Document analysis and Textual philology: A Formal Perspective

Riccardo Del Gratta[1,✉] , Federico Boschetti[1,2], Luigi Bambaci[3] and Francesco Sarnari[4]
[1]Institute for Computational Linguistics "A. Zampolli", Italian National Research Council, Pisa, Italy
Email: riccardo.delgratta@ilc.cnr.it, federico.boschetti@ilc.cnr.it
[2]Venice Centre for Digital and Public Humanities (VePDH)
Email: federico.boschetti@unive.it
[3]Department of Cultural Heritage, University of Bologna
Email: luigi.bambaci2@unibo.it
[4]Biophysics Institute, Italian National Research Council, Pisa, Italy
Email: francesco.sarnari@ibf.cnr.it

***Abstract*—We introduce a formal approach to document and text analysis. The method proposed herein results in a mathematical model/framework which can formalize different challenges in research fields such as computational linguistics, digital philology, and software engineering, principally if applied to document and text analysis. We examine texts and documents from an evolutionary perspective, where both corruption and correction are involved. We describe document evolution via fibre bundles formalism. We also provide other examples to demonstrate the capabilities of the model.**

***Index Terms*—Formal model, document analysis, evolutionary approach, fibre bundles**

I. Introduction

MANY attempts to formalize the methodology of textual philology (or textual criticism) have been made, some of which have proved to be quite useful (see Section II). However, since textual philology is a historical-critical discipline dealing with specific textual traditions, these attempts suffered from being too heavily conditioned by the particular perspectives typical of the different literary domains (classical philology, biblical philology, romance philology, etc.).

Lachmann's method, perhaps the most popular and accomplished as far as formalization is concerned, is an example of this issue: since it is shaped on the traditions of Greek and Latin texts, it is hardly applicable outside the domain of classical philology [5].

It is important, therefore, to foster a unitary vision of textual philology that goes beyond the borders existing between the different sub-domains. This implies the formalization of phenomena which are common to the different textual traditions (e. g. phenomena of unconscious or mechanical corruption) as well as of phenomena which are different in each domain but are the result of general scribal habits and approaches (conscious innovations such as exegetical reworkings, interpolations due to moral or theological concerns, etc.). Such a process of formalization must be based on entities, properties, and relations characterizing the textual-philological domain in the strict sense, but it must also take into account the relationship existing between textual philology and other forms of criticism, such as literary criticism, source criticism, and redaction criticism. This means that it is important to consider not only texts and documents but also the agents or actors responsible for their creation and transmission such as authors, editors, scribes, and translators.

In this article, we introduce a formal approach to text analysis that takes into account some of the aspects mentioned above. The case study we present is artificial and hence simplified in many respects. Nonetheless, it highlights some relevant phenomena occurring in textual transmission and serves as a starting point for the application of the model to real cases (see Section III).

We propose herein a definition of documents and texts which we believe useful for digital and computational philology since it can model variants, conjectures, source reconstruction, and document evolution. The same definition turns out to be also useful in software engineering. The more formal the description of documents and texts, the easier the work for software engineers since they can design general objects to address different issues. Moreover, this helps to keep well-designed software separate from ad-hoc scripts.

The last section before the appendices (Section VII) is dedicated to describe several possible avenues for future research from the introduction of dynamics to the interpretation of the text, independently on *who* interpret it [6], either automatic tools or humans.

II. Background

The first systematic attempts to formalize an approach to texts and documents according to a formal model can be traced back to the end of the nineteenth century within the field of

The authors of this article have developed together the themes here discussed within the Laboratory for Collaborative and Cooperative Philology (CoPhiLab, https://cophilab.ilc.cnr.it/). The authorship of the different sections is specified below: Riccardo Del Gratta is responsible for Sections IV and V, and the Appendices A and B. Federico Boschetti is responsible for Section I. Luigi Bambaci is responsible for Sections II and III. Francesco Sarnari is a cowriter of the Appendices A and B. All the authors are cowriters of Section VI. Additional information about the authors are available at [1]–[4].

textual criticism and textual history. The main goal of these attempts was to reconstruct the history of texts (i.e., their tradition) and in so doing, to uncover the earliest authorial version, the so-called *Original* or *Ur-text*. The dominant idea at that time, influenced both by the emergence of Darwinian theory and by the Romantic school, was that texts develop according to an evolutionary model characterized by a process of progressive deterioration: from earlier, uncontaminated forms down to later and corrupted copies. A close analysis of the differences existing between documents (the variants), and in particular of the errors made by the actors of the textual transmission (the copyists or scribes), allowed scholars to restore a text which was presumed to best represent the author's original intentions, as well as to portray the entire evolutionary process in the form of a tree-like graph (the *stemma codicum*). These assumptions formed what is currently known as Lachmann's method (or genealogical or stemmatological method) [7], [8]. At that time, the main concepts and assumptions forming the grounds of the model were formulated and developed so that textual philology has arisen as science on its own within the realm of historical-critical disciplines. Thanks to the contributions of several scholars, mostly in the fields of New Testament and Classics studies, textual philology underwent a process of systematization culminating, as Timpanaro states, in the "scientific foundation of *recensio*"[1] or text genealogy [9], [10].

The evolutionist approach to texts and documents came to prevail again in the late 1960s, contemporaneously with the birth of the phylogenetic and cladistic schools in evolutionary biology [11]. This led, in the early 1990s, to a revision of the classical Lachmannian genealogical model and to the development of various computational techniques, which now take the name of computer-assisted stemmatology [12], [13]. Since then, various attempts have been made to apply phylogenetic methods to real textual traditions [14]–[16] as well as to artificial ones [17], [18], and computational models have continued to grow in number and specialize over the years [19]–[21]. At the same time, a debate has arisen between supporters and opponents of phylogenetic methods as applied to text traditions, which has had the advantage of highlighting both affinities and differences between the evolution of texts and living organisms [22]–[27].

By contrast, other approaches to the formalization of documents besides graph theory and the evolutionary model exist, among them studies based on set theory [28], [29] and on taxonomic methods such as those developed by Greg [30], Dearing [31] Griffith [32] and others. These last, in particular, tend to reject any attempt to build genealogical hypotheses and principally rely upon techniques borrowed from the fields of algebraic logic and numerical taxonomy.

As to the background of such researches, the last years have seen an increasing interest in epistemological problems. The traditional view stating that texts are the product of human thought and, as such, cannot be reduced to any pre-established formalism [33], [34] has been recently challenged by other proposals defining the theoretical underpinnings which stand behind the methods of textual philology [35]–[39]. These are, however, scanty when compared to other fields such as history or social sciences, which have been better investigated by both scholars of the field [40], [41] and philosophers of science [42], [43].

[1] [9, page 43]: "Of the two parts into which Lachmann divided textual criticism – *recensio* [recension] and *emendatio* [emendation] – the second had been practiced since antiquity. [...] In the nineteenth century, methods of emendation were refined further (this was especially due to progress in the study of the language and style of various epochs and authors), but were not transformed in a revolutionary way [...]. Instead, the great novelty of nineteenth-century textual criticism was the scientific foundation of *recensio*."

III. A "thought" case study

In this section, we present an artificial case study. We imagine a closed collection (C) consisting of different books (B). Each book derives from one or more sources (S) containing the text, such as manuscripts or printed editions.

We imagine that many different copies of the collection were made by scribes in the course of time and that translations (T) were made of these copies by different translators.

We suppose, on one hand, that the copies were made with great care by the scribes and that they consequently reflect a unitary text type. The variations occurring between the different copies are of little import, consisting, for example, of mechanical corruptions that frequently arise during the copying process (unconscious variants).

The translations, on the other hand, are more differentiated, not only because they are written in different languages, but also because they capture different aspects of the source text: while some can be very literal, others are more literary or paraphrastic, and each is the product of the historical context as well as of the cultural background of the individual translator. Besides unconscious variants, therefore, the translations will reveal more important kinds of variation arising from deliberate intervention of the scribe (conscious variants), such as stylistic reworkings, exegetical interpretations and ideologically motivated interpolations.

The example we present includes many typical aspects of (computational) philology: from lost sources to available witnesses, from corruptions caused by both conscious and unconscious variations to corrections made in the attempt of restoring an original text.

