Metadata & Discovery Group

Envisioning Dante

a metadata journey from Inferno to Paradiso

Case study: *Le terze rime di Dante*: from Alma MARC (18873) to MDC TEI (PR-ALDI-18873)

Ourania Karapasia

Digital Metadata Specialist, Metadata & Discovery, Collection Strategies Directorate, The University of Manchester Library

Received: 11 August 2025 | Published: 22 September 2025

ABSTRACT

Through the case study of Le terze rime di Dante (PR-ALDI-18873), this article charts the progression from a MARC bibliographic record in Ex Libris Alma | Primo to an enriched TEI-XML file that forms a meaningful digital object in Manchester Digital Collections (MDC). The case study outlines the automated and editorial stages of the workflow, highlights key modelling decisions and challenges, and reflects on TEI's value as a flexible, expressive framework for describing digitised early printed books whose complexity exceeds the limits of standard bibliographic control. It demonstrates that while automation enables efficiency, human-led editorial intervention remains essential to produce accurate, semantically rich, and reusable metadata for sustainable digital collections.

KEYWORDS metadata transformations; early printed books; TEI-XML metadata; MARC to TEI; Dante Alighieri

CONTACT Ourania Karapasia ourania.karapasia@manchester.ac.uk The University of Manchester Library

On 29 May 2025, the first instalment of the Dante Digital Library[1] was published on Manchester Digital Collections (MDC)[2]. This major digital resource features some of the rarest and most significant early printed editions of Dante Alighieri's *Divine Comedy*, many of which are being made available digitally for the first time. The release coincided with a two-day academic conference held at The John Rylands Research Institute and Library, home to the original volumes now featured in the digital collection.

The Dante Digital Library forms a central output of *Envisioning Dante, c. 1472–c. 1630: Seeing and Reading the Early Printed Page*, a digital humanities research project funded by the Arts and Humanities Research Council (AHRC) and based at The University of Manchester. Encompassing the digitisation of 99 editions printed between 1472 and

[1] https://www.digitalcollections.manchester.ac.uk/collections/dante

[2] https://www.digitalcollections.manchester.ac.uk/

1629, the project represents both a major scholarly resource and a significant advancement in metadata practices for early printed materials in digital collections. The first release features 20 editions, with further material scheduled for release after 2025.

Beyond its scholarly contributions, the Dante Early Printed collection[3] within the Dante Digital Library also served as a platform for experimentation and refinement in metadata transformation. At the heart of this work lies the transformation of MARC 21 records into structured TEI-conformant files, aligning with FAIR principles to support metadata that is *findable*, *accessible*, *interoperable* and *reusable*.

This article uses *Le terze rime di Dante* (PR-ALDI-18873) as a case study and examines that transformation in detail. It considers the technical workflow, editorial modelling, and key TEI decisions, and reflects on the challenges, outcomes, and possibilities for further development. Through this example, it demonstrates how TEI can serve as a flexible and semantically rich framework for representing the descriptive complexity of digitised early printed books beyond the reach of conventional bibliographic control.

From MARC to TEI: overview of the workflow

The process begins with a MARC 21 record catalogued in Alma, which is exported to MARCXML using MarcEdit. An automated script then generates a preliminary TEI-conformant file, based on a template aligned with MDC conventions. This TEI file is committed to the `mdc-contents` Bitbucket repository via GitHub Desktop, enabling version control and collaborative editing. A Python script enriches the file with references to digital images, utilising **`<facsimile>`**, **`<surface>`**, and **`<graphic>`**. The result, a complete digital object (facsimile files plus TEI metadata), is then uploaded to the MDC test environment for internal review. At this point, the TEI file represents the initial structured output of the conversion workflow. Following conversion, the TEI file undergoes editorial refinement in Oxygen XML Editor. A spreadsheet-based tool is used to review the output and apply targeted encoding adjustments in line with MDC's TEI guidelines, to produce a publication-ready file.

For brevity, from here on MARC 21 is referred to as MARC, and TEI-XML and TEI-conformant file(s) as TEI file(s).

Stage 1: Conversion

From Primo | Alma MARC to MDC | TEI: mapping metadata

The conversion stage involves the automated mapping of MARC fields to corresponding TEI elements. While MARC already limits semantic expression, the conversion process further flattens distinctions between types of bibliographic description and contributor roles. The following examples illustrate how semantic detail is lost or obscured in practice.

