ADVANCES IN KNOWLEDGE ORGANIZATION
Proceedings of International ISKO Conference - ISSN: 0938-5495 - 2026 - São Paulo

Knowledge Organization for Literary Datafication: A Ten-Stage Workflow for Semantic Nanopublication Applied to Selected Works by Camilo Castelo Branco

Henrique Monteiro Cristovão
Department of Archival Science, Graduate Program in Information Science, Federal University of Espírito Santo, Brazil

Carlos Guardado da Silva
Faculty of Letters, Centre for Classical Studies, University of Lisbon, Portugal

Abstract

Camilo Castelo Branco's works are widely available in digital form and remain central to Portuguese literary heritage, yet digital availability does not by itself yield structured, auditable, and interoperable narrative data. This short paper presents a ten-stage methodological workflow, developed within an ongoing research project, for converting digitized literary works into traceable narrative data and semantically publishable claims, using selected works by Camilo Castelo Branco as a case of application. The workflow combines corpus delimitation, metadata modeling, LLM-assisted schema-guided extraction, manual verification, data cleaning, rule-based movement derivation, network modeling, information network analysis, geospatial enrichment, cartographic visualization, and semantic publication through nanopublications. Each extracted record was required, whenever available, to include chapter number, page number, and textual excerpt; human verification was conducted on randomly selected samples. The contribution lies in framing literary datafication as a Knowledge Organization process in which metadata, controlled vocabularies, evidence fields, curation procedures, analytical models, and nanopublications operate as complementary devices for transforming digitized literary heritage into reusable, provenance-aware information resources.

Keywords: Camilo Castelo Branco; literary datafication; information network analysis; nanopublications; literary cartography; literary tourism.

1.0 Introduction

Camilo Castelo Branco occupies a central place in nineteenth-century Portuguese literature and continues to attract renewed scholarly, editorial, and patrimonial attention (Sobral et al. 2024). Although a substantial part of his work is available through critical editions, national repositories, and open digital collections, digital access does not automatically produce structured, auditable, and interoperable narrative data. Literary works may be digitized, indexed, or incorporated into cultural mediation initiatives and still lack traceable representations of characters, relationships, events, places, and movements.

Recent research shows that semi-structured extraction from narrative texts is feasible, especially for character networks and social relations (Elson et al. 2010; Kubis 2021). LLM-based approaches expand this potential, but they also require validation, human curation, and explicit treatment of uncertainty (Celli and Mingazov 2024). At the same time, literary datafication cannot be reduced to quantification; it requires theoretically informed modeling and attention to the interpretive consequences of formalization (Piper 2018; Piper et al. 2021).

From a Knowledge Organization perspective, the problem is to build a controlled representational chain linking literary text, extracted evidence, tabular structures, transformation logs, analytical models, and reusable semantic outputs. This concern resonates with the informational bond, which extends archival contextual integrity into digital information environments (Pacheco and Silva 2023), and with literary cartography and literary tourism, since places and movements must first be formalized before they can be compared, mapped, and reused (Moretti 2007; Herbert 2001).

Nanopublications are small, machine-actionable publication units that separate an assertion from its provenance and publication information (Mons and Velterop 2009; Gray et al. 2026). For Knowledge Organization, they transform selected interpretive or analytical findings into citable, provenance-aware knowledge units, connecting representation, accountability, interoperability, and reuse. This orientation also dialogues with literary Linked Data research, such as BookSampo, which shows how semantic portals and knowledge graphs can support data-analytic literary research through interoperable semantic resources (Ahola et al. 2025).

The research question is: How can a traceable and interoperable methodological workflow be designed to convert digitized literary works into structured narrative data suitable for information network analysis, literary cartography, and publication as nanopublications? The objective is to present a ten-stage workflow for literary datafication and semantic publication through nanopublications, illustrated through selected works by Camilo Castelo Branco. Rather than reporting the full empirical interpretation of the corpus, this short paper focuses on the methodological architecture of the workflow as a contribution to Knowledge Organization.

2.0 Methodological Workflow

The workflow is organized as a single methodological chain to prevent disconnection between extraction, cleaning, analysis, mapping, and semantic publication. In this short paper, the ten stages are presented as a reusable methodological architecture for literary datafication and semantic publication, while selected works by Camilo Castelo Branco function as a case of application. Table 1 summarizes the stages, operational procedures, tools or technologies, and human validation mechanisms.