The features described in our model are common to different textual traditions. In particular, they fit well with textual traditions characterized by the presence of a *textus receptus* (for example traditions of printed texts) and traditions in which translations play an important role (such as Old Testament philology).

The next two sections will introduce the model (Section IV) and the application of the model (Section V) to the presented case study.

IV. Definitions and Axioms

In this section, we provide definitions for concepts and objects that will be used throughout the paper.

We define the document D as an object consisting of three components: a content, i. e. the informative message, a format, i. e. how the content is arranged in the document, and a set of para-textual layers, i. e. notes, glosses etc.

The proposed definition of documents and texts is useful for Digital and Computational Philology as it models well known aspects such as variants, conjectures, the reconstruction of the relationships between sources, and so on. We introduce also *Cartesian* (or separable) documents, i. e. documents where the three components are easily identified. Cartesian documents simplify the proposed model, since, within Cartesian documents, it is easier to keep separated (for instance) operations performed on the content from those on the format. However, we are aware that it is quite rare that such a simplification constitutes a perfect fit for real documents since, in real documents, content, format, and para-textual layers are difficult to be separated from each other (for instance, italics can be used with the same meaning of quotation marks and footnotes can be authorial). To make it easier for the reader to understand the defined concepts and objects, we will start from *electronic* documents for which to have intuitions about the entities defined is more straightforward.
We also give a list of axioms (see Section IV-B needed to make the model sound.
Additional information concerning definitions and axioms are provided in Appendices A and B.

A. Definitions

D 1. An *electronic* textual document D is an conceptual object which consists, at least, of the following three components:

i a content c
ii a format f;
iii a set of para-textual layers $\{p_0 \dots p_k\}$:

$$D = D(c, f, \{p_0 \dots p_k\}) \tag{1}$$

The content c is the amount of information (its informative message) carried by document D, f is the format used to *formalize* such information, and the set $\{p_0 \dots p_k\}$ represents additional para-textual layers[2] added to document D.

D 2. A Cartesian (or separable) document D_c is a document D where c, f, and $\{p\}$ are independent of each other. In this simplified model, D is somehow presumed to be a separable object:

$$D = c \times f \times \{p\} \tag{2}$$

where $\times$ is the "Cartesian product"[3]. In other words, D_c is described by saying something about c, something about f, and something about set $\{p\}$.

D 3. An extended electronic textual document D_{ex} is an abstract object which contains additional non-mandatory elements[4]:

$$D_{ex} = D_{ex}(c, f, \{p_0 \dots p_k\}, [\{m\}]) = D \times \{m\} \tag{3}$$

[2]So far, we may assume the format of the para-textual layers be the same as document's. Indeed, it's fascinating that the para-textual layer are documents (according to (1)). This aspect adds "recursion" to the theory.

[3]The analogy is with the Cartesian place: content, format, and para-textual layers play the role of the coordinates of a point in a Cartesian plane. As a point is identified by its coordinates, so a Cartesian document is identified by its components.

[4]Where [a,b,...] in (3) means that elements a, b are optional.

Equation (3) means that such additional non-mandatory elements of a given document D — for example, historical period, language, provenance, etc. — must be kept logically separated from the other para-textual layers.

D 4. A base-space $\mathscr{B}$ is a synchronously-accessed collection of documents $D_1, D_2, \dots, D_n$ at a time, $t = \tau$:

$$\mathscr{B}(\{D_1, \dots, D_n\}; \tau) = \{D\}\Big|_\tau =: \mathscr{B}_\tau(\{D\}) \tag{4}$$

The short notation $\mathscr{B}_\tau(\{D\})$ means that to identify different collections of documents at different times τ and t, $\{D\}\Big|_\tau \neq \{D\}\Big|_t$, the difference is highlighted by the suffixes τ or t.

D 5. An operation op is an action belonging to a finite set $\{\mathbb{O}\}$. This set is supposed not to be empty, $\{\mathbb{O}\} \neq \emptyset$.
An operation op takes a subset of documents $\{D_{i_1}, \dots, D_{i_n}\} \subseteq \mathscr{B}_t(\{D\})$ as input and produces a new **single** document $D_j \in \mathscr{B}_\tau(\{D\}), \tau > t$ as output.
Let op be an action in $\mathbb{O}$. The result of op on a set of documents is again a document:

$$op(\underbrace{\{D_{i_1}, \dots, D_{i_n}\}}_{\in \mathscr{B}_t(\{D\})}) = \underbrace{D_j}_{\in \mathscr{B}_\tau(\{D\})} \tag{5}$$

Operations can be *unary*, if they accept a single document in input, *binary* if they accept two, and up to *n-nary* if they accept n.

D 6. A Cartesian (or separable) operation is an operation op which can be decomposed:

$$op = op_c \times op_f \times op_{\{p\}} \tag{6}$$

where op_c acts on the content only, op_f on the format, and $op_{\{p\}}$ on the para-textual layers only.
In the case of Cartesian documents, operations can be considered separable as well.

D 7. An actor ac is either a human or an automatic *tool* belonging to a finite set $\{\mathbb{A}\}$ that performs one or more operations $op \in \{\mathbb{O}\}$.

D 8. An evolution $\mathcal{E}$ is a graph with documents D as nodes and actors performing operations as edges.

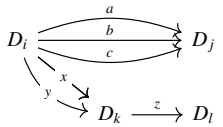

Figure 1: An *Evolution* graph from D_i to D_j and to D_l (through D_k.)

D 9. An evolution space $\mathscr{H}$ is a collection of evolutions, one for each document $D_l \in \mathscr{B}_\tau(\{D\})$, such that for $t < \tau$, $\mathscr{H}(D_l; t) = \{D_{j_1} \dots D_{j_k}\} \subseteq \mathscr{B}_t(\{D\})$ and $\mathscr{H}(D_i; \tau) = \{D_i\} \in \mathscr{B}_\tau(\{D\})$.
To be more concise, we often write $\mathscr{H}(D_i; \tau) = D_i$, meaning by D_i the proper subset $\{D_i\}$ of $\mathscr{B}_\tau(\{D\})$ with the one element[5]:

$$\mathscr{H}(D_i; t) = \begin{cases} \{D_j, \dots, D_{j_k}\} \subseteq \mathscr{B}_t(\{D\}), & \mathrm{t} < \tau \\ D_i \in \mathscr{B}_\tau(\{D\}), & \mathrm{t} = \tau \end{cases} \tag{7}$$

[5]The cardinality is $\big|\{D_i\}\big| = 1$.

We define the elements ev of $\mathscr{H}$ as:

$$ev(\{D_j, \ldots, D_{j_k}\}, D_i, t) \in \mathscr{H}(D_i; t) \tag{8a}$$

s.t.

$$ev(\{D_j, \ldots, D_{j_k}\}, D_i, t) = \{D_m\} \text{ with } m \in \{j, \ldots, k\} \tag{8b}$$

$$ev(\{D_j, \ldots, D_{j_k}\}, D_i, \tau) = \{D_i\} \tag{8c}$$

Each element ev -in (8c)- is a "constrained map", that is, a map "constrained" between a set of documents (initial state) and a single document (final state), while the paths between initial and final states remain unspecified.

D 10. An (operative) total space $\mathscr{E}$ is a *bundle* obtained by placing every document $D_i \in \mathscr{B}_\tau(\{D\})$ into relation with its evolution $ev(\{D_j, \ldots, D_{j_k}\}, D_i, t)$:

$$\mathscr{E} = \mathscr{B} \times \mathscr{H} \tag{9}$$

An element of $\mathscr{E}$ is a pair $b = (D, ev)$ such that:

$$(D, ev) = \begin{cases} (D_i, \{D_j, \ldots, D_{j_k}\}), & \mathrm{t} < \tau \\ (D_i, D_i), & \mathrm{t} = \tau \end{cases} \tag{10}$$

We also need a projection map π (cf. (A.9) in Appendix IV-B) such that:

$$\pi(D, ev) = D \tag{11}$$

B. Axioms

In this section we provide three axioms needed to make sound the part of the model we presented.

Axiom 1. The empty document D_e is a document.

