New technologies for the study of Euclid's *Elements*

Mark J. Schiefsky

February 1, 2007

1 Introduction

The specific purpose of this paper is to describe a set of new software tools and some of their applications to the study of Euclid's *Elements*. More generally, it is intended as a case study to illustrate some of the ways in which recent developments in information technology can open up new perspectives for the study of source materials in the history of mathematics and science. I argue that the creative and judicious use of such technology can make important contributions to historical scholarship, both by making it possible to pursue old questions in new ways and by raising new questions that cannot easily be addressed using traditional means of investigation.[1]

I begin with a brief description of three areas of current scholarly research on Euclid's *Elements*.

1. **Language and argument**. The language of ancient Greek mathematical texts is a highly specialized *Fachsprache*, which is sharply distinguished from literary prose in terms of lexical choice, syntax, and the repeated use of formulaic expressions. The

[1]The software tools described in this paper were developed in the course of the Archimedes Project (`http://archimedes.fas.harvard.edu`), an ongoing international collaboration between the Department of the Classics at Harvard University, the Max Planck Institute for the History of Science in Berlin, and the Perseus Project at Tufts University, and funded by the National Science Foundation (grant IIS-0085960, 2001-2004) and the Deutsche Forschungsgemeinschaft. Of the many individuals who have contributed in crucial ways to this project, special acknowledgment is due here to Malcolm Hyman, without whose insight and formidable programming skills the software I describe in section 2 would never have been developed, and to Peter Damerow, whose vision has shaped and guided my thinking on the role of information technology in the history of science from the beginning. I would also like to thank Jim Carlson for the invitation to participate in the Oxford conference and all the participants who discussed an earlier version of this paper with me.

mathematical lexicon, conceived of as the sum total of these distinctive elements, is the fundamental tool for the expression of mathematical argumentation. A recent and welcome trend in scholarship on Greek mathematics has been an increased level of attention to the characterization of the mathematical lexicon.[2] But the study of the linguistic expression of mathematical argumentation is still very much in its early stages, and stands to benefit greatly from a more thorough and systematic approach.[3]

2. **Transmission studies.** As other contributions to this volume amply illustrate, the transmission of the *Elements* from Greek antiquity to the modern world raises complex problems which can only be addressed by a combination of manuscript studies, detailed philological work, and the study of cultural contexts of reception. The investigation of this process — of the ways in which the *Elements* was received, transmitted, and transformed in the civilizations of medieval Islam, the Latin West, and early modern China — is a major field of inquiry in its own right. The Arabic tradition, in particular, is especially rich and complex, and presents a diverse array of edited versions and adaptations of the *Elements* whose relations to one another are far from clear.[4] Furthermore, recent work has shown that study of the Arabic transmission is essential to the constitution of the Greek text itself: for it is by no means obvious that the text of the Greek manuscripts is always to be preferred to that of the Arabic tradition.[5] Thus, there is an urgent need for more detailed studies of the transmission of the *Elements*, even if the goal is limited to improving the Greek text.

3. **The study of deductive structures.** The problem here is to characterize both the structure of argumentation within specific propositions and the complex networks of deductive relationships between different propositions.[6] On the level of the individual proposition, this investigation is closely connected with the study of the relationship between language and argument discussed above. On a larger scale, the mapping of deductive relationships between propositions is crucial to understanding the organization of individual books and the *Elements* as a whole. It is also fundamental to

[2]See especially Netz (1999, chh. 3-4), building on Aujac (1984).

[3]Netz (1999, ch. 4) gives a fairly lengthy catalog of formulaic expressions in Greek mathematics, but does not pursue the analysis with anything like the level of detail needed to draw convincing conclusions.

[4]See the contribution of Sonja Brentjes to the present volume, as well as Brentjes (1994), Brentjes (2001), DeYoung (1981), DeYoung (1984), Engroff (1980). In light of these analyses, the longstanding hypothesis of a simple bifurcation in the tradition between the descendants of the so-called al-Hajjaj and Thabit/Ishaq recensions (cf. Heath (1956, 1:75-90)) has been rendered untenable.

[5]Knorr (1996) argues convincingly against the assumption, dominant since Heiberg, that the Greek manuscript tradition is more reliable than the Arabic, *tout court*. But the situation is a highly complex one, and the problems must be approached individually for specific books and indeed specific propositions within them. Cf. Vitrac (1990-2001, 4:32-71) and Acerbi (2003).

[6]The latter approach is fundamental to Mueller (1981); Netz (1999, ch. 5) devotes more attention to argumentation within propositions.

> any attempt to identify the mathematical 'toolbox', the set of results that because of their especially frequent application in mathematical investigations can be viewed as making up the basic conceptual armory of the Greek mathematician.[7]

The problems raised by these areas of scholarly inquiry into the *Elements* pose substantial challenges for information technology. Reflection on these challenges suggests that a software platform that can assist with scholarship of this kind should be designed with the following three goals in mind.

First, it should provide access to the large and growing array of electronic resources that have been developed for the automatic analysis of languages such as ancient Greek, Latin, Arabic, and Chinese. The two primary types of tool are morphological analyzers — in which a word as it appears in a text is automatically analyzed into its dictionary form and part of speech — and online dictionaries, which provide access to entries via the dictionary form of the word.[8] The software platform should offer a unified yet flexible way of accessing these tools and applying them to a wide range of historical sources.

Second, the software should aim to integrate browsing and annotation. The advent of the World Wide Web has undoubtedly made a large amount of information widely accessible to scholars, including large text collections as well as the linguistic tools mentioned in the previous paragraph. But access to these resources tends to be passive, as reflected in the term web *browser*: the user accesses the information without having the opportunity to create any comments or annotations on it. For example, a user might call up the image of a manuscript from a digital collection and have the ability to zoom in on particular parts of it as desired, but there is in general no way to mark a particular spot on the image as being of special significance (e.g. the beginning of a book or proposition). Alternatively, one might click on a word in a text and be taken automatically to the entry for the word in an online dictionary, but there is no way to indicate, for example, that the word in question is a technical term in relation to a particular discipline. The annotation of images and texts is the fundamental activity required by philologically rigorous scholarship of the kind in question here. Such scholarship demands a new kind of software that will both provide access to the vast resources of the Web and also make it possible to comment on the information that these resources provide.

[7] The notion of the mathematical toolbox was introduced by K. Saito and taken up by Netz (1999, 216-235). See Saito's online indices of the propositions used in Apollonius and Pappus 7, available at `http://www.hs.osakafu-u.ac.jp/~ken.saito/`.

[8] A brief note may be helpful for readers unfamiliar with highly inflected languages such as Greek and Latin. In such languages, the actual word that appears in a text often differs considerably from the dictionary form. For example, the Greek word *isos* 'equal' can appear in a wide variety of forms including *ison, isa, isē,* and *isai*, among others, depending on the grammatical factors of gender, number, and case. A morphological analyzer can take each one of these forms and reduce it to the dictionary form *isos*.

Third and most important, the software should not only allow for the creation of annotations, but also provide mechanisms by which those annotations can become the starting point for further analysis. To take a simple example, one might search in a text for all the instances of a particular term, then use those search results as a basis for navigating through the text, performing further searches, and so on. The need to create annotations on annotations is a reflection of a more general point: software should be designed to maximize the *interaction* between scholarly judgment and the results reached by automatic methods, for it is in such interaction that the greatest potential of information technology to contribute to the humanities lies. There is no question of computer analysis *replacing* scholarly judgment and the rigors of traditional philology. No software for morphological analysis will identify every form in a text with perfect accuracy, and no search routine will provide results that are 100% accurate on every occasion. But it would be a serious mistake to reject all the information that such technology can provide out of hand just for this reason. What is needed are more tools that enable such information to be assessed and used by scholars, and a primary feature of such tools will be the kind of interactivity whereby the results provided by the software can be evaluated and used as the starting point for further analysis.

It is with these three goals in mind that the tools which I will describe in the next section were conceived and developed. To this description I now turn.

2 New technologies

2.1 XML

The technological approach which I will describe here is based on the use of the so-called Extensible Markup Language or XML. I therefore begin with a brief discussion of XML, its usefulness for scholarly purposes, and the technical challenges it poses.

The basic idea of a *markup language* such as XML is to combine text with information about the text in a single document. The information, which serves to structure the text, is encoded in *markup*. This is best illustrated by means of an example (fig. 1). Here we see the beginning of an XML version of book 1 of the *Elements* in Heath's translation. The XML text consists of a set of nested *elements*, such as `book`, `p`, and `s`; the beginning of each element is marked by a *start tag* (e.g. `<book n="1">`), and the end by an *end tag* (e.g. `</book>`). These tags constitute the *markup* (shown in bold), and enclose the *content* of the XML document (shown in regular type). Each element may have associated with it one or more *attributes*, which contain further information about that element. For example, each of the `s` elements in the above document carries an `id` attribute, whose value serves

```
<text>
   <author>Euclid</author>
   <title>Elements</title>
   <book n="1">
     <p n="Prop1">
       <s id="000001">On a given finite straight line
          to construct an equilateral triangle.</s>
       <s id="000002">Let AB be the given finite
          straight line.</s>
       <s id="000003">Thus it is required to construct
          an equilateral triangle on the straight line
          AB.</s>
       ...
     </p>
     ...
    </book>
</text>
```

Figure 1: XML version of Euclid's *Elements*, book 1.

as a unique identifier of that particular element within the document.

A great advantage of XML for scholarly purposes is that it allows for highly flexible structuring of documents; the user may specify arbitrary element names and relations between them, the only absolute requirement being that the elements must be nested hierarchically (i.e. the content of elements cannot overlap). In addition to this, XML provides a number of other advantages. (1) It is fully compatible with the Unicode standard, which enables documents to be encoded in virtually all the world's languages.[9] (2) It is the emerging standard for data exchange in the World Wide Web, where it should eventually replace the now standard HTML (Hypertext Markup Language). (3) Sophisticated queries can be performed on XML documents, such as 'find the fifth sentence of the third paragraph with attribute `type` equal to `'Prop.'`[10] (4) Finally, the so-called Extensible Stylesheet Language or XSL provides a standard method of transforming XML documents into one another or into other formats such as HTML.[11]

[9] See `http://www.unicode.org`.

