

### 3 Towards the construction of a stemma

Introductory remarks by the chapter editor, Marina Buzzoni

The elaboration of a *stemma codicum*, representing the filiation between the witnesses that transmit a text whose original is lost, is the core of the genealogical method: on the one hand, only once these relationships have been determined can text restoration be tackled; on the other hand, the stemma may be the goal of the work of synthesising a certain textual tradition. In order to construct a stemma, some preliminary steps are needed; these steps are specifically treated in the sections of the present chapter.

The first step of the stemmatic workflow – namely, the identification of both direct and indirect witnesses (technically: heuristics) – is the subject of Gabriel Viehhauser’s contribution (3.1). After sketching a brief history of the concept, he addresses the issue of how the heuristic process is carried out after the material turn in the twentieth century, providing useful information about both the traditional and the more recent tools that researchers have at their disposal. Particularly relevant is the advent of digital catalogues and digital facsimiles, which can offer easier and faster access to primary sources. This development has profound consequences for framing the history of transmission of a text, as shown in the critical review of various *Parzival* editorial projects based on different heuristic approaches.

Caroline Macé (3.2) deals with a frequently neglected aspect of editorial practice: the use of the indirect tradition of a given text (e.g. translations and rewritings, quotations, interpolations, glosses, and marginal notes) for stemmatological purposes. The conclusion reached, namely that “the main point of using indirect witnesses is that their text has been preserved ‘outside’ of the main tradition; they can therefore be used as an ‘outgroup’ [...] to orientate the stemma”, is central from a methodological point of view. The indirect tradition can also be used to document the early history of textual traditions – especially when indirect witnesses are older than the oldest extant direct ones of a given work – as well as the appearance of (hyp)archetypes. Despite their relevance for stemmatic analysis, she warns us to use indirect witnesses with great caution due to the methodological difficulties inherent to them.

In her section (3.3), Tara Andrews addresses the problems of transcribing and then comparing (technically: collating) the different instances of a text preserved in several witnesses. In so doing, she presents both non-digital and digital ways of transcribing and collating witnesses, providing also some insights into the current theoretical debate on what these processes and the results they produce mean to different scholars and scholarly communities. She offers a definition of the central notion of a “variant location”, which arises when different witnesses show different readings at a point that can be considered “the same place” in the text. The discovery of these places is key to the establishment of a stemma, as the set of variant locations is the information with which a stemmatic analysis is performed. In a

traditional perspective, a distinction is primarily to be made between substantial readings and formal ones: usually, only the former are clues for determining the genealogical relationships between witnesses (see, among many others, Stussi 2006, 9–10). Andrews, however, discusses all variation (close to the traditional notion of *varia lectio*) – including, for example, spelling differences, abbreviation marks, and different letter forms – that may or may not later undergo a process of normalisation for the purposes of publication or for the purposes of stemmatic analysis, or both. The extent of normalisation, as well as the rules followed by the editors, depends on their judgement and the methods they adopt.

Once stemmatologically relevant data have been produced, they need to be represented, a need which is particularly acute when the editor chooses to take recourse to computational methods. Joris van Zundert’s section (3.4) focuses mainly on the representation in various digital forms of both input and output information for computational stemmatological analysis. This is highly relevant, for the aim of data formats is not just to ensure the proper storage of data but also to favour its processing by algorithms specific to the data they represent. Besides, van Zundert turns his attention to two further key points: (i) whether the chosen format is best suited to the type of analysis the editor wants to perform, and (ii) interoperability, since “the scholar should also consider how other scholars and other software may want to reuse the data, and whether the chosen format supports such reuse well”. Finally, he underlines that the choice of a specific data format may be influenced by considerations about the presentation of the data, either in separate form or within the broader context of a digital scholarly edition.

The four sections that make up this chapter demonstrate that even the steps that at first sight may appear merely descriptive or mechanical (e.g. the transcription of witnesses and their encoding using a given markup language) are actually always interpretative. In fact, they depend on the methods adopted for the analysis of the text and its witnesses, as well as ultimately on the very idea of textuality the editor embraces and intends to foster. The methods adopted may in turn be based on the type of textual tradition under inspection (e.g. an active tradition usually requires a different approach than a quiescent one; see “1970” in 2.4.3), as well as on the language of the text.

### 3.1 Heuristics of witnesses

Gabriel Viehhauser

In textual scholarship, heuristics is the identification and collection of direct and indirect witnesses (on the latter, see 3.2) of a text or a text corpus. Although often only discussed in the context of *recensio*, heuristics precedes *collatio*, *examinatio*, and *emendatio* in traditional outlines of textual criticism (see 6.2), and is commonly regarded as the first step of the editorial workflow.

Different philologies and disciplines arguably have specific perspectives on heuristics, mainly because of diverse research traditions but also because of the differences in the amount of extant witnesses that have to be dealt with (e.g. between Latin and vernacular traditions). This means that, although the following account aims at a comprehensive overview, it necessarily has to work with discipline-specific examples.

### 3.1.1 History

The idea of collecting witnesses to reconstruct a text can be traced back to the φιλόλογοι of the library of Alexandria, where, for instance, Callimachus of Cyrene (ca. 310–240 BC) compiled a catalogue of 120 volumes or Aristophanes of Byzantium (ca. 257–180 BC) established a bibliography of canonical Greek writers that had a decisive impact on their later transmission (Greetham 1994, 14–15). The library also collected different manuscripts of the same texts as a basis for the efforts of the φιλόλογοι (Greetham 1994, 15; on the case of Homer, see Plachta and van Vliet 2000, 15). However, a systematic concept of heuristics did not gain major importance until the emergence of critical philology in the nineteenth century (consider also the simultaneous development of the tripartite division between “heuristics”, “source criticism”, and “interpretation” in the “historical method” of the nineteenth century in historiography; Lorenz 2002, 139). In textual criticism, the insistence on a full survey of the extant transmission was based in particular on the rejection of the common practice of editing texts only on the basis of a single manuscript (especially the oldest manuscript or the vulgate version). According to Lachmann, an edition had to be built on a “hinreichende Menge an guten Handschriften” (Lachmann 1876, 1:82) [sufficient quantity of good manuscripts] (on Lachmann’s predecessors in this respect, see Timpanaro 2005, 115), which served as the foundation for a critical examination of the transmission (*recensio*). Therefore, it was not enough to consult the witnesses only occasionally (for the correction of individual errors); this had to be done systematically in order to gain an overview of the genealogy of the manuscripts beforehand from which to build the basis for all future editorial decisions. Besides direct witnesses, this also includes indirect witnesses (translations or quotations of the text to be edited), fragments, and anthologies that contain the text (on which, see 3.2).

However, in Lachmann’s conception, the manuscripts were of interest only insofar as they fostered the reconstruction of the archetype; Lachmann himself did not base all of his editions on the full range of known manuscripts because he did not see the need to go too far into the details of the transmission. Consequently, manuscripts were only relevant as witnesses of the text, but not in their importance as historical documents of their time, in other words in their materiality. This clearly changed with the material turn of philology in the twentieth century (see

Bein 2010). Against this backdrop, manuscript or print catalogues, which are of major importance as a tool for heuristics (see 3.1.3), obtain an interesting intermediate position between abstract indexes and detailed descriptions of the transmission, for they do not only register shelfmarks but also data about the provenance, language, layout, and material aspects of a manuscript. Thus, these traditional heuristic tools can also be useful for a kind of philology with a stronger orientation towards material aspects, one that is not only interested in the reconstruction of an archetype but also in the transmission history of a text. These two functions of a catalogue correspond to the distinction between the concepts of an enumerative vs an analytical bibliography: whereas the former confines itself to a list of sources, the latter also provides information with which to examine the sources as material artefacts (Greetham 1994, 7). The shift towards the materiality of texts is substantially helped by the advent of digital catalogues, which can be linked to digital facsimiles and therefore offer a more detailed picture of the manuscript cultures of the past. Thus, the same tendency that can be observed in the case of digital editions, namely the tendency towards a broadening of contexts (Sahle 2013, 2:168–172) fostered by the openness and the limitlessness of the digital medium, also holds true for digital catalogues: since catalogues do not have to be confined to printed book pages any longer, they can be enriched with various kinds of metadata and hyperlinks pointing to a huge amount of different online resources.

### 3.1.2 Implications of heuristics for building a stemma – an example

The different historical phases of attitudes towards heuristics, as outlined in section 3.1.1, have consequences for the devising of a stemma. In this respect, four phases may be discerned: (i) a pre-Lachmannian one, where the edition of a text did not necessarily imply a systematic pursuit of heuristics; (ii) an early phase of heuristics that meets Lachmann's stipulation to consider a sufficient basis of good manuscripts, but is as yet unable to draw on comprehensive catalogues of witnesses and on easily accessible sources; (iii) a phase where the heuristic work can rely on printed library catalogues and is thus based in principle on the whole transmission, but is sometimes still hampered by poor accessibility of the sources; and (iv) a phase that is shaped by the seemingly unlimited possibilities of the Internet and its digital resources. It may be added that the general approach, namely that of undertaking a study of the entire extant transmission, has, in theory, remained the same in phases (ii) to (iv).

Therefore, in the following, these four phases will each be characterised by a case study from the edition history of the Middle High German Grail romance *Parzival*, by Wolfram von Eschenbach, from the beginning of the thirteenth century. The text was one of the most successful German courtly romances, if its transmission is anything to go by. Today, sixteen complete manuscripts, one incunabulum from the year 1477, and around seventy fragments are known to be extant.

(i) The first modern print edition was established by Christoph Heinrich Myller, a student of the famous Swiss scholar Johann Jakob Bodmer, in 1784. The edition was based on a copy of St. Gallen, Stiftsbibliothek, Cod. Sang. 857 (the *St. Galler Epenhandschrift*) that Myller received from Bodmer. Bodmer himself knew two sources of the *Parzival* text: an exemplar of the incunabulum (which is now in the Zentralbibliothek Zürich, 2.103) and the St Gall codex. It is quite likely that Bodmer compared these two sources for his own works, which include modern adaptations of selected parts of *Parzival*, since it appears that the text of his adaptations is based on variant readings from both sources (Mertens 2011, 723). However, Myller, Bodmer's student, obviously did not strive to collect different exemplars for his edition, let alone to construct a stemma of the text, and only used Bodmer's modern copy of the St Gall manuscript.

(ii) Before Karl Lachmann, the first scholarly editor of *Parzival*, established his famous critical *Wolfram-Ausgabe* of 1833, he published an anthology of mediaeval texts which also included parts of Myller's edition, not without criticising the earlier editor for basing it solely on one manuscript (Lachmann 1820, viii; see Mertens 2011, 726). In order to prepare his own edition of 1833, Lachmann used two copies of Myller's edition, which he took with him on his travels to the libraries of St Gall, Heidelberg, and Munich. In order to collate the text, Lachmann inscribed the variant readings of the manuscripts into those copies (McCulloh 1983). Although Lachmann knew by this time of thirteen manuscripts as well as the incunabulum of *Parzival* (from a catalogue created by Friedrich Heinrich von der Hagen which was established in 1812), he himself did not use or even examine all of these sources for his edition, because he thought that, in the case of the *Parzival* transmission, three manuscripts were often reliable and representative enough to create his critical text (Schirok 1999, lix). Lachmann never devised a stemma of the *Parzival* tradition, but he claimed that the extant manuscripts can be grouped into two classes which are in principle "von gleichem werth" (Schirok 1999, xix) [of the same value]. These classes are known in *Parzival* philology as classes \*D and \*G. For the greater part of his text, he followed a representative of class \*D, namely the St Gall codex which had already been the basis for Myller's edition. In fact, it appears that, because his workflow relied heavily on the two exemplars of Myller's edition, Lachmann's text even inherited some of the errors that Myller had made in the reproduction of the manuscript (McCulloh 1983; see also below 7.4.1).

(iii) Since Lachmann was only interested in the tradition insofar as it (according to him) justified the text of his edition, more precise research on the stemmatic relationships remained to be undertaken by later scholars. Eduard Hartl, who was responsible for the sixth and seventh editions of Lachmann's *Wolfram-Ausgabe*, was the first of Lachmann's successors to try to reconsider Lachmann's findings on the basis of the whole manuscript tradition (of which by then all sixteen manuscripts and a large amount of fragments were known). Although Hartl was only able to publish one volume of the comprehensive *Textgeschichte* he had in mind (Hartl

1928), he identified four manuscripts which constitute a stemmatic group of their own (in Hartl's terminology, class *\*W*, now *\*T*). Lachmann did not know, or did not take into account, any of the manuscripts of this group for his edition. Even if Hartl was not very clear about it, it seems that he considered this group *\*T* to be a subgroup of *\*G*, but thought that it was heavily contaminated with *\*D*. The most striking evidence for this are twenty-two passages where *\*G* lacks lines compared to *\*D* (they are not necessary for the comprehension of the text and therefore cannot be considered *Bindefehler*). *\*T* partly shares this loss of verses, but only in eight of the twenty-two passages. The far more obvious stemmatic explanation for this observation, namely that *\*G* and *\*T* are both descendants of a group *\*GT*, was ruled out as unlikely by Gesa Bonath (1970). However, Bonath could base her judgement only on the variant readings of the first quarter of *Parzival* because she had to rely on Hartl's studies that remained incomplete (for details, see Chlench and Viehhauser 2014). Thus, it seems that there are mainly two reasons why the position of *\*T* in the stemma of *Parzival* was obviously misjudged by Hartl and Bonath: first, the reductive approach of Lachmann fostered a canonical notion of the *Parzival* transmission as split into the two groups, *\*D* and *\*G*, which was hard to overcome; and second, despite knowing all the extant manuscripts, Hartl and Bonath obviously did not have the resources to consider the relatively wide manuscript tradition in its entirety. Even if in the times of Hartl and Bonath printed catalogues provided potential support for a heuristics that enabled scholars to find all the known manuscripts, those manuscripts could not always be easily accessed and considered in practice.

(iv) A thorough examination of the *Parzival* tradition has been made possible by the digital *Parzival* project ([parzival.unibe.ch](http://parzival.unibe.ch)). The project aims at a digital edition of the text that considers all of the extant witnesses and provides digital transcriptions of them (Stolz 2002). In the project, digital phylogenetic methods have been used to visualise the stemmatic relations of the manuscripts (Stolz 2003). Along with the use of new methods, the project also offers a new attitude towards the transmission: instead of reconstructing an "original" text, it focuses on tracing the outlines of the three-centuries-long transmission history of *Parzival*. This also includes a new assessment of the classes of the text's witnesses. While Hartl and Bonath in principle considered *\*D*, *\*G*, and *\*T* as subordinated groups of the archetype, the *Parzival* project is based on four versions which are treated as manifestations of the text in their own right (on *\*T*, see esp. Schöller 2009). Besides *\*D*, *\*G*, and *\*T*, it was possible to identify a further class, *\*m*. While *\*m* is mainly transmitted in three codices of the fifteenth century produced in the workshop of Diebold Lauber, and shares a single reading with a very short (and therefore not very indicative) fragment from the thirteenth century (*F* 6), the discovery of a longer fragment from the fourteenth century in 2006 (*F* 69; see Schneider 2006) corroborated the evidence that the group is not a late redaction from the workshop but dates back to earlier times. As this example demonstrates, not only the availability of comprehensive catalogues but also the accessibility of the sources is crucial for heuristics.

This exemplary review of the history of *Parzival* philology shows that the assessment of the stemmatic relationships of a text sometimes cannot be seen independently from the material basis that underlies the philological endeavour. Whereas Myller only had a very constrained knowledge of the transmission and used a modern copy of a manuscript text for his edition, Lachmann could in principle have drawn on catalogues for his heuristic work; however, since he did not yet have microfiche copies or facsimiles of the texts at hand, he had to undertake demanding journeys to see the manuscripts, which he then had to collate in a way that consumed as little time as possible. It seems that his lack of interest in the details of the transmission goes hand in hand with the need to employ a practical approach towards the collection of the witnesses. While Bonath and Hartl could rely on more modern tools for heuristics, they too did not have unlimited access to the transmission. It could be argued that the picture of *Parzival* transmission in the first one hundred years of editorial attention was strongly shaped by insufficient means to pursue the ideal of a complete heuristics of the whole textual transmission, which is most strikingly illustrated by the fact that printing errors in Myller's edition can be found even in the later revised editions of Lachmann's *Wolfram-Ausgabe*. In the case of *Parzival*, a comprehensive view of the transmission was only achieved using the possibilities of a digital edition that includes electronic facsimiles and transcriptions of the text.

Of course, a case study like this can only show tendencies and should not be overgeneralised. In contrast to *Parzival*, in many other traditions it was possible to establish reliable editions on the basis of complete heuristics of witnesses even before the advent of digital methods. Furthermore, the example of the newly found fragment of class \**m* shows that, even in digitally informed times, it is conceivable that the discovery of hitherto unknown witnesses can change the assessment of the transmission.

### 3.1.3 Old and new tools for heuristics

Since there is no single printed bibliography that can cover all existing books or manuscripts, and bibliographies therefore necessarily have to be selective (see Greetham 1994, 5), the heuristics of manuscript witnesses very often has to be based on a variety of sources. A first starting point is provided by libraries and their catalogues (see 1.3). In the modern period, libraries began systematically collecting books in the thirteenth century. Prominent early examples of catalogues that exceed the scope of individual libraries by uniting different collections are the *Registrum librorum Angliae* and the *Catalogus scriptorum ecclesiae* (Bischoff 1990, 203; Russell 2001, 27–28; Greetham 1994, 18). Besides the emerging public or semi-public libraries, private collections of the new humanist scholars provided valuable resources (Greetham 1994, 18), but according to Greetham it was not until 1627 and Gabriel Naudé's theoretical treatise *Avis pour dresser une bibliothèque* “that a true systematic enumera-

- 39 **Bibl. Nationale, Fonds latin cod. 13955:** membr., 218×190, cc. 169 num. Minuscola della fine del secolo IX, a linee piene, senza elementi rubricati, fitta e sbiadita. Nei margini si leggono alcune glosse e fra esse una altotedesca: (c. 142v) *artemisia] bibodis*, e un'altra francese: (c. 144v) *rubarba*, le quali sembrano porre il volume in un ambiente bilingue. Il Jones (*The scriptorium at Corbie*, 390) lo elenca fra i manoscritti presenti in quell'abbazia. A c. 1r, di mano settecentesca, sono la nota di provenienza: *Si Germani a Pratis*, e le segnature: *olim 544, n. 1094, 16* (cfr. OMONT, *Concordances*, 93); ma non è indicato nel catalogo del monastero del 1677. Legatura in pergamena.

Contiene una miscellanea per lo studio delle arti liberali e specialmente di quelle del quadrivio con aggiunte sull'agricoltura e sulla medicina ed appunti di teologia. Così dopo un gruppo di estratti da Columella seguono:

1. <Antonio Musa, De herba vettonica liber> (cc. 137v-138r). È soltanto il trattatello con i sinonimi, la descrizione e gli usi: *Bettonica a grecis dicitur cestros — I. Ad capitis fracturam et ossa extrahenda. Herba bettonica tunsia et vulneribus capitis imposita — (XLVI. Ad podagram) ipsamque tritam et impositam dolorem lenire experti affirmant.*

2. <Apuleio Platonico, Herbarius, exc.> (cc. 138r-145r): *Plantago maior a romanis dicitur — estratti saltuari e alquanto rimaneggiati — (XLVIII. Petroselinum) nervorum dolores sedat.*

3. Estratti dal Liber medicinae ex herbis feminis attribuito a Dioscoride, dai Dynamidia e da altre fonti (cc. 145r-146r): *XLVIII. Abrotanum vel aeracion. Huius genera sunt duo — poi viola purpurea, elleborum nigrum, samsucus, yppericon, satireia, eruca, urtica, urtica cantirina, rubus, cicuta, fenum grecum — (Verbena) ad quartanas autem quattuor.*

4. Ricette (cc. 146r-147v): *Ut pili ENR — Apum percussus malvarum folia imposita continuo curant.*

DELISLE, *Inventaire des mss. de St. Germain des Prés*, 123: sec. X. G. SCHEPSS, *Zu Columella, Julius Victor, Macrobius-Plinius, Martianus Capella und PseudoApuleius in Blätter für das GymnasialSchulwesen* (Monaco), XXXII (1896), 407-08: sec. X. SANFORD, *The use of classical Latin authors in the Libri manuales*, 214, n° 187: sec. X. A. MUSAE *de herba vettonica*, PSEUDO-APULEI *herbarius* etc. ed. HOWALD e SIGERIST, XIV: sec. X.

