Bulletin for Spanish and Portuguese Historical Studies
Journal of the Association for Spanish and Portuguese Historical Studies

Volume 43
Issue 1 *Digital Humanities*

Article 2

2018

Deciphering Secrets of Medieval Cathedrals: Crowdsourced Manuscript Transcriptions and Modern Digital Editions

Roger Louis Martinez-Davila
Universidad Carlos III de Madrid, rogerlmartinez@gmail.com

Sean Perrone
Saint Anselm College, sperrone@anselm.edu

Francisco Garcia Serrano-Nebras
Saint Louis University-Madrid Campus, francisco.garciaserrano@slu.edu

Maria Martin de Vidales Garcia
Universidad Carlos III de Madrid, mariamvidalesg@gmail.com

Deciphering Secrets of Medieval Cathedrals: Crowdsourced Manuscript Transcriptions and Modern Digital Editions

Cover Page Footnote
The Deciphering Secrets project has received funding from the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement nº 600371, el Ministerio de Economía y Competitividad (COFUND2013-51509) and Banco Santander. More information at http://www.decipheringsecrets.com.

Deciphering Secrets of Medieval Cathedrals: Crowdsourced Manuscript Transcriptions and Modern Digital Editions

Roger Louis Martinez-Davila, **Sean Perrone**, **Francisco Garcia Serrano-Nebras**, **Maria Martin de Vidales Garcia**

Introduction

Spanish historians are familiar with the great editorial projects of the late nineteenth and early twentieth century, such as *Colección de documentos inéditos para la historia de España*, 112 vols. (1842-1895). By transcribing and printing documents, these projects made primary sources easily accessible to scholars across the globe and thereby advanced scholarship. Today, we are entering a new era of large-scale editorial projects. The digitalization of archives and new information technologies makes it easier to transcribe and edit primary sources and subsequently make those sources accessible to scholars and students in machine-readable file formats.

These new editorial projects rely heavily on crowdsourcing, or the practice of soliciting and employing the assistance of large groups of persons, to transcribe the text. Some well-known historical transcription projects include George Mason University's *Scripto* project (engages citizens in the process of transcribing the *Papers of the War Department, 1784-1800*); *Transcribe Bentham: A Participatory Initiative*; New York Public Library's *Scribe*; *The Old Weather Project*, St. Louis University's *T-PEN Tool*, and University of Saskatchewan's *Textual Communities*.[1] Crowdsourcing allows text to be transcribed more quickly and more cost-effectively than hiring professional transcribers. Crowdsourcing, of course, raises concerns, especially regarding the accuracy of transcriptions. However, Tim Causer and Valerie Wallace determined that utilizing volunteer citizens as transcription specialists was both effective and accurate.[2]

When queried about their perspective on the value of "crowdsourced research", students in our *Deciphering Secrets* Massive Open Online Courses (MOOCs) and transcription project resoundingly supported it.[3] Specifically, 82.5%

---

[1] See http://wardepartmentpapers.org/; https://www.ucl.ac.uk/bentham-project/transcribe-bentham; http://scribeproject.github.io/; https://www.oldweather.org/; http://t-pen.org/TPEN/; and https://textualcommunities.org/app/home (accessed November 29, 2018).

[2] Tim Causer and Valerie Wallace, "Building a Volunteer Community: Results and Findings from *Transcribe Bentham*," *Digital Humanities Quarterly* 6, no. 2 (2012): 1-84, http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html (accessed November 29, 2018).

[3] This student survey data collected during the summer 2016 edition of the *Deciphering Secrets: Unlocking the Manuscripts of Medieval Burgos* MOOC on edX.org at https://www.edx.org/course/deciphering-secrets-unlocking-uc3mx-hga-2-1x-0 (accessed December 16, 2018). The current version of the MOOC on coursera.org is accessible at

of our students believe it is *very important* for academic researchers to incorporate the public into scholarly investigations. A full 100% believe these scholarly-public collaborations are *very or somewhat important*. When asked if they wanted to participate in other types of research:

- 85% indicated they wanted to contribute to additional paleography transcription projects,
- 65% stated they wanted to assist with editing and creating final versions of transcribed manuscripts, and
- 75% wanted to help index the contents of manuscripts that have not been abstracted.

After completing our course, 77.5% of our students stated they were *very likely* to return and take additional *Deciphering Secrets* MOOCs. Therefore, it appears that this student body is highly motivated and is an important, untapped cooperative resource for scholars.

However, there are ethical issues we must consider when tapping into this potential resource for our research. For example, Amazon's Mechanical Turk (https://www.mturk.com/), a crowdsourcing platform employing paid participants, has raised ethical concerns about fair compensation. Our project is not a crowdsourcing marketplace along those lines. Rather, we incorporate nonprofessional student volunteers, who are learning paleography in an online course, into the project. Most students take the course for free, and those who take it for an edX certificate pay a nominal fee.[4] Through course context (described below), we try to create an academic community, where student success is determined by acquiring a skill and advancing scholarship. In this sense, there is an implicit trade that occurs between the research professor and their students: a no-cost, well-designed, and skill-building course is exchanged for student assistance with manuscript transcriptions. The collaborative nature of these transcriptions often means that multiple students transcribed different parts of a single document. Consequently, it is difficult to give individual students attribution in such cases. Instead, we will acknowledge the cohort that transcribed the specific documents. However, if individual students make substantive contributions to

https://www.coursera.org/learn/burgos-deciphering-secrets-medieval-spain (accessed December 16, 2018).

[4] In the spring 2017 cohort of students in Deciphering Secrets: Unlocking the Secrets of Medieval Toledo, for example, only thirty-six of 1,957 persons paid to take the course (98.2% receive the course for free).

transcribing, marking up, or tagging final documents for posting on the web, those students will receive proper recognition for their work on the project.[5]

Crowdsourced transcription is also a process that we have all already wittingly or unwittingly participated in every time we have deciphered a handwritten text to complete an online transaction. This test was created by Louis von Ahn to prove registrants to online forums were human and not "spambots" in 2000. This test is paleography, and once von Ahn realized that these mini-transcriptions could be used to decipher words from "text-scanning projects that a computer's optical-character-recognition (OCR) program couldn't understand," he updated the program, renamed ReCAPTA, for that purpose. Now, when five people correctly identify the same set of squiggly lines at the end of a transaction, the computer knows what those lines signify.[6] Essentially millions of users daily help to transcribe texts without even realizing it and thereby improve OCR software.[7] Online Spanish paleography courses will never reach such large numbers of people, but they can potentially serve the same purpose. Once we can correlate the squiggly lines from medieval manuscripts with words, we can potentially teach the computer to transcribe the text for us through OCR software or the more recently developed Handwritten Text Recognition (HTR) software.[8] Such possibilities are not whimsically flights of fancy either. A project along similar lines is currently underway in Venice.[9]

This paper discusses *Deciphering Secrets*, a large-scale transcription project linked to a MOOC to teach Spanish paleography and the SILReST paleography method used in the course. It also addresses issues of accuracy, particularly how

---

[5] For more on the ethical dimensions of crowdsourcing see Tom Bivins, "Crowds, Community, and Jeremy Bentham," *Journal of Media Ethics* 31, no. 4 (2016): 266-268; Vanessa Williamson, "On the Ethics of Crowdsourced Research," *PS: Political Science & Politics* 49, no. 1 (January 2016): pp.77-81, https://doi.org/10.1017/S104909651500116X (accessed November 16, 2018). UCLA's Digital Humanities program has also developed a useful Collaborators' Bill of Rights for students participating in digital humanities research https://humtech.ucla.edu/news/a-student-collaborators-bill-of-rights/ (accessed November 16, 2018). See also *Transcribe Bentham* at http://transcribe-bentham.ucl.ac.uk/td/Transcribe_Bentham (accessed December 16, 2018).

[6] See https://www.google.com/recaptcha/intro/v3.html (accessed December 16, 2018).

[7] Viktor Mayer-Schönberg and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Boston: Mariner Books, 2013), 98-99.

[8] Michael Widner, "Toward Text-Mining the Middle Ages: Digital Scriptoria and Networks of Labor," in *The Routledge Handbook of Digital Medieval Literature*, eds. Jennifer E. Boyle and Helen J. Burgess (London: Routledge, 2018), 132-133; Uwe Springmann and Anke Lüdeling, "OCR of Historical Printings with an Application to Building Diachronic Corpora: A Case Study Using the RIDGES Herbal Corpus," *Digital Humanities Quarterly* 11, no. 2 (2017). http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html (accessed September 10, 2018).

[9] Alison Abbott, "The 'Time Machine' Reconstructing Ancient Venice's Social Networks," *Nature*, no. 546 (15 June 2017): 341-344.

connecting crowdsourced transcription to class assignments will increase the accuracy of transcriptions and quickly make available to the scholarly community many manuscript transcriptions from the cathedral chapters of Burgos, Plasencia, and Toledo. It then discusses the refinement of editorial techniques in the migration of archival materials to digital format, creating standards for text encoding, and postulates the next steps of building databases. In the future, this will include the ability to toggle back-and-forth so that we can view orthographical distinctions between transcriptions. The paper finally shows how new knowledge came to light through the transcription process and highlights possibilities for future research. Moreover, crowdsourcing makes students active participants in the creation and preservation of cultural materials. It engages students – making history and the humanities more relevant to them. [10]

Cathedral Chapters, Digitalization of Documents, and Possibilities

The historiography on cathedral chapters has increased dramatically in recent years,[11] and better hours at many cathedral archives make investigation easier. Still, the dearth of printed sources has limited research on cathedral chapters compared to other corporate bodies.[12] Consequently, the ability to access and search for vital data from cathedral archives online will allow for more thorough research on cathedral chapters. Scholars can better understand the place of chapters within local and national affairs as well as within the ecclesiastical estate. The records can also reveal interreligious history as thousands of residential leases, business documents, and inter-religious agreements are found in cathedral archives. For instance, the Muslim, Audalla de Valladolid, leased homes in Burgos's Muslim quarter from the cathedral in 1431.[13] These same records can help us to understand rental markets across Spain, the role of the church in financing the crown, elite networks, and many other subjects.[14]

---

[10] Anne Brudrick et al. *Digital_Humanities* (Cambridge: MIT Press, 2012), 6, 23.