[10] See the descriptions of the so-called XPath and XQuery standards at `http://www.w3.org/TR/xquery` and `http://www.w3.org/TR/xpath`.

[11] See `http://www.w3.org/Style/XSL`.

Despite these advantages, the use of XML also poses a number of technical challenges that are not met by standard software. Two of these stand out in particular. First, the hierarchical organization of XML documents makes it impossible to encode overlapping structures in a single document. Second, although standard software tools provide excellent support for accessing XML documents at the level of individual elements, they do not support access below the level of individual elements nearly as well. This raises major problems for the kind of philologically rigorous scholarship described in the Introduction, which depends crucially on the ability to search for and annotate individual words or groups of words.

In the remainder of this section I will describe a software platform that attempts to overcome these technical challenges while keeping in view the goals identified in the Introduction.

2.2 Linguistic services: Pollux and Donatus

The Pollux and Donatus systems provide morphological analysis and dictionary lookup as services accessible via the World Wide Web. The basic principle of such web services is simple: the server receives a query in a standard format and returns a result encoded in XML. The power lies in the unification of the various resources on the server. Instead of consulting a variety of morphological analyzers and online dictionaries, the user can go to a single Web address that provides direct access to all these resources. The technicalities of implementing the morphological analyzers and dictionaries are handled on the server and thus shielded from the user's view.

The Pollux service accepts a dictionary keyword as well as the selected dictionary as input and returns the relevant entry; it can be accessed via a simple Web interface (`http://archimedes.fas.harvard.edu/pollux`) or else through the Arboreal software (below, section 2.4). Donatus accepts a single XML document as input and returns an XML file containing morphological analyses of all words in the document as well as a list of unanalyzed forms. Documents may contain text in multiple languages; language is specified in a standardized fashion in the XML markup. Donatus also provides an interface for uploading forms not recognized by the morphological analyzers installed on the server; these forms will be analyzed correctly when the service is subsequently queried, making the system fully extensible. The languages for which morphological analysis is currently available include Arabic, Dutch, English, French, German, Greek, Italian, and Latin; further languages are being added as morphological analyzers become available. Like Pollux, Donatus may be accessed via a Web interface (`http://archimedes.fas.harvard.edu/cgi-bin/donatus`) or through the Arboreal software (below, section 2.4).

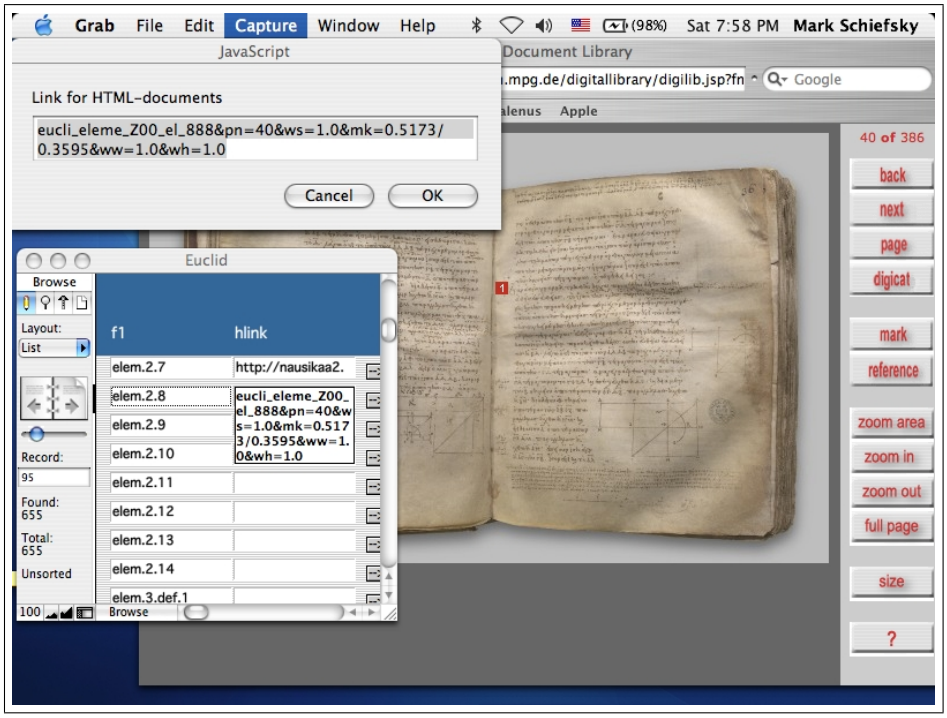

Figure 2: Use of the Digilib software for image browsing and annotation.

2.3 Image browsing and annotation: Digilib

The Digilib software is designed to address the problem of integrating the browsing and annotation of digital images. The user can browse flexibly through an image collection stored on a server and zoom in on particular images or parts of images. Annotations can be created by using the mouse to place marks on the image; the software then supplies a stable link that points to the image at the selected resolution, together with its annotations. By storing these links in an external file or database, it is possible to build up a set of links for navigating through a manuscript. Digilib is implemented in JavaScript, and can be accessed with any standard Web browser such as Internet Explorer or Mozilla Firefox.[12]

Figure 2 illustrates the Digilib software in use. In the background we see a page of the D'Orville manuscript of Euclid's *Elements*[13] stored in a Digilib collection and opened in

[12]For further information on Digilib, developed at the University of Bern in cooperation with the Max Planck Institute for the History of Science in Berlin, see `http://developer.berlios.de/projects/digilib`.

[13]Codex Bodleianus Dorvillianus X, 1 inf. 2, 30 = MS D'Orville 301, Bodleian Library, Oxford; designated

a standard Web browser. The user has placed a mark on the image at the beginning of proposition 8 of Book 2. The dialog box on the upper left shows the link to the image with annotation, which is being pasted into a FileMaker Pro database on the lower left; once the database is complete, the user will be able to use it to navigate through the manuscript proposition by proposition.[14]

2.4 Arboreal: an XML-based browsing and annotation environment

The Arboreal software provides a unified platform for accessing the resources described in the previous two subsections, as well as sophisticated functions for the annotation of source texts encoded in XML format. It is designed as a prototype of a next-generation Web browser for XML documents, and is the linchpin of the software platform described here. Arboreal is a standalone Java program and runs on all platforms on which Java is implemented (including Mac OSX, Windows, and Linux); it is continually being extended and improved, and the latest version can be downloaded at `http://archimedes.fas.harvard.edu/arboreal`.

Basic navigation and linguistic services. Figure 3 shows an XML text of Euclid's *Elements* open in Arboreal. The main window is split into two parts, the *tree pane* on the top and the *content pane* on the bottom. The user navigates through the text by selecting (clicking on) particular elements in the tree pane; each of the selected elements is then displayed in the content pane. Text in languages such as Greek or Arabic can be rendered in an underlying transcription such as Beta Code, in transliteration (as here), or in the original script. Arboreal is fully integrated with both the Pollux and Donatus web services: the user can access morphological analyses and dictionary entries for any word shown in the content pane by clicking on the word and selecting the appropriate option from a pulldown menu. In fig. 3, the user has clicked on the word *isopleuron*; morphological information for this word is displayed in the 'Morphology' window at the upper right hand corner of the screen. Morphological data can be read either from an XML file stored on the user's computer or directly from Donatus; additionally, data for unanalyzed forms can be added and uploaded to Donatus by selecting various menu options. Arboreal is also fully integrated with the Digilib software. The pulldown menu which appears when the user clicks anywhere in the content pane provides an option to open the page image corresponding to the selected point in the document; the image then opens in a standard Web browser. The link to the page image is contained in the XML markup of the document, and integration with Digilib is achieved simply by making this link point to an image stored in a Digilib

'B' in Heiberg's edition (Heiberg & Menge 1883-1916).

[14]For an index allowing this kind of navigation through the D'Orville manuscript in Digilib see `http://archimedes.fas.harvard.edu/euclid/digilib.html`.

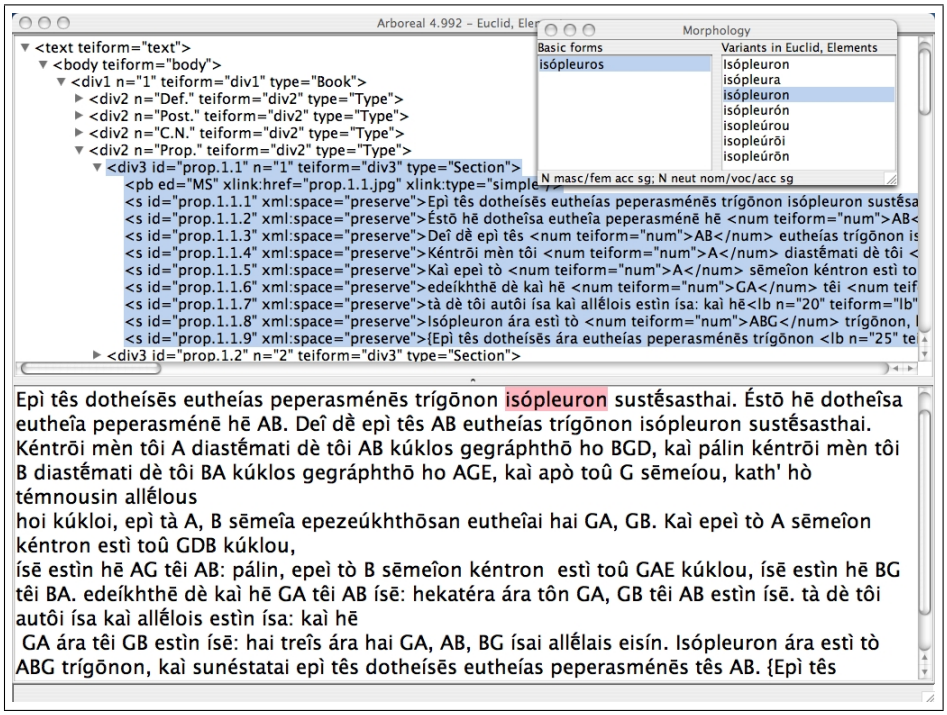

Figure 3: Arboreal: basic navigation and morphology.

