Università di Catania

ASSOCIAZIONE per l'INFORMATICA UMANISTICA e la CULTURA DIGITALE

Consiglio Nazionale delle Ricerche

ME.TE. DIGITALI

MEDITERRANEO IN RETE TRA TESTI E CONTESTI

ATTI DEL XIII CONVEGNO ANNUALE
AIUCD 2024

28 - 30 MAGGIO

MONASTERO DEI BENEDETTINI

P.ZZA DANTE, 32 CATANIA

ISBN 978-88-942535-8-0

A cura di: Di Silvestro Antonio; Spampinato Daria (2024). Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, Proceedings del XIII Convegno Annuale AIUCD, Catania 28-30 maggio 2024, Università di Catania.

Editing: Denise Bruno; Christian D'Agata; Laura Mazzagufo; Francesca Prado; Eliana Vitale; Alessandro Zammataro.

Ultimo accesso agli URL in data 15 maggio 2024.

Si prega di notificare all'editore ogni omissione o errore si riscontri: aiucd.segreteria [at] aiucd.org
Please notify the publisher of any omissions or errors found: aiucd.segreteria [at] aiucd.org

Il programma della conferenza AIUCD 2024 è disponibile online https://aiucd2024.unict.it/programma/
The AIUCD 2024 Conference Program is available online https://aiucd2024.unict.it/programma/

I contributi pubblicati nel presente volume hanno ottenuto il parere favorevole da parte di valutatori esperti della materia, attraverso un processo di revisione anonima mediante *double-blind peer review* sotto la responsabilità del Comitato di Programma di AIUCD 2024.
All the papers published in this volume have received favourable reviews by experts in the field of DH, through an anonymous double-blind peer review process under the responsibility of the AIUCD 2024 Program Committee.

Chair

Antonio Di Silvestro (Università di Catania)
Daria Spampinato (CNR Istituto di Scienze e Tecnologie della Cognizione)

Comitato di programma / *Program committee*

Emmanuela Carbé (Università di Siena)
Massimo Cultraro (CNR Istituto di Scienze del Patrimonio Culturale)
Christian D'Agata (Università di Catania)
Antonio Di Silvestro (Università di Catania)
Greta Franzini (Eurac Research)
Maurizio Lana (Università del Piemonte Orientale)
Cristina Marras (CNR Istituto del Lessico intellettuale europeo e Storia delle Idee)
Marco Mazzone (Università di Catania)
Ouafae Nahli (CNR Istituto di Linguistica Computazionale "Antonio Zampolli")
Marianna Nicolosi-Asmundo (Università di Catania)
Marina Paino (Università di Catania)
Giuseppe Palazzolo (Università di Catania)
Jonathan Prag (University of Oxford Merton College)
Daria Spampinato (CNR Istituto di Scienze e Tecnologie della Cognizione)
Rachele Sprugnoli (Università di Parma)
Francesco Stella (Università di Siena)

Segreteria scientifica / *Scientific Secretariat*

Liborio Barbarino (Università di Catania)
Denise Bruno (Università di Catania)
Giulia Cacciatore (Università di Catania)
Giuseppe Canzoneri (Università di Catania)
Elisa Conti (Università di Catania)
Milena Giuffrida (Università di Catania)
Miryam Grasso (Università di Catania)
Francesca Prado (Università di Catania)
Emilio M. Sanfilippo (CNR Istituto di Scienze e Tecnologie della Cognizione)
Eliana Vitale (Università di Catania)
Alessandro Zammataro (Università di Catania)

Comunicazione istituzionale: Claudia Cantale (Università di Catania) e Area Per la Comunicazione dell'Università di Catania (ACOM).
Institutional communication: Claudia Cantale (University of Catania) and the Area for Communication of the University of Catania (ACOM)

Supporto tecnico: Rosario Agrò, Area della Terza Missione dell'Università di Catania, per la consulenza e la progettazione grafica dei materiali informativi del convegno.
Technical support: Rosario Agrò, Third Mission Area of the University of Catania, for advice and graphic design of the conference information materials.

Enti organizzatori / *Organisers*
AIUCD; Università di Catania: Dipartimento di Scienze Umanistiche; CNR Istituto di Scienze e Tecnologie della Cognizione; CINUM: Centro di Informatica Umanistica dell'Università di Catania.