Axiom 2. Let $D_1 = D_1(c_1, f_c, \{p_1\})$ and $D_2 = D_2(c_2, f_c, \{p_2\})$ be two documents, then the union $D = \cup(D_1, D_2)$ is also a document. As a consequence, $\cup \in \{\mathbb{O}\}$.

Axiom 3. Let $D_1 = D_1(c_1, f_c, \{p_1\})$ and $D_2 = D_2(c_2, f_c, \{p_2\})$ be two documents, then the intersection $D = \cap(D_1, D_2)$ is also a document. As a consequence, $\cap \in \{\mathbb{OP}\}$.

V. The model in action

We now move on to employ the concepts defined in Section IV to the case study introduced in Section III. We consider the simplest possible system and suppose that Collection C has been independently compiled by 2 scribes, say a, b. Both a and b had 3 prior sources (S) available to them, to form the foundation upon which the 4 single books (B) constituting C are based. We assume 2 translators of C, t_1 and t_2, who then interpret C (R_{t_i}) and produce 2 final translations: T_{t_1} and T_{t_2}. A possible scenario which models the case study is reported in Fig. 2:

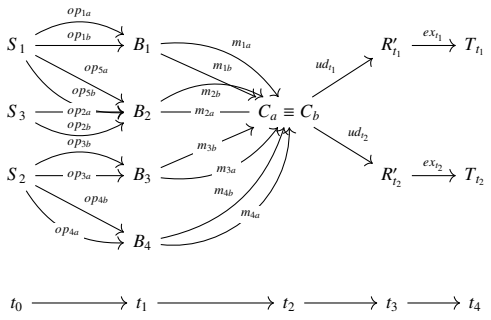

Figure 2: Complete picture: from sources to final translations.

According to base-space definition (4), the set of base-spaces is reported as follows:

$$\mathscr{B} = \begin{cases} \mathscr{B}_{t_0}(\{D\}) := \mathscr{B}(\{S_1, S_2, S_3\}; t_0), & \text{at } t = t_0 \\ \mathscr{B}_{t_1}(\{D\}) := \mathscr{B}(\{B_1, B_2, B_3, B_4\}; t_1), & \text{at } t = t_1 \\ \mathscr{B}_{t_2}(\{D\}) := \mathscr{B}(\{C\}; t_2), & \text{at } t = t_2 \\ \mathscr{B}_{t_3}(\{D\}) := \mathscr{B}(\{R_{t_1}, R_{t_2}\}; t_3), & \text{at } t = t_3 \\ \mathscr{B}_{t_4}(\{D\}) := \mathscr{B}(\{T_{r_1}, T_{r_2}\}; t_4), & \text{at } t = t_4 \end{cases} \tag{12}$$

where each S_i, B_j, C, R_{t_k}, and T_{t_l} are documents (with different content, format, and para-textual layers at different times) in the sense of **D**. (1). Scribes a and b access the documents in $\mathscr{B}_{t_0}(\{D\})$ synchronously. This means that the sources S_{1-3} are available simultaneously, regardless of S_1 being produced years before or after S_2. Similarly, the books in $\mathscr{B}_{t_1}(\{D\})$ are synchronously accessed by both scribes to generate collection C.

Scribes a and b are *actors* ($ac_a, ac_b \in \{\mathbb{A}\}$) who operate ($op$) on documents belonging to different base-spaces $\mathscr{B}_{t_i}$ in order to produce new documents, cf. (5).

If we focus on time lapse $t_0 - t_1$, we may use (5) to formalize the combination of actors and operations[6]:

$$op_{1a}(S_1) = B_{1a}, op_{5a}(S_1) = B_{2a}$$

$$op_{3a}(S_2) = B_{3a}, op_{4a}(S_2) = B_{4a}, op_{2a}(S_3) = B_{2a}$$

Collection C is thus created by merging books B_{1-4}[7]:

$$merging(B_1, \ldots, B_4) = C$$

Fig. 3 describes the combined effort of actors and operations:

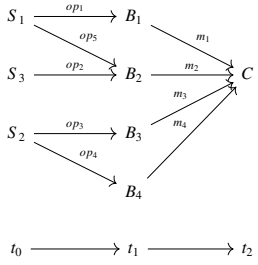

Figure 3: Schematic flows from sources S to Collection C.

[6]Operations are generally written as op_{na}, where n is a counter and index i identifies either scribe a or b.

[7]Please note that *merging* is a valid operation according to definition **D**. (5) and **A**. (2).

We observe that Fig. 3 is very similar to Fig. 1. It represents the *evolution* of the base-spaces: from $B_{t_0}(\{D\})$ to $B_{t_1}(\{D\})$ and from $B_{t_1}(\{D\})$ to $B_{t_2}(\{D\})$. We can use evolution space $\mathscr{H}$ (cf. **D**. (9)) and its elements *ev* to model such evolution. Looking at Fig. 3, there are 5 arrows which couple the elements of $B_{t_0}(\{D\})$ with those of $B_{t_1}(\{D\})$ and 4 arrows which couple $B_{t_1}(\{D\})$ with collection C. The arrows can be labeled with the operations responsible for *evolving* the sources into the targets:

$$\begin{aligned}\Big\{e_1 = (S_1, B_1)_{op_1}, e_2 = (S_3, B_2)_{op_2}, e_3 = (S_2, B_3)_{op_3}\\ e_4 = (S_2, B_4)_{op_4}, e_5 = (S_1, B_2)_{op_5}\\ e_6 = (B_1, C)_{m_1}, e_7 = (B_2, C)_{m_2},\\ e_8 = (B_3, C)_{m_3}, e_9 = (B_4, C)_{m_4}\Big\}\end{aligned} \tag{13}$$

By grouping each individual target with its sources, we can rewrite (13) in a more compact way[8]:

$$\begin{aligned}\mathscr{H}_{01} &= \Big\{H_1 = (\{S_1\}, B_1), H_2 = (\{S_3, S_1\}, B_2),\\ &\quad H_3 = (\{S_2\}, B_3), H_4 = (\{S_2\}, B_4)\Big\}\\ \mathscr{H}_{12} &= \Big\{H_5 = (\{B_1, B_2, B_3, B_4\}, C)\Big\}\end{aligned} \tag{14a}$$

Both B_i and S_k are subsets of the corresponding base-spaces: $\{S_3, S_1\} \subset \mathscr{B}_{t_0}(\{D\}), \{B_1, B_2, B_3, B_4\} \subseteq B_{t_1}(\{D\}), \{C\} \subseteq \mathscr{B}_{t_2}(\{D\})$
$\mathscr{H}_{01}, \mathscr{H}_{12}$ in (14) are realizations of (7), and H_i are the "constrained maps" *ev* of **D**. (9).

We add the constraints to each element of H:

$$H_i\Big|_{t=t_0} = \{S_i \ldots\};\ H_i\Big|_{t=t_1} = \{B_k\}$$

For example, for element $H_2 \in H_{01}$ and $H_5 \in H_{12}$, we have

$$\begin{aligned}H_2\Big|_{t=t_0} &= \Big\{S_1, S_3\Big\}, H_2\Big|_{t=t_1} = \{B_2\}\\ H_5\Big|_{t=t_1} &= \Big\{B_1, B_2, B_3, B_4\Big\}, H_5\Big|_{t=t_2} = \{C\}\end{aligned} \tag{15a}$$

Fig. 4 shows the similar evolution from C to the final translations T_k:

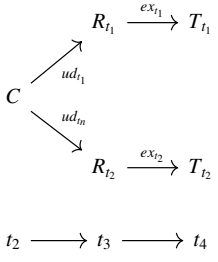

Figure 4: Different interpretations generate different translations.

[8] $\mathscr{H}_{ij}$ is the evolution space from t_i to t_j.