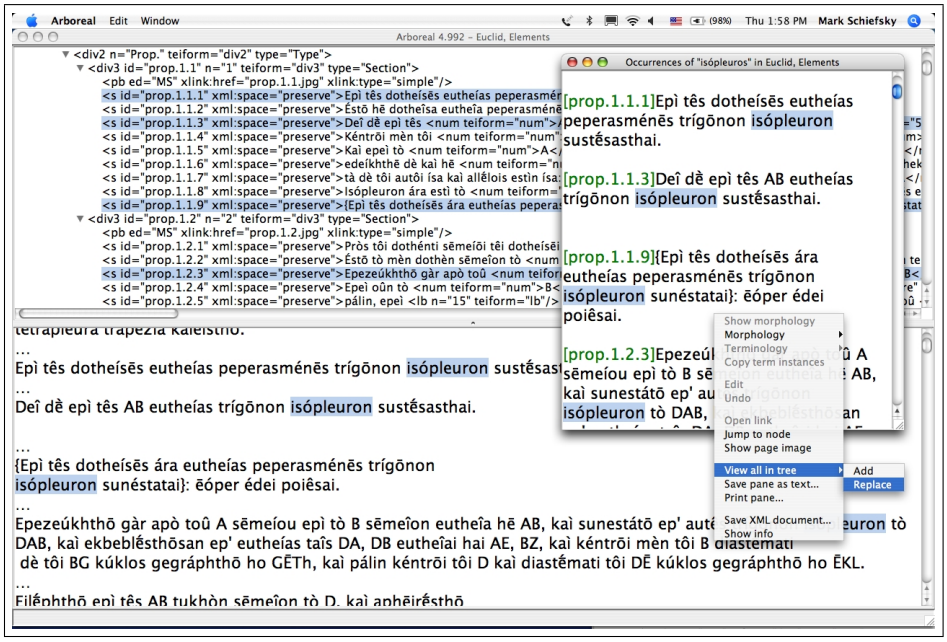

Figure 4: Search results and navigation in Arboreal.

collection.

Searching. Arboreal provides a variety of sophisticated search capabilities, including searching by wildcards or regular expressions and lexical searching (where the program finds all morphological variants of a given lexical form). As well as searching for text, the user may also search for XML elements or particular attributes, or make more complex queries of the document using the XPath syntax. Search results can be displayed either in the main window or a new independent window; in the latter case they can be used to navigate through the text displayed in the main window. This procedure is illustrated in figure 4. The user has performed a lexical search for all forms of the word *isopleuros* in the *Elements*; the sentences containing forms of this word are displayed in the smaller window at the upper right, with the forms themselves highlighted. By clicking anywhere in this window and selecting the appropriate option from a pulldown menu, the user can add the search results to the selection of elements in the main window or replace that selection entirely (as here). In this way search results become the starting point for further analysis of the text.

Annotation of terminology. A basic feature of Arboreal is the ability to annotate individual words or groups of words as *terms*. Each term has a single name, language,

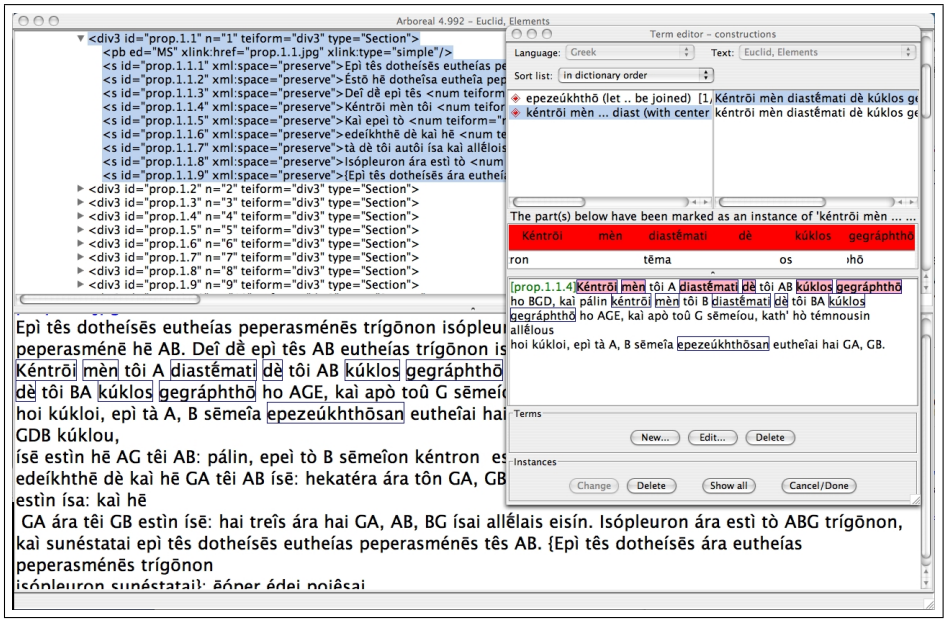

Figure 5: Annotation of terminology in Arboreal.

and translation, as well as a fixed number of parts. Terms may have multiple *instances*, depending on the number of times they occur in a given text or set of texts; each instance consists of a set of *parts*, i.e. words in the XML text. Terms are stored in termlists, which can be saved as XML files. When a termlist is loaded into Arboreal, every instance of every term in the list is highlighted in the content pane. The handling of terminology is facilitated by the term editor, a secondary window which displays all the terms in a given termlist along with references to their instances in the text. The user can add instances of a term either by clicking on words in the content pane and selecting the appropriate option from the pulldown menu or by dragging search results directly into the term editor; the latter option makes it easier to create extensive termlists for a given text or set of texts. Figure 5 illustrates some of these features. The user has marked two instances of the formulaic phrase for the construction of a circle: 'with center ... and distance ... let the circle ... be drawn' (*Kentrōi men ... diastēmati de ... kuklos gegraphthō*); these are highlighted in the content pane. The term editor displays both the terms (in the left hand column) and references to their instances (in the right hand column); the user has clicked on a term instance, which is displayed and highlighted in the lower part of the term editor. Just as in the case of search results, the user can add the elements containing the selected instances to the selection in the main window or replace that selection entirely.

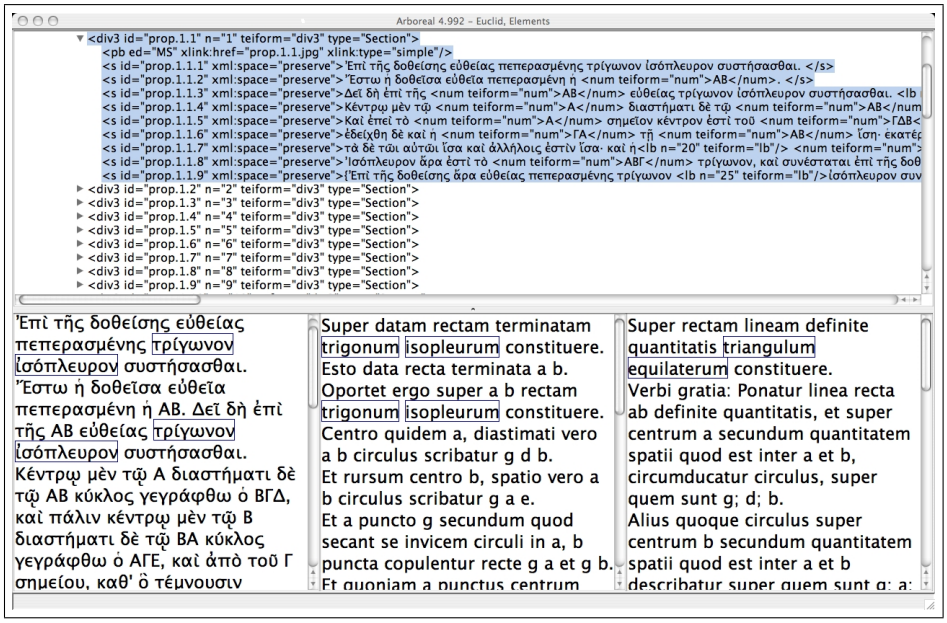

Figure 6: Using Arboreal to work with parallel texts.

Work on parallel texts. Arboreal allows multiple parallel or *slave* texts to be loaded for a single main or *master* text. As the user navigates through the master text, the panes displaying the slave texts automatically jump to the corresponding elements of those texts. The correspondence between master and slave texts is specified in an XML matching file, which is loaded along with the slave text; all linguistic services as well as searching and annotation functions are available for the slave texts. Figure 6 illustrates the use of Arboreal to work with parallel versions of the *Elements*. The master text is the Greek of Heiberg's edition, displayed in the content pane on the bottom left; the other two content panes display the text of the anonymous Latin version made directly from the Greek (the so-called 'Graeco-Latinus'; see below, section 3.2) and that of Gerard of Cremona. Some correponding terms have been marked in all three versions.

Editing and XSL transformations. Any content pane in Arboreal can be made editable by clicking on the pane and selecting an option from the pulldown menu. The user can edit the content of the XML document at will, though no changes can be made to the XML structure itself (i.e. elements and attributes cannot be changed). When combined with the ability to work with parallel texts, such editing functionality has a number of important applications, including the writing of translations or commentaries and the preparation of critical editions. Finally, Arboreal allows an arbitrary XSL transformation to be applied to

an XML document, generating a new document that opens in its own main window. With this feature Arboreal becomes a highly extensible tool for the creation of XML texts; the possibilities are limited only by those of the XSL specification itself.