**Fig. 3.1-1:** Page from a thematic manuscript catalogue (A. Beccaria 1956, 176) listing the content of a medical miscellany.

tive bibliography as related to the organization of book collections got under way” (Greetham 1994, 18). Catalogues may focus on manuscripts from single libraries or on specific languages (e.g. the catalogue of German manuscripts from the Universitätsbibliothek Heidelberg by M. Miller and Zimmermann 2007) as well as on specific temporal or thematic constellations (e.g. the catalogue of illuminated manuscripts of the thirteenth century from the Staatsbibliothek München by Klemm 1998, or A. Beccaria 1956 on pre-Salernitan Latin medical manuscripts; see fig. 3.1-1). Greetham (1994, 24–46), provides an extensive list of national and regional catalogues (especially for the United States, the United Kingdom, and France) and other bibliographical resources. An example of an important metacatalogue which assembles catalogues, inventories, and other resources for Latin manuscripts is Kristeller’s *Latin Manuscript Books before 1600: A List of the Printed Catalogues and Unpublished Inventories of Extant Collections* (Kristeller and Krämer 1993; Krämer 2007).



In recent times, access to these catalogues and other resources on textual witnesses has been substantially facilitated by the retro-digitisation of catalogues and resources (e.g. the online version of Kristeller and Krämer 1993 and Krämer 2007 on [mgh-bibliothek.de/kristeller](http://mgh-bibliothek.de/kristeller), or the extensive list of retro-digitised catalogues on [manuscripta-mediaevalia.de](http://manuscripta-mediaevalia.de)) and the advent of a vast amount of digital search tools on the Internet. The interlinking of different resources also opens up new possibilities for comprehensive research and the combination of hitherto separated knowledge bases. However, interoperability can only be achieved on the back of standardised metadata descriptions for entries (e.g. Dublin Core, [dublincore.org](http://dublincore.org); TEI, [tei-c.org](http://tei-c.org); OAI-PMH, [openarchives.org/pmh/](http://openarchives.org/pmh/); see S. J. Miller 2011). Digital catalogues, therefore, have to be diligently built according to such standards to reach their full potential and to increase their chances of long-term sustainability. In particular, techniques related to the Semantic Web and linked data appear to be promising for this endeavour (Burrows 2010; Baierer et. al. 2016). Once these standards are met, more detailed analyses and visualisations of the material also become conceivable: provenances of manuscripts (for instance) could be geolocated on a map, which might also lead to new insights that can be used for the heuristics of witnesses.

Since the online resources for manuscript research are manifold, divergent in scope, quality, methods, and aspirations, and – due to the fluctuation of the Internet – also sometimes only short-lived (see e.g. the overview of German portals in Stäcker 2010), it is not feasible to give a comprehensive list of all digital catalogues for all traditions in this contribution. Instead, the potential of online resources will be illustrated by an example, namely [handschriftencensus.de](http://handschriftencensus.de), which strives to list all German-language manuscripts and fragments from AD 750 to 1520 on a single website. The scope of the project hence encompasses approximately 26,000 witnesses that are held in over 1,500 libraries. The *Handschriftencensus* continues the efforts of the handwritten *Handschriftenarchiv* of the Berlin-Brandenburgische Akademie der Wissenschaften, which began a systematic list of German manuscripts in the early twentieth century (see Wolf 2007) and is now also available in a retro-digitised form ([bbaw.de/forschung/dtm/HSA/hsa-index.html](http://bbaw.de/forschung/dtm/HSA/hsa-index.html)). Compared to the handwritten catalogue cards of the *Handschriftenarchiv*, a digital collection like the *Handschriftencensus* offers refined search functions (manuscripts can be listed by authors, works, or libraries) and the possibility of linking catalogue descriptions with digital facsimiles. The website also includes a list of manuscript catalogues and a bibliography that can – like all the resources of the website – be continually updated.

As an example, a search for manuscripts of the *Parzival* tradition in the *Handschriftencensus* will outline a possible workflow for heuristics in the digital age. On its starting page, the *Handschriftencensus* offers two of the above-mentioned possibilities for accessing the database of witnesses: either by sorting the manuscripts according to the libraries that hold them (“Verzeichnisse” > “Handschriften”) or by searching for authors or works (“Verzeichnisse” > “Autoren/Werke”). With the latter approach, “Wolfram von Eschenbach” and “*Parzival*” can be searched for or



**Fig. 3.1-2:** The beginning of the list of *Parzival* manuscripts in the *Handschriftencensus* (handschriftencensus.de, accessed October 15, 2019).

selected from an alphabetical index. This search leads to a list of eighty-seven results that indicate the libraries and the shelfmarks of the known witnesses (fig. 3.1-2). Full codices are marked with a black square bullet point, fragments with a white one. Also, fragments that originally belonged to the same codex, but are now preserved in different libraries, are grouped together. By clicking on an entry, a full catalogue description of the witness can be obtained. It encompasses codicological details such as the number of folios, the size of the codex, a possible dating, and so on; the content (and context) of the codex; and finally a bibliography. If there are facsimiles (or parts of the bibliography) available online, the respective websites are linked. Due to the possibility of adding new entries to the list in the digital medium, the *Handschriftencensus* remains updated and also includes witnesses found only recently (e.g. the above-mentioned *F* 69, which plays an important role in the assessment of the *\*m* version).

## 3.2 Indirect tradition

Caroline Macé

Apart from direct witnesses containing a work (most usually manuscripts), the editor will be well advised to make an inventory of the indirect tradition, that is, of any other works or versions of a work that can bear witness to the history of the textual tradition in question, to the establishment of the stemma, and finally to the establishment of the text itself. This inventory is important for the history of the reception of the work, but may often yield some insights into the history of the tradition as well. Of course, these indirect witnesses are themselves generally preserved in manuscripts and have their own textual histories and editorial problems.

### 3.2.1 Types of indirect witnesses

The indirect tradition of a work may consist of

- (i) ancient or mediaeval translations of that work into other languages;
- (ii) quotations of longer or shorter portions of the text, especially in florilegia or in commentaries;
- (iii) interpolations into other works;
- (iv) adaptations of the text (epitomes, paraphrases, other recensions or redactions of the same work, and so on); and
- (v) paratextual elements (glosses, marginal notes, and so on).

In addition to this, direct witnesses preserved in other media than manuscripts (e.g. graffiti or papyri) or as underwriting in palimpsests may be considered similar to indirect witnesses since the text preserved in them may have followed different paths of transmission than the usual direct tradition.

The importance of indirect witnesses for the establishment of the stemma and of the edition will depend mostly on their antiquity and, more importantly, on their position in the stemma, on their fidelity to the original work (otherwise they may not be usable), and on the reliability of the editions through which they are accessible (or, failing that, of the witnesses transmitting them). In general, Dekkers and Hoste showed that, even in the case of well-preserved late antique works, the indirect tradition is crucial for establishing the text (*constitutio textus*; see 6.2 below): “Les citations anciennes sont une véritable pierre de touche pour distinguer les bons mss. des mss. corrompus” (Dekkers and Hoste 1980, 36) [Ancient citations are truly a touchstone for distinguishing the good manuscripts from the corrupted ones].

Some ancient and mediaeval works have an exclusively “indirect” existence, as all direct witnesses have disappeared and the work is known only through translations or citations. This is the case, for example, with three treatises on providence, free will, and evil by the Neo-Platonist Proclus Diadochus (fifth century CE; see 4.5.2), which are preserved in a thirteenth-century Latin translation and in citations (or plagiarism) by Isaac Komnenos the *sebastocrator* from the twelfth century (Isaac 1977, 22–25). See also the case discussed in section 4.5.4 below.

### 3.2.2 Translations

Late antique and mediaeval translations are potentially precious witnesses to the work from which they are translated, especially when they were made before the time of the oldest manuscripts of that work in its original language that are preserved. In the case of Greek and, to a lesser extent, Latin texts, manuscripts earlier than the beginning of the ninth century, that is, before the change from majuscule to minuscule, are relatively rare (see 1.2.3), and palimpsests or translations made before the ninth century are therefore very valuable.

When dealing with translations, scholars will face different types of problems. First, it is not so easy to find translations of a given work in languages with which the editor is not familiar. For Greek patristic works, the *Clavis Patrum Graecorum* (Geerard and Noret 1984–2018) often mentions translations into Latin (see also Siegmund 1949), the languages of the Christian Orient (Arabic, Armenian, Coptic, Ethiopic, Georgian, Syriac), and Old Slavonic. This is not done systematically, but it is nevertheless a valuable help. For several languages, scholars have provided lists and bibliographical tools, such as, for example, Graf (1944) for Arabic, Thomson (1995, 29–88) for Armenian, and the ongoing *Catalogus translationum et commentariorum* (Kristeller 1960–2003; Dinkova-Bruun 2014–2016) for Latin. Second, the translation might not be edited at all, as translations are often considered less important in the literary canon of a language, or edited in a way which does not meet modern standards (see Macé et al. 2015, 374, 435–439, on nineteenth-century editions of Armenian texts and editions of Syriac texts in the twentieth century respectively). An exemplary enterprise, but certainly not the only one, is represented by the critical editions of Arabic, Armenian, Georgian, and Syriac translations of Gregory of Nazianzus’ homilies (fourth century CE) in the *Corpus Nazianzenum*, with a special apparatus highlighting the differences between the translations and the Greek originals (see e.g. Coulie 1994; fig. 3.2-1 below). One important methodological rule for editions of translations is that one should not correct the translator’s mistakes, but only mistakes that may have appeared in the manuscript tradition of the translated text. In other words, the editor of a translation must attempt to reconstruct the text of the archetype of the tradition in the translation’s language, using the text in the source language as a hint for orientating the stemma, but should resist the temptation of “correcting” the translator’s text on the basis of that source. Moreover, the exact source used by the translator is often difficult to assess, and may not exist any longer.

### Example 1

In figure 3.2-1, the difference between the edited Armenian text and the Greek original, as indicated in note 10, is most probably due to a confusion of two words which are graphically close in Greek, but not synonyms (“εὐσεβῶν” [pious] and “εὐσεβειῶν” [piety], both in the genitive plural), by the Armenian translator; the editor of the Armenian text kept the translator’s mistake in the edited text. Divergences from the original can also point to original (primary) readings that have disappeared from the direct tradition, as with the reading given in note 8 in figure 3.2-1, reflecting the Greek “Βοσόρ”, the name of a city in the Old Testament whose “garments” are “red” (Isaiah 63:1, Septuagint), that is, stained with wine or blood, and therefore impure. This reading, also present in the Latin translation, is not found in any of the Greek manuscripts, which all have “βόρβορον” [filth] instead, obviously a simplification (see Dubuisson and Macé 2003, 307–308). In this case, “correcting” the seemingly strange reading “Bosor” in the Armenian text, to make it conform to the Greek text as it exists today, would have made a likely primary reading disappear from the indirect witness, where that reading is preserved as a kind of fossil.

զմանկականն թոթովէն, նախ քան զանցանկէ յաստուածային գաւթսն,  
յառաջ քան զաստուածային զրոյն ճանաչել եւ<sup>6</sup> զանուանսն, նախ քան  
զնորոյս եւ զՆորն զվերապարսն<sup>7</sup> ճանաչել եւ զվերակացուս, վասն զի  
նշ եւս ասեմ՝ նախ քան զրոտր<sup>8</sup> լուանալ եւ զզուոյն զազրութիւն  
որովք չարութիւն զմեզ կանխաւ չաղախեաց<sup>9</sup>, եթէ երկուս կամ երիւ  
բանս ի բարեպաշտութեանցն<sup>10</sup> կրթիցեմք եւ զնոսին ի լրոյ ոչ ի  
Հանդիպելոյ, կամ եթէ գԴաւթայսն<sup>11</sup> դուզնաբեայս խաւիցիմք, եւ  
կամ բանկոնաւ բարուք ինչ պանծնիցիմք, եւ կամ մինչեւ ի գաւտին  
իմաստասիրիցեմք, բարեպաշտութեան իմն ստեղծուածս եւ դէմս մեզ  
ինքեանց գունելով. բարձ գաւերիցութեան եւ իմաստութեան: Քաւա-  
նայ<sup>12</sup> ի խանձարոց Սամուէլ. վաղվաղակի եմք իմաստունք եւ  
վարդապետք եւ բարձրագոյնք յաստուածայինս եւ զզպրացն<sup>13</sup> եւ  
աւրինականացն առաջինք<sup>14</sup>, եւ ձեռնադրեմք զմեզ ինքեանս երկ-  
նայինս, եւ կոչել ի մարդկանէ ռարբի<sup>15</sup> Հայցեմք, եւ ոչ ուրեք զիր եւ  
ամենայն որ ինչ պարտ իցէ իմանալ Հոգեւորապէս, եւ աղճատանք

7 զմանկականն] զմանկան  $MM^1M^2$  թոթովէն] զթոթովէն  $JJ^1$ , թոթովէ  
 $MM^1M^2$  զանցանկէ] յանցանկէ  $JM^1M^2$ , անցանկէ  $N$  8 զաստուածային]  
յաստուածային  $J$  9 զնորոյս] զնորոյն  $M^2M^2$  եւ<sup>1</sup>] *om.*  $J^1$  զվերապարսն]  
զվերապարսն  $J$ , վերապարսն  $MM^1M^2$  10 ասեմ] ասեմն  $JJ^1M^2$  11 կանխաւ]  
ի կանխաւ  $M^2$  13 դուզնաբեայս] դուզնաբեա  $N$  եւ] *om.*  $M^1M^2$  18 բարձրա-  
գոյնք] բարձրագոյնս  $MM^1$  յաստուածայինս եւ] յաստուածայինսն եւ  $N$ , յաստու-  
ածայինսն  $MM^1M^2$ , յաստուածայինս  $J^1$  զզպրացն] վարդապետացն  $MM^1M^2$   
20 կոչել] կոչիլ  $JJ^1M^2N$  ռարբի] ռարբ  $J^1$ , Հարի  $JN$

6) *kān Gr.* (457 B 5); *kai Arm.*, cfr *PG* 35, col. 457 n. 79 et *BERNARDI* ed., p. 154.

7) *χαρκτήρα Gr.* (457 B 6); *χαρκτήρας Arm.*, cfr *PG* 35, col. 457 n. 80 et *BERNARDI* ed., p. 154 (*codd.*  $S^2DPC$ : accord de l'arménien avec la famille M).

8) *βόρβορον Gr.* (457 B 7); *βοσόρ Arm.*; sur ce mot, réf. à *Is.* 63, 1 dans *N.B.H.*, I, p. 505 et *STEPHANUS, Thesaurus*, III, col. 335.

9) *ὅσα ἡ κακία ἡμῖν προσεμάζετο Gr.* (457 B 8-9); *ὅσαις ἡ κακία ἡμῶς προσεμάζετο Arm.*

10) *τὸν εὐσεβῶν Gr.* (457 B 9-10); *«pietatum» (τὸν εὐσεβειῶν) Arm.*

11) *τῷ Δαβὶδ Gr.* (457 B 11); *τὰ (τοῦ) Δαβὶδ Arm.*

12) *Ἰερὸς καὶ Gr.* (457 B 15); *Ἰερὸς vel Ἱερεὺς Arm.*

13) La variante *վարդապետացն* (qui correspondrait au grec *διδασκάλων*) de  $MM^1M^2$  résulte d'une confusion entre *զպրացն* (*γραμματέων*) et la forme abrégée de la variante *զպրացն*.

14) *γραμματέων τὰ πρῶτα καὶ νομικῶν Gr.* (457 C 2-3); *γραμματέων καὶ νομικῶν οἱ πρῶτοι Arm.*

15) Cfr *Mt.* 23, 7: *եւ կոչել ի մարդկանէ ռարբի...*

Fig. 3.2-1: Special apparatus comparing an Armenian translation with a Greek original (Coulie 1994, 49). © Brepols Publishers.

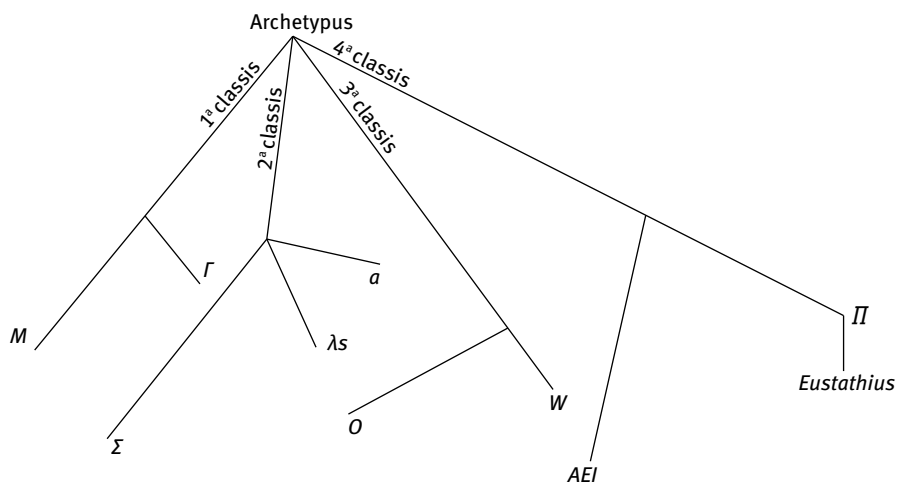
For the purpose of comparing them with the original works, translations can be divided into two types: *ad verbum* (according to the wording, i.e. literal) and *ad sensum* (according to the meaning, i.e. free; P. Chiesa 1987). The spectrum is continuous between extremely literal (to the point of becoming almost unintelligible; see Forrai 2012, 296) and extremely creative translations. There might be several translations of a given work even in the same language, or subsequent revisions of a given translation, or translations made not from the original directly but from

another pre-existing translation (this is very common in Latin; see e.g. Dolbeau 1989).

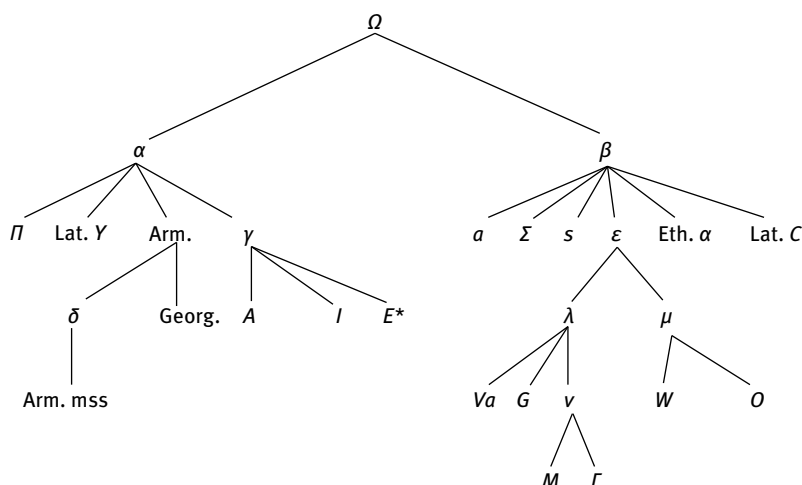
## Example 2

There are at least two different early Latin translations of the Greek *Physiologus* (an early Christian text moralising animal behaviour, edited by Sbordone 1936; see Pakis 2010 for an excellent *status quaestionis*). One of them (version C) exists only in two manuscripts (one of which is the famous *Physiologus Bernensis*; e-codices.unifr.ch/de/list/one/bbb/0318); the other (version  $\gamma$ ) was more widespread (Carmody 1941) and has an extremely large diffusion through adaptations in Latin and in vernacular languages (Henkel 1976; Orlandi 1985). An early mediaeval Armenian translation of the *Physiologus* exists as well, and was in turn translated into Georgian before the tenth century (Muradyan 2005, 5). In this case, the Georgian translation, being preserved in a much older manuscript than all extant Armenian ones, is an important indirect witness for the establishment of the Armenian text. In a review of Peeters (1898), Gottheil (1899, 120) drew a “pedigree of the *Physiologus* literature” (see fig. 3.2-2 below). Although outdated and now known to be wrong on several points, this diagram provides a good picture of the spreading of this work through many languages and over a large timespan. As defective as it may be, Gottheil’s bird’s-eye view of this tradition has not been replaced yet; research has made progress on some parts of the diagram, but not on all of them. Except for the Armenian translation mentioned above, the *status quaestionis* is no better today than in Gottheil’s time for any of the “oriental” translations, and Carmody’s edition (1941) of the Latin texts (of versions  $\gamma$  and B, the latter probably derived from the former) did not really improve our knowledge of the early stages of the Latin tradition. Sbordone’s edition of the Greek text (1936) took all known Greek manuscripts into account (and only a few more have been discovered since then), but neglected the ancient translations altogether in its stemma (see fig. 3.2-3) and critical text. Sbordone did, however, consider the Greek indirect tradition, especially a curious work attributed to Eustathius of Antioch (fifth century) and entitled *Commentary to the Hexaemeron* (the six days of creation), edited in a seventeenth-century edition (still the only one existing; *Patrologia Graeca*, 18:707–794). Even though this work is not an exegetical commentary, but probably part of a chronicle, and was not written by Eustathius of Antioch but dates from some time between the sixth and the eighth centuries, it is still valuable as an indirect witness because it quotes, more or less exactly, many passages from the *Physiologus* (see Macé forthcoming). Sbordone was able to locate the quotations from pseudo-Eustathius in his stemma (see fig. 3.2-3 below), but he gave too much credit to manuscript M (Milano, Biblioteca Ambrosiana, A 45 sup.), which he believed to be the oldest preserved Greek manuscript. In fact, earlier manuscripts exist (some known to Sbordone) but were at that time wrongly dated on palaeographical grounds; but this is not the real point. According to my own analysis of the tradition, M must actually be located rather “low” in the





**Fig. 3.2-3:** Stemma of the manuscript tradition of the oldest recension of the Greek *Physiologus* (Sbordone 1936, lxxix, redrawn and simplified).