[3] https://www.digitalcollections.manchester.ac.uk/collections/danteearlyprinted

General notes | MARC 500 → TEI <note>

In MARC, separate 500 fields are used to record discrete bibliographic observations. For example, collation, typographic features, or editorial responsibility, each appear “independently” in Alma and in Primo’s public interface[4].

Description Signatures: a-z^{8} A-F^{8} G^{12}. Leaf l2 is a blank.
A Lyonese counterfeit of the Aldine edition of 1502, omitting Aldus' preface.
No named printer, printer’s device, date or colophon.
Imprint from UCLA Online Catalog.
Printed in italic type; capital spaces with guide letters at the beginning of the three parts; no catchwords.
Without preface or notes.
On title-page verso: Lo'nferno e'l Purgatorio e'l Paradiso di Dante Alaghieri.
Edited by Pietro Bembo.

```
500 __ |a Signatures: a-z⁸ A-F⁸ G¹².  Leaf l2 is a blank.
500 __ |a A Lyonese counterfeit of the Aldine edition of 1502, omitting
Aldus' preface.
500 __ |a No named printer, printer’s device, date or colophon.
500 __ |a Imprint from UCLA Online Catalog.
500 __ |a Printed in italic type; capital spaces with guide letters at
the beginning of the three parts; no catchwords.
500 __ |a Without preface or notes.
500 __ |a On title-page verso: Lo'nferno e'l Purgatorio e'l Paradiso di
Dante Alaghieri. owner. |5 UkMaJRU
500 __ |a Edited by Pietro Bembo.
```

During conversion, these fields are merged into a single composite structure. In the TEI output, all notes are rendered within a single **<note>**, structured only by multiple undifferentiated **<p>**.

• Note(s):
Signatures: a-z^{8} A-F^{8} G^{12}. Leaf l2 is a blank.
A Lyonese counterfeit of the Aldine edition of 1502, omitting Aldus' preface.
No named printer, printer’s device, date or colophon.
Imprint from UCLA Online Catalog.
Printed in italic type; capital spaces with guide letters at the beginning of the three parts; no catchwords.
Without preface or notes.
On title-page verso: Lo'nferno e'l Purgatorio e'l Paradiso di Dante Alaghieri.
Edited by Pietro Bembo.

[4] https://www.librarysearch.manchester.ac.uk/permalink/44MAN_INST/1lr7mpn/alma9912616334401631 The fonts used in Alma are not encoded in XML. They pertain to the visual presentation layer and are not part of the structured metadata. As such, they are not retained or represented in the XML outputs, which focus solely on the underlying bibliographic data.

```
<note>
<p>Signatures: a-z⁸ A-F⁸ G¹². Leaf l2 is a blank.</p>
<p>A Lyonese counterfeit of the Aldine edition of 1502, omitting Aldus'
preface.</p>
<p>No named printer, printer’s device, date or colophon.</p>
<p>Imprint from UCLA Online Catalog.</p>
<p>Printed in italic type; capital spaces with guide letters at the
beginning of the three parts; no catchwords.</p>
<p>Without preface or notes.</p>
<p>On title-page verso: Lo'nferno e'l Purgatorio e'l Paradiso di Dante
Alaghieri.</p>
<p>Edited by Pietro Bembo.</p>
</note>
```

Although syntactically valid, this structure reproduces the MARC display logic without reintroducing any semantic differentiation. Collation, publication context, and descriptive notes are flattened into generic prose, limiting interpretability and complicating downstream processes such as filtering, indexing. These notes require disaggregation and more detailed encoding in Stage 2.

Personal names and roles | MARC 100, 700 → TEI <author>|<editor>

In MARC, field 100 records the primary creator of a work, typically an author, while field 700 records additional individuals who have contributed in other capacities. Relator terms in subfield ‡e, such as ‡e editor, ‡e printer, or ‡e former owner, specify the nature of each individual’s role. Subfield ‡5 indicates that the information applies only to a specific institutional copy. Together, these fields and subfields help identify who created, contributed to, or owned a given copy of a work.

