

Modeling the Letters of Nobel Laureate Ivo Andrić as a Knowledge Graph

Marija Đokić Petrović^{1,*†}, Mihajlo Stojanović^{2,†} and Branko Tošović^{3,†}

¹Department of Computer Science, The School of Computing, Union University, Kneza Mihaila VI/6, 11000, Belgrade, Serbia

²Department of Slavic Studies, The University of Graz, Merangasse 70/I, 8010, Graz, Styria, Austria

Abstract

This paper presents a workflow for transforming a selected subset of Nobel laureate Ivo Andrić's letters into a Knowledge Graph (KG). Building on previous work that combined Optical Character Recognition (OCR), Named Entity Recognition (NER), and SPARQL-based contextualization of Andrić's letters, the present study extends this methodology by introducing an explicit RDF representation of letters. A small corpus of 15 letters written by Ivo Andrić to Zdenka Marković between 1925 and 1939 was first processed using Transkribus. The PAGE XML output was then converted into structured records representing individual letters. These records were enriched with automatically extracted named entities, normalized through post-processing rules, and linked to Wikidata. Finally, the structured and enriched data were transformed into an RDF graph. The resulting KG contains 1,667 RDF triples representing letters, persons, places, organizations, dates, and links to Wikidata. The final output is a queryable RDF graph that supports the structured exploration, analysis, and visualization of Andrić's letters, providing a more nuanced view of his personal experiences and social networks. This offers a basis for exploring how these factors may have informed Andrić's literary work. The developed KG can serve as a reusable resource for further research in Digital Humanities.

Keywords

Knowledge Graph, Linked Data, Named Entity Recognition, RDF, Digital Humanities, Transkribus, Ivo Andrić

1. Introduction

Personal letters are an important source for literary, cultural, and historical research [1]. They contain information on communication networks, locations, dates, personal relationships, intellectual exchange, and broader cultural contexts. However, such material is often available only in printed form, which makes systematic computational analysis difficult. Therefore, transforming letters into structured and machine-readable data opens new possibilities for querying, visualizing, and contextualizing these valuable documents [2].

This paper focuses on a selected subset of letters written by Ivo Andrić [3], a Yugoslav writer, diplomat, and Nobel laureate. Previous work [4] applied OCR and NER to extract persons, places, and dates from Andrić's letters and used them in predefined SPARQL [5] queries over DBpedia to uncover important information about the socio-political climate and Andrić's social milieu that could influence his work. The present work extends this approach by constructing an RDF [6] Knowledge Graph (KG) from the letters, their metadata, and extracted entities. This enables the letters to be explored as structured Linked Data and supports the analysis of entities and their relations across the corpus.

The main contribution of this paper is the design and implementation of a reproducible workflow that transforms Ivo Andrić's printed letters into a queryable and visualizable KG. The resulting resource

Posters, Demos, Blue Sky, and Tutorials at SEMANTICS 2026, Sep 2026, Ghent, Belgium

*Corresponding author.

†These authors contributed equally.

✉ mdjokicpetrovic@raf.rs (M. Đokić Petrović); m_stojanovic@raf.rs (M. Stojanović); branko.tosovic@uni-graz.at (B. Tošović)

🌐 <https://marijadjokicpetrovic.com/> (M. Đokić Petrović);

<https://raf.edu.rs/o-nama/nastavnici-i-saradnici/mihajlo-stojanovic/> (M. Stojanović);

<https://homepage.uni-graz.at/de/branko.tosovic/> (B. Tošović)

🆔 0000-0002-0568-5219 (M. Đokić Petrović); 0009-0006-8823-9906 (M. Stojanović); 0000-0002-6970-671X (B. Tošović)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

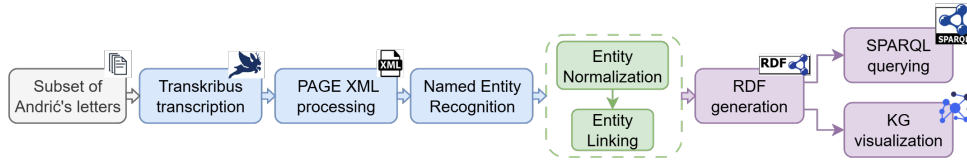


Figure 1: Overview of the semi-automatic pipeline for creating a knowledge graph from a subset of Ivo Andrić's letters

provides a basis for broader exploration of Andrić's social networks and historical context and can be reused in future Digital Humanities research.

