

Human-Centered TEI Encoding as Infrastructure for Meaning-Driven AI in Portuguese Literature

Diego Emanuel Giménez¹  

¹ University of Macau

Abstract

This paper argues that the alignment of artificial intelligence with human values begins not at the level of algorithms but at the level of knowledge modeling. Drawing on the ongoing TEI encoding of *Os Lusíadas* by Luís de Camões, it introduces metadata as a form of epistemic governance that structures what AI systems can recognize, process, and reproduce as knowledge. Rather than positioning AI as an interpretative agent, the study frames it as operating within a semantic architecture designed by human editors. TEI encoding is therefore approached not merely as technical markup but as an interpretative infrastructure in which scholarly values, transparency, plurality, and cultural responsibility, are formalized into structured data. The paper proposes a human-centered workflow in which AI assists editorial practices through schema-constrained suggestions, consistency checks, and semantic pattern detection while preserving human authority over interpretative decisions. Attention is given to the ethical implications of encoding a canonical early modern epic whose textual tradition intersects with mythology, religion, and historical memory. By conceptualizing metadata as a mechanism of epistemic governance, this study shifts the discussion of meaning-driven AI upstream and argues that responsible machine participation in cultural knowledge depends primarily on the design of metadata environments rather than on post-hoc algorithmic correction.

1. Introduction

Artificial intelligence has renewed scholarly attention to metadata as a mediating layer between computational systems and human meaning-making. While current debates often frame AI alignment as a downstream technical challenge, this paper argues that alignment can also begin upstream — at the level of knowledge modeling.

Metadata structures define what can be identified, categorized, and operationalized within a dataset. In this sense, they function as mechanisms of epistemic governance, shaping the conditions under which AI systems participate in knowledge production. The term "epistemic governance" is used here in a descriptive sense to refer to the structuring of knowledge through metadata, rather than in a normative sense.

The digital encoding of literary texts offers a particular environment for examining this relationship. Unlike purely descriptive metadata, Text Encoding Initiative (TEI) markup embeds scholarly interpretation directly into structured data.

Using the ongoing TEI encoding of *Os Lusíadas* as a case study, this paper proposes that human-centered metadata environments can serve as infrastructures for meaning-driven AI.

The contribution of this paper is primarily conceptual and methodological. It does not propose a fully automated encoding system or a complete description of the project schema. Rather, it argues that TEI-based semantic modeling, authority reconciliation, and AI-assisted validation can operate together as an infrastructure of epistemic gover-

Short Paper

Available Aug 01, 2026

Correspondence to
Diego Emanuel Giménez
dgimenez@um.edu.mo



Copyright © 2026 Giménez. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license, which enables reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator.

nance, in which machine participation is constrained by human-defined interpretative categories.

2. Metadata and Epistemic Governance

Metadata is often treated as technical scaffolding for information retrieval. However, in scholarly editing it performs a far more consequential role: it formalizes interpretative categories and constrains the conditions under which knowledge can be represented and processed.

Encoding decisions establish textual hierarchies, semantic distinctions, ontological commitments, and boundaries of interpretability. AI systems do not operate on texts themselves but on these epistemically governed representations.

Within the TEI framework [1], these commitments become explicit through schema design and encoding practices. Metadata is therefore not neutral but reflects specific epistemological positions (Gilliland [2]). As Coyle notes in "Metadata Models of the World Wide Web," XML provides rules for encoding documents and data electronically, whereas machine-actionable semantics depend on RDF, persistent identifiers, and explicit relations among entities [3].

This approach builds on existing traditions in digital scholarly editing, where encoding is not merely a mechanism for transcription but a means of representing textual structure, interpretation, and editorial responsibility [4]. In relation to metadata studies, the paper extends the view of metadata as a descriptive or retrieval-oriented apparatus by emphasizing its role in shaping the conditions under which AI systems encounter literary knowledge. The case of *Os Lusíadas* therefore connects TEI-based scholarly modeling with current debates on machine-actionable metadata, authority control, and responsible AI.

Meaning-driven AI depends less on model sophistication than on the semantic richness of the metadata structures that constrain it.

3. Case Study: Encoding *Os Lusíadas*

As a foundational work of Portuguese literature, *Os Lusíadas* presents interpretative challenges that exceed structural markup. Its textual fabric intersects with mythology, historical discourse, religion, and early modern cosmology.

The encoding project adopts a taxonomy-driven approach to entity annotation, explicitly modeling categories of referents within the TEI header:

```
<taxonomy xml:id="entityType">
  <category xml:id="historical">
    <catDesc>Figuras históricas reais</catDesc>
  </category>
  <category xml:id="mythological">
    <catDesc>Personagens da mitologia</catDesc>
  </category>
  <category xml:id="ethnicity">
```

```

    <catDesc>Grupos étnicos ou povos</catDesc>
  </category>
  <category xml:id="religion">
    <catDesc>Conceitos e entidades religiosas</catDesc>
  </category>
</taxonomy>

```

This taxonomy governs the semantic interpretation of entities encoded throughout the text. Distinctions such as mythological, historical, ethnic, and religious entities are not intrinsic to the text but emerge from editorial modeling.