Supporter
CLARIN-IT; Neperia Group; Storage; programma Piaceri 2020-2022, Linea 1; Parmalat-Sole.

Chair di area/ Track chair

Le culture digitali nel Mediterraneo
Cristina Marras (CNR Istituto del Lessico intellettuale europeo e Storia delle Idee)
Paola Moscati (CNR Istituto di Scienze del Patrimonio Culturale)
Archivi ed edizioni digitali
Christian D'Agata (Università di Catania)
Greta Franzini (Eurac Research)
Analisi computazionale dei testi
Angelo Mario Del Grosso (CNR Istituto di Linguistica Computazionale "Antonio Zampolli")
Simone Rebora (Università di Verona)
Ontologie e Semantic Web
Marianna Nicolosi Asmundo (Università di Catania)
Francesca Tomasi (Università di Bologna)
Preservazione della memoria e del patrimonio digitale
Fabio Ciracì (Università del Salento)
Anna Maria Marras (Università di Torino)

Lista dei revisori /List of reviewers

Maristella Agosti (Università di Padova), **Stefano Allegrezza** (Università di Bologna), **Chiara Alzetta** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Liborio Barbarino** (Università di Catania), **Nicola Barbuti** (Università di Bari Aldo Moro), **Stefano Bazzaco** (Università di Verona), **Benedetta Bessi** (Università Ca' Foscari di Venezia), **Andrea Bolioli** (ricercatore indipendente), **Paolo Bonora** (Università di Bologna), **Federico Boschetti** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Dominique Brunato** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Flavia Bruni** (Università Gabriele D'Annunzio di Chieti-Pescara), **Marina Buzzoni** (Università Ca' Foscari di Venezia), **Alberto Campagnolo** (Universitè Catholique de Louvain/KULeuven), **Anna Cappellotto** (Università di Verona), **Emmanuela Carbé** (Università di Siena), **Vittore Casarosa** (CNR Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" – Università di Pisa), **Fabio Ciotti** (Università di Roma "Tor Vergata"), **Fabio Ciracì** (Università del Salento), **Elisa Conti** (Università di Catania), **Salvatore Cristofaro** (CNR Istituto per il Lessico Intellettuale Europeo e Storia delle Idee), **Christian D'Agata** (Università di Catania), **Elisa D'Argenio** (HUN-REN Hungarian Research Centre for Linguistics), **Mauro De Bari** (Università di Bari Aldo Moro), **Riccardo Del Gratta** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Angelo Mario Del Grosso** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Saulo Delle Donne** (Università del Salento), **Giorgio Maria Di Nunzio** (Università di Padova), **Antonio Di Silvestro** (Università di Catania), **Filippo Diara** (Università di Torino), **Giulia Fabbris** (Università Ca' Foscari di Venezia), **Riccardo Fedriga** (Università di Bologna), **Franz Fischer** (Università Ca' Foscari di Venezia), **Greta Franzini** (Eurac Research), **Francesca Frontini** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Daniele Fusi** (Stuttgart University & VeDPH – Università Ca' Foscari di Venezia), **Carola Gatto** (Università del Salento), **Lucia Giagnolini** (Università di Bologna), **Emiliano Giovannetti** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Milena Giuffrida** (Università di Catania), **Edmondo Grassi** (Università San Raffaele di Roma), **Miryam Grasso** (Università di Catania), **Alessandro Iannella** (Università di Cagliari - Università di Pisa – Università di Torino), **Paola Italia** (Università di Bologna), **Maurizio Lana** (Università del Piemonte Orientale), **Pietro Maria Liuzzo** (Bibliotheca Hertziana), **Dominique Longrée** (Université de Liège), **Francesco Mambrini** (Università Cattolica del Sacro Cuore di Milano), **Tiziana Mancinelli** (Istituto Italiano di Studi Germanici), **Anna Maria Marras** (Università di Torino), **Cristina Marras** (CNR Istituto del Lessico intellettuale europeo e Storia delle Idee), **Federico Meschini** (Università della Tuscia), **Alessio Miaschi** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Paolo Monella** (Università Sapienza di Roma), **Ouafae Nahli** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Marianna Nicolosi-Asmundo** (Università di Catania), **Giuseppe Palazzolo** (Università di Catania), **Valentina Pasqual** (Università di Bologna), **Gianluca Pavani** (Università di Roma "Tor Vergata"), **Giulia Pedonese** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Jonathan Prag** (University of Oxford Merton College), **Simone Rebora** (Università di Verona), **Giulia Renda** (Università di Bologna), **Roberto Rosselli Del Turco** (Università di Torino), **Enrica Salvatori** (Università di Pisa), **Emilio M. Sanfilippo** (CNR Istituto di Scienze e Tecnologie della Cognizione), **Eva Sassolini** (CNR Istituto di Linguistica Computazionale "Antonio Zampolli"), **Pietro Sichera** (CNR Istituto di Scienze e Tecnologie della Cognizione), **Daniele Silvi** (Università di Roma "Tor Vergata"), **Elena Spadini** (University of Basel), **Daria Spampinato** (CNR Istituto di Scienze e Tecnologie della Cognizione), **Linda Spinazzè** (Università Ca' Foscari di Venezia), **Rachele Sprugnoli** (Università di Parma), **Francesco Stella** (Università di Siena), **Cecilia Tamagnini** (Università di Bologna), **Timothy Tambassi** (Università Ca' Foscari di Venezia), **Francesca Tomasi** (Università di Bologna), **Marco Venuti** (Università di Catania), **Fabio Vitali** (Università di Bologna).