[11] Antonio J. Díaz Rodríguez, "Cabildos catedralicios y clero capitular en le antiguo régimen: Estado de la cuestión," *Revista de Historiografía* 7 (2010): 82-99.

[12] One notable exception is the summaries of the minutes from Avila prepared by Andrés Sánchez Sánchez. See Andrés Sánchez Sánchez, *Resumen de actas del Cabildo Catedralicio de Avila*, vol. 1 (1511-1522), vol. 2 (1522-1533), vol. 3 (1534-1541), vol. 4 (1542-1550) (Avila: Ediciones de la Institución "Gran Duque de Alba" de la Excma. Diputación Provincial de Avila, 1995-2009).

[13] Manuscript Archivo de la Catedral de Burgos (ACB). Registro 9, Folio 136v.

[14] Mauricio Drelichman and David González Agudo, "Housing and the Cost of Living in Early Modern Toledo," *Explorations in Economic History* 54 (2014): 27-47; Sean T. Perrone, *Charles V and the Castilian Assembly of the Clergy: Negotiations for the Ecclesiastical Subsidy* (Leiden & Boston: Brill Academic Publishers, 2008); Susana Guijarro González, "Monarquía, carreras eclesiásticas y cabildos catedralicios en la castilla bajomedieval: Los capellanes reales en la cathedral de Burgos," in *Ecclesiastics and political state building in the Iberian monarchies 13th-15th centuries*, eds. Hermina Vasconcelos Vilar and Maria João Branco (Lisbon: Publicações do Cidehus, 2016).

Cathedral chapters then are a critical historical source that detail how ecclesiastical institutions functioned, describe the detailed social and economic threads that connected community members to one another, and offer opportunities to evaluate significant thematic issues like the formation of early modern identities. From the Cathedral of Plasencia's *Actas Capitulares*, the chapter's administrative tomes, we can appreciate how conversos like the fifteenth-century Carvajal-Santa Maria family of churchmen captured the bishop's mitre and monopolized the position of archdeacon and other chapter positions to enhance the clan's wealth, institutional power, and to re-cast themselves into pious Catholic leaders, such as Cardinal Bernardino Lopez de Carvajal, who resided at the center of Spanish and papal affairs.[15]

The first step to any large crowdsourcing transcription project is the digitalization of primary sources. Archives and libraries have been digitalizing text for conservation and internal use for decades now. Probably the most well-known early digitalization project in Spain involved the Archivo General de las Indias, which began to digitalize its most frequently used collections in the early 1990s. Ecclesiastical archives have also received funding for digitalization projects, such as Fundación Cajacírculo's ongoing contributions to the Archivo Historico de la Catedral de Burgos to create electronic indexes for over 150,000 documents.[16] However, these digitalized sources are often available for in-house use only. They make research easier, especially the ability to magnify difficult to read passages, but they don't allow users to search the text for particular words or phrases. That is, digitalization of the text is not "datafication" of the text. Transcriptions of manuscripts, on the other hand, will make it possible to analyze word usage over time (i.e., text-mining) and make text into searchable data.[17]

Transcription, however, is just the first step in the creation of data. Once the manuscripts are transcribed, the transcriptions will need to be processed so that the data can be analyzed. For example, the data could be visualized as maps, graphs, or word clouds.[18] Computers can also generate concordance and indices to the manuscripts by automating tasks (e.g., word searching, sorting, counting, and listing). Information technology's ability to read transcriptions much faster than humans will make it easier to identify speech patterns or changing word usage within a cathedral chapter or between cathedral chapters. Such distant reading

---

[15] Roger L. Martinez-Davila, *Creating Conversos: The Carvajal-Santa Maria Family in Early Modern Spain* (Notre Dame: University of Notre Dame Press, 2018).

[16] For more information on this collaboration between the Cathedral of Burgos and Fundación Cajacírculo see: https://www.fundacioncajacirculo.es/AHCB.php (accessed September 18, 2018). This information is reported under the sub-title, "Normas de busqueda".

[17] For more on digitalization vs. datafication see Mayer-Schönberg and Cukier, *Big Data*, 83-85.

[18] For an excellent appraisal of different ways to visualize text and a comprehensive survey of the literature, see S. Jänicke, G. Franzini, M.F. Cheema, and G. Scheuermann, "Visual Text Analysis in Digital Humanities," *Computer Graphics Forum* 36, no. 6 (2017): 226-250.

might give us insights into the material that the typical close readings of scholars would miss.[19] That is, the ability to detect/identify linguistic tendencies, changes, and diffusion in archival records might provide us with new avenues for close readings. The larger the dataset will also give us even more verifiable conclusions.[20] Google's Ngram viewer already makes such analysis possible in Google books, and scholars are becoming familiar with these types of "big data" searches.[21] The ability to find "statistically significant words" across transcribed manuscript collections should allow for similar types of analysis.

Moreover, the ability potentially to merge distinct projects' databases in the future will allow searches and cross-referencing across multiple archives and types of archives. Currently, most printed editions of documents focus on a specific archive or a determined period/theme/person, and thus the collections maintain a certain fragmented quality. There are innumerable documents and many interesting stories to be told, but the inability to see large-scale patterns can limit the usefulness of such collections. Online document collections can potentially merge with other collections as well as be hyperlinked to Google Books or other websites to provide the breath necessary to do text searches for patterns as well as bring to the fore ordinary people hidden in obscure manuscripts.[22] This is especially true for scholars seeking intriguing details of Jewish, Christian, and Muslim relationships in cathedral archives. Most cathedral manuscripts were never indexed or cataloged with other religious group's history in mind. Thus, these medieval sources on interfaith interactions are an unknown treasure.

Of course, any editorial project will be limited by the documents that archives are willing to digitalize, or allow scholars to digitally photograph, and make available for transcription. For example, during summer 2016, the Archive of the Cathedral of Burgos granted our Deciphering Secrets MOOCs unprecedented access and permission to photograph and digitally-distribute approximately 1,000+ pages of medieval manuscripts that pertain to religious coexistence. Don Matías Vicario Santamaría, a cathedral canon and the archivist of the Cathedral of Burgos, embraced the MOOC and our effort to bring these manuscripts, some of which had

---

[19] Alexander T.J. Barron, Jerry Huang, Rebecca L Spang, and Simon DeDeo "Individuals, Institutions, and Innovation in the Debates of the French Revolution," *Proceedings of National Academy of Science* 115 (2018): 4607-4612 http://www.pnas.org/cgi/doi/10.1073/pnas.1717729115 (accessed June 29, 2018); see also Brudrick et. al., *Digital_Humanities,* 39, 123.

[20] Brudrick et. al., *Digital_Humanities*, 37; Elena Pierazzo, "Digital Documentary Editions and the Others," *Scholarly Editing: The Annual of the Association for Documentary Editing* 25 (2014) http://www.scholarlyediting.org/2014/essays/essay.pierazzo.html (accessed August 25, 2018).

[21] Mayer-Schönberg and Cukier, *Big Data*, 84. See http://books.google.com/ngrams (accessed December 16, 2018).

[22] For more on the new types of knowledge generation made possible by technology, see Lara Putman, "The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast," *American Historical Review* 121, no. 2 (April 2016): 377-402.

not been viewed for several hundred years, to the public for closer evaluation. Key to the collaboration was our focus on the protecting the rights and privileges of the institutions, which included watermarking all manuscript images used in the course, as well as requiring all participating students to abide by a Student Code of Conduct that prohibited them from distributing or reproducing manuscript images outside of the course. The watermark includes a statement reserving the rights of the cathedral as well as blue lines that subdivide the image into eight sections labeled "a" through the letter "h" for the purposes of student transcription. (See Figure 1) Thanks to collaboration with the cathedral archives of Burgos, Plasencia, and Toledo, several distinct types of materials have been transcribed from broad records like cathedral minutes to more specific business records like those from Burgos. For example, Burgos' archive includes two critical collections of volumes, *Libros* and *Registros*, that pertain to the daily administrative operations of the cathedral and the diocese. Within these chronologically-ordered volumes, the daily decision-making and agreements of the cathedral chapter are in full view, and the volumes reveal members and chapter roles, salaries and payments to contractors, church property holdings, leases and sales of real property and goods, and rules governing the chapter organization. In Plasencia, the corollary collection is the *Actas Capitulares*. Toledo is similarly organized, with the important caveat of the *Obras y Fabricas* collection, which specifically details the collection of rents from church properties as well as payments to contractors, such as Muslim carpenters, for the construction and repair of church properties.

A further limitation of all transcription projects is that most originate as part of larger projects with specific ends. Consequently, the selection of data and the organization of data shape "what we and other scholars ask and see."[23] *Deciphering Secrets*, for example, is the crowdsourcing component of our larger effort, the *Revealing Cooperation and Conflict Project (RCCP)* (http://revealingcooperationandconflict.com), an international collaboration, to study interfaith relations in Plasencia between roughly 1391 and 1609.[24] Clearly, the paleography MOOC has gone well beyond that, but it is important to know the initial research interests of the project members to understand the scope and context of such transcription project, and why at times the material will seem fragmented.

---

[23] Katherine Rawson, "Curating Menus: Digesting Data for Critical Humanistic Inquiry," in *Laying the Foundations: Digital Humanities in Academic Libraries*, eds. John W. White and Heather Gilbert (Purdue University Press, 2016), 62 http://www.jstor.org/stable/j.ctt163t7kq.7 (accessed June 21, 2018).