Name	Dante Alighieri, 1265-1321, author. Manuzio, Aldo, 1449 or 1450-1515, associated name. Bembo, Pietro, 1470-1547, editor. Gabiano, Balthazard de, -approximately 1517, printer. Mequignon, Cavaliere, former owner. Soma. Luigi, Magioniere. former owner. Christie, Richard Copley, 1830-1901, former owner. Spencer, George John Spencer, Earl, 1758-1834, former owner. Fréret, Nicolas, 1688-1749, former owner. Nicolet, Jean Baptiste Thomas. former owner.

```
100 0_ |a Dante Alighieri, |d 1265-1321, |e author.

700 1_ |a Manuzio, Aldo, |d 1449 or 1450-1515, |e associated name.
700 1_ |a Bembo, Pietro, |d 1470-1547, |e editor.
700 1_ |a Mequignon, |c Cavaliere, |e former owner. |5 UkMaJRU
700 1_ |a Soma. Luigi, |c Magioniere. |e former owner. |5 UkMaJRU
```

```
700 1_ |a Christie, Richard Copley, |d 1830-1901, |e former owner. |5
UkMaJRU
700 1_ |a Spencer, George John Spencer, |c Earl, |d 1758-1834, |e
former owner. |5 UkMaJRU
700 1_ |a Fréret, Nicolas, |d 1688-1749, |e former owner. |5 UkMaJRU
700 1_ |a Nicolet, Jean Baptiste Thomas. |e former owner. |5 UkMaJRU
```

However, MARC's structure limits copy-level specificity. When multiple former owners are recorded using subfield ‡5, such as ‡5 UkMaJRU, it implies that all are associated with the same institutional copy. If more than one copy of the work exists, MARC cannot reliably indicate which name corresponds to which copy, as in the case of copy R213785.

In TEI the **<author>** is reserved for primary authorship derived from MARC field 100. All other contributors from field 700 are rendered as **<editor>**, with **@role** preserving the relator term through its corresponding relator code: **role="prt"** for printer, **role="fmo"** for former owner.

• **Author(s):** Dante Alighieri, 1265-1321
• **Editor(s):** Bembo, Pietro, 1470-1547
• **Printer(s):** Gabiano, Balthazard de, -approximately 1517
—
• **Former Owner(s):** Mequignon, Cavaliere; Soma. Luigi, Magioniere.; Christie, Richard Copley, 1830-1901; Spencer, George John Spencer, Earl, 1758-1834; Fréret, Nicolas, 1688-1749; Nicolet, Jean Baptiste Thomas.
—
• **Associated Person(s):** Manuzio, Aldo, 1449 or 1450-1515; Gabiano, Balthazard de, -approximately 1517

```
<author key="person_v00000000" ref=" http://viaf.org/viaf/
00000000">Dante Alighieri, 1265-1321, author.</author>
<editor role="asn">Manuzio, Aldo, 1449 or 1450-1515</editor>
<editor role="edt">Bembo, Pietro, 1470-1547</editor>
<editor role="prt">Gabiano, Balthazard de, -approximately 1517</editor>
<editor role="prt">Gabiano, Balthazard de, -approximately 1517</editor>
<editor role="fmo">Mequignon, Cavaliere</editor>
<editor role="fmo">Soma. Luigi, Magioniere.</editor>
<editor role="fmo">Christie, Richard Copley, 1830-1901</editor>
<editor role="fmo">Spencer, George John Spencer, Earl, 1758-1834</
editor>
<editor role="fmo">Fréret, Nicolas, 1688-1749</editor>
<editor role="fmo">Nicolet, Jean Baptiste Thomas.</editor>
```

This results in a semantic collapse: editorial, physical production, and provenance roles are conflated under **<editor>**, an element conventionally associated with editorial responsibility. While encoding printers as **<editor role="prt">** is justifiable in some contexts, extending the same structure to former owners and associated names conflates fundamentally distinct roles.

Although accepted in Stage 1 as a pragmatic interim measure, this approach has clear limitations. In Stage 2, only roles relevant to this copy are retained. Former owners will be encoded using more appropriate elements, such as **<provenance>** within **<history>**, while references to other copies will be removed.

The absence of authority-controlled identifiers, such as VIAF IDs, limits the file's capacity for disambiguation and interoperability. Where available, they will be added in later stages to support consistent, reusable metadata.