Representing texts as LOD: a Systematic Literature Review

Michela Bandini[1], Valeria Quochi[2]
[1]CNR Istituto di Linguistica Computazionale "A. Zampolli", Italia - michela.bandini@ilc.cnr.it
[2]CNR Istituto di Linguistica Computazionale "A. Zampolli", Italia - valeria.quochi@ilc.cnr.it

ABSTRACT
Despite the growing interest in publishing linguistic data as Linked Open Data, the publishing of ancient language corpora for the Semantic Web is still challenging. This contribution describes a systematic literature review on the representation of corpus data as Linguistic Linked Open Data, focusing especially on models and (data) granularity. Our goal is to gain insights into the advantages and disadvantages of the different approaches. Here we present our systematic review methodology and some initial results.

KEYWORDS
Linked Open Data; corpora; ancient languages; systematic literature review.

1. INTRODUCTION
A trend has gained increasing attention in the representation and publication of language datasets as Linked Open Data (LOD), primarily for Knowledge Representation (KR) and Natural Language Processing (NLP) purposes. In recent times, LOD and Semantic Web (SW) technologies have also captured the interest of digital humanists, with an expanding variety of data sources being published as LOD. The vast majority of linguistic resources in the LOD cloud [5] are dictionaries, lexica, thesauri, terminologies, and (controlled) vocabularies. There is a recent growing interest in publishing text corpora as Linguistic LOD (LLOD) both in NLP and in Digital Humanities (DH) communities. Linked data, in fact, "allows better data integration than existing models of linguistic data, due to the ecosystem of tools provided by the Semantic Web" and "enable[s] better resource interoperability" [11: 315]. As part of an ongoing project on the digital representation of scholarly knowledge about archaic languages and cultures, one of the main goals of the present contribution is to explore the possibilities for representing and publishing corpora of ancient inscriptions (mostly available in `XML` TEI-Epidoc or in database formats) as LLOD. We thus started with a systematic literature review on this topic. The review's methodology follows 3 generic steps, described in detail in the rest of the paper.

2. REVIEWING METHODOLOGY
Defining terms and questions of the research
This systematic review addresses works and projects in which text corpora are represented as Linked Open Data. We, therefore, developed this review to understand:

- what are the most relevant works that have already attempted to transform - or represent - (text) corpora as LOD;
- what are the models and formats already in use by the digital humanities community to represent corpora as LOD;
- how can we classify projects according to the model used for representation, to the granularity of the representation and to the purpose (or research sub-community).

Given the high number of studies and articles regarding the publication of data following LOD principles, we decided to adopt a systematic approach to help us focus on relevant studies. We searched specific digital libraries, used seed keywords and authors, and applied several criteria to filter the results to concentrate on the most relevant for our purposes. Figure 1 shows the workflow which explains the steps followed for the selection of works used in this review.

