Editing in the age of AI: stratified establishment of medieval texts

Ariane Pinche

Editing in the Age of AI: Stratified establishment of Medieval Texts

Ariane Pinche, CNRS, CIHAM UMR 5648

December 22, 2025

Introduction

Editorial practices are rooted in a long history of attempts to reconstruct lost authorial originals or to establish canonical versions, working with texts such as the Bible, Homer, Cicero, Vergil, Chrétien de Troyes, and Shakespeare J. J. McGann 2001. Since the 16th century, scholars have sought ways to represent the plurality of texts and their material instantiation, ranging from diplomatic transcription to the reconstruction of an archetype as close as possible to the authorial version. Critical editions aimed to establish the relationships between different realisations of a text, sometimes building a hypothetical archetype from the manifestations that remain accessible today. Since the 20th century, the authority of the archetype as the sole version of value has been challenged. With Cerquiglini 1989 *Éloge de la variante*, witnesses and textual variations have regained scholarly attention, giving rise to the new philology and influencing modern approaches to editing corpora, including genetic philology. Then, Shillingsburg 1996 defined the role of the editor as representing the textual process rather than producing a static text, while Bryant 2007 stated that 'the only "definitive text" is a multiplicity of texts, or rather, the fluid text'. Yet printed editions remain constrained by the physical limitations of the page and cannot fully convey these processes or offer an exhaustive overview. Digital editions have emerged as part of the solution, enabling the integration of digitized sources, multiple reading paths, and immersive networks of contextualisation through hyperlinks. Some projects adopt a document-oriented approach, providing access to complete texts of multiple witnesses Pinche, Bureau, and Nicolas

2016; others reconstruct an archetype through intermediary steps, resulting in an established text by textual tradition branch Trachsler and Leonardi 2015. Other editions focus on creating networks of hyperlinks to situate the text within broader historical, literary, or cultural contexts.[1] However, as far as we know, relatively little attention has been paid to representing the process of text establishment itself.

For thirty years, the Text Encoding Initiative (TEI) has further expanded possibilities for text annotations, enabling markups for material features, linguistic information, structure, and metadata. This has not only enhanced the ways information can be represented, but also fostered a community of shared editorial practices. However, the rise of AI, and more specifically of automated transcription of historical documents, have transformed access to medieval texts. Tasks that once required months or even years of transcriptions and annotations can now be completed in a matter of days, sometimes even hours. This acceleration is not only a question of productivity, it also reshapes how scholars engage with texts, the workflows through which these texts are acquired, and the modes of their representation and dissemination. In traditional editorial practices, transcription was often treated as a preliminary, mechanical, and largely invisible step, intended only to prepare the manuscript for subsequent critical work. The emergence of automated processes invites us to reconsider the epistemic status of transcription itself.

In the following sections, we will outline why transcriptions should therefore be regarded not as a preliminary step, but as a scientific object in its own right: a space for interpretation, where each layer of the text: raw automatic output, corrected transcription, normalized forms, or enriched semantic annotations, carries distinct epistemic value. This layered approach places transcription at the centre of a new editorial ecology, which simultaneously supports a return to the materiality of the manuscript and enables a synthesis of textual, linguistic, and computational analysis. Therefore, we will explore the new possibilities opened by automatic text recognition (ATR),[2] and how these tools offer ways to add textual layers of information, sharing the process of text establishment and enrichment. We will begin with the digital capture of the text contained in the physical object (transcription),

[1] See projects such as the *Walt whitman Archive*, the *Rosetti Archive*, or even the *Woolf online*.

[2] We use the acronym ATR (Automated Text Recognition) to encompass both HTR (Handwritten Text Recognition) and OCR (Optical Character Recognition) in order to refer to processes applicable to both printed and manuscript sources.

starting point for any editorial work. Then, we will show how the ability to acquire and to share large amounts of data online, thanks to the use of AI, reshapes the role of transcription and enrichment, granting them a central place in the editorial process. Finally, we will examine the role of the TEI as a mediating framework, capable of integrating machine-generated data with philological insight and editorial practice, and facilitating the creation of digital editions that are both reusable and meaningful.

1 Transcription in Digital Editorial Methods

For centuries, transcription has been relegated to the status of a pre-editorial, technical task, one of the many preliminary drafts before the 'real' work of editing begins. In the tradition of French medieval editing, as said by Duval 2012, transcription has remained an unexamined assumption; current practices still follow rules inherited from figures such as Paul Meyer and Mario Roques, which continue to shape what is preserved and what is discarded. These conventions aim to facilitate access to medieval texts for modern readers by balancing fidelity to the source with contemporary reading expectations. Yet, the establishment of a text is never neutral: it has direct consequences for its future interpretation and reuse. As Pierazzo 2011 has underlined, the final product presented to readers was, for a long time, the printed edition, a constrained object shaped by the physical limits of the page. Within this framework, transcription was deliberately rendered invisible beneath a normalized, smooth surface, designed to be easily readable by an audience sometimes unfamiliar with messy medieval sources. This smoothing process, however, came at the cost of erasing linguistic and palaeographic features, with abbreviation expansions, punctuation and spacing silently standardized, sometimes introducing interpretative biases. Thus, despite its central role in preparing the edited text, transcription remained an invisible labour that nevertheless structured every subsequent stage of textual establishment.

This invisibilisation makes it necessary to pause and clarify what we mean by 'transcription.' Far from being a neutral transfer of signs from manuscript to page, transcription is always a transformation: it turns the complex content of a source document into a more readable form, often, today, into a digital plain text usable for keyword searching. It is therefore not a mirror image of the manuscript, but a selective, interpretative process. As Pier-

azzo 2011 reminds us, any transcription encodes only a selective subset of the potentially infinite 'facts' present in a manuscript. In this sense, it is inherently interpretative and necessarily incomplete. This view is reinforced by Huitfeldt and Sperberg-McQueen 2008, who argue that transcription is not a neutral mapping of graphemes[3] where each sign in the source has an exact equivalent, but rather a 'systematic program of selective alteration and preservation.' However, different levels of granularity can be chosen when representing the source. As Robinson and Solopova 1993 described, the different levels of transcription include:

- Regularized level: This level propose a normalized version of the text according to an editorial norm, where all manuscript spellings are regularized, offering ease of reading but at the cost of erasing graphical diversity.

