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, 8–15. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.iwltp-1.2.

[13] Chiarcos, Christian, Émilie Pagé-Perron, Ilya Khait, Niko Schenk, and Lucas Reckling. 2018. 'Towards a Linked Open Data Edition of Sumerian Corpora'. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA). https://aclanthology.org/L18-1387.

[14] Chiarcos, Christian, and Antonio Pareja-Lora. 2020. 'Open Data—Linked Data—Linked Open Data—Linguistic Linked Open Data (LLOD): A General Introduction'. In *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*, edited by Antonio Pareja-Lora, María Blume, Barbara C. Lust, and Christian Chiarcos, 1–18. The MIT Press. https://doi.org/10.7551/mitpress/10990.003.0003.

[15] Cimiano, Philipp, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. 'Linguistic Linked Data in Digital Humanities'. In *Linguistic Linked Data*, by Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia, 229–62. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-30225-2_13.

[16] Decker, S., S. Melnik, F. van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, and I. Horrocks. 2000. 'The Semantic Web: The Roles of XML and RDF'. *IEEE Internet Computing* 4 (5): 63–73. https://doi.org/10.1109/4236.877487.

[17] Fantoli, Margherita, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. 2022. 'Linking the LASLA Corpus in the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin'. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference,* 26–34. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.ldl-1.4

[18] Forstall, Christopher W, Simone Finkmann, and Berenice Verhelst. 2022. 'Towards a Linked Open Data Resource for Direct Speech Acts in Greek and Latin Epic'. *Digital Scholarship in the Humanities* 37 (4): 972–81. https://doi.org/10.1093/llc/fqac006.

[19] Hellmann, Sebastian, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. 'Integrating NLP Using Linked Data'. In *Advanced Information Systems Engineering*, edited by Camille Salinesi, Moira C. Norrie, and Óscar Pastor, 7908:98–113. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41338-4_7.

[20] Hyvönen, Eero, Esko Ikkala, Mikko Koho, Jouni Tuominen, Toby Burrows, Lynn Ransom, and Hanno Wijsman. 2021. 'Mapping Manuscript Migrations on the

Semantic Web: A Semantic Portal and Linked Open Data Service for Premodern Manuscript Research'. In *The Semantic Web – ISWC 2021*, edited by Andreas Hotho, Eva Blomqvist, Stefan Dietze, Achille Fokoue, Ying Ding, Payam Barnaghi, Armin Haller, Mauro Dragoni, and Harith Alani, 12922:615–30. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-88361-4_36.

[21] Khan, Anas, Christian Chiarcos, Thierry Declerck, Daniela Gîfu, Elena García, Jorge Gracia, Maxim Ionov, et al. 2022. 'When Linguistics Meets Web Technologies. Recent Advances in Modelling Linguistic Linked Data'. *Semantic Web* 13 (June):1–64. https://doi.org/10.3233/SW-222859.

[22] Mambrini, Francesco, and Marco Passarotti. 2019. 'Linked Open Treebanks. Interlinking Syntactically Annotated Corpora in the LiLa Knowledge Base of Linguistic Resources for Latin'. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, 74–81. Paris, France: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-7808.

[23] McCrae, John P., Steven Moran, Sebastian Hellmann, and Martin Brümmer. 2015. 'Multilingual Linked Data'. *Semantic Web* 6 (4): 315–17. https://doi.org/10.3233/SW-150178.

[24] Navigli, Roberto, and Simone Paolo Ponzetto. 2012. 'BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network'. *Artificial Intelligence* 193:217–50. https://doi.org/10.1016/j.artint.2012.07.001.

[25] Paré, Guy, Marie-Claude Trudel, Mirou Jaana, and Spyros Kitsiou. 2015. 'Synthesizing Information Systems Knowledge: A Typology of Literature Reviews'. *Information & Management* 52 (2): 183–99. https://doi.org/10.1016/j.im.2014.08.008.

[26] Passarotti, Marco, Eleonora Litta, Flavio Massimiliano Cecchini, Matteo Pellegrini, Giovanni Moretti, Paolo Ruffolo, and Giulia Pedonese. 2022. 'The LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. Architecture and Current State'.

[27] Passarotti, Marco, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. 'Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin'. *Studi e Saggi Linguistici* 58 (1): 177–212. https://doi.org/10.4454/ssl.v58i1.277.

[28] Platas, María Luisa Diez, Salvador Ros, Elena González-Blanco, Helena Bermúdez, and Oscar Corcho. 2019. 'The POSTDATA Network of Ontologies for European Poetry.'