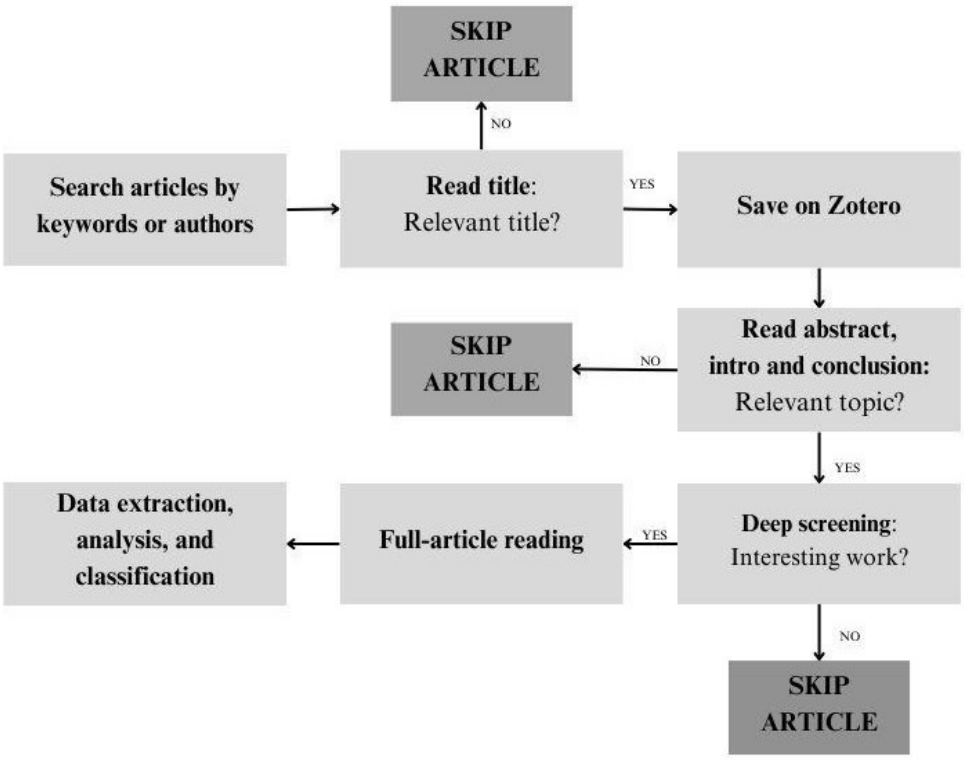

Figure 1. Reviewing workflow

Defining criteria: sources, keywords/seed terms and authors

We mainly applied regular and advanced online searches on digital libraries such as *ACL Anthology, DBPL, Google Scholar, IEEE Xplore, and Semantic Scholar*. We used 3 different seed keywords combined with other terms as reported in Table 1. We opted for 2 groups of terms to be inclusive and relevant. Terms in the first group were considered seed keywords (second column), main and generic words concerning core or general topics such as "Linked Open Data", "LOD", or "XML/RDF"; in the second group terms were related to more specific topics strictly linked with our project such as "ancient languages" or "corpora". The idea was to use a seed keyword combined with a more detailed term to gather projects as relevant as possible to our interests. We, indeed, tried to broaden the research with a multi-term search using words closely related to the main focus of our project. Other extra keywords, such as "POWLA" or "NIF", as well as seed authors[1] were used for targeting known formats/models directly. The search conducted on digital libraries was also filtered by date from 2000 up to today and by language (we only opted for English and Italian works). The results were sorted by relevance[2] and/or by the number of citations (when possible). Whenever the number of results pages was high, we only focused on the first 15 pages per search.

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

[1] These are authors strictly related to LOD works on corpora and are the most cited authors in articles about this topic.

[2] It refers to the plug-in functionality of the respective digital libraries where it's possible to reorder search results by relevance or pertinence.

3. LITERATURE SCREENING AND ASSESSMENT

Screening for Inclusion

Title Reading. We focused on conference papers, journal articles, extended abstracts, dissertations, specific case studies, and book chapters. For each search, we read through all the resulting titles and ignored all those papers which clearly were not relevant (e.g. generic and theoretical papers related to topics such as what is LOD, the RDF format, etc.). All other articles, 219, were passed to the next stage and saved in a dedicated Zotero library[3].