2. Semi-Automatic Pipeline for Knowledge Graph Creation

The transformation of Andrić's letters into a KG was performed using a semi-automated pipeline that combines Transkribus-based transcription [7], PAGE XML processing, Named Entity Recognition (NER) [8], entity normalization, Wikidata [9] linking, RDF [6] generation, SPARQL [5] querying, and KG visualization. Figure 1 provides an overview of the complete workflow, while the individual steps are described in the following subsections.

2.1. Corpus and Data Preparation

The experimental corpus consists of 15 pre-selected letters written by Ivo Andrić to Zdenka Marković, a Croatian writer and translator, between 1925 and 1939. The letters were taken from a printed edition of Andrić's private letters [3]. As a proof-of-concept subset, they were chosen to cover different years, places of writing, and entity-rich texts, with the aim of testing the feasibility of representing letter metadata and named entities in a structured RDF model before scaling the workflow to a larger corpus.

The selected letters were uploaded to Transkribus [7], where several models were tested, including Text Titan ¹, Transkribus Print M1², and Burgenland Croatian Typewritten 2010-2019³. The Burgenland Croatian model [10] produced the best results for the mixed Serbian-Croatian Latin-script text, with occasional German and Serbian Cyrillic words, and was therefore used for transcription. The output was lightly corrected in Transkribus, focusing only on errors relevant for metadata extraction and NER, such as personal names, locations, organizations, and dates.

The corrected transcription was exported from Transkribus as PAGE XML (a standardized XML-based format that stores recognized text together with its spatial layout) and processed with a custom Python pipeline. The pipeline converted PAGE XML lines into structured letter records and segmented the letters mainly by place-and-date headers, such as *Beograd 2. IV. 25.* or *Marseille 2 februara 1927.* Additional normalization repaired words split across line breaks, for example *Polj- sku* into *Poljsku*. The resulting output contains 15 structured records, including a letter identifier, the source PAGE XML file, the raw header, the full text, the sender, the recipient, the normalized date, and the place of writing. In this selected corpus, the sender was normalized as *Ivo Andrić* and the recipient as *Zdenka Marković*.

2.2. Named Entity Recognition and Entity Linking

NER was applied to the letter texts using the BERTiC (classla/bcms-bertic-ner) model [11]. The model was used to identify core named-entity categories, such as person (PER), location (LOC), organization (ORG), and miscellaneous entities (MISC) - including publications and literary works. Dates (DATE) were extracted separately using regular expressions, as the NER model does not provide a dedicated label for this entity category. The NER output was processed using predefined rules to reduce

¹<https://www.transkribus.org/models/the-text-titan-i-super-model>

²<https://www.transkribus.org/models/transkribus-print-multi-language-dutch-german-english-finnish-french-swedish-etc>

³<https://www.transkribus.org/models/burgenland-croatian-typewritten-2010-2019>

errors associated with OCR-derived text and the model tokenizer. WordPiece token fragments were merged, and entity surface forms were reconstructed from their character offsets in the source text. For example, the location name *Avignona*, initially tokenized as *Avi*, *##gno*, and *##na*, was reconstructed as a single entity mention, *Avignona*. Rule-based post-processing subsequently removed false-positive and out-of-scope candidates, such as *Bog* (God), *Uskrs* (Easter), and the spurious fragment *Ma*, as well as candidates assigned an incorrect entity type, such as *Francusku* (France) classified as a person. Selected entity-boundary errors were also corrected. Finally, predefined alias mappings normalized abbreviated, inflected, and script variants to canonical forms. For example, *I. Andrić* was normalized to *Ivo Andrić*, while the inflected form *Beogradu* and the Cyrillic variant *Београд* were normalized to *Beograd*. The NER model output contains 99 entities: 32 PER, 59 LOC, 5 ORG, and 3 MISC; and 28 extracted DATE. Among the most frequent entities are *Ivo Andrić*, *Beograd*, *Zagreb*, *Ženeva*, *Zdenka Marković*, and *Marseille*.

