##DIGITAL CLASSICS ONLINE##

Digital Texts and Diagrams: Representing the Transmission of Euclid's Elements

Christine Roughan

Abstract: The Digital Euclid project aims to publish an open, digital edition of every extant witness to the text and diagrams of Euclid's *Elements*. This paper discusses the required groundwork and is divided in two parts. It first covers a survey of the surviving manuscript and print sources for the *Elements* that intends to identify the extent of these materials, how many of these works have already been digitally imaged, and what challenges they pose to current data extraction methods. The latter part of the paper discusses the methods used to produce machine-actionable texts and diagrams and focuses especially on the development of tools for the identification and extraction of diagrammatic data.

Introduction

The *Elements* of Euclid is a text which has received continuous publication and use since it was first authored in the third century B.C.E. This geometrical text appears in hundreds of manuscripts; combined its manuscript and print editions number well over one thousand and span languages from across the globe.

The many and evolving forms that the *Elements* has taken throughout its lengthy transmission history have been a challenge to detail in their entirety. This is a text which has been well-studied: I. L. Heiberg for instance provides the current critical edition of the Greek text,[1] and editions even exist for specific translations of the *Elements* as well (consider H. L. L. Busard's critical editions of various medieval Latin translations).[2] The original Greek, after all, is only one form that the text has taken, and it cannot answer any questions about how Euclid was read and understood in — for example — the medieval Latin West. The sheer volume of material leaves the transmission of this text difficult to navigate and grasp in its entirety. In the case of just one translation within this transmission, Busard himself states that „131 manuscripts of Campanus' version of Euclid's *Elements* are known. Thus it was impossible to collate all of them."[3]

When dealing with a large set of varied texts, print is not the ideal medium to convey it in its entirety. Navigating and analyzing such a dataset of the *Elements* in print would be slow and unwieldy, requiring thousands of pages. A second limitation of print appears in the case of the diagrams. Each proposition of the *Elements*, after all, is accompanied by a mathematical diagram, which itself contains essential information that cannot always be gleaned from the

[1] Heiberg (1883–6).

[2] Busard (1968), (1983), (1984), (1987), (1992), (1996), (2001), (2005).

[3] Busard (2005), S. 46.

text alone.[4] Variation appears during the transmission of these figures just as it does during the transmission of the text; however, they have received little scholarly attention. Until recently, in the many critical editions of Greek mathematical texts, there have been no *apparati critici* detailing the diagrams — Reviel Netz is perhaps the first to take steps towards such an apparatus by providing written explanations of variation between Archimedes manuscripts and occasional thumbnails showing alternate diagrams.[5] This is an important start, but this approach does consume space rapidly, especially when attempting to provide alternate diagrams from a transmission as wide as that of the *Elements*.

The Digital Euclid project takes a digital approach to the transmission of the *Elements* in order to present for scholarly access and reuse the text and diagrams from these many witnesses. What would take thousands of pages in print and either lengthy or incomplete *apparati critici* can be published much more efficiently in an electronic format.

The project takes its cues from University of Leipzig's Open Greek and Latin Project, which will publish at least one version of all extant Greek and Latin sources and which ultimately aims to represent every surviving version of these texts.[6] The Digital Euclid project keeps in mind the latter goal and strives to represent every edition of the *Elements* in a form that is open, machine-actionable, and annotated. This paper discusses the initial work that is necessary for a project of this kind. It has two parts: firstly, a survey of the manuscript and print transmission of Euclid's *Elements*; secondly, the testing of extant tools and the development of new ones to aid in digitization.

A survey on the transmission of the elements

The survey comprises various editions, translations, revisions, and recensions of the *Elements*, as well as its adaptions into school texts. Commentaries have not yet been included. For the most part, separate works which only quote the *Elements* are not included – the current exceptions to this are texts which preserve fragments of Boethius's Latin translation.

When considering the extent of the survey, one must acknowledge that only a portion of material transmitted as Euclid's *Elements* can be attributed to him (the apocryphal Books XIV and XV were once considered Euclidean, for example). However, the Digital Euclid project does include this material in its entirety, not solely that which scholars presently attribute to Euclid. The goal of doing so is to provide a fuller picture of how the *Elements* was read and understood throughout history. A side result of this is that the survey does include manuscripts and texts that contain only apocryphal material.

Multiple authors have produced bibliographies of Euclid's works, including Pietro Riccardi (1887), Georges J. Kayas (1977), and Max Steck (1981). These bibliographies are of varying comprehensiveness; the Digital Euclid survey takes Riccardi's bibliography as its starting point, which offers a very thorough listing of print editions of the *Elements* up through 1887. This date is also convenient for the purposes of the project, as material published before 1887 is public domain and is therefore available for scholarly reuse and republication. The list totals

[4] Saito (2009), S. 817.

[5] Netz (2004).

[6] http://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/.

over a thousand editions. Riccardi also describes more than 180 Euclidean manuscripts, but notes that this is an incomplete list. The project therefore turned to manuscript lists provided by other scholars (such as the formerly mentioned Heiberg and Busard, as well as Folkerts and Lo Bello, etc.)[7] as well as manuscript catalogs.

The goals of the survey were threefold. For each extant version, the survey would:

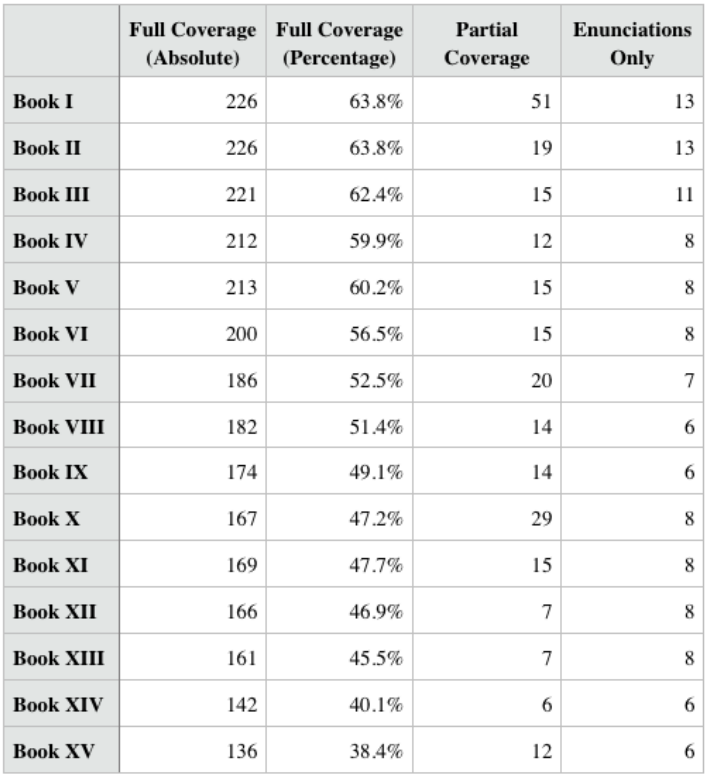
1) record the relevant bibliographic data and assign identifiers both to the physical codex or book and the abstract work
2) note the current status of digitization, current copyright on the book itself, and – where applicable – current copyright on digital images of the version in question
3) record factors that could impact automated text and diagram extraction workflows (page layouts, languages or fonts, and quality of the page images)

The survey covered 415 scanned printed editions and 477 manuscripts, spanning fourteen languages and the first through the nineteenth century C.E.

Manuscript editions in the survey

The survey presently contains 477 codices, papyri, and fragments, which appear in ten languages and date from the first through the eighteenth century C.E. These materials preserve 522 distinct versions and translations of the *Elements* text, 84 of which are Greek, 51 of which are Arabic, and 344 of which are Latin. The Latin portion of the survey is currently the most complete, followed by the Greek. Other languages include French and Middle French, Hebrew, Italian, Modern Greek, Persian, and Turkish.