- Graphetic level: This level captures every distinct letter form, called allograph (e.g. long <s> and round <s>, various forms of <a> or <d>, pointed <i> and not, etc.). This level of precision is useful for identifying multiple scribal hands or classifying scripts across large corpora. However, it is extremely time-consuming and often difficult to apply consistently, since the boundaries between allographs are not always clear and differ among scholars.

- Graphemic level: Between the graphetic and normalized levels, the graphemic level preserves spelling distinctions (e.g. she vs. sche) without encoding individual letter forms. It includes brevigraphs, original punctuation, special signs such as insertion marks, and layout features like line breaks or folio changes. This level of granularity may roughly correspond to what is traditionally called a 'diplomatic edition.' Yet it often occupies an ambiguous position, blurring the boundary between transcription and edition, since both are the results of editorial decisions and belong to the same spectrum of textual establishment Shillingsburg 1996.

The very existence of these levels demonstrates that transcription is not a uniform practice, but the result of choices. The distinction between edition and transcription therefore lies not only in the degree of fidelity, but

[3]Grapheme refers to the abstract unit that connects a sign in the source to a standardized form in transcription, usually corresponding to a phonetic representation.

Numerisation of the source¶	 Geneva, ms. Comites Latentes, 102, fol. 122 ra, image from e-Codices¶
Normalized transcription¶	Ci conmance la vie saint Lambert l'evesque et le martir¶
Graphetic transcription¶	Cı ꝯmance la uíe ſaínt lambert leueſq̃ ꝛ le marꞇ¶
Graphemic transcription¶	Ci ꝯmance la uie saint lambert leuesq̃ ⁊ le marꞇ¶

Figure 1: Examples of Different Levels of Transcription of ms. Comites Latentes, 102, fol. 122 ra, numerisation from e-Codices

also in the purpose, intended audience, and epistemic status. As Pierazzo 2011 has argued, diplomatic transcriptions often function as internal working documents, whereas editions are public outputs governed by explicit conventions, with sometimes the addition of editorial notes to explicit context. The fact that transcription is an interpretative act becomes even clearer when we turn to its limits. No digital system, however sophisticated, can capture every feature of the source: ink density, colour variations or quire structure remain beyond the reach of automated tools. Experience shows that even at the level of graphemes, mapping messy real-world shapes to abstract types requires interpretative simplifications. While alphabetic characters can be more or less easily standardized through Unicode, relying on a common representation of what a letter is in the alphabetic system, the representation of brevigraphs, scribal abbreviations, or layout features is far more complex. No international consensus currently exists on how to encode medieval abbreviations, and practices vary significantly across time and scholar traditions. As a result, producing graphemic transcriptions is less a mechanical operation than a modelling task. The purpose of the transcription becomes decisive. A palaeographer seeking to classify scripts, a linguist studying variation, or a digital humanist preparing data for ATR training will not encode the same features, nor with the same granularity. Transcription thus operates at the intersection of theory and practice: it is constrained by the material richness of manuscripts, yet guided by scholarly priorities, project-specific needs and technological considerations.

The rise of digital editing frameworks can make editorial choices explicit. The TEI, for example, provides tools to annotate a vast array of textual

features: abbreviations and expansions, original and normalized punctuation, line and folio breaks, and more. The graphemic transcription in figure 1 could be rendered in TEI with detailed markup, reporting normalisation and sign in the source:

```
<head><hi rend="red">
Ci <choice><abbr>9</abbr><expan>con</expan></choice>mance la
<choice><orig>u</orig><reg>v</reg></choice>ie saint
<choice><orig>l</orig><reg>L</reg></choice>ambert
l<pc type="reg">’<pc><choice>e<orig>u</orig><reg>v</reg></choice>es
<choice><abbr>q~</abbr><expan>que</expan></choice>
<choice><abbr>7</abbr><expan>et</expan></choice> le
ma<choice><abbr>r'</abbr><expan>rtir</expan></choice>
</hi>.<lb/></head>
```

Yet this richness comes at a price. Encoding manually at such granularity is extremely time-consuming, and its results are often bound to the logic of the edition: organisation in paragraphs, apparatuses, variant readings, and linguistic annotations. All information is treated as if it were on the same plane, ‘inherently’ attached to the editorial object, resulting in a dense markup that is difficult to reuse or repurpose for different scholarly goals. In fact, most digital editing frameworks still tend to privilege a linear vision of the text, moving directly from the manuscript to the edited version, without fully acknowledging the intermediate layers that shape this transformation. Rarely do they envisage the text as a stratified object, composed of successive layers of establishment that extend from the source to the edited form.[4] If yet such stratification has been conceptualized for the reconstruction of textual histories across multiple witnesses, in these contexts, each strata of text establishment: raw ATR output, graphemic transcription, normalized forms, edited texts constitute distinct epistemic objects, documenting both the material history of the text and the interpretative processes. Applying a similar perspective to digital editing would make the different levels of transcriptions visible as layers building a network of textual representation. TEI’s <sourceDoc> element can provide a framework, as it allows transcriptions to be represented outside the constraints of a single ‘final’ edition in

[4] As counterexample, the Menota project for icelandic texts has adapted TEI with three levels of textual represention: facsimile, diplomatic and normalised level, see Menota handbook: Levels of text representation [https://www.menota.org/HB3_ch4.xml\#sec4.6], accessed 08/25/2025

the <body> element, preserving them as reusable scholarly artefacts and allowing another text hierarchy.[5]

Taken together, these reflections suggest that transcription can be conceived not as an invisible, preparatory stage, but as a scholarly object in its own right. It is a locus of interpretation, where the balance between fidelity and readability, materiality and abstraction, automation and human judgment is constantly negotiated. For a long time, transcription stayed a hidden preliminary step, but in the digital age, especially with ATR and flexible encoding frameworks such as TEI, its epistemic weight has become impossible to ignore. Transcription requires more than technical solutions and should be established as citable, layered, and reusable scholarly objects, governed by conventions. In this sense, transcription is not a silent prelude to editing, but a central moment of editorial decision-making, an interpretative act that structures all subsequent uses of the text.