Abstract, introduction, and conclusion reading. We proceeded with the screening of the abstract, introduction, and conclusion of the items saved in the Zotero library, to decide which ones were truly relevant for our research. At this stage, we only included works that described some model or approach to represent or convert (possibly annotated) texts in compliance with L(O)D principles, i.e. papers that even generically described the representation of textual data for the Semantic Web. At the end of this step, we were left with 136 relevant papers, and excluded 83 papers which:

- illustrated general and theoretical topics (e.g. description of formats such as XML or RDF, etc.). Yet, we took into account theoretical papers regarding corpora (e.g. state-of-art about POWLA [2] or Ligt [4]);
- provided not interesting works or not relevant topics (e.g. describing the process used in the digitization of data, not related to LOD);
- provided a LLOD-compliant model or activity clearly not related to texts (e.g. OntoLex Lemon, SKOS, etc.[4]).

Deep reading. The remaining papers were skimmed through entirely to determine whether they were actually relevant. This time we also focused on specific keywords present inside the texts such as "XML", "text", "corpora" or "corpus", "sentences", "books", "manuscripts", etc. We excluded another 59 articles which:

- did not target text-corpora, but rather corpus-derived data represented as lexicons, or CSV / TSV data.
- mentioned textual data, but in fact dealt with platforms, website implementation, or specific ontologies.

Quality and Eligibility Assessment

Full-article reading. Following the second screening, 77 articles remained for deep full-text reading aimed at further assessing the quality and eligibility of the works and eventually excluded a few other non-relevant works. 12 relevant works were theoretical papers on LOD representation models for texts. The rest focused on case studies or project-related works. The assessment and analysis were performed independently by the authors of this paper, and disagreement was resolved through discussion.

4. DATA ANALYSIS

The 77 remaining papers were analyzed and categorized based on two key criteria: the level of granularity in data representation and the models and formats used for representing texts within the Semantic Web. Our primary goal was 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.

Granularity of the data representation. From the perspective of granularity representation of data, many surveyed papers represent datasets at a document level, i.e. as bibliographic entities or cultural objects, without detailing the representation of the textual data thereby contained, i.e. sequences of linguistic signs. For example, in the *Mapping Manuscript Migrations (MMM) project* [9] manuscripts are represented as bibliographic entries, i.e. at metadata level (e.g. author, production place, production data, language, etc.).

Most 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, in [1] tokens are extracted from Arabic sentences and automatically mapped to DBPedia for generating semantic triples as enrichments of the original text, with text documents and triple datasets remaining distinct.

In a few other projects, mainly those related to `NIF` and `CoNLL-RDF`, the representation is more granular. The *Machine Translation and Automated Analysis of Cuneiform Languages project (MTAAC project)* [6], for instance, employs the

[3] The exported and versioned dataset of the Zotero library discussed in this paper is available on Zenodo for reference and reproducibility purposes. See https://zenodo.org/doi/10.5281/zenodo.10978178. The bibliographic entries within the library are categorized into "relevant" and "not-relevant" works. All tags used for classification and analysis have been preserved in the exported dataset.

[4] Yet, some papers were retained if they discussed projects or platforms that integrate various types of data, including text corpora; for instance the *LiLa: Linking Latin* project [13].

CoNLL-RDF [3] 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 and word order, as shown in Figure 2, extracted from an actual example provided in [6: 2441][5]. 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's order thanks to specific CoNLL-RDF attributes. To interpret the code provided in the figure below, 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 provide, concluding with 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.

```
@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/ell66>; 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 2. Example of CoNLL-RDF representation of textual corpora representation in MTAAC project

Lastly, the representation of linguistic corpora according to POWLA [2] generally also extends from sentence to token level and can include linking to external resources to provide richer morphological and linguistic information. As shown in Figure 3, in the LASLA corpus of the *LiLa: Linking Latin* project [7][6], the document has different layers of representation to encode sentences and tokens. Specific properties are used to specify detailed information about the structure of the text, such as the beginning or the ending of the sentences and their token order.

Models and Formats. Regarding this criterion, a number of surveyed works can be considered precursors to LOD representation either because they predate the definition of LLOD or because they rely on customizations of the XML formats, allowing direct use of RDF within TEI documents by exploiting **RDFa**. In such cases, RDF triples are directly encoded inline in the XML documents. For instance, the *Diachronic Spanish Sonnet Corpus (DISCO)* [16] makes use of TEI/XML for the digital edition of poems 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 4 below displays a simplification of

[5] This code-snippet is a re-elaboration of the provided in Figure 1 of [6: 2441].
[6] This code-snipped is extracted from the "Catullus Catullis" 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

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.