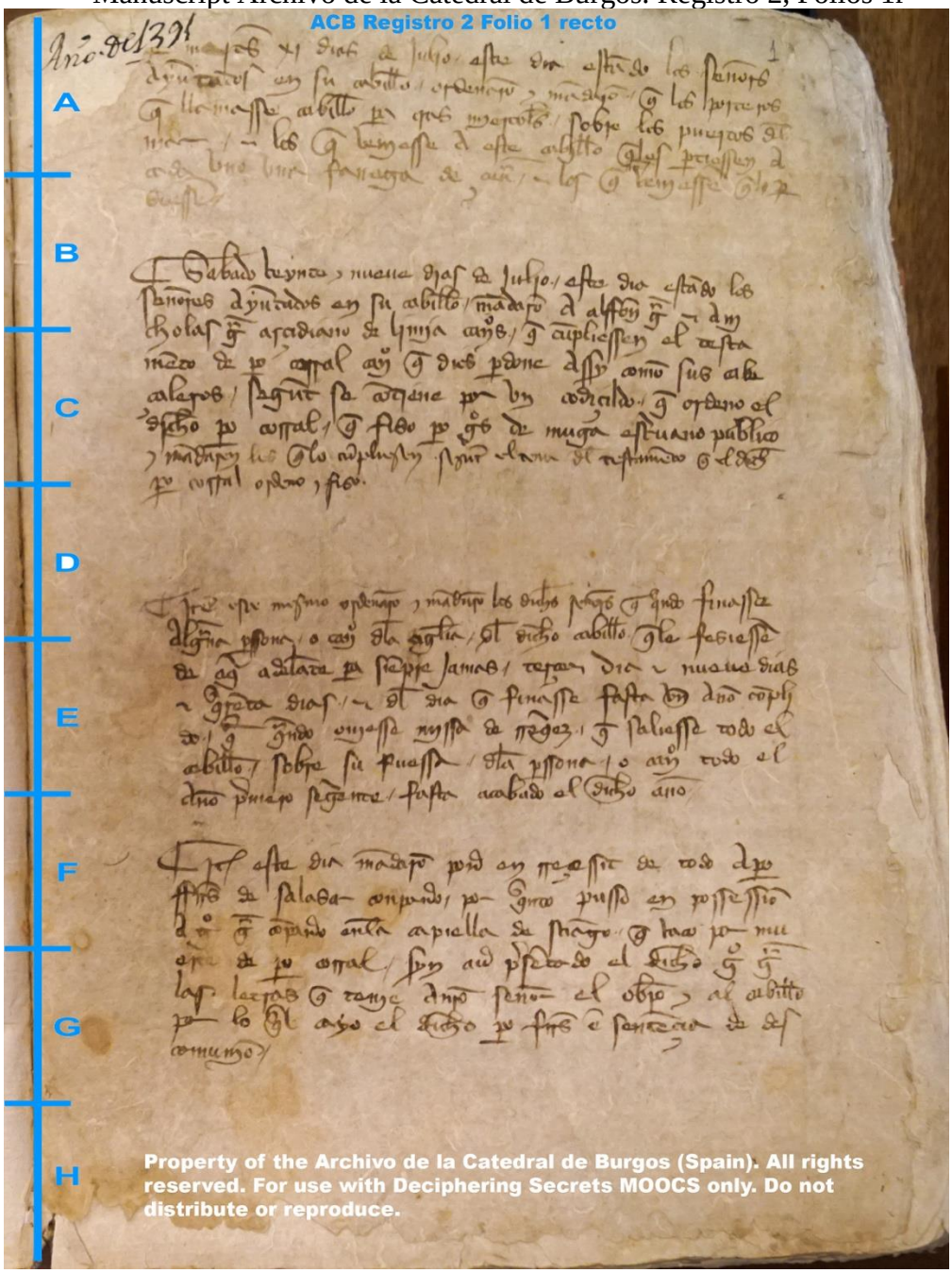
[24] Additional details about the RCCP and its virtual world, Virtual Plasencia, are published in Roger Louis Martínez-Dávila, Paddington Hodza, Mubbasir Kapadia, Sean T. Perrone, Christoph Hölscher, and Victor R. Schinazi, "Telling Stories: Historical Narratives in Virtual Reality," in *The Routledge Digital Medieval Literature and Culture*, eds. Jennifer E. Boyle and Helen J. Burgess (London: Routledge, 2018), 107-130.

Yet, as stated above, the ability to connect to other sources online will likely make these transcriptions even more fruitful for research purposes than printed editions.

MOOCs, Paleography, Crowdsourcing manuscripts

Our first MOOC integrating online education with crowdsourced transcription of manuscripts for research purposes took place in the summer of 2014. Since then Dr. Martinez-Davila has delivered several MOOCs via the Coursera.org-Univer-sity of Colorado System and the edX.org-Universidad Carlos III de Madrid col-laborations. As of early fall 2018, almost 22,000 students from over 140 nations have participated in the MOOCs. Similar to other MOOCs, student completion rates for the courses range from 8% to 19%. Demographically our students our older, better-educated, more often to be women, and English-speakers. These types of students seem well positioned to take on the challenge of not only learning about religious diversity during the Spanish Middle Ages but also well suited to the difficult task of learning to perform manuscript analysis.

Figure 1:
Manuscript Archivo de la Catedral de Burgos. Registro 2, Folios 1r



Source: Roger Louis Martinez-Davila with permission of the Archivo de la Catedral de Burgos, 2018. Creative Commons license BY-NC-ND 4.0.

Over four years, these citizen scholars transcribed most of the six hundred plus page *Libro I* of the *Actas Capitulares* of the Cathedral Chapter of Plasencia as well as dozens of documents from the cathedral archives of Burgos and Toledo.[25] Table 1 displays the seven MOOC course editions we authored as well as student enrollments and the specific manuscript collections we investigated. It is worth noting that starting in spring 2018 each Coursera.org MOOC enrolls a new cohort of students every six weeks, or about eight-course sessions each calendar year.

Table 1
Deciphering Secrets MOOCs: Editions, Enrollments, Manuscript Collections

MOOC	Edition	Enrollment	Manuscript Collections
Unlocking the Manuscripts of Medieval Spain (Coursera)	Summer 2014	10,600 students	19th-century copy of the 600+ page *Book One (1399-1453) of the Capitulary Acts* of the Archivo de la Catedral de Plasencia (Spain)
Unlocking the Manuscripts of Medieval Spain (Coursera)	Spring 2016	6,000 students	Same as prior MOOC.
Unlocking the Manuscripts of Medieval Burgos (edX)	Summer 2016	1,700 students	13th, 14th and 15th-century manuscripts from the Archivo de la Catedral de Burgos (Spain) and the Archivo Municipal de Burgos (Spain).
Unlocking the Manuscripts of Medieval Burgos (edX)	Spring 2017	1,400 students	Similar materials to the prior MOOC on Burgos.
Unlocking the Manuscripts of Medieval Toledo (edX)	Fall 2017	1,300 students	13th, 14th and 18th/19th-century manuscripts from the Archivo y Biblioteca de la Catedral de Toledo (Spain), Archivo Municipal de Toledo (Spain), and the Archivo Historico de la Nobleza (Toledo, Spain)
Burgos: Deciphering Secrets of Medieval Spain (Coursera)	Continuous enrollment since Spring 2018	1,500 students	Similar materials to the prior MOOC on Burgos.
Toledo: Deciphering Secrets of Medieval Spain (Coursera)	Continuous enrollment since Summer 2018	800 students	Similar materials to the prior MOOC on Toledo.

[25] Many of Deciphering Secrets' crowdsourced manuscript collections are routinely updated on the research section of our website. See http://www.decipheringsecrets.com/research/ (accessed December 16, 2018).

In terms of content, the MOOCs provide online history education and paleographic instruction. In five and seven-week courses, students learn about the complex nature of Jewish, Christian, and Muslim relations through brief Spanish and English documentary-style videos with English and Spanish-subtitling,[26] readings (including original sources), and participatory discussions via online forums and social media (e.g., *Facebook's Revealing Cooperation and Conflict Project* group page). These activities challenge students to examine the significance of medieval material culture and manuscripts. For example, how was it possible that an eleventh-century Islamic ivory chest was transformed into a "Christian" one to hold a Christian saint's relics? (See Figure 2.) Or, how should we understand an eleventh-century royal privilege that made Jewish communities contribute monies for the care of Christian pilgrims traveling on the Camino de Santiago?[27] By using engaging materials and raising intriguing questions, these MOOCs entice and encourage students to take on the challenge of learning medieval Spanish paleography and transcribing original manuscripts.[28]

After reviewing the history of medieval Spain, the course shifts its focus to paleographical instruction. While we teach using medieval Spanish manuscripts, very little or no knowledge of the Spanish language is necessary to complete the course. Using an intensive array of paleography practices, exams, independent projects, and collaborative efforts, students garner the essential skills to interpret any medieval European handwriting. To demonstrate their mastery of paleography students (1) create alphabet, numeral, and abbreviation guides using manuscript images, (2) complete paleographic exams, and (3) transcribe selections from a medieval manuscript.

---

[26] For example, the following *Deciphering Secrets* videos that feature local museum directors (i.e., Museo de Burgos), archivists (i.e., Archivo Municipal de Burgos), cultural delegates (i.e., Centro Sefarad Israel, and Casa Arabe), and scholars (i.e., Consejo Superior de Investigaciones Científicas) hosted on Youtube.com. See https://youtu.be/SQDVIhhc8GY, https://youtu.be/ZZQHu2WyvvI, https://youtu.be/3z5s7PuiZEY), https://youtu.be/o076lixXduU, https://youtu.be/pcp_MwT_paE, https://youtu.be/8qeEqJ5FKLI, and https://youtu.be/2rTC47_iwmg (accessed December 16, 2018).

[27] The manuscript described is cataloged as Archivo Municipal de Burgos SJ 1/1. In this 1091 royal donation, King Alfonso VI concedes valuable resources (a grazing reserve, a mill, a communal oven, and Jewish tax collections) to the Monastery of San Roberto de Casa Dei, located in the vicinity of the city of Burgos. See the *Deciphering Secrets* Youtube.com video https://youtu.be/jV5AWmjBddg (accessed December 16, 2018).

[28] The present Courersa.org-University of Colorado MOOC titled, *Burgos: Deciphering Secrets of Medieval Spain*, is accessible at https://www.coursera.org/learn/burgos-deciphering-secrets-medieval-spain (accessed December 16, 2018).

Figure 2:
Arqueta de Marfil y Esmaltes (Ivory chest with enamel).
Monasterio de Santo Domingo de Silos (Burgos). Date: 1026.

Source: Roger Louis Martinez-Davila with permission of the Museo de Burgos, 2018. Creative Commons license BY-NC-ND 4.0.