Table 1 Ten-stage workflow for literary datafication and semantic publication

Stage and output	**Operationalization and tools/technologies**	**Human validation and KO control**
1. Corpus delimitation	Selection according to research scope, documentary availability, narrative suitability, and comparative potential.	Inclusion/exclusion criteria; assessment of source reliability and suitability for extracting characters, events, places, and movements.
2. Work-level metadata	Metadata table with bibliographic, narratological, thematic, and access-related fields.	Verification of source identification, edition consistency, metadata values, and documentary links.
3. Schema-guided extraction	LLM-assisted generation of candidate Character, Relationship, and Event tables using predefined schemas, controlled vocabularies, and evidence fields.	Each record must include, whenever available, chapter number, page number, and textual excerpt; sample-based manual review of randomly selected records.
4. Cleaning and normalization	OpenRefine-based cleaning, clustering, normalization, disambiguation, and JSON transformation history.	Correction of duplicate entities, variant names, inconsistent values, ambiguous labels, and deviations from schemas and vocabularies.
5. Movement derivation	Rule-based derivation of origin-destination movements from ordered character-linked events using a Python algorithm.	Control that movements are derived from event order and spatial references, not treated as directly extracted facts.
6. Network modeling	Preparation of curated tables in OpenRefine and conversion into GML network structures.	Control of node and edge semantics, aggregation level, directionality, weighting, and consistency with tabular data.
7. Information network analysis	Network inspection and analysis in Gephi, including centrality, brokerage, communities, cohesive patterns, and recurrent trajectories.	Interpretation restricted to patterns supported by curated data, modelling choices, and corpus scope.
8. Geospatial	Reconciliation of	Manual assessment of candidate

Stage and output	Operationalization and tools/technologies	Human validation and KO control
enrichment	place/location references through OpenRefine services, with coordinates, identifiers, and matching-status fields.	matches, approximate identifications, ambiguous references, and uncertain geographic alignments.
9. Cartographic visualization	Georeferenced visualization of places, locations, and trajectories in QGIS.	Visual inspection of spatial consistency, coordinates, aggregation, and cartographic plausibility.
10. Semantic publication as nanopublications	RDF/TriG representation of selected findings according to assertion, provenance, and publication-info layers.	Selection of claims only when scope, evidence, analytical derivation, and publication metadata can be documented.

In the Camilo Castelo Branco application, Stage 1 resulted in a purposive corpus of 33 works. This corpus is not a statistically representative sample of the author's complete production, but a methodological corpus selected to test the workflow under conditions of literary-cultural relevance, documentary availability, narrative density, suitability for extracting characters, events, places, and movements, and potential for cross-work comparison.

The LLM was used to generate candidate records, not final scholarly annotations. Each extracted record was required, whenever available, to include chapter number, page number, and textual excerpt, preserving a direct evidential link to the literary text. Because exhaustive verification would be impracticable at corpus scale, human verification was conducted on randomly selected samples, focusing on textual traceability, identifiers, event boundaries, relationship categories, spatial references, evidence fields, uncertainty markers, and controlled-vocabulary values.

Stages 5–9 transform curated records into analytical and geospatial structures: movements are derived through a Python algorithm; network data are prepared in OpenRefine and exported in GML; network analysis is conducted in Gephi; geospatial enrichment uses OpenRefine reconciliation services; and maps are produced in QGIS. Stage 10 formalizes selected findings as nanopublications. Its function is not merely to export results in RDF/TriG, but to reorganize selected findings as accountable knowledge units whose assertion, provenance, and publication information can be inspected independently (Mons and Velterop 2009; Gray et al. 2026). When explicit provenance modelling is required, this layer can be aligned with PROV-O (Lebo et al. 2013). In KO terms, nanopublications extend the workflow from data organization to semantic dissemination and reuse, also resonating with discussions on documentary granularity and monographic representation (Corrêa and Sales Marques 2025).

3.0 Methodological scope and Knowledge Organization contribution

This short paper deliberately focuses on the methodological and infrastructural contribution of the workflow rather than on the full empirical interpretation of the corpus. In the Camilo Castelo Branco application, the workflow generated a work-level metadata table, narrative tables organized by characters, relationships, and events, consolidated curated datasets, movement records derived from ordered event sequences, and network-ready GML files. These outputs document a representational chain from literary evidence to reusable analytical and semantic structures.

The contribution can be framed within Knowledge Organization because the workflow makes explicit the representational devices that mediate between literary text and reusable data. Metadata contextualizes the works (Gilliland 2016; Zeng and Qin 2022); controlled vocabularies define categories for relationships and events; evidence fields preserve links to textual passages; cleaning logs document transformations; network models formalize relations among entities; and nanopublications reorganize selected findings as provenance-aware semantic claims. In this configuration, KO is the organizing principle through which narrative evidence becomes inspectable, comparable, interoperable, and reusable, in line with semantic and interoperability concerns in KO (Hjørland 2007, 2018; Zeng 2019).

4.0 Final remarks

This short paper presented a ten-stage workflow for converting digitized literary works into structured narrative data and citable semantic assertions. Although illustrated through selected works by Camilo Castelo Branco, its methodological scope is broader: it can support literary datafication in other corpora where characters, relationships, events, places, and movements are embedded in unstructured or semi-structured textual evidence. The paper deliberately limits its scope to the workflow architecture and its Knowledge Organization implications. The broader research is still in progress, especially regarding the expansion, validation, and empirical interpretation of network, geospatial, and nanopublication outputs. The main result is therefore a traceable, KO-oriented workflow connecting literary interpretation, data curation, network analysis, spatial representation, and semantic publication.