The coverage of the *Elements* is known for 354 of these texts; perhaps unsurprisingly, the books that receive the most coverage are Books I and II (the tendency of geometrical manuscripts to provide Book I's definitions results in Book I also receiving the most partial coverage). Figure 1 provides an illustration of coverage across books. With only two exceptions, later books appear less often than the ones preceding them. Nevertheless, between the ninth and the eighteenth centuries C.E. even Book XV appears in five languages and over 130 manuscripts — more than enough examples of a lengthy and far-ranging transmission.



	Full Coverage (Absolute)	Full Coverage (Percentage)	Partial Coverage	Enunciations Only
Book I	226	63.8%	51	13
Book II	226	63.8%	19	13
Book III	221	62.4%	15	11
Book IV	212	59.9%	12	8
Book V	213	60.2%	15	8
Book VI	200	56.5%	15	8
Book VII	186	52.5%	20	7
Book VIII	182	51.4%	14	6
Book IX	174	49.1%	14	6
Book X	167	47.2%	29	8
Book XI	169	47.7%	15	8
Book XII	166	46.9%	7	8
Book XIII	161	45.5%	7	8
Book XIV	142	40.1%	6	6
Book XV	136	38.4%	12	6

Abb. 1: Coverage of the Elements in 354 manuscript versions of the text.

[7] Folkerts (1989), Lo Bello (2003).

As of July 2015, 112 of these manuscript texts have been at least partially imaged and made available online; 82 have been imaged in full. Of the texts with complete digital publication, 59 are available either as public domain images or under Creative Commons licenses that would allow for scholarly reuse. These 59 manuscript texts still provide a wide selection of *Elements* editions and translations: they represent examples from the first to the seventeenth century C.E. and span six languages (although the overwhelming majority are Latin). Most of these digital manuscripts are provided by the Münchener Digitalisierungszentrum and Gallica.[8] A little over half of these are presented in greyscale, but the remainder are full-color images. Nearly all of the manuscripts provided by Gallica are digitized microfilm copies.

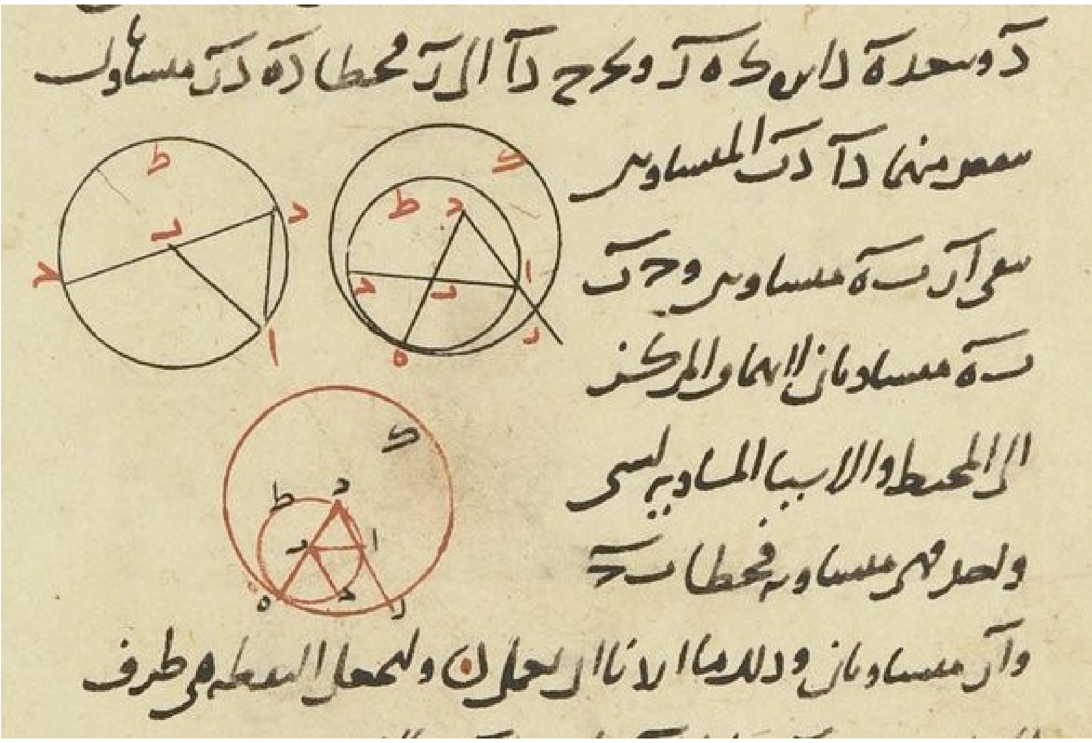
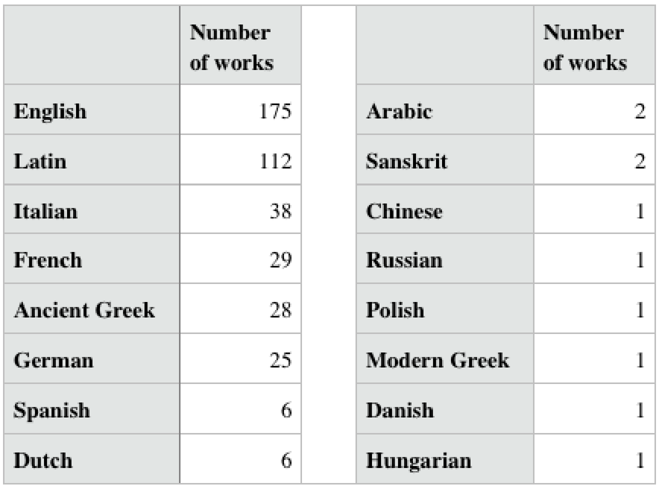


Abb. 2: A selection of folio 145v of BnF Arabe 2484, showing diagrams relevant to an Arabic version of Book I, proposition 2. Source: gallica.bnf.fr.

In surveying these manuscripts, the Digital Euclid project determined some categories in which to organize the mathematical diagrams. The project encountered three types of layouts: inline, where the diagrams appear within the margins of the main text; marginal, where they remain outside these borders; and a combination, where they appear in both spaces throughout the work. In this initial survey, the majority of the manuscripts viewed contain diagrams in both locations. Inline diagrams appear second most often. Where full-color texts were available, the Digital Euclid project also recorded when differently colored inks were used between texts and diagrams with the hope that this distinction might aid in automated attempts to locate the diagrams. The overwhelming majority uses ink that is the same color, but not all: figure 2 offers one example. The diagrams are drawn with both red and brown ink, while text is in a brown ink. The diagram layout illustrated here is inline.

[8] MDZ: http://www.digitale-sammlungen.de/. Gallica: http://gallica.bnf.fr/.

Print editions in the survey

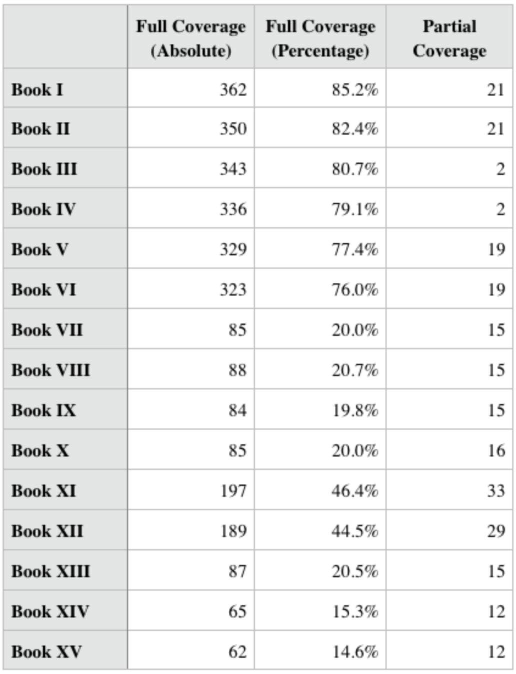


	Number of works		Number of works
English	175	Arabic	2
Latin	112	Sanskrit	2
Italian	38	Chinese	1
French	29	Russian	1
Ancient Greek	28	Polish	1
German	25	Modern Greek	1
Spanish	6	Danish	1
Dutch	6	Hungarian	1

Abb. 3: Languages represented in the Digital Euclid survey of print editions as of July 2015.