**Fig. 3.2-4:** My own stemma of the *Physiologus* tradition, taking the ancient translations into account (previously unpublished, the Greek letters in lower case represent postulated hyparchetypes).

stemma (see fig. 3.2-4). In critically evaluating this manuscript tradition, the ancient translations, which were made a few centuries earlier than the Greek manuscripts, prove to be crucial. The agreements between the Armenian and Latin  $\gamma$  translations and the Greek manuscript  $\Pi$  (Moskva, Gosudarstvennyj Istoričeskij Muzej, Sinod. Gr., 467, dated to the eleventh century), as well as some more recent manuscripts, on the one hand, and the agreements between the Ethiopic and Latin  $C$  translations



and the other Greek manuscripts on the other hand, are a very strong argument in favour of a split of the tradition between two main branches (and not four, as Sbordone thought). The differences between these two branches are such that they cannot be explained simply as copyists' mistakes or involuntary interventions; they are traces of the existence of two recensions or redactions (within the oldest recension singled out by Sbordone) very early in the history of the tradition (see fig. 3.2-4). Sbordone's second family ("2a classis" in fig. 3.2-3) cannot be confirmed on the basis of common mistakes, and therefore it cannot be a family at all. Two further manuscripts (*G* and *Va*) can be added to the branch formed by *M* and *I*; they are both older than *M* and both from southern Italy (like *M*), thus clearly showing that *M*, which contains many singular mistakes, is but a member of a family of manuscripts which is not situated very high in the stemma.

### 3.2.3 Quotations/(auto-)plagiarism

In Antiquity and the Middle Ages, texts were "recycled", often without explicitly crediting the original author, sometimes to such an extent that the new work does not contain anything (or not much) else than excerpts from one or more previous writers (see P. Chiesa 2012, 381: "gran parte di tale letteratura è compilativa" [a large part of such literature [of the early Middle Ages] is compilatory]). One out of very many examples would be the letters that the monk Jacob Kokkinobaphos addressed to the *sebastokratorissa* Irene around 1040 (edited by Jeffreys and Jeffreys 2009). These letters are actually a *cento*, or a "tapestry of quotations" (Jeffreys 2012), taken from a large number of Greek Church Fathers. As Cassin (2018) has shown, those quotations are an important indirect witness, at least for the history of the reception, but also potentially for the history of the textual tradition, of Gregory of Nyssa's commentary on the Song of Songs. Compare also the case of pseudo-Eustathius quoting the *Physiologus*, as discussed above (3.2.2 – example 2). To provide another similar example from the Latin world, editors of Augustine of Hippo (354–430) are compelled to make use of Florus of Lyon's (first two thirds of the ninth century) compilations of extracts (Chambert-Protat 2014) because Florus had access to old manuscripts containing Augustine's works which are no longer extant. Yet another famous example is the citation of verses from the *Poetic Edda* in Snorri Sturluson's *Prose Edda* (beginning of the thirteenth century). Sometimes, these verses are preserved only there and not in the direct tradition, represented mainly by a thirteenth-century manuscript, the *Codex Regius* (*Konungsbók*) [Book of Kings], Reykjavík, Stofnun Árna Magnússonar í íslenskum fræðum, GKS 2365 4<sup>o</sup>.

An author may also reuse his own text in different places. For example, in the homilies of Gregory of Nazianzus, nine chapters are shared between homily 38 (on Christmas) and homily 45 (on Easter; Trisoglio 1965). For our purposes, it does not matter if Gregory himself did this or someone else interpolated the chapters of one homily into the other, because these chapters are present in both homilies in the

whole tradition. At any event, the same text in one homily can be used as an indirect witness for the other homily (see Dubuisson and Macé 2003, 315–317).

Of a different kind are *florilegia*, in which longer or shorter excerpts of works are not reworked to form a new work but displayed as such, often with the name of their author, sometimes even with the name of the excerpted work. Those *florilegia* are organised thematically or alphabetically, and transmitted through a more or less broad manuscript tradition. In Greek patristics, one of the most important of these *florilegia* is the so-called *Sacra parallela* attributed to John of Damascus (Thum 2018), preserved in several recensions (one manuscript, Paris, Bibliothèque nationale de France, gr. 923, is illustrated). When editing such a *florilegium*, the danger is the same as when editing a translation: that the editor may hypercorrect the text on the basis of the source (De Vos et al. 2008, 179).

In *mediaeval commentaries* (see 1.2.1), smaller or larger portions of the commented text are quoted, either as a lemma or in the body of the commentary, which constitutes another type of indirect witness. The commentaries may be transmitted as works in themselves or as *scholia* accompanying the commented work (on *mediaeval commentaries* and glosses in general, see Copeland 2012). As an example, one can mention the lemmata of Proclus' commentary on Plato's *Parmenides*, which are one of the oldest witnesses, albeit an indirect one, to Plato's text (see 4.5.2). The ancient *scholia* (from the Alexandrian school or from late Antiquity; see 1.2.1) accompanying the text of Homer or of the Greek tragedies in papyri or in Byzantine manuscripts can help in restoring the oldest layer of those texts (see e.g. the project of an online edition of Euripides *scholia*: [euripidesscholia.org](http://euripidesscholia.org)). See also Browning (1960) for the importance of marginal variants and *scholia* to classical literature sometimes preserved in recent manuscripts.

### 3.2.4 Interpolations

The term “interpolation” is sometimes used to designate the process of reusing and reworking previous works in a new one, but for this we prefer the term “excerpting” (see 3.2.3). By interpolation we mean the introduction into a text of a portion of text foreign to it. This is different from *gloss-incorporation* (see 4.3.2), which is usually unintentional; interpolation normally happens intentionally.

One problem is that interpolations are normally removed from the edited text of a given work (see 6.2.3) as foreign to that work, and might not even be mentioned in the introduction to the edition, and so in this way they remain out of reach for scholars. If the interpolation is interesting, it may be edited for its own sake. For example, a passage present in some manuscripts of homily 38 by Gregory of Nazianzus, which was obviously introduced at some point in the transmission process, has been edited in an article; unfortunately, it was impossible to identify its author (Macé 2004). Some works considered “heretical” by the official Church were preserved only as interpolations in orthodox works (Tuilier 1987).

If the interpolated piece of text belongs to a known work, it can be considered an indirect witness to the corresponding part of that work because it was transmitted outside of it. Unfortunately, such cases are rarely documented. When collating the Armenian text of Pseudo-Dionysius' *Epistula de morte apostolorum Petri et Pauli* (see 4.5.4) in the manuscript Erevan, Matenadaran 993 (a hagiographical-homiletic collection copied in 1456), I discovered that the copyist (or his model) had interpolated into Dionysius' text a passage which belongs to the *Martyrium Pauli*, a second-century apocryphal text existing also in Armenian translation and transmitted in, among other manuscripts, Matenadaran 993. The text of the interpolation offers a variant reading (*ew asē pawłos c'neron* և ասէ պաւլոսս ցնէրոն [and Paul speaks to Nero]) which is not found in the direct tradition of the *Martyrium*, where all manuscripts read "and he speaks to Caesar"; as a direct witness to the *Martyrium*, manuscript Matenadaran 993 presents a rather long omission including the passage in question (Calzolari 2017, 637).

### 3.2.5 Adaptations

Ancient and mediaeval texts were not only often reused; they were also often reworked: abridged (*recensio brevis* or *brevior*); summarised (epitome); expanded (*recensio fusior*); transposed into another genre, typically from poetry to prose or vice versa (paraphrase); rephrased (recension/redaction); and so on (on the Byzantine vocabulary and practice of rewriting, see Signes Codoñer 2014).

Depending on how deep the changes reach, another recension can or cannot be used as a direct or indirect witness to the transmission of the work. The Greek *Physiologus* (see 3.2.2), for example, is known in several recensions: three prose works in learned Greek, one verse adaptation, and one rewriting in "vulgar" mediaeval Greek (Sbordone 1936). Even in the prose adaptations, the text of the oldest recension was so much altered that the other two recensions cannot be used to establish the text of the first one. However, as far as the first recension is concerned, I was able to determine that the Greek tradition and the ancient translations are divided into two main branches, representing two different redactions of the same work. It is therefore possible to use the witnesses of one redaction to polarise some variant locations in the other redaction. For example, the adverb "καλῶς" [well] is present in manuscripts *α* *ς* *Σ* of redaction *β* (see fig. 3.2-4), but omitted in *Γ* *Μ* *Γ* *Ο* *W*. In the corresponding passage of redaction *α*, the same adverb is present. Therefore (unless one supposes a contamination of one redaction by the other), the omission in *Γ* *Μ* *Γ* *Ο* *W* must be secondary, and if this is confirmed by other cases, it points to the existence of a hyparchetype common to these five manuscripts (*Va* is lacunary at this place). Similarly, Godfried Croenen has shown that it is possible to use one (authorial) recension of Jean Froissart's *Chronicles* (fourteenth century) to orientate the stemma of the manuscripts of the main recension (for a short methodological discussion, see Croenen 2010).

### 3.2.6 Paratextual elements

Marginal or interlinear corrections or indications of co-occurring variants in manuscripts can be the result either of philological emendation by mediaeval readers or of collation with other witnesses. In the latter case, these variants are sometimes accompanied in Greek manuscripts by γράφεται, “it is written [elsewhere]”; see fig. 3.2-5) or ἐν ἄλλῳ, “in another [witness]”. For an example of a thorough collation of the text of a manuscript against another manuscript, lost in the meantime, see section 4.5.2 on Bessarion’s (fifteenth-century) corrections to his exemplar of Proclus’ commentary on Plato’s *Parmenides*. In figure 3.2-6, the tenth-century copyist of the text (homily 38 by Gregory of Nazianzus, on Christmas) wrote in the margin next to the words “τὸ θεῖον” [the divine]: “ἐν ἄλλῳ τοὺς θεοὺς γραφὲν εὖρον” [in another [manuscript] I found “τοὺς θεοὺς” [“the gods”] written]. Interestingly, this variant is not found in the text of any still-extant manuscript, only in the margin of Paris, Bibliothèque nationale de France, gr. 515 and also in the margin of the codex Milano, Biblioteca Ambrosiana, E 50 inf., one of the two remaining illustrated uncial manuscripts of Gregory of Nazianzus’ homilies.

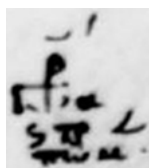


Fig. 3.2-5: Sinai, St Catherine’s Monastery, gr. 399, f. 115r, marginal note: “γράφεται καὶ τραπῶμεν” [it is also written “τραπῶμεν”].

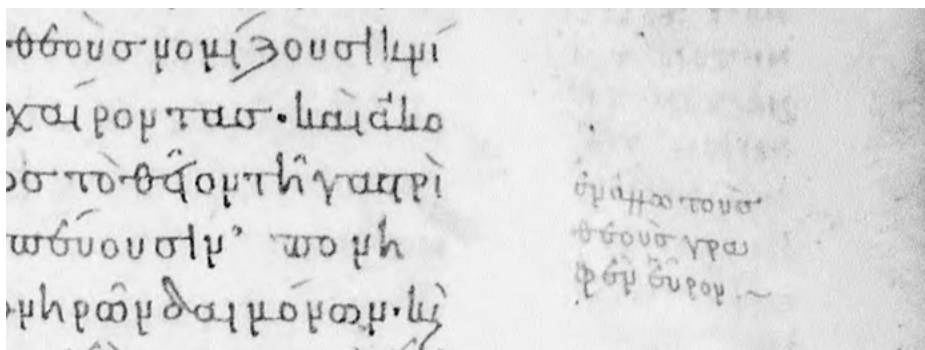


Fig. 3.2-6: Paris, Bibliothèque nationale de France, gr. 515, f. 124r, marginal note. Source: Gallica, Bibliothèque nationale de France, [gallica.bnf.fr/ark:/12148/btv1b107215420/f128.image.r=grec%20515](https://gallica.bnf.fr/ark:/12148/btv1b107215420/f128.image.r=grec%20515). Image: CC-BY-NC.

In the same way, glosses and other types of marginalia can be used as indirect witnesses: see Buzzoni (2011) for an example of the use of glosses in the study of the *Heliand* tradition.

### 3.2.7 Direct witnesses preserved in other media

Palimpsests (see 1.2.2) and papyri have often preserved works, especially from Greek and Latin literature from Antiquity, which would otherwise have been lost (Reynolds and Wilson 2013, 195–199). The text preserved on palimpsests or papyri is usually fragmentary because the material support may be heavily damaged and because the script has vanished (especially in the case of palimpsests) and may be hard to decipher. Because of this, the texts found in papyri and palimpsests are usually the object of a documentary edition and require a specific methodology (see Gippert 2015).

If they are not unique witnesses, palimpsests and papyri will usually be the oldest direct witnesses to works otherwise preserved. For example, the papyrus codex edited by Capron (2013) contains saints' lives of which there are also manuscripts and ancient translations – this allows a comparison between the fragments of the text preserved on the papyrus and the other witnesses. Because papyrus was a relatively cheaper material than parchment, papyri often preserve types of texts which may be characterised either as *Gebrauchsliteratur* or as documentary in nature (especially with commercial content), or sometimes as “school books” (see Turner 1968). For this reason, and not only because of their age, papyri are potentially very interesting witnesses, precisely because they did not necessarily follow the same “literary” path of transmission as manuscripts. There is an interesting methodological discussion about the place of papyri (the same could more or less be said about palimpsests as well) in a stemma. Collomp (1929) refutes the idea of the “eclecticism” of papyri (a theory according to which papyri would often contain readings from several families of mediaeval manuscripts), arguing that, because they are much older than the manuscripts, the variants they attest may often go back to the archetype or even predate it. The same line of argument was often taken by Irigoien (e.g. Irigoien 1968–1969, 138). In unpublished papers translated by Most, Timpanaro returns to this question as well (2005, 207–215).

Although they are much rarer, graffiti can also shed light on the history of a text otherwise transmitted in manuscripts. One famous example is the mediaeval (thirteenth- or fourteenth-century) graffito found in a cave near Vardzia (Georgia) containing two strophes from Shota Rustaveli's epos *The Knight in the Panther's Skin*, which is preserved in manuscripts, none of which are older than the sixteenth or seventeenth century (Gippert 2018, 157). Graffiti preserved in Pompeii also offer their share of literary verses and epigrams, for which they are amongst the oldest witnesses (see Milnor 2019).

Inscriptions can also serve as indirect tradition for literary works otherwise transmitted through manuscript tradition. For example, De Simini showed that extracts from two treatises written in Sanskrit (*Śivadharmaśāstra* and *Śivadharmaṭṭara*) and known through late manuscripts are quoted in mediaeval works and also known through inscriptions from as early as the eleventh century, much earlier than the manuscripts (De Simini 2016, esp. 237). Conversely, manuscripts can be used as an indirect tradition for existing but damaged or lost inscriptions: in his edition of Byzantine stone epigrams, for example, Rhoby (2014) more than once uses transcriptions of the inscriptions preserved in manuscripts.

### 3.2.8 Using indirect witnesses

In all the cases mentioned here, the main point of using indirect witnesses is that their text has been preserved “outside” the main tradition; they can therefore be used as an “outgroup” (see 5.2.1, 8.1.3.4) to orientate the stemma and document the early history of the textual tradition and the appearance of (hyp)archetypes. These witnesses should be used with great caution, however, and some of the methodological difficulties inherent to them have been highlighted above. Nevertheless, when they exist, indirect witnesses are indispensable for gaining access to the earliest stages of a tradition because direct witnesses to these earliest stages are usually missing. In this way, they often provide a clue for understanding how the tradition developed from the earliest stages on, and they will often help orientate the stemma, as illustrated in 3.2.2 (example 2 and fig. 3.2-4). Ancient translations are especially important in this respect because they have preserved larger portions of text than other types of indirect witnesses. Unfortunately, translations are not always considered in the process of editing, for reasons that have been explained above (3.2.2). It may indeed be perfectly justified not to make use of indirect translations; but, as a rule, one should never believe that direct witnesses alone are enough, because there is always much to be gained by looking at the indirect tradition of any work.

## 3.3 Transcription and collation

Tara Andrews

Once the manuscript witnesses to a text have been gathered and scrutinised for the clues they might give about the transmission history of a text, the individual text instances must be compared so that their similarities and differences may be analysed. This is the phase of textual criticism known as *collatio* (see 2.2, 6.2). In a digital environment, it is increasingly common to separate this phase into two distinct steps, transcription and collation. When the collation is made, a list of variant locations, or in certain cases *loci critici*, can be produced for further analysis. It

should be noted that, although stemmatic analysis eventually requires the editor to distinguish significant from insignificant variation, this cannot be done before the texts are compared in the first place; the question of how to make that distinction will therefore not be treated in depth in this section.

### 3.3.1 Definition of terms

Transcription is the act of transferring a text from one carrier to another. Normally, this refers to a transfer from one medium to another: for instance, the transcription of a recorded speech, or the transcription of a handwritten document into a corresponding digital form. The word may also refer to the textual version, or document, that results from this act. Transcription may also be said to happen in the process of collation if the editor chooses to collate texts without digital assistance (Nury 2018, 109–111). Such collations, however, are not normally considered “transcriptions” in the usual sense of the word.

Collation is the act of comparing different instances of a text; a collation is a document that contains the result of this comparison. A collation can take a number of different forms. Non-digital forms can include marginal notes on a physical version of a text, or a series of tabular records (fig. 3.3-1).

Digital forms of a collation can include a spreadsheet that mimics or extends the non-digital form of tabular collation (normally referred to as an alignment table), an XML document (Andrews 2009) or relational database (Robinson 1989) that stores a list of textual variants, or other less commonly used data structures such as the “multi-version document” advocated by Desmond Schmidt and Robert Colomb (2009). The advantage of a digitally stored collation is that, under most circumstances, it can be transformed more or less automatically into an apparatus of variants, an alignment table, or a variant graph (see 3.4) for display and examination. This is true no matter which format has been chosen to store the collation, although the particular mechanics of the transformation will vary.

A variant location arises when different manuscripts show different readings at a point that can be considered “the same place” in the text. Figures 3.3-2a–c show, in each of the various visualisations, an example of a variant location – the point in the collated text where “ἡκριβωκόντων” (perfect passive participle of ἀκριβόω, “to make exact or accurate”) appears in most manuscripts but an alternative, “ἡκριβηκόντων”, appears in manuscripts *P* and *S*. Variant locations are the units of change upon which almost all methods for stemma construction operate.