We build the corresponding evolution spaces:

$$\begin{aligned}\mathscr{H}_{23} &= \Big\{H_{t_1} = (\{C\}, R_{t_1}), H_{t_2} = (\{C\}, R_{t_2})\Big\}\\ \mathscr{H}_{34} &= \Big\{H_{t_1} = (\{R_{t_1}\}, T_{t_1}), H_{t_2} = (\{R_{t_2}\}, T_{t_2})\Big\}\end{aligned} \tag{16a}$$

Collection C, produced independently by scribes a and b, consists in two physically distinct objects, but logically they are the same. This is very subtle: when the translators begin their task, it is implicitly assumed that the author of the Collection does not play a relevant role. We identify the interpreted Collection by its content rather than by other information, such as the identity of the author. Extended documents (cf. **D**. (3)) are used when other aspects become relevant. The application of (3) to R_{t_i} and T_{t_i} leads[9] to the following:

$$R_{t_1} = R \times \{author = t_1\}, R_{t_2} = R \times \{author = t_2\} \tag{17a}$$

$$T_{t_1} = T \times \{author = t_1\}, T_{t_2} = T \times \{author = t_2\} \tag{17b}$$

Extended documents in (17), in conjunction with evolution space, play an important role in tracking the process from C to T_{t_i}.
There are two possible reasons for ending up with translations T that differ from each other, even when translating into the same language: Either the respective translators misinterpreted the underlying meaning of Collection C, or each was accustomed to using different terms and concepts from those habitually used by the other.

Fig. 5 shows two possible processes: either (i) the reconstruction of R_{t_1} from T_{t_1}[10] to understand how from R_{t_1} we obtained T_{t_1}, or to evaluate the skills of t_1[11]; or (ii) the other way around: C is compared to R_{t_1} (the σ operation) and the complete process is replicated in order to understand whether T_{t_1} contains errors.

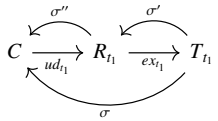

Figure 5: Evolution spaces in action.

According to definition (10), the total space of T_{t_1} consists of T_{t_1} and all possible evolutions that led to T_{t_1}, specifically $ex_{t_1}(R_{t_1}) = T_{t_1}$. Domain R_{t_1} has as its total space R_{t_1} along with $ud_{t_1}(C) = R_{t_1}$. In these cases, σ', σ'' play a role similar to the inverse of *ex* and *ud* respectively, in the sense that $ex_{t_1}(\sigma'(T_{t_1})) = T_{t_1}$.

We can use the same strategy between books (B) of C and their initial sources (S). In Fig. 6, we explicitly show the effect produced by operations σ, op_2 and op_5.

[9] Please note that R_{t_i} and T_{t_i} are both operations according to definition (5), since they accept documents and return a document.

[10] In this case σ' is the translator's identification which is possible since T_{t_1} is an extended document.

[11] σ'' is a comparison of C vs. R_{t_1}.

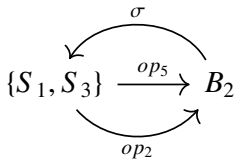

Figure 6: Going back from B_2 to the subset of sources S=$\{S_1, S_3\}$.

The σ operation goes back from B_2 to $S = \{S_1, S_3\}$. Since we know that op_5 applies to S_1 and op_2 to S_3 only, the *composition* of op_3 and op_5 after σ makes sense. When we compose σ with either op_5 or op_2, we obtain B_2 again[12]. However, there can be cases where we do not know op_2 and op_5. In these cases, we can suppose the existence of a function π from $S = \{S_1, S_2, S_3\}$ to B'_2 such that the obtained B'_2 is equal to B_2, see Fig. 7.

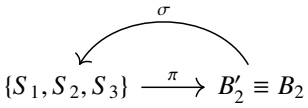

Figure 7: Project $\{S_1, S_2, S_3\}$ to re-obtain B_2.

A. The Case Study and its Bundle

In this section we recap some basic concepts from bundles and use them to justify their use in the model[13].
In Fig. 8 we sketch a possible evolution graph that from the base-space $\mathscr{B}_0$ bring to $\mathscr{B}_1$ and in fig. 9 its translation into bundles. The point $D_a \in \mathscr{B}_1$ is connected to its fibre $\mathscr{H}$; the three points $ev_{1-3}(D_1, D_2)$ in $\mathscr{H}$ are three different evolutions that from the set $\{D_1, D_2\}$ (in the base-space $\mathscr{B}_0$) lead to D_a.

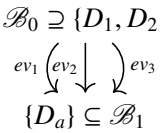

Figure 8: D_1 and D_2 in $\mathscr{B}_0$ evolve into D_a in $\mathscr{B}_1$.

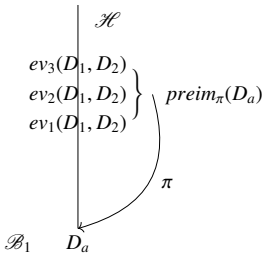

Figure 9: The point D_a and its fibre.

[12]For clarity: from B_2 we go back to the set of its sources, then we apply the right op_i operation to obtain B_2 again. $\sigma : B_2 \to \{S_1, S_3\}$ and $op_2(\{S_1, S_3\}) \equiv op_2(S_3) = B_2$. $op_2 \circ \sigma(B_2) = op_2(\{S_1, S_3\}) = op_2(S_3) = B_2$. In other word, we rebuild the sources.

[13]cf. appendix A for additional details on bundles, section bundles and product bundles.

In bundle terms, $ev_{1-3}(D_1, D_2)$ is the preimage of the point D_a. Fig. 9 shows a product fibre bundle whose points are identified by the pair (D_a, ev_{i-3}). According to (A.9), there is a projection map π such that $\pi(D_a, ev_{i-3}) \to D_a$.
The meaning is the following: from a philological point of view, especially for the reconstruction of sources, we do not focus on the path or paths that lead from the sources to the analyzed document (D_a), but only on the fact that a path can be reconstructed. Much of the philological research lies exactly in the reconstruction of this path. In terms of bundles, it doesn't matter which point in the preimage leads to D_a, it only matters that a map from one of these points to D_a exists.

VI. Critical Editions as Cartesian Documents and other examples

In section V, we described the bundle formalism to focus on reconstructing the sources, but we left out other features of the model itself.
In this section, we show other applications of the proposed model which may result interesting under many points of views.

A. Critical Editions

Critical editions are examples of Cartesian documents, as the content, format, and para-textual levels are identifiable and separable. The text established by the editor is identifiable with the content c, the critical apparatus of variants together with the apparatus of sources, historical notes, etc. with the para-textual levels $\{p_i\}$ and the publishing conventions for the "mise en page" with the format f.
Fig. 10 shows a sample of a critical edition [44][14] with f, c, and $\{p_i\}$ highlighted. The (brown) arrow connects the content to the para-textual layers.

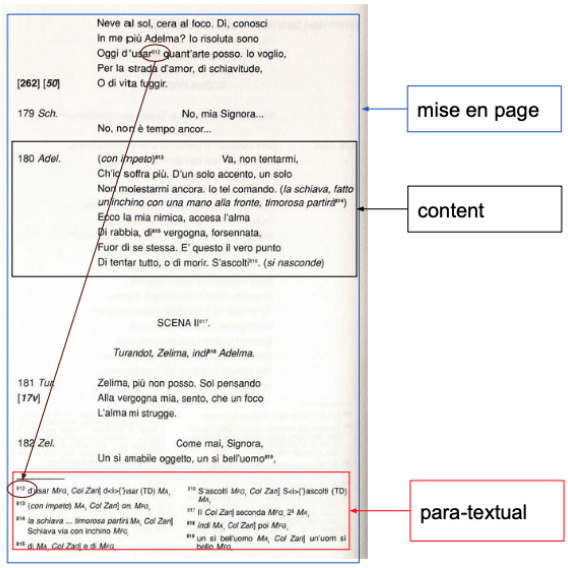

Figure 10: Example of a critical edition.

[14]Courtesy of http://www.usc.es/gozzi/arch/turandot.html

The critical edition of Fig. 10 is then formalized as (cf. (3))

$$D_{ce} = c \times f \times \{p_i\} \tag{18}$$

The (brown) arrow in figure, connects the accepted reading *d'usar* with the refused variant *di usar* in the critical apparatus. However, it may be the case that the role of the variant changes, if it is accepted in the critical text by another editor. According to the proposed model, this change is formalized as an operation from D_{ce} to D'_{ce}:

$$\tilde{op} : D_{ce} \to D'_{ce} \tag{19a}$$

$$\text{under } \tilde{op} \begin{cases} op_c : c \to c' \\ op_f : f \to f' \\ op_{\{p\}} : \{p\} \to \{p'\} \end{cases} \tag{19b}$$

We may suppose that the "mise in page" of D'_{ce} is the same as D_{ce}, so that $f = f'$ and $op_f = Id_f$.
Moreover, c' differs from c because of the different accepted reading and $\{p_i\}'$ differs from $\{p_i\}$ because of the refused variant. To change the role of the variants, means extracting an information from the para-textual layers and inserting it into the content. As described in appendix A, this is a typical function of a special class of operations called *extractors*.