Riccardi lists over one thousand publications of the *Elements* between 1482 and 1887. As of July 2015, the project's survey has covered 614 unique examples of the various editions, or 415 unique editions. Several of these editions either were multivolume works or contained multiple translations, so the total number of abstract works surveyed comes to 429. All of these have been imaged and made available online: the project's initial focus was on works that can be located in one of four databases: Google Books (contains 45.4% of the 614 versions), the Internet Archive (38.7%), HathiTrust (30.0%), and SLUB Dresden (13.4%).[9] The works contained in these databases provide a wide selection already: they date from 1482 to 1908 and span sixteen languages.

Figure 3 outlines the distribution of works across languages currently in the survey. The counts consider not physical books, but abstract versions of a text (and so treat a text published across multiple volumes as one). While Euclid has certainly been published far more often in English than in Modern Greek, for example, the lack of representation in the survey for certain languages is partially caused by the focus so far on scanned copies present in the four databases mentioned above.



	Full Coverage (Absolute)	Full Coverage (Percentage)	Partial Coverage
Book I	362	85.2%	21
Book II	350	82.4%	21
Book III	343	80.7%	2
Book IV	336	79.1%	2
Book V	329	77.4%	19
Book VI	323	76.0%	19
Book VII	85	20.0%	15
Book VIII	88	20.7%	15
Book IX	84	19.8%	15
Book X	85	20.0%	16
Book XI	197	46.4%	33
Book XII	189	44.5%	29
Book XIII	87	20.5%	15
Book XIV	65	15.3%	12
Book XV	62	14.6%	12

Abb. 4: Coverage of the Elements in 425 print versions of the text.

In comparison to the manuscript situation, a far more dramatic preference for certain books crystalizes during the print transmission. Again, Book I takes the lead, appearing in 85.2% of the surveyed texts. More notably, Books I–VI all appear in more than 75% of these texts, Books XI and XII in around 45%, and the rest in approximately 20% or less. The first six books were published frequently, often with the eleventh and twelfth attached. The survey reveals that the prevalence of this combination varies across languages: about two thirds of the surveyed English texts comprise Books I–VI, sometimes with Books XI and/or XII added. The same is true for less than half of the Latin texts and less than a fifth of the Greek ones.

[9] Google Books: https://books.google.com/; Internet Archive: https://archive.org/; HathiTrust: https://www.hathitrust.org/; SLUB Dresden: http://www.slub-dresden.de/startseite/.

As noted, all of the print editions included thus far in the survey have already been made available as digital scans, and all surveyed materials from the Internet Archive, Google Books, HathiTrust, and Dresden SLUB are public domain. The majority of the scans are provided already binarized: 73.9% in Google Books, 87.9% in HathiTrust, and 77.3% in the Internet Archive (Dresden SLUB provides full-color scans).

During the survey, four potential manuscript layouts were recorded. Like manuscript diagrams, print diagrams might have inline or marginal layouts. A third option is provided by those texts that consistently locate the diagram at the top of the page. Lastly, many print editions did not locate the diagrams beside their proposition, but rather placed diagrams together on pages that folded out, usually from the back of the book. The layout represented most often in the survey is inline, appearing over 70% of the time. Foldout layouts are the second most common and are used in about 15% of the editions. Very few editions use marginal or top-of-page layouts.

The Digital Euclid project found one major issue in the scanning process for texts with foldout diagrams: across 138 scans of books with that layout, only 6 of them actually imaged the diagrams. This is a serious omission. While the overwhelming majority of printed diagrams are imaged successfully because they appear inline with the text, the failure to image diagram foldout pages means that digital methods currently cannot be used to analyze an entire tradition of print diagrams.

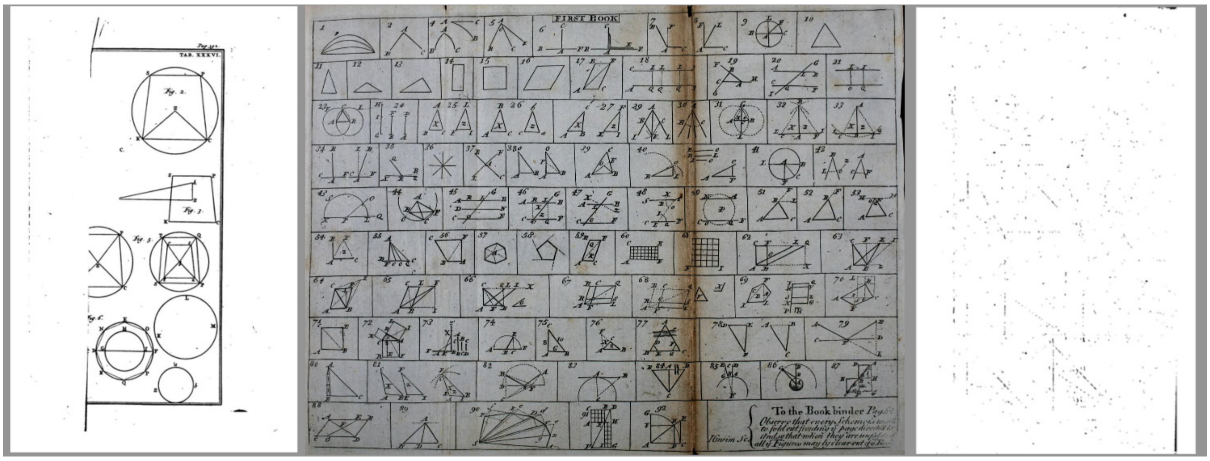


Abb. 5: Three examples of scanned diagram foldouts. The leftmost page was partially imaged, but not unfolded. The middle is an example of an unfolded page. The rightmost page remained folded, and nothing of the diagrams was captured. Source for all three images: the Internet Archive.

Final remarks on the survey

Work with the survey, even in its initial form, indicates that it is a useful dataset: Digital Euclid has made great use of it during development of appropriate workflows for data extraction across the *Elements*. As this survey approaches completion, it is hoped that it can be of continuing use. With this goal in mind, a portion of the survey will be published as part of the Perseus Catalog, a digital catalog that unites classical bibliographies and metadata.[10] The full dataset of the survey is located in the Github repository for the Digital Euclid project.[11]

[10] http://catalog.perseus.org/.

[11] https://github.com/cmroughan/digital-euclid.

The survey already shows comparative potential, as demonstrated by the quick analyses on each book above. As more texts are transcribed and the details of their content are described by CTS URNs (which are discussed below), it will become possible to investigate questions like these with far deeper levels of granularity.

This year's usage of the survey has also suggested several possibilities for expansion. Continuing along the coverage example, it would be beneficial to note where lacunae are caused by a physical absence of material (and therefore it is unknown whether or not the absent portion of the text was originally included or not) and where the absent text is recognized to have never been included in the first place. Company is also a feature that would be illuminating to trace through the history of a text — how often does the *Elements* appear alongside another particular text? How is this affected across time, languages, and the transition from manuscript to print? In a similar vein, when do scholia or other scholarly commentaries accompany the text? Another dimension could eventually be added to the survey by noting scholarship on the editions; for example, to whom is a manuscript attributed? Has this answer changed over time? In the future, as the survey is completed for versions of the *Elements*, it will come to comprise further works. The Digital Euclid project will include commentaries on the *Elements* in later versions, and other Euclidean texts are likely eventual candidates as well.