The key to the course is the SILReST Paleographic Method developed by Dr. Martinez-Davila for online audiences. This method teaches paleography to non-specialists in two to three weeks of instruction. SILReST is an initialism, each letter represents one of six strategies. Briefly, we explain these strategies to students as follows:

- S. Strategy #1 - Scan the entire document before attempting to transcribe it. It is important for you to become familiar with how the scribe writes. Repeatedly scanning a document will accustom your eyes to the "hand" of the scribe.
- I. Strategy #2 - Identify those letters, abbreviations, and numbers that you can immediately recognize. This is very straightforward, but it is the beginning of finding your way into the document. Finding easy-to-recognize letters will help you to appreciate how much you can already see and spur you along to uncover other letters and words.
- L. Strategy #3 - Locate common words to (a) understand how the scribe connects their letters together and (b) recognize other alphabetical letters and numbers. This strategy helps you identify letters that are hard to recognize. If you are flexible in terms of how a common word might be spelled, then, you will be able to see many curious spellings of words you know. More importantly, you can find new letters using this strategy.

- R. Strategy #4 - Recognize the abbreviations used in the document and if they vary within the document. Finding and marking abbreviations makes your task easier because it reminds you that some words on the page are not complete words at all. Rather, they are almost nonsensical connections of letters. Find the abbreviations so that your eyes and mind do not attempt to create words that do not exist on the page.
- S. Strategy #5 - Search for English-Spanish cognates (those words that share similar meanings and spellings in English and Spanish) to identify more letters and connections. Cognates are helpful because you can work "backward" into reading letters on the page. For example, if you know the word might be "jurisdiction" in English and therefore is 'jurisdicción" in Spanish, then you can begin to identify hard to read letters within the word on the page.
- T. Strategy #6 - Type or write your transcription and leave plenty of room to add edits. Creating a transcript will help you fill in the blanks as you work through those last, hard to read letters and words.

The SILReST method is now integrated into each *Deciphering Secrets* MOOC's toolkit of practice exercises, examinations, and transcription projects. These MOOCs often attract former students to re-enroll and thus their paleography proficiency undoubtedly improves with each course.[29]

To ensure that students create manuscript transcriptions that are standardized, but not overly complicated to record, we employ a basic transcription approach.[30] Students prepare individual transcriptions of selections of manuscripts, denoted as text blocks, and then share these in online discussion forums for peer-review. Therefore, students can comment and assist one in another in exploring the complicated nature of more illegible pen markings.

The instructions we provide to students read as follows:

- Choose 4 Text Blocks. For your manuscript image, try to transcribe the text in four (4) of the blocked areas as best as you can. For example, areas A, B, C, and D. Or, for example, E, F, G, and H.

[29] A video overview of the SILReST method, which is presented to the students, is viewable at https://youtu.be/N4PgmILwaKw (accessed December 16, 2018). A video tutorial demonstration of how to use SILReST with a fifteenth century Spanish manuscript is available at https://youtu.be/aI9nS98H1Is ("Advanced Paleography: Learning a Script from the 15th-century using SILReST, Part 1") and https://youtu.be/iy4wDa9oXX0 ("Advanced Paleography: Learning a Script from the fifteenth century using SILReST, Part 2") (accessed December 16, 2018).

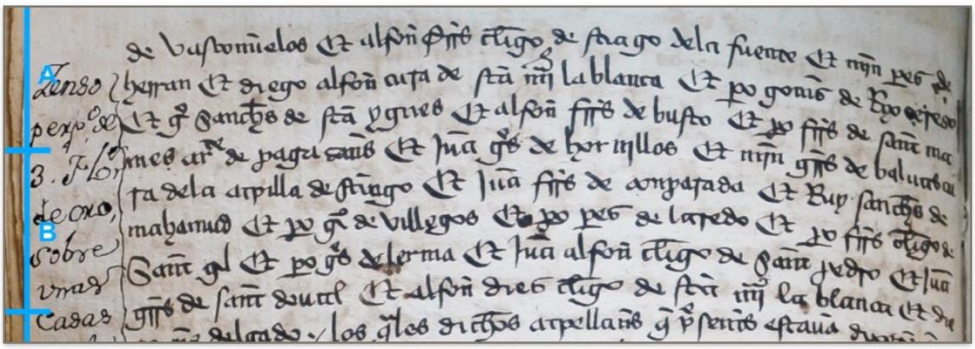
[30] For a discussion of various norms within Spanish transcription see José Miguel López Villalba, "Normas españolas para la transcripción y edición de collecciones diplomáticas," *Espacio, Tiempo y Forma. Serie III, Historia Medieval,* no. 11 (1998): 285-306.

- Formatting your transcription:
    - First Line. The first line should indicate what text blocks you transcribed. For example: "Blocks A, B, C, and D"
    - Use Line-Breaks. Type in your transcription as a line-by-line recording. In other words, use line-breaks.
    - Marginalia. You do not need to type in marginalia (the writing on either side of the main section of text). If you choose to record marginalia, include the word and a colon, "Marginalia:", in front of the text.
    - Abbreviations. If you know the abbreviation, please type the abbreviation and then spell the entire word out. The spelled-out word should appear inside of (parentheses). For example: "dho (dicho)" or "t* (testigo)".
    - Missing Letters and Words. If you cannot read a section of the manuscript text, then you should type one period " . " for a missing individual letter. If you cannot read more than one letter, or an entire word, then you should record three periods " …".
- Post Your Transcription in the Forum.
- Provide Feedback to Other Students. Comment on the transcriptions of your fellow students in that same forum. Please be kind and offer constructive criticism.

In their transcription formatting, then, students have begun to encode the text, which will be necessary for developing finished transcriptions.[31] For example, Figure 3 presents a manuscript selection and a portion of a student transcription from the summer 2016 MOOC. Although manuscripts often follow formulaic patterns, in the MOOCs, we only provide students with essential information about the manuscript, such as the date and a short abstract, and simple transcription instructions. Student paleographic accuracy might improve with better instructions regarding each documents key components.

[31] For more on the importance of clear, simple guidelines for crowdsourcing projects see Polly Duxfield, "The Practicalities of Collaboratively Digitally Editing Medieval Prose: The *Estoria de Espanna* Digital Project as a Case Study," *Digital Philology* 7, no. 1 (Spring 2018): 78.

Figure 3:
Sample Student Transcription from Deciphering Secrets MOOCs



ACB Registro 9 f.136v - Blocks A, B, C, and D

discussion posted 6 months ago by [redacted]

Blocks A, B, C, and D

A.

1. de vascomielos et alfon* (alfonso) frr*s (ferrandes) cl*igo (clerigo) de stiago dela fuente et mi*n (martin) p*es (peres) de

2. herran et diego alfon* (alfonso) cura de st*a (santa) mi* (maria) la blanca et p*o (pero) gom*s (gomes) de Eyo sereso

3. et que sanch*s (sanches) de sta* (santa) ygnes et alfon* frr*s (ferrandes) de busto et p*o (pero) frr*s (ferrandes) de sant* (santa) ma

B.

4. mes ar* de paga san*s et ju*a (juan) g*s (gonsales) de hornillos et mi*n (martin) grr*s (gutierres) de baluasai

5. ra dela atpilla de s*tiago (Santiago) et jua* frr*s (ferrandes) de conparada et Euy sanch*s (sanches) de

6. mahamud et p*o (pero) gr* de Villegos et p*o (pero) p*rs (peres) de laredo et p*o (pero) frr*s (ferrandes) cl*igo (clerigo) de

7. Sant* g.l et p*o (pero) g*s (gonsales) de lerma et jua* (juan) alfo* (alfonso) cl*igo (clerigo) de sant* pedro et jua* (juan)

Source: Roger Louis Martinez-Davila, 2018. Creative Commons license BY-NC-ND 4.0.

Student Paleographic Accuracy: Paleographic Exams and Transcriptions

Most scholars have difficulty with medieval transcription work because of the perception that paleography is more "art" than "method". However, our students repeatedly demonstrate that they can understand a substantial amount of medieval handwriting when they use SILReST. Our method consequently shows that paleography is a skill that can be acquired by non-specialist rather quickly. Still, the question of accuracy is the bugaboo of crowdsourcing projects. For this reason, we employ two different evaluative methods (machine-grade paleographic examinations and peer-review of student transcriptions) to assess students' paleographic proficiency so that we can measure the effectiveness of SILReST as an online paleography pedagogy and to assess the quality of students' crowdsourced transcriptions.

For this present paper, we also performed an intensive expert analysis of a representative sample of forty-eight students from our summer 2016 MOOC. This student cohort was one of five in the course. Each cohort completed all paleography exams ***and*** transcribed four blocks of text from fourteenth or fifteenth-century manuscripts from the Cathedral of Burgos. The anonymized summer 2016 student performance data indicates *Deciphering Secrets* MOOCs students are highly accurate paleographers.

Paleographic Exams

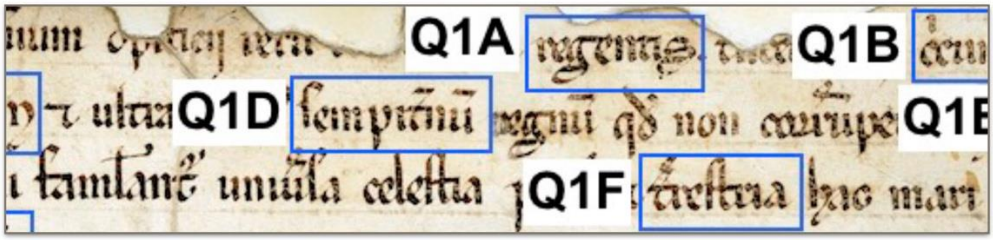
First, in each course, we utilize examinations to evaluate the caliber of each student's individual ability to perform paleographic interpretation. The paleography exams used five quantitatively-evaluated machine-graded instruments to gauge student proficiency in terms of:

- Ability to distinguish individual letters from one another,
- Ability to recognize individual letters within words,
- Ability to identify abbreviations, and
- Ability to transcribe small sections of texts.