These two examples exemplify broader challenges in automated conversion: flattened descriptive content, ambiguous role encoding, and copy-specific metadata collapse into generic structures. Related MARC fields such as **561**, **562**, and **563** are mapped to high-level TEI wrapper elements such as **<provenance>**, **<physDesc>**, and **<bindingDesc>**, each containing generic **<p>**.

From syntax to semantics

At first glance, the TEI file derived from the MARC record appears to be a successful conversion. The XML is valid, well-formed, and legible. However, this surface-level success conceals a deeper issue: the file conforms to XML syntax but reproduces the logic of MARC rather than engaging TEI's modelling potential. Descriptive richness is often lost in translation.

In effect, the TEI file reproduces the structure of the Alma catalogue record. It presents MARC-encoded metadata reformatted using TEI elements, without engaging TEI's underlying model. Although the XML is technically sound, it lacks the interpretative structure that gives TEI its descriptive power. XML provides structural rules; TEI offers a modelling vocabulary that defines how parts of a resource function and relate. This supports meaningful, contextual representation.

TEI is not just a structural wrapper for metadata. It is a modelling language designed to express not only what a source is, but also how and why it holds meaning in historical, material, and intellectual terms. The automated conversion follows the letter of the standard but not its intent. Valid XML does not, in itself, constitute valid TEI in any intellectually rigorous sense. The challenge lies not in conversion but in interpretation. It is not a matter of syntax, but of semantics.

Stage 2: Transformation

TEI editorial refinement and semantic modelling

The transformation stage introduces both standardisation and interpretive encoding. At this point, MARC-derived metadata is aligned with TEI not only structurally but also conceptually. The focus shifts from syntax to semantics, from automated output to expert-led modelling.

Editorial refinement addresses structure, terminology, and descriptive content to ensure clarity, consistency, and accuracy. It corrects misaligned mappings, applies normalisation rules, and disaggregates descriptive, structural, and copy-specific information, ensuring that each element conveys precisely the relationship it is meant to express.

A key challenge identified in Stage 1 was the generic mapping of MARC fields to high-level TEI wrapper elements, which prompted the remapping of fields to more semantically appropriate TEI elements and the re-encoding of their content to accurately reflect the structure and meaning of the source. This work was underpinned by a spreadsheet-based mapping tool. It captures element-level markup, editorial rationale, and proposed schema enhancements, while supporting metadata enrichment through authority control and linked data.

The following examples revisit the same cases presented in Stage 1, repeated here to illustrate how they are structurally and semantically transformed during Stage 2.

Personal names and roles | MARC 100, 700 → TEI `<author>|<editor>|<provenance>`

Stage 2 of the editorial workflow introduces a clearer, semantically grounded treatment of names imported from MARC records. In pre-editorial TEI, personal names, regardless of their role, were rendered as **`<editor>`**, distinguished only by relator codes such as **`edt`** (editor), **`prt`** (printer), **`asn`** (associated name), or **`fmo`** (former owner). This flattening obscured meaningful distinctions and often introduced inconsistencies, including duplication.

Contributor roles are disaggregated by assigning each name to the appropriate TEI element based on function: the author is re-encoded using **`<author>`**, while only printers and editors are encoded using **`<editor>`** but are distinguished using specific **`@role`** values. VIAF identifiers are added via **`@key`** to support authority control and disambiguation.