Data Extraction Workflow in the Digital Euclid project

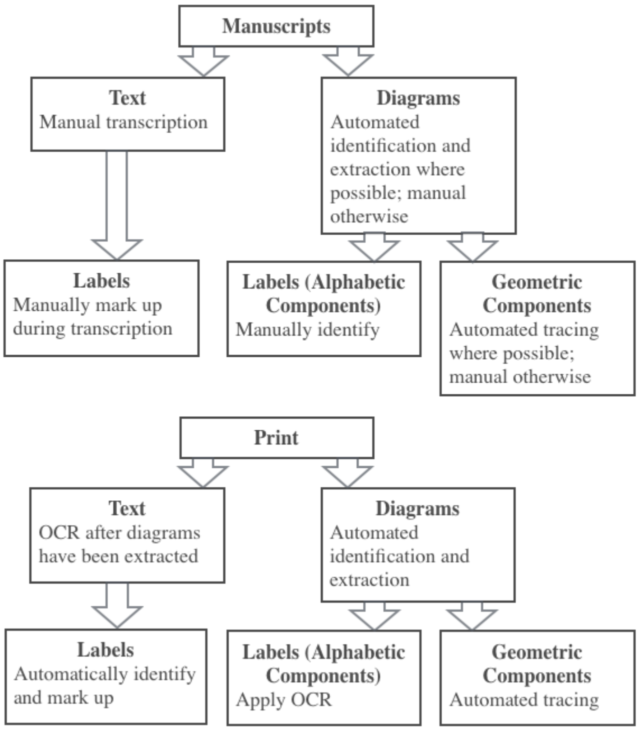

Manuscripts
Text
Manual transcription
Diagrams
Automated identification and extraction where possible; manual otherwise
Labels
Manually mark up during transcription
Labels (Alphabetic Components)
Manually identify
Geometric Components
Automated tracing where possible; manual otherwise
Print
Text
OCR after diagrams have been extracted
Diagrams
Automated identification and extraction
Labels
Automatically identify and mark up
Labels (Alphabetic Components)
Apply OCR
Geometric Components
Automated tracing

Abb. 6: Workflows for extraction of textual and diagrammatic data from manuscript and print editions.

Like the Open Greek and Latin Project, Digital Euclid has as its end goal the publication of digital, annotated EpiDoc- and CTS-compliant XML texts. The Digital Euclid project additionally notes the importance of the mathematical diagrams, and so aims to produce digital, annotated SVG traces of the mathematical figure.

With a multitude of print editions and manuscripts revealed by the survey to have been digitally imaged already, Digital Euclid began to develop workflows and to test tools for the extraction of both textual and diagrammatic information. The second part of this paper will discuss these efforts.
The different challenges that manuscripts and print editions pose necessitate two separate approaches. This is already accepted in the case of texts: optical character recognition (OCR) for manuscripts still has a significant way to go. Manuscript diagrams also pose slightly different challenges than print ones, as will be discussed below. Figure 6 outlines the two workflows.
Like the survey of *Elements* editions, the resultant transcriptions and traces are being published in the project's Github repository.

Regarding the Text of the *Elements*

The Digital Euclid project will produce texts that are machine-actionable and CTS- and EpiDoc-compliant, following the lead of Open Greek and Latin. CTS URNs serve as a means of identifying texts and selections of those texts unambiguously; one advantage this provides is that CTS texts are automatically aligned.[12] EpiDoc is a subset of the Text Encoding Initiative's guidelines that is especially suitable for primary sources.[13]

Two tags used by the Digital Euclid project in the markup of the text are worthy of note, since they highlight information particular to mathematical texts. Firstly, because the labels function as identifiers linking text and diagram, these receive referencing string tags. A label or cluster of labels is tagged as `<rs type="labels"></rs>`, making it easy to point a computer either towards a structure that is identified with multiple labels or towards the individual point or shape that corresponds to a single label (once these are automatically generated in the diagram portion of the workflow, that is.). By looking between a list of labels or label groupings in the text and labelled points or objects in the diagram, a computer can automatically create explicit links between the text and the diagram. It must be recognized that this automated method cannot work where there are disagreements between text and diagram (which might occur due to an error within one or the other). The second element used in the Digital Euclid project is the `<figure/>` tag. This marks the existence of a diagram in the text that accompanies the section in question, even if the diagram is not physically near that section, as would be the case for foldout diagrams. Figure tags reference the unique identifier assigned by the project for the diagram in question.

The question of how to extract text from page images has already received a great deal of work, and the current answer — optical character recognition (OCR) — is generally sufficient for the many printed texts of the *Elements*. Certain editions do contain mathematical notation (and these are noted in the survey) and so require OCR that can handle those characters. Such software does exist — a math detection module has been developed for Tesseract OCR, for example.[14] Manuscript OCR is an area of ongoing research, but for now, manual transcription (usually accomplished through large citizen science projects) has proved to be successful, if slower.

[12] http://cite-architecture.github.io/ctsurn/overview/.

[13] https://sourceforge.net/p/epidoc/wiki/Home/.

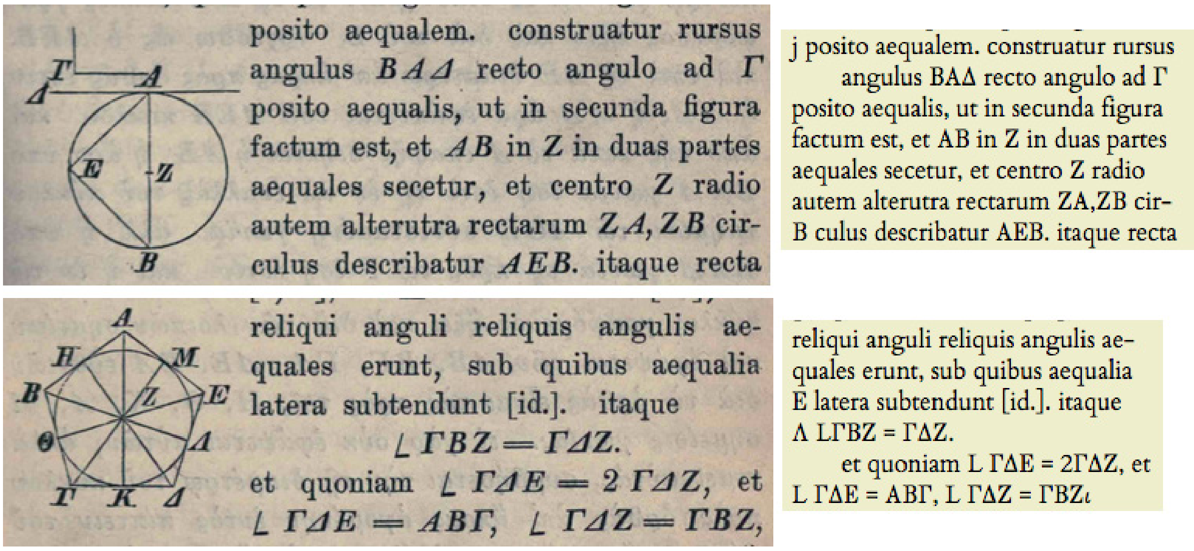
[14] https://github.com/tesseract-ocr.

Markup can also be accomplished automatically or manually, depending on the text in question. For manuscript editions, the labels would be tagged as such during the manual transcription process. For printed texts, labels are usually distinct enough to be automatically tagged, although this process does require review. In most cases, a group of capitalized letters that are separated from the others by periods or spaces can be identified as a group of labels, especially when they do not spell out a word in the language of the text. Figure elements can be added to the text's XML once diagrams and text have been aligned.

Regarding the Diagrams of the *Elements*

The major challenge to automated data extraction in the case of Euclid's *Elements* comes in the form of the mathematical diagrams. These are worthy of study in their own right: they might contain information not in the text, reveal insights into their function in ancient and medieval times, or offer their own clues regarding the process of manuscript creation and transmission.