[29] Rezk, Martín, Jungyeul Park, Yoon Yongun, Kyungtae Lim, John Larsen, YoungGyun Hahm, and Key-Sun Choi. 2013. 'Korean Linked Data on the Web: Text to RDF'. In *Semantic Technology*, edited by Hideaki Takeda, Yuzhong Qu,

Riichiro Mizoguchi, and Yoshinobu Kitamura, 7774:368–74. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37996-3_31.

[30] Ruiz Fabo, Pablo, Helena Bermúdez Sabel, Clara Martínez Cantón, and Elena González-Blanco. 2021. 'The Diachronic Spanish Sonnet Corpus: TEI and Linked Open Data Encoding, Data Distribution, and Metrical Findings'. *Digital Scholarship in the Humanities* 36 (Supplement_1): i68–80. https://doi.org/10.1093/llc/fqaa035.

[31] Schiavone, Luisa, Federico Morando, and The CoBis Communication Working Group. 2018. 'The CoBiS Linked Open Data Project and Portal'. Edited by S. Lesteven, B. Kern, R. D'Abrusco, and B. Dorch. *EPJ Web of Conferences* 186:12013. https://doi.org/10.1051/epjconf/201818612013.

[32] Siemoneit, Benjamin, John Philip McCrae, and Philipp Cimiano. 2015. 'Linking Four Heterogeneous Language Resources as Linked Data'. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, 59–63. Beijing, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-4207.

[33] Simpson, John, and Susan Brown. 2013. 'From XML to RDF in the Orlando Project'. In *2013 International Conference on Culture and Computing*, 194–95. https://doi.org/10.1109/CultureComputing.2013.61.

[34] Sobhy, Asmaa, Mahmoud Helmy, Michael Khalil, Sarah Elmasry, Youtham Boules, and Nermin Negied. 2023. 'An AI Based Automatic Translator for Ancient Hieroglyphic Language—From Scanned Images to English Text'. *IEEE Access* 11:38796–804. https://doi.org/10.1109/ACCESS.2023.3267981.

[35] Templier, Mathieu, and Guy Paré. 2015. 'A Framework for Guiding and Evaluating Literature Reviews'. *Communications of the Association for Information Systems* 37. https://doi.org/10.17705/1CAIS.03706.

[36] Tittel, Sabine, and Christian Chiarcos. 2018. 'Historical Lexicography of Old French and Linked Open Data: Transforming the Resources of the *Dictionnaire Étymologique de l'ancien Français* with OntoLex-Lemon'. *Proceedings of Globalex 2018: Lexicography & Wordnets,* edited by Ilan Kernerman and Simon Krek. 8 May 2018, Miyazaki, Japan. European Language Resources Association (ELRA).

[37] Tupman, Charlotte. 2021. 'Where Can Our Inscriptions Take Us?: Harnessing the Potential of Linked Open Data for Epigraphy'. In E*pigraphy in the Digital Age: Opportunities and Challenges in the Recording, Analysis and Dissemination of Inscriptions*, 115–28. Archaeopress. http://www.jstor.org/stable/j.ctv1xsm8s5.15.

[38] Xiao, Yu, and Maria Watson. 2019. 'Guidance on Conducting a Systematic Literature Review'. *Journal of Planning Education and Research* 39 (1): 93–112. https://doi.org/10.1177/0739456X17723971.

[39] Zinn, Claus, Marie Hinrichs, and Erhard Hinrichs. 2022. 'Adapting GermaNet for the Semantic Web'. In *Proceedings of the 18th Conference on Natural Language Processing*

(KONVENS 2022), 41–47. Potsdam, Germany: KONVENS 2022 Organizers. https://aclanthology.org/2022.konvens-1.6.

[40] Hawkins, Ashleigh. 2022. 'Archives, Linked Data and the Digital Humanities: Increasing Access to Digitised and Born-Digital Archives via the Semantic Web.' *Archival Science* 22 (3): 319–44. https://doi.org/10.1007/s10502-021-09381-0.

[41] Kiefer, Ferenc. 1988. 'Linguistic, conceptual and encyclopedic knowledge: Some implications for lexicography'. In *Proceedings of the 3rd EURALEX International Congress,* 1-10. Budapest: Akadémiai Kiadó.