Furthermore, selected entities were linked to Wikidata [9] using the Wikidata API and manually controlled linking rules. For example, *Ivo Andrić* entity was linked to the corresponding Wikidata resource - <http://www.wikidata.org/entity/Q47561>. DATE entities were excluded from the NEL.

To obtain an initial estimate of the quality of the information extraction pipeline, we performed a preliminary evaluation of NER and NEL on the 15-letter corpus (Table 1). NER was evaluated at the entity-string level, ignoring span offsets. The results indicate that both NER and NEL achieve solid overall performance (micro F1 = 0.78).

Table 1

Preliminary evaluation setup for NER and entity linking

Task	Type	Gold	Pred.	Correct	Metric (P/R/F1)
NER	PER	29	32	28	0.88/0.97/0.92
	LOC	71	59	46	0.78/0.65/0.71
	ORG	5	5	5	1.00/1.00/1.00
	MISC	1	3	1	0.33/1.00/0.50
	micro avg.	106	99	80	0.81/0.76/0.78
NEL	PER	11	11	10	0.91/0.91/0.91
	LOC	24	26	18	0.69/0.75/0.72
	ORG	5	5	4	0.80/0.80/0.80
	micro avg.	40	42	32	0.76/0.80/0.78

2.3. RDF Modelling

The extracted and normalized data were transformed into an RDF [6] graph using RDFLib⁴ library. The KG uses a lightweight project-specific vocabulary (<https://example.org/andric-letters/>), indicated by the prefix *andric*: together with established vocabularies such as Schema.org, Dublin Core Terms, PROV-O, RDFS, SKOS, and XSD. In the current proof-of-concept version, the project namespace is temporary and will be replaced by a stable registered URI in future work.

Each letter is represented as an instance of *andric:Letter* and *schema:Message* (Listing 1). The model includes basic correspondence metadata, such as sender, recipient, date sent, place of writing, source file, pagination, and full transcription text. The graph distinguishes between normalized entity resources and entity-mention nodes. Normalized entity resources represent labeled entity categories, while entity-mentioned nodes represent concrete occurrences of these entities in individual letters. The use of explicit mention nodes is important because the same normalized entity can occur several times, in different forms, and in different letters. For example, a place may be linked to a normalized entity resource, while each textual occurrence keeps its own surface form, offsets, NER type, confidence score, and provenance.

```
@prefix andric: <https://example.org/andric-letters/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix schema: <https://schema.org/> .
```

⁴<https://rdflib.readthedocs.io/en/stable/>

```

@prefix prov: <http://www.w3.org/ns/prov#> .
andric:letter_0001 a andric:Letter,
    schema:Message ;
    dcterm:identifier "letter_0001" ;
    dcterm:isPartOf andric:andric-private-letters-1925-1941 ;
    prov:hadPrimarySource "0001_p001.xml" ;
    schema:pagination "1" ;
    schema:dateSent "1925-02-11"^^xsd:date ;
    schema:locationCreated andric:loc_sarajevo ;
    schema:sender andric:per_ivo_andric ;
    schema:recipient andric:per_zdenka_markovic ;
    schema:mentions andric:date_11_ii_25, andric:date_1925,
        andric:loc_sarajevo, andric:per_ivo_andric ;
    schema:text ""Capajebo, 11. II 1925.

Draga gospodice,
vraćajući se sa izbora sjećam Vas se i pozdravljam
Vaš Ivo Andrić
11. II. 25
Mama Vam zahvaljuje na brizi i pozdravlja Vas."" .

```

Listing 1: A simplified RDF representation of Ivo Andrić’s letter

The resulting graph⁵ contains 1,667 triples, 15 letter resources, 121 entity mentions, and 73 unique entities. To assess the quality of the generated KG, we performed a preliminary quality check combining automatic structural validation and manual inspection (Table 2). The graph was checked for syntactic validity, completeness of letter metadata, and consistency of entity mention nodes.