The data, the diagrams themselves contain thus makes it desirable to produce digital and annotated traces of diagrams similar to the digital and annotated transcriptions of text. Since OCR is able to rapidly obtain textual data from page images, the lack of similar tools to handle diagrams is a major speed bump in the digitization and extraction process for a complete edition. Furthermore, the diagrams are best located before any OCR of the text takes place: the alphabetic characters[15] — or simply geometric components that a computer might mistake for a character — can be picked up during OCR and introduce errors into the resultant output, as seen in Figure 7. It is therefore worthwhile to seek automated means of identifying diagrammatic regions on a page or folio, both to remove non-textual data before later OCR workflows and to link images of the diagram with the appropriate proposition.



posito aequalem. construatur rursus angulus BAΔ recto angulo ad Γ posito aequalis, ut in secunda figura factum est, et AB in Z in duas partes aequales secetur, et centro Z radio autem alterutra rectarum ZA, ZB circulus describatur AEB. itaque recta

j posito aequalem. construatur rursus
angulus BAΔ recto angulo ad Γ
posito aequalis, ut in secunda figura
factum est, et AB in Z in duas partes
aequales secetur, et centro Z radio
autem alterutra rectarum ZA,ZB cir-
B culus describatur AEB. itaque recta

reliqui anguli reliquis angulis aequales erunt, sub quibus aequalia latera subtendunt [id.]. itaque

∟ ΓBZ = ΓΔZ.

et quoniam ∟ ΓΔE = 2 ΓΔZ, et ∟ ΓΔE = ABΓ, ∟ ΓΔZ = ΓBZ,

reliqui anguli reliquis angulis ae-
quales erunt, sub quibus aequalia
E latera subtendunt [id.]. itaque
Λ LΓBZ = ΓΔZ.
et quoniam L ΓΔE = 2ΓΔZ, et
L ΓΔE = ABΓ, L ΓΔZ = ΓBZι

Abb. 7: Screenshots of OCR output from the Lace Greek OCR project website. In the output for the first example, the 'j' at the start of the first line and 'B' in the last were introduced from the diagram. The second includes 'Λ' at the start of the fourth line. Source: heml.mta.ca/lace.

[15] The mathematical diagrams in the Elements, like many Ancient Greek mathematical figures, contain both alphabetic and geometric content. Most, though not all, consist of an arrangement of shapes and the alphabetic labels that identify them.

Several layout analysis tools were tested to determine whether extant software would be capable of distinguishing diagrams from text. Digital Euclid first investigated the open source document analysis system OCRopus and the layout analysis component of the Tesseract OCR system.[16] Neither were successful in recognizing diagrams.

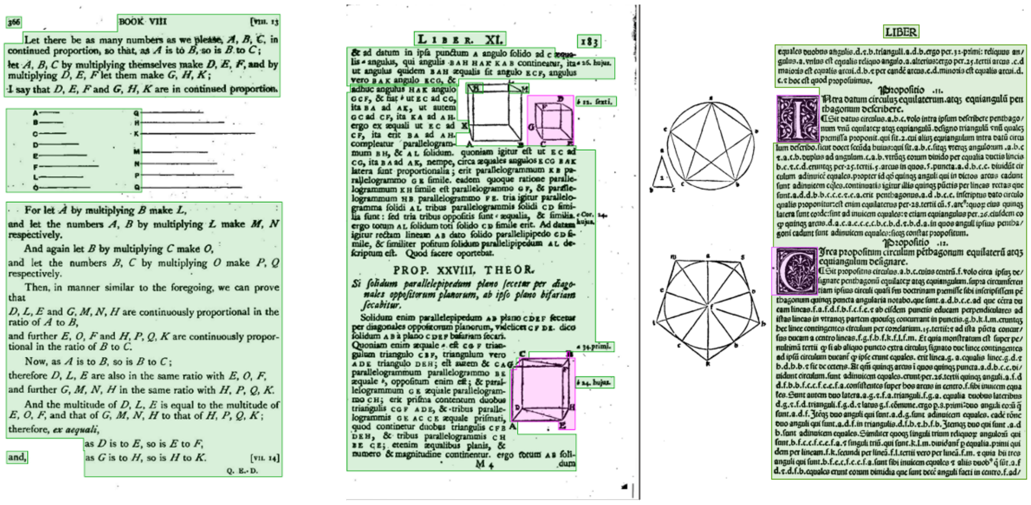


Abb. 8: Three sample pages with incorrect layout analysis by ABBYY FineReader. Source for all three original images: the Internet Archive.

The project then tested out ABBYY FineReader[17] and found that while the program did have some success, there was considerable room for improvement. The software was tested on 93 diagrams, in which were represented all varieties of diagram layout except for foldout, and it properly identified 18 of them (a further 11 were poorly or incompletely captured).

While this initial sample was small, the output ABBYY FineReader provided from it was illuminating. As the rightmost page in figure 8 illustrates, the program often returned false positives when drop caps were present, and this is to be expected: the program was not trained to distinguish between different kinds of 'illustrations'. Additionally, in the case of line diagrams like the ones seen on the leftmost page of figure 8, only 7.4% were recognized. ABBYY FineReader was inconsistent in how it handled the diagrams' labels and showed no preference towards handling them as text, handling them as part of an illustration, or skipping them entirely.

Identification and Extraction of Diagram Data

Since extant tools proved inapt to handle the mathematical diagrams, new ones were necessary. The Digital Euclid project therefore used the Gamera framework to build tools that would recognize diagrams in the *Elements*.[18] Automated diagram recognition posed a different challenge than automated character recognition. Texts are composed of a limited set of characters, in many languages separated from each other by whitespace — it therefore makes sense to train a computer to locate connected components and recognize the individual characters. Diagrams can contain both geometric and alphabetic content and can comprise multiple elements

[16] OCRopus: https://github.com/tmbdev/ocropy, Tesseract: https://github.com/tesseract-ocr.

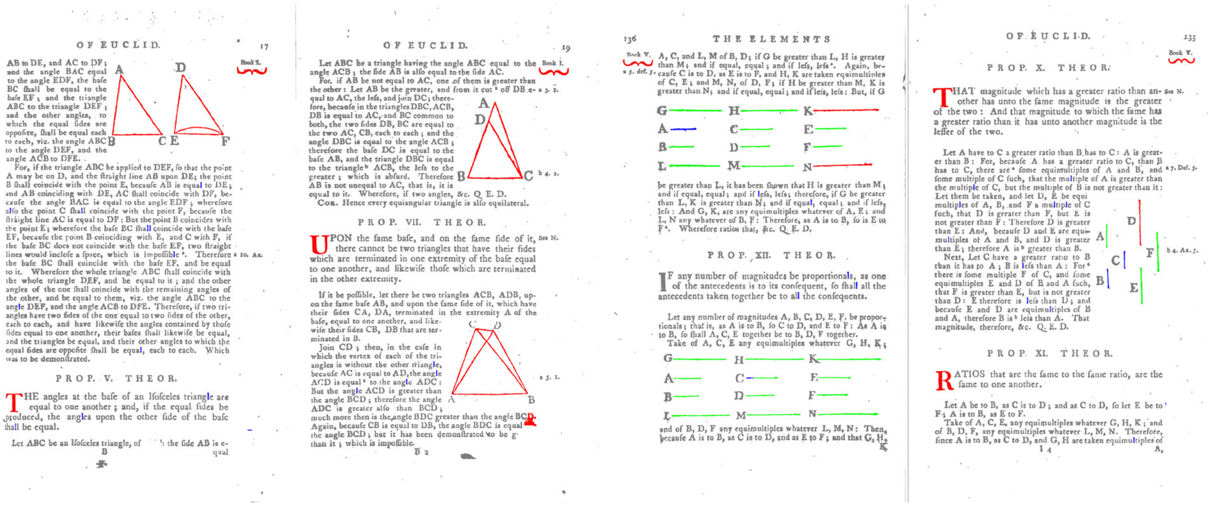
[17] http://www.abbyy.com/finereader/.