Let me now return to the technological challenges raised above in section 2.1 and explain how they are addressed by this software platform. (1) The first challenge, the problem of the hierarchical nature of XML documents, is handled via the notion of *overlay tagging*, whereby overlapping structures are represented as collections of links or pointers into the source XML text rather than as elements within the text itself. This is the strategy underlying terminology annotation in Arboreal. Since multi-word terms may be discontinuous and interlaced with one another, they cannot be tagged as XML elements in the source text itself; but there is no problem if the information is stored as links in separate XML documents. Similarly, different attempts to divide the Euclidean proposition into constituent parts could easily result in overlapping structures; the only way to allow for such possibilities is via overlay tagging.[15] (2) The second challenge posed by the use of XML is access to document content at the level of individual words; this problem is largely resolved by Arboreal's linguistic architecture.[16]

That the software described here goes a long way towards meeting the first two of the three challenges to information technology discussed in the Introduction— unified access to linguistic services and the integration of browsing and annotation — should by now be clear. It is, however, worth reflecting on the way in which it addresses the third challenge, that of interactivity. To some extent interactivity in Arboreal arises out of specifically programmed functions, as when search results are used to navigate through a document. But the more fundamental consideration is that Arboreal is a highly general tool that can be used to browse and annotate any XML document; since the annotations it creates can themselves be saved as XML documents, Arboreal has a built-in capacity to create annotations on annotations. When this is combined with the power of XSL transformations to generate new XML documents, the range of possibilities is wide indeed.

In the Introduction I stressed the challenges of scholarship to information technology. Now that a first step towards the realization of these goals has been achieved, it is time to consider the challenge posed by the technology to scholarship and to look at what these tools can accomplish.

[15] Overlay tagging has the further great advantage of separating annotation from changes to the source text; this enables multiple users to make different annotations to the same sources without conflicting with one another.

[16] Difficulties remain, however, in the case of languages such as Chinese where the concept of 'word' is problematic; the development of a generalized architecture for word segmentation is an area where further work is needed.

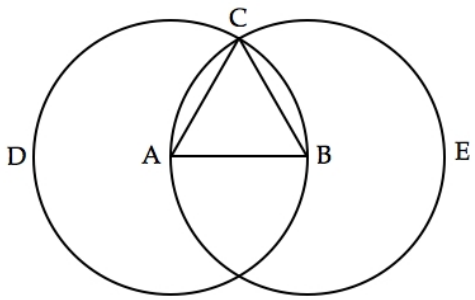

Figure 7: Euclid, *Elements* 1.1.

3 Applications and Results

I return to the three areas of scholarly investigation into the *Elements* mentioned in the Introduction and present some applications and results.

3.1 Language and argument

The software platform described in the previous section is ideally suited to the identification and study of formulaic language in the *Elements*. Using Arboreal's searching and annotation facilities, regularly recurring words and combinations of words can easily be identified, annotated, and their frequencies determined. While this is in principle no different from a traditional philological approach, the software can assist in carrying out such investigations in a large-scale and systematic way. Furthermore, since the results are stored in XML files with a standard format, it is relatively easy to correct, extend, and share them with other scholars.

The issues are best approached by taking the example of a specific proposition. The table below gives the standard Greek text (Heiberg) of proposition 1 of book 1 of the *Elements* along with Heath's translation (slightly modified); formulaic elements have been underlined. The text is divided according to the standard scheme given by the ancient commentator Proclus; this provides a convenient way of referring to the parts of the proposition and will be the basis of my exposition, though it is by no means rigidly adhered to even in book 1.[17] For the diagram see fig. 7.

[17] On the formal divisions of the Euclidean proposition see Heath (1956, 1:129-131).

Enunciation (*protasis*)	*Epi tēs dotheisēs eutheias peperasmenēs trigōnon isopleuron sustēsasthai.*	On a given finite straight line to construct an equilateral triangle.
Setting-out (*ekthesis*)	*Estō hē dotheisa eutheia peperasmenē hē AB.*	Let AB be the given finite straight line.
Specification (*diorismos*)	*Dei dē epi tēs AB eutheias trigōnon isopleuron sustēsasthai.*	Thus it is required to construct an equilateral triangle on the straight line AB.
Construction (*kataskeuē*)	*Kentrōi men tōi A diastēmati de tōi AB kuklos gegraphthō ho BGD, kai palin kentrōi men tōi B diastēmati de tōi BA kuklos gegraphthō ho AGE, kai apo tou G sēmeiou, kath' ho temnousin allēlous hoi kukloi, epi ta A, B sēmeia epezeukhthōsan eutheiai hai GA, GB.*	With center A and distance AB let the circle BCD be drawn; again, with center B and distance BA let the circle ACE be drawn; and from the point C, at which the circles cut one another, to the points A, B let the straight lines CA, CB be joined.
Proof (*apodeixis*)	*Kai epei to A sēmeion kentron esti tou GDB kuklou, isē estin hē AG tēi AB: palin, epei to B sēmeion kentron esti tou GAE kuklou, isē estin hē BG tēi BA. edeikhthē de kai hē GA tēi AB isē: hekatera ara tōn GA, GB tēi AB estin isē. ta de tōi autōi isa kai allēlois estin isa: kai hē GA ara tēi GB estin isē: hai treis ara hai GA, AB, BG isai allēlais eisin.*	And since the point A is the center of the circle CDB, AC is equal to AB. Again, since the point B is the center of the circle CAE, BC is equal to BA. But it was shown that CA is equal to AB; therefore each of the straight lines CA, CB is equal to AB. And things which are equal to the same thing are also equal to one another; therefore CA is also equal to CB. Therefore the three straight lines CA, AB, BC are equal to one another.
Conclusion (*sumperasma*)	*Isopleuron ara esti to ABG trigōnon, kai sunestatai epi tēs dotheisēs eutheias peperasmenēs tēs AB. hoper edei poiēsai.*	Therefore the triangle ABC is equilateral; and it has been constructed on the given finite straight line AB. Which it was necessary to do.

I shall go through the parts of the proposition one by one, commenting on the formulaic elements as they come up.

First the *enunciation* states what is given as well as the result to be proved or the procedure to be carried out: 'on a given finite straight line to construct an equilateral triangle'. This is an example of what the ancient commentators classified as a 'problem', since it requires that something be *done* rather than *shown* or demonstrated (the distinguishing feature of 'theorems', the other main type of proposition).[18] The enunciation of problems typically makes use of the bare infinitive, here *sustēsasthai* 'to construct'. Theorems are typically stated using finite verbs in the indicative mood, as in the enunciation of 1.5: 'In isosceles triangles the angles at the base are equal to one another, and, if the equal straight lines are produced further, the angles under the base will be equal to one another'.

While the propositions of the *Elements* are stated in general terms, the arguments in the text deal with particular mathematical objects. The move from general to particular is made in the *setting-out*, which introduces the particular objects with which the argument will be concerned. The form *estō* 'let ... be', third person imperative of the verb *eimi* 'to be', marks the setting-out in 42 of the 48 propositions of book 1, including proposition 1. In most cases *estō* occurs at the beginning of the setting-out (propositions 1-6, 8-12, 16-20, 22-26, 29-48; in propositions 41 and 48 it is postponed). Elsewhere, other third person imperative forms mark the setting-out (propositions 7, 13-15, 21, 27-8).

In book 1, the *estō* of the setting-out is regularly followed by two other expressions introducing the *definition* or *specification* (*diorismos*). This part of the proposition states what it is necessary to do or show in the case of the particular mathematical object specified in the setting-out. In the case of problems such as proposition 1, the specification is regularly marked by the phrase 'thus it is required' (*dei dē*) (14 cases in book 1). In the case of theorems, it is marked by 'I say that' (*legō hoti*) (34 cases in book 1). In book 1 *dei dē* occurs only in specifications (14 cases), with one interesting exception.[19] However, we find *legō hoti* in both the 'construction' and 'proof' sections of a number of propositions (e.g. 9-12, 22, 26, 34, 39, 40, and 46); clearly its use was not *restricted* to specifications.

The specification is typically followed by the *construction* (*kataskeuē*), in which the geometrical constructions necessary for the proof are carried out. These operations are regularly expressed using the third person perfect passive imperative, which carries a strong sense of completed action ('let ... have been constructed'); it is as though the author wanted to emphasize that the construction has already been carried out by the time one gets to the proof. In the construction of proposition 1 we find two instances of the highly formulaic

[18] On the distinction between theorems and problems see Heath (1956, 1:124-129) (drawing heavily on Proclus).

[19] In 1.22, where *dei dē* is found in the enunciation as well as the specification. Heiberg, following ancient commentators such as Proclus and Eutocius, originally emended the text of the enunciation to *dei de*; later, however, he retracted the emendation (see Heath (1956, 1:131n1), who accepts *dei dē*). Cf. Vitrac (1990-2001, 1:237n107).

expression for constructing a circle:

with center {A} and distance {AB} let the circle {BCD} be (lit. 'have been') drawn

Kentrōi men {*tōi A*} *diastêmati de* {*tōi AB*} *kuklos gegraphthō* {*ho BGD*}

This formula occurs some eight times in Book 1, each time in identical form; the elements within the braces are variable, but everything else is fixed. Also formulaic in the construction for proposition 1 is *epezeukhthōsan*, 'let there be (have been) joined'; it occurs 30 times in Book 1, always with straight lines as the subject.

The *proof* (*apodeixis*) supplies the argument that the objects in the construction meet the requirements of the specification. The proof in 1.1 is largely made up of assertions of equality between different mathematical objects, expressed using the formula:

isos esti {*A*} {*B*}[20]

{A} is equal to {B}

This is by far the most common relational formula in book 1; it occurs over 300 times and often makes up the core of the proof. There is significant variation in the order of the elements: in the present proposition alone, the references to the objects whose equality is being asserted sometimes precede the words *isos esti* and sometimes follow them, and we find both the order *isos esti* and *esti isos*.