2 Transcription and the Emergence of Automatic Text Recognition

The emergence of ATR invites us to reconsider the role of transcription in the editorial process. While traditional philological models tended to treat transcription as a silent and intermediate stage in the production of a critical edition, the integration of ATR into scholarly workflows reveals a more dynamic and distributed ecosystem. In the digital context, transcription is no longer limited to a single purpose. It can serve as a first step toward an edition, but also circulate independently as a valuable digital object. Raw automatic transcriptions or manually revised versions can be shared online and reused in various ways. Because digital texts can be evolutive, one can envision portals where transcriptions are updated, corrected, and enriched over time. Even imperfect or partially corrected transcriptions can prove useful: for instance, as preliminary material for research or as input for quantitative methods capable of accommodating a degree of noise Eder 2013. Furthermore, transcriptions can enter new cycles of reuse as training data for ATR models, if their are produced with clear guidelines and documentation ensuring internal consistency and interpretability. This multiplicity of functions demands that we treat transcriptions as a central and shareable component

[5]See project Gallicorpora, Pinche, Christensen, and Gabay 2022.

of scholarly work, even more with the emergence of AI and deep learning, as it is training model material.

One of the most transformative aspects of ATR lies in its ability to generate transcriptions as autonomous textual objects guided by their own internal logic rather than necessarily aiming toward a critical edition. This shift allows transcription guidelines to focus on the characteristics of the source and compatibility with born-digital text acquisition pipelines. Within the framework of the *Consistent Approaches to Transcribing Manuscripts* (CATMuS) project Clérice, Pinche, et al. 2024; Pinche, Clérice, et al. 2024, a collective effort has been undertaken to produce multilingual and diachronic datasets for ATR. The goal is to harmonize transcription practices in order to enhance model performance while also fostering data interoperability. The CATMuS-medieval dataset brings together data from various languages (from the 9$^{\text{th}}$ to the 16$^{\text{th}}$ centuries) and across a broad spectrum of textual genres.[6] What unites these diverse sources is the use of shared guidelines that standardize transcriptions in a way that supports the training of generic ATR models. These guidelines are grounded in a graphemic approach, which has proven the most viable path toward consistency in collaborative transcriptions.

In fact, implementing a fully allographic method across such a corpus would pose nearly insurmountable challenges. First, establishing an inventory of allographs across multiple languages and time periods, each mapped to Unicode characters, would be a massive and complex undertaking despite the foundational work of the Medieval Unicode Font Initiative (MUFI).[7] Second, even if such a mapping existed, it would be difficult to ensure consistent interpretations of allographic boundaries by different transcribers. For example, while distinctions between round and long <s> or common and rotonda <r> are relatively straightforward, the same cannot be said of ligatures or subtle variations in module size. Devoting so much attention to distinguishing allographs cannot only introduce heterogeneity into the corpus, but also result in basic transcription errors, as it diverts focus from more fundamental aspects of the text Robinson and Solopova 1993. Third, opting for a highly detailed level of transcription would exponentially increase the number of character classes to be learned by ATR models, thus requiring more training data.

[6]"CATMuS/medieval · Datasets at Hugging Face",`https://huggingface.co/datasets/CATMuS/medieval`, accessed 08/19/2025.

[7]Odd Einar Haugen, Alec McAllister and Tarrin Wills, *The Medieval Unicode Font Initiative*, [`https://www.mufi.info/q.php?p=mufi`, accessed 11/05/2025].

Instead, the *CATMuS guidelines* aim for a balance between formal rigour and practical feasibility. They are built around a systemic logic and a limited set of straightforward rules designed to facilitate collaborative work. We assume that our model will necessarily be a simplification of the real object, but, as said by Pierazzo and P. A. Stokes 2011, it is necessary to make it useful. Each sign in the source is transcribed using a standardized Unicode character based on its phonetic realisation and visual form. Certain exceptions remain, however. In most of the medieval corpora, the characters <u> and <v> or <i> and <j> are not distinct letters, but positional allographs. They do not mark a phonetic opposition (vocalic versus consonantal realisation); rather, their distribution depends largely on their position within the word, and sometimes their use appears almost arbitrary. However, for non-alphabetic symbols, the task becomes significantly more complex and subjective. The diversity of signs through languages, periods, and document types makes it difficult, if not impossible, to identify a universally applicable system. Punctuation, for example, poses a serious challenge: medieval texts possess a wide range of signs (median punctus, punctus elevatus, punctus versus, double points, high or low points etc.), often used in ways that diverge from modern habits. Since their usage is generally not governed by a clear system applicable to a large range of documents, CATMuS adopts a simplifying approach, reducing point representation to a binary of simple and double points to ensure consistency across the datasets. Abbreviations are similarly difficult to manage. The line between allographic variation and distinct signs is often blurry. To maintain progress, CATMuS has adopted a pragmatic set of decisions based on two criteria: visual similarity and functional equivalence. For instance, both tildes and macrons, frequently used to mark nasalisation or suspension, are represented with the tilde sign easy to access on a keyboard. The Tironian <7> (et) is treated as a single character, and stroke-based variants are considered allographs. Structural signs like the pilcrow <¶> or insertion marks are also addressed, with compromises struck between fidelity to the source and a pragmatic approach facing the wide number of signs used across sources. Unreadable parts of the text are marked with square brackets <[ ]>, and parts of the text corrected in the manuscripts (ratura, exponctuation, etc.) are encoded using <⟦⟧> brackets with the text inside, following Leiden conventions.[8] These two cases are the only instances

[8]Sterling Dow, "Conventions in Editing; a Suggested Reformulation of the Leiden System", *Internet Archive* ([Cambridge, Mass.] Duke University, 1969),

in which the guidelines explicitly privilege interpretation over strict visual representation, with the latter case leaning more toward semantic markup than pure transcription.[9] To this day, they remain largely theoretical and are applied only in manually transcribed data, as they do not yet occur in sufficient quantities to train a model dealing efficiently with these phenomenons. Regarding the representation of abbreviations, CATMuS anticipates updates aimed at refining their representation in future corpora and ensuring closer alignment with palaeographic criteria, while maintaining compatibility with previous corpora. The first draft was grounded in existing guidelines for French vernacular manuscripts from the 10^{th} to the 15^{th} centuries, which provided a practical starting point, but did not encompass multilingual and cross-script perspectives Pinche 2022.

These reflections illustrate the complexity of defining transcription guidelines in a collaborative, multilingual, and computational context. Many of the CATMuS decisions diverge from traditional scholarly habits, because they serve a different purpose: the creation of reusable, shared, and scalable transcription data.[10] This approach is a pragmatic compromise that enables progress and collaboration in a context where it serves both humans and AI. Projects with more specific goals, for example, a homogeneous corpus with a focus on punctuation practices, may adopt more granular transcription rules. If properly documented and systematically applied, such data can still be converted into a more generic form to be used with standard models. Likewise, projects wishing to distinguish between phenomena not marked in the generic models (e.g. u/v or i/j distinctions, abbreviation expansion) may begin by using a generic CATMuS model, and then fine-tune it according to their own requirements. In this way, the cost of model development, whether measured in computational resources or annotation time, can be significantly reduced. Transcription, when undertaken thoughtfully in a digital context and in dialogue with deep learning tools, is not a shadowy or auxiliary task but a constitutive act of knowledge production. It establishes the conditions of possibility for future research by inscribing texts into systems that are sustainable, interoperable, and collective. It facilitates faster project start-ups

http://archive.org/details/conventionsinedi0000dows.

[9]The second case was introduced at the request of transcribers and is used less as a feature for model training than as a means of signalling and retrieving those passages for researchers.

[10]Those kind of data are needed to help digital approaches to rely on standardized datasets, allowing benchmarking, P. Stokes 2018.

and promotes a more responsible use of technological and human resources. In this sense, transcription shifts from the register of individual labour to that of the scientific commons, where its value lies in the creation of shared, cumulative, and durable knowledge.

ATR also introduces a new layer of information that is rarely preserved in transcription practices. Scholars often produce editions that present the text independently of its manuscript context, but this material dimension can be crucial for understanding the text. In this regard, ATR offers a new way of addressing this issue. Indeed, most of the time, before textual recognition itself, the image is segmented into distinct zones and lines, and if needed characters. It generally follows a two-step process: first, the identification of text zones, and then the mapping of text lines within those zones. These elements are preserved in the XML output with their coordinates on the image. Thus, ATR maintains a record of the page's spatial organisation, information traditionally neglected in most transcription workflows. Most of the material manifestations can be translated into TEI, as line, column, page beginnings, features like numbering, running title in <fw> and the <sourceDoc> element enables a comprehensive mapping between image and text. As manual segmentation remains extremely time-consuming, automating this task is therefore a real game changer.

That information is of particular interest because it captures the material structure of the source document, which is not simply a container for text, but also an object with semantic and functional clues. As Jacques Monfrin in *Mise en page et mise en texte du livre manuscrit* Martin, Vezin, and Monfrin 1990 noted, the layout is designed for the human eye: it reflects textual hierarchy with the support of visual artefacts. In this sense, layout is not merely an accessory but a semiotic artefact, guiding the reader's understanding of the text. Digital transcription that preserves segmentation, thus, participates in the building of a model where the visual form of the text is considered an integral part of its meaning. ATR not only captures where the text is on the page, it also enables the labelling of zones and lines. Lines can be annotated as headings, rubrics, or marginal notes, while zones may be identified as columns, running titles, decorative initials, or gloss areas. Modelling layout is, however, a complex task. It requires a conceptual and technical framework adapted to computer vision to detect and classify layout structures based on visual cues. The SegmOnto project addresses this challenge by offering a controlled vocabulary for layout description, spanning documents from the medieval period to the 20$^{\text{th}}$ century Gabay, Pinche, et al.

2024. It includes zone types such as *MainZone*, *RunningTitleZone*, *DropCapitalZone*, *DecorationZone*, *NumberingZone*, and line types like *DefaultLine* or *HeadingLine*. SegmOnto remains general enough for broad applicability, while allowing fine-grained customisation. Second-level taxonomies can be developed to suit specific corpora or research questions, as illustrated by the LADAS project, which incorporates textual structures like <p>, <lg>, and <sp> from the TEI into the layout model Clérice, Janès, et al. 2025.

Segmentation can be especially valuable in computational analysis, providing new measurable features: line length, interlinear spacing, proportions of white spaces, or average letter size. It enables the tagging of codicologic features (e.g. red ink, marginal glosses, decorated initials), and can be used in computer vision workflows for letter form analysis or script classification Kestemont, Christlein, and Stutzmann 2017. Segmentation thus opens up the possibility of automated palaeographic or codicological studies, such as identifying allographs or distinguishing scribal hands based on spatial or visual parameters. Despite its potential, segmentation still presents challenges. Automated segmentation can be difficult when faced with non-standard or degraded layouts (e.g. papyri). There is also a noticeable lack of sharing and reuse when it comes to layout models. Training such models, at least with tools like Kraken Kiessling 2022, requires significant amounts of data and computing power. Moreover, the absence of common practices exacerbates this fragmentation: vocabularies differ, levels of granularity diverge, and reuse objectives often pull in different directions. For instance, Kraken's general model (BLLA) does include only one class, while the PageXML standard defines around fifteen region types, such as *TextRegion*, *ImageRegion*, *GraphicRegion*, *ChartRegion*, *LineDrawingRegion*, *SeparatorRegion*, or *TableRegion*, which are tailored to modern printed documents. On the other hand, the M6Doc dataset includes 76 classes, also designed for modern layouts, but sometimes leading to ambiguous categories, for example *footer* and *footnote* Cheng et al. 2023. Yet well-mastered segmentation could represent a shift toward material description and open new methodological avenues, going finally further than what could have been in printed editions. The information captured through segmentation is not merely useful for measurement or statistical analysis; it can also feed into editorial workflows, enabling what might be termed a *pre-editorialisation* of the text based on this step Gabay, Pinche, et al. 2024. Far from being a by-product of ATR, segmentation is an epistemic resource, enabling a more grounded, multidimensional understanding of the textual artefact. Then, it offers a foundation not just for

document description, but also for automatic text structure tagging.

In the age of AI, transcription is no longer a private stage in the editorial process. It involves a series of deliberate choices about *what to transcribe*, *how to render it*, and *to what extent*. These choices are not only technical; they shape the epistemological status of the text and condition its future uses, whether in critical editions, computational models, or distant reading. In this sense, transcription should be understood not as a preliminary step but as a constitutive act of interpretation and modelling. Segmentation, likewise, is far more than a technical prerequisite for ATR. It functions as a descriptive and epistemic operation, allowing us to capture the spatial organisation and material logic of the page. Through the classification of visual zones, segmentation preserves its functional structures, often essential for understanding medieval textuality. The very act of segmenting implies a model of how the document's materiality carries meaning. Then, this ability to produce this wealth of information automatically invites us to think about digital philology as a layered process, where textual and material dimensions are articulated.

3 New Editorial Horizons for Medieval Texts ?

Automatic process in text acquisition has lead to a major acceleration of the task of transcription and annotation. Yet this acceleration is not simply about increasing productivity: it fundamentally reshapes how scholars engage with texts, as well as the workflows through which they are constructed, represented, and disseminated. What was traditionally conceived as a linear movement from manuscript to edition now appears as a stratified process, where successive layers of transcriptions, annotations, and interpretations coexist and remain accessible to future reuses. Even though ATR remains an imperfect process that inevitably produces texts with errors, it fundamentally transforms the way we work with medieval manuscripts. Its value extends beyond producing 'perfect' scholarly editions and even noisy outputs can be leveraged for quantitative analysis, text mining, and exploratory reading D. A. Smith et al. 2023. By converting manuscripts into a searchable digital form, ATR not only accelerates the transcription process, but also enables new approaches to textual scholarship, allowing to engage with larger

corpora, trace patterns across manuscripts, and to experiment with methods that were previously impossible at this scale. By enabling large-scale transcription, it has encouraged the development of document-oriented pipelines with workflows that combine segmentation, recognition, normalisation, and sometimes alignment with other textual witnesses. Since 2020, such pipelines have multiplied across projects, reflecting a growing ambition not merely to transcribe texts, but to structure entire corpora for subsequent computational analysis, as seen in projects such as DAHN Chiffoleau, Baillot, and Ovide 2021, AGODA Bourgeois et al. 2022 or Gallic(orpor)a Christensen, Pinche, and Gabay 2022).

Most of the time, the process starts with layout analysis. Then, ATR produces the transcription, which can be further enriched by abbreviation expansion, normalisation, or lemmatisation. But each layer of enrichment raises the challenge of linking the information to the original text. When one token corresponds to one word or punctuation mark, alignment is straightforward. But in many cases, words are split at line breaks, column changes, page boundaries, and in the case of medieval manuscripts, words can be agglutinated. Normalisation is then required before annotation and alignment can proceed. The problem is further complicated in medieval corpora, as graphic normalisation is highly dependent on language, genre, and editorial choice and difficult to automate, as it depends on context and scientific objectives. Each project has to rely on its own *ad hoc* solutions, ranging from rule-based scripts to deep learning models or LLMs. In fact, we are still far from a 'prêt-à-porter' solution, immediately usable by humanities scholars without technical support Pierazzo 2019. What is emerging instead are hybrid workflows, where automation and scholarly expertise are interwoven. Step by step, tools are developed to facilitate the editor's work and to provide semi-automated processes in which human expertise remains essential to guide, correct, and validate the automatic output. The document-oriented pipeline thus acts for medieval documents as a pre-edition, a structured intermediary that shapes the possibilities of the critical edition to come.

One of the limitations is that most of those pipelines are designed to handle document-oriented production and do not directly address the issue of text alignment between witnesses. Algorithms and tools to assist alignment are available today, such as CollateX. This tool enables the comparison of texts with minimal normalisation, using alignment processes inspired by bioinformatics. Its output can be either a TEI critical apparatus or a graph representation of textual variation. However, one of its limits lies in the

difficulty of handling medieval vernacular variations and large manuscript traditions Haentjens Dekker et al. 2015. FALCON has adapted CollateX to medieval French by aligning linguistic lemmas rather than word forms, in order to cope with graphic variation, and by proposing a classification of variants Camps, Spadini, and Ing 2019. Still, macro-alignment remains a challenge for vernacular traditions, where textual transmission is versatile and structures can be modified. Frequently, text reuse techniques, as implemented in PASSIM D. Smith 2025, or large language models are employed to align texts, whether mono- or multilingual Levenson, Ing, and Camps 2024. This step remains at the heart of current research and is still in need of improvement. Some of these approaches have even been integrated into text acquisition pipelines starting from ATR output, such as HTR2CritEd for Hebrew Ezra et al. 2022 or tei-collator relying on the preceding work done for FALCON Levenson 2025. Yet the question of alignment between witnesses persists, especially when working with noisy ATR data. One of the technical constraints that has emerged in automation is the limitation of the lemma length in the critical apparatus to the word level. While technically efficient, this choice is debatable. In many traditions, the meaningful unit for comparison is not the isolated word, but a larger segment, and it often falls to the editor to decide the appropriate lemma length according to the textual tradition and the unit of sense. Defining what constitutes a unit of variation in medieval texts is not merely a technical matter, but a challenge that calls for contextualisation and expert judgment. The critical apparatus must not only document textual variation, but also render it intelligible. This requires moving beyond the mere accumulation of divergent readings to a structured representation of variations that reflects editorial interpretation and facilitates the reader's engagement with the text.

In my view, automatic collation remains, for the time being, exploratory. Semi-automation undoubtedly accelerates the processing of texts, helping to identify which passages to collate in order to establish a stemma, or even to begin sketching the branches of a manuscript tradition. Yet it also raises fundamental editorial questions: what, precisely, should be compared? Where does the boundary lie between two versions of the same text and two distinct texts derived from a common narrative source? The change of scale introduced by digital corpora, capable of handling hundreds of manuscripts both within the same language and across different linguistic traditions, makes these questions all the more urgent. The size of these corpora also demands precise metadata, making it essential to encode such distinctions explicitly, so

that databases remain not only technically functional, but also philologically meaningful.

In fact, the sheer volume of text now producible by ATR forces us to reconsider how editions are structured. Rather than offering a single authoritative text, editors can increasingly provide multiple textual layers that reflect different stages of processing and interpretation. At the same time, collective work is needed: effective digital philology requires collaborative expertise, combining philology, palaeography, codicology, and other traditional fields. Scholars such as J. McGann 2006 had anticipated this approach through the concept of the social text, and advances in AI now make it reachable. Yet, there is a danger of overwhelming readers with unstructured information, a 'forest' of raw textual data, which will lead to refusal to edit, as warned by Driscoll and Pierazzo 2017. Then, the editor's role remains essential to trace coherent paths, organize the data patiently, and make sense of the accumulation of information.

I would suggest here the notion of 'stratified editing': a medieval text is not a single, static object, but a set of layers, each reflecting a stage in its editorial establishment. From transcription to edition, multiple versions of a text coexist: raw automatic transcriptions, semi-corrected outputs, and fully revised human-checked transcriptions, edited text, as well document-oriented or auctorial-oriented. Each of these states carries a distinct epistemic value. The raw ATR output reflects the machine's interpretation of the manuscript, including machine errors and palaeographic ambiguities. Corrected transcriptions integrate human judgments, resolving ATR errors and palaeographic ambiguities. Normalized transcriptions expand abbreviations and tokenize words semantically, offering easier reading and more computationally useful text. The reference edition smooths the text further, modernises punctuation, presents editorial amendments and incorporates ecdotic or explicative notes to enhance clarity, usability, and interpretative guidance. Recognising these distinct states encourages transparent editorial practices and allows revisiting interpretative decisions at any stage. One of the major advantages of digital objects is their ability to evolve over time and to accumulate layers of information that far exceed what is possible in printed editions, breaking temporal and space limitations. This capacity allows a transcription to exist as an autonomous scholarly object, rather than only a preliminary step toward editing. Transcriptions can serve multiple audiences: linguists seeking more direct access to a version of the text that has really circulated in a spatial and time context, scholars analysing divergences

between versions, computational researchers requiring raw textual data, and editors preparing the foundation for a critical edition. Reference editions themselves may be multiple, reflecting one or several witnesses, branches of a reception tradition, or a reconstructed authorial text. Many of the further uses are unpredictable. Building layered resources is a long-term collaborative task, but the digital environment enables their continuous evolution, making stratified publishing both more transparent and more versatile for collective reuse.

Beyond textual stratification, ATR output can also be enriched with linguistic tagging, named-entity recognition, or codicological features extracted from layout analysis. Here the notion of 'layer' extends to informational overlays, turning the edition into a multi-dimensional object where text, metadata, and analysis coexist. Texts can then be automatically enriched through deep learning lemmatizers, or large language models, which, unlike rule-based systems, can accommodate medieval graphic variation. These models can add linguistic annotations, including lexical lemmas, grammatical categories (number, case, gender), and verbal features (tense, mode, person) or even named entity recognition (NER), ideally linked to external reference databases such as Wikidata or Pleiades. Such enrichments open the way not only for traditional linguistic analysis but also for quantitative approaches like lexicometry, stylometry, diachronic analysis Romanello, Gabay, and Carboni 2025. Although NER are not yet fully optimized for medieval corpora, they enable the generation of indexes, prosopographical data, or geographic mappings, facilitating deeper exploration of the textual content.

This stratified approach illustrates how digital workflows can transform medieval texts into multi-dimensional objects. Each layer from raw ATR transcription to corrected, normalized, and semantically enriched annotations functions as an autonomous, reusable resource, while together they form a coherent scholarly edition. Yet layering alone is not sufficient. Without a shared encoding framework, these strata remain difficult to integrate. What makes a layered edition truly usable is not just the data itself, but its modelling and the interpretative, philological insights embedded within it, a challenge that leads us to the TEI.[11] The accumulation of such a wealth of

[11]For this entire section, I am deeply indebted to my colleague Simon Gabay, with whom I have had long debates on the subject and with whom I am currently collaborating on a collective project devoted to guidelines for the treatment of digital corpora. His insights have provided me with a broader perspective on the issues discussed here, see Gabay and Pinche 2025.

information thanks to AI offers the possibility of shifting the scale of humanities research. It allows us to work on more representative samples, to detect broader trends, and perhaps to overcome the prism of canonical texts transmitted through generations. In doing so, it can help bring to light less well known, but widely circulated works, such as Books of Hours or legendaries, which remain comparatively underexplored next to chivalric literature. Yet, as mentioned earlier, ATR and automatic annotations, if not interconnected, often remain confined to separate files, reused only by small research communities already accustomed to working with this kind of textual data. The risk is to miss the text as a complex whole and to work instead with fragmented pieces. As P. Stokes 2018 and Robinson 2010 have rightly argued, it is crucial to reconnect these pieces within a larger scholarly framework.

The TEI provides the most widely adopted framework for representing digital editions. Its role has become crucial in mediating between automated outputs and philological interpretation. TEI offers a rich vocabulary to describe textual phenomena: from abbreviations to complex apparatus structures. It allows for the representation of multiple dimensions of the text: material, linguistic, and hierarchical, through elements such as <sourceDoc>, <w> with attributes like @lemma, @pos, @msd, or more structural tags such as and <p> . In this sense, TEI can serve as a pivot framework. It allows ATR-generated data, normalized forms, and interpretative texts to be mapped into a common structure. This makes corpora interoperable and facilitates their long-term sustainability. At the same time, the TEI model is not without its challenges. Rooted in an Ordered Hierarchy of Content Objects (OHCO), it presupposes a single, linear hierarchy, whereas whole textual representation often requires multiple overlapping hierarchies, such as material layout versus editorial interpretation. As Pierazzo and P. A. Stokes 2011 point out, the TEI privileges textual structure over material structure: page, line, or column (<pb/>, <lb/>, <cb/>) are relegated to milestones, while the 'substantial' text encoded in the <body> takes precedence. Balancing these hierarchies is crucial to preserve both the philological and codicological dimensions of manuscripts. While TEI does offer workarounds: milestone elements, stand-off markups, or even multiple files, each project must define its own methodological priorities: What is the unit of analysis? Which layer takes precedence? How to ensure interpretative transparency?

Another unresolved question concerns visualisation. Can all this information be meaningfully displayed at once? Should a digital edition present entire documentation or selected parts? How should physical and semantic

objects be linked for both reading and computational analysis? As Rasmussen 2017 notes, digital editions are both read and used, and TEI enable this dual role by providing structured, reusable objects. Yet not every layer must be made visible in the interface: what matters most is that the files themselves remain accessible, intelligible, reusable and organized in a database, while the data are used to produce analysis and commentaries. From this perspective, the digital edition remains a coherent scholarly object, structured by editorial choices, while also serving as a gateway into a larger corpus of data that AI and digital infrastructures open up for reuse, reanalysis, and new interpretations.

In sum, despite some limits, the TEI should serve as a pivotal framework that integrates machine-generated data, philological insight, and editorial practices. It supports both the stratified layering of text establishment and the enrichment of informational content. By mediating between these layers, TEI enables visualisation, exploration, and publication while ensuring that digital editions remain intelligible, reusable, and philologically meaningful. Yet TEI is not a ready-made solution: projects that integrate AI and computational methods still require significant technical expertise and carefully designed workflows Pierazzo 2025. Representing multiple hierarchies, maintaining interoperability, and ensuring long-term sustainability remain major challenges. Nevertheless, TEI still provides the common denominator and the most robust framework currently available to mediate between automated outputs and scholarly interpretation.

Conclusion

Transcription can no longer be only considered a preliminary step leading mechanically to the edition. In the digital age, and especially with the rise of ATR, it emerges as a place where scholarly decisions, tensions, and interpretations are made explicit. With the accumulation of successive transcriptions, corrections, normalisations, and enriched annotations, the text is rendered as a multi-dimensional object, where each layer preserves distinct epistemic value and each supports different research goals and interpretative perspectives. Far from erasing philologic methodology, this practice encourages an understanding of editing as a processual and stratified practice, where machine-generated output, scholarly modelling, and editorial expertise converge. A layered editorial approach thus enables both transparency and

flexibility, allowing texts to evolve over time while preserving the material, linguistic, and interpretative dimensions that support scholarly understanding. Ultimately, the layering of editorial work transforms transcription into a reflective, collaborative, and knowledge-generating act, positioning it at the heart of an editorial ecology, where the stratified text becomes simultaneously a research object, a computational resource, and a vehicle for new forms of scholarly synthesis and insight. Yet sustaining these practices is challenging. It requires databases, interoperable formats, and platforms that allow layers to be published, cited, and reused. It also demands recognition of transcription as a scientific endeavour and disciplinary legitimacy. Only then can stratified editorial practices fully unfold their potential, opening the way to edit and engage with medieval texts.

References

Bourgeois, Nicolas et al. (June 2022). "Le projet AGODA. Annoter et publier les débats parlementaires français de la fin du XIXe siècle : défis et solutions". In: *Présentation des projets AGODA et Gallicorpora, Bibliothèque nationale de France*. Paris, France. URL: https://hal.science/hal-03762957 (visited on 03/13/2024).

Bryant, John (2007). "Witness and Access: The Uses of the Fluid Text". In: *Textual Cultures* 2.1, pp. 16–42. ISSN: 1559-2936. (Visited on 12/18/2023).

Camps, Jean-Baptiste, Elena Spadini, and Lucence Ing (July 2019). "Collating Medieval Vernacular Texts. Aligning Witnesses, Classifying Variants". In: *DH2019 Digital Humanities Conference 2019*. Utrecht, Netherlands. URL: https://hal.science/hal-02268348 (visited on 08/20/2025).

Cerquiglini, Bernard (1989). *Éloge de la variante: histoire critique de la philologie*. Paris, France: Éd. du Seuil. ISBN: 978-2-02-010433-3.

Cheng, Hiuyi et al. (May 2023). "M6Doc: A Large-Scale Multi-Format, Multi-Type, Multi-Layout, Multi-Language, Multi-Annotation Category Dataset for Modern Document Layout Analysis". In: *CVPR 2023*. arXiv:2305.08719. Vancouver, Canada: arXiv. DOI: 10.48550/arXiv.2305.08719. (Visited on 11/11/2024).

Chiffoleau, Floriane, Anne Baillot, and Manon Ovide (Oct. 2021). "A TEI-based publication pipeline for historical egodocuments - the DAHN project". In: *Next Gen TEI, 2021 - TEI Conference and Members' Meeting*. Vir-

tual, United States. URL: https://hal.science/hal-03451421 (visited on 12/20/2023).

Christensen, Kelly, Ariane Pinche, and Simon Gabay (June 2022). "Gallic(or-por)a: Traitement des sources textuelles en diachronie longue de Gallica". In: *DataLab de la BnF*. Paris, France. URL: https://hal.science/hal-03716534 (visited on 12/21/2024).

Clérice, Thibault, Juliette Janès, et al. (July 2025). *Layout Analysis Dataset with SegmOnto (LADaS)*. URL: https://github.com/DEFI-COLaF/LADaS (visited on 07/29/2025).

Clérice, Thibault, Ariane Pinche, et al. (2024). "CATMuS Medieval: A multilingual large-scale cross-century dataset in Latin script for handwritten text recognition and beyond". In: *International Conference on Document Analysis and Recognition (ICDAR)*. URL: https://inria.hal.science/hal-04453952 (visited on 12/21/2024).

Driscoll, Matthew James and Elena Pierazzo (July 2017). "1. Introduction: Old Wine in New Bottles?" In: *Digital Scholarly Editing : Theories and Practices*. Digital Humanities Series. Cambridge: Open Book Publishers, pp. 1–15. ISBN: 978-2-8218-8400-7. URL: http://books.openedition.org/obp/3394 (visited on 03/16/2018).

Duval, Frédéric (2012). "Transcrire le français médiéval: De l' "Instruction" de Paul Meyer à la description linguistique contemporaine". In: *Bibliothèque de l'école des chartes* 170, pp. 321–342. ISSN: 0373-6237. (Visited on 07/10/2025).

Eder, Maciej (Dec. 2013). "Mind your corpus: systematic errors in authorship attribution". In: *Literary and Linguistic Computing* 28.4, pp. 603–614. ISSN: 0268-1145. DOI: 10.1093/llc/fqt039. (Visited on 10/11/2021).

Ezra, Daniel Stoekl Ben et al. (2022). "HTR2CritEd: A Semi-Automatic Pipeline to Produce a Critical Digital Edition of Literary Texts with Multiple Witnesses out of Text Created through Handwritten Text Recognition". In: *DH2022 Digital Humanities Conference 2022*. URL: https://dh-abstracts.library.cmu.edu/works/11895 (visited on 08/20/2025).

Gabay, Simon and Ariane Pinche (June 2025). "Fluidités graphiques : des images à la transcription". In: *Les Réformes orthographiques à la Renaissance entre Humanisme et imprimerie: une perspective européenne, 67e Colloque international d'études humanistes*. Tours, France. DOI: 10.5281/zenodo.10599911]. (Visited on 08/22/2025).

Gabay, Simon, Ariane Pinche, et al. (Dec. 2024). "SegmOnto: A Controlled Vocabulary to Describe and Process Digital Facsimiles". In: *Journal of*

Data Mining & Digital Humanities. ISSN: 2416-5999. DOI: 10.46298/jdmdh.12689. (Visited on 08/26/2025).

Haentjens Dekker, Ronald et al. (Sept. 2015). "Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project". In: *Digital Scholarship in the Humanities* 30.3, pp. 452–470. ISSN: 2055-7671. DOI: 10.1093/llc/fqu007. (Visited on 05/09/2023).

Huitfeldt, Claus and C. M. Sperberg-McQueen (Sept. 2008). "What is transcription?" In: *Literary and Linguistic Computing* 23.3, pp. 295–310. ISSN: 0268-1145. DOI: 10.1093/llc/fqn013. (Visited on 07/10/2025).

Kestemont, Mike, Vincent Christlein, and Dominique Stutzmann (Oct. 2017). "Artificial Paleography: Computational Approaches to Identifying Script Types in Medieval Manuscripts". In: *Speculum* 92.S1, S86–S109. ISSN: 0038-7134. DOI: 10.1086/694112. (Visited on 09/20/2021).

Kiessling, Benjamin (Apr. 2022). *The Kraken OCR system*. URL: https://kraken.re (visited on 07/29/2025).

Levenson, Matthias Gille (May 2025). *matgille/tei-collator*. URL: https://github.com/matgille/tei-collator (visited on 08/21/2025).

Levenson, Matthias Gille, Lucence Ing, and Jean-Baptiste Camps (Nov. 2024). "Textual Transmission without Borders: Multiple Multilingual Alignment and Stemmatology of the "Lancelot en prose" (Medieval French, Castilian, Italian)". In: *Computational Humanities Research 2024*. Aarhus, Danemark, p. 65. URL: https://enc.hal.science/hal-04759151 (visited on 08/20/2025).

Martin, Henri-Jean, Jean Vezin, and Jacques Préfacier Monfrin (1990). *Mise en page et mise en texte du livre manuscrit*. Paris, France: Éditions du Cercle de la librairie : Promodis. ISBN: 978-2-7654-0446-0.

McGann, Jerome (2006). "From Text to Work: Digital Tools and the Emergence of the Social Text". In: *Text* 16, pp. 49–62. ISSN: 0736-3974. (Visited on 07/30/2025).

McGann, Jerome J. (2001). *Radiant textuality: literature after the World Wide Web*. New York, Etats-Unis d'Amérique: Palgrave. ISBN: 978-1-4039-6436-6.

Pierazzo, Elena (Dec. 2011). "A rationale of digital documentary editions". In: *Literary and Linguistic Computing* 26.4, pp. 463–477. ISSN: 0268-1145. DOI: 10.1093/llc/fqr033. (Visited on 07/10/2025).

— (May 2019). "What future for digital scholarly editions? From Haute Couture to Prêt-à-Porter". In: *International Journal for Digital Humanities* 1, pp. 1–12. DOI: 10.1007/s42803-019-00019-3. (Visited on 08/20/2025).

Pierazzo, Elena (Jan. 2025). "Distant Editing: The Challenges of Computational Methods to the Theory and Practice of Textual Scholarship". In: *Futures of Digital Scholarly Editing*. U of Minnesota Press. ISBN: 978-1-4529-7253-4.

Pierazzo, Elena and Peter A. Stokes (2011). "Putting the Text back into Context: A Codicological Approach to Manuscript Transcription". In: *Codicology and Palaeography in the Digital Age 2*. Ed. by Franz Fischer, Christiane Fritze, and Georg Vogeler. Vol. 3. Norderstedt: Books on Demand (BoD), pp. 397–429. ISBN: 978-3-8423-5032-8.

Pinche, Ariane (June 2022). *Guide de transcription pour les manuscrits du Xe au XVe siècle*. URL: https://hal.archives-ouvertes.fr/hal-03697382 (visited on 07/20/2022).

Pinche, Ariane, Bruno Bureau, and Christian Nicolas (Sept. 2016). "Hyperdonat, digital edition project". In: *TEI Conference and Members' Meeting 2016*. URL: https://hal-univ-lyon3.archives-ouvertes.fr/hal-01413479 (visited on 09/14/2021).

Pinche, Ariane, Kelly Christensen, and Simon Gabay (Sept. 2022). "Between automatic and manual encoding". In: *TEI 2022 conference : Text as data*. Newcastle, United Kingdom. DOI: 10.5281/zenodo.7092214. URL: https://hal.science/hal-03780302 (visited on 12/21/2024).

Pinche, Ariane, Thibault Clérice, et al. (Aug. 2024). "CATMuS-Medieval: Consistent Approaches to Transcribing ManuScripts". In: *DH2024 Digital Humanities Conference 2024*. URL: https://inria.hal.science/hal-04346939 (visited on 12/21/2024).

Rasmussen, Krista Stinne Greve (July 2017). "7. Reading or Using a Digital Edition? Reader Roles in Scholarly Editions". In: *Digital Scholarly Editing : Theories and Practices*. Ed. by Matthew James Driscoll and Elena Pierazzo. Digital Humanities Series. Cambridge: Open Book Publishers. ISBN: 978-2-8218-8400-7. URL: http://books.openedition.org/obp/3406 (visited on 03/19/2018).

Robinson, Peter (2010). "Editing Without Walls". In: *Literature Compass* 7.2, pp. 57–61. ISSN: 1741-4113. DOI: 10.1111/j.1741-4113.2009.00676.x.

Robinson, Peter and Elizabeth Solopova (July 1993). *Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue*. DOI: 10.5281/zenodo.4050360. (Visited on 12/12/2021).

Romanello, Matteo, Simon Gabay, and Nicola Carboni (Sept. 2025). "Predicting the Fictional Time and Space of French Theatre Plays by Using Large Language Models". In: *IEEE International Conference on Cyber*

Humanities. Florence, Italy. URL: https://hal.science/hal-05157140 (visited on 09/01/2025).

Shillingsburg, Peter L. (1996). *Scholarly Editing in the Computer Age: Theory and Practice*. University of Michigan Press. ISBN: 978-0-472-06600-1.

Smith, David (Aug. 2025). *dasmiq/passim*. URL: https://github.com/dasmiq/passim (visited on 08/21/2025).

Smith, David A et al. (Dec. 2023). "Automatic Collation for Diversifying Corpora: Commonly Copied Texts as Distant Supervision for Handwritten Text Recognition". In: *CHR 2023: Computational Humanities Research Conference*. Paris, France.

Stokes, Peter (2018). "Digital and Computatinal approaches to paleography". In: *Manuscript Cultures*. Studies in manuscript cultures volume 22. Berlin Boston: De Gruyter. ISBN: 978-3-11-072349-6.

Trachsler, Richard and Lino Leonardi (2015). "L'édition critique des romans en prose : le cas de Guiron le Courtois". In: *Manuel de la philologie de l'édition*. Ed. by David Trotter. Berlin/Boston, Allemagne, Etats-Unis d'Amérique: De Gruyter, pp. 44–80.