[18] http://gamera.informatik.hsnr.de/.

separated by whitespace; they are more akin to a larger structure of text such as a word or a line rather than a character. Furthermore, the connected components that form the geometric portion of the diagram vary dramatically, with many being unique.
The constructed tools currently used by the project work best when input is provided regarding the page layout of the diagrams. Diagrams that are consistently located in the margins, for example, or at the top of the page can be identified easily. Some editions have diagrams consistently located on the outer edge of the page, which again simplifies the search. Editions with inline or mixed-layout diagrams are more complicated to handle.

The Digital Euclid project has also found success by distinguishing between two types of connected components, here termed ‘line’ and ‘quad’. The project considers line components to be those portions of a diagram whose bounding box fits the profile of a line, exceeding certain aspect ratio thresholds. Quad components are bounded by boxes with less extreme aspect ratios and which exceed certain area thresholds. In figure 8, the diagram on the leftmost page is composed of line components, while the rest consist of quad components.
The first challenge is how to identify lines reliably when 1) line length varies, so there are no consistent dimensions to look for and 2) line cross-sections (since the printed line is not truly one-dimensional), are not consistent across different editions. Digital Euclid first identifies ‘definite lines’[19] as connected components that surpass one of two extreme aspect ratios corresponding to horizontal or vertical orientations. Meanwhile, ‘possible lines’ are identified as connected components that surpass more moderate aspect ratios. This first pass records all possibilities, as well as data on the definite lines. The second pass learns from the first and considers various factors (heights of definite horizontal lines, widths of definite vertical ones, and nearest neighbors for each possibility) to remove false positives from the initial results.

There are some remaining issues with lines. Accidental intersection between lines and labels occurs often, causing the bounding boxes to not match the profile for a line. Future work will test projection analysis as a means of identifying these problematic components. Dotted and other non-continuous lines are also an issue. The segments of broken lines still usually result in matches, but dotted lines must be handled manually.



OF EUCLID. 17

Book I.

AB to DE, and AC to DF; and the angle BAC equal to the angle EDF, the baſe BC ſhall be equal to the baſe EF; and the triangle ABC to the triangle DEF; and the other angles, to which the equal ſides are oppoſite, ſhall be equal each to each, viz. the angle ABC to the angle DEF, and the angle ACB to DFE.

For, if the triangle ABC be applied to DEF, ſo that the point A may be on D, and the ſtraight line AB upon DE; the point B ſhall coincide with the point E, becauſe AB is equal to DE; and AB coinciding with DE, AC ſhall coincide with DF, becauſe the angle BAC is equal to the angle EDF; wherefore alſo the point C ſhall coincide with the point F, becauſe the ſtraight line AC is equal to DF: But the point B coincides with the point E; wherefore the baſe BC ſhall coincide with the baſe EF, becauſe the point B coinciding with E, and C with F, if the baſe BC does not coincide with the baſe EF, two ſtraight lines would incloſe a ſpace, which is impoſſible[a]. Therefore the baſe BC ſhall coincide with the baſe EF, and be equal to it. Wherefore the whole triangle ABC ſhall coincide with the whole triangle DEF, and be equal to it; and the other angles of the one ſhall coincide with the remaining angles of the other, and be equal to them, viz. the angle ABC to the angle DEF, and the angle ACB to DFE. Therefore, if two triangles have two ſides of the one equal to two ſides of the other, each to each, and have likewiſe the angles contained by thoſe ſides equal to one another, their baſes ſhall likewiſe be equal, and the triangles be equal, and their other angles to which the equal ſides are oppoſite ſhall be equal, each to each. Which was to be demonſtrated. [a] 10. Ax.

PROP. V. THEOR.

THE angles at the baſe of an Iſoſceles triangle are equal to one another; and, if the equal ſides be produced, the angles upon the other ſide of the baſe ſhall be equal.

Let ABC be an Iſoſceles triangle, of h the ſide AB is e-

B qual

OF EUCLID. 19

Book I.

Let ABC be a triangle having the angle ABC equal to the angle ACB; the ſide AB is alſo equal to the ſide AC.

For, if AB be not equal to AC, one of them is greater than the other: Let AB be the greater, and from it cut[a] off DB equal to AC, the leſs, and join DC; therefore, becauſe in the triangles DBC, ACB, DB is equal to AC, and BC common to both, the two ſides DB, BC are equal to the two AC, CB, each to each; and the angle DBC is equal to the angle ACB; therefore the baſe DC is equal to the baſe AB, and the triangle DBC is equal to the triangle[b] ACB, the leſs to the greater; which is abſurd. Therefore AB is not unequal to AC, that is, it is equal to it. Wherefore, if two angles, &c. Q. E. D. [a] 3. I. [b] 4. I.

Cor. Hence every equiangular triangle is alſo equilateral.

PROP. VII. THEOR.

UPON the ſame baſe, and on the ſame ſide of it, there cannot be two triangles that have their ſides which are terminated in one extremity of the baſe equal to one another, and likewiſe thoſe which are terminated in the other extremity. See N.

If it be poſſible, let there be two triangles ACB, ADB, upon the ſame baſe AB, and upon the ſame ſide of it, which have their ſides CA, DA, terminated in the extremity A of the baſe, equal to one another, and likewiſe their ſides CB, DB that are terminated in B.

Join CD; then, in the caſe in which the vertex of each of the triangles is without the other triangle, becauſe AC is equal to AD, the angle ACD is equal[a] to the angle ADC: But the angle ACD is greater than the angle BCD; therefore the angle ADC is greater alſo than BCD; much more then is the angle BDC greater than the angle BCD. Again, becauſe CB is equal to DB, the angle BDC is equal the angle BCD; but it has been demonſtrated to be g than it; which is impoſſible. [a] 5. I.

B 2

136 THE ELEMENTS

Book V.

A, C, and L, M of B, D; if G be greater than L, H is greater than M; and if equal, equal; and if leſs, leſs[a]. Again, becauſe C is to D, as E is to F, and H, K are taken equimultiples of C, E; and M, N, of D, F; if H be greater than M, K is greater than N; and if equal, equal; and if leſs, leſs: But, if G be greater than L, it has been ſhown that H is greater than M; and if equal, equal; and if leſs, leſs; therefore, if G be greater than L, K is greater than N; and if equal, equal; and if leſs, leſs: And G, K, are any equimultiples whatever of A, E; and L, N any whatever of B, F: Therefore, as A is to B, ſo is E to F[a]. Wherefore ratios that, &c. Q. E. D. [a] 5. def. 5.

PROP. XII. THEOR.

IF any number of magnitudes be proportionals, as one of the antecedents is to its conſequent, ſo ſhall all the antecedents taken together be to all the conſequents.

Let any number of magnitudes A, B, C, D, E, F, be proportionals; that is, as A is to B, ſo C to D, and E to F: As A is to B, ſo ſhall A, C, E together be to B, D, F together.

Take of A, C, E any equimultiples whatever G, H, K; and of B, D, F any equimultiples whatever L, M, N: Then, becauſe A is to B, as C is to D, and as E to F; and that G, H,

K

OF EUCLID. 135

Book V.

PROP. X. THEOR.

THAT magnitude which has a greater ratio than another has unto the ſame magnitude is the greater of the two: And that magnitude to which the ſame has a greater ratio than it has unto another magnitude is the leſſer of the two. See N.

Let A have to C a greater ratio than B has to C: A is greater than B: For, becauſe A has a greater ratio to C, than B has to C, there are[a] ſome equimultiples of A and B, and ſome multiple of C ſuch, that the multiple of A is greater than the multiple of C, but the multiple of B is not greater than it: Let them be taken, and let D, E be equimultiples of A, B, and F a multiple of C ſuch, that D is greater than F, but E is not greater than F: Therefore D is greater than E: And, becauſe D and E are equimultiples of A and B, and D is greater than E; therefore A is[b] greater than B. [a] 7. Def. 5.

Next, Let C have a greater ratio to B than it has to A; B is leſs than A: For[a] there is ſome multiple F of C, and ſome equimultiples E and D of B and A ſuch, that F is greater than E, but is not greater than D: E therefore is leſs than D; and becauſe E and D are equimultiples of B and A, therefore B is[b] leſs than A. That magnitude, therefore, &c. Q. E. D. [b] 4. Ax. 5.