Let c_1 be the content c without the original accepted reading *d'usar* and let $ex_{\{p\}} : \{p_i\} \to c_2$ the operation of extracting the variant *di usar* from the apparatus, then, according to **A**. (2):

$$c' = \cup(c_1, ex_{\{p\}}) = \cup(c_1, c_2) \tag{20}$$

and op_c results:

$$op_c = \cup(\bullet) \tag{21}$$

Similar considerations can be made about the apparatus[15]:

$$\{p'\} = \cup(p_1, ex_c) = \cup(p_1, p_2) \tag{22}$$

and $op_{\{p\}}$ results:

$$op_{\{p\}} = \cup(\diamond) \tag{23}$$

Finally:

$$\tilde{op} = \cup(\bullet) \times Id_f \times \cup(\diamond) \tag{24}$$

which is (6) in **D**. (6).

B. Isomorphic Transformations

An isomorphism is a bijection between two sets, A, B. The bijective map is also structure-preserving.
Suppose the set $A = \{a_1, a_2\} = \{$'Cp x,y', '||'$\}$ and $B = \{b_1, b_2\} = \{$'x:y', ' • '$\}$, we map $a_1 \to b_1$ and $a_2 \to b_2$. We call such bijection a *format change*, f_{ch}:

$$f_{ch} = \begin{cases} \text{'Cp x,y'} \leftrightarrow \text{'x:y'} \\ \text{'||'} \leftrightarrow \text{' • '} \end{cases} \tag{25}$$

Fig. 11 shows an example[16] of apparatus from two editions of the biblical book of Qohelet: the *Biblia Hebraica Stuttgartensia* (BHS) [45] and the *Biblia Hebraica Quinta* (BHQ) [46]

10 [a] l c nonn Mss יְהִי; cf 2,7 [b] ‖ **13** [a] Or אֲשֶׁר; > 𝔠 ‖ [b] 𝔠 mlt Mss 𝔖𝔗𝔙 הַשָּׁמֶשׁ; cf 1,3.9 etc ‖ **15** [a] l לְהִתְקֵן cf 𝔊 ‖ [b] prp לְהַמְּלוֹת (cf bBer 16b) ‖ **16** [a] > 𝔠 ‖ [b] prp גָּדַלְתִּי cf 𝔊 *ἐμεγαλύνθην* = וגדלתי et 𝔊 2,9 ‖ [c–c] mlt Mss Vrs בִּירוּשׁ cf 2,7 ‖ **17** [a] 𝔊𝔠 cj c aα ‖ [b] nonn Mss הוֹלֵלוּת cf 10,13 ‖ [c] mlt Mss וְסִ ‖ **18** [a] 𝔊 *γνώσεως* = דַּעַת ‖ **Cp 2,1** [a] sic L, mlt Mss Edd כָּה—; prp אֲנַסֶּנָּה ‖ **3** [a] l לִשְׁמוֹךְ cf Cant 2,5 ‖ [b] l frt וְלֹא אֹחַז ‖ [c] 2 Mss 𝔊𝔖 הַשָּׁמֶשׁ.

(a) BHS, Qohelet 1:10-2:1

1:1 :מֶלֶךְ בִּירוּשָׁלִָם T ‖ βασιλέως Ἰσραὴλ ἐν Ἰερουσαλήμ G (facil) ‖ *regis Hierusalem* V Hie[Iem] S (facil) ✣ • **2** קֹהֶלֶת (assim-1:1) ‖ ὁ ἐκκλησιαστής G ‖ V S T (indet) ‖ pref הַקֹּהֶלֶת see G ✣ • **3** עָמָל֔וֹ G V T (assim-usu) ‖ > sfx α' S ‖ pref הֶעָמָל see α' S ✣ • הַשָּׁמֶשׁ G V S[Mss] T (assim-usu?) ‖ [illegible] S* ✣ • **5** חָזָרָה ⟨α'⟩ ‖ καὶ ἀνατέλλει G ⟨σ'⟩ ‖ > cj V Hie[Iem] S ‖ וְיָרַח T ‖ pref חֹרֵחַ see G ⟨σ'⟩ ✣ •

(b) BHQ, Qohelet 1:1-5

Figure 11: Examples of apparatus from critical editions

In BHS (Fig. 11a), places of variation (number of chapter and verse) are identified by the annotation 'Cp', followed by the chapter and the verse separated by a comma (in the example: 'Cp 2,1'). In the BHQ (Fig. 11b), the same information is expressed as two numbers separated by a colon (eġ. '1:1'). Similarly, in the former, the apparatus entries are separated by two vertical strokes ('||'), while in the latter each apparatus entry ends with a circle ('•')[17].

The operation op : BHS → BHQ is an isomorphism if we use (25), with (6) which becomes:

$$op = Id_c \times f_{ch} \times Id_{\{p\}} \tag{26}$$

Isomorphisms must preserve the structure as well. The canonical example is the isomorphism between two ordered sets, X and Y.
If $x_1, x_2 \in X$ with $x_1 \leq_X x_2$, m is an isomorphism from X to Y is a bisection $m : X \xrightarrow{Y}$ such that:

$$x_1 \leq_X x_2 \iff y_1 \leq_Y y_2, m(x_{1|2}) = y_{1|2}$$

In the specific case, the structure of the elements of the set A is the "meaning" of the symbol. For example, 'Cp x,y' → *place* associates to the symbol 'Cp x,y' the meaning of *place*, see Fig. 12.

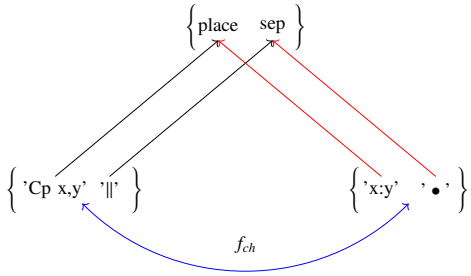

Figure 12: The meaning of the symbols is preserved under f_{ch}.

Since f_{ch} of (25) acts (by construction) on the elements only the meaning of the symbol is preserved:

$$f_{ch}(\text{'Cp x,y'} \to place) = f_{ch}(\text{'Cp x,y'}) \to place = \text{'x:y'} \to place$$

[15] p_1 is the apparatus without the refused variant and $ex_c : c \to p_2$ plays the role of extracting the accepted reading from the original content.

[16] Please note that Figs. 11b and 11a do not represent the same excerpt of text in different editions, they have been chosen to demonstrate the use of various typographical marks to identify identical concepts such as the reference to places of variation and separation in different editions.

[17] Which are the entries of A and B defined at the beginning of the section.

VII. Conclusion and Future Research Lines

We have introduced a formal model to approach various aspects of document analysis. We first considered an electronic document, in the belief that readers of an electronic document might more readily intuit things about its content, format, and para-textual layers than they would in the case of a manuscript, for example.

Our model defines documents, operations, and actors, in order to formalize the evolution of documents, their corruption, and their reconstruction.

Historically, Mathematics has been employed in the field of Linguistics, especially to model sentence structure [47] and distributional semantics [48], but our article has had as its intent an exploration of the use of (section) fibre bundles in examining the evolution of documents in (Computational) Philology. (Section) fibre bundles are appropriate in connecting a document D with its sources. Pairing each document in a base-space with those that led to D in a prior base-space allows our model to manage D and its history simultaneously.

Total spaces represent the "Cartesian product" of a document with its history, in the sense that, from the documents belonging to a base-space $B_t(\{D\}), t < t_0$, we are able to recover (project) document D at $t = t_0$. Total spaces can be nested in such a way that we are able not only to backtrace a document all the way to its original sources but also to reconstruct D from its sources at various intermediate times.