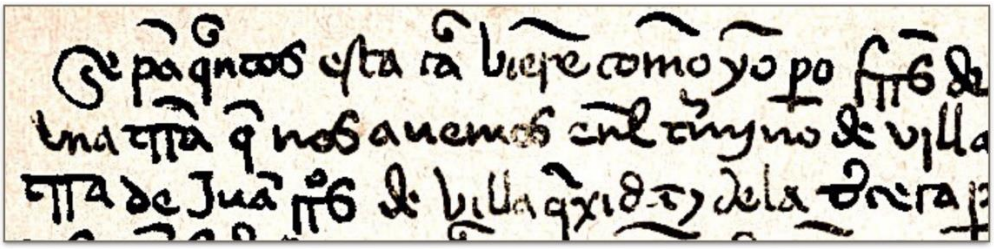
Exams #1 and #2 were introductory thirteenth-century Spanish paleography, Exam #3 was intermediate fifteenth-century paleography, and Exams #4 and #5 were advanced fifteenth-century paleography. (Figure 4 presents sample manuscript images from the paleography exams.)

Figure 4: Paleography Exam Samples (Summer 2016)

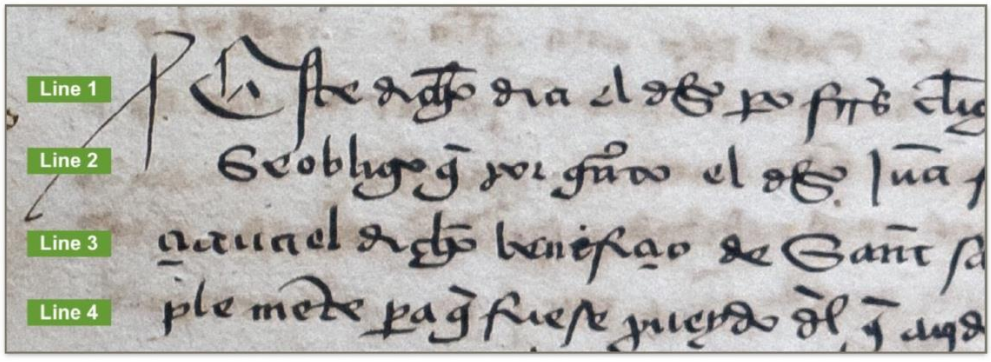
Exam #1 - Introductory Paleography - 13th Century Letters



Exam #3 - Intermediate Paleography - 15th Century Letters and Abbreviations



Exam #5 - Advanced Paleography - 15th Century Letters and Abbreviations

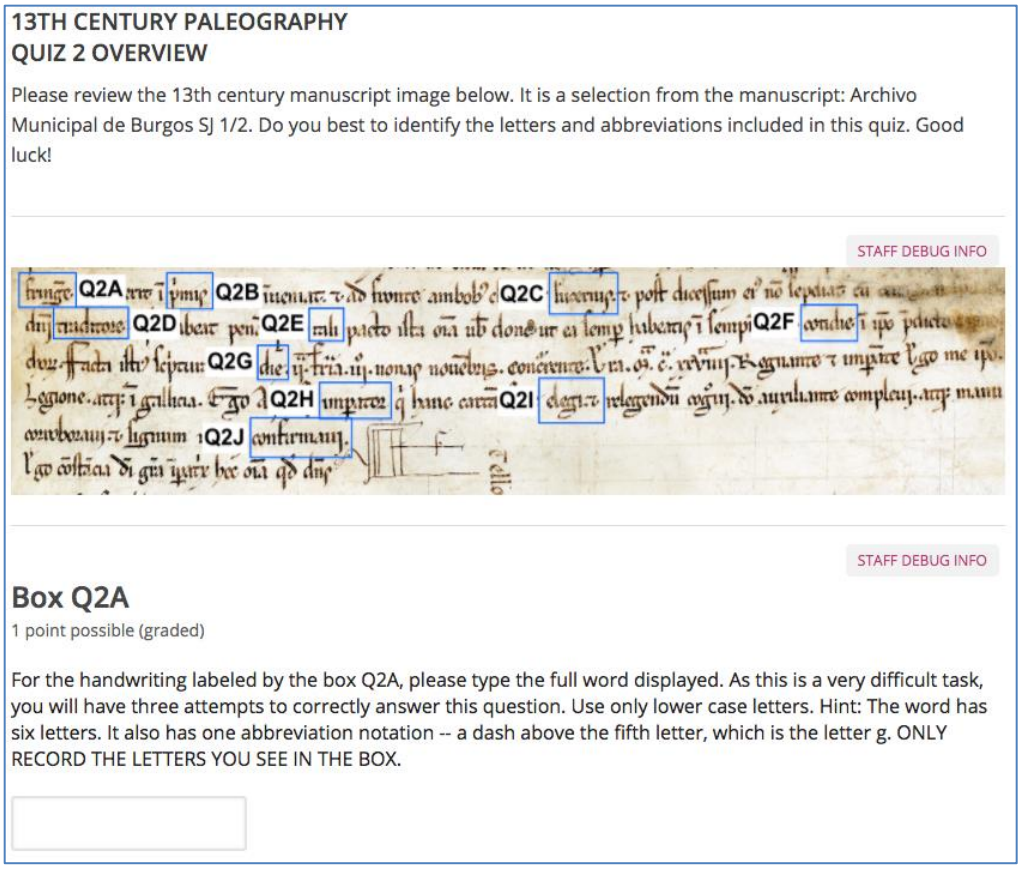


Source: Roger Louis Martinez-Davila, 2018. Creative Commons license BY-NC-ND 4.0.

To rate student paleographical proficiency, we primarily presented students with small selections of manuscripts and queried students to type in the letters they could identify. Figure 5 presents a sample question from Exam #2, which requires the student to identify a word that is in the box labeled "Box Q2A". As students are just beginning their paleography studies, which quickly ramp up in difficulty, we often provide some guidance. In this example, which is after students have been evaluated on their ability to count the number of letters in a word, we inform them that the word has six letters. Although this might appear to be an unnecessary and unfair aid for students, our interest is in simultaneously teaching and evaluating skill levels. In this example, even if the student knows there are six letters, there are 308,915,776 possible twenty-six-letter permutations for a six-letter word (i.e. 26 x 26 x 26 x 26 x 26 x 26 = 308,915, 776). In other words, these types of paleography "hints" do not undermine the credibility or efficacy of the exam questions. The answer to this exam question is: "fringe". It is an abbreviated word. The unabbreviated word is "infringere". For this specific question, students performed "excellent" and the average student score was 96.3%. However, student performance diminished as students continued the exam. On the final question (#10), which we rated as "very difficult", students earned an average score of 53.6%. Therefore, there appeared to be learner fatigue as the test became progressively more difficult. Nonetheless, the average score for Exam #2 was still 77.7%. For Exams #3, #4, and #5 (administered during the second week of paleography instruction), students evaluate fifteenth-century Spanish letters, abbreviations, and words. Fifteenth-century "hands", as they are commonly described by medievalists, are substantively more difficult to read because letters are connected into cursive scripts, whereas thirteenth-century hands feature individual unconnected letters. The average score for Exam #3 was 79.9%, Exam #4 was 67.3%, and Exam #5 was 69.4%; these are impressive paleography marks. For instance, let us consider a sample question from Exam #5 (Figure 6). It asks the student to transcribe the first three words that appear in Line 1 and offers the advice, "Please do not include the first letter that looks like a P. Hint: The first word has an uppercase letter. The second word has five letters. The third word is a common word." Again, we offer guidance to students to help them to learn paleography – rapidly. At this point, students will have only used the SILReST method for less than two weeks and we do not want them to abandon their studies. Instead, we supportively encourage their mastery of basic paleography. The answer to this question is the phrase, "Este dicho dia." For this specific question (#1), the first on this exam, students performed "below average" and the average score was just 57.6%. However, by the final question (#10) of the exam, students performed "excellent" and the average student score for this question was 91.1%. By the fifth paleography exam, instead of incurring the learner fatigue seen in Exam #2, student test performance generally improved. Thus, our paleography exams simultaneously

test proficiency and also continue to teach and enhance students' ability to identify words and letters.

Figure 5:
Sample Thirteenth-Century Paleography Exam Question (Summer 2016)



Source: Roger Louis Martinez-Davila, 2018. Creative Commons license BY-NC-ND 4.0.

Figure 6:
Sample Fifteenth-Century Paleography Exam Question (Summer 2016)

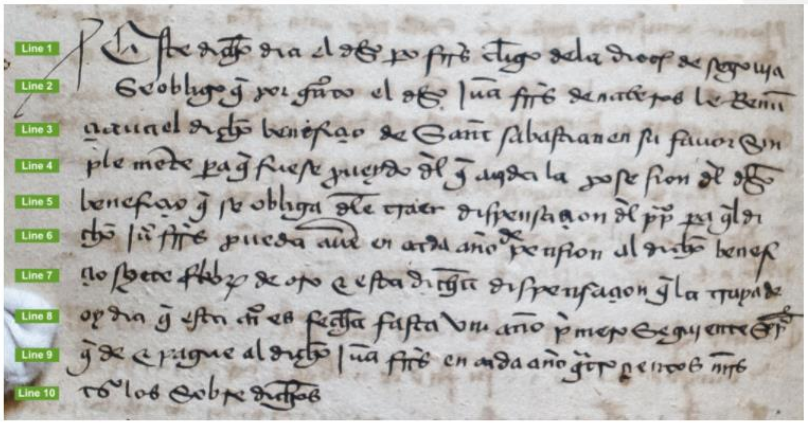
15TH CENTURY PALEOGRAPHY
QUIZ 5 OVERVIEW

Please review the 15th century manuscript image below. This selection is from manuscript: Archivo de la Catedral de Burgos, Registro 9. Do you best to identify the letters, words, and abbreviations included in this quiz. Good luck!

Hints:

- Remember to use the six strategies:
  - Strategy #1 - Scan the entire document before attempting to transcribe it.
  - Strategy #2 - Identify those letters, abbreviations, and numbers that you can immediately recognize.
  - Strategy #3 - Locate common words to (a) understand how the scribe connects their letters together and (b) recognize other alphabetical letters and numbers.
  - Strategy #4 - Recognize the abbreviations used in the document and if they vary within the document.
  - Strategy #5 - Search for English-Spanish cognates (those words that share similar meanings and spellings in English and Spanish) to identify more letters and connections.
  - Strategy #6 - Type or write your transcription and leave plenty of room to add edits.