```
<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 3. Example of POWLA representation of LASLA corpus in LiLa project

```
<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 4. Example of TEI/XML-RDFa representation of bibliographical information in DISCO project

Other projects explicitly rely on domain-specific RDF models and/or vocabularies, such as **CoNLL-RDF** [3], used to represent linguistically annotated natural languages. The *MTAAC project* [6] is a nice example of the application of this model to represent linguistically annotated text as LOD. According to this first and broad analysis, this model can be considered one of the most convenient solutions for representing textual data or sentences, given the detailed possibilities in granularity representation.

[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

CoNLL-RDF was developed based on the **NLP Interchange Format (NIF)** [8], a stand-off representation model for the integration of corpus data into the Semantic Web, especially devised to leverage NLP tools over L(O)D. [17], for example, successfully employs it to convert and publish the "Manually Annotated Subcorpus (MASC) of the American National Corpus" at a good granular level, as stated previously. However, it looks like this format is not very popular: we found only one other work that exploits it [15]. As also stated in [10], there seems to be a sort of dispreference of NIF over other solutions such as CoNLL-RDF, apparently because it does not provide sufficient support for the annotation of morphological traits and for the internal structure of words. In detail, NIF is an RDF-based format for describing strings in text documents, and its classes and properties are defined in the NIF Core Ontology. **Figure 5** is a code excerpt extracted from the NIF edition of the Brown corpus published in 2015 [10][8]. As shown in the figure, the context, which usually describes the whole document's text, is created with nif:Context. Then each node is associated with a nif:String to represent a textual chunk, in our example: sentence (nif:Sentence) and words (nif:Word). For each textual chunk, properties are assigned to provide information about its beginning and the ending, and the representation of the actual string.

Several papers in our review, instead, represent linguistic corpora according to **POWLA,** an OWL2/DL vocabulary for linguistic annotations based on the LAF ISO standard, made to support any kind of text-oriented annotation [2]. It is exploited in many projects throughout the digital humanity community; we count at least 10 papers in our screening, 7 of which 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 [12].

Other relevant projects do not adhere to any of the previously mentioned models and describe other, mostly custom or proprietary, models and formats. For instance, the *Poetry Standardization and Linked Open Data* project develops a specific ontology for the semantic representation of European poetry [14]. Other works, like the project *Orlando: Women's Writing in the British Isles Project* [18], provide some solutions to partially convert XML documents to RDF.

```
<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 5. Example of NIF representation of the Brown corpus

5. CONCLUSION

This contribution presented the initial results of a systematic literature review on models and representation formats of linguistic textual data as LLOD. Although the interest in the task seems to be growing, in practice the actual adoption of these practices remains limited, with few projects publishing linguistic text corpora as LOD. This raises concerns about the reasons and their real-world viability or utility.

Our exploration of models and formats for LOD representation for linguistic corpora reveals a restricted range of existing approaches. Three models, in particular, stand out as the most interesting and used: CoNLL-RDF and NIF, which seem to demonstrate their effectiveness in representing linguistically annotated corpora, especially within NLP; and the POWLA ontology/model, which is a preferred choice in digital humanities projects, notably in the *LiLa Knowledge Base*.

[8] The snippet in Figure 5 is taken from https://bpmlod.github.io/report/nif-corpus/index.html.

Looking ahead, we plan to expand our review to include not only published papers but also datasets themselves, potentially enriching our understanding of the field. Many datasets, as a matter of fact, may not be described in publications, but could instead be available in discipline-specific or institutional data repositories, such as Zenodo and Research Infrastructure repositories. This broader approach may provide a more comprehensive picture of how linguistic data is represented and utilized in current research.