[42] Sowa, John F. 1993. 'Lexical Structures and Conceptual Structures'. In *Semantics and the Lexicon*, edited by J. Pustejovsky, 231-263. *Studies in Linguistics and Philosophy*, vol. 49. Springer, Dordrecht. https://doi.org/10.1007/978-94-011-1972-6_12

[43] Tomasi, Francesca, Marilena Daquino and Lucia Giagnolini. 2021. 'ARTchives: a linked open data native catalogue of art historians' archives'. In *Proceedings of Linked Archives International Workshop 2021, co-located with 25th International Conference on Theory and Practice of Digital Libraries (TPDL 2021).* Online, September 13th, 2021, edited by Carla Teixeira Lopes, Cristina Ribeiro, Franco Niccolucci, Irene Rodrigues, and Nuno Freire.

[44] Baillot, Anne, Marie Puren, Charles Riondet, Dorian Seillier, and Laurent Romary. 2017. 'Access to cultural heritage data. A challenge for digital humanities. *Proceedings of the Digital Humanities Conference 2017*'. Aug 2017, Montréal, Canada. https://dh2017.adho.org/abstracts/DH2017-abstracts.pdf

[45] Lambon, Ralph Matthew A. 2014. 'Neurocognitive insights on conceptual knowledge and its breakdown'. *Philoosophical Transaction of the Royal Society B* 369: 2012. http://doi.org/10.1098/rstb.2012.0392.

[46] Armaselu, Florentina , Chaya Liebeskind, Paola Marongiu, Barbara McGillivray, Giedre Valunaite Oleskeviciene, Elena-Simona Apostol, Ciprian-Octavian Truica, and Daniela Gifu. 2024. 'LLODIA: A Linguistic Linked Open Data Model for Diachronic Analysis'. In *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*, pages 1–10, Torino, Italia. ELRA and ICCL. https://aclanthology.org/2024.ldl-1.1/

[47] Stanković, Ranka, Milica Ikonić Nešić, Mihailo Škorić, Olja Perišić, and Olivera Kitanović. 2024. 'Towards Semantic Interoperability: Parallel Corpora as Linked Data Incorporating Named Entity Linking'. In *Proceedings of the 9th Workshop on Linked Data in Linguistics: Resources, Applications, Best Practices*, Turin, 25 May 2024. ACL Anthology.

[48] Caporossi, Guillaume, and Cédric Leblay. 2011. 'Online Writing Data Representation: A Graph Theory Approach'. In *International Symposium on Intelligent Data Analysis*, 80–89. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-24800-9_10

[49] Michele Mallia, Michela Bandini, Andrea Bellandi, Francesca Murano, Silvia Piccini, Luca Rigobianco, Alessandro Tommasi, Cesare Zavattari, Mariarosaria Zinzi, and Valeria Quochi. 2024. 'DigItAnt: a platform for creating, linking and

exploiting LOD lexica with heterogeneous resources'. In *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*, pages 55–65, Torino, Italia. ELRA and ICCL.

[50] Murano, Francesca, Valeria Quochi, Angelo Mario Del Grosso, Luca Rigobianco, and Mariarosaria Zinzi. 2023. 'Describing Inscriptions of Ancient Italy. The ItAnt Project and Its Information Encoding Process.' *ACM Journal on Computing and Cultural Heritage* Vol. 16, no. 3 (September 2023): 53. https://doi.org/10.1145/3606703

[51] Hyvönen, Eero. 2023. 'Digital Humanities on the Semantic Web: Sampo Model and Portal Series.' *Semantic Web* Vol. 14, no. 4: 729–744. https://doi.org/10.3233/SW-223034

[52] Siemoneit, Benjamin , John Philip McCrae, and Philipp Cimiano. 2015. 'Linking Four Heterogeneous Language Resources as Linked Data'. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 59–63, Beijing, China. Association for Computational Linguistics. https://aclanthology.org/W15-4207/

[53] Del Grosso, A. M., Capizzi, E., Cristofaro, S., De Luca, M. R., Giovannetti, E., Marchi, S., Seminara, G and Spampinato, D. (2019). Bellini's Correspondence: a Digital Scholarly Edition for a Multimedia Museum. Umanistica Digitale, 3(7). https://doi.org/10.6092/issn.2532-8816/9162

[54] Daquino, M., Giovannetti, F., & Tomasi, F. (2019). 'Linked Data per le edizioni scientifiche digitali. Il workflow di pubblicazione dell'edizione semantica del quaderno di appunti di Paolo Bufalini'. *Umanistica Digitale*, 3(7). https://doi.org/10.6092/issn.2532-8816/9091