### 3.3.2 Transcription

One of the first decisions that must be made by the philologist who works with a particular text is to determine the extent to which transcription of that text is neces-

1. 4264  
stanza 610

HOCCELEVE RP COLLATION SECS. 11 and 14  
(Hoc<sup>2</sup> lacks these sections so omitted)

A:	And	opneth	hys	dore	and	doun	goth	hys	way
Ar:	✓	openyth	his	doore.	✓	✓	✓	his	wey
Ad:	✓	opned	his	✓	✓	gop dōw	←	his	wai
As:	✓	openeth	his	✓	✓	dōw	✓	his	✓
Bo:	✓	openyd	his	✓	✓	dōwne	✓	his	✓
Co:		looks his pence							
Co:	✓	opned	his	✓	✓	dōw	goth	his	wey
Di:	✓	openyth	his	✓	✓	dōw	goth	his	wey
Do:	✓	✓	his	✓	✓	dōwne	goth	his	wey
Du:	✓	open	his	✓	✓	dōwne	gop	his	✓
Ed:	✓	opned	his	✓	✓	dōwne	goth	his	✓
Fi <sup>1</sup> :	✓	openeth	his	✓	✓	dōw	✓	his	✓
Fi <sup>2</sup> :	✓	openeth	his	✓	✓	dōw	goth	his	wey
Ge:	✓	openeth	his	✓	✓	dōw	goth	his	wey
Ge:		MS looks not 1 gōh 1 cōh							
Ha <sup>1</sup> :	✓	openeth	his	✓	✓	dōwne	✓	his	wey
Ha <sup>2</sup> :	✓	openyth	his	doore.	✓	dōw	✓	his	wey
Ha <sup>3</sup> :	He	opnyd	his	✓	✓	dōw	went	his	wey
Ha <sup>4</sup> :	✓	opned	his	✓	✓	dōwne	✓	his	wey
Hh:	Aftir	opnyd	his	✓	✓	dōw	gop	his	wey
Hh <sup>1</sup> :	✓	opned	his	✓	✓	dōw	✓	his	wey
Hh <sup>2</sup> :	✓	openeth	his	✓	✓	dōwne	✓	his	wey
Kk:	withre	opnyth	his	✓	✓	dōw	goth	his	wey
La:	✓	openyd	his	✓	✓	dōw	✓	his	wey
La:	✓	openeth	his	✓	✓	dōw	✓	his	wey
Le:	✓	✓	his	✓	✓	dōwne	✓	his	wey
La:	✓	openeth	his	doore.	✓	dōwne	✓	his	wey
Qu:	✓	opned	his	✓	✓	dōwne	✓	his	wey
Ra <sup>1</sup> :	✓	✓	his	✓	✓	dōwne	✓	his	wey
Ra <sup>2</sup> :	✓	opned	his	✓	✓	dōwne	goth	his	wey
Ro:	✓	openeth	his	✓	✓	dōw	✓	his	wey
Ry <sup>1</sup> :	✓	openyth	his	door/	✓	dōw	goth	his	wey
Ry <sup>2</sup> :	✓	openeth	his	✓	✓	dōw	✓	his	wey
Ry <sup>3</sup> :	✓	open	his	✓	✓	dōw	goth	his	wey
Ry <sup>4</sup> :	✓	✓	his	✓	✓	dōw	✓	his	wey
Se:	✓	openeth	his	✓	✓	dōw	✓	his	wey
Sj:		MS looks not 1 S 1 C 1 h							
Sl <sup>1</sup> :	✓	✓	his	✓	✓	dōw	goth	his	wey
Sl <sup>2</sup> :	✓	openyth	his	✓	✓	dōw	goth	his	wey
So:	✓	openeth	his	✓	✓	dōw	gop	his	wey
Tc:	✓	openeth	his	✓	✓	dōw	✓	his	wey
Ya:	✓	opnyd	his	✓	✓	dōw	✓	his	wey

NOTES:

opneth } Ad-As-Bo-Co-Di-Fi<sup>1</sup>-Fi<sup>2</sup>-Ha<sup>2</sup>-Ha<sup>4</sup>-Hh-Hh<sup>1</sup>-La<sup>2</sup>-Ra<sup>2</sup>-Ry<sup>2</sup>-Se-Ya  
 = Ox 8 + Co-Di-Ry<sup>2</sup> + Ad + opned (these only, not syllables marked)  
 Hz<sup>2</sup> + Hz<sup>4</sup> + Hh<sup>1</sup> + Ra<sup>2</sup>

Fig. 3.3-1: Example of a tabular collation: Thomas Hoccleve's *Regiment of Princes*, line 4264. Hoccleve Archive, University of Texas Libraries. Image: CC-BY-NC-SA.





choose to transcribe only one text in full. This would then become the “base text”, against which all other texts are compared. The relative trade-offs of computer-assisted vs manual collation will be discussed below, in section 3.3.3.

### Digital transcription

Insofar as the vast majority of critical editions produced nowadays are done with the computer in some form, the focus here is on modes of digital transcription. There are several possibilities for how to transcribe a manuscript text; the editor’s choice will depend on the later use to which the transcription will be put. Perhaps the simplest option is to make a plain text transcription; this entails typing the text of the manuscript into a text editor or word processor, and saving it in plain text format (see 3.4.5). The primary advantage of this approach is its simplicity. Many philologists, however, will quickly discover that the inability to use more than the most basic formatting becomes more of a hindrance than a help.

At this point, many philologists will be tempted to use the more advanced formatting features provided by word processing software – to change the font size, include footnotes, use colour or superscript formatting to represent additions or deletions, and so on. This must be avoided, unless the philologist intends that the transcription should never be imported into another tool! Hardly any word processor or file formats can be read reliably by other programs; if the transcription is to be used further, it would need to be saved as plain text, and the formatting features in question would be lost.

### Markup languages and markup schemes

To address this problem, the best solution currently available is to use a markup scheme. By far the most well known of these is the XML scheme provided by the TEI consortium and described in the TEI guidelines ([tei-c.org/p5](http://tei-c.org/p5); see also 3.4 below). These guidelines provide a way to describe, in a form that is more or less machine-readable, the vast majority of textual and palaeographical phenomena that occur in manuscript texts. TEI XML has been the transcription format of choice for the vast majority of digital edition projects since the early 1990s, and has a large community behind its use. Users of TEI can also draw on a well-developed ecosystem of tools and programming libraries to parse XML documents, search and query them, and transform them into common online display formats such as HTML, EPUB, and PDF.

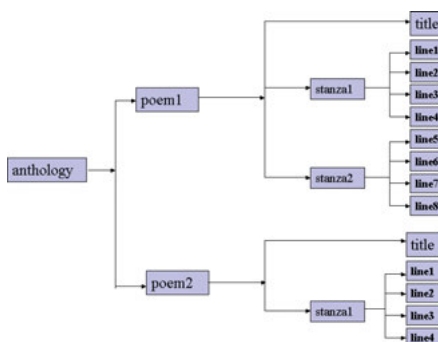
XML-based markup of text is justified by the OHCO model – the idea that text can be expressed as an “ordered hierarchy of content objects” (DeRose et al. 1990). The hierarchy imposed by XML syntax is a strict one: a text must be modelled, conceptually, as a branching (but never merging) tree (see figs 3.3-3a–b for an example). A text, for instance, can contain front matter, main body, and back matter; the main body can contain chapters, which contain paragraphs, which contain sentences, and so on.

```

<anthology>
  <poem>
    <heading>The SICK ROSE</heading>
    <stanza>
      <line>O Rose thou art sick.</line>
      <line>The invisible worm,</line>
      <line>That flies in the night</line>
      <line>In the howling storm:</line>
    </stanza>
    <stanza>
      <line>Has found out thy bed</line>
      <line>Of crimson joy:</line>
      <line>And his dark secret love</line>
      <line>Does thy life destroy.</line>
    </stanza>
  </poem>
  <!-- more poems go here -->
</anthology>

```

**Fig. 3.3-3a:** Example XML markup for a poem.  
Source: [tei-c.org/release/doc/tei-p5-doc/en/html/SG.html](http://tei-c.org/release/doc/tei-p5-doc/en/html/SG.html).



**Fig. 3.3-3b:** Corresponding hierarchy model for the poem in fig. 3.3-3a. Source: [tei-c.org/release/doc/tei-p5-doc/en/html/SG.html](http://tei-c.org/release/doc/tei-p5-doc/en/html/SG.html).

Alongside the increasingly widespread adoption of XML for text transcription came the realisation that the OHCO model is not always entirely adequate to describe a text (e.g. Renear, Mylonas, and Durand 1996). How, for instance, should the scholar deal with a quotation that begins in the middle of a paragraph and continues to the next paragraph? How should a manuscript text be made to fit into a strict hierarchy that its author, or its scribe, had no conception of when the text was written, and would therefore quite often violate? One can imagine, for example, an authorial rewrite of three and a half lines of text that cross a chapter boundary, or an annotation added to the margin of a manuscript that refers to a portion of the text not precisely defined.

These objections to the OHCO model have led some scholars to propose alternative schemes for text markup; perhaps the best known of these is LMNL (Piez 2014), which rejects the idea of a strict hierarchy, allowing arbitrary regions of the text to be annotated without regard to their place in the overall text structure. LMNL is not widely used, however, owing primarily to the lack of the technical infrastructure that makes XML so popular.

### Normalisation for transcription

Alongside choosing a format, the next decision that a scholarly editor must make is the extent to which the transcription should be normalised for spelling, punctuation, layout, and so on. Here, the editor places the transcription on a continuum between the idea of a documentary transcription (Pierazzo 2011), in which every feature of the manuscript is represented as faithfully as possible in the chosen medium, and an interpretative transcription, in which the text of the manuscript is represented in a way that minimises the differences between versions.

There is no one “correct” level of normalisation to be observed in the transcription phase. The extent to which a text is normalised will greatly affect the possible

results of collation and identification of variants, which will in turn have an impact on any stemmatic analysis to be done. If the editor chooses the more labour-intensive documentary approach at the transcription phase, there remains the opportunity to apply normalisation techniques in a later phase of text collation. If, on the other hand, the editor chooses at the outset to produce normalised transcriptions, the collation can never be made to reflect any manuscript variation that was omitted at the transcription stage. In making this decision, scholars should carefully consider their overall purpose in editing the text, as well as any material or time constraints on the project.

### 3.3.3 Collation

Although the acts of transcription and collation are often regarded as separate steps in digital workflows for critical editing, many textual scholars regard the collation as a distinct entity in its own right, comprising the text of the individual witnesses and the correspondence between them, inseparable from the acts that go into its creation. The collation is not only the centrepiece of a critical edition of a text, but also what makes any sort of analysis of the transmission of a text possible. Without a collation, there can be no stemma. We therefore need to understand what a collation is and how this might vary depending on context.

In recent decades, the concept of what a collation is has evolved, and varied, according to the aims of the editor whose definition is used and according to the capabilities of the time. Into and beyond the 1960s, one conceived of a collation as a process carried out with reference to a base text, usually some kind of norm such as a published edition (Colwell and Tune 1964, 253). By the early 1990s, perhaps spurred on by the adoption of computer technology, the relative ease of splitting text automatically into individual words based on the spaces between them, and the wide availability of algorithms for pairwise comparison, collation was described as the comparison of “two genetic states or two versions [...] of a text” (Grésillon 1994, 242) and something that was done “word for word” (Stussi 1994, 123), albeit still with respect to a reference text. Computational methods allowed this precision to be taken farther still, as is demonstrated by another definition of collation as an act that was carried out “character for character” (Shillingsburg 1996, 134). This definition is striking in another aspect: rather than referring to comparison with a base text, its author calls for the comparison of “all versions that could conceivably have been authoritatively revised or corrected”. It is around this time that the notion of the base text ceases to be a central part of the definition of the collation. Later scholars define collation as an act whose purpose is to find agreements and divergences between witnesses (Plachta 1997, 137) or explicitly to track the descent of a text (Kline 1998, 270); they differentiate between collation as a process of comparison (carried out “word-for-word and comma-for-comma”; Eggert 2013, 103) and the

result of comparison, which is known as the “historical collation” (Greetham 1994, 4); or they describe collation again as a process, whose result is described simply as lists of variant readings (Greetham 2013, 21).

From these descriptions, it is possible to detect a converging (though also evolving) definition of collation, and a distinction between the act and its result. Collation may be carried out against a reference text, pairwise, or as a many-to-many comparison. The comparison may be done at the word level, at the character level, or at another unspecified syntactic or semantic level, according to the sensibilities of the editor. The question of authority enters the picture with Shillingsburg’s definition (1996); this arises more in modern genetic criticism than in classical or mediaeval textual criticism, but conveys the idea that some manuscripts may represent definite departures from the “original”, “authorial”, or “main” text and that these might therefore be left out of a collation. The purpose of collation is usually given as being the discovery of where witnesses to a text converge and diverge; one might also claim that its purpose is to track the descent or the genesis of a text.

The act of collation produces a result, also known as a collation. Although the term “collation” can be used for the set of data that results from the process in any of its forms (whether that be a spreadsheet based on a copy text, a list of variants keyed on an existing edition, or even a digital object such as a JSON-format alignment table produced by collation software programs), it usually has a more specific meaning. Eggert (2013, 103) uses for this the term “historical collation”, by which he means “an extended report” on the substantive variants between the texts. It is important to note here that the historical collation is almost always a curated and pruned version of the results of comparison of the text, a fact to which Eggert also alludes when he writes that the historical collation “is often restricted to [...] ‘substantives’, leaving the now-orphaned commas and other ‘accidentals’ to look after themselves”. In that sense, the collation, as many textual scholars understand it, is a document that reflects not only the “raw” results of comparing a text but also the scholarly work of interpreting these results into a particular argument about the constitution and history of that text.

Here, however, it would be useful to draw a distinction between the collation and the critical apparatus. These things can easily be conflated; for example, Greetham (1994, 4) refers to the *apparatus criticus* and historical collation as a representation of a “collation and the results of emendation”. A reader might deduce from this that, for Greetham, a “historical collation” is the *apparatus criticus* of an edition minus any emendations. This is, however, almost certainly a misinterpretation of his words. Whereas a collation is a catalogue of variant readings in a text and may or may not be constructed with reference to a base text, an *apparatus criticus*, as its name implies, is a record of variants that takes the critically established text as its point of reference. In fact, the *apparatus criticus* may restrict itself to those variants judged to be genealogically revealing, that is, “significant errors”. Maas (1960, 8) even goes so far to say that only the non-mechanically decidable readings of the

archetype, which he calls “variant-carriers”, deserve a place in the critical apparatus; in this case, even the substantive readings would be omitted if they were clearly secondary. Since a collation is a necessary prerequisite to the *constitutio textus*, and the *apparatus criticus* is a result of this process, it is clear that they cannot be the same thing. This distinction also serves to explain why, contrary to the expectations of many users of a critical edition, textual witnesses can almost never be reconstructed in full from the edited text and its apparatus.

### Manual collation

A collation can, naturally, be made without the use of automated alignment tools. In this case, the scholar will follow the advice of West (1973, 66): write down the differences between each manuscript and a reference text. West recommends the use of a printed edition for this; if no edition is yet in print, the scholar can choose a manuscript copy of the text that seems well suited for the purpose. According to West, the collator should record even apparent trivialities in orthography, as they may be unexpectedly useful in constructing the stemma or otherwise understanding the relationship between manuscripts; this is, in essence, an argument for keeping normalisation to a minimum at the transcription phase. West also recommends including information in the collation about page divisions, scribal or second-hand corrections, and so on.

### Automatic collation

In order to use any sort of automated collation software, every manuscript witness needs to be transcribed in full; the software operates on the basis of these transcriptions to identify and align the readings they contain. The author of one of the first well-known text-collation tools was initially taken with “the notion of feeding these manuscripts into one end of the computer, which would then extrude a critical apparatus on the other” (Robinson 1989, 99). His tool, COLLATE, was eventually designed to work interactively and closely with the editor. Robinson included the facility not only to align variant readings, but also to normalise selected readings and to choose the readings that should constitute the edited text, so that the result was not merely a collation but essentially a fully constituted text and its *apparatus criticus*.

The current generation of collation tools, on the other hand, limit themselves strictly to the act of comparison; the authors of the CollateX tool describe collation simply as text comparison and refer to it as a process (Haentjens Dekker et al. 2015, 453). The process of collation around which these tools are based, also known as the collation workflow, is known as the “Gothenburg model” after its definition there at a workshop in 2009. The workflow is composed of discrete steps – tokenisation, normalisation, alignment, analysis, and visualisation – which, taken together, form the process by which a scholarly collation artefact is generally produced.

**Tokenisation** refers to the subdivision of a text into discrete units suitable for comparison. Normally this is done word for word, but depending on the language, structure, or grammatical rules of a text, the units might comprise multiple words (e.g. “et cetera”, “sine qua non”) or, on the other hand, might split words apart (e.g. “filio-que”).

**Normalisation** refers to the decision, for each token in the text, about whether to compare it to other tokens in its precise literal form, or whether to treat it as being a version of another known word for the sake of alignment. If spelling normalisation was not incorporated into a transcription process, it is often done here. Other examples of normalisation include the use of morphological analysis tools such as stemmers (which produce the root stem of a word, so that, for example, “give” and “given” are recognised as corresponding readings), the conversion of spelled-out numbers into their modern numerical equivalents (e.g. representing both “forty-two” and “XLII” as “42”), or the use of sound-value software such as SoundEx to account for shifts in spelling (as in Birnbaum 2014). It is important to realise that, in terms of automatic collation, the purpose of this normalisation is *not* to produce a canonical version of each reading, but merely to provide hints for a better alignment of the variant texts.

**Alignment** is the meat of an automatic collation process, whereby the (normalised) tokens of each text are compared with each other, and a proposal is produced for how they correspond to each other. The result of an alignment most often takes the form of a table, as described above; it may also take the form of a variant graph (see below).

**Analysis and visualisation** must follow any automatically produced text alignment. A good visualisation of results allows for a meaningful analysis, which is the scrutiny of a proposed alignment by the textual scholar. The purpose here is to evaluate the overall correctness of a given alignment. A scholar may choose to adjust the approach taken to tokenisation or normalisation until a satisfactory alignment is produced; alternatively, the scholar may wish to use the alignment as a starting point for producing a satisfactory collation without rerunning the automated steps.

### **Manual vs automatic collation**

The choice of automatic or manual collation is a topic on which most scholars will eventually develop strong preferences, as well as strong opinions on which is faster or more efficient. In the case of automatic collation, the bulk of the work is in transcription of the source witnesses – a task that for some edition projects is too daunting to contemplate, but is perfectly feasible for others. Depending on how normalisation is handled during the transcription process, the collation workflow steps described above usually progress very rapidly once the transcriptions are finished. The use of automatic collation has two advantages: first, that the scholar emerges from the process with detailed transcriptions of each manuscript witness, and sec-

ond, that there is no need to preselect a base text for comparison of witnesses. This allows for an easier and more flexible construction of an *apparatus criticus* in the final edition.

For manual collation, on the other hand, the bulk of the work is in the meticulous comparison and alignment of, and record-keeping about, a succession of witnesses. In practice, a manual collation must be done with reference to a base text chosen for the purpose at the outset. The scholar must therefore be very sure of the suitability of the chosen base text; once the collation has begun, a change of base can mean the repetition of an enormous amount of work. Moreover, with manual collation there can be an increased incentive for the scholar to speed up the process by disregarding the advice of West and omitting variation that is deemed to be trivial. This is particularly true for edition projects where full transcriptions were considered to be impractical, since a manual collation is essentially a codified form of normalised transcription. Insofar as a collation is meticulous and complete, it is possible in theory to reconstruct individual witness transcriptions from the collation itself; however, any variation omitted from the manually produced collation cannot be reconstructed at a later stage.

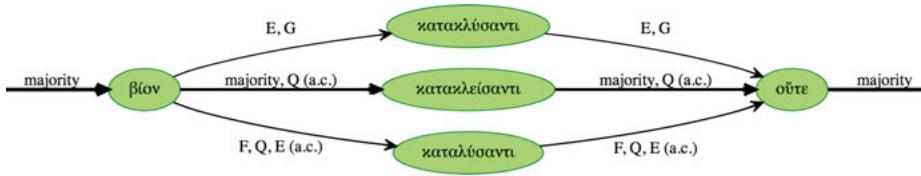
Where the text is unusually long and the number of its manuscripts is unusually large, as with, for instance the *Divina Commedia* of Dante (see the discussion of overabundant traditions in 6.2.2.1), an alternative to collating the full text is to select a certain number of *loci critici* (see 3.3.4 for a full definition) which are considered representative of the manuscript tradition and which will be used as the basis of comparison and stemma creation. In order to use this method successfully, the editor must be able to justify the selection of particular passages; for this, it is necessary to have a thorough and detailed grasp of the text, its manuscripts, and the sorts of variation they display.