The argument itself may be schematized as follows:[21]

Step	**Assertion**	**Justification**
1	$AC \sim AB$	Def. 15
2	$BC \sim AB$	Def. 15
3	$AC \sim AB$	(restatement of 1)
4	$AC, BC \sim AB$	2, 3
5	$AC \sim BC$	CN 1
6	$AC \sim AB \sim BC$	CN 1

[20]The form *isos* is nominative singular masculine; depending on such features as the gender and number of the objects whose equality is being asserted, it may occur in the forms *isos, isa, isē* or *isai*, among others. Similarly the verb 'is' (*estin*) may also appear in the plural 'are' (*eisin*).

[21]I use $\sim$ to represent the relation conveyed by the Greek words *isos esti* 'is equal', and ignore variation in the references to lettered objects between (e.g.) AB and BA.

Fundamental to the proof is the application of Common Notion 1, 'things which are equal to the same thing are also equal to each other' (*ta de tōi autōi isa kai allēlois estin isa*), which is explicitly quoted in the move from step 4 to step 5. The core of the argument, which runs from steps 2 to 5, can be summarized as follows: $A \sim B, C \sim B \rightarrow C \sim A$ (where A stands for BC, B for AB, and C for AC); this in turn may be abbreviated $ABCBCA$. The question then arises to what extent this particular pattern of inference is typical. A survey of all the applications of CN 1 in Book I reveals the following distribution of cases:

Pattern	**Frequency**
$ABCBCA$	4
$ABCBAC$	4
$ABACCB$	4
$ABACBC$	9
$ABBCAC$	4
$ABCACB$	1
$ABCABC$	1
Total	27

The data show a striking degree of flexibility. The last two elements in each group are evidently interchangeable. As far as the first four elements are concerned, there is a slight preference for the order $ABAC$ over $ABCB$, and a more marked preference for both of these over $ABBC$ and $ABCA$, which correspond more closely to the modern assertion of transitivity ($A \sim B, B \sim C \rightarrow A \sim C$). This situation presumably reflects the force of the verbal formulation of CN 1, which is not a statement of transitivity as such; indeed the pattern followed is $ABCB$ in the only three passages where CN 1 is actually quoted according to Heiberg's text (1.1, 1.2, and 1.13). Nevertheless, the data clearly indicate that the author of the *Elements* did not associate the principle formulated in CN 1 with a rigidly formalized schema of argument; rather, it is applied in a highly flexible way that implies a conception of equality as a reflexive, symmetric, and transitive relation. However striking the formulaic character of the language of the *Elements* may be in certain respects, it does not extend to the level of rigid standardization of either the linguistic expression of notions such as equality or of the forms of argument associated with such notions.

The final part of the proposition, the *conclusion* (*sumperasma*), asserts that what the enunciation said was to be shown or done has indeed been achieved. Characteristic linguistic features here are both the particle *ara* ('therefore'), which occurs in the conclusion of every proposition in book 1, and a brief concluding phrase, either *hoper edei poiēsai* ('which it was necessary to do') in the case of problems or *hoper edei dei* ('which it was necessary to show') in the case of theorems.

We have seen that various linguistic elements are characteristic of the different parts of the Euclidean proposition. But the structure is a flexible one. Many propositions have no construction, and the distinction between the parts is not always clear cut; moreover some formulae (like the phrase *legō hoti* 'I say that') are used freely in different parts of the proposition. The Euclidean proposition is best thought of as a creative combination of different linguistic formulae and argument forms rather than fixed building blocks. Furthermore, although some operations such as geometrical constructions are expressed in a highly formulaic way, the expression of relations such as equality shows substantial variation and flexibility. There is no question of the *Elements* simply being reducible to a sequence of linguistic formulae; rather, what is needed is further study of the concrete ways in which different structures of argument are expressed in language. I hope to have shown in this section that technology offers a promising approach to such study.

3.2 Transmission studies

With its built-in facilities for working on parallel texts, the Arboreal software greatly enhances the possibilities for making the kinds of systematic comparisons between different versions of a text in the same language, or between texts and translations, which are fundamental to the study of transmission and reception. A crucial problem in this connection is the study of terminological correspondences between texts and translations. Such correspondences often provide considerable insight into translation methods, and the choice of terminological equivalents is sometimes an important criterion for attributing a translation to a particular translator. Again, Arboreal provides a framework in which these correspondences can be identified and studied in a systematic way.

As an example I present a comparison of proposition 1.1 according to (a) the Greek text of Heiberg, (b) the single anonymous Latin version known to have been made directly from the Greek (henceforth the 'Graeco-Latinus'), and (c) the Latin version of Gerard of Cremona.[22] That the Graeco-Latinus was indeed made directly from the Greek is clear both from the translator's frequent use of transliterated Greek words and from the close, word-for-word manner with which he follows the Greek text; various details suggest that he used a manuscript closely related to the D'Orville Euclid.[23] Gerard's version was made from the Arabic, not from the Greek; it thus provides information about the state of the Arabic tradition in the twelfth century AD as well as the incipient Latin traditions of the *Elements*.

How closely the translator of the Graeco-Latinus follows the Greek is illustrated by the

[22]For the Graeco-Latinus see Murdoch (1966) and the critical edition of Busard (1987); for Gerard's text see Busard (1984).

[23]See Busard (1987, 7-11), building on arguments of Murdoch (1966, 260-263).

following table, which gives the opening phrase of the construction of proposition 1.1 ('with center A and distance AB let the circle BCD be drawn') in the three versions (**H** = Heiberg, **GL** = Graeco-Latinus, and **G** = Gerard):[24]

H	*Kentrōi men*	*tōi A*	*diastēmati de*	*tōi AB*	*kuklos gegraphthō*	*ho BGD*
GL	*Centro quidem*	*a*	*diastimati vero*	*a b*	*circulus scribatur*	*g d b*
G	*super centrum*	*a*	*secundum quantitatem spatii*	*quod est inter a et b*	*circumducatur circulus*	*super quem sunt g d b*

The Graeco-Latinus follows the Greek down to the level of individual words, including the transliteration *diastimati* for the Greek *diastēmati* 'distance'. Gerard's version, on the other hand, has the relatively lengthy periphrasis 'according to the quantity of space between a and b' (*secundum quantitatem spatii quod est inter a et b*) for the Greek 'with distance AB' (*diastēmati de tōi AB*), and refers to the circle itself as 'the circle, on which are g d b' (*circulus, super quem sunt g d b*) instead of 'the (circle) BGD' in the Greek and the Graeco-Latinus. This way of referring to geometrical objects (e.g. 'the point, on which is A') is found elsewhere in Gerard's version, and is especially interesting because it corresponds to a usage found in the mathematical passages of the Aristotelian corpus (e.g. *to eph' hou A* instead of the normal Euclidean *to A* for 'the point A').[25] Further terminological differences between Gerard and the Graeco-Latinus in 1.1 are summarized in the following table; the most striking pattern that emerges is the tendency of the Graeco-Latinus to use transliterations (e.g. *trigonus, isopleurus*) where Gerard has Latin equivalents (*triangulus, equilaterus*):

Heiberg	**Graeco-Latinus**	**Gerard**
trigōnon 'triangle' (3)	*trigonus* (3)	*triangulus* (3)
eutheia 'straight line' (5)	*recta* (5)	*linea* (3), *recta linea* (2)
isopleuros 'equilateral' (3)	*isopleurus* (2), *equilaterus* (1)	*equilaterus* (2), *equalium laterum* (1)
estō 'let ... be' (1)	*esto* (1)	*ponatur* (1)

[24] Cf. Murdoch (1966, 253-254).

[25] See e.g. Aristotle *Physics* 7.5, 249b27-250a19; *Mechanical Problems* 848a30, 848b16-17, 849a6, 849b12, 850a11, 850a16.

As well as these terminological divergences, Gerard's version also shows some larger differences from the Greek. (1) In his version the specification is postponed until after the construction, and it is stated in the form appropriate to a theorem rather than a problem: 'I say then that we have now made an equilateral triangle on the given line ab' (*Dico igitur quia iam fecimus triangulum equilaterum super lineam ab datam*). (2) In keeping with this, the proof concludes with the phrase 'And this is what we wished to demonstrate' (*Et hoc est quod demonstrare voluimus*) as opposed to 'which it was necessary to do' in the Greek and the Graeco-Latinus (*hoper edei poiēsai, quod oportebat facere*). (3) In the proof Gerard omits the backward reference 'but it was shown that CA is equal to AB', present in both the Greek and the Graeco-Latinus.

A variety of factors might be invoked to explain such differences between Gerard and the Greek tradition, including (1) redaction by the editor of the Arabic version on which the Latin is based; (2) the process of translation from Greek to Arabic and/or Arabic to Latin; or (3) the possibility that the Arabic text on which Gerard's version is based represents an earlier stage in the textual history of the *Elements* than the Greek manuscript tradition. The relative importance of these factors can only be assessed through a systematic study and comparison of the different versions; the software tools I have described provide a new framework for carrying out such investigations.