Table 2
Preliminary KG quality assessment

Dimension	Scope	Check	Result
RDF validity	Full graph	Turtle parsing	Valid RDF graph
Letter metadata	15 letters	Sender, recipient, date, place, source	Complete for all letters
Entity mentions	121 mentions	Type, offsets, confidence, provenance	Structurally complete

2.4. Querying and Visualization

Once represented as RDF, the letters can be queried using SPARQL [5]. The current query set, available in the project repository⁶, supports several exploratory tasks. For example, querying the graph for the most frequently mentioned persons highlights the central role of *Ivo Andrić* and *Zdenka Marković*, while additional person entities reveal the broader social and intellectual network surrounding the letters. Similarly, querying places of writing and mentioned locations makes it possible to distinguish between the geography of letters production and the geography evoked within the letters. These query results demonstrate that the KG is not only a storage format, but also an analytical layer for exploring letters as structured cultural-historical data.

The KG can also be visualized as a network in which letters are connected to persons, places, organizations, literary works, and dates (Figure 2). This visualization is important because it makes the relational structure of the letters visible: instead of reading the letters only as isolated textual documents, researchers can explore them as a network of social, geographical, and temporal connections. Such views support exploratory analysis and may help identify patterns in Andrić’s personal networks and historical context that are not immediately apparent through close reading alone.

⁵https://github.com/marijadjokic/andrics_correspondence_as_kg/blob/master/data/output/andric_letters_kg.ttl

⁶https://github.com/marijadjokic/andrics_correspondence_as_kg/tree/master/queries

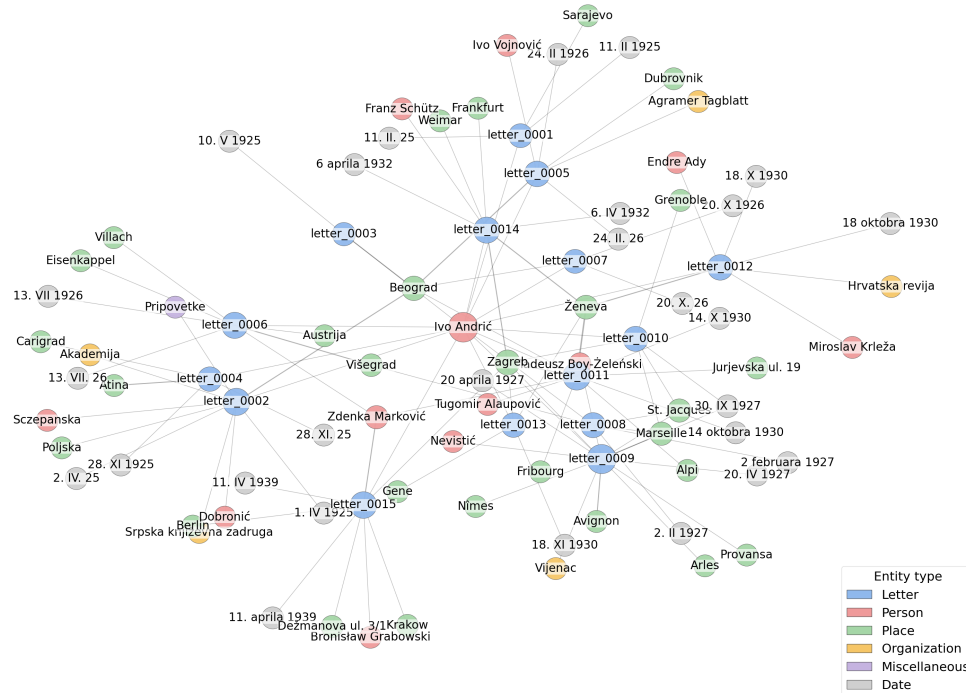


Figure 2: Knowledge graph visualization of the selected subset of Ivo Andrić's letters

3. Discussion and Limitations

The experiment confirms that a semi-automatic workflow from Transkribus to RDF is feasible for the selected subset of Andrić’s letters. However, KG quality depends on OCR accuracy, the performance of the NER model, and the effectiveness of entity normalization and entity linking.

The current corpus is limited to 15 letters and should be understood as a proof-of-concept rather than a basis for broader literary-historical interpretation. While manually creating post-processing rules was manageable at this scale, applying the workflow to a larger corpus will require automated normalization rules and iterative evaluation.