### 3.3.4 Variant, variant location, and *locus criticus*

When a collation has been made, the scholar is left with information on where the witnesses to the text can be seen to differ, and what those differences are. A text, as carried in a particular witness, can be thought of as a series of readings – that is, a series of lexical units that are the reader’s interpretation of the marks upon the page. An individual reading is often equivalent to a word, but in certain contexts might be multiple words (see the discussion of tokenisation above), or in other contexts might be suffixes (e.g. “-que”). Since each reading has its particular place in the text sequence in a given witness, each reading can be thought of as having a location.

To collate a text, then, is to align these sequences of readings. Once that is done, the collation (particularly in its form as an alignment table) contains an overall sequence of locations and, for each location, a set of readings that occur in that





**Fig. 3.3-4:** A variant location as represented in a graph. This location includes corrections made by the scribe at the time of copying; the state of the text before these corrections has been denoted with the abbreviation “a.c.” (*ante correctionem*). In this example, witness *E* has been corrected from “καταλύσαντι” to “κατακλύσαντι”, and witness *Q* from “κατακλείσαντι” to “καταλύσαντι”.

Referentie Variant (regel)	nr.	Griekse tekst	
		Basistekst	Variant
1.41.2.	38	κατακλείσαντι	κατακλύσαντι (1) καταλύσαντι (2)

**Fig. 3.3-5a:** A variant location with a lemma specified (here as “Basistekst”).

location across the collated witnesses. When this set contains more than one reading, those readings are known as *variants*, and the place where they occur is their *variant location* (see figs 3.3-4–3.3-5). The discovery and definition of these locations is key to the establishment of a *stemma*, no matter the method used for the *stemma* construction. The set of variant locations is the information, deriving from a collation, with which a *stemmatic* analysis is done.

Depending on the methods adopted for *stemmatic* analysis of the text, the editor may designate a subset of these variant locations to be *loci critici*: those places in the text where the variation is believed to betray information about the copying (i.e. text-genealogical) relationships between the manuscripts – that is to say, those places that show “significant” variation – and on which construction of the *stemmatic* tree should be based (see 2.2.5, 4.3.1). In some cases (see 3.3.3) the *loci critici* will be chosen prior to collation; they can also be chosen based on the results of collation, either of the whole text or of samples from it.

Whatever means is chosen to create the *stemma*, the editor will eventually use it to work through all variant locations in the text and make a choice about which (if any) of the extant readings should become part of the critical text (see 6.2 for a fuller description of this process). This reading will be designated as the *lemma*, and from that point on the term “variant” will refer specifically to the readings at the given location that differ from the lemma.

While most textual variation represented in a collation will concern the set of variants at a single location within the text, there are a few sorts of variation that comprise multiple locations. One common example of variation across locations is

1.27 ὥς...οἶδε] om. P<sup>ac</sup> (add. in mg. P<sup>ead</sup>. manu) 28 καθαίρει  
 αὐτόν] καθεαυτὸν S | εἰ] ἡ T, ὁ P 29 εὐχαριστία] εὐχαριστεία A PS  
 G 31–32 ἀχαριστίας] ἀχαριστείας A P F 35 πολλάς] πολλάκις  
 CPS 37 εὐ ποιῶν] εὐποιῶν PS 37–38 ὑπεραπελογήσατο ] ὑπερα-  
 πολογήσατο S 39 ἀπιστία] ἀπιστεία A T | ἑτέραν] om. PS 40 οὖν]  
 om. EGK | τῶ] τῶν K | ἀπιστία] ἀπιστεία A P D 41 κατακλείσαντι]  
 καταλύσαντι E<sup>ac</sup>FQ<sup>sl</sup>, κατακλύσαντι E<sup>sl</sup>G

Fig. 3.3-5b: The same variant location as represented in an *apparatus criticus* (highlighted).

textual repetition, for example when a scribe copies the same line of a manuscript twice in a row. As another example, when a reading has been moved in the text relative to other manuscripts, we speak of transposition. Some editors, however, refer to this as translocation; for them, transposition refers only to the situation where two readings have been swapped with each other. That is, given a base text that reads:

The **white** cat played with the dirty ball,

an example of translocation would be:

The cat played with the dirty **white** ball,

and an example of “true” transposition would be:

The dirty cat played with the **white** ball.

Editors may also speak of inversion, which is when the transposition involves two contiguous words, for example if one manuscript reads “ἔστῳσα βοτάνη” where the others read “βοτάνη ἔστῳσα” [grass standing].

The question of how to represent transposition (or translocation) adequately in a collation is a complex one. For the case where the transposition or translocation is not a simple inversion but is still relatively isolated (i.e. it comprises only one or a few contiguous readings), it usually suffices to add a row to the collation table (assuming one is collating manually). The collation software CollateX can also attempt to detect and mark translocations in its output, which works reasonably well for small and isolated cases. However the collation is rendered, the presence of a transposition will usually lead to overlapping entries in the resulting *apparatus criticus* in order to accommodate variation within the component readings as well as the transposition itself (see figs 3.3-6a–c).

In general, if the text to be collated includes substantial dislocation of text, the resulting collation – and the resulting apparatus – can quickly become very intricate and complex, especially if the editor also wishes to note variants of individual readings within the dislocated segment. One strategy is to collate these segments in a separate table and indicate in the main collation where this segment is located, in which witnesses, with reference to the rest of the text. If the editor is using auto-



Fig. 3.3-6a: A transposition as rendered in a CollateX result graph. Note that the words “βέλος έστι” [is an arrow] appear twice.

Referentie Variant (regel)	nr.	Griekse tekst	
		Basistekst	Variant
9.1.1.	60	βέλος	μέλος
9.1.2.	61	βέλος έστι πεφραμακευμένον	πεφραμακευμένον βέλος έστι

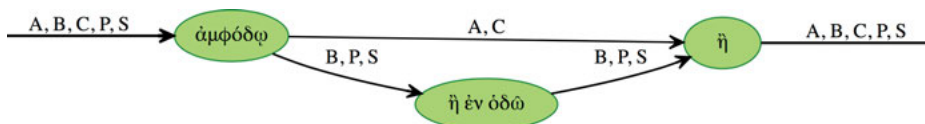
Fig. 3.3-6b: The same transposition marked in a tabular collation. Note the need for the reading “βέλος” [arrow] to appear in both collation rows.

9.1 Ὅψις] ή praem. H	βέλος] μέλος T	βέλος έστι] post πεφραμα-
κευμένον transp. BPS	4 οὐ] οὔτε Q	πανδήμοις] δήμοις DEGHK,
οὐκ έμφιλοχωρεῖ add. EGK	οὐδὲ] οὔτε Q	5 περιάγει] παραβάλλει
Q 7 πάρεργον] παρανάλωμα H	τῶν έχθρῶν] τὸν έχθρόν A,	τὸν έχ-
θρῶν T <sup>a.c.</sup> (τῶν T <sup>p.c.</sup> )	8 μὴ δῶς] μηδαμῶς S	αὐταῖς] αὐτῆς P
12 λα-		
λοῦσι] καὶ praem. Q	13 σεμνῶς] σεμναὶ P	13–14 άγνείας] άγνοίας P
15 άνένευσεν] άνένευσον P		

Fig. 3.3-6c: The same transposition represented in an apparatus criticus.

mated collation tools, it is often a good idea in any case to break the text into discrete logical segments, as the collation results improve markedly. If some of these segments are dislocated in some witnesses, a record will need to be kept of the respective order of collated segments across all witnesses.

Additions to, and omissions from, a text are also spoken of in reference to a variant location. These arise from the situation where, at a particular location, certain witnesses have no reading at all. In the absence of a base text and before the establishment of at least a preliminary stemma, it is impossible to label the missing readings as omissions, or conversely to label the readings that do exist at that location as additions. The terms addition and omission therefore gain their meaning only when a particular reading at the location (or, indeed, the absence of a reading there) is given some form of authority. Moreover, in the particular case of additions and omissions, the added or omitted text is usually considered a single reading, no matter its size (see fig. 3.3-7).



**Fig. 3.3-7:** An example of a reading that has been either added or omitted. The editors eventually chose to regard this as an addition (which is not expressed in the graph).

### 3.3.5 Normalisation for collation

Given the wide variety in ancient and mediaeval orthography, a full and exact collation of diplomatically transcribed witnesses will often show a great deal of variation that does not seem to impact the sense of the text; this may include spelling variants, abbreviation marks, accents, or variant letter forms. While in some cases, for instance in mediaeval works such as the *Hildebrandslied* (Baesecke 1945) where the dialect is itself an object of study, the editor may wish to retain every orthographical detail, in other cases the editor, or the reader, would prefer to reduce the amount of variation to be dealt with. In this case the editor will subject the text to a second process of *normalisation*, which usually involves the substitution of a reading with a canonically written form of that reading. The obvious challenge, then, is to define what constitutes “canonical” in each particular situation. The particular conventions will depend as much on the text in question and the degree to which differences must be scrutinised as on the history and norms of the language and writing system that is employed in the text.

For the purposes of publication, the editor will usually choose a single set of orthographical conventions to use; each reading that is recognised as carrying the same text will then be written in the same manner. For example, the article written as “τον” in a manuscript may be given its usual accent and be rendered as “τὸν” in its normalised form; spelling will also be normalised. It follows naturally from this that the question of whether two readings carry the same text is in most cases up to the judgement of the editor. For example, if the two readings “κρίνεται” and “κρίντετε” are seen, which are both forms of the verb κρίνω, “to separate; to choose”, the editor may choose to regard the second as a spelling variation of the first, given the context of the surrounding sentence and the fact that the pronunciation of αἰ and ε in Greek had already stopped being distinguishable in the classical period. On the other hand, it is also possible to conclude that, since these are two distinct recognisable conjugations of κρίνω (third-person present singular middle/passive indicative and second-person present plural active indicative, respectively), one should not be normalised to the other.

For the purposes of stemmatic analysis, entirely separate rules for normalisation may be needed. Here, we come back to the advice of West to note even apparent trivialities; it may be that a peculiar spelling or abbreviation of a word, or even a strange shape of a glyph, turns out to explain the emergence of an otherwise inex-

plicable copying error in the textual tradition and thereby shed light on the stemma, even if it is correctly interpreted in a different branch of the tradition. An example of sorts can be found in book 25 of the *Speculum historiale* of Vincent of Beauvais; the vulgate printing of 1624 (Vincentius Bellovacensis 1624) erroneously numbers as “Cap. CXIX” a chapter which is marked merely as “C.XIX” in its manuscript exemplars but is indicated more clearly as “Capitulum XIX” in other witnesses (e.g. Roma, Archivium Generale Ordinis Praedicatorum MS XIV.28b). It is, strictly speaking, debatable whether this error constitutes a significant one in the Lachmannian sense, and it concerns a reading that is almost paratextual in nature, which an editor may well be tempted to skip altogether in the collation; nevertheless, it does reveal some information concerning the textual transmission. Textual scholars will thus, in their analysis, often want to take into account variation, such as spelling, peculiar orthography, or even ink highlights, that is unlikely to be desired in a printed apparatus. To complicate the matter even further, some automated collation tools provide a string substitution feature so that a given string can be collated in place of the reading itself. This is also referred to as “normalisation”. For example, given the two sentence fragments “the lazy dogs” and “the lazy sleeping dog”, the user of a collation tool will want to ensure that “dog” aligns with “dogs”, and may achieve this by normalising both words to “dog”, “ANIMAL”, or even “NOUN”.

It is usually at the normalisation stage that editors must confront the question of how to handle punctuation within the text (see also 4.3.4). This first requires an answer to the question of whether punctuation marks should be regarded as readings in their own right, or simply as aids to the interpretation of the words and sentences they accompany. The latter point of view provides many editors with the justification necessary to discard entirely (or almost entirely) the punctuation of the manuscript witnesses once it has played its role in the interpretation of that witness’s readings; punctuation is then reintroduced in the finished edition based on the conventions that modern readers expect (see 4.3.4). If, instead, punctuation is treated at the level of readings (either as independent reading tokens or combined with the readings it accompanies), then it too must at some stage undergo normalisation, and the considerations set out here concerning normalisation of readings in general also apply.

### 3.4 Data representation

Joris van Zundert

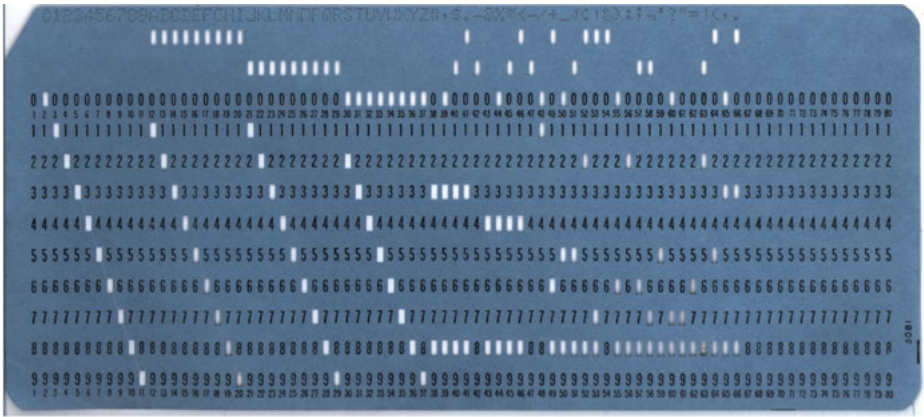
These days, as an editor, it may be hard to avoid the use of computers altogether. Even scholars aiming exclusively at a printed book edition will usually prepare such an edition using a computer. When inferring a stemma, the use of computers becomes even more likely. If a textual scholar opts to use digital tools for this, once

direct (3.1) and indirect (3.2) witnesses have been found and collated (3.3), the resulting data needs to be represented and stored in a digital environment. Computational stemmatological analysis requires variant information as input and results in various kinds of output. Both input and output information can be represented in various digital forms. This section provides an introduction to the more technical sides of dealing with digital data representing texts available in different versions: the make-up and pros and cons of various digital data formats that scholarly editors may encounter when they start working with computational means and digital methods. The actual critical study of the textual variation, often leading to a critical edition, will follow in the next three chapters.

### 3.4.1 A tiny history of the genesis of storing text as digital data

As soon as digital computing became practicable for researchers in the 1950s and 1960s, textual scholars started to move text into the digital environment. Often, Father Roberto Busa is pointed to as a founding figure (Jones 2016), but many other examples of early work exist (Raben 1991; Nyhan and Flinn 2016, 2–4). In particular, tasks in concordancing and lexicography – tedious, repetitive, error-prone – lent themselves to the convenience of automation by computer. Another strand of work that applied computational means early on was stylometry. Stylometry is the study of quantitative aspects of style, mostly known for its frequent successes in authorship attribution – for example the identification of J. K. Rowling as the person behind the pseudonym Robert Galbraith (Juola 2013). Stylometry developed from earlier painstaking statistical work without the aid of computers. George Zipf, for instance, found in the 1930s that there is an inverse and roughly logarithmic relationship between a word’s rank in a frequency table and the times it appears in a text (D. Holmes 1998, 112). That is, the most frequent word will occur almost twice as often as the second most frequent word, and three times as often as the third most frequent word, and so forth.

Computer-aided analysis of text required these early scholars to figure out how to actually record text in such a way that it could be both digitally stored and made processable for computers (i.e. made machine-readable). The first “digital” scholars therefore followed the computer engineers that were already familiar with storing text encoded as numbers. At a most basic level, the central processing units (CPUs) of computers process tiny voltage changes as discrete signals that represent binary states. In human terms, the central chip in a computer “listens” to voltage changes in the tiny electric currents that run through it. A higher voltage level is associated with a 1 or the logical value “true”; a lower voltage level is accepted as representing 0 or the logical value “false” (Crosley 2015). These bits, as atomic units of digital computing, can be used to encode higher-level representations. For instance, a set of 8 bits (a byte) can be used to represent numbers from 0 (all bits zero: 00000000), through 1 (one bit turns to 1: 00000001), to 255 (all bits 1: 11111111). Then, if we



**Fig. 3.4-1:** Standard IBM punch card used commonly in the 1960s. Source: commons.wikimedia.org/w/index.php?title=File:Blue-punch-card-front-horiz.png&oldid=241324408.

agree on a certain translation table, it is possible to have specific numbers represent characters (Null and Lobur 2003, 62–76), as for example is depicted in table 1.

Table 1: Translation table encoding characters as (binary) numbers.

character	A	B	C	D	E	F	G	H	...
decimal	65	66	67	68	69	70	71	72	...
binary	1000001	1000010	1000011	1000100	1000101	1000110	1000111	1001000	...

In the early days of digital computing, such numerical values representing characters would be recorded on punched cards. Punch card systems to record information had been around for a long time. They were already used, for instance, in the first half of the nineteenth century to have Jaquard looms weave the same patterns into cloth (Ceruzzi 2012, 7–9) and to direct the play of pianolas (Petzold 2000, 239). Punch card systems similar to these were developed to feed numerical information into the early computers of the twentieth century via standardised punch cards (fig. 3.4-1). Meticulously standardised, precise places in the columns on a punch card corresponded to particular numerical values. Punched with a hole at those particular places, a column could thus represent or “hold” a numerical value; several such column values together would then, in turn, represent a character. Whole stacks of cards could, in this way, record a complete text.

Thus, the first ways of recording – and more importantly storing – digital data were very much through analogue carriers: straightforward tabular cards of sturdy paper. Since that time, the carriers have changed quite a bit. First, stiff paper punch cards were replaced by magnetic tape (which recorded magnetic “stripes” rather

text stream	meaning of control command
.ce	center next line
Introduction	
.ju	justify right margin
.ll 39	line length to 39 columns
.ss	use single spacing
The work of authors, publishers, and researchers involves, in varying degrees, three recognizable text	

**Fig. 3.4-2:** Example of typesetting commands inserted in a text stream (adapted from Goldfarb 1997, 659).

than punch holes). Tape cylinders were replaced by magnetic hard disks. Hard drives are still by far the most-used digital storage medium today, but they are finding “competition” in solid state disks (SSDs). SSDs are electronic chips recording digital bits in flash cells, microscopic containers that can hold electrons or not, each cell simply representing again a 1 or a 0.

**3.4.2 From storing character streams to textual formats**

It is a bit of a no-brainer for textual scholars that text is not just a linear and one-dimensional series of characters (see e.g. Buzzetti and McGann 2006; DeRose et al. 1990), but the first computers and computer languages offered few possibilities for capturing, expressing, and handling more elaborate text structures. The amount of information that could be handled was severely limited. Univac 9000 systems of the mid-1960s typically took up the space of a medium-sized conference room, and could store about 32 kilobytes of information. A typical modern smartphone may well boast 64 gigabytes of memory, which would be two million times more than such a Univac computer. The average novel contains 90,000 words, while the Univac could, with the best compression achievable, represent perhaps some 10,000 words. For early computational textual scholarship, cleverly encoding and storing information was an important challenge in itself, let alone representing complex text structures.

The practical need to meet the challenge of storing and representing more structure came from the publishing world. Typesetters had a very concrete need not just for electronically representing the characters of text in the right order, but also to know what needed to be in bold print or italics, where a page number went, and so forth. This was solved by inserting typesetting control codes into the linear character stream (Goldfarb 1997; see fig. 3.4-2 below).

More elaborate markup systems like SGML (Standard Generalised Markup Language) and XML (eXtensible Markup Language) were eventually developed from



```
<sentence>Many years later, as he faced the firing squad,
<protagonist><mil_rank>Colonel</mil_rank> <name>Aureliano
Buendía</name></protagonist> was to remember that distant
afternoon when his father took him to discover ice.</sentence>
```

**Fig. 3.4-3:** A gentle example of XML marked-up text with opening and closing tags, for instance “<name>” and “</name>”.

these typesetting languages. These markup languages use a controlled vocabulary and grammar to express information about the text. Usually, angle-bracketed “tags” are then used to point out what part of the text that information pertains to (fig. 3.4-3).

### 3.4.3 What makes a good data format?

This tiny history serves to point to three foremost tenets of digital formats: formalisation, storage, and processing capabilities. As the roots of both *formal* and *format* suggest, formats are concerned with formalising the form in which we record information. It is only by conforming exactly to such an agreed-upon form that we can make information machine-readable. Computer programs are extremely bad at lenient interpretation. Consider the following markup examples: “My name is <name>Ismael</name>” and “My name is (name)Ismael</name>”. A human reader would have little trouble inferring the intent of the markup, even with the “bracket typo” being present. Most computer programs developed to process XML markup, however, would simply choke on this typo, error out, and stop processing the data.