• **Author(s):** Dante Alighieri, 1265-1321
• **Editor(s):** Bembo, Pietro, 1470-1547
• **Printer(s):** Gabiano, Balthazard de, -approximately 1517
—
• **Former Owner(s):** Nicolet, Jean Baptiste Thomas; Fréret, Nicolas, 1688-1749; Spencer, George John Spencer, Earl, 1758-1834; Spencer, John Poyntz Spencer, Earl, 1835-1910; Rylands, Enriqueta, 1843-1908
—
• **Associated Person(s):** Manuzio, Aldo, 1449 or 1450-1515
—
• **Provenance:**
▪ Jean Baptiste Thomas Nicolet
Bookplate on the inner back cover: *Ex Libris Joannis Baptistæ Thomæ Nicolet*.
▪ Nicolas Fréret (1688-1749), historian and linguist
Possibly Nicolas Fréret. Manuscript inscription on the upper right corner of a1r: *N. Freret | # | 1710*.

```
<author key="viaf:97105654">Dante Alighieri, 1265-1321</author>
<editor key="viaf:54144140" role="edt">Bembo, Pietro, 1470-1547</
editor>
<editor role="prt">Gabiano, Balthazar de, -approximately 1517</editor>
<editor key="viaf:96325760" role="asn">Manuzio, Aldo, 1449 or 1450-
1515</editor>
<provenance>
<p> <name role="fmo"><persName type="display">Jean Baptiste Thomas
Nicolet</persName><persName type="standard">Nicolet, Jean Baptiste
Thomas</persName></name></p>
<p>Bookplate on the <locus from="Inner_back_cover" to="Inner_back_
cover">inner back cover</locus>: <hi rend="italic">Ex Libris Joannis
Baptistæ Thomæ Nicolet</hi>.</p>
</provenance>
<provenance>
<p><name key="viaf:34456779" role="fmo"><persName
type="display">Nicolas Fréret (1688-1749), historian and linguist</
persName><persName type="standard">Fréret, Nicolas, 1688-1749</
persName></name></p>
<p>Possibly Nicolas Fréret. Manuscript inscription on the upper right
corner of <locus from="a1r" to="a1r">a1r</locus>: <hi rend="italic">Ex
N. Freret | # | 1710 </hi>.</p>
 </provenance>
—
```

Former owners, previously conflated with contributors, are now correctly encoded in **<provenance>**, using **<name role="fmo">** with structured **<persName>** content.

Physical traces, such as bookplates, inscriptions, and annotations, are anchored via **<locus>** and **<hi>**, allowing evidence of ownership to be linked precisely to locations within the item. For example, Jean Baptiste Thomas Nicolet is associated with a bookplate on the inner back cover, while Nicolas Fréret is linked to a marginal inscription on folio a1r. These granular encodings preserve the material history of the object without distorting intellectual attribution.

General notes | MARC 500 → TEI <note>|<physDesc>|<additions>

A broad range of information, such as collation, edition history, typographic features, and editorial attributions, was previously grouped within a single **<note>** and embedded in **<physDesc>**, expressed through multiple **<p>**. This content is now distributed and re-encoded across semantically appropriate TEI elements.

General remarks about the edition's character, such as its identification as a Lyonese counterfeit of the Aldine edition, are retained in **<note>**, but only when they serve as interpretive or contextual commentary rather than formal description.

- **Note(s):**

A Lyonese counterfeit of the Aldine edition of 1502, omitting Aldus' preface.

```
<note>
<p>A Lyonese counterfeit of the Aldine edition of 1502, omitting Aldus'
preface.</p>
<note>
```

Within **<physDesc>**, collation details such as signature sequences and blank leaves are encoded to **<collation>**; layout and line count are expressed using **<layoutDesc>**; and typographic features, including typeface and guide letters, are captured in **<typeDesc>** and **<decoDesc>**. Where applicable, controlled vocabulary terms, such as material types from the Getty AAT identifiers, are added via **@key** to improve semantic precision and support interoperability.

- **Format:** Codex
- **Material(s):** Paper
- **Extent:** 244 leaves
- **Collation:** Signatures: a-z^{8} A-F^{8} G^{12}
  Leaf l2 is blank.
- **Layout:**

Printed in one column, 28-30 lines.

- **Typeface:**
  - Printed in *italic type*.
- **Decoration:**

Blank spaces for initials are marked with printed guide letters.

```
<physDesc>
<objectDesc form="codex">
<supportDesc material="paper">
<support><material key="aat:300014109">Paper</material>, folded in
<measure type="folding">inner 8vo (octavo)</measure>.</support>
<extent><measure quantity="244" type="leaf">244</measure> leaves</
extent>
<collation>Signatures: <signatures>a-z<hi rend="superscript">8</hi> A-
F<hi rend="superscript">8</hi> G<hi rend="superscript">12</hi></
signatures>
<p>Leaf <locus from="Blank_folio_[l2r]" to="Blank_folio_[l2r]">l2</
locus> is blank.</p>
</collation>
</supportDesc>
<layoutDesc>
<layout columns="1">
<p>Printed in one column, 28-30 lines.</p>
</layout>
</layoutDesc>
</objectDesc>
<typeDesc>
<typeNote scope="sole">Printed in <term>italic type</term>.</typeNote>
</typeDesc>
<decoDesc>
<decoNote type="initial">
<p>Blank spaces for initials are marked with printed guide letters.</
p>
</decoNote>
</decoDesc>
```

Copy-specific features, such as manuscript inscriptions and pencil annotations, previously embedded as multiple **<p>** within generic **<physDesc>** content, are now encoded in **<additions>**. The **<locus>** indicates their physical location within the item, while **<hi>** is used to preserve stylistic or visual emphasis.