PROP. XI. THEOR.

RATIOS that are the ſame to the ſame ratio, are the ſame to one another.

Let A be to B, as C is to D; and as C to D, ſo let E be to F; A is to B, as E to F.

Take of A, C, E, any equimultiples whatever G, H, K; and of B, D, F, any equimultiples whatever L, M, N. Therefore, ſince A is to B, as C to D, and G, H are taken equimultiples of

I 4 A,

Abb. 9: Four pages from Simson’s edition put through the first phase of the identification process. Potential diagram components are colorized for manual review. Red: quad component; green: definite line component; blue: possible line component. Source for all four images: the Internet Archive.

[19] Definite lines are not necessarily components of diagrams: the three non-diagram examples that are often picked up are 1) page edges, 2) characters such as the letter ‘l’, the number ‘1’, or the symbol ‘=’, and 3) lines used elsewhere as part of the page layout (such as separation of main text and critical apparatus).

Meanwhile, the added dimension causes quad components to have much more variety in comparison to line components. Where lines can generally be distinguished through the aspect ratios of their bounding boxes, quad components can be distinguished primarily through the area of theirs. With the exception of dots in dotted lines, these components are always larger than the regular characters that make up the text. Digital Euclid's initial approach is therefore as follows: remove noise from the page, then analyze the connected components. Assuming the page contains text, the majority of these will be textual characters. Any connected components that have areas significantly larger than the text are potentially full or partial diagrams.

This method results in its own share of false positives, for example large drop caps or character clusters that are not properly separated by whitespace. These can be cut down through knowledge of the layout, analysis of nearest neighbors, or determining the percentage of black pixels within the bounding box. In the future, the Digital Euclid project plans to test the use of distance transforms to improve this automated correction step. While manual review is still necessary to handle the false positives, these methods do shorten the time it takes to identify diagrams across works.

The method works best in print editions. Within manuscripts, ligatures, abbreviations, and scripts where the characters are not separated by whitespace result in connected components that can be larger than geometric components. While this alone isn't necessarily an insurmountable barrier, intersection between the diagrams and either the *Elements* text or scholia text is also frequent. Further work is needed to develop tools specifically for manuscript diagrams. While more time-consuming, identifying manuscript diagrams manually is still a reasonable option today.

Once the diagrams have been identified, they can be separated from the text, allowing for diagram images to continue down their path in the workflow and for OCR to be applied to the remaining page. The approach discussed above finds the geometric components, but the separate alphabetic labels must be captured as well, both to create complete traces and to prevent them from interfering with the OCR of the text. Digital Euclid locates these by considering all material within the bounding boxes of the geometric components part of the diagram, and by expanding the bounding boxes by a certain input amount in order to locate outer labels.

Producing Machine-actionable Diagram Data

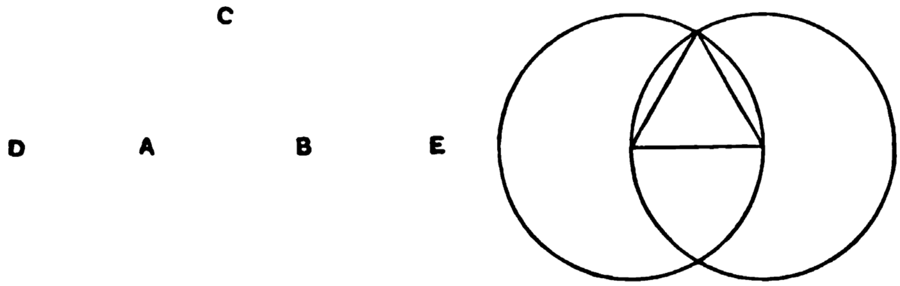


Abb. 10: Diagram for Elements Book I proposition 1, separated into two images containing either alphabetic or geometric components. Source: the Internet Archive.

Once diagram selections have been extracted, they can be split into two images that will go through separate workflows – figure 10 provides an example. The prior identification process already found most of the geometric material, so it is not complicated to separate this from the labels.

Currently, this step cannot be completed solely through automated means because of the high tendency of labels to intersect with the geometric portions. How often this occurs varies across editions: in Heath's translation, only 25 of the 494 diagrams have this issue. In some cases this intersection is accidental, the result of poor binarization for instance, but in others it is simply how the diagram was produced (an example can be seen in figure 11).

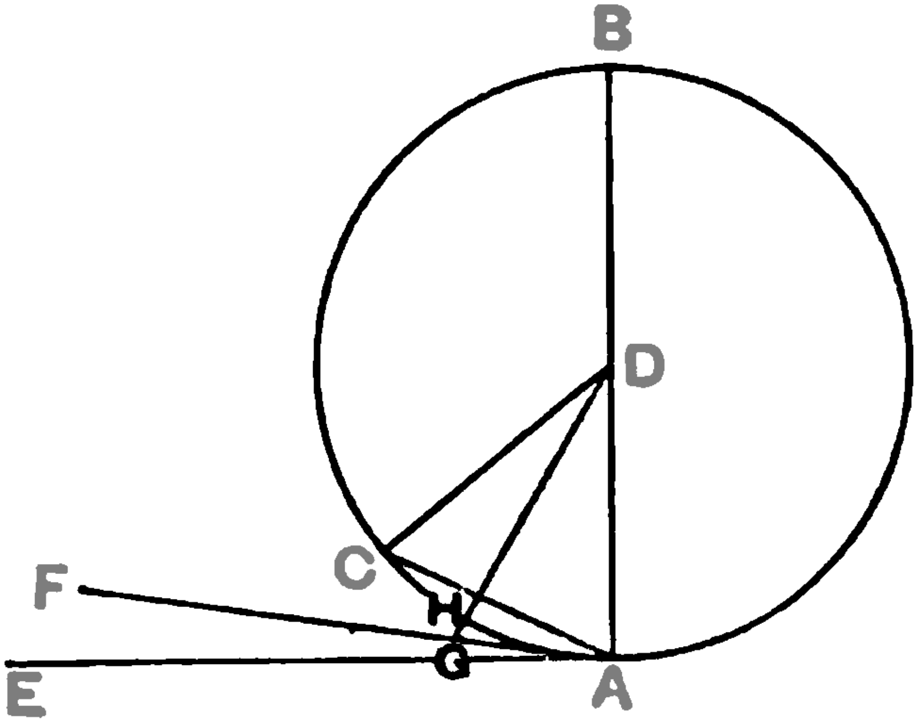


Abb. 11: Diagram from Heath's English edition. Connected components that are identified as labels appear in grey. This diagram contains two labels, H and G, that overlap with the figure and that must be removed manually. Source: the Internet Archive.

OCR can be used to transcribe the labels. Although most out-of-the-box systems will return errors on these images because they expect lines, words, or at least a significant number of characters, Gamera for instance can be used to train a basic OCR system that will function on images with scattered characters like figure 10. The OCR process also provides a bounding box defining the region of interest for each label, which will be useful in later steps when traced diagrams are annotated according to their labels.

As with the main text, OCR has difficulty dealing with manuscript labels, and these are currently handled manually.