The exact formal structure of a data format is put into a written technical specification. In the case of XML, for instance, the W3C (World Wide Web Consortium) is the ruling body that has issued the precise specification for the data format or markup language (see [w3.org/TR/xml](http://w3.org/TR/xml)). Anyone wanting to implement a computer tool that is going to process XML in some way should take these exact specifications into account. Technical specifications, however, are usually not the most gentle introduction to digital formats for users who are not highly specialised programmers. In the case of XML, a very successful and widely used data format, the last two decades have therefore seen a bewildering flood of handbooks and tutorials, from technical “bibles” such as *XML Unleashed* (Morrison 1999), geared towards professional programmers, to the proverbial *XML for Dummies* for a more general public (Dykes and Tittel 2005). Obviously, similar documentation on most of the other widely used data formats is available for the novice user.

Next to formalisation, the point of many data formats is simply to be able to store data for later use. When a computer program quits, it is removed from the memory of the computer, including any possible data that was associated with it. If such data were not stored safely somewhere, it would just evaporate. From the point of view of textual scholars, there is a point to be made that storing digital text should be done in a fashion that warrants some long-term sustainability. The mere

```
"Name","Sex","Age","Height(in)","Weight(lbs)"~
"Alex","M",41,74,170~
"Bert","M",42,68,166~
"Carl","M",32,70,155~
"Dave","M",39,72,167~
```

**Fig. 3.4-4:** Example of CSV data (after Burkardt 2016).

formalisation of a data format according to a technical specification at least goes some way to guaranteeing this, but there are other aspects to consider. A rule of thumb is that, the simpler and the more open a format is, the more it will be resistant to changes in digital platforms, operating systems, programs, and so forth. A very complicated format requires equally complex programs to read the data and make it usable for a user/reader. An open format means a format of which the specification is shared publicly (such as with the XML standard of the W3C). This enables and allows any programmer to create software that will read the data format. In contrast to open standards, there are also closed or proprietary formats, whose internal structure is known only to the original creators and is legally protected. In the realm of textual studies, Adobe's PDF (Portable Document Format) and .doc (the Microsoft Word document format) used to be the best-known and most widely used closed proprietary formats. This meant that all works stored as such files could only be read via software originating from the vendors in question. Fortunately, nowadays it is pretty rare to come across digital text formats that are not open. Even PDF and .doc (supplanted mostly by the Office Open XML or .docx format) are now open formats. Obviously, non-open formats make it hard to recover data in the event that the original software developer is no longer able or willing to maintain the software. This should be reason enough for scholars to privilege open formats.

The aim of data formats is not just to ensure the proper storage and “memorisation” of data; usually, their structure is also geared towards ease of processing by algorithms specific to the data they represent. Thus, data for calculations usually takes a different form than data for text processing. Numerical and tabular data can, for instance, often be found stored as CSV (Comma Separated Values) files, of which an example is given in figure 3.4-4. Tabular formats are the bread and butter of spreadsheet programs and statistical computer languages such as R.

### 3.4.4 Structured and unstructured data

CSV data is what is known as structured data (data with a clear and predefined structure). Another such format is JSON (short for “JavaScript Object Notation”). JSON (see the example in fig. 3.4-5) is currently a rather popular format for storing structured (meta)data, though it should be pointed out that plenty of equally valid

alternatives exist. A plain text file, in contrast, is an example of unstructured data (that is to say, there is no clear predefined formal structure to prose text). XML, especially TEI XML, of which more below, represents something of a compromise between structured and unstructured formats. It attempts to capture an interpreted structure of a text in a formal manner. Each in its own way, these formats capture different aspects of data, sometimes even of the same data, and make it easily processable for specific processing by computer algorithms. In the text-oriented domain of stemmatology, one is likely to encounter certain structured and semi-structured formats more than others. The most common ones are detailed below.

```
[{
  "Name": "Alex",
  "Sex": "M",
  "Age": "41",
  "Height(in)": "74",
  "Weight(lbs)": "170"
}, {
  "Name": "Bert",
  "Sex": "M",
  "Age": "42",
  "Height(in)": "68",
  "Weight(lbs)": "166"
}, {
  "Name": "Carl",
  "Sex": "M",
  "Age": "32",
  "Height(in)": "70",
  "Weight(lbs)": "155"
}]
```

Fig. 3.4-5: Example of JSON data (reworked from Burkardt 2016).

#### 3.4.4.1 Plain text (.txt)

All the above suggests that the simplest possible open data format is a sensible choice for storing digital information. The simplest is known simply as plain text and is usually recognisable by the filename extension “.txt”. Basically, .txt files only store unstructured data as the series of bytes that represent characters. Complicating matters is the fact that there are several possible encodings for how these bytes should get translated into characters. One of the earliest encoding standards was ASCII (American Standard Code for Information Interchange), the first version of which defined 127 characters (Mackenzie 1980). Obviously, this is far too few to represent, for instance, a Chinese character system with over 50,000 different characters. Therefore, several encoding standards have developed over time, with the

Unicode standard ([unicode.org/standard/WhatIsUnicode.html](https://unicode.org/standard/WhatIsUnicode.html)) as the most comprehensive one, which even leaves room for a user's own encoding if needed. At present, the most widely used technical implementation of this standard is arguably UTF-8, which covers a large variety of character systems, including Latin, Greek, Cyrillic, historical scripts (runes, Ogham, polytonic Greek, and others), Asian scripts, mathematical symbols, and even emojis. Being sure of the long-term preservability of text thus means ensuring one uses an open and simple text format in the right encoding.

[illegible]

**Fig. 3.4-6:** Example of various scripts encoded in Unicode in a plain text file.

#### 3.4.4.2 Text as structured data

The most widely advocated data format in textual scholarship today is without doubt TEI XML, the XML grammar developed by the Text Encoding Initiative (see 3.3.2). One way of describing TEI XML is to say it is a particular dialect of XML, one that is specifically aimed at describing the structure of texts and documents that can be of interest to scholarship (prose, verse, stage-play scripts, historical documents, charters, editions, letters, and so on). TEI XML has a defined grammar that describes what elements (tags) can be used and in what combinations. This grammar is maintained by the TEI Consortium, which publishes the guidelines containing the description and explanation of all TEI tags and attributes on its website ([tei-c.org/index.xml](http://tei-c.org/index.xml)). A brief example and its visualisation in a browser are shown in figures 3.4-7-8. TEI XML covers an extensive set of types and genres of text, but not all textual eventualities are accounted for. This is why TEI XML also allows the extension of the standard in both informal and formal ways. If a certain textual phenomenon requires some description for which there is no element to be found in the TEI grammar, an editor or scholar can choose to add an arbitrary tag. The TEI grammar is a community-maintained project, which means that tags that are not represented yet can become part of the officially accepted guidelines. The TEI Consortium boasts a number of special interest groups (SIGs), of which the Special Interest Group on Manuscripts will probably be of most interests to scholars and scholarly editors working on traditions and stemmata ([tei-c.org/activities/sig/manuscript](http://tei-c.org/activities/sig/manuscript)).

```

<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="#"?>

<xsl:stylesheet version="1.0"
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  xmlns:TEI="http://www.tei-c.org/ns/1.0">

  <!-- TEI-XML Document start -->
  <TEI:data>
    <text>
      <body>
        <pb/>
        <cb/>
        <head>Van den Vos Reynaerde</head>
        <l>Willem die
          <app>
            <rdg wit="#Co">Vele boucke</rdg>
            <rdg wit="#Dy">Madocke</rdg>
          </app>
          maecte</l>
        <l>daer hi dicken omme waecte,</l>
        <l>hem vernoyde so haerde</l>
        <l>dat die avonture van Reynaerde</l>
        <l n="5">in Dietsche onghemaket bleven</l>
        <l>- die Arnout niet hevet vulscreven -</l>
      </body>
    </text>
  </TEI:data>
  <!-- TEI-XML Document end -->

  <!-- XSLT Stylesheet templates start -->
  <xsl:template match="/xsl:stylesheet">
    <xsl:apply-templates select="TEI:data/*" />
  </xsl:template>

  <xsl:template match="text/body">
    <html>
      <head>
        <title>Example of HTML resulting from TEI-XML with an XSLT stylesheet transformation</title>
      </head>
      <body>
        <div style="font-size: 180%; margin-bottom: 12pt;"><xsl:value-of select="head" /></div>
        <xsl:apply-templates select="l" />
      </body>
    </html>
  </xsl:template>

  <xsl:template match="l">
    <div style="width: 28pt; float: left;">&#160;<xsl:value-of select="@n" /></div>
    <div style=""><xsl:apply-templates/></div>
  </xsl:template>

  <xsl:template match="app">
    <xsl:apply-templates select="rdg[@wit='#Dy']"/>
  </xsl:template>

  <!-- XSLT Stylesheet templates end -->

</xsl:stylesheet>

```

**Fig. 3.4-7:** Example of a TEI XML document and an XSLT stylesheet combined in a single plain text file. TEI XML document above, stylesheet template specifications below.

## Van den Vos Reynaerde

Willem die Madocke maecte  
 daer hi dicken omme waecte,  
 hem vernoyde so haerde  
 dat die avonture van Reynaerde  
 5 in Dietsche onghemaket bleven  
 – die Arnout niet hevet vulscreven –

**Fig. 3.4-8:** The resulting HTML file viewed in a Web browser after it has performed the XSLT stylesheet template transformations on the TEI document.

Most scholarly editors will take a special interest in chapter 12 of the guidelines, which describes the elements and procedures that pertain to the encoding of a critical apparatus for a (digital) scholarly edition. The markup devised by the TEI community for structuring the apparatus is founded on the idea that an edition takes the classic form of a base text with apparatus. The document model of TEI is thus conventional, and its concept of editions should hold few surprises for scholarly editors. The apparatus is obviously the place to account for different readings in various witnesses. TEI XML offers three ways of linking items in the apparatus to the actual text: the location reference method, double end-point attachment, and parallel segmentation. Location reference and double end-point attachment have the most resemblance to the conventional use of footnotes in editions of texts. They describe variants in separate blocks of the document and link them to either a specific point in the text or a specific part of the text (location reference and double end-point attachment, respectively). Parallel segmentation is different in that variants are coded inline (i.e. at the point in the text where they occur) and that all variants are considered as variants of one another, although a preferred or primary reading can still be indicated. From the point of view of automated stemma construction, parallel segmentation is the preferable choice as variants are unambiguously marked up and easier to parse from the XML source.

Although TEI XML is mostly accepted as a de facto standard for digital scholarly edition data, it has not been without its critics. On the theoretical level, TEI as a document model, and especially the strong hierarchical view of text inherent in XML in general, have drawn criticism. XML demands that all elements be neatly nested in other elements. This means that it is rather awkward and quirky, for example, to mark up the fact that a paragraph spans a page break. Paragraphs and pages are two different dimensions of the same text that do not mix neatly according to the hierarchical formalisation that XML demands; you can express either of them very well in a separate hierarchy of nested XML elements, but you cannot combine the two without violating XML's nesting rules. Because text is in fact multidimensional (think only of the lexical, syntactic, and semantic levels, and how they overlap, and of how material and typographical aspects can be intertwined, and how narrative structures can function on several interconnected levels), many text theo-

```
ad impossibilia nemo tenetur
ad impossibilia *nemo* tenetur
ad impossibilia <emph>nemo</emph> tenetur
```

**Fig. 3.4-9:** A single text in three different formats: as simple plain text, with boldface or emphasis according to Markdown/AsciiDoc markup, and as TEI XML.

rists have drawn attention to the rather limited conception of text that TEI offers (see e.g. Buzzetti and McGann 2006; Huitfeldt 1995). On a more practical level, these theoretical problems express themselves as the “problem of overlap” for which XML does not really offer an adequate solution. Several solutions for overlapping structures have been proposed (DeRose 2004), but none seem to have solved these problems satisfactorily in a fundamental way. Several scholars have therefore raised the question of whether the strongly document-oriented models of XML and TEI XML should be supplanted by more versatile digital models (e.g. Haentjens Dekker and Birnbaum 2017; van Zundert and Andrews 2017).

What can be gauged from the above is that there is no data format that “does it all”. There is no format that will work for all purposes under all conditions, and choosing a proper format thus requires an understanding of the aims a scholar is pursuing and under what conditions he or she is pursuing them. A short description of various formats that a scholar may encounter while at work in the realm of digital text may serve to help facilitate such an understanding. What follows is categorised according to tasks scholars can reasonably expect to be executing in that realm, but note should be taken that a data format is rarely developed to address only one category. In practice, formats facilitate multiple tasks.

### 3.4.5 Transcribing and storing text digitally

One of the most basic tasks of any scholarly editor is transcribing the source (see 3.3.2). Even though this may look and feel a mundane and self-evident scholarly act, when a digital tool is used something quite exciting happens: for the first time, perhaps in centuries, the text is being made part of a new medium and becomes computer-processable. Scholarly editors may not always be as conscious as they could be about the fact that their digitally produced transcriptions are not only transcriptions but also imply a change of medium that creates new ways to use and experience the text (Karlsson and Malm 2004). As argued above, scholarship may be best served by keeping transcription formats as straightforward and open as possible. This favours the plain text format (.txt), as depicted in the first line of figure 3.4-9.

Plain text files can be created using editors purposefully tailored to that end, such as Notepad++ ([notepad-plus-plus.org](http://notepad-plus-plus.org)) for Windows or Atom ([atom.io](http://atom.io)) on the Mac. One could even go as low-level as the command line and use a Unix tool like VIM ([vim.org](http://vim.org)) or Emacs ([gnu.org/software/emacs](http://gnu.org/software/emacs)) to edit plain text files, though most editors will probably prefer some form of graphical interface. Many varieties of text editors exist, often also as open source and free alternatives to commercial tools. It does not really matter what editor is used for creating plain text files, as long as the editor checks that the actual file is a text file encoded in UTF-8. Most current text editors already store plain text files natively in that encoding. Many programs boast useful search and replace functions, scripting, and so forth.

One step up from plain text, a scholarly editor enters the domain of markup formats. As detailed above, XML is the most widely adopted form of markup in the scholarly editing landscape, but many forms of markup exist. For the initial transcription of the text, it may be useful to prefer a more lightweight markup language such as Markdown (Gruber 2004) or AsciiDoc ([asciidoc.org](http://asciidoc.org)). Such lightweight markup languages try to minimise the invasiveness of the markup, which may be less distracting when focusing on transcription: contrast lines 2 and 3 in figure 3.4-9, for example.

Other formats that may be considered when transcribing a text would be those known from word processing programs. Formats such as .doc (Microsoft Word document file), .docx (Office Open document format), .odt (Open Document Text, used e.g. by LibreOffice and OpenOffice), .rtf (Rich Text Format), and so on. Although often used, scholars and developers that have worked with digital text data mostly consider employing programs such as Microsoft Word, OpenOffice, Pages, Scrivener, and so on bad practice. Such word processors do a lot to facilitate the work of the writer, but they store all information in formats that are cumbersome to read and error-prone in processing by computer. From a point of view of information integrity, these programs are best avoided.

At some point, a scholarly editor will probably want to express or describe more of the structure of the text. This is the point where straightforward text will not suffice and some formalism will be needed to differentiate between source text and metadata. Those who want to rely on a digital formalisation that can count on some proven continuity may well opt for TEI XML. The Oxygen XML Editor, although not open and not free, is currently the go-to editor for many ([oxygenxml.com/xml\\_editor.html](http://oxygenxml.com/xml_editor.html)). Some other text editing programs do have support for XML authoring and validating (e.g. Atom). The Text Encoding Consortium offers helpful information on how adjusting grammars and validation works in the case of TEI XML ([tei-c.org/Support/Learn](http://tei-c.org/Support/Learn)), and there are plenty of spaces on the Web that offer introductions to XML in general (e.g. [w3schools.com/xml/default.asp](http://w3schools.com/xml/default.asp)).

As has been shown above, many alternatives to XML exist. It is doubtful if XML will be around for many more decades. It is more likely that it will be supplanted by some other markup language or an altogether different technology in a number



of years. Much is expected from Semantic Web technologies such as RDF (w3.org/RDF) and HTML-RDfA (w3.org/TR/rdfa-lite, 2nd ed. 2015) as successors to XML. Although the latter are themselves XML-based, a transcription in RDF would look rather different from one in “plain” XML. It is unclear, however, how successful RDF and related technologies will be. XML has long been seen as the ultimate data exchange format, but JSON and other standards have found considerable success as well. Some scholars and technologists are looking into standoff markup, which is a technology that keeps the markup separated from the transcription text; by doing so, many problems associated with XML are resolved (Haentjens Dekker and Birnbaum 2017; Spadini, Turska, and Broughton 2015). These technologies, promising as they are, are still very immature. What is important to remember, however, is that if the formalisation used is open, strict, and consistent, all formats can be transformed into other formats. This may require particular software, but if a very successful new format should emerge, such “porting” software is very likely to come into existence as well.

TEI XML is not lightweight; it is rather aimed at very full and very detailed textual criticism. To illustrate this point, let us look at the very minimum that is required to encode the example from figure 3.4-9 in full so that it is a valid TEI XML document (see fig. 3.4-10). Normally, in the description there would also be detailed bibliographical descriptions of witnesses and other sources (in the header), which have been omitted in the case of this one-line example. Obviously, the length of the header is a bit ridiculous for a one-sentence example. Most of an XML document is made up of metadata that will not be repeated, and when transcribing 13,000 lines rather than 1, the ratio between metadata overhead and text becomes rather more reasonable.

The salient point is that, due to its verbosity, TEI (and XML in general) may not be the most convenient format when transcription work and data for stemmatological analysis are still in a very volatile state. Although opinions among the technically informed may differ, TEI XML might best be regarded as a final format for digital publication, but transcriptions and data for stemmatological analysis may best be kept in more lightweight formats, such as plain text with a touch of idiosyncratic markup. Once all scholarly preparation is done, the text and its metadata can then be gathered and formed into a TEI XML online publication.

It may be less obvious that tabular data in CSV format may be of use to scholars and philologists as well. However, if fine-grained manual control and overview of word-based alignment, variant detection, and annotating is of the utmost importance, then the view of CSV data offered by spreadsheet programs may be very helpful (see fig. 3.4-11). Spreadsheets safeguard openness and reasonable longevity of the data as long as files are stored in CSV format. Another advantage is that most analytical software packages are very well equipped to use CSV files as a data source.