Such encoding constructs a semantic ontology through which the text becomes machine-readable and creates better conditions for AI systems to process and interpret the data. The following examples illustrate how locally encoded entities are linked to Wikidata identifiers:

```

<persName type="mythological" ref="https://www.wikidata.org/entity/Q
3954">Neptuno</persName>
<persName type="historical" ref="https://www.wikidata.org/entity/Q
7328">Gama</persName>
<persName type="religion" ref="https://www.wikidata.org/entity/Q302">Jesu
Cristo</persName>
<rs type="ethnicity" ref="https://www.wikidata.org/entity/Q
837549">Lusitanos</rs>
<placeName ref="https://www.wikidata.org/entity/Q80338">Bengala</
placeName>

```

These links are not generated or accepted automatically. Epistemic governance is implemented through the combination of a TEI taxonomy, controlled entity annotation, and an authority-control register. This register records the local textual form, the corresponding TEI element, the semantic type assigned by the editor, the Wikidata identifier, the authority label, and possible editorial notes. It functions as a reconciliation and remediation layer between the literary text, the TEI document, and external knowledge graphs. It ensures that ambiguous or culturally complex cases are reviewed, adjusted, or documented before they are exposed to downstream computational workflows. This linkage transforms the TEI document from a closed encoding system into an interoperable semantic resource.¹

The use of persistent identifiers from Wikidata enables the alignment of locally encoded entities with external knowledge graphs, allowing AI systems to operate across datasets while preserving semantic distinctions defined in the encoding.

4. Toward a Human-Centered AI-Assisted Workflow

Rather than treating AI as an interpretative authority, this study conceptualizes it as operating within a human-designed semantic framework.

A semi-automatic workflow is proposed in which AI supports editorial practices through five controlled operations:

¹At the current stage of the project, the encoding guidelines and entity reference record function as internal documentation; selected sample files and documentation are being prepared for public release after editorial validation.

- Structural segmentation: AI suggests canto, stanza, and verse divisions; the editor reviews, adjusts, and validates the final structure.
- Entity detection: AI suggests candidate persons, places, mythological figures, and religious entities; the editor reviews, confirms, rejects, or adds entities.
- Semantic classification: AI suggests @type values such as historical, mythological, religion, ethnicity, or geographical; the editor reviews, corrects, adds, and assigns the final taxonomy.
- Reconciliation: Because automatically suggested authority links may be unreliable, the editor searches Wikidata, checks candidate entities against scholarly literature, verifies the authority record, and documents ambiguous cases.
- Consistency checking: AI detects divergent encodings of the same entity; the editor normalizes the encoding or records the reason for variation.

These tasks can support processes such as entity disambiguation, consistency checking across encoding layers, and semantic alignment with external knowledge graphs. Rather than generating interpretations, AI operates within constraints defined by the metadata and encoding schema so that it ensures alignment with editorial decisions.

For this reason, the project conceptualizes AI-readiness in terms of metadata rather than exclusively as a technical export issue. Responsible reuse of TEI-derived datasets in downstream AI workflows depends on documentation that includes structure, provenance, licensing, and editorial constraints.

This approach resonates with critiques of large language models as lacking genuine understanding (Bender et al. [5]), which reinforces the need for structured semantic guidance.

5. Ethical and Cultural Implications

Encoding *Os Lusíadas* involves representing a text embedded in historical, mythological, and religious discourses. Without structured metadata, AI systems risk flattening these distinctions and reproducing reductive interpretations.

Human-centered metadata mitigates these risks by making interpretative decisions explicit, auditable, and contextually grounded.

This aligns with broader frameworks for ethical AI that emphasize transparency and accountability (Floridi et al. [6]): "the need to understand and hold to account the decision-making processes of AI".

6. Discussion: Alignment Begins in Metadata

This paper examines how meaning-driven AI can be conceptualized in the context of literary works. It argues that alignment should not be treated as a corrective layer applied after model deployment, but as a consequence of epistemic design.

Metadata defines the space of possible meanings within which AI systems operate. The TEI encoding of *Os Lusíadas* demonstrates how scholarly modeling constrains machine behavior while preserving interpretative complexity.

In this framework, meaning-driven AI emerges from semantic infrastructures rather than from autonomous systems.

7. Conclusion

As AI becomes embedded in scholarly workflows, the question is no longer whether machines will participate in knowledge production, but under what epistemic conditions.

Human-centered metadata provides such conditions by embedding scholarly values directly into structured data. TEI encoding can thus be understood as a form of epistemic governance that aligns computational processes with human interpretative frameworks.

Acknowledgements

Project "Portuguese Literary Studies: Texts, Readings, and Digital Approaches" (SRG2026-00007-FAH), University of Macau; Camões Lab – University of Coimbra, University of Macau, Calouste Gulbenkian Foundation, "la Caixa" Foundation, Imprensa Nacional-Casa da Moeda. I also thank Bruno Ministro and the members of the Camões Lab team for discussions related to the entity encoding process.

References

- [1] TEI Consortium, "TEI P5: Guidelines for Electronic Text Encoding and Interchange," 2023. [Online]. Available: <https://tei-c.org/>
- [2] A. J. Gilliland, "Setting the Stage," in *Introduction to Metadata*, Getty Research Institute, 2016.
- [3] K. K. Coyle, "Metadata Models of the World Wide Web," in *RDA Vocabularies for a Twenty-First-Century Data Environment*.
- [4] J. Flanders and F. Jannidis, *The Shape of Data in Digital Humanities*. Routledge, 2019.
- [5] E. M. Bender and et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*. in Proceedings of FAccT 2021.
- [6] L. Floridi and et al., "AI4People—An Ethical Framework for a Good AI Society," *Minds and Machines*, vol. 28, 2018.