The correlation of terminology between different versions is a labor intensive and time consuming task; it would therefore be highly desirable to develop a technological strategy to assist in the process. The basic idea behind such a strategy is simple. Given a source text and translation aligned sentence-by-sentence, a computer should be able to find the term in the translation that is most strongly correlated with the term in the source text. Assuming that the translation is consistent and one-to-one, this should yield the translation of the source term directly. If, for example, *trigonus* appears in every Latin sentence of the translation where *trigōnon* appears in the corresponding Greek sentence and nowhere else, it is a reasonable guess that *trigonus* is the translation of *trigōnon*. Of course it is unrealistic to expect that translations will in fact be consistent and one-to-one in this way; simply looking for term correlations will therefore yield results of only limited usefulness. The technological challenge is to develop statistical techniques that can take account of the fact that a single term in the source text may have multiple translations and, conversely, that a single term in a translation may correspond to multiple terms in the source. Fortunately, a good deal of recent work in the fields of cross-language information retrieval and machine translation has been concerned with just these issues; a number of techniques developed in these areas can be applied directly to the problems of transmission studies.[26] While

[26]The two main methods are so-called Latent Semantic Analysis (LSA) (`http://lsa.colorado.edu`) and the IBM statistical machine translation models as implemented, for example, by the GIZA software package developed at Johns Hopkins University (`http://www.fjoch.com/GIZA++.html`). Both these methods use

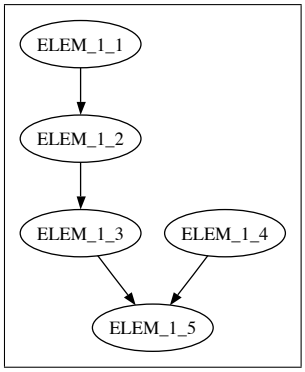

Figure 8: Deductive relationships between propositions 1-5 of book 1.

much more work is needed, preliminary experiments suggest that such methods will yield excellent results for languages such as Greek, Latin, and Arabic, and thus make it possible to analyze a much larger sample of data than would be possible by traditional methods given a realistic view of current scholarly priorities.

3.3 The study of deductive structures

I shall be concerned here with relationships between propositions rather than between the steps within individual propositions. The modes of citing propositions in the *Elements* vary from nearly word-for-word quotation to use without any explicit reference in the text; the identification of proposition use is thus to some extent a matter of interpretation.[27] The identifications in Heath's edition, though a helpful starting point, are far from reliable; I have therefore based the following analysis on the work of Neuenschwander (1973), who provides an exhaustive census of proposition use in books 1-4.[28]

Once the relationships between propositions have been determined, standard software can be used to visualize them in graphical form. For example, the graph in figure 8 shows the relationship between the first five propositions of book 1.[29] Graphs such as these, which

sophisticated statistical procedures to take account of the tendency of translations to deviate from perfect consistency and one-to-oneness, and to match texts with translations at the level of individual words or groups of words. Both methods are purely statistical and do not rely on any specific lexical or syntactic information about the languages in question.

[27]For a detailed study of the various modes of citation in books 1-4 see Neuenschwander (1973, 339-352).

[28]Mueller (1981, 52) accepts Neuenschwander's analysis of proposition usage in books 1-4 as definitive. Cf. Vitrac (1990-2001, 1:513-517).

[29]The graphs in this paper were produced using GraphViz, an open source graph visualization program

can easily be generated for arbitrary collections of propositions as well as entire books, are of obvious utility in the study of deductive relationships. Thus the important place of 1.45 in the deductive structure of book 1 is suggested by its placement at the bottom center of the graph of deductive relationships for that book (fig. 9). Proposition 1.45 shows how to construct a parallelogram 'in' a given angle (i.e. with a given angle between two of its sides) and equal in area to a given rectilineal area. Its importance lies in the fact that it enables any rectilineal area to be represented as a rectangle; the further step of showing that any rectilineal area can be represented as a *square* is taken in 2.14, for the proof of which 1.45 is essential. 1.45 is thus crucial to showing that any rectilineal figure can be 'squared'. From the graph it is evident that 1.45 draws on seven earlier propositions (14, 29, 30, 33, 34, 42, and 44), each of which itself depends on a number of earlier propositions.[30]

Graphs of deductive relationships can also reveal the general character of the deductive organization of particular books or sequences of propositions. Thus the shape of the graph for book 2 (fig. 10) is dramatically different from that for book 1 (fig. 9). This is in part due to the different number of starting points (propositions that make no use of any earlier proposition).[31] In the graphs, starting points are colored light gray and end points (propositions from which nothing is deduced in the book in question) are dark grey. There are exactly two starting points among the propositions of Book 1, 1.1 and 1.4, in relation to the 48 propositions of the book. In book 2, however, there are 19 starting points (most drawn from book 1) for only 14 propositions proved. Books 3 and 4 have a similar ratio of starting points to propositions proved; the data are summarized in the following table:

Book	**Starting points**	**Propositions**
1	2	48
2	19	14
3	25	37
4	34	16

available online at `www.graphviz.org`. In this software, graphs are encoded in the DOT language, in which each vertex is named by a string and edges are indicated by the symbol `'->'`. Thus the connection between propositions 1 and 2 in the above graph would be represented by `ELEM_1_1 -> ELEM_1_2;` and so forth.

[30]Mueller (1981, 16-27) argues that the need to prove 1.45 is the most important consideration in determining the content and order of propositions in Book 1, in the sense that the analysis of the conditions of solution of the problem posed in 1.45 leads naturally back to the earlier propositions and theorems in the book. This is not to deny, of course, the fundamental importance of other results proved in book 1, especially 1.47 (the Pythagorean theorem).

[31]By 'starting point' I mean simply a proposition that is not based on a prior *proposition*; I am not considering the Postulates, Definitions, or Common Notions in this study, though in principle there is no difficulty in doing so.

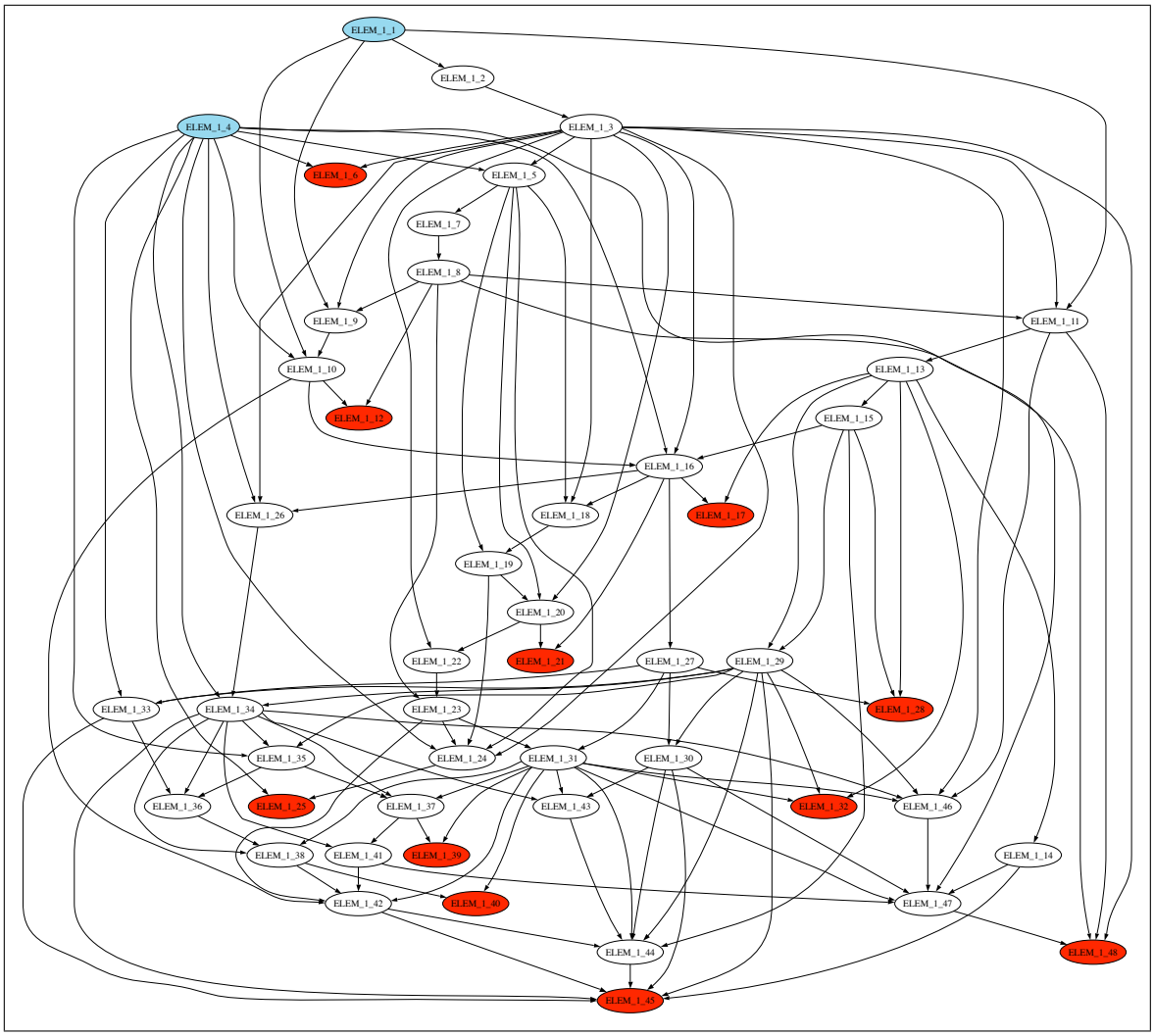

Figure 9: Deductive relationships in book 1. Proposition 1.45 is at the bottom center; starting points are colored in light gray and end points in dark grey.

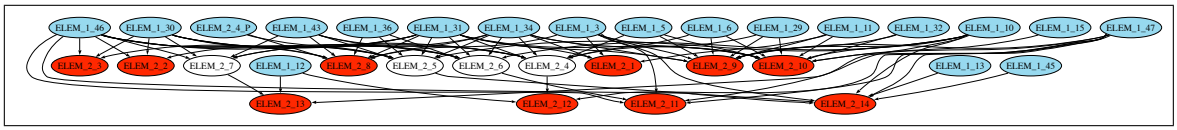

Figure 10: Deductive relationships in book 2.

A further major difference evident from these graphs is the length of the longest path from proposition to proposition; for book 2 the longest path is 2 (e.g. the path joining 1.30, 2.7, and 2.13), whereas the graph for book 1 contains much longer paths (e.g. the path of length 10 joining 1.1, 1.2, 1.3, 1.5, 1.7, 1.8, 1.23, 1.31, 1.43, 1.44, 1.45).