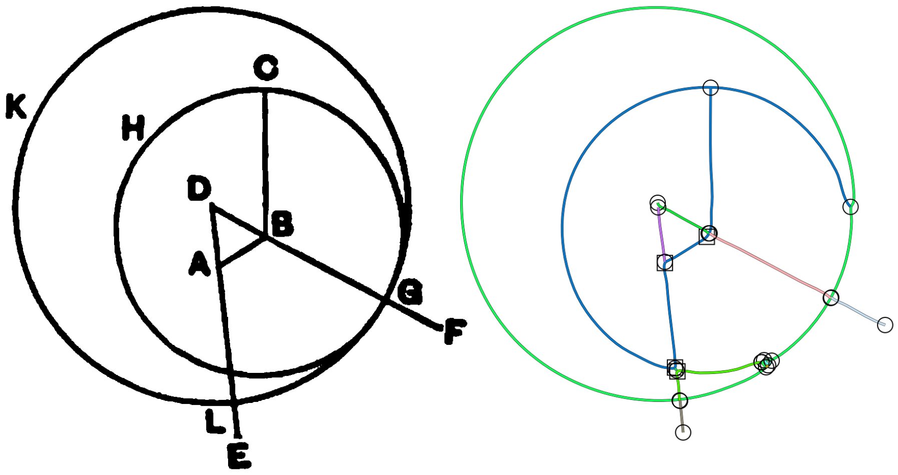


Abb. 12: Left: diagram for Elements Book I proposition 2 from Heath's edition. Right: a segmented autotraced SVG version with endpoints and corners located and marked for manual review. Source for original image: the Internet Archive.

Tracing the geometric portion requires a different approach. Autotrace, a command line utility, can produce a rough SVG version of a raster file.[20] The resultant output is not precise enough to serve as the final digital trace, but does contain enough data to more easily locate certain points of interest, namely corners, endpoints, and intersections. By using these points and referring back to the original raster image, the Digital Euclid project hopes to produce cleaner SVG traces. Figure 12 shows a diagram that has been analyzed in this way, and which has also been segmented into the smallest elements of the overall shape.

Additionally, traces can be produced manually using vector-editing software. For this the Digital Euclid project uses Inkscape.[21]

Recombining the data

Lastly, the workflow turns to the recombination of the now-extracted data. Within diagrams, the transcribed labels are reunited with the SVG traces. This is accomplished by analyzing the distance between each label and each point of interest to find pairs that are in closest proximity to each other. In the Elements of Euclid, labels are most often located at these points. For labels that are not within the expected distance to one of these points, the method checks for the nearest arbitrary point on the diagram and assigns the label to that point.

Assigning labels to points allows for the automatic creation of identifiers for each segment of the diagram. Sections that contain labelled points can be identified as 'line.AB', for example, or 'arc.ΓΔ'. These identifiers are added to the diagram's SVG file.

It is also necessary to unite the diagram with the proposition it accompanies. Except in the case of fold-out diagrams, this can be accomplished by comparing the regions of interest for

[20] http://autotrace.sourceforge.net/.

[21] https://inkscape.org/en/.

the diagrams and propositions. Whether appearing at the beginning, middle, or end of a proposition, the diagram is generally located in close vicinity. The Digital Euclid project preserves bounding box information obtained during the OCR workflow for the main text, so once the text is structured in CTS, the region of interest for the propositions can be determined. Even in a manual transcription or tracing process, the region of interest for that text or diagram is recorded and can be used at this stage.

Conclusion

This initial phase of the Digital Euclid project investigated the potential for a digital representation of the transmission of Euclid's *Elements*, and this goal is certainly feasible. Institutional and large-scale digitization efforts have produced high quality images of hundreds of editions. This is important: further steps along the digitization process are dependent upon the first step of imaging the texts. While print editions currently dominate the digitized set, more and more manuscripts are being made available for scholarly use.

The project used a combination of extant tools and experimental new ones to test out identifying and extracting the textual and diagrammatic data. The text of the *Elements* can for the most part be approached through typical OCR workflows. Diagrams were the major challenge in the groundwork for the Digital Euclid project: especially in comparison to the text, they were the speed bump to the extraction process. However, this past year with Digital Euclid has demonstrated that this was the case primarily due to lack of work on the problem. Ancient mathematical diagrams are not unapproachable by automated means.

When looking forward, plenty of work remains to be done. The survey is not yet comprehensive: the project will continue to add to this and publish updates to the Digital Euclid Github repository. Similarly, the identification and extraction tools and methods that were newly developed this year will be improved before the project begins to use them in earnest. Ultimately these approaches will allow the Digital Euclid project to publish numerous machine-actionable texts, diagrams, and datasets from and regarding the *Elements* that can serve as a flexible resource for further scholarship.

Literatur:

Busard (1968): H. L. L. Busard, The Translation of the Elements of Euclid from Arabic into Latin by Hermann of Carinthia(?), Leiden 1968.

Busard (1983): H. L. L. Busard, The First Latin Translation of Euclid‘s Elements commonly ascribed to Adelard of Bath. Books I–VIII and Books X.36–XV.2, Toronto1983.

Busard (1984): H. L. L. Busard, The Latin Translation of the Arabic Version of Euclid’s Elements commonly ascribed to Gerard of Cremona, Leiden 1984.

Busard (1987): H. L. L. Busard, The Mediaeval Latin translation of Euclid’s Elements Made Directly from the Greek, Stuttgart 1987.

Busard (1992): H. L. L. Busard, Robert of Chester’s Redaction of Euclid’s Elements, the so-called Adelard II Version, Birkhäuser 1992.

Busard (1996): H. L. L. Busard, A Thirteenth-Century Adaption of Robert of Chester‘s Version of Euclid‘s Elements, München 1996.

Busard (2001): H. L. L. Busard, Johannes de Tinemue‘s redaction of Euclid‘s „Elements“, the so-called Adelard III version, Stuttgart 2001.

Busard (2005): H. L. L. Busard, Campanus of Novara and Euclid‘s Elements, Stuttgart 2005.

Folkerts (1989): M. Folkerts, Euclid in Medieval Europe, The Benjamin Catalogue 1989.

Heiberg (1883–6): I. L. Heiberg, “Euclidis Elementa”, Euclidis Opera Omnia Vol.1–4, Leipzig 1883–6.

Kayas (1977): G. J. Kayas, Vingt-trois siecles de tradition euclidienne: essai bibliographique, Palaiseau 1977.

Lo Bello (2003): A. Lo Bello, Gerard of Cremona‘s Translation of the Commentary of Al-Nayrizi on Book I of Euclid‘s Elements of Geometry, Leiden 2003.

Netz (2004): R. Netz, The Works of Archimedes: Translation and Commentary, Cambridge 2004.

Riccardi (1887): P. Riccardi, Saggio di una bibliografia euclidea, Tipografia Gamberini e Parmeggiani 1887.

Saito (2009): K. Saito, “Reading ancient Greek mathematics”, The Oxford Handbook of the History of Mathematics, Oxford 2009.

Steck (1981): M. Steck, Bibliographia Euclideana: Die Geisteslinien der Tradition in den Editionen der ‘Elemente’ des Euklid um 365–300, Arbor scientiarum, Hildesheim 1981.

Weitere Ressourcen (Zuletzt aufgerufen am 29.12.2015):

ABBYY FineReader:
http://www.abbyy.com/finereader/

Autotrace:
http://autotrace.sourceforge.net/

Inkscape:
https://inkscape.org/en/

Gallica:
http://gallica.bnf.fr/

Google Books:
https://books.google.com/

HathiTrust:
https://www.hathitrust.org/

Internet Archive:
https://archive.org/

Münchner Digitalisierungszentrum (MDZ):
http://www.digitale-sammlungen.de/

OCRopus:
https://github.com/tmbdev/ocropy

Open Greek and Latin Project of the Open Philology Project:
http://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/

Perseus:
http://catalog.perseus.org/

SLUB Dresden:
http://www.slub-dresden.de/startseite/

Sourceforge (Find, Create, and Publish Open Source Software):
epidoc.sourceforge.net/.

Tesseract:
https://github.com/tesseract-ocr
https://code.google.com/p/tesseract-ocr/

The Gamera Project:
http://gamera.informatik.hsnr.de/

The CITE Architecture:
http://cite-architecture.github.io/ctsurn/overview/

Autorenkontakt[22]

Christine Roughan
Universität Leipzig
Lehrstuhl für Digital Humanities
Email: cmroughan@gmail.com

[22] Die Rechte für Inhalt, Texte, Graphiken und Abbildungen liegen, wenn nicht anders vermerkt, bei den Autoren.