6. ACKNOWLEDGEMENTS

This work is carried out in the context of the PRIN 2017 "Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models" (no. 2017XJLE8J) funded by the Italian Ministry of University and Research. The DigItAnt platform is also supported by CLARIN-IT.

REFERENCES

[1] Bouziane, Abdelghani, Bouchiha Djelloul, and Doumi Noureddine. '"Annotating Arabic Texts with Linked Data."' 2020 4th International Symposium on Informatics and Its Applications (ISIA), 2020, 1–5.

[2] Chiarcos, Christian. '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, 7295:225–239. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2012.

[3] Chiarcos, Christian, and Luis Glaser. '"A Tree Extension for CoNLL-RDF."*' In Proceedings of the Twelfth Language Resources and Evaluation Conference*, 7161–7169. European Language Resources Association, 2020.

[4] Chiarcos, Christian, and Maxim Ionov. 'Ligt: An LLOD-Native Vocabulary for Representing Interlinear Glossed Text as RDF', 2019.

[5] Chiarcos, Christian, Bettina Klimek, Christian Fäth, Thierry Declerck, and John Philip McCrae. '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, 2020.

[6] Chiarcos, Christian, Émilie Pagé-Perron, Ilya Khait, Niko Schenk, and Lucas Reckling. '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), 2018.

[7] Fantoli, Margherita, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. '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, 2022.

[8] Hellmann, S., J. Lehmann, S. Auer, and M. Brümmer. 'Integrating NLP Using Linked Data'. In The *Semantic Web – ISWC*, Vol. 8219. Berlin, Heidelberg: Springer, 2013.

[9] Hyvönen, Eero, Esko Ikkala, Mikko Koho, Jouni Tuominen, Toby Burrows, Lynn Ransom, and Hanno Wijsman. 'Mapping Manuscript Migrations on the Semantic Web: A Semantic Portal and Linked Open Data Service for Premodern Manuscript Research'. In *The Semantic Web – ISWC*, edited by Andreas Hotho, Eva Blomqvist, Stefan Dietze, Achille Fokoue, Ying Ding, Payam Barnaghi, Armin Haller, Mauro Dragoni, and Harith Alani, 12922:615–630. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021.

[10] Khan, Fahad Anas, Christian Chiarcos, Thierry Declerck, Daniela Gîfu, Elena González-Blanco García, Jorge Gracia, Maxim Ionov, et al. 'When Linguistics Meets Web Technologies. Recent Advances in Modelling Linguistic Linked Data'. *Semantic Web* 13 (2022): 1–64.

[11] McCrae, John P., Steven Moran, Sebastian Hellmann, and M. Brümmer. 'Multilingual Linked Data'. *SemanticWeb* 6 (2015): 315–317.

[12] Passarotti, Marco, Elena Litta, Flavio Massimiliano Cecchini, Matteo Pellegrini, Giovanni Moretti, Paolo Ruffolo, and Giulia Pedonese. 'The LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. Architecture and Current State'. In *Elsevier Guest Seminar Series*, 2022.

[13] Passarotti, Marco, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 'Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin'. *Studi e Saggi Linguistici* 58, no. 1 (2020): 177–212. https://doi.org/10.4454/ssl.v58i1.277

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

[15] Rezk, Martín, Jungyeul Park, Yoon Yongun, Kyungtae Lim, John Larsen, Young Gyun Hahm, and Key-Sun Choi. 'Korean Linked Data on the Web: Text to RDF'. In *Semantic Technology*, edited by Hideaki Takeda, Yuzhong Qu, Riichiro Mizoguchi, and Yoshinobu Kitamura, 7774:368–374. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[16] Ruiz, Fabo Pablo, Helena Bermúdez Sabel, Clara Martínez Cantón, e Elena González-Blanco. 'The Diachronic Spanish Sonnet Corpus: TEI and Linked Open Data Encoding, Data Distribution, and Metrical Findings'. *Digital Scholarship in the Humanities* 26, no. Supplement 1 (2021): i68-i80.

[17] Siemoneit, Benjamin, John Philip McCrae, and Philipp Cimiano. ‘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, 2015.

[18] Simpson, John, and Susan Brown. ‘From XML to RDF in the Orlando Project’. In *2013 International Conference on Culture and Computing*, 194–195, 2013.