These impressions, derived partly from inspection of the graphs, can be made more precise by applying some graph theoretical methods. In the terminology of graph theory, maps of deductive relationships are *directed graphs*; that is, they are collections of *vertices* (i.e. propositions) and *edges* (i.e. the lines connecting propositions) in which each edge expresses a one-way or directed relationship ('is used in') between vertices. With each graph is associated a unique *adjacency matrix* (A). This is a square $n \times n$ matrix, where n is the number of vertices, whose values are defined as follows: if vertex i is joined to vertex j by an edge in the graph then $A(i, j) = 1$; otherwise $A(i, j) = 0$. For example, the adjacency matrix for the graph of deductive relationships between the first five propositions of book 1 (fig. 8) is:

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

For a given row i of the adjacency matrix, the row vector $A(i,)$ indicates which propositions proposition i is used in. Thus, the vector corresponding to row 1 of the adjacency matrix (0 1 0 0 0) has value 1 only in position 2, because proposition 1 is used only in proposition 2. Similarly, for a given column j of the matrix, the column vector $A(, j)$ indicates which propositions are used in proposition j. The adjacency matrix thus gives a simple method for indexing the use of propositions throughout the *Elements*.

It is also possible to use the adjacency matrix to compute both the length of the longest path through the graph and the number of paths connecting any two vertices.[32] In the case of book 1, this yields a length of 20 for the longest path (between propositions 1 and 45); the maximal number of paths linking any two propositions is a remarkably high 558, again between propositions 1 and 45. As well as confirming the importance of proposition 45 in the deductive structure of book 1, these figures provide a quantitative measure of the

[32]These calculations are based on the following properties of an adjacency matrix A (Chartrand 1985, 217-222). (1) The i, jth element of the matrix A^n gives the number of paths of length n between vertices i and j. (2) Let I be the $n \times n$ identity matrix (i.e. the $n \times n$ matrix in which all elements on the diagonal are 1 and all other elements are 0). Then the i, jth element of the matrix $(I - A)^{-1}$ gives the total number of paths between i and j. The matrix computations in this paper were performed using the R statistical language, freely available online at `http://www.r-project.org`.

distinctiveness of book 1 in comparison to other books. The following table sets out the maximal path length and maximal number of paths between any two propositions in the case of the first four books:

Book	**Longest path**	**Max. no. of paths**
1	20	558
2	2	2
3	6	9
4	5	5

As noted above, indices of proposition usage — giving both the propositions used in a given proposition and the propositions in which it is used — can easily be generated from the data represented in these graphs. Such methods make it possible to investigate the contents of the mathematical 'toolbox', i.e. the set of propositions which are used most repeatedly in a particular book or collection of books. Thus out of the 115 propositions in books 1-4, some 17 (15%) are used in ten or more propositions:

Proposition	**Number of propositions used in, in *Elements* 1-4**
1.3	20
1.4	22
1.5	15
1.8	12
1.10	17
1.11	19
1.12	10
1.13	10
1.23	11
1.29	12
1.30	13
1.31	21
1.32	10
1.34	18
1.46	10
1.47	11
3.1	23

A scatter plot of proposition usage for books 1-4 (fig. 11) reveals that the majority of these frequently used propositions are from the first half of the proposition set. Strictly speak-

ing, the investigation of the toolbox should be based on a complete census of proposition *citations*; Neuenschwander's tables, however, give only the propositions used in a given proposition, and the table above as well as the associated plot is based on his data. To the extent that the number of citations for some of the propositions listed will be higher than the number of propositions they are used in, the table is only a first approximation to the contents of the toolbox. The important point is that further work in the identification of proposition usage could immediately be plugged into the computational framework developed here to yield improved results.

In studying the deductive structure of the *Elements*, it is interesting to consider indirect as well as direct proposition use. By this I mean that in determining which propositions are used by a given proposition, we should include not only those propositions which are directly used, but also those which are implied by the propositions directly used. Thus in fig. 8 above, while propositions 3 and 4 of book 1 are both directly used in proposition 5, propositions 1 and 2 are indirectly used in 5 as well, since 2 is used in 3 and 1 in 2; thus the set of propositions directly and indirectly used by proposition 5 is 1, 2, 3, 4. In terms of the graphical representation of deductive relationships, considering indirectly as well as directly used propositions is equivalent to taking all propositions lying on all paths through the graph that lead to the given proposition. Again, graph theoretical methods provide a convenient method of deriving such information.[33]

When indirect as well as direct use is considered, it turns out that of all the propositions in book 1, 1.45 makes use of the largest number of propositions (34) — a further indication of its importance in the deductive structure of the book. Plotting the number of directly and indirectly used propositions against the proposition number gives a sense of the degree to which each new proposition builds on what has already been established. Figure 12 shows such a plot for book 1. The progression is very close to linear up to 1.25, then drops off abruptly and is much less regular through the remainder of the book. This reflects the fact that the deductive organization of 1.1-25 is very tight: with each new proposition, there is roughly the same increase in the number of propositions used, indicating that each new proposition makes use of most of the results previously established. Proposition 1.26, which is the final result on triangle congruence, makes use only of 1.3, 1.4, and 1.16; proposition 1.27 marks a new start in the book, the beginning of the theory of parallels. Figure 13 shows the result when books 2-4 are added to the picture; it is clear that while there are some stretches where the plot is approximately linear, there is no sequence comparable to 1.1-25 in length. This provides a further indication of the distinctive character of book 1, and especially the first part of the book, in comparison to other books. Still, within each

[33] Let A be an adjacency matrix of direct proposition usage and I the corresponding adjacency matrix of direct and indirect proposition usage. Let $I(,n)$ denote the nth column vector of I, and similarly for A. Then for any n, $I(,n)$ = union of $A(,n)$ and $I(,m)$ for all m such that $A(m,n) = 1$.

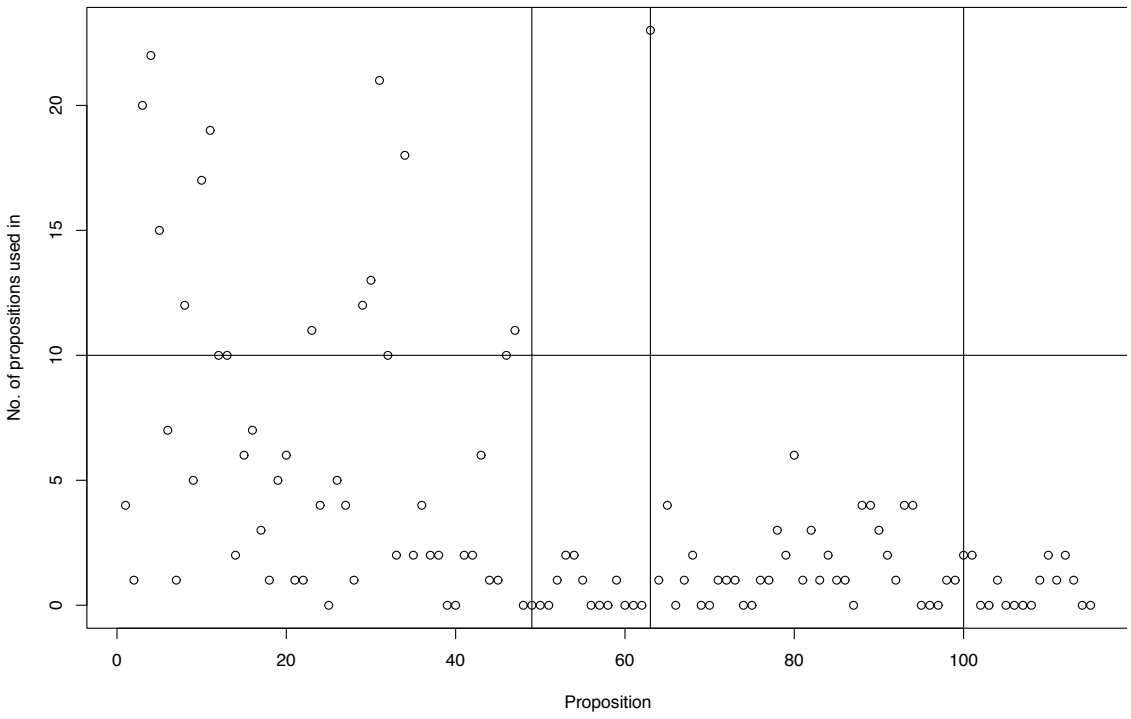

Figure 11: Proposition usage in books 1-4. Vertical lines mark the beginning of books 2, 3, and 4; the horizontal line is the cutoff for inclusion in the table on the previous page.