The (collaborative) work of transcribing and annotating can easily involve maintaining multiple versions of a text in multiple stages of editing. It may be easy

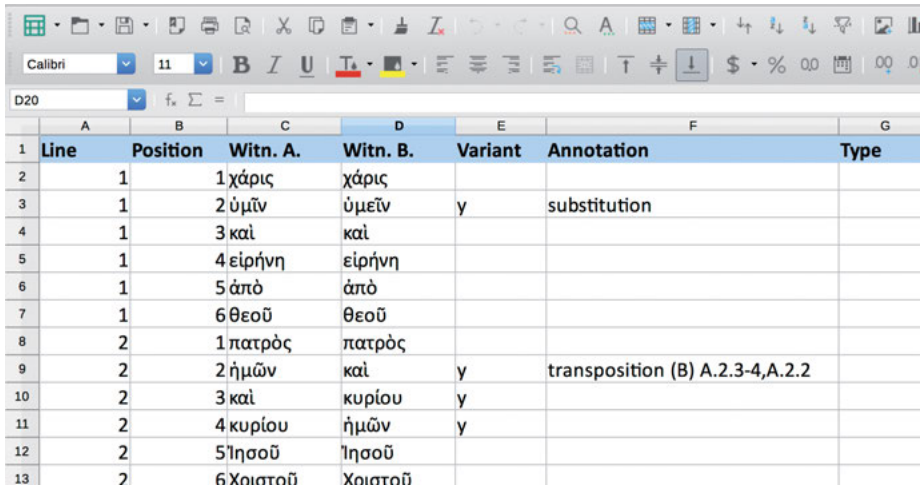
```

<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Example of TEI-XML file</title>
        <author>Joris J. van Zundert</author>
      </titleStmt>
      <editionStmt>
        <edition>
          <date>Thursday 8 March 2018</date>
        </edition>
      </editionStmt>
      <publicationStmt>
        <distributor>The Huygens Institute for the History of the Netherlands</distributor>
        <address>
          <name type="institution">The Huygens Institute for the History of the Netherlands</name>
          <street>Oudezijds Achterburgwal 185</street>
          <postCode>1012 DK</postCode>
          <name type="city">Amsterdam</name>
          <name type="country">The Netherlands</name>
        </address>
        <availability>
          <p>CC BY-SA 3.0
            <ref target="https://creativecommons.org/licenses/by-sa/3.0/deed.en">
              (https://creativecommons.org/licenses/by-sa/3.0/deed.en)
            </ref>
          </p>
        </availability>
      </publicationStmt>
      <sourceDesc>
        <listWit>
          </listWit>
        </sourceDesc>
      </fileDesc>
      <encodingDesc>
        <variantEncoding method="parallel-segmentation" location="internal"/>
      </encodingDesc>
      <revisionDesc>
        <change when="2018-03-08" who="Joris J van Zundert">Transcribed</change>
      </revisionDesc>
    </teiHeader>
    <text>
      <body>
        <p>ad impossibilia <emph>nemo</emph> tenetur</p>
      </body>
    </text>
  </TEI>

```

**Fig. 3.4-10:** The minimal TEI XML structure needed to describe our one-line example text.

to lose track of the status of transcription work, and of all the places that need to be checked, compared, and commented on. Apart from this, there is not enough commiseration in this world for all the times that months if not years of work have gone missing because of a careless click on a “delete” button, a failed hard drive, or a stolen laptop. This is therefore the place to point out the importance as well as convenience of backups and version management. When working with many (versions of) files, a convenient way of storing files safely is to use a code repository such as GitHub (github.com) or Sourceforge (sourceforge.net). Apart from guaran-



	A	B	C	D	E	F	G
1	Line	Position	Witn. A.	Witn. B.	Variant	Annotation	Type
2		1	1 χάρις	χάρις			
3		1	2 ὑμῖν	ὕμειν	γ	substitution	
4		1	3 καὶ	καὶ			
5		1	4 εἰρήνη	εἰρήνη			
6		1	5 ἀπὸ	ἀπὸ			
7		1	6 θεοῦ	θεοῦ			
8		2	1 πατρός	πατρός			
9		2	2 ἡμῶν	καὶ	γ	transposition (B) A.2.3-4,A.2.2	
10		2	3 καὶ	κυρίου	γ		
11		2	4 κυρίου	ἡμῶν	γ		
12		2	5 Ἰησοῦ	Ἰησοῦ			
13		2	6 Χριστοῦ	Χριστοῦ			

**Fig. 3.4-11:** Using LibreOffice to create transcriptions, alignment, and metadata as CSV files.

teeing a safe external backup location, such repositories tirelessly keep track of all changes and versions that happen to emerge during an edition project. If privacy is of importance, private repositories can be created online, or the software can be installed locally (e.g. in the digital infrastructure of an institution). Working with version management software and repositories does require an additional learning curve, but it protects the scholar from the dreaded prospect of errors that collating various versions by hand can present. The *Programming Historian* (van Strien 2016) has a helpful introduction to version management with GitHub, a must-read for any scholar wanting to avoid the mess of version confusion and lost files.

It is very possible for a scholarly edition project to end up with many files and many versions of files, certainly if one is working on larger traditions with many witnesses. In that case, it may become useful to keep track of all versions and text files with a database (e.g. mysql.com). A relational database can be used to keep track of the metadata of documents, files, texts, traditions, and so on, and how they relate to one another. How a database is designed is not covered here, as many useful resources exist. Stephen Ramsay (2004) offers a good introduction oriented towards humanities scholars and their work.

### 3.4.6 Representing variants

The transcription of different witnesses of the same work will generally lead to the identification of variant readings. In the above, we have already met some ways of representing these variants. When working with TEI XML files as a format for representing the text and apparatus, software engineers and scholars who need to

parse text files to aggregate variants in order to compute a stemma will generally prefer parallel segmentation because of its unambiguous identification (oversights on the part of the editor aside) of variant readings. See figure 3.4-7 for an example of how parallel segmentation is used to express variant readings. However, it may very well be that a TEI XML file is not considered a suitable form of input for a stemmatic analysis. Automated stemmatic analysis requires only knowledge of the variant readings and does not (usually) take into account any other text. In fact, if one looks at an input file for PAUP\* (one of the most-used programs for inferring phylogenetic trees), one might wonder if the data is related to text at all (see fig. 3.4-12).

Let us walk through the example in figure 3.4-12 to get a feel for the information that phylogenetic algorithms and software generally require, and to understand how this sort of file can be representative of texts and variant readings at all. First of all, all NEXUS files are simply plain text files. Although NEXUS files often have the file extension “.nex”, PAUP\* reads any file put to it as a plain text file and will error out if it does not recognise the formalities it expects. Internally, each NEXUS file starts with the marker “#nexus”, which primarily just testifies to the human reader that it is meant to be a NEXUS file. Any information and comments that the author of the file wants to supply can be put anywhere in the file between square brackets, as shown in figure 3.4-12. PAUP\* simply ignores anything between such brackets. All information really relevant to phylogenetic analysis and trees in NEXUS files is put in “blocks” that start with the word “start” and end with the word “end”. The word after “start” identifies the type of block, that is, the type of information that is contained inside the block. Usually one will find here the definition of a matrix, specifying the number of rows (taxa, “ntax”) and the number of columns (“nchar”). Thus, the line starting with “dimensions” just describes the length and width of the matrix that one finds further on after the command “matrix” and its closing semicolon. The next line describes how the information in the matrix is formatted. Many possibilities exist, but in the case of computing stemmata one usually only encounters a format described as “symbols=“01””, which denotes that the values in the matrix should only be 1 or 0. In this case, the values for “missing”

```
#nexus
[Variant information (20160314): Maerlant Rhyme Bible, Book of Judith, Mss. Br1, Dh1, Dh2, Le1, Lol.]
begin data;
  dimensions ntax=6 nchar=54;
  format symbols="01" missing=? gap=-;
  matrix
    Br1 110001101010101110001010001100111100010010000110100010
    Br2 11000110101010111010101000110011-000010010000110100010
    Dh1 11000110101010101000101000110011-000010010000110100010
    Dh2 110001101010101010001010001111110100011010000110100010
    Le1 110001101010101110001010001100110100011010000110100010
    Lol 100001101010101010001010001111110100011010000110100010
  ;
end;
```

Fig. 3.4-12: Example of a NEXUS file used as input for PAUP\*.

ms. A	ms. B	ms. C	ms. D	ms. F	ms. G	ms. E
Echt	Echt	Echt	Echt	Echt	echt	
daden	dadensi	daden	daden	deden	deden	
si		si	si	si	zi	
na	na	na	na	na	na	Doe
sanghers	sangers	SANGHERS	sangers	sanghers	sanghers	sanger
						die
						rechtre
						was
doot	doet	dod.	doet	doot	doet	doot
						Daden
						si
Dien	Die	Die	Dien	Die	die	dien
van	van	van	van	van	van	van
israel	israhel	YSRAHEL	israel	ysrahel	israhel	ysrahel
mesdaet	anxt	sonde	anxt	sonde	zonde	anxt
groot	groet	grod.	groet	groot	groet	groot

**Fig. 3.4-13:** Seven witnesses for a biblical verse in a Middle Dutch translation of Petrus Comestor’s *Historia scholastica*.

and “gap” indicate that the symbols “?” and “–” may also occur in the matrix, identifying places where it is not known whether a 1 or 0 occurred (“?”) or if there is just no value in that place in a certain row (“–”).

But how do these rows of 0s and 1s represent witnesses? The truth is: they do not. They represent only very reduced information about variant readings in witnesses. To understand this, we need to look at how the matrix information can be derived from a representation of actual text. For this we need to take a look (again) at how different witnesses may be aligned using a spreadsheet or CSV file. Figure 3.4-13 lists a verse from the biblical tale of Deborah according to seven different witnesses from a Middle Dutch rhymed translation of Petrus Commestor’s *Historia scholastica* (van Maerlant, 1858). The verse reads “Echt daden si na sanghers doot / Dien van israel mesdaet groot” [But after Shamgar’s death the Israelites sinned greatly].

The tabulator format in fig. 3.4-13 is known as an alignment table, where all identical words of all witnesses are lined up in the same row (or column, if the text of witnesses is put in individual rows instead of, as in this case, columns). Let us consider first the witnesses that seem less problematic, those that are depicted in columns 1–6 (A, B, C, D, F, G). We can encode the variant readings of these witnesses by listing each different reading we find in another table, and we can then indicate column-wise whether a certain witness has that reading (1) or not (0). The results obtained by doing so for the table in figure 3.4-13 are depicted in figure 3.4-14. Manuscript A has the reading “daden si”, as one can gauge from the first column in figure 3.4-14; therefore, in the row for that particular reading in the table in figure 3.4-15,

reading	ms. A	ms. B	ms. C	ms. D	ms. F	ms. G	ms. E
daden si	1	1	1	1	0	0	–
deden si	0	0	0	0	1	1	–
doe daden si	0	0	0	0	0	0	1
d[e a]den si na	1	1	1	1	1	1	0
dien	1	0	0	1	0	0	1
die	0	1	1	0	1	1	0
mesdaet	1	0	0	0	0	0	0
anxt	0	1	0	1	0	0	1
sonde	0	0	1	0	1	1	0

Fig. 3.4-14: Encoding of readings for the same six witnesses as those in figure 3.4-13.

we find “1” in the column for that manuscript. Manuscript *F*, however, has a different reading, “deden si” (a linguistic variant of the Middle Dutch past tense of “to do”), so we find “0” in the relevant column in the same row. The reverse situation is found in the following row, which encodes whether a manuscript has the reading “deden si” or not.

As the reader will probably have noticed, there is more variation between the witnesses than there is encoded in the readings table of figure 3.4-14. This is a result of the particular choices an editor makes about which variants are to be encoded and which are not. In this instance, the editor decided that the spelling difference between “Echt”, “EChT”, and “echt” was not genealogically relevant, and thus that particular variant was not encoded. There is no absolute consensus between scholars on what type of variants reveal genealogical relationships and what types do not. Some hold that spelling variants are never interesting, some point out that spelling variation in vernacular manuscripts may well indicate geographical and, following from that, genealogical proximity. For a more thorough discussion of this issue, see sections 4.3 and 6.2.

Variant readings can become pretty complex. Let us now consider witness manuscript *E* in the last column of figure 3.4-13. In this case, witness *E* is still unmistakably a cousin of the original translation of Petrus Comestor’s *Historia scholastica*, but because it was done from memory or by a very liberal copyist, witness *E* contains all kinds of wonderfully exotic readings. First of all, this makes alignment at various points a matter of debate and interpretation (in the example above, for instance, should “doot” be aligned with “doot”, or rather “Daden si” with “daden si”? – one cannot have it both ways). Such variation also makes it a matter of interpretation how variants should be encoded. As can be inferred from figure 3.4-14, in this case the editor decided to treat the complicated variant as a transposition of “Doe [...] Daden si” [when [...] did they] in *E* with regard to the readings in the other manuscripts (“daden si na [...]” [did they after [...]]). Again, once it has been decided

```
#nexus
[Variant information (20160315): Maerlant Rhyme Bible, Judges: Deborah, mss. A, B, C, D, E, F, and G.]
begin data;
  dimensions ntax=7 nchar=54;
  format symbols="01" missing=? gap=-;
  matrix
    ms.A 100110100
    ms.B 100101010
    ms.C 100101001
    ms.D 100110010
    ms.F 010101001
    ms.G 010101001
    ms.E --1010010
  ;
end;
```

**Fig. 3.4-15:** Example NEXUS file encoding for the variant readings from figure 3.4-14.

what the possible readings are, it merely remains to note down which manuscripts have which reading. Because *E* has neither the reading “daden si” nor “deden si”, we find the gap indicator (“-”) here.

The table that results from this process can subsequently be made into a NEXUS file by transposing it. A matrix transposition results in a new matrix whose rows are the columns of the original. After this operation, we end up with the representation in figure 3.4-15.

### 3.4.7 Representing alignment

We have already spoken (3.4.6) about aligning witnesses, which is the task of trying to line up the individual matching words in various witnesses. This work allows an editor to meticulously compare and examine variant readings. There are various ways in which this aligning can be achieved, and there is no “best” one because what works best is very much dependent on the context and preferences of the editor. There are, however, helpful tools and formats that facilitate the job. We have already seen above how CSV files, as simple text files, might be useful as a storage format for aligned texts. CSV files guarantee, most of the time, flawless processing and interchange between programs and tools. Typing and working in a CSV file via a plain text editor may turn out to be cumbersome, and most users will probably prefer to use a spreadsheet program that will provide convenient navigation, overviews, search and replace functions, and so forth. However, it should be noted that typographical information is never stored in CSV files, so it is not advisable to use elaborate typography or colour coding to encode any essential information about witnesses, because such information will evaporate when the content of a spreadsheet is stored as a CSV file. Here too, “less is more” counts.

Apart from CSV and spreadsheets, another possible way of denoting alignment is by way of a variant graph (see 3.3). This is a fairly new way of representing align-

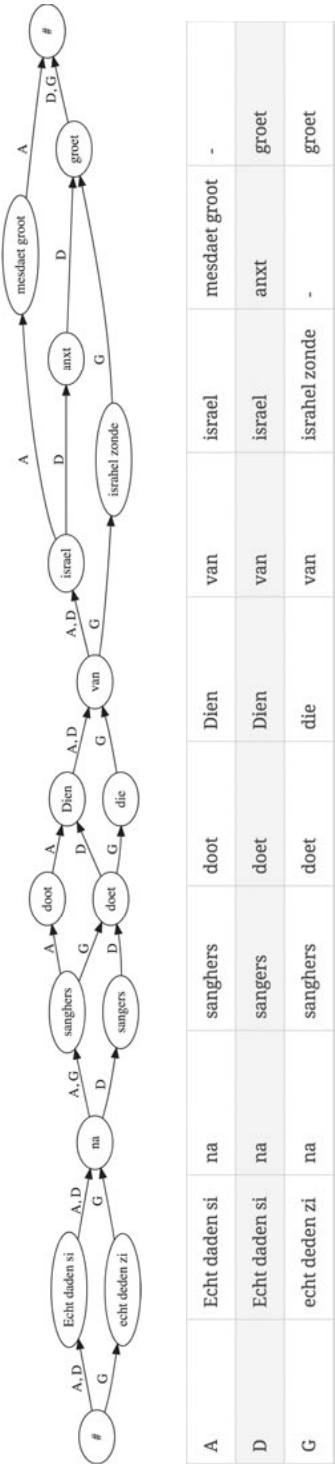


Fig. 3.4-16: A variant graph (top), reducing redundancy in an alignment table (bottom).



ment. An example is given in figure 3.4-16. Essentially, a variant graph collapses all redundant information from an alignment table or spreadsheet. If the columns of a table or spreadsheet represent the linear positions in witness texts, then wherever the same value (reading) is found in cells in the same column, those cells can be collapsed into a node of the graph.

### 3.4.8 Representing trees, networks, and graphs

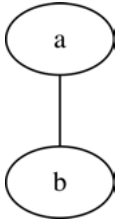
The variant graph takes us into the realm of the representation of relational data (also called linked data or networked data). CSV files are very good at storing factual tabular data, but they do not store any information about the relations between values in particular rows, columns, or cells in a table or spreadsheet. Let us return to the example in figure 3.4-13. Some may argue that there is relational information in that table because it clearly depicts that manuscript *A* contains the reading “dien”. However, this relation is not formally expressed at all in the CSV file. We can assume that the relation exists because we follow a convention for how rows and columns relate in a table and what their headers may mean, but all that information is in fact assumed by the reader/user; it is not formally noted in the CSV file. Thus, if we need to explicitly describe and sustain such relational information, we need to put more information into the file. Again, many plain text-based file formats, such as XML or JSON, would allow us to do this. For XML and JSON, some illustrations of expressing structures in and relations between data can be found in the examples above. Here, we limit ourselves to examples of file formats customarily used to describe trees, networks, and graphs, as these are the ones most often used to store variant graphs and stemmata of traditions (and possibly, related to them, correspondence networks and social networks, trees or networks of provenance, and so on).

A very basic yet powerful language for expressing graphs is DOT. Graphs consist of nodes connected by edges. Arguably the simplest graph consists, therefore, of two nodes and one edge between them. In the DOT language, this would be described as depicted in figure 3.4-17. This description corresponds to the graph in figure 3.4-18.

The graph in figure 3.4-18 is an undirected graph, which means the edges have no direction (in this case, node *A* does not “point to” node *B*, they “just relate”). However, edges can also have a direction, and both edges and nodes can have attributes,

```
graph SimplestGraph {
    a -- b;
}
```

**Fig. 3.4-17:** Description of a rudimentary graph in DOT.



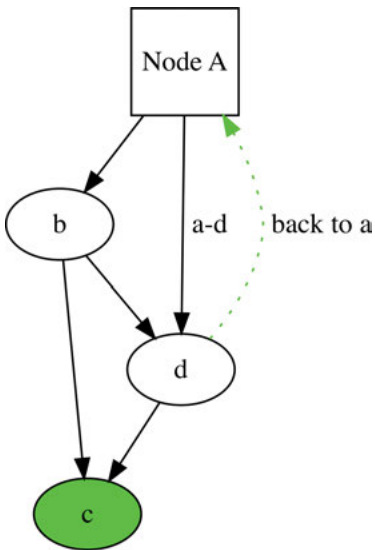
**Fig. 3.4-18:** The simplest graph.

```

digraph MoreComplexGraph {
    a [label="Node A",shape=box,width=0.75,height=0.75,fixedsize=true]
    c [fillcolor=green style=filled]
    a -> b;
    b -> c;
    b -> d;
    a -> d [label=" a-d  " ];
    d -> a [label=" back to a " color=green style=dotted];
    d -> c;
}

```

**Fig. 3.4-19:** A more complex graph description in DOT.



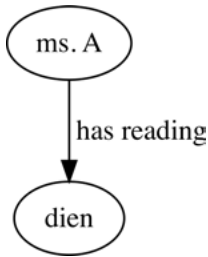
**Fig. 3.4-20:** The graph resulting from the formal description in figure 3.4-19.

```

digraph ReadingRelation {
    a [label="ms. A"]
    b [label="dien"]
    a -> b [label=" has reading"];
}

```

**Fig. 3.4-21:** Graph description in DOT of the relation between manuscript *A* and the reading “dien”.



**Fig. 3.4-22:** Graph resulting from the description in figure 3.4-21.

which can all be described as text in the DOT language, leading to elaborate graphs. In figure 3.4-19, the DOT description of the graph in figure 3.4-20 is given. Note the notation for a directed graph (“digraph”) and directed edges (e.g. “ $b \rightarrow c$ ”), and that properties of nodes and edges are added between square brackets. Using a relatively uncomplicated textual description, the DOT language thus offers a versatile way of describing very complex networks, graphs, and phylogenetic trees. A concise reference for the language is offered by Hayes-Sheen (2017); more comprehensive documentation is given by John Ellson et al. (graphviz.org).

Using DOT, we can formally capture the relation between pieces of data. Suppose we wanted to formally and explicitly express the relation between manuscript *A* and the reading “dien” in the table in figure 13. We could do so as in figure 3.4-21 (but note that there are also many other ways the relations could be expressed), with the resulting graph in figure 3.4-22.

The information in the table in figure 3.4-13 (excluding manuscript *E*) could then be captured in a DOT file as depicted in figures 3.4-23–24.

Of course, there is no reason why exactly the same information could not be expressed in other file formats. A graph can be described perfectly well in JSON, or in XML for that matter. In fact, there is a somewhat limited XML dialect especially geared towards describing graphs (graphml.graphdrawing.org/index.html). The benefit of using DOT is that it can be read by one of the most popular open source visualisation tools for graphs and networks, GraphViz (see graphviz.org). Downloading and installing GraphViz allows one to execute commands such as “dot -Tpng -Gdpi=600 graph.dot -o graph.png”, which means that the program will take as input the file “graph.dot” (e.g. a text file like that in fig. 3.4-23), translate it into

```

digraph ReadingRelation {
    A [label="ms. A"]
    B [label="ms. B"]
    C [label="ms. C"]
    D [label="ms. D"]
    F [label="ms. F"]
    G [label="ms. G"]
    a [label="daden si"]
    b [label="deden si"]
    c [label="dien"]
    d [label="die"]
    e [label="mesdaet"]
    f [label="anxt"]
    g [label="sonde"]
    A -> a [relation="has reading"]
    A -> c [relation="has reading"]
    A -> e [relation="has reading"]
    B -> a [relation="has reading"]
    B -> d [relation="has reading"]
    B -> f [relation="has reading"]
    C -> a [relation="has reading"]
    C -> d [relation="has reading"]
    C -> g [relation="has reading"]
    D -> a [relation="has reading"]
    D -> c [relation="has reading"]
    D -> f [relation="has reading"]
    F -> b [relation="has reading"]
    F -> d [relation="has reading"]
    F -> g [relation="has reading"]
    G -> b [relation="has reading"]
    G -> d [relation="has reading"]
    G -> g [relation="has reading"]
}

```

**Fig. 3.4-23:** Full graph description in DOT of the information in figure 3.4-13.

the PNG graphics format with a resolution of 600 dots per inch, and store the resulting picture in a file named “graph.png”.