STAFF DEBUG INFO



STAFF DEBUG INFO

Line 1

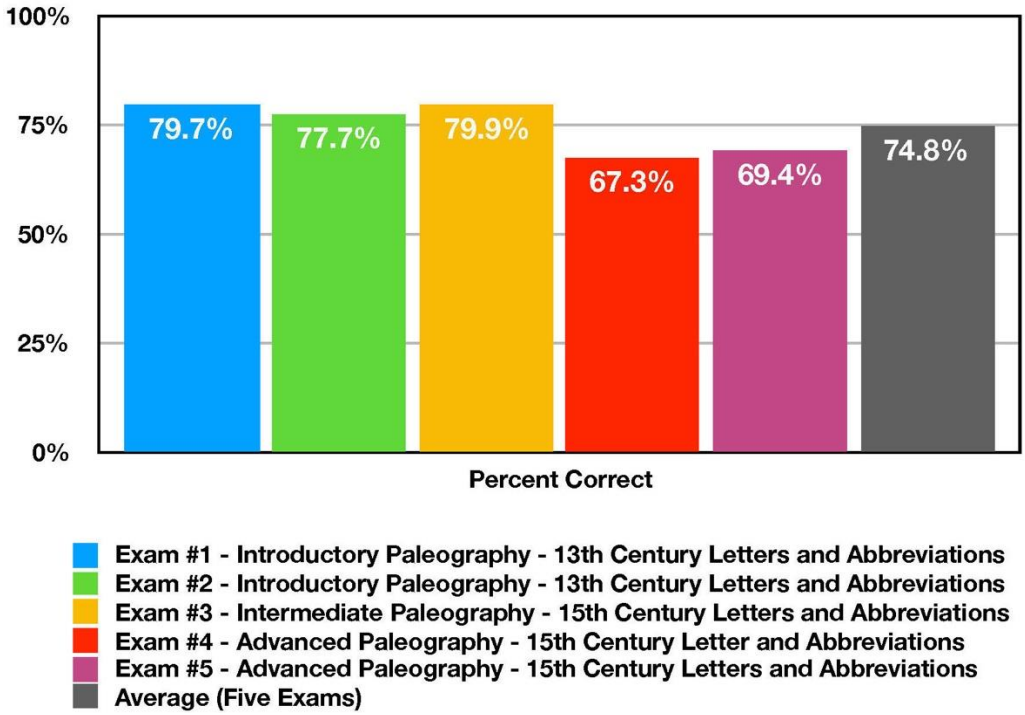
1 point possible (graded)

Please review Line 1. Please transcribe the first three words that appears in Line 1. Please do not include the first letter that looks like a P. Hint: The first word has an uppercase letter. The second word has five letters. The third word is a common word.

Source: Roger Louis Martinez-Davila, 2018. Creative Commons license BY-NC-ND 4.0.

Our evaluation of the forty-eight students' performance indicates that on average students recognized 74.8% of medieval Spanish handwriting (thirteenth and fifteenth-century scripts) when tested using five machine-graded paleographic exams (Figure 7). At the introductory level, or thirteenth-century script, students scored very high on their exams (79.7% and 77.7%). At the intermediate level, which was a fairly readable fifteenth-century handwriting that students specifically studied in practice lessons, exam scores remained high (79.9%). However, when presented with an entirely novel fifteenth-century hand, what we considered advanced paleography, students average scores dropped by about 10 percentage points (67.3% and 69.4%). Given students only engaged in two weeks of paleographical training, we consider this level of comprehension and accuracy to be quite good for non-specialists.

Figure 7:
Average Student Paleography Exam Scores



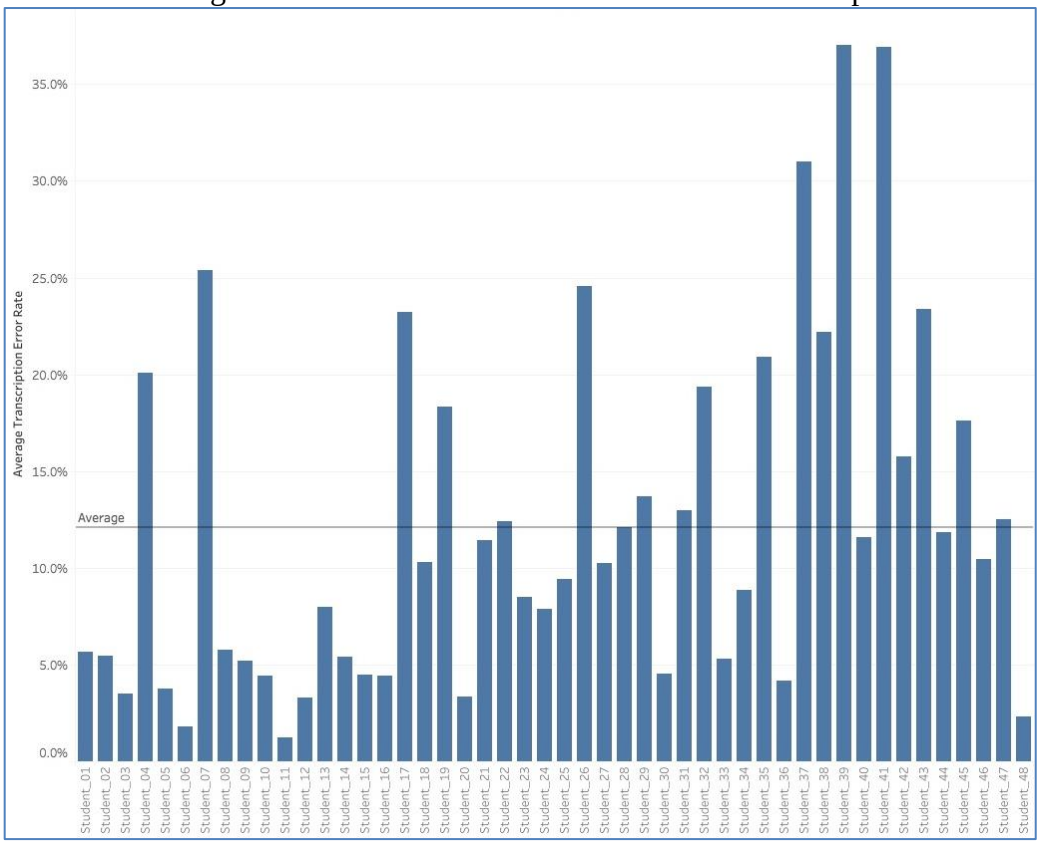
Peer-Reviewed Individual Transcriptions

One of the most important findings of our crowdsourcing effort is solid evidence demonstrating student paleographic accuracy substantially improves when students collaborate and use peer-review for preparing transcriptions of novel blocks of manuscript text. At the conclusion of each *Deciphering Secrets* MOOC, our class transcribes several manuscript selections. Students prepare their own individual

transcriptions and then post their transcriptions online in the course's discussion forums for peer-review. Students are then encouraged to enhance their original transcriptions using their classmates' feedback that includes suggestions and notations about problematic sections of transcriptions.

For the same group of forty-eight students discussed above, we performed an expert review of their individual transcriptions of four blocks of text from a cathedral manuscript. Specifically, our assessment of student transcriptions revealed a 12.1% average error rate; students correctly identify 87.9% of all the letters, numbers, and abbreviations from fifteenth-century manuscripts. (See Figure 8)

Figure 8:
Average Student Error Rate for Peer-Reviewed Transcriptions



Our analysis of each of the individual student transcriptions of manuscript required a character-by-character appraisal of three sections of the student's work. That is, we checked each student's first line of transcription, last line of transcription, and a machine-selected random line of transcription. The types of errors we noted were incorrectly identified or missing letters, numbers,

abbreviations, and spaces. Students were given the option of spelling out abbreviations or denoting them with an asterisk. To calculate an error rate, we first counted the number of characters (letters, numbers, and spaces) in each line and created a model transcription. Subsequently, we counted the number of student errors or deviations from the model transcription.

For example, the model first line transcription of Folio 136 verso without abbreviations was recorded as including eighty-seven characters. (See Table 2) However, if the student chose to use asterisks where abbreviations appeared in the manuscript, then the model first line with abbreviations was recorded as including seventy-six characters. As capitalization of letters in medieval manuscripts can be haphazard, students were not evaluated using this metric. To earn a 0.0% error rate a student would be required to transcribe the first line with no errors. Unsurprisingly, no student generated a perfect transcription. For example, Student #30's transcription included four errors and therefore corresponded eighty-three *correct* characters out of eighty-seven *model* characters. The error rate is therefore calculated with the following formula: Error rate = | (Student Transcription - Model Transcription without Abbreviations) / Model Transcription without Abbreviations | = | (83 – 87) / 87 | = 4.6%.

Table 2: Sample Evaluation of Student #30 Transcription Errors in Relationship to Model Transcription of Line 1 of Mss. ACB Registro 9, Folio 136 verso

Model Transcription without Abbreviations (Mwo)	Model Transcription with Abbreviations (Mw)	Student #30 Transcription (St)	Error Rate = | (St-Mwo)/Mwo | or Error Rate = | (St-Mw)/Mw |
de Vascomielos et Alfonso Fernandez clerigo de Santiago de la Fuente et Martin Perez de	de vascomi*elos et alfon* fr*rs cl*igo de stiago de la fuente et mi*n p*es de	de Vascocnielos et Alfon Fernandez clerigo de Santiago de la Fuente et Martin Perez de	= | (83-87)/87 | = 4.6% error rate
Model Character Count = 87	Model Character Count = 76	Character Evaluation using no abbreviations: Student used "c " and "n" incorrectly for the letter "m". Minus 2 points. Student did not see abbreviation for "Alfonso" and used "Alfon". Minus 2. Total missed is 4 points. Correct Character Count = 87 - 4 = 83	

In this manner, student error rates were calculated for the first, last, and random lines of each transcription. Finally, a cumulative error rate was calculated for each student. (See Table 3) The results of our study indicate that student error rates were quite consistent across their entire transcriptions and on average were only 12.1%.