Funding sources

This work was financed with National Funds through FCT (Foundation for Science and Technology), through the project UID/00019/2025, Centro de Estudos Clássicos (https://doi.org/10.54499/UID/00019/2025); and was supported by the National Council for Scientific and Technological Development (CNPq) and the Espírito Santo Research and Innovation Foundation (FAPES) [grant number 1011/2025].

Declaration of AI use

OpenAI ChatGPT was used to support schema-guided extraction of candidate records from literary texts and to assist with language editing, including grammar correction and manuscript refinement. The methodological use of AI-assisted extraction is described in the Methodological Workflow section. All AI-assisted outputs were reviewed, curated, and interpreted by the authors, who take full responsibility for the final content of the paper.

References

Ahola, Aleksi, Tuomas Peura, and Eero Hyvönen. 2025. "Using Linked Data for Data Analytic Literary Research: Case BookSampo—Finnish Fiction Literature on the Semantic Web." Journal of the Association for Information Science and Technology 76: 937–958. https://doi.org/10.1002/asi.24984 .

Celli, Fabio, and Dmitry Mingazov. 2024. "Knowledge Extraction from LLMs for Scalable Historical Data Annotation." Electronics 13, no. 24: 4990. https://doi.org/10.3390/electronics13244990 .

Corrêa, José Gustavo Moura, and Luana Farias Sales Marques. 2025. "O princípio monográfico de Paul Otlet e as nanopublicações: tecendo relações teóricas." In Anais do XXIV Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação (XXIV ENANCIB), vol. 4, no. 1. UFES; ANCIB. https://enancib.ancib.org/enancib/pt_BR/article/view/984 .

Elson, David K., Nicholas Dames, and Kathleen R. McKeown. 2010. "Extracting Social Networks from Literary Fiction." In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, edited by Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre, 138–147. Uppsala, Sweden: Association for Computational Linguistics. https://aclanthology.org/P10-1015/ .

Gilliland, Anne J. 2016. "Setting the Stage." In Introduction to Metadata, 3rd ed., edited by Murtha Baca. Los Angeles: Getty Research Institute.

Gray, Alasdair J. G., Kees Burger, Spyros Kotoulas, Antonis Loizou, Valery Tkachenko, Andra Waagmeester, Sune Askjaer, et al. 2026. Nanopublication Guidelines. Working draft. https://nanopub.net/guidelines/working_draft/ .

Herbert, David. 2001. "Literary Places, Tourism and the Heritage Experience." Annals of Tourism Research 28: 312–333. https://doi.org/10.1016/S0160-7383(00)00048-7 .

Hjørland, Birger. 2007. "Semantics and Knowledge Organization." Annual Review of Information Science and Technology 41: 367–405.

Hjørland, Birger. 2018. "Data (with Big Data and Database Semantics)." Knowledge Organization 45: 643–652.

Kubis, Marek. 2021. “Quantitative Analysis of Character Networks in Polish 19th- and 20th-Century Novels.” Digital Scholarship in the Humanities 36, suppl. 2: ii175–ii181. https://doi.org/10.1093/llc/fqab012 .

Lebo, Timothy, Satya Sahoo, and Deborah McGuinness, eds. 2013. PROV-O: The PROV Ontology. W3C Recommendation. World Wide Web Consortium. https://www.w3.org/TR/2013/REC-prov-o-20130430/ .

Mons, Barend, and Jan Velterop. 2009. “Nano-Publication in the e-Science Era.” In Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009), edited by Tim Clark, Joanne S. Luciano, M. Scott Marshall, Eric Prud’hommeaux, and Susie Stephens. CEUR Workshop Proceedings 523. https://ceur-ws.org/Vol-523/Mons.pdf .

Moretti, Franco. 2007. Graphs, Maps, Trees: Abstract Models for Literary History. London and New York: Verso.

Pacheco, André, and Carlos Guardado da Silva. 2023. “From the Archival Bond to the Informational Bond.” Preservation, Digital Technology & Culture 52: 49–58. https://doi.org/10.1515/pdtc-2023-0004 .

Piper, Andrew. 2018. Enumerations: Data and Literary Study. Chicago: University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/E/bo28465405.html .

Piper, Andrew, Richard Jean So, and David Bamman. 2021. “Narrative Theory for Computational Narrative Understanding.” In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, edited by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, 298–311. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.26 .

Sobral, Cristina, Serafina Martins, and Carlota Pimenta, eds. 2024. Estados da Arte sobre Camilo Castelo Branco. Lisboa: Imprensa Nacional. https://imprensanacional.pt/wp-content/uploads/2025/02/EstudosCamilianos_IN.pdf?btn=red .

Zeng, Marcia Lei. 2019. “Interoperability.” Knowledge Organization 46: 122–146.

Zeng, Marcia Lei, and Jian Qin. 2022. Metadata. 3rd ed. Chicago: Neal-Schuman.