We have hinted throughout at several possible avenues for future research. The first is connected to separability. In Section VI-A, we showed that a critical edition can be modeled with Cartesian documents and used special classes of operations, the *extractors*, to move information among the different constituents of a document. Since the documents are separable, even the operations are separable, but what is the extent to which a function $f(x, y, z) \equiv f_1(x) \times f_2(y) \times f_3(z)$? What makes it possible, as regards documents, to *separate* a document D into its constituent parts? What would such a separation imply? In document and text analysis, separating content from format and para-textual layers is neither easy nor clear.

A second promising research line might be extended documents $D_{ex} = D \times \{m\}$. Here, we would want to investigate the boundaries between para-textual layers and other additional information. It is clear that D_{ex} are useful for source reconstructions, since they allow the identification of the author of a document, but in other situations, it is not so direct. For example, is the historical period of a document an $\{m\}$ or a $\{p\}$? The answer is: "It depends on the operation we have to perform". As future work we would expect to formalize the "it depends" as much as possible.

Yet another interesting possibility might be the application of our model to non-textual documents. To apply our definition of D to images leads to a formalization of *lossy* and *lossyless* transformations between images.

Finally, we might add the element of dynamics. We have defined evolution, corruption, and reconstruction of documents in our models, implying that documents D_i are dynamical variables. And at first sight, this is true. But it is also the result of the dynamics of c, f, and $\{p\}$ which creates the dynamics of D. We would therefore need to define the dynamical variables very clearly.

A further aspect of dynamics is related to the human agent as it is in our case study, where translation T depends on interpretation R. In such cases, the historical context and cultural background of the agent are of primary importance. The same agent may repeat the identical operation at different moments, producing different results, cf. Fig. 13. Or two distinct agents may begin from the same initial document and create different outputs, cf. Fig. 14.

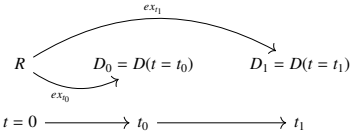

Figure 13: How does D_0 differ from D_1?

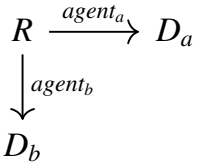

Figure 14: How does D_a differ from D_b?

Our research hypothesis is that the difference between the outputs is a "measure" of the relevance of the historical context and the cultural background of the agents in the process.

Appendix A
Explanation for Definition

In **D**. (1), document D is assumed to be composed of three identifiable parts: c is the content i.e. the amount of information carried by document D, f is the format used to *formalize* such information, and the set $\{p_0 \dots p_k\}$ represents additional para-textual layers added to document D. However, at least in principle, every component might depend on the others. Equation (1) should be read as the following:

$$D(c, f, \{p_0 \dots p_k\}) = D(c(f, \{p_i\}), f(c, \{p_i\}), \{p_i(c, f)\}) \quad \text{(A.1)}$$

$c(f, \{p_i\})$ in (A.1) means that the informative message, the content c, depends on both the format f and the para-textual layers $\{p_i\}$. For example, many web pages convey information (c) thanks to specific fonts, such as bold and italic (f), as well as thanks to additional messages included in different areas of the pages, such as side notes or footnotes ($\{p_i\}$). Similar considerations can be done for f and $\{p_i\}$ in (A.1).

D. (5), through (5), affirms that an operation op is an action that takes documents and produces a document. According to **D**. (5), we can define a specific class of operations (see below) called extractors. Extractors are used to manage interdependencies among the components.

As stated above, if $c = c(f, \{p_i\})$, we can assume that there is a pair of extractors, $e_f : c(f, \{p_i\} \to f$ and $e_p : c(f, \{p_i\}) \to \{p_i\}$,

able to extract from the *content* the format and the para-textual layers respectively. If such extractors exist, then a document D can be transformed into a Cartesian document D_c as in **D**. (2). Cartesian documents are a subclass of documents: the class of documents which either a set of extractors is available,

$$ex : D \rightarrow D_c \tag{A.2}$$

or they are born separable. Let $D = D(c, f, \{p\})$ a generic document and op an operation acting on D to produce $D'(c', f', \{p'\})$, $op : D \rightarrow D'$ according to (5). We cannot generally assume that:

$$\text{under } op \begin{cases} op_c : c \rightarrow c' \\ op_f : f \rightarrow f' \\ op_{\{p\}} : \{p\} \rightarrow \{p'\} \end{cases} \tag{A.3}$$

Or, in other words: it is not possible to make *a-priori* assumptions to guarantee that op factorizes on component-based sub-operations, op_c, op_f, and $op_{\{p\}}$.
Indeed, it is not generally true that:

$$op^t = op^c \times op^f \times op^{\{p\}}$$

However, (A.4) holds for Cartesian documents. As we can see in Figure 15, D and D' are Cartesian documents, the operation op factorizes according to (A.4):

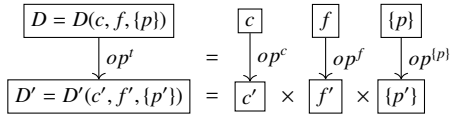

Figure 15: From separable D to D' using a separable operation op.

One may say that op^t acts on each component "at the same time".

Similarly to (A.2), there might be extractors[18] g_i that transform documents into extended documents as in **D**. (3) (cf. (3)):

$$g : D_{ex}(c, f, \{p_i\}, [\{m\}]) \rightarrow D_{ex} = D(c, f, \{p_i\}) \times \{m\} \tag{A.4}$$

Extended documents D_{ex} are Cartesian documents in the sense that additional metadata $\{m\}$ are kept logically separated from D[19], so that D_{ex} is identified by the pair $(D, \{m\})$.

Since we have cited many times the term operation, it is necessary to justify a couple of notions we inserted in **D**. (5), namely that $\mathbb{O}$ is a finite not empty set.
In **D**. (5), we say that operations are actions that take a set of documents to produce a single document in output. We now know that the input documents belong to a subset of the input base-space $\mathscr{B}_0$, while the output document is a subset of the output base-space, $\mathscr{B}_1$. Operations act on a finite set of documents, more precisely, they act on elements of the powerset[20] of $\{\mathscr{B}_0\}, P(\{\mathscr{B}_0\})$, which is a finite set[21]. When we come at defining the set of operations $\{\mathbb{O}\}$ on $\{\mathscr{B}_0\}$, we can implicitly define its elements as:

$$\mathbb{O} = \{op | op \text{ takes s } \subseteq P(\mathscr{B}_0) \text{ in input}\}. \tag{A.5}$$

The main question for (A.5) is :"how many elements does the set $\mathbb{O}$? contain?" Potentially, they are infinite even if the input set $P(\mathscr{B}_0)$ contains a finite number of elements. This because since there is no *a-priori* preclusion about possible operations on documents. However, is reasonable, when we come at listing the actual operations performed in NLP and philology, to limit the number of the elements of $\mathbb{O}$ to a finite number N[22].
In **D**. (5) we specify that $\{\mathbb{O}\}$ is also non empty. The reason why $\{\mathbb{O}\} \neq \emptyset$ is because for every document D it is always possible *do nothing*. We call these *do-nothing* operations *Identities*.

$$Id_D : D \rightarrow D \tag{A.6a}$$

$$Id_D \equiv Id_c \times Id_f \times Id_{\{p\}} \tag{A.6b}$$

The meaning of (A.6b), when compared to (A.6a) is that for Cartesian documents, *do-nothing* on D, Id_D, is the same as *do-nothing* on $c, f, \{p\}$ *at the same time*.A final remark on extractors. Extractors are operations because they send documents into documents. Indeed, if at $t = t_0$ we have $D_0 \in \mathscr{B}_0$ and we apply for example ex_f, then we obtain at $t = t_1 > t_0$ a new document D_1 which, by definition belongs to $\mathscr{B}_1$. In conclusion, $ex_f \in \mathbb{O}$.

In **D**. (7), as for operations, we assume the set of actors $\{\mathbb{A}\}$ to be finite and not empty. While the reason why it is finite is straightforward (the number of human is finite, as it is the number of NLP tools), the motivations behind being non empty are more subtle.
We may say that it is the combination between the actors and the operations that bring from a document to another, as in (A.7) where $ac \in \{\mathbb{A}\}$ is an actor:

$$ac(D_i)\big|_{op \in \{\mathbb{O}\}} = D_j \tag{A.7}$$

The meaning of (A.7) is that the actor ac performs a finite number of operations op on D_i to produce D_j. To make documents evolve, as in **D**. (8) and **D**. (9), we must assume that at least an operation exists, regardless of D_j being equal to or different from D_i[23].