Thus, the lion share of student transcriptions produced using SILReST are highly accurate. As the project progresses and matures, we will investigate how students' lack of knowledge of formulaic document structures and other idiosyncrasies affects error rates. As of yet, we have not evaluated this issue but we imagine it does contribute to error rates and therefore we will need to adapt our pedagogy.

Table 3:
Average Transcription Error Rates for 48 Student Transcriptions of
Mss. ACB Registro 9, Folio 136 verso

Transcription Line Analyzed	Error Rate
First Line Analysis	10.7%
Last Line Analysis	13.6%
Random Line Analysis	12.0%
Cumulative Analysis	12.1%

We believe that an overall 74.8% level of accuracy for average paleography exam scores and an average error rate of 12.1% for peer-reviewed transcriptions for large classes is acceptable. Some students are scoring much higher, and those more complete transcriptions prepared by exceptional students will eventually be used for text-mining and preparing diplomatic transcriptions. Again, von Ahn's program ReCAPTCHA only needs five correct deciphers to confirm a word for the computer. We are reaching such numbers through the MOOC in many cases. At the same time, for large-scale big data searches, we are open to some level of incompleteness or messiness to get results, knowing that as the data is fine-tuned the results will become more exact. Essentially, with thousands of transcriptions in various states of preparedness, we need to just analyze them. By using even unstructured data we can gain insights now that would take many years to gain by waiting to structure the data completely. Simply getting a sense of the general trends or broad outlines over time will allow us to better structure the fully processed data. The larger the corpus or database (even with incomplete transcriptions) will allow machine-learning algorithms to accurately calculate trends than smaller more accurate data sets. However, as historians, we believe broad and narrow analyses can positively reinforce each type of finding.[32]

To underline the point, the key is not necessarily that the transcriptions are 100% methodologically sound, but rather that there is enough text available for fruitful searches. As Patrick Sahle's notes:

[32] For more on these themes, see Mayer-Schönberg and Cukier, *Big Data*, 35-36, 45-48.

> The criteria here must be content and usability. As soon as the publication makes a substantial amount or percentage of the intended documents or text available so that it can be fruitfully used in research, we may call it an edition. The question of quality is even harder to answer, particularly in times of upcoming public, social, crowdsourced editions. … again the question of usefulness is crucial. Does the edition provide a reliable proxy for the documents? Can scholarly research be trustworthily based on the edition without the need to go back to the originals?[33]

We hope that scholars will eventually find our transcriptions sufficiently accurate that they can just as confidently use them for research as transcriptions from the *Colección de documentos inéditos*. Crowdsourced transcription projects and digitalized manuscripts posted online will make more information readily available to scholars across the globe. However, such efforts are often limited to only a portion of a larger collection or codex. And, the amount of medieval manuscripts that have been digitalized is very small. In 2011, Stanford University estimated that less than half a percentage had been digitalized.[34] Consequently, historians will still need to go to the archives, especially considering that certain material aspects of documents will never be able to be conveyed digitally (e.g., historians of the book will want to examine the physical form of the book not just its content).[35]

Processing text

One challenge for all digital transcription projects is to determine the editorial framework appropriate for the type of text and/or for the type of research questions.[36] The framework will help team members to figure out how to address in their transcriptions orthographic inconsistencies, multiple abbreviations for the same word, variations in scribal hands, lack of punctuation, and so forth. For instance, scribal variations might allow us to identify individual scribes, but has little bearing on interfaith relations in medieval Spain. Consequently, is keeping intact these distinctions necessary? What other information embedded in the manuscript is meaningful and should be maintained in transcription and what textual variations are unnecessary to maintain? For example, it is important to document and record the individual church leaders' and scribes roles in shaping the historical record. One specific case from the Cathedral of Burgos relates to Juan Diez de Mahumud, who served as a chaplain and member of the cathedral chapter

[33] Patrick Sahle, "What is a Scholarly Digital Edition?" in *Digital Scholarly Editing: Theories and Practices*, eds. Matthew James Driscoll and Elena Pierazzo (Open Book Publishers, 2016), 35-37.
[34] Toby Burrows, "Medieval Manuscripts and Their (Digital) Afterlives," in *The Routledge Handbook of Digital Medieval Literature*, eds. Jen Boyle and Helen Burgess (London: Routledge, 2018), 159.
[35] For more on this theme, see John Palfrey, *BiblioTECH: Why Libraries Matter More Than Ever in the Age of Google* (New York: Basic Books, 2015), 70-71.
[36] Pierazzo, "Digital Documentary Editions," 15.

in 1391.[37] Diez appears to have been a *morisco*, or Muslim convert to Christianity, and in the early fifteenth century he assumed more prominent roles on the chapter. Throughout his career, Diez facilitated the work of many Muslim carpenters on church projects. However, as we move to evaluate larger thematic issues, such as residential patterns via big data, a leader's or scribe's identity and role may be less important. In either case, we are capturing all of the data by recording all of the text on the page.

In the past, with printed documentary editions, these questions were more urgent, and the answers were often determined by the research agenda of the original project members. Fortunately, today, with digital transcriptions, we don't need to worry as much about how to maintain variation without limiting the ability of the data to be machine queried. New digital technology allows us to maintain multiple layers within a document (normalized text, semi-diplomatic text, and diplomatic text) that can be turned on and off as needed by individual researchers.[38] Simply being aware of potential other uses for the data and being able to keep in place as much meaningful variation as possible within the data sets will help make the metadata more useful to people with other research questions or interests.[39]

We also need to be aware of the corrections and emendations that we make to the text, and how in the process of transcription we are developing the first draft of a "critical edition" of these documents. At the same time, these digital editions are always open to change and amendments. These are open enterprises.[40]

Text mining – Data from Plasencia

Text-mining is probably one of the more popular options for analyzing written texts. This approach harnesses the computer to read and process large amounts of text and to transform that text into data. There are limitations to the computer's ability to count words, such as orthographic variations over time. Though, programs, such as VARD 2, can address some issues arising from spelling variations. Nonetheless, this imperfect process still produces results. The first step is simply to count the words, and frequency of occurrence. Then, to remove, so-called "stop words," before analyzing the corpus for themes. This process is especially useful in looking at homogeneous manuscripts such as the Actas Capitulares. The creation of dictionaries will be necessary for text-mining across medieval collections.[41]

---

[37] Manuscript Archivo de la Catedral de Burgos (ACB). Registro 2, Folis 1r-1v.

[38] Elena Pierazzo provides a more thorough discussion of the multifaceted nature of digital documentary editions in Pierazzo, "Digital Documentary Editions", and Elena Pierazzo, "A Rationale of Digital Documentary Editions," *Literary and Linguistic Computing* 26, no. 4 (2011): 463-477.

[39] See also Rawson, "Curating Menus," 64-66.

[40] Sahle, "What is a Scholarly Digital Edition?" 24-26, 29. Accessed June 21, 2018.

[41] Widner, "Toward Text-Mining," 136-138.

Even with incomplete transcriptions, large amounts of text can be mined to find big data correlations that capture broad patterns and point the way towards new areas for more narrow case studies.[42] For example, via international relationships created through our first MOOC on Plasencia, Dr. Ignacio Perez, a professor of industrial engineering and statistics, and Ms. Piedad Catalina Gonzalez Orjuela, a Master of Science candidate, at the Universidad Tecnologica de Pereira (Colombia) applied text-mining to the transcriptions from the Actas Capitulares of Plasencia. In her thesis, Ms. Gonzalez Orjuela attempted to reconstruct the roles of women in thirteenth- and fourteenth-century Plasencia, as well as to understand the nature of the cathedral's administrative functions.[43] Through cluster analysis she discovered within the institutional framework of the cathedral, "the only reference to women was as a member of a family, charged to look after children or as the spouse of an influential man."[44] Although data was limited about women, Gonzalez Orjuela was able to construct extensive "word clouds" that revealed the fundamental structures that ordered the life of a medieval cathedral. These word clouds exposed repeated terms in the manuscripts that showed categories of governing officials, church parishes, agricultural production, property boundaries, given and surnames, real property and possessions, income streams, and time delineations.[45] For example, Figure 9 is a word cloud displaying the most commonly used words, arranged by size and color-coding, in the Actas Capitulares. This cloud demonstrates that the Actas Capitulares speak extensively about church properties, namely houses (casas) and vineyards (viñas).

[42] For more on the benefits of using big, messy datasets, see Mayer-Schönberg and Cukier, *Big Data,* 191.
[43] Piedad Catalina Gonzalez Orjuela, "Aplicación de minería de texto a documentos históricos: Reconstrucción del rol de las mujeres del siglo xiv y xv en Plasencia-España, a partir de documentos históricos," Master's thesis, Universidad Tecnologica de Pereira (Colombia), 2017.
[44] Gonzalez Orjuela, "Aplicación de minería de texto a documentos históricos," 88.
[45] Ibid., 88-89. For more about cluster analysis, see Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques* (Burlington: Morgan Kaufmann, 2011).

Figure 9:
Word Cloud of Most Commonly Used Words in the Actas Capitulares

Source: Piedad Catalina Gonzalez Orjuela, "Aplicación de minería de texto a documentos históricos: Reconstrucción del rol de las mujeres del siglo xiv y xv en Plasencia-España, a partir de documentos históricos," Master's thesis, Universidad Tecnologica de Pereira (Colombia), 2017: 80.