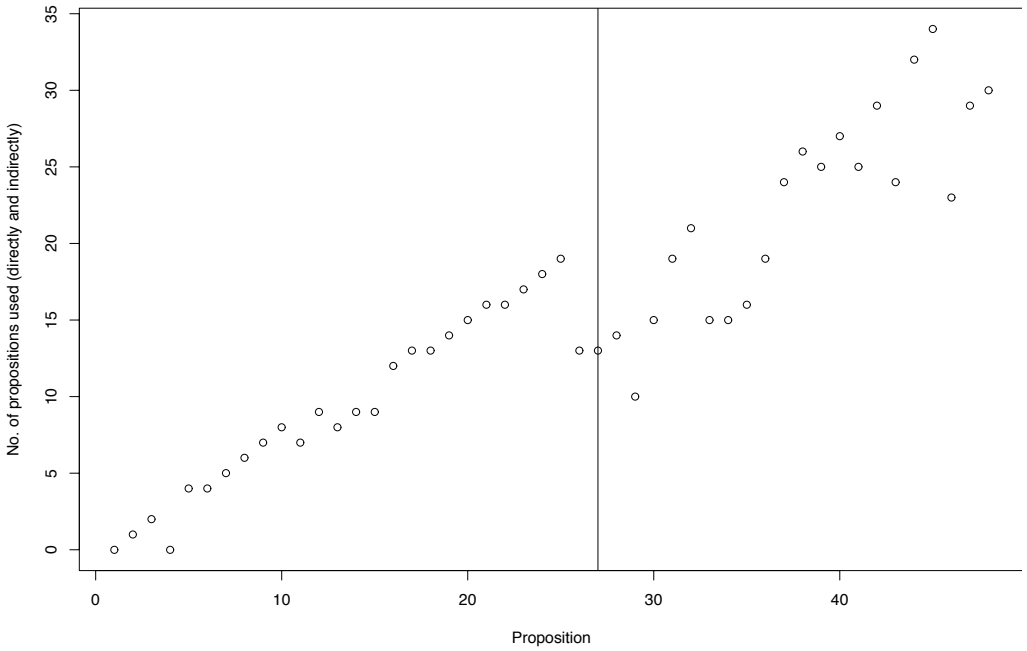

Figure 12: Direct and indirect usage of propositions in book 1. The vertical line marks proposition 1.27.

book, there is an overall tendency for the number of propositions used to increase with the proposition number. Moreover the last proposition of book 4, 4.16, makes use of the largest number of propositions (72) of any proposition in books 1-4. Whatever variation there may be in the organization of particular sequences of propositions, *overall* the trend in the first four books is towards the use of increasing numbers of previously established results.

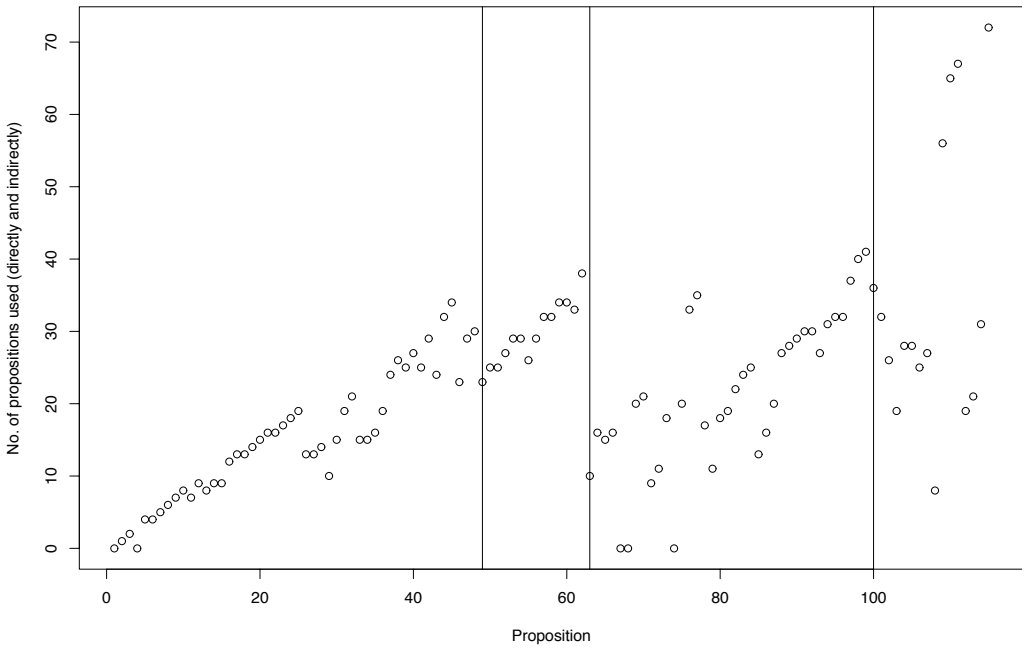

Figure 13: Direct and indirect usage of propositions in books 1-4. Vertical lines mark the beginning of books 2, 3, and 4.

4 Conclusion

These results show that information technology can contribute to the understanding of the *Elements* both by offering new approaches to traditional questions and by raising new questions that cannot easily be addressed using traditional means. Still, it is clear that much work remains to be done in all three of the areas discussed in the previous section. (1) The identification of formulaic language is a special case of the more general problem of term discovery, which has been the focus of a good deal of recent work in the fields of information retrieval and natural language processing. Here the Arboreal software provides the essentials: a basic framework for annotation and sophisticated searching capabilities. But technology has much more to contribute to solving the problem of term discovery, especially in the case of multi-word terms. One approach, which is purely statistical, is based on the co-occurrence of words: the idea is that words that are frequently found together in the same sentence are likely to comprise a single compound term. A second approach involves rule-based systems that specify the basic forms of multi-word terms and the possible grammatical transformations between them.[34] Both of these methods offer the hope of analyzing much larger amounts of data than would realistically be possible even with software that provides advanced searching capabilities. (2) A similar point applies in the area of transmission studies, where new techniques for the automatic correlation of terminology across different languages have shown promising results (see above, end of section 3.2). (3) In the study of deductive structures, further work is needed both to improve and extend the data set (e.g. recording proposition citations and including more books of the *Elements* as well as other mathematical texts) and to develop and evaluate new techniques for analyzing the data. With a larger data set, it may be possible to characterize precisely the differences in deductive organization between different mathematical texts, to examine how such organization changes over time, and possibly to develop criteria for judging the authenticity of disputed texts.

Each of these problems —the discovery of technical terminology, the correlation of terminology between different languages, and the representation and analysis of networks such as deductive structures — is both highly general and relevant to the study of a wide range of sources in the history of science, not just the *Elements* or other mathematical texts. The present paper thus stands as a case study of the use of information technology in the analysis of sources in the history of mathematics and science. It reveals a twofold challenge, one to technology and one to scholarship. The key challenge to technology lies in the

[34] A major problem for the first method (based on co-occurrence of terms) is synonymy (multiple terms with the same sense) and polysemy (a single term with multiple senses). A promising approach to addressing these problems lies in application of the technique of Latent Semantic Analysis (LSA) (`http://lsa.colorado.edu`). For an example of the second, rule-based approach to term discovery, see Jacquemin (2001).

development of software that maximizes interactivity, making it possible to evaluate the results provided by automatic methods and to subject the results to further analysis. The challenge to scholars is to make full use of the range of tools that now go some way towards meeting this requirement and to contribute to the further development of such tools. The potential certainly exists for information technology to deepen and even to transform our understanding of historical sources, but only when scholars seize the opportunities that technology offers will this potential come closer to being realized.[35]

References

Acerbi, F. (2003). Drowning by multiples: remarks on the fifth book of Euclid's *Elements* with special emphasis on prop. 8, *Archive for History of Exact Sciences* **57**: 175–242.

Aujac, G. (1984). Le langage formulaire dans la géometrie grécque, *Revue d'Histoire des Sciences* **XXXVII**: 97–109.

Brentjes, S. (1994). Textzeugen und Hypothesen zum arabischen Euklid in der Überlieferung von al-Haggag b. Yusuf b. Matar (zwischen 786 und 833), *Archive for History of Exact Sciences* **47**: 53–92.

Brentjes, S. (2001). Observations on Hermann of Carinthia's version of the *Elements* and its relation to the Arabic transmission, *Science in Context* **14**: 39–84.

Busard, H. L. L. (1984). *The Latin Translation of the Arabic Version of Euclid's Elements Commonly Ascribed to Gerard of Cremona*, Leiden: Brill.

Busard, H. L. L. (1987). *The Mediaeval Latin Translation of Euclid's Elements Made Directly from the Greek*, Vol. 15 of *Boethius*, Wiesbaden: Steiner.

Chartrand, G. (1985). *Introductory Graph Theory*, New York: Dover.

DeYoung, G. (1981). *The Arithmetic Books of Euclid's Elements in the Arabic Tradition: an Edition, Translation, and Commentary*, PhD thesis, Harvard University, Cambridge, Mass.

DeYoung, G. (1984). The Arabic textual traditions of Euclid's *Elements*, *Historia Mathematica* **11**: 147–160.

[35]This paper was completed during a sabbatical leave at the Institute for Advanced Study in Princeton, New Jersey; I am deeply grateful to the Institute for granting me a Martin L. and Sarah F. Leibowitz Membership for Fall term 2006.

Engroff, J. W. (1980). *The Arabic Tradition of Euclid's Elements, Book V*, PhD thesis, Harvard University, Cambridge, Mass.

Heath, T. L. (1956). *The Thirteen Books of Euclid's Elements*, New York: Dover. 3 vols.

Heiberg, J. L. & Menge, H. (1883-1916). *Euclidis Opera Omnia*, Leipzig: Teubner. 8 vols.

Jacquemin, C. (2001). *Spotting and Discovering Terms Through Natural Language Processing*, Cambridge, Mass.: MIT Press.

Knorr, W. R. (1996). The wrong text of Euclid: On Heiberg's text and its alternatives, *Centaurus* **38**: 208–276.

Mueller, I. (1981). *Philosophy of Mathematics and Deductive Structure in Euclid's Elements*, Cambridge, Mass.: MIT Press.

Murdoch, J. (1966). Euclides Graeco-Latinus, a hitherto unknown medieval Latin translation of the *Elements* made directly from the Greek, *Harvard Studies in Classical Philology* **71**: 249–302.

Netz, R. (1999). *The Shaping of Deduction in Greek Mathematics*, Cambridge: Cambridge University Press.

Neuenschwander, E. A. (1973). *Die ersten vier Bücher der Elemente Euklids: Untersuchungen über den mathematischen Aufbau, die Zitierweise und die Entstehungsgeschichte*, Zürich: Universitätsdruckerei H. Stürtz. Offprint from *Archive for History of Exact Sciences*, Volume 9, Number 4/5, 1973, pp. 325-380.

Vitrac, B. (1990-2001). *Les Éléments. Traduction et commentaires par Bernard Vitrac*, Paris: Presses Universitaires de France. 4 vols.