4. Conclusion

This paper presented a workflow for modeling a selected subset of Ivo Andrić’s letters as a KG. Starting from the Transkribus transcription and PAGE XML export, the workflow performs NER, entity normalization, Wikidata linking, RDF generation, SPARQL querying, and graph visualization. The resulting graph contains 1,667 RDF triples and represents 15 letters, their metadata, entity mentions, and links to external knowledge bases such as Wikidata.

The main value of the approach lies in transforming Andrić's letters into a structured, queryable, extensible, and reusable semantic resource, providing a basis for exploring Andrić's correspondence networks and their broader cultural and historical context. Future work will extend the corpus to different correspondents and periods, allowing the analysis of changes in Andrić's social, geographical, and intellectual networks over time.

5. Acknowledgments

This work is supported by COST Action CA23147 GOBLIN - Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in

Science and Technology, <https://www.cost.eu>).

This work uses the Transkribus model Burgenland Croatian Typewritten 2010–2019, developed with the support of the Transkribus Scholarship program, by READ-COOP SCE.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly to check grammar and spelling. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

- [1] G. D. L. Camiciotti, Letters and Letter Writing in Early Modern Culture: An Introduction, *Journal of Early Modern Studies* 3 (2014) 17–35. doi:10.13128/JEMS-2279-7149-14163.
- [2] E. Hyvönen, P. Leskinen, J. Tuominen, LetterSampo–Historical Letters on the Semantic Web: A Framework and Its Application to Publishing and Using Epistolary Data, *ACM Journal on Computing and Cultural Heritage* 16 (2023). URL: <https://doi.org/10.1145/3569372>. doi:10.1145/3569372.
- [3] I. Andrić, *Pisma: (1912–1973): privatna pošta*, edited by Miroslav Karaulac, Matica srpska, Novi Sad, 2000. COBISS.BH-ID 11313414.
- [4] M. Đokić Petrović, D. Bečejski-Vujaklija, J. Đorđević, J. Mitrović, Unveiling Literary Legacies: Integrating Named Entity Recognition and SPARQL to Analyse Andrić's Letters, in: B. Tošović (Ed.), *Andrić's Brief*, Institute of Slavic Studies at the Karl-Franzens University of Graz, National and University Library of the Republic of Srpska in Banjaluka, Commission for Stylistics of the International Slavic Committee, World of Book Belgrade, Aletia Belgrade, Graz - Belgrade - Banjaluka, 2025, pp. 231–247.
- [5] World Wide Web Consortium (W3C), SPARQL 1.1 Query Language, W3C Recommendation, 21 March 2013. URL: <https://www.w3.org/TR/sparql11-query/>.
- [6] World Wide Web Consortium (W3C), RDF 1.1 Concepts and Abstract Syntax, W3C Recommendation, 25 February 2014. URL: <https://www.w3.org/TR/rdf11-concepts/>.
- [7] READ-COOP SCE, Transkribus, Innsbruck: READ-COOP SCE. <https://transkribus.eu/>, 2026. Accessed: 10 June 2026.
- [8] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, A. Doucet, Named Entity Recognition and Classification in Historical Documents: A Survey, *ACM Comput. Surv.* 56 (2023). URL: <https://doi.org/10.1145/3604931>. doi:10.1145/3604931.
- [9] Wikimedia Foundation, Wikidata, 2026. URL: <https://www.wikidata.org/>, accessed: 10 June 2026.
- [10] M. Đokić Petrović, M. Simonović, Automatic Text Recognition for Burgenland Croatian Printed Materials: Developing a Custom Model on Transkribus, *Research Square preprint*, 2026. URL: <https://doi.org/10.21203/rs.3.rs-9957704/v1>. doi:10.21203/rs.3.rs-9957704/v1.
- [11] N. Ljubešić, D. Lauc, BERTiC - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian, in: B. Babych, O. Kanishcheva, P. Nakov, J. Piskorski, L. Pivovarov, V. Starko, J. Steinberger, R. Yangarber, M. Marcińczuk, S. Pollak, P. Přibán, M. Robnik-Šikonja (Eds.), *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, Association for Computational Linguistics, Kyiv, Ukraine, 2021, pp. 37–42. URL: <https://aclanthology.org/2021.bsnlp-1.5/>.

6. Online Resources

The source code and supporting materials required to reproduce the pipeline described in this paper are available on [GitHub](#).