Diplomatic and Semi-diplomatic Transcriptions – Data from Burgos

We also need to prepare finished texts that include ancillary material such as hyperlinks and annotations.[46] This process is time-consuming, but it is crucial to study real people: how they lived, what they valued, what they believed, and how people's lives, values, and beliefs changed over time. Organizing the material though can be challenging. Current taxonomies often are too narrow (though we plan to use them, especially to create geospatial databases for mapping) for text analysis and linking vast quantities of text. Tagging, therefore, is becoming more popular with big data sets, because it is more flexible and makes it easier to distinguish material and to find similar types of material. Tagging consequently might be especially useful early in the process despite its imprecision.[47] Our transcription guidelines already provide for some tagging by students (e.g., the

[46] Joris van Zundert, "Barely Beyond the Book?" in *Digital Scholarly Editing: Theories and Practices*, eds. Matthew James Driscoll and Elena Pierazzo (Open Book Publishers, 2016), 90-91. http://www.jstor.org/stable/j.ctt1fzhh6v.9 (accessed June 21, 2018).
[47] Mayer-Schönberg and Cukier, *Big Data*, 42-43.

placement of abbreviation marks). Tagging of individual documents might also help us to find metadata previously buried by the methods used to prepare the card catalog of particular collections. Similarly, we hope to mimic or employ the Textual Communities system of capturing variations and revisions of transcription as is used with the *Canterbury Tales*.[48]

The potential to shift between the digital image of a document (facsimile), diplomatic transcription, and annotated version opens the door to different types of reading and interpretation, and thus offsets some of the limitations placed on the data by the project originators. For each of our crowdsourced transcriptions, we intend to use our website, http://www.decipheringsecrets.com, to display the digital image of the manuscript alongside of:

- the finalized diplomatic transcription (characters and abbreviations as they appear in the manuscript),
- semi-diplomatic transcription (full-text without abbreviations), and
- annotated/tagged documents with economic, religious, social, familial, geographic, and other historical dimensions.

One initial example of a complete student transcription from Burgos with limited corrections will provide some idea of this process. The document for 24 April 1431 explains how the cathedral chapter met to confirm the lease of certain homes in the Muslim quarter to the Muslim, Audalla de Valladolid.[49] (See Figure 10 and accompanying transcriptions.)

At this point, team members will continue to facilitate new transcriptions as our MOOCs continue to be taught over the next several years as well as make final edits and tag documents for use in our relational database. However, the interest of citizen-scholars to do more than transcribe might open the door to additional collaboration on this end of the process as well.[50] We are also very cognizant of the need to preserve digital materials.[51] Consequently, for the long-term preservation

---

[48] Textual Communities is located at https://textualcommunities.org/app (accessed December 16, 2018). Textual Communities' system is discussed in Peter Robinson, "Some principles for making collaborative scholarly editions in digital form" *Digital Humanities Quarterly* 11, no. 2 (2017) http://digitalhumanities.org:8081/dhq/vol/11/2/000293/000293.html (accessed November 29, 2018).

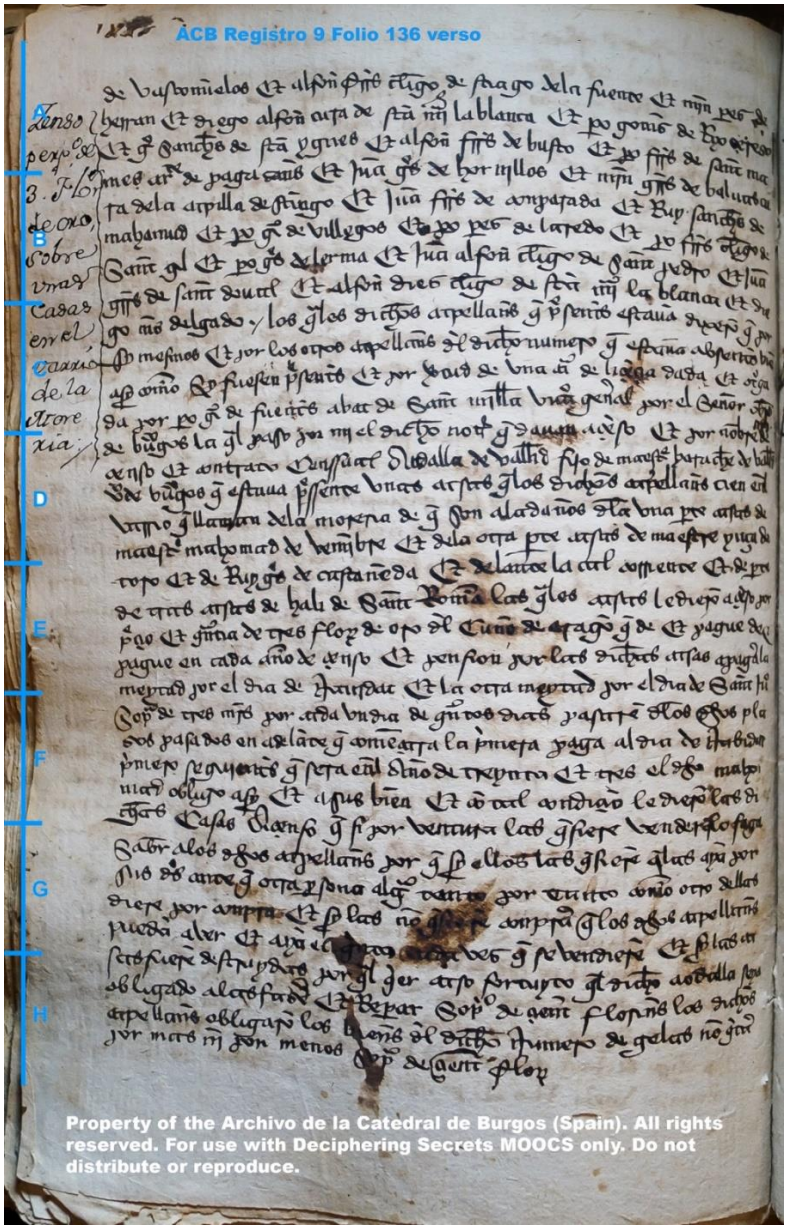
[49] Manuscript Archivo de la Catedral de Burgos (ACB). Registro 9, Folio 136v.

[50] Some volunteers have successfully undertaken the preparation of mark-ups for the *Transcribe Bentham* project. See Tim Causer and Melissa Terras, "Crowdsourcing Bentham: Beyond the Traditional Boundaries of Academic History," *International Journal of Humanities and Arts Computing* 8, no. 1 (2014): 52 https://doi.org/10.3366/ijhac.2014.0119 (accessed December 16, 2018).

[51] For a succinct discussion of various problems with preserving digital material from "data rot" to changing formats, see John Palfry, *BiblioTech*, 110ff.

of the finished transcriptions, we are working with the UCCS library system to store hard electronic copies of the materials generated in the MOOC.

Figure 10:
Manuscript ACB Registro 9, Folio 136v.



Source: Roger Louis Martinez-Davila with permission of the Archivo de la Catedral de Burgos, 2018. Creative Commons license BY-NC-ND 4.0.

Diplomatic Transcription of ACB Registro 9, Folio 136v.

BLOCK A

de Vascomielos et Alfon* Frr*s cl*igo de Stiago de la Fuente et My*n P*es de
Herran et Diego Alfon* cura de St*a M*ri la Blanca et P*o Gom*s de Eyo Cereso
et G* Sanch*s de Sta* Ygnes et Alfon* Frr*s de Busto et P*o Frr*s de San*t Ma

BLOCK B

mes ar* de paga san*s et Jua* G*s de Hornillos et My*n Gr*rs de baluuscu
ra de la capilla de Sti*ago et Jua* Frr*s de conparada et Ruy Sanch*s de
Mahamud et P*o Gi* de Villegos et P*o P*es de Laredo et P*o Frr*s cl*igo de
San*t Gil et P*o G*s de Lerma et Ju*a Alfon* cl*go de San*t Pedro et Jua*

BLOCK C

G*rrs de Sant*doval et Alfon* Dies cl*go de St*a M*ri la Blanca et Die
go M*s Delgado y los q*les dichos capellan*s q* p*sents estaba* dixero* q* por
sy mesmos et por los otros capellan*s d*l dicho numero q* estava* absent*s via*
asy como sy fuesen p*esents et por virtud de una ca* de lice*cia dada et ot*ga
da por P*o G*i de Fuent*s de Abat de San*t Arilla vica* gen*al por el señor ob*p

BLOCK D

de Bu*gos la q*l paso por mi el dicho not* q* davan a ce*so et por no*bre de
censo et contrato menssual Audalla de Vall*id fijo del maestr* Harache de Vall*id
v* de Bu*gos q* estava p*sente unas casas q* los dichos capellan*s tien en*l
varrio q* llaman de la Moreria de q* son aladaños d*la una p*re casas de
maestr* Mahomad de Vem*ibre et de la otra p*re casas de Maestre Yuça de

BLOCK E

Toro et de Ruy G*s de Castañeda et delante de la cal corriente et de p*re
de tras casas Hali de San*t Roma* las q*les casas le diero* a ce*so por
p*cio et qu*tia de tres flori* de oro d*l Cuño de Arago* q* de et pague des
pague en cada año de censo et pension por las dichas casas a paga la
meytad por el dia de Navidat et la otra meytad por el dia de San*t Ju*

BLOCK F
24. sop* de tres m*rs por cada un dia de qu*tos dias pasare d*los d*hos pla
25. sos pasados en adela*te q* come*zarra la p*mera paga al dia de Nabidat
26. p*mero seguientes q* sera en*l año de treynta et tres el d*ho Maho
27. mad obligo asy et a sus bien* et co tal condicio* le diero* las di
28. chas casas si censo q* si por ventura las q*siere vender a lo faga

BLOCK G
29. sap*r a los d*hos capellan*s por q* sy ellos las q*siere q* las ay*a por
30. sus d*s ante que otra p*sona alg* tanto por quanto como otro dellas
31. diere por compra et si las no q*siere compra* q* los d*hos capellan*s
32. pueda* aver et ayan el q*nta cada vez q* se vendiere* et sy las ca

BLOCK H
33. sas fuere* destrydas por q*lq*er caso fortuyto q*l dicho Audalla seya
34. obligado a las fase et reparar sop* de cient* florin*s los dichos
35. capellane* obligaro* los bien*s del dicho numero de que las no* q*ta
36. por mas ni por menos sop* de cient* flori*

Semi-Diplomatic Transcription of ACB Registro 9, Folio 136v.
BLOCK A
1. de Vascomielos et Alfonso Fernandes clerigo de Santiago de la Fuente et Myrtin Peres de
2. Herran et Diego Alfonso cura de Santa Maria la Blanca et Pero Gomes de Eyo Cereso
3. et Garcia Sanches de Santa Ygnes et Alfonso Fernandes de Busto et Pero