Journal of the Text Encoding Initiative

Issue 18 | 2024
Selected Papers from the 2023 joint MEC and TEI Conference

Applied Text as Graph (ATAG) in an Algorithmic Edition

Andreas Kuczera

Electronic version
URL: https://journals.openedition.org/jtei/6027
DOI: 10.4000/14vyl
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Andreas Kuczera, “Applied Text as Graph (ATAG) in an Algorithmic Edition”, *Journal of the Text Encoding Initiative* [Online], Issue 18 | 2024, Online since 18 July 2025, connection on 08 November 2025. URL: http://journals.openedition.org/jtei/6027 ; DOI: https://doi.org/10.4000/14vyl

Applied Text as Graph (ATAG) in an Algorithmic Edition

Andreas Kuczera

ABSTRACT

The “Applied Text as Graph” (ATAG) methodology introduces a transformative framework for digital scholarly editing that addresses key challenges in text modeling, annotation, and citation with exceptional granularity (https://github.com/THM-Graphs/). By representing texts as interconnected labeled-property-graphs down to the character level, ATAG enables dynamic text analysis, flexible annotation hierarchies, and editable text structures without compromising their integrity.[1]

ATAG has been successfully applied to projects such as Liber Epistolarum (http://www.liberepistolarum.de/) and the Socinian Letter Project (http://www.sozinianer.de/), demonstrating its adaptability to diverse editorial needs. A pivotal advance is ATAG’s integration of the ENC-link citation model, which provides a precise and reliable method for scholarly referencing. By making every character uniquely addressable, ENC-links generate URL-based citations that pinpoint specific text segments, streamlining verification and enhancing academic transparency. This approach ensures robust citations even as texts evolve, through versioning mechanisms like character-node UUIDs and text hashes.

Furthermore, ATAG lays the groundwork for *algorithmic editions*, which are designed to be equally navigable by humans and computational systems. By enabling seamless access to uniquely addressable texts, annotations, and metadata, these editions transform digital texts into comprehensive, interdisciplinary resources. The dual emphasis on human readability and machine accessibility fosters new opportunities for textual analysis, intertextual exploration, and computational workflows.

Ultimately, ATAG and its associated advances in ENC-link citation and algorithmic editions redefine digital scholarship by creating dynamic, integrative systems for the modeling, analysis, and application of texts, ensuring their relevance in both traditional and data-driven humanities contexts.

INDEX

Keywords: digital scholarly edition, annotation, citation, TextAsGraph, algorithmic editions

1. Introduction to the Current State of Text Modeling

1 The field of digital scholarly editing has seen significant advances, particularly with the adoption of standards like TEI XML (Cummings 2012) and adherence to FAIR principles.[2] There has been an evolution in hardware, cloud-based infrastructure, and data connectivity through RDF and graph technologies. Along with the impact of virtualization (Rosenthal 2015) on the longevity and accessibility of digital texts, these developments have led to significant improvements.

2 However, digital scholarly editing continues to face technical challenges in text modeling (Cugliana et al. 2024). Critics such as Buzzetti (2002), DeRose (2004), and Haentjens Dekker and Birnbaum (2017) highlight problems of addressability, persistence of text elements, and digital representation of small elements. In addition, digital scholarly editions (DSE) focus on technologies for annotating strings of text, but the existing approaches fail to capture the full range of scholarly activities and observations.

3 One of the challenges in text modeling lies in the duality of models of and for text.[3] In some cases, a description of a text becomes a model of an observable thing in the real world, whereas in other contexts models are built to decode, encode, represent, or remedialize text in other formats. The challenge in text modeling lies in covering various perspectives—from genesis and production to reception and interpretation—and integrating these perspectives into a cohesive framework.[4] But it is important to state that in every case the text reflects what the editor has read or wants to state about the manuscript.

4 While different approaches and encoding strategies have emerged over time, text encoding has in the past treated text as an Ordered Hierarchy of Content Objects (OHCO). Descriptive markup identifies the textual hierarchy and makes it explicit for systematic processing. Although this approach has limitations in managing large numbers of annotations, especially when they overlap, the annotated text remains editable.

5 The field of computational linguistics, by contrast, views text primarily as a sequence of characters, often ignoring other textual features like layout or written language modes (bold, italics, etc.). Annotations are stored in stand-off files (Pagel et al. 2020). This approach has led to significant advances in text handling, transmission, translation, analysis, and use, but the base text cannot be changed after having been indexed.

2. Related Work and Comparative Approaches

6 The ATAG model distinguishes itself from earlier graph-based approaches through its integration of editable base text, recursive annotation, and character-level granularity with persistent UUID-based addressability. Unlike the *Linguistic Annotation Framework* (LAF), which pioneered a graph-based standard for stand-off linguistic annotation (Ide et al. 2003), ATAG supports dynamic updates to the base text while preserving the integrity of annotations—an aspect not addressed by LAF. The Structured Text Annotation Model (*STAM*; Gompel 2023) introduces character-based addressability and recursive annotations with a focus on TEI compatibility, sharing several conceptual foundations with ATAG. However, it does not offer versioning mechanisms or editorial tool support like ATAG's WYSIWYM editor. The *Salt* model, which underpins the ANNIS annotation tool (Zeldes et al. 2009), provides a flexible schema for multilayer annotations but does not model editable text directly, nor does it support persistent, addressable character-level nodes (Zipser

and Romary 2010). The *TextFrame* model from the *Unlocking Digital Texts* initiative (Hughes et al. 2020) takes a different approach by applying IIIF-inspired addressability to textual segments, aligning conceptually with ATAG's URL-based citation scheme, but it focuses more on reference structures than on granular text modeling in a graph. Finally, the W3C *Web Annotation Model* (Sanderson et al. 2017), while offering broad interoperability and character-offset referencing via TextPositionSelector, does not feature recursive annotation structures or persistent identifiers for fine-grained citation under evolving versions. In contrast to all these models, ATAG tightly integrates character-level editability, recursive and multidimensional annotation, and robust versionable citation within a single labeled-property graph framework. This makes it particularly well suited for algorithmic editions, where the interplay between stable addressability, editorial flexibility, and computational readability is essential.

3. Transition to Applied Text as Graph (ATAG)

7 The proposed Applied Text as Graph (ATAG) methodology extends annotations to cover both super- and substructures of texts, using addressable structures for specific graphical phenomena. Unlike TEI XML, which uses hierarchical tree structures for textual representation, ATAG's labeled-property-graph–based model offers distinct advantages in handling overlapping hierarchies and multidimensional annotations. For example, while TEI can represent multiple layers of annotations through stand-off markup, maintaining these layers becomes increasingly complex as the annotation depth grows. In contrast, ATAG's flexible node-based architecture seamlessly integrates multiple annotation layers, allowing for addressability at the character level. Additionally, ATAG inherently supports dynamic updates to the base text without invalidating existing annotations, a feature difficult to achieve in conventional TEI-based workflows.

8 In addition, the new approach provides greater formalization and operationalization of editorial tasks. It supports multiple annotation hierarchies and improves the transparency of the editorial process. In practice, this approach involves making textual data more addressable and using formal languages for tasks such as grapheme normalization.

9 Starting with glyphs is not a universal solution, but has specific advantages, particularly for medieval studies, as medieval texts often exhibit complex abbreviation practices, inconsistent orthography, and visually distinct letterforms. Modeling at the glyph level enables precise

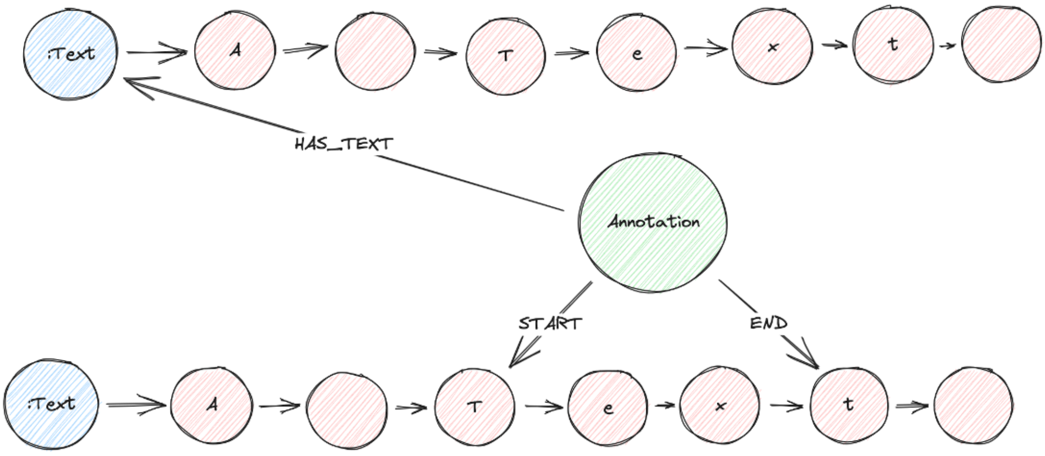
representation of these features before any editorial normalization occurs. UUIDs are applied to each character node, an approach that emphasizes the importance of the grapheme base in textual interpretation and focuses on the lowest level of input. These UUIDs help keep track of the broader narrative and interpretations derived from the text. ATAG's focus on character-level detail and its ability to dynamically update the base text without losing associated annotations is an important step in the field of text modeling. But even the granularity is flexible. In different research contexts with focus on the token level, the text could be modeled in a chain of token nodes, or both character and token nodes could be integrated in one graph.

10 Figure 1 shows an example where each individual character in a text is represented by a "character node" (orange). These nodes are connected in sequence, highlighting the linear arrangement of the text, called the ATAG chain. Each block of text, whether it is a word or a paragraph, begins with a "text node" (blue), which acts as a starting point and root element. These blocks of text can be linked together to suggest a reading order.

11 The framework extends beyond merely representing individual characters; it also allows annotations to be added using "annotation nodes" (green). These annotations can provide semantic information, such as identification of a person, or layout information such as bold or italics. Each annotation node has multiple properties and can be qualified by different annotation systems, for example TEI (Kuczera 2022).[5] The annotation node is associated with the specific characters to which it belongs. Annotations are not isolated. They can be intricately linked to other text nodes; any of these can have comments as text with another character chain, which in turn may be the subject of further annotations, and so on. This recursive capability introduces a multidimensional aspect to text analysis, allowing for a depth of exploration.

12 Consequently, the concept of "Applied Text as Graph" may be envisaged as a dynamic, interactive cartography of text. This model not only assists in navigating through textual content but also facilitates comprehensive analysis and the application of multifaceted annotations.

Figure 1. The basic pattern of ATAG.



13 At the heart of this model lies linear text, which is dissected into discrete units. Each character within this linear continuum is encapsulated within a character node, distinguished by an orange hue, establishing a sequential link to either its predecessor or successor character node. This sequentiality is crucial, as it mirrors the natural flow of reading, thereby preserving the inherent order of the text. ATAG focuses on the text as it is read and interpreted, rather than replicating the layout of the source page as often done in TEI XML. Additional details, such as layout or formatting, are stored in annotations, allowing the core text to remain accessible and editable. This reader-centric approach enables scholars to engage with the text's content directly while preserving contextual information in a structured and flexible manner.

14 Branched texts—such as alternative readings—are represented as annotations rather than integrated directly into the linear text. These annotations are linked to specific character nodes within the primary sequence, ensuring that their association with the main text is precise and contextually meaningful. This approach allows the primary text to remain uninterrupted and readable while providing seamless access to additional layers of information. Scholars can explore these additional layers as needed, and even switch the readings between the main text and the branched texts in the annotation.

15 Beyond individual characters, the model extends to encompass larger textual fragments. Each piece of text is inherently connected to adjacent pieces, thereby suggesting a proposed reading order. This connectivity begins at the text node, marking the starting point for each textual segment. The meticulous granularity of this model, extending down to the individual character level, ensures high precision in text representation.

16 Each character is addressed by one UUID, ensuring that every character of the text is distinctly identifiable and accessible. This feature is pivotal for detailed analysis and referencing, allowing scholars to pinpoint and interact with the smallest components of the text. For further processing, the chain of character nodes is collected and stored in the text property of the text node. This is the basis for the web publishing system (Neill and Kuczera 2019) used in the Socinian Letter (http://www.sozinianer.de/) and the Liber Epistolarum projects (http://www.liberepistolarum.de/).

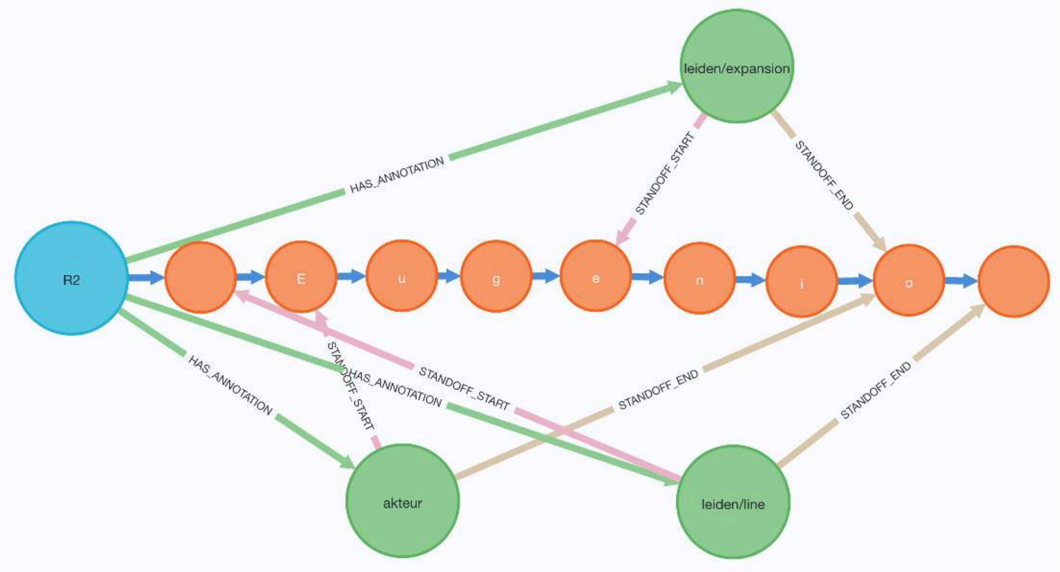
4. Implementation

4.1 Character Structures

4.1.1 An Example from the Liber Epistolarum Project

17 In figure 2 we see the text node of letter R2 of the Liber Epistolarum of Hildegard of Bingen in the graph database Neo4j, with the subsequent chain of character nodes.[6] The text in the chain of character nodes is also stored in the text property of the text node, and the text nodes are used for the publication environment at https://liberepistolarum.mni.thm.de/.

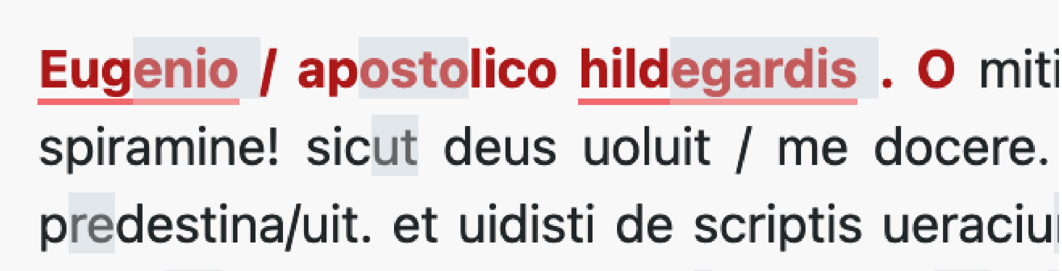
Figure 2. The text node of letter R2 of the Liber Epistolarum of Hildegard of Bingen in the graph database Neo4j, with the first character nodes and annotations.



18 So the ATAG-Chain is crucial for editing, annotating, and for algorithmic research, whereas the text node with the according corresponding text property is used for the publication system. Our system makes sure that the character chain and the text property are always in sync.

19 From the text node, the chain of character nodes starts with the sequence of characters of the token “Eugenio” (figure 3). The green annotation nodes show that the “enio” part of the first token has been expanded by the editor. In the original there are only the first three characters. The two lower annotation nodes show that Eugenio is annotated as an actor in the DSE, and the second shows a line annotation.

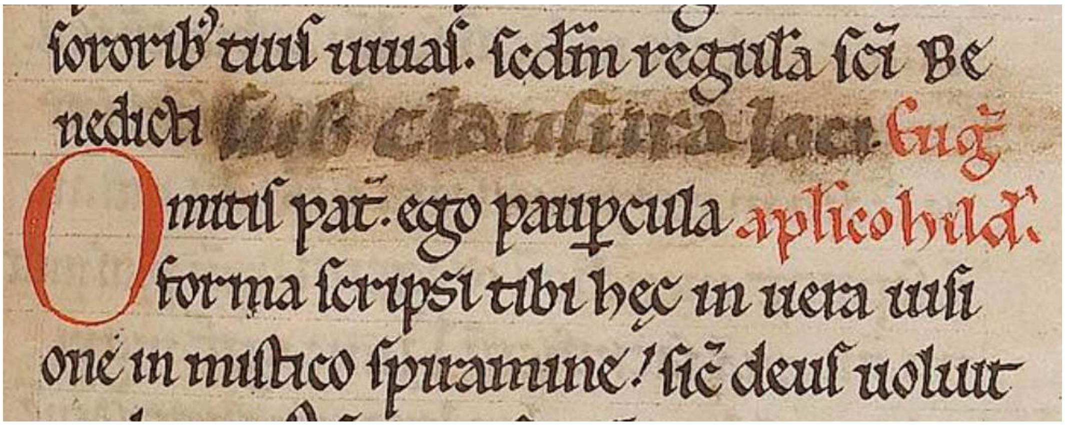
Figure 3. The first words of letter R2 in the web publication environment.



Eugenio / apostolico hildegardis . O miti
spiramine! sicut deus uoluit / me docere.
predestina/uit. et uidisti de scriptis ueraciu

20 It is important to emphasize that the main body of the text serves as the normalized version; that is, it contains the standard or extended form of the content. This arrangement allows for a clean and readable base text, while still providing a way to include additional forms or information via annotations (figure 4).

Figure 4. First words of letter R2 in the original (Wiesbaden, Hochschul- und Landesbibliothek RheinMain [formerly Wiesbaden, Hessische Landesbibliothek]), Hs 2 (“Riesenkodex”). Parchment, 481 folios, 490 × 320 mm, 2 cols., Rupertsberg (doi: 10.25534/tudigit-133).

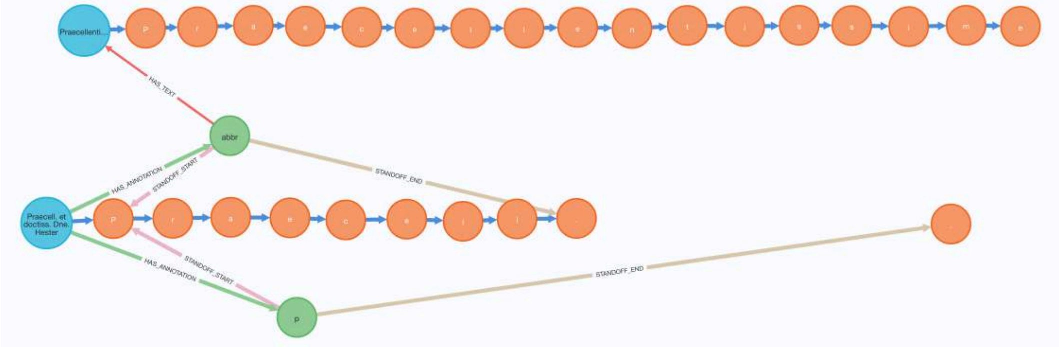


sororib' tuis uiuas. secdm regula sci Be
nedicti sub clausura loci. Eug
Omniu pat. ego paupcula aplico hild.
forma scripsi tibi hec in uera uisi
one in mistico spiramine. sic deus uoluit

4.1.2 Character Structure in a Socinian Letter

21 In the ATAG framework used in the Socinian Letters project, the base text consists of the direct transcription of the original version, capturing the content in its original form. Figure 5 illustrates the character structure of a Socinian letter by Johannes Müller, highlighting the base text node (blue) and associated annotation nodes (green) used to capture semantic elements such as abbreviations (https://sozinianer.mni.thm.de/view/ed_m3l_ldy_vkb). The character chain starts from this node, and the chain is annotated by an SPO node of type `"abbr"`, which covers the semantics of the TEI element `<abbreviation>` (https://tei-c.org/Vault/P5/2.2.0/doc/tei-p5-doc/en/html/ref-abbr.html).

Figure 5. The letter by Johannes Müller (Hamburg) to Stanisław Lubieniecki (Hamburg), 5 January 1665 (https://sozinianer.mni.thm.de/view/ed_m3l_ldy_vkb), in the graph database. Every text node has a text property which contains the content of the character chain. This property is updated after every change of the chain. It is used to speed things up in the publication environment.



22 The normalized version of the abbreviated text in the base text chain is covered by the text node in the upper left corner and the associated character chain. Any extensions, elaborations, or additional details are thus stored in the annotations. This setup allows for a clean and unaltered base text, while still providing the ability to add more information or context through annotations. In this way, the reader can engage with the original text while having the opportunity to explore the expanded material, all without cluttering up the main text.

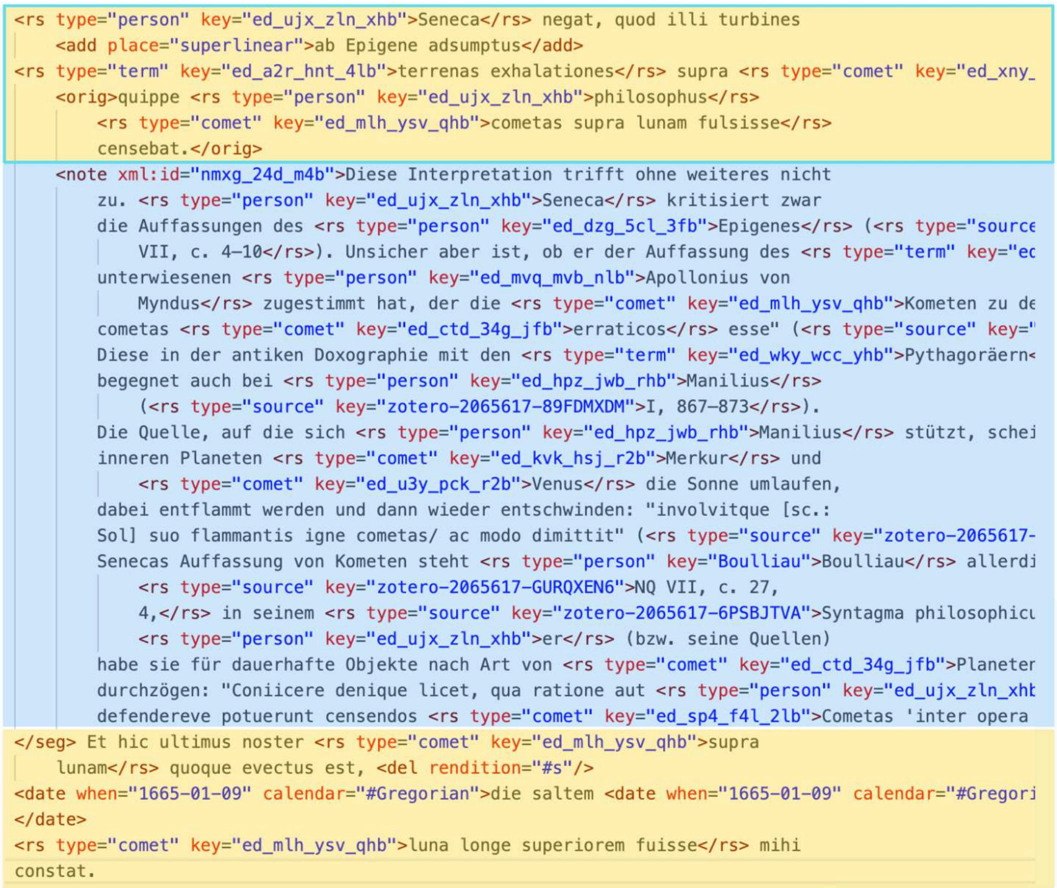
23 It is important to note that the two projects follow different editorial rules. The Liber Epistolarum Project has the normalized text in the base text chain with the expansions annotated, whereas the Socinian Letter Project has the original text in the base text chain and the normalized version in an annotation. The ATAG system does not force either of these solutions. It can handle both; it is up to the editors of the projects to decide.

24 One of the most flexible features of ATAG is its ability to dynamically edit the base text, even after annotations have been added. This makes it easy to update or change the text as required without losing the associated annotations. This dynamic nature of the base text is particularly useful in evolving projects or collaborative environments where the text may need to be adapted frequently. The user must be careful that changes to the text do not affect the meaning and structure of the annotations.

4.2 Text Structures

25 This section illustrates the structures and characteristics of the ATAG with a letter by Ismaël Boulliau (Paris) to Stanisław Lubieniecki (Hamburg), 6 March 1665 (https://sozinianer.mni.thm.de/id/MAIN_ed_kbj_wfw_xmb), from XML to the labeled property graph database Neo4j to the project's web publication system (see figure 6).

Figure 6. XML fragment of the letter by Ismaël Boulliau (Paris) to Stanisław Lubieniecki (Hamburg), 6 March 1665, with a comment in a note element in XML (https://www.sozinianer.de/id/MAIN_ed_kbj_wfw_xmb). The lines are truncated for readability.

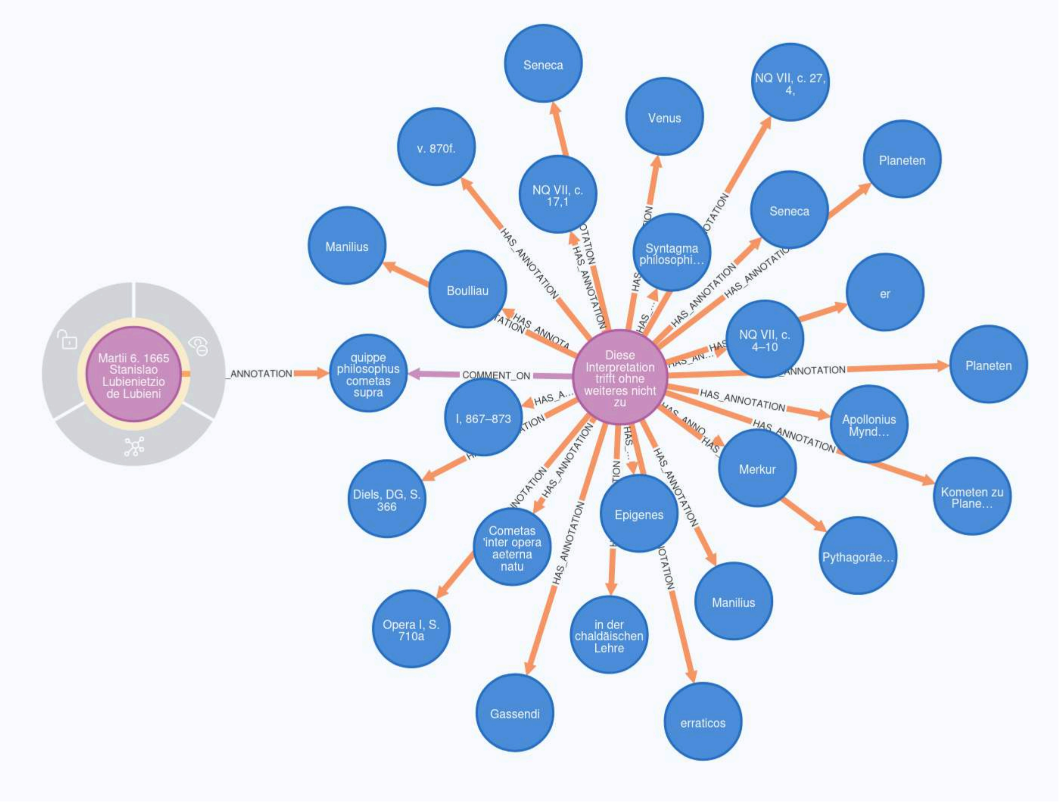


```
<rs type="person" key="ed_ujx_zln_xhb">Seneca</rs> negat, quod illi turbines
    <add place="superlinear">ab Epigene adsumptus</add>
<rs type="term" key="ed_a2r_hnt_4lb">terrenas exhalationes</rs> supra <rs type="comet" key="ed_xny_
    <orig>quippe <rs type="person" key="ed_ujx_zln_xhb">philosophus</rs>
        <rs type="comet" key="ed_mlh_ysv_qhb">cometas supra lunam fulsisse</rs>
        censebat.</orig>
    <note xml:id="nmxg_24d_m4b">Diese Interpretation trifft ohne weiteres nicht
        zu. <rs type="person" key="ed_ujx_zln_xhb">Seneca</rs> kritisiert zwar
        die Auffassungen des <rs type="person" key="ed_dzg_5cl_3fb">Epigenes</rs> (<rs type="source
            VII, c. 4–10</rs>). Unsicher aber ist, ob er der Auffassung des <rs type="term" key="ec
        unterwiesenen <rs type="person" key="ed_mvq_mvb_nlb">Apollonius von
            Myndus</rs> zugestimmt hat, der die <rs type="comet" key="ed_mlh_ysv_qhb">Kometen zu de
        cometas <rs type="comet" key="ed_ctd_34g_jfb">erraticos</rs> esse" (<rs type="source" key="
        Diese in der antiken Doxographie mit den <rs type="term" key="ed_wky_wcc_yhb">Pythagoräern<
        begegnet auch bei <rs type="person" key="ed_hpz_jwb_rhb">Manilius</rs>
            (<rs type="source" key="zotero-2065617-89FDMXDM">I, 867–873</rs>).
        Die Quelle, auf die sich <rs type="person" key="ed_hpz_jwb_rhb">Manilius</rs> stützt, schei
        inneren Planeten <rs type="comet" key="ed_kvk_hsj_r2b">Merkur</rs> und
            <rs type="comet" key="ed_u3y_pck_r2b">Venus</rs> die Sonne umlaufen,
        dabei entflammt werden und dann wieder entschwinden: "involvitque [sc.:
        Sol] suo flammantis igne cometas/ ac modo dimittit" (<rs type="source" key="zotero-2065617-
        Senecas Auffassung von Kometen steht <rs type="person" key="Boulliau">Boulliau</rs> allerdi
            <rs type="source" key="zotero-2065617-GURQXEN6">NQ VII, c. 27,
            4,</rs> in seinem <rs type="source" key="zotero-2065617-6PSBJTVA">Syntagma philosophicu
            <rs type="person" key="ed_ujx_zln_xhb">er</rs> (bzw. seine Quellen)
        habe sie für dauerhafte Objekte nach Art von <rs type="comet" key="ed_ctd_34g_jfb">Planeter
        durchzögen: "Coniicere denique licet, qua ratione aut <rs type="person" key="ed_ujx_zln_xht
        defendereve potuerunt censendos <rs type="comet" key="ed_sp4_f4l_2lb">Cometas 'inter opera
</seg> Et hic ultimus noster <rs type="comet" key="ed_mlh_ysv_qhb">supra
    lunam</rs> quoque evectus est, <del rendition="#s"/>
<date when="1665-01-09" calendar="#Gregorian">die saltem <date when="1665-01-09" calendar="#Gregori
</date>
<rs type="comet" key="ed_mlh_ysv_qhb">luna longe superiorem fuisse</rs> mihi
constat.
```

26 The XML fragment begins with the text of the letter: "Seneca negat, quod illi turbines ...," marked in yellow. In the seventh line, a <note> element marked in blue indicates the beginning of a commentary on the text "quippe philosophus cometas supra lunam fulsisse censebat." After the comment, the letter text continues with "Et hic ultimus noster supra luna" marked in yellow again. The example shows how TEI/ XML works with its inline markup. The XML file itself is a stream

of Unicode characters that only makes sense as a hierarchical XML structure. It mixes different hierarchies of information, in our example the text of the transcription of the letter and the text of the editor's commentary. In ATAG, these different hierarchies of information are kept separate. The next figure shows the same part of the letter and the annotation in the graph. The explicit chains of characters are not shown in order not to lose the overview.

Figure 7. Graph model of letter (left), the annotation node, the text node of the annotation with the commenting text (center), and the annotations in the annotation text (blue).

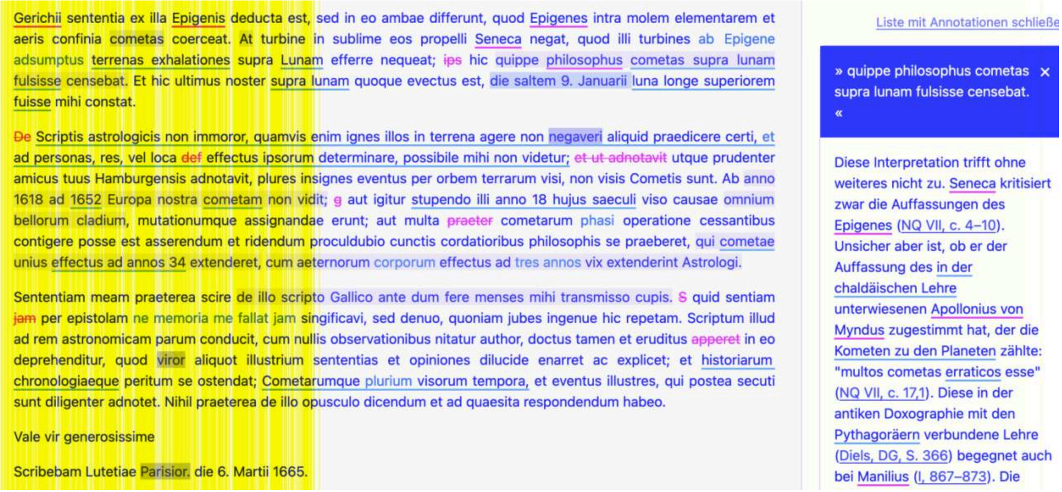


27 In figure 7 the left node is the text node of the letter, which has the base text of the letter's transcription in its text property. From the text node a HAS_ANNOTATION edge goes to a blue annotation node. This annotation node has the annotated text ("quippe philosophus cometas supra lunam fulsisse censebat") in its text property and the start and end index of the annotation in the letter text. From the annotation node a COMMENT_ON-edge goes to another text node with the text of the comment. This comment is then annotated with the other annotation nodes around. All

annotation nodes linked to the annotation text node represent further annotations on the content of the comment itself. Of course, these annotations can be annotated, and so on, and then we are in a network of texts.

28 The graph model is not well suited to reading and understanding the text and structures by humans, but is ideal for computers and algorithms, which can transform the material into presentations that are human-readable. Figure 8 shows the same letter with the annotation and the comment in the project’s web publishing environment.

Figure 8. Letter from Ismaël Boulliau (Paris) to Stanisław Lubieniecki (Hamburg), 6 March 1665 in the project’s web-publication system. (https://sozinianer.mni.thm.de/id/MAIN_ed_kbj_wfw_xmb).



Gerichii sententia ex illa Epigenis deducta est, sed in eo ambae differunt, quod Epigenes intra molem elementarem et aeris confinia cometas coerceat. At turbine in sublime eos propelli Seneca negat, quod illi turbines ab Epigene adsumptus terrenas exhalationes supra Lunam efferre nequeat; ~~ips~~ hic quippe philosophus cometas supra lunam fulsisse censebat. Et hic ultimus noster supra lunam quoque evectus est, die saltem 9. Januarii luna longe superiorem fuisse mihi constat.

~~De~~ Scriptis astrologicis non immoror, quamvis enim ignes illos in terrena agere non negaveri aliquid praedicere certi, et ad personas, res, vel loca ~~def~~ effectus ipsorum determinare, possibile mihi non videtur; ~~et ut adnotavit~~ utque prudenter amicus tuus Hamburgensis adnotavit, plures insignes eventus per orbem terrarum visi, non visis Cometis sunt. Ab anno 1618 ad 1652 Europa nostra cometam non vidit; ~~g~~ aut igitur stupendo illi anno 18 hujus saeculi viso causae omnium bellorum cladium, mutationumque assignandae erunt; aut multa ~~praeter~~ cometarum phasi operatione cessantibus contigere posse est asserendum et ridendum proculdubio cunctis cordatioribus philosophis se praeberet, qui cometae unius effectus ad annos 34 extenderet, cum aeternorum corporum effectus ad tres annos vix extenderint Astrologi.

Sententiam meam praeterea scire de illo scripto Gallico ante dum fere menses mihi transmisso cupis. ~~S~~ quid sentiam ~~jam~~ per epistolam ne memoria me fallat jam singificavi, sed denuo, quoniam jubes ingenue hic repetam. Scriptum illud ad rem astronomicam parum conducit, cum nullis observationibus nitatur author, doctus tamen et eruditus ~~apperet~~ in eo deprehenditur, quod viror aliquot illustrium sententias et opiniones dilucide enarret ac explicet; et historiarum chronologiaeque peritum se ostendat; Cometarumque plurium visorum tempora, et eventus illustres, qui postea secuti sunt diligenter adnotet. Nihil praeterea de illo opusculo dicendum et ad quaesita respondendum habeo.

Vale vir generosissime

Scribebam Lutetiae Parisior. die 6. Martii 1665.

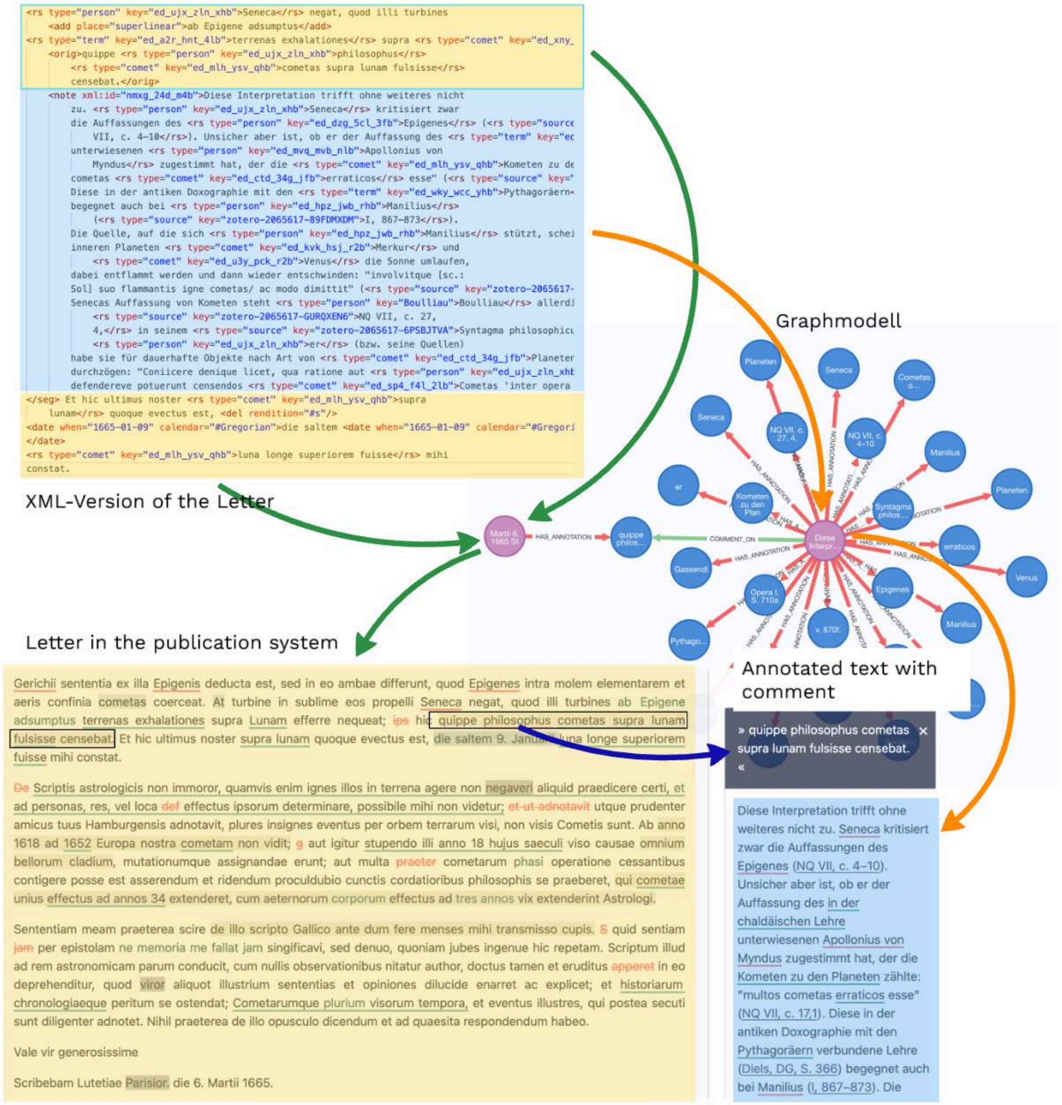
» quippe philosophus cometas supra lunam fulsisse censebat. « ×

Diese Interpretation trifft ohne weiteres nicht zu. Seneca kritisiert zwar die Auffassungen des Epigenes (NQ VII, c. 4–10). Unsicher aber ist, ob er der Auffassung des in der chaldäischen Lehre unterwiesenen Apollonius von Myndus zugestimmt hat, der die Kometen zu den Planeten zählte: "multos cometas erraticos esse" (NQ VII, c. 17,1). Diese in der antiken Doxographie mit den Pythagoräern verbundene Lehre (Diels, DG, S. 366) begegnet auch bei Manilius (I, 867–873). Die

29 On the left is the basic text of the transcription. In the first paragraph we see the annotated text “quippe philosophus cometas supra lunam fulsisse censebat.” In the right column, the annotation is opened in an annotation box, starting with a fragment of the annotated text in the dark blue block, followed by the comment, which is also annotated.

30 Figure 9 shows the combined scenarios. The arrows show which structures in one model belong to which structures in the other.

Figure 9. Collage of the three examples mentioned above. The yellow boxes cover the basic text of the letter, while the blue boxes show the text of the comment. The graph model in between shows the structure of the edition’s graph database.



```
<rs type="person" key="ed_ujx_zln_xhb">Seneca</rs> negat, quod illi turbines
    <add place="superlinear">ab Epigene adsumptus</add>
<rs type="term" key="ed_a2r_hnt_4lb">terrenas exhalationes</rs> supra <rs type="comet" key="ed_xny_
    <orig>quippe <rs type="person" key="ed_ujx_zln_xhb">philosophus</rs>
        <rs type="comet" key="ed_mlh_ysv_qhb">cometas supra lunam fulsisse</rs>
        censebat.</orig>
    <note xml:id="nmxg_24d_m4b">Diese Interpretation trifft ohne weiteres nicht
        zu. <rs type="person" key="ed_ujx_zln_xhb">Seneca</rs> kritisiert zwar
        die Auffassungen des <rs type="person" key="ed_dzg_5cl_3fb">Epigenes</rs> (<rs type="source
            VII, c. 4–10</rs>). Unsicher aber ist, ob er der Auffassung des <rs type="term" key="ec
        unterwiesenen <rs type="person" key="ed_mvq_mvb_nlb">Apollonius von
            Myndus</rs> zugestimmt hat, der die <rs type="comet" key="ed_mlh_ysv_qhb">Kometen zu de
        cometas <rs type="comet" key="ed_ctd_34g_jfb">erraticos</rs> esse" (<rs type="source" key="
        Diese in der antiken Doxographie mit den <rs type="term" key="ed_wky_wcc_yhb">Pythagoräern<
        begegnet auch bei <rs type="person" key="ed_hpz_jwb_rhb">Manilius</rs>
            (<rs type="source" key="zotero-2065617-89FDMXDM">I, 867–873</rs>).
        Die Quelle, auf die sich <rs type="person" key="ed_hpz_jwb_rhb">Manilius</rs> stützt, schei
        inneren Planeten <rs type="comet" key="ed_kvk_hsj_r2b">Merkur</rs> und
            <rs type="comet" key="ed_u3y_pck_r2b">Venus</rs> die Sonne umlaufen,
        dabei entflammt werden und dann wieder entschwinden: "involvitque [sc.:
        Sol] suo flammantis igne cometas/ ac modo dimittit" (<rs type="source" key="zotero-2065617-
        Senecas Auffassung von Kometen steht <rs type="person" key="Boulliau">Boulliau</rs> allerdi
            <rs type="source" key="zotero-2065617-GURQXEN6">NQ VII, c. 27,
            4,</rs> in seinem <rs type="source" key="zotero-2065617-6PSBJTVA">Syntagma philosophicu
            <rs type="person" key="ed_ujx_zln_xhb">er</rs> (bzw. seine Quellen)
        habe sie für dauerhafte Objekte nach Art von <rs type="comet" key="ed_ctd_34g_jfb">Planeter
        durchzögen: "Coniicere denique licet, qua ratione aut <rs type="person" key="ed_ujx_zln_xht
        defendereve potuerunt censendos <rs type="comet" key="ed_sp4_f4l_2lb">Cometas 'inter opera
</seg> Et hic ultimus noster <rs type="comet" key="ed_mlh_ysv_qhb">supra
    lunam</rs> quoque evectus est, <del rendition="#s"/>
<date when="1665-01-09" calendar="#Gregorian">die saltem <date when="1665-01-09" calendar="#Gregori
</date>
<rs type="comet" key="ed_mlh_ysv_qhb">luna longe superiorem fuisse</rs> mihi
constat.
```

Gerichii sententia ex illa Epigenis deducta est, sed in eo ambae differunt, quod Epigenes intra molem elementarem et aeris confinia cometas coerceat. At turbine in sublime eos propelli Seneca negat, quod illi turbines ab Epigene adsumptus terrenas exhalationes supra Lunam efferre nequeat; ~~ips~~ hic quippe philosophus cometas supra lunam fulsisse censebat. Et hic ultimus noster supra lunam quoque evectus est, die saltem 9. Januarii luna longe superiorem fuisse mihi constat.

~~De~~ Scriptis astrologicis non immoror, quamvis enim ignes illos in terrena agere non negaveri aliquid praedicere certi, et ad personas, res, vel loca ~~def~~ effectus ipsorum determinare, possibile mihi non videtur; ~~et ut adnotavit~~ utque prudenter amicus tuus Hamburgensis adnotavit, plures insignes eventus per orbem terrarum visi, non visis Cometis sunt. Ab anno 1618 ad 1652 Europa nostra cometam non vidit; g aut igitur stupendo illi anno 18 hujus saeculi viso causae omnium bellorum cladium, mutationumque assignandae erunt; aut multa ~~praeter~~ cometarum phasi operatione cessantibus contigere posse est asserendum et ridendum proculdubio cunctis cordatioribus philosophis se praeberet, qui cometae unius effectus ad annos 34 extenderet, cum aeternorum corporum effectus ad tres annos vix extenderint Astrologi.

Sententiam meam praeterea scire de illo scripto Gallico ante dum fere menses mihi transmisso cupis. ~~S~~ quid sentiam ~~jam~~ per epistolam ne memoria me fallat jam singificavi, sed denuo, quoniam jubes ingenue hic repetam. Scriptum illud ad rem astronomicam parum conducit, cum nullis observationibus nitatur author, doctus tamen et eruditus ~~apperet~~ in eo deprehenditur, quod viror aliquot illustrium sententias et opiniones dilucide enarret ac explicet; et historiarum chronologiaeque peritum se ostendat; Cometarumque plurium visorum tempora, et eventus illustres, qui postea secuti sunt diligenter adnotet. Nihil praeterea de illo opusculo dicendum et ad quaesita respondendum habeo.

Vale vir generosissime

Scribebam Lutetiae Parisior. die 6. Martii 1665.

» quippe philosophus cometas supra lunam fulsisse censebat. « ×

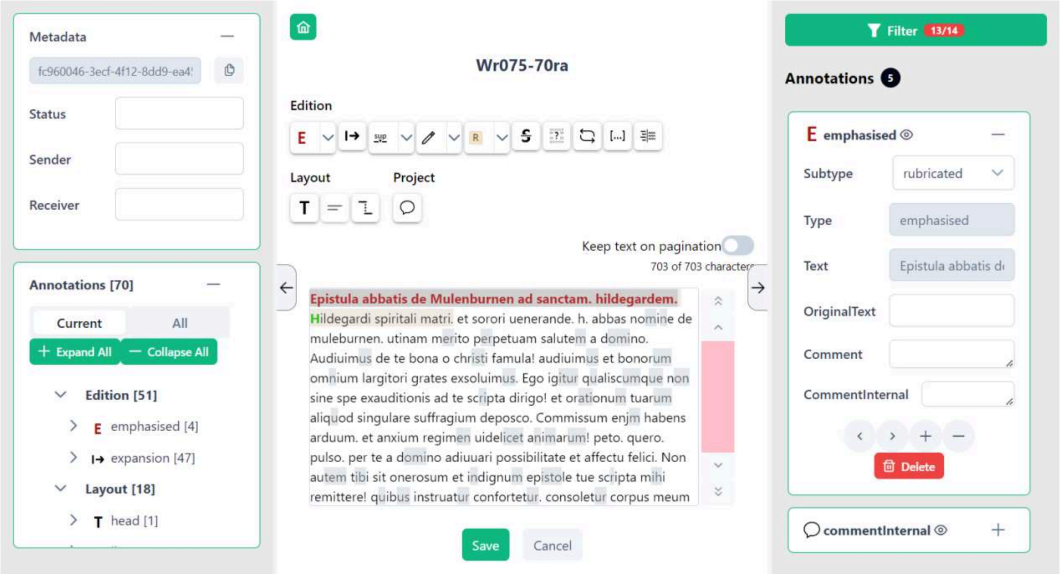
Diese Interpretation trifft ohne weiteres nicht zu. Seneca kritisiert zwar die Auffassungen des Epigenes (NQ VII, c. 4–10). Unsicher aber ist, ob er der Auffassung des in der chaldäischen Lehre unterwiesenen Apollonius von Myndus zugestimmt hat, der die Kometen zu den Planeten zählte: "multos cometas erraticos esse" (NQ VII, c. 17,1). Diese in der antiken Doxographie mit den Pythagoräern verbundene Lehre (Diels, DG, S. 366) begegnet auch bei Manilius (I, 867–873). Die

5. The Editor

31 Since there was no suitable tool for creating, editing, annotating, and saving texts modeled directly in the graph, the web-based ATAG text editor was implemented. In contrast to previous tools such as the Standoff Property Editor SPEEDy (Neill and Kuczera 2019, Neill and Schmidt 2021), which reached its limits, the ATAG editor can handle text volumes of any size efficiently. While ATAG

assigns UUIDs to every character node to ensure precise and persistent addressability, this level of granularity is used primarily during the editorial process. For example, a 50 MB TEI XML file results in approximately 7 GB of graph storage with full annotation structures and UUIDs at the character level. For publication and long-term storage, however, only the compacted graph version —typically around 1.1 GB—is used, which includes character-level linking but omits transient editing metadata. Additionally, UUIDs are only instantiated when annotations or versioning requires them, significantly reducing overhead in annotation-light sections. This tiered storage model ensures that the system remains efficient and scalable, even with larger corpora. The previous editor loaded the whole text in the browser DOM, and became very slow when dealing with long texts (>5,000 characters) with many annotations.

Figure 10. Screen shot of the ATAG-Editor.



Metadata: fc960046-3ecf-4f12-8dd9-ea4!

Wr075-70ra

Keep text on pagination

703 of 703 character

Epistula abbatis de Mulenburnen ad sanctam. hildegardem. Hildegardi spiritali matri. et sorori uenerande. h. abbas nomine de muleburnen. utinam merito perpetuam salutem a domino. Audiuimus de te bona o christi famula! audiuimus et bonorum omnium largitori grates exsoluimus. Ego igitur qualiscumque non sine spe exauditionis ad te scripta dirigo! et orationum tuarum aliquod singulare suffragium deposco. Commissum enjm habens arduum. et anxium regimen uidelicet animarum! peto. quero. pulso. per te a domino adiuuari possibilitate et affectu felici. Non autem tibi sit onerosum et indignum epistole tue scripta mihi remittere! quibus instruatur confortetur. consoletur corpus meum

E emphasised

Subtype: rubricated

Type: emphasised

Text: Epistula abbatis d

32 With the experience gained from the first project phase of the Liber Epistolarum project, the ATAG editor was created as a WYSIWYM (what-you-see-is-what-you-mean) editor (figure 10), which loads only chunks of text, pulled from the Character-Chain in the graph db. The size of the chunk can be changed according to the editor's need. Annotations are marked up using visual elements such as underlining, font color, or background color, so that the base text remains legible. Such approaches are already used by annotation tools such as CATMA (https://catma.de/) or INCEpTION (https://inception-project.github.io/), but the ATAG editor makes it possible for the

first time to annotate and edit the underlying text simultaneously. The intuitive user interface, in which annotations can be added by highlighting the text and clicking buttons, makes it particularly user-friendly. In the next stage, a Context Management Tool (CMT) will be implemented to organize the connections between the different ATAG nodes and other contextual information such as metadata and norm data.

6. ENC-links citations with ATAG

33 Citations are fundamental to scholarly work, serving both as a means to credit sources and as a tool for tracing the origins of ideas. In the humanities and increasingly in the digital humanities, precise referencing is vital. This practice not only maintains academic integrity but also aids in uncovering intertextual relationships across disciplines. Traditionally, citations in scholarly editions were tied to the physical structure of texts, referencing specific page numbers. While effective for printed works, this method proves limiting in digital formats, where dynamic layouts often render the concept of "pages" irrelevant. To address these challenges, the W3C Web Annotation Model (https://www.w3.org/TR/annotation-model/) was developed, enabling annotations across digital documents in a standardized, interoperable framework. This model supports diverse annotation types, from text comments to multimedia, embracing the decentralized nature of the web. Its flexibility, however, comes at the cost of granularity, which can be essential for academic citations.

34 ATAG (Applied Text as Graph) overcomes this limitation by making every character within a text uniquely addressable. This fine-grained approach, called ENC-links (Explicitly Notated Citations) allows for precise citations, offering unparalleled accuracy for scholarly referencing. ENC-links use URL-based citation structures to directly reference specific text segments. This approach ensures transparency and ease of verification. For example: https://www.sozinianer.de/id/MAIN_ed_kbj_wfw_xmb?guid=ed_kbj_wfw_xmb&s=7261&e=7324.

Address of the edition	https://www.sozinianer.de/id/
ID of the communication	MAIN_ed_kbj_wfw_xmb
ID of the variant	ed_kbj_wfw_xmb

Startindex	7261
Endindex	7324

35 In this URL:

- The base address identifies the edition.
- Unique IDs specify the communication context and text variant.
- Start and end indices pinpoint the exact cited text.
- Opening this link in a browser highlights the cited text, streamlining access and validation.

A key feature of ATAG is its versioning capability. Future citations will leverage character-node UUIDs by including versioning the text with git and commit hashes, ensuring that citations remain valid even if the base text evolves. Additionally, incorporating text hashes in citation links will allow integrity checks. These measures ensure that citations remain robust and trustworthy. The citation function works on the website sozinianer.de and can be tested there. The steps to a citation link are explained step by step at https://sozinianer.de/project/digital-citations.

36 ATAG also supports multiple serialization formats, enabling both human-readable and machine-actionable citation outputs. By fostering both user accessibility and algorithmic exploration, this dual capability marks a step toward creating *algorithmic editions*—resources equally navigable by humans and machines—and thus broadening the scope and depth of digital scholarship.

7. Toward an Algorithmic Edition

37 An algorithmic edition emphasizes the structuring of texts and annotations for computational engagement while maintaining accessibility for traditional scholarly use. Unlike conventional digital editions, which often rely on visual interfaces for user interaction, algorithmic editions prioritize the precise structuring of textual elements and their interrelations. The Applied Text as Graph (ATAG) methodology, combined with ENC-link citations, provides the foundation for algorithmic editions. ATAG's granular text representation, which uniquely addresses each character, ensures that textual components, annotations, and contextual metadata are directly

accessible. ENC-links extend this functionality by offering precise, URL-based references to specific text segments. These citations integrate seamlessly into computational workflows, enabling algorithmic analysis of both texts and their annotations. By leveraging these innovations, algorithmic editions transform texts into dynamic, machine-readable resources. They allow for the retrieval of specific text segments, along with associated metadata such as semantic annotations and linguistic information. This ensures that the edition is as usable for algorithms as it is for human readers, supporting interdisciplinary research and detailed textual analysis. Unlike traditional editions, algorithmic editions do not depend on visual interfaces like websites; they can exist purely as structured, machine-actionable data. This flexibility enables their integration into diverse digital ecosystems, fostering advanced research applications and interdisciplinary collaboration.

38 In essence, ATAG and ENC-links together make algorithmic editions a practical reality, bridging the gap between human-centered scholarship and computational analysis.

8. Conclusion

39 Digital scholarly editing has traditionally centered on the digitization and representation of texts, with an emphasis on visual presentation and human interaction. The introduction of the “Applied Text as Graph” (ATAG) methodology, the ENC-links, and the emerging concept of algorithmic editions, however, signals a transformative shift in the field. These innovations redefine how texts are modeled, analyzed, and used in the digital humanities.

40 ATAG revolutionizes text modeling by breaking down texts into their smallest units, making every character uniquely addressable. This granular approach allows for precise annotations and dynamic editing without compromising the integrity of the text or its associated metadata. The result is a highly flexible and interoperable framework that supports both human-readable editions and machine-actionable data. ATAG integrates multiple annotation hierarchies and fosters a multidimensional understanding of texts, enabling scholars to traverse and analyze textual content with unprecedented depth.

41 The concept of algorithmic editions complements and extends this vision. By prioritizing machine readability, algorithmic editions enable texts, annotations, and metadata to be seamlessly integrated into computational workflows. These editions are not limited to visual interfaces, but

exist as networks of interconnected and uniquely addressable data points. ENC-links provide precise access to specific text segments and their contextual information, such as semantic annotations and linguistic metadata, making these editions valuable for both scholarly and algorithmic exploration.

42 Together, these developments mark a shift in focus from static representations of texts to dynamic systems that accommodate evolving scholarly needs.

BIBLIOGRAPHY

Buzzetti, Dino. 2002. "Digital Representation and the Text Model." *New Literary History* 33 (1): 61–88. https://doi.org/10.1353/nlh.2002.0003.

Cugliana, E., A. Ward, J. J. van Zundert, A. Kuczera, and M. Grüntgens. 2024. "Computational Approaches and the Epistemology of Scholarly Editing." *International Journal of Digital Humanities* 6: 169–188. https://doi.org/10.1007/s42803-024-00088-z.

Cummings, James. 2012. "The Compromises and Flexibility of TEI Customisation." *Proceedings of the Digital Humanities Congress*. Sheffield. https://www.dhi.ac.uk/books/dhc2012.

DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear. 1990. "What Is Text, Really?" *Journal of Computing in Higher Education* 1 (2): 3–26. https://doi.org/10.1007/BF02941632.

Gompel, Maarten van. 2023. *STAM*. CLARIAH/Huygens Instituut: https://annotation.github.io/stam/.

Haentjens Dekker, Ronald, and David J. Birnbaum. 2017. "It's More Than Just Overlap: Text As Graph." In *Balisage, The Markup Conference 2017*. Washington, DC. Accessed 17 February 2024. https://doi.org/10.4242/BalisageVol19.Dekker01.

Hughes, Lorna M., Pip Willcox, and Matthew McGrattan. 2020. *Unlocking Digital Texts: A Framework for Addressable Textual Data*. https://textframe.io/.

Ide, N., L. Romary, and E. de la Clergerie. 2003. *International Standard for a Linguistic Annotation Framework*. ISO/TC 37/SC 4 Working Group Report. https://aclanthology.org/W03-0804.pdf.

Kuczera, Andreas. 2022. "TEI Beyond XML—Digital Scholarly Editions as Provenance Knowledge Graphs." In *Graph Technologies in the Humanities—Proceedings 2020*, ed. Tara Andrews, Franziska Diehr, Thomas Efer, Andreas Kuczera, and Joris van Zundert. http://ceur-ws.org/Vol-3110/paper6.pdf.

Neill, Iian, and Andreas Kuczera. 2019. “The Codex—An Atlas of Relations.” In *Die Modelierung des Zweifels —Schlüsselideen und -konzepte zur graphbasierten Modelierung von Unsicherheiten*, ed. Andreas Kuczera, Thorsten Wübbena, and Thomas Kollatz. Wolfenbüttel (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände: 4). https://doi.org/10.17175/SB004_008.

Neill, Iian, and Desmond Schmidt. 2021. “SPEEDy: A Practical Editor for Texts Annotated with Standoff Properties.” In Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing, 45–54 (http://nbn-resolving.de/urn:nbn:de:hbz:38-552241).

Pagel, Janis, Nils Reiter, Ina Rösiger, and Sarah Schulz. 2020. “Annotation als flexibel einsetzbare Methode.” In *Reflektierte algorithmische Textanalyse*, ed. Nils Reiter, Axel Pichler, and Jonas Kuhn, 125–42. De Gruyter. https://doi.org/10.1515/9783110693973-006.

Pierazzo, Elena. 2016. *Digital Scholarly Editing: Theories, Models and Methods.* London: Routledge.

Rosenthal, David S. H. 2015. *Emulation and Virtualization as Preservation Strategies.* University of North Texas Libraries, UNT Digital Library. https://digital.library.unt.edu/ark:/67531/metadc799755/.

Sanderson, Robert, Paolo Ciccarese, and Benjamin Young, eds. 2017. *Web Annotation Data Model.* W3C Recommendation. https://www.w3.org/TR/annotation-model/.

Zeldes, A., J. Ritz, A. Lüdeling, and C. Chiarcos. 2009. “ANNIS: A Search Tool for Multi-Layer Annotated Corpora.” In *Proceedings of Corpus Linguistics* 2009, July 20–23, Liverpool. https://ucrel.lancs.ac.uk/publications/CL2009/358_FullPaper.doc.

Zipser, Florian, and Laurent Romary. 2010. “A Model-Oriented Approach to the Mapping of Annotation Formats Using Standards.” Workshop on Language Resource and Language Technology Standards, LREC 2010, May 2010, La Valette, Malta. inria-00527799. https://inria.hal.science/inria-00527799v1.

NOTES

[1] https://en.wikipedia.org/wiki/Graph_database#Labeled-property_graph.

[2] The Socinian Letter project adopts the semantics of TEI to ensure standardized and interoperable annotations across projects, leveraging its robust framework widely recognized in the digital humanities.

[3] Cf. DeRose et al. 1990 for an early and influential discussion on the ontological status of text and its structural representation.

[4] Cf. Pierazzo 2016 on the complexity of editorial modeling and interpretation in digital editions.

[5] https://liberepistolarum.mni.thm.de/id/300dcfe1-9f1a-4e21-914d-4730fd85f1d2.

[6] https://liberepistolarum.mni.thm.de/id/300dcfe1-9f1a-4e21-914d-4730fd85f1d2.

AUTHOR

ANDREAS KUCZERA

Prof. Dr. Andreas Kuczera is Professor for Applied Digital Methods in Digital Humanities at the University of Applied Sciences in Gießen, Germany