It is likely that, for practical purposes, editors will prefer the CSV format for capturing variant readings, because writing a DOT or JSON file involves more and quite tedious work. The salient point in showing the capabilities of these formats is to demonstrate the various levels and differences of formal explicitness that can be

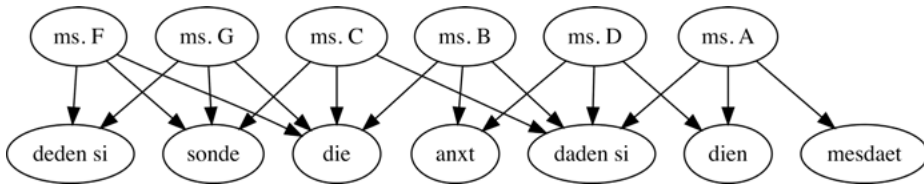


Fig. 3.4-24: Graph resulting from the description in figure 3.4-23.

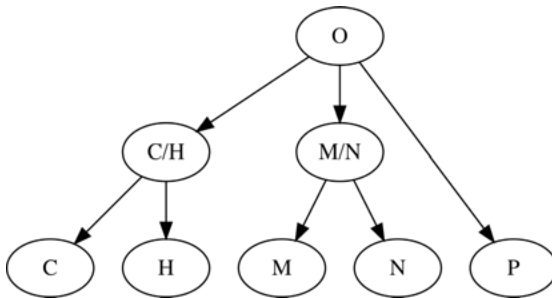


Fig. 3.4-25: Stemma of sources of the Middle Dutch *Reis van Sente Brandane*, adapted from Strijbosch (1995, 21).

achieved with them. However, one should bear in mind that a lot of the dullness of writing JSON, XML, or DOT files can be overcome by automating the boilerplate parts of the formatting.

When working with stemmata, it will be useful to express them in a machine-readable way; this is obviously a case where the application of a format such as DOT is warranted. For instance, a stemma inferred through conventional non-computational means for the sources of the Middle Dutch translation of the *Voyage of Saint Brendan* (in Middle Dutch, the *Reis van Sente Brandane*) could be captured as depicted in figure 3.4-25. The DOT description for this stemma is given in figure 3.4-26.

NEXUS files also support describing phylogenetic trees through the Newick format, which uses the nested parentheses approach to indicate branches ([scikit-bio.org/docs/0.2.2/generated/skbio.io.newick.html](http://scikit-bio.org/docs/0.2.2/generated/skbio.io.newick.html)). Those scholars with a linguistics background may be familiar with this format as it is similar to the labelled bracketing method of describing the linguistic tree structure assumed in sentences (see Kerstens, Ruys, and Zwarts 1996). The Newick expression describing the same stemma as that in figure 3.4-25 would be  $((C,H)C/H,(M,N)M/N,P)O$ . The NEXUS file format (see also above) just embeds this in some notation to indicate that it is indeed a tree (see fig. 3.4-27).

Another XML variant for describing cladistic trees that has been adopted by several popular phylogenetic tools and software libraries is PhyloXML (Han and Zmasek 2009; [phyloxml.org](http://phyloxml.org)). A file depicting the same tree as that given in figure 3.4-25 would look like figure 3.4-28.

```

diGraph StemmaBrandane {
    A [label="O"]
    B [label="C/H"]
    C [label="M/N"]
    D [label="C"]
    E [label="H"]
    F [label="M"]
    G [label="N"]
    H [label="P"]
    A -> B
    A -> C
    A -> H
    B -> D
    B -> E
    C -> F
    C -> G
    { rank="same"; D, E, F, G, H }
}

```

**Fig. 3.4-26:** DOT description of the stemma in figure 3.4-25.

```

#nexus
begin trees;
    Tree StemmaBrandane = ((C,H)C/H,(M,N)M/N,P)O;
end;

```

**Fig. 3.4-27:** NEXUS description of the stemma in figure 3.4-25.

The list of possible plain text-based formal description formats for graphs, trees, and stemmata is potentially much longer than the few formats shown here. The salient point to keep in mind, however, is that they are all able to express the same basic information about tree and network structures. Some support many custom attributes and visualisation properties (e.g. DOT), others solely capture the basic tree structure (e.g. Newick). One way or another, it is therefore possible to transform one format into another. Sometimes, tools for this even exist (see e.g. [graphviz.org](http://graphviz.org), s. v. “[graphml2gv](http://graphml2gv)”), though the reader/user should be warned that support for migration can be dodgy or even completely lacking. Care should be taken, when migrating a graph description from one format to another, that all the properties available in one format also exist in another. If this is not the case, some information will get lost. A transformation of a file from DOT to Newick, for instance, would clearly not be without loss of information.

```

<?xml version="1.0" encoding="UTF-8"?>
<phy:Phyloxml xmlns:phy="http://www.phyloxml.org/1.10/phyloxml.xsd">
  <phy:phylogeny>
    <phy:name>StemmaBrandane</phy:name>
    <phy:clade>
      <phy:name>O</phy:name>
      <phy:branch_length>0</phy:branch_length>
      <phy:clade>
        <phy:name>P</phy:name>
        <phy:branch_length>10</phy:branch_length>
      </phy:clade>
    </phy:clade>
    <phy:clade>
      <phy:name>M/N</phy:name>
      <phy:branch_length>5</phy:branch_length>
      <phy:clade>
        <phy:name>N</phy:name>
        <phy:branch_length>5</phy:branch_length>
      </phy:clade>
      <phy:clade>
        <phy:name>M</phy:name>
        <phy:branch_length>5</phy:branch_length>
      </phy:clade>
    </phy:clade>
    <phy:clade>
      <phy:name>C/H</phy:name>
      <phy:branch_length>5</phy:branch_length>
      <phy:clade>
        <phy:name>H</phy:name>
        <phy:branch_length>5</phy:branch_length>
      </phy:clade>
      <phy:clade>
        <phy:name>C</phy:name>
        <phy:branch_length>5</phy:branch_length>
      </phy:clade>
    </phy:clade>
  </phy:phylogeny>
</phy:Phyloxml>

```

Fig. 3.4-28: PhyloXML description of the stemma in figure 3.4-25.

### 3.4.9 Representing the edition

Stemmatology is usually undertaken as a subtask during research that should eventually lead to the publication of a scholarly edition or a work synthesising a certain

textual tradition. Increasingly, we see such works also being published as digital works. In such cases, the issue of digital format obviously also applies to the eventual publication itself. More importantly, the digital medium allows us to model and produce anything we can think up as long as it can be depicted on a screen. With this re-mediation, therefore, comes also a potential renegotiation of the digital scholarly edition and its related scholarly processes (see Bolter and Grusin 2000 on the topic of re-mediation). What constitutes an adequate digital scholarly edition is a much-debated issue (for more on this, see 6.3). Different scholars have arrived at different conclusions on this matter. Pierazzo (2015), for instance, seems to conclude that a digital scholarly edition should be a digitised form of an edition that is in all respects created the same as a printed edition, but with specific digital means and technologies. According to Bordalejo (2018), too, digital scholarship has in no sense changed the goals and methods of the scholarly editor. These attitudes could be called “mimetic”, “conventional”, or “conservative”. Others hold that a medium shift also necessarily involves in part rethinking and reshaping the object that flows from one medium to another and is thus re-mediated. Sahle (2013) argues, for instance, that a digital scholarly edition is defined precisely by being inalienably digital. Thus, for Sahle, a digital scholarly edition is defined mostly by those aspects that would be lost if the edition were published as printed edition. More radical perspectives are offered by, for example, van Zundert and Andrews (2017), who argue that digital texts should indeed be regarded first and foremost as digital objects. A digital scholarly edition could thus, for instance, be equivalent to a database or graph model representing the text rather than to the visual, derived representation of them in a graphical user interface.

Databases and graphs, regarded as versatile modelling tools for text, certainly enrich the ability of scholars to express the multiple dimensions of a text and the different perspectives on it. Databases and graphs, understood in this way, do not merely provide a container for a collection of flat transcriptions and the additional information needed to create a critical edition. Beyond that, they allow the editor to augment such material with various interpretations, perspectives, and additional digital (or digitised) objects. Together with software to query and present the material thus stored in databases and graphs, these technologies allow for an unparalleled richness in representing editions. Not just the edition authorised by the editor can be visualised, but also other critical interpretations, as well as their constituent material. Specific dimensions or aspects that are of interest to a certain user or reader can be dynamically inferred and presented (e.g. social relations between persons or characters, a chronology of events, a histogram of topics related in a text, and so on). Creating such advanced digital scholarly editions requires assiduous effort by both scholar and programmer, for software and program code do not redefine the scholarly edition by themselves, nor do they create scholarly editions automatically – this all, obviously, remains human scholarly work. But, with the new possibilities they open up, digital scholarly editions do invite us to rethink what a scholarly



edition could or should be. The volatility of the debate surrounding the digital scholarly edition may give scholars wanting to produce an edition cause to seek orientation about the various text-philosophical approaches towards digital textual scholarship. This specific debate is not covered in the present handbook, but good entry points into it may be Thaller (2004), Deegan and Sutherland (2008), Driscoll and Pierazzo (2016), Andrews (2013), Fischer (2013), and Robinson (2013a). Here, we limit ourselves to considering some of the formats that may be encountered when a scholar wishes to represent the results of editorial work (including stemmatology) digitally.

In a sense, any JSON, TEI XML, or DOT file can be made to contain the bare essential information that describes the results of scholarly work. Thus, when they represent all the information pertaining to a digital scholarly edition, such files can be said to represent that edition. The base data that CSV, JSON, XML, and other types of files contain is, however, usually only regarded as storage data. Additional processes are applied to derive visual representations of that data that cater to the user reading or viewing the edition. The data files, for instance, may be stored on the hard drive of a Web server where custom-made or out-of-the box Web software transforms it on request into files that are viewable by Web browsers. In other cases, formats will be derived that port the base data into file formats that are accessible via a tablet or e-reader. The most common file types for this visualisation are listed in what follows.

#### 3.4.9.1 HTML

HTML is the first language that was used to create Web pages. It is a markup language that allows one to indicate how specific text should be visualised and how documents are linked. A comprehensive overview of its history and technology is offered by Shannon (2019). The Web itself offers many helpful introductions to crafting HTML and integrating related technologies for Web publishing (e.g. [htmlprimer.com/htmlprimer/html-beginners](http://htmlprimer.com/htmlprimer/html-beginners); [w3schools.com/html](http://w3schools.com/html)). An especially gentle entry-level introduction to HTML and related technologies for Web publishing is offered by Robert Mening (2018). HTML works in exactly the same way as XML and is thus again plain text interspersed with markup codes between angle brackets. In contrast to XML, HTML is concerned with layout and typography rather than with structure or content. HTML has meanwhile progressed to a fifth version (HTML5) that integrates support for many other types of output than text (audio, video, pictures, screen readers). HTML is still the basic fabric of most Web pages, but it is today often combined with many other technologies to provide elaborate styling (CSS, or Cascading Stylesheets; see Mills et al. n.d.), interaction and dynamic presentations (JavaScript; see [javascript.info/intro](http://javascript.info/intro)), scalable graphics (for a comprehensive overview, see [inkscape.org/develop/about-svg](http://inkscape.org/develop/about-svg)), fonts (e.g. Web fonts; see [developer.mozilla.org/en-US/docs/Learn/CSS/Styling\\_text/Web\\_fonts](http://developer.mozilla.org/en-US/docs/Learn/CSS/Styling_text/Web_fonts)), and so forth. The drawback of using (“stacking”) many such technologies on top of base-data files to

produce nice-looking visualisations is that the combined data and software that produce the digital edition may become hard to maintain and sustain over time. A responsible scholarly editor will therefore always make sure that the base data is also archived for perpetuity in some specialised institution, such as a digital library or an institutional data repository.

### 3.4.9.2 PDF

The Portable Document Format, developed by Adobe, is arguably the most commonly used file format for storing the visual representation of a document. PDF ensures that a document will look exactly the same whatever device is used to view it ([w3.org/TR/WCAG-TECHS/pdf.html#pdf\\_notes](http://w3.org/TR/WCAG-TECHS/pdf.html#pdf_notes)). Most software that can be used to create texts and documents also supports exporting documents as PDF files. For scholarly editors wanting to produce a fully controlled document-style digital edition, PDFs can thus be a reliable solution. The downside of the PDF, however, is that it is a binary format: it stores all textual and layout information as a series of zeroes and ones, that is, it is not human-readable. This may be a hazard for long-term storage, as the format specification could change in future. Another drawback used to be that PDF was a proprietary format, that is, the specification and the related software technology were solely owned by the Adobe company. This made it difficult or impossible for software engineers other than those working for that company to do anything effectively with the format. These days, however, PDF is an open format and the specifications have been published for everyone to read and use. PDF is geared heavily to representing print-like documents. This means that interaction with a scholarly edition as a PDF will be limited almost completely to reading and searching. If more dynamic representation is required, editors would be better off looking into other formats and software.

### 3.4.9.3 EPUB

EPUB ([idpf.org/epub](http://idpf.org/epub)) is a widely used format for publishing e-books that can be read both on computer and tablet screens. EPUB is basically a packaged form of HTML. Various HTML files are contained in a larger container file. A number of special files are used to describe indexes, chapter structure, front matter, and so on. As has already been said, under the hood EPUB is relatively “plain” HTML5 with the same possibilities for styling. EPUB should, in theory, also be able to support interaction and multimedia, but device support for this is sketchy at best. A common misconception is that HTML/EPUB does not support page numbers because of its responsive design (i.e. scaling fonts and reflowing text to fit different window and tablet sizes). It is certainly possible to anchor page numbers to the text, but publishers mostly choose not to do so because it is highly likely that page breaks due to reflowing content will not neatly coincide with the bottom of a reading frame. This choice makes reliable referencing inside an EPUB text a considerable pain, and an issue that is in urgent need of being solved by future technology.

#### 3.4.9.4 LaTeX

LaTeX is a verbose document description language written by Leslie Lamport (for a history and technical details, consult [latex-project.org](http://latex-project.org)). It runs atop a typesetting system called TeX developed largely by Donald Knuth (see [tug.org/whatis.html](http://tug.org/whatis.html)). LaTeX and TeX have an important focus on publishing scientific papers containing complicated formulae. The LaTeX format has therefore found widespread adoption in academic publishing, both among researchers themselves as well as publishing houses. There are dedicated websites that support the authoring of LaTeX (e.g. [overleaf.com](http://overleaf.com)). Like XML, LaTeX uses markup codes. These codes can indicate what a part of a text represents – a section called “Introduction” for instance: “`\section{Introduction}`”. Such codes result in a fitting layout. Codes can also be more typographically specific, such as “`{\large This Text Will Be Large}`”. Because LaTeX was one of the first general-purpose document typesetting languages, it has evolved considerably over time. It now supports various dialects, modules, and libraries, often offering multiple paths towards the same end. For this reason, LaTeX is generally seen as a powerful but not easy-to-use or intuitive solution to document production. This notwithstanding, it has found a very large community of users and support (e.g. [tex.stackexchange.com](http://tex.stackexchange.com); [sharelatex.com/learn/latex/Main\\_Page](http://sharelatex.com/learn/latex/Main_Page)), especially in the academic context.

#### 3.4.9.5 XML and XSLT

If the base data of scholarly output is in the form of XML, scholars may choose to transform the XML into presentable HTML by using the specially designed templating language XSLT. XSLT is short for “eXtensible Stylesheet Language Transformations”. XSLT is a standard technology for transforming XML documents into XML with a different structure, but also into different documents altogether. It is most often used to transform some XML as a data source into an HTML form to represent that data visually. A comprehensive example of this was given in figures 7–8.

The popularity of (TEI) XML, especially in the scholarly community, has given rise to a number of software applications that facilitate the publishing of XML data as HTML. Noteworthy are especially Edition Visualization Technology (EVT; [evt.labcd.unipi.it](http://evt.labcd.unipi.it)) and the Versioning Machine (Schreibman 2016). In essence, these applications make the work of writing an XSLT stylesheet less cumbersome through clear tutorials and examples, and by abstracting away a bit from the most basic level of angle brackets and code verbs.

#### 3.4.9.6 On stacks, chains, pipelines, and sustainability

As mentioned, Web technologies are seldom used in isolation. Unless a scholar writes HTML directly, there will always be software, processes, templates, and transformations involved with publishing data in an electronic form. The full array of specific technologies that in a certain context is needed to produce a visualisation

of some source of data on the Web is often called a stack, a technology chain, or a pipeline. Usually, a Web framework will also be part of such a stack, a Web framework itself being a combination of various Web-oriented languages and technologies for creating Web applications.

Consider the case where a scholar has produced an XML description of a certain physical manuscript. Although it is possible to place this XML file on a server and open it to the world, the viewing of the XML itself would probably not satisfy either scholar or reader. So, at the very least, the scholar will also compose an XSLT stylesheet to present the XML in some more conveniently readable form. This XML plus XSLT combination amounts to the minimum stack that is needed for Web publishing an edition. But with every further requirement (paginating, searching, comparing, annotating), more styling templates, software, and components will be needed. Digital scholarly editions can therefore grow into large, intricate software machinery that requires sophisticated software engineering knowledge, enduring maintenance, and careful balancing of all the integrated components. The more elaborate such structures are, the more questionable the longevity of the edition tends to become. Unless maintenance can be guaranteed institutionally for a very long period, it would seem that keeping things as lean and as simple as possible offers a scholar the best chances of seeing an edition survive the constantly changing turmoil of digital environments. A base format for the data (such as JSON) that is then transformed into a Web publication consisting of HTML, CSS, and JavaScript would arguably be a good, lean choice from the perspective of persistence.

#### **3.4.10 Some concluding remarks**

The key question with respect to data formats is how we store our data. How do we best inscribe in a digital medium textual data and the critical observations we have made about such textual data, and how do we ensure the longevity of the variants we have examined and identified; our alignments; the collations we have produced; and the stemmata and, possibly, other networks and graphs that are the results of our analyses? Scholars should want to know enough about how data is or can be stored to judge the adequacy of the storage in terms of precision, and indeed representing what was meant, and to judge the potential sustainability of the chosen technological solution(s). The potential for sustainability and preservation is probably key when it comes to which formats are chosen by a scholar. But, secondly, a scholar should always consider whether a format is suited to the type of analysis he or she wants to perform – in other words, does the format formalise the data adequately and does it enforce some consistency that will allow sufficient (computational) analysis? Third, it would be good to bring interoperability into the equation – that is, the scholar should also consider how other scholars and other software may want to reuse the data, and whether the chosen format supports such reuse well. Finally, the choice of format may be influenced by considerations about

how an eventual presentation of the data or a digital scholarly edition as a whole might be published.

Unfortunately – or maybe not – there is no single format that does it all. A vast amount of work in the digital realm is to do with transforming data from one form to another to appropriate it for some other purpose, simply because not all formats are suited to all purposes. Some are better geared towards one function or another (Vitali 2016). The choice and use of a format should always be carefully evaluated, with respect both to the format's ability to store the information needed and to the ability of the format to be transformed (migrated) easily to other formats because of later and different requirements. In practice, different formats have different purposes and versatility, and turning one into another may affect readability or may lead to cumbersome and error-prone handling of information. For all of these reasons, good care should be taken when choosing formats, for these digital technologies do impact our ability to analyse historic texts, both in good and bad ways.

Finally, we may note that digital files require care to ensure their sustainability. Backups remain important, and any digital data or edition should be hosted on a server that is regularly maintained. Lastly, because digital data can still be vulnerable, and especially as institutional support can be fleeting, it may be sensible, while making a digital edition, to also provide for a print equivalent of some kind (e.g. PDF) to ensure longevity along both digital and analogue lines.