• **Additions:**
Manuscript pencil note on flyleaf 1v: *Character of G. Huyon A | 1520?*.
Various manuscript pencil prices[?] on flyleaf 1v: *£2.2.0 | 0 | 3/[?] | 1/4/2[?]*.
Manuscript inscription, mostly illegible, on the lower margin of a1r: *[?]De[?]dras[?]*.

```
<additions>
<p>Manuscript pencil note on <locus from="Flyleaf_1v" to="Flyleaf_
1v">flyleaf 1v</locus>: <hi rend="italic">Character of G. Huyon A |
1520?</hi>.</p>
```

```
<p>Various manuscript pencil prices[?] on <locus from="Flyleaf_
1v" to="Flyleaf_1v">flyleaf 1v</locus>: <hi rend="italic">£2.2.0 | 0 |
3/[?] | 1/4/2[?]</hi>.</p>
<p>Manuscript inscription, mostly illegible, on the lower margin
of <locus from="a1r" to="a1r">a1r</locus>: <hi
rend="italic">[?]De[?]dras[?]</hi>.</p>
</additions>
```

Through these interventions, the TEI file moves beyond formal validity to become a meaningful digital object in MDC, representing the physical item through metadata, displayed alongside its digital facsimile.

Reflections on metadata conversion and transformation

The *Envisioning Dante* project provided a timely opportunity to trial and refine a metadata workflow not bound by the structural constraints of MARC. Drawing on both automation and editorial expertise, it explored scalable methods for producing semantically rich, structurally coherent TEI files for publication in MDC as digital objects.

In Stage 1, an automated conversion model was used to generate a baseline TEI from MARC metadata. This reliable and repeatable workflow reduced the need for manual intervention and enabled efficient early-stage processing of a large number of files.

In Stage 2, the spreadsheet-based mapping tool guided encoding decisions and established a uniform encoding practice, ensuring consistency across approximately one hundred TEI files. Used collaboratively by team members, it proved effective, sustainable, and adaptable within a distributed editorial workflow. Its successful application demonstrated that it supported a practical, repeatable process, well-suited to both current and future projects. Beyond refining individual files, the tool also contributed to the development of the MDC TEI guidelines, both informing and being shaped by editorial decisions grounded in practical encoding challenges.

TEI files were further enriched with persistent identifiers (e.g. VIAF, Getty TGN, and AAT), refined encoding of names and roles, and additional copy-specific descriptive features such as bindings, provenance, and annotations. These enhancements strengthened authority control and improved interoperability with linked data frameworks.

As part of this work, additional TEI elements were introduced into our practice to support the description of printed materials. Elements such as **<typeDesc>** and **<typeNote>** were used to encode and display as '*Typeface*', in parallel with **<handDesc>** and **<handNote>** for manuscripts displaying as '*Script*'.

A customised **<note type="printersDevice">** was added to encode and display as '*Printer's Device*', enabling the description of devices used by printers in early printed books.

To support the structured encoding of bibliographic formats such as '*folded in 8vo (octavo)'*, which is currently recorded using **<measure type="folding">** within **<support>**, a proposal was submitted to the TEI community to extend the schema to allow **<objectType>** as a valid child of **<objectDesc>**. This change should enable more precise representation of an item's functional or semantic category, along with its physical form.

In addition, the customised element **<note type="publisher">**, introduced earlier in local practice, was used to encode and display as '*Publication*', supporting the representation of publication, distribution, and related information (imprint), as recorded in MARC field 260 or 264.

The **<filiation>** was also employed to record bibliographic relationships between related printed editions. In this case, linking to the Aldine copy R213785 to reflect shared printing history or edition lineage.

Despite these successes, significant challenges emerged when working with MARC-derived metadata. Records created under varying cataloguing standards and levels of detail often exhibited semantic ambiguity, inconsistent field usage, and limited granularity. These issues frequently hindered automation and required careful editorial judgement to harmonise descriptive content and ensure alignment with MDC's standards for display and discovery.

This became particularly evident during the processing of the case study file, *PR-ALDI-18873*[5]. As the first TEI file generated, it included minimal descriptive content, making it relatively straightforward to convert and helping to establish a practical benchmark for the workflow. By contrast, later files presented richer but more ambiguous content that could not always be confidently interpreted. This highlighted the value of working with colleagues who have expertise in early printed books. Future projects would benefit from deeper cross-disciplinary engagement to address uncertainty and support the creation of more consistent, higher-quality TEI files.

Due to its bibliographic orientation, MARC's flat, field-based structure often lacks the contextual depth required for the kind of expressive, semantically rich markup supported by TEI. Such a structure obscures relationships between entities, restricts item-level attribution, and excludes interpretative elements such as meaningful summaries or tables of contents. These limitations reinforce the need for human-led editorial intervention to produce metadata that is semantically precise, structurally coherent, and optimised for digital presentation and discovery.

[5] https://www.digitalcollections.manchester.ac.uk/view/PR-ALDI-18873/1

A hybrid approach, combining automation, editorial expertise, and collaborative practice, can deliver high-quality, non-MARC metadata at scale. The editorial model developed in this project offers a practical foundation for future initiatives aiming to produce TEI metadata that meets both structural and descriptive standards and supports long-term discovery and use of early printed books in digital collections.

Conclusion

Through the lens of a case study, this article has examined the evolving potential of TEI to enhance and contextualise MARC-based metadata for digitised early printed books. TEI's semantic flexibility, extensibility, and use of structured XML designed for human interpretation make it particularly well suited to describing materials that exceed the descriptive and structural limits of MARC.

The Library's public-facing interfaces, Primo and MDC, manifest the pivot from MARC's catalogue-centred logic to TEI's semantically driven model. Primo draws on MARC records to support catalogue search, inventory management, and authority control. These are tasks aligned with traditional library workflows. However, MARC is not designed to deliver the layered, interpretative access required for digitised early printed books.

Transforming MARC into TEI entails more than field mapping. It requires interpretation, structural modelling, and content enrichment to ensure that the resulting metadata is accurate, interoperable, and suitable for digital presentation and discovery. In short, MARC provides bibliographic data, whereas TEI enables it to be expressed as meaningful, structured, and contextualised digital object metadata.

The MDC Dante items, comprising both facsimile images and their corresponding TEI files, now benefit from richer, semantically enhanced descriptions. This pairing of object and metadata reflects MDC's core definition of a digital object: a unified entity that enables access, interpretation, and future reuse.

Working across the full set of records revealed the limits of automation and the importance of flexibility in metadata creation. While defined workflows are necessary, they must be adaptable enough to accommodate the descriptive complexity of early printed books. Moving beyond MARC requires more than conversion. It is a process shaped by interpretation, editorial judgement, and contextual understanding.

Emerging AI tools are beginning to reshape the metadata landscape. Steven Hartshorne (2025) recently conducted a small-scale experiment that used ChatGPT to enhance existing MARC records. The trial suggests that AI can streamline certain metadata upgrade tasks by providing lightweight suggestions that still require human input. Crucially, the semantic depth, structural complexity and editorial nuance of TEI continue to demand interpretive decisions that current AI cannot match.

Our metadata journey echoes Dante's ascent from *Inferno to Paradiso*: moving from inherited constraints toward a realm of expressive possibility. The progression from initial conversion to enriched transformation was shaped by human judgement, editorial expertise and care. At its core, the process remains interpretative and constructive and, for the foreseeable future, "*unmistakably*" human.

Both PR-ALDI-18873 (https://www.digitalcollections.manchester.ac.uk/view/PR-ALDI-18873/1) and PR-ALDI-R-00002-13785 (https://www.digitalcollections.manchester.ac.uk/view/PR-ALDI-R-00002-13785/1) are now accessible for viewing on MDC.

References

Hartshorne, Steven (2025) Manipulating Rare Print Metadata with ChatGPT. *Catalogue & Index*, 211, pp. 9-19. Available at: https://journals.cilip.org.uk/catalogue-and-index/article/view/748