Base-spaces, evolution spaces, and total space, which are defined in **D**. (4) and from **D**. (8) to **D**. (10), are terms derived from the bundle theory which has loads of applications, especially in topology [49]. Without pretending to be rigorous, a bundle (of topological manifolds) is a triple (E, π, M) where E is called the *total space*, M the base space, and π is a surjective map called the *projection map*, $\pi : E \rightarrow M$. If p

[18] For some $\{m_i\}$ there are $g_i : m \rightarrow D$.

[19] D can or can not be Cartesian.

[20] The powerset of a set S is the set of all possible subsets of S, including S and $\emptyset$.

[21] If $|\{\mathscr{B}_0\}| = n$, the the number of elements of its powerset is 2^n, $|P(\mathscr{B}_0)| = 2^n$.

[22] At least is reasonable to limit the number of *type* of operations: parsing, merging, interpretations etc.

[23] We have just seen that identities are operations that leave the document unchanged, that is to say $D_j = D_i$.

is a point in M, then in E there is a set of points called the $preim(\{p\})$[24] so that:

$$preim_\pi(\{p\}) = \pi(preim(\{p\})) = F_p \tag{A.8}$$

F_p is called the "fibre at point p", kind of all the points $x \in E$ mapped to $p \in M$ by the map π.
Is very intuitive to image the total space E as a "product", in the sense that to each point of the base-space M is attached its fibre F. Indeed, the product bundle is the most trivial example of bundle. The total space is the Cartesian product $M \times F$, so that $\pi : M \times F \to M$ is defined as a proper projection (on the first factor):

$$\pi(p, f) \to p \tag{A.9}$$

where $p \in M$ and $f \in F$.
Another important concept is the "section" (fibre) bundle[25]. A "section" (fibre) bundle is a tuple (E, π, M, σ) where σ is a map from the base space to the total space, $\sigma : M \to E$. σ maps a point $p \in M$ to (some) point(s) in E, but for the bundle to be a "section" (fibre) bundle it must be that when we apply π after σ, we go back again to the same p:

$$\sigma(p) = \{f\} \tag{A.10a}$$

$$\pi(\{f\}) = p \tag{A.10b}$$

$$\pi \circ \sigma = Id_p \tag{A.10c}$$

Consider a sheet of paper, and draw a horizontal line (h) just in the middle. Take a couple of points (p_1, p_2) on such line and draw two vertical lines, v_1, v_2. The triple total space E (generated by connecting every point of h with the corresponding vertical lines), the map π, and h as the base space, is a "section" fibre bundle if the map σ maps $p_{1(2)}$ to $v_{1(2)}$ so that projecting with $\pi, \pi(v_{1(2)})$, we obtain $p_{1(2)}$ again. In other words, σ lowers and raises $p_{1(2)}$ on its vertical line. $v_{1(2)}$ is the section of $p_{1(2)}$. Section fibre bundles do not allow $\sigma(p_{1(2)}) \to v_{2(1)}$. Figure 16 reports on the left the correct way and of mapping between σ and π, while on the right the wrong one.

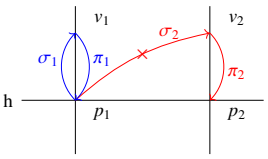

Figure 16: $\pi_1 \circ \sigma_1$ is Id_{p_1} while $\pi_2 \circ \sigma_2$ moves p_1 to p_2 which is not allowed in section fibre bundles.

We have defined the base-space $\mathscr{B}_i$ as the set of documents at time $t = t_i$ and the evolution space $\mathscr{H}_i$ as the collection of all evolutions that starting from a previous base-space bring to D_i.
Both $\mathscr{B}$ and $\mathscr{H}$ are (finite) sets, so they can be equipped with the discrete topology. This makes it possible to use the concepts of bundles. It is intuitive to define $\mathscr{H}_i$ as the fibre of D_i, since under the projection map π (the element ev in **D**. (9)), every element of $\mathscr{H}_i$ goes back to D_i attached to each document D in $\mathscr{B}_0$. The triple $(\mathscr{E} = \mathscr{B} \times \mathscr{H}, ev, \mathscr{B})$ is a fibre bundle, more precisely a product bundle. We may impose the bundle to be a section fiber bundle to guarantee that, starting from D_i we can go back to the documents whose evolutions brought to D_i, which is a strong philological constraint.

[24]For the curious readers, we call it *preim*, because the map π is not invertible. Given two sets A, B, and a surjective map m from B to A, the $preim(x)$, $x \in A$ is the subset of elements $y_i \in B$ sent to x by m, $m(y_i) = x$.

[25]A product bundle is a special case of fiber bundle, see for example [50].

Appendix B
Explanation for Axioms

In **A**. (1), we state that the empty document is a document. This is implicit also in **D**. (1) if we define the empty document as follows:

$$D_e = D_e(c_e, f, \emptyset) \tag{B.1}$$

where c_e means that D_e has no content, no informative message. Format f is the only component left unspecified. Empty files are valid examples of empty documents[26]. The empty document is needed since it may be the output of an operation op, for example "extract all number" from a document that contains no numbers produces D_e. This specific kind of operations is assimilated to *intersection*: according to **A**. (3), if D_1, D_2 are disjoint documents $\cap(D_1, D_2) = D_e$.
The *merging* operation used in Section V is a union (cf. **A**. (2)) of two (or even more) documents: $\cup(D_i, D_e) = D_i$.

Acknowledgment

The authors would like to thank Angelo Mario Del Grosso for directing us to extremely helpful articles on document models and abstractions [51] used by the most important programming languages, [52], [53]. And for guaranteeing that the proposed definition of documents is, some way, compatible [54], [55].

References

[1] R. Del Gratta, "Institute for Computational Linguistics "A. Zampolli" Personal Page: http://www.ilc.cnr.it/en/content/riccardo-del-gratta," 2021.
[2] F. Boschetti, "Institute for Computational Linguistics "A. Zampolli" Personal Page: http://www.ilc.cnr.it/en/content/federico-boschetti," 2021.
[3] L. Bambaci, "University of Bologna Personal Page: https://www.unibo.it/sitoweb/luigi.bambaci2," 2021.
[4] F. Sarnari, "Institute of Biophysics Secondary Seat of Pisa: http://www.pi.ibf.cnr.it/?page_id=3353&lang=en," 2021.
[5] M. Weitzman, "The Analysis of Open Traditions," *Studies in Bibliography*, vol. 38, pp. 82–120, 1985.
[6] A. Di Iorio, G. Spinaci, and F. Vitali, "Multi-layered edits for meaningful interpretation of textual differences," in *Proceedings of the ACM Symposium on Document Engineering 2019*, DocEng '19, (New York, NY, USA), Association for Computing Machinery, 2019.
[7] P. Maas, *Textkritik*. Leipzig: B. G. Teubner Verlagsgesellschaft, 2 ed., 1950.
[8] P. Trovato, *Everything You always wanted to know about Lachmann's Method — A Non-Standard Handbook of genealogical textual Criticism in the Age of Post-Structuralism, Cladistics and Copy-Text*. Storie e linguaggi, Padova: Libreriauniversitaria, 2014.
[9] S. Timpanaro, *The Genesis of Lachmann's Method*. Chicago / London: University of Chicago Press, 2005.

[26]We prefer not to go into details on how much information is contained into f. Just think that, to some extent, empty document carry no information.

[10] E. J. Kenney, *The Classical Text – Aspects of Editing in the Age of the printed Book*. Berkeley/Los Angeles/London: University of California Press, 1974.
[11] W. Robins, "Editing and Evolution," *Literatur Compass*, vol. 3, pp. 89–120, 2006.
[12] P. van Reenen and M. van Mulken, eds., *Studies in Stemmatology*, vol. 1. Amsterdam/Philadelphia: John Benjamins Publishing, 1996.
[13] P. van Reenen, A. den Hollander, and M. van Mulken, eds., *Studies in Stemmatology*, vol. 2. Amsterdam/Philadelphia: John Benjamins Publishing, 2004.
[14] P. Robinson, "A Stemmatic Analysis of the Fifteenth-Century Witnesses to The Wife of Bath's Prologue," in *Canterbury Tales Project Occasional Papers* (P. Robinson and N. F. Blake, eds.), vol. II, (London), pp. 69–132, 1997.
[15] L. R. Mooney, A. Barbrook, C. Howe, and M. Spencer, "Stemmatic Analysis of Lydgate's Kings of England: A Test Case for the Application of Software Developed for Evolutionary Biology to Manuscript Stemmatics," *Revue d'Histoire des Textes*, vol. 31, pp. 275–297, 2001.
[16] A.-C. Lantin, P. Baret, and C. Macé, "Phylogenetic Analysis of Gregory of Nazianzus' Homily 27," in *Le poids des mots. Acted des 7èmes Journées Internationales d'Analyse statistique des Donées Textuelles*, vol. 2, (Louvain-la-Neuve), pp. 700–7, 2004.
[17] P. V. Baret, P. Robinson, and C. Macé, "Testing Methods on an Artificially Created Textual Tradition," in *Proceedings of the international workshop held in Louvain-la-Neuve on September 1 - 2, 2004*, (Pisa), pp. 255–83, Ist. Ed. e Poligrafici Internazionli, 2004.
[18] T. T. Roos and T. Heikkilä, "Evaluating Methods for Computer-Assisted Stemmatology Using Artificial Benchmark Data Sets," *Literary and Linguistic Computing*, vol. 24, no. 4, pp. 417–433, 2009.
[19] P. J. Gurry, "How Your Greek New Testament is Changing: A Simple Introduction To The Coherence-Based Genealogical Method (CBGM)," *Journal of the Evangelical Theological Society*, vol. 59, no. 4, 2016.
[20] A. Hoenen, "An Open Problem in Computational Stemmatology: A Model for Contamination," *Umanistica Digitale*, vol. 3, May 2019. Number: 5.
[21] P. Roelli, *Handbook of Stemmatology*. Berlin, Boston: De Gruyter, 2020.
[22] R. Hanna, "The Application of Thought to Textual Criticism in All Modes — with Apologies to A. E. Housman," *Studies in Bibliography*, vol. 53, pp. 163–172, 2000.
[23] N. Caetlidge, "The Canterbury Tales and Cladistics," *Neuphilologische Mitteilungen*, vol. 102, no. 2, pp. 135–150, 2001.
[24] P. V. Baret and C. Macé, "Why Phylogenetic Methods Work : The Theory of Evolution and Textual Criticism," *Linguistica computazionale*, vol. 24, pp. 89–108, 2006.
[25] C. Howe, R. Connoly, and H. Windram, "Responding to Criticism of phylogenetic Methods in Stemmatology," *Studies in English Literature*, vol. 52, pp. 51–67, 2012.
[26] M. Buzzoni, E. Burgio, M. Modena, and S. Simion, "Open versus Closed Recensions (Pasquali): Pros and Cons of Some Methods for Computer-Assisted Stemmatology," *Digital Scholarship in the Humanities*, vol. 31, pp. 652–669, 2016.
[27] B. Alexanderson, "Why Phylogenetic Methods Do Not Work Very Well in Textual Transmission," *Revue d'Histoire des Textes*, vol. 13, pp. 383–410, Jan. 2018. Publisher: Brepols Publishers.
[28] J. Froger, *La Critique des Textes et son Automatisation*. Paris: Dunod, 1968.
[29] D. S. Avalle, *Principî di Critica Testuale*. No. 7 in Vulgares eloquentes, Roma / Padova: Antenore, 2 ed., 1978.
[30] W. W. Greg, *The Calculus of Variants — An Essay on Textual Criticism*. Oxford: Clarendon, 1927.
[31] V. A. Dearing, *Principles and Practice of Textual Analysis*. Berkeley: University of California Press, 1974.
[32] J. G. Griffith, "A Taxonomic Study of the Manuscript Tradition of Juvenal," *Museum Helveticum*, vol. 25, no. 2, pp. 101–138, 1968.
[33] A. E. Housman, "The Application of Thought to Textual Criticism," in *Selected Prose* (J. Carter, ed.), pp. 131–50, Cambridge: Cambridge University Press, 1961.
[34] E. Tov, "The Relevance of Textual Theories for the Praxis of Textual Criticism," in *A Teacher for All Generations: Essay in Honor of James C. Vanderkam* (E. F. Mason, ed.), no. 153 in Journal for the Study of Judaism, Supplements, pp. 23–25, Leiden: Brill, 1982.
[35] J. Mackie, "Scientific Method in Textual Criticism," *Australasian Journal of Philosophy*, vol. 25, no. 1-2, pp. 53–80, 1947.
[36] D. F. McKenzie, "Printers of the Mind: Some Notes on Bibliographical Theories and Printing-House Practices," *Studies in Bibliography*, vol. 22, pp. 1–75, 1969.
[37] G. T. Tanselle, "Bibliography and Science," *Studies in Bibliography*, vol. 27, pp. 55–89, 1974.
[38] V. Fano and G. Bomprezzi, "Considerazioni Epistemologiche Sulle Scienze Storico-Filologiche," *Studi Urbinati B*, vol. 81, pp. 21–44, 2011.
[39] R. Hendel, "The Epistemology of Textual Criticism," in *Reading the Bible in Ancient Traditions and Modern Editions* (A. B. Perrin, K. S. Baek, and D. K. Falk, eds.), pp. 245–67, Atlanta: Society of Biblical Literature, 2017.
[40] M. Bloch, *Apologie pour l'Histoire ou Métier d'Historien*. No. 3 in Cahiers des Annales, Paris: Librairie Armand Colin, 1949.
[41] C. Ginzburg, "Clues: Roots of an Evidential Paradigm," in *Clues, Myths, and the Historical Method*, pp. 96–125, JHU Press, 1989.
[42] K. R. Popper, *The Poverty of Historicism*. Routledge, 1957.
[43] E. Nagel, *The Structure of Science: Problems in the Logic of Scientific Explanation*. New York: Harcourt, Brace & World, 1961.
[44] J. G. Carou, *Carlo Gozzi, Turandot (texto crítico italiano, tradución galega, introdución e notas de Javier Gutiérrez Carou)*. La Coruña: Biblioteca Arquivo-teatral Francisco Pillado Mayor, 2007.
[45] F. Horst, "Ecclesiastes," in *Biblia Hebraica Stuttgartensia* (K. Elliger and W. Rudorf, eds.), Stuttgart: Deutsche Bibelgesellschaft, 5 ed., 1975.
[46] Y. A. P. Goldman, "Qoheleth," in *Biblia Hebraica Quinta: Megilloth: Ruth, Canticles, Qoheleth, Lamentations, Esther*, Stuttgart: Deutsche Bibelgesellschaft, 2004.
[47] J. Lambek, *From Word to Sentence: A Computational Algebraic Approach to Grammar*. Open access publications, Polimetrica, 2008.
[48] B. Coecke, E. Grefenstette, and M. Sadrzadeh, "Lambek vs. lambek: Functorial vector space semantics and string diagrams for lambek calculus," *Ann. Pure Appl. Log.*, vol. 164, no. 11, pp. 1079–1100, 2013.
[49] N. Steenrod, *The topology of fibre bundles*. Princeton university press, 1999.
[50] V. A. Vasiliev and V. A. Vasiliev, *Introduction to topology*. No. 14, American Mathematical Soc., 2001.
[51] B. Liskov and J. Guttag, *Program Development in Java: Abstraction, Specification, and Object-oriented Design*. Computer programming, Addison-Wesley, 2001.
[52] F. Carrano, *Data Structures and Abstractions with Java*. USA: Prentice Hall Press, 3rd ed., 2011.
[53] J. Friesen, *Java XML and JSON: Document Processing for Java SE*. Apress, 2019.
[54] F. Boschetti and A. M. Del Grosso, "TeiCoPhiLib: A Library of Components for the Domain of Collaborative Philology," *Journal of the Text Encoding Initiative*, Jan. 2015. [on line journal].
[55] A. M. Del Grosso, D. Albanesi, E. Giovannetti, and S. Marchi, "Defining the Core Entities of an Environment for Textual Processing in Literary Computing," in *DH2016 Abstracts*, (Kraków), pp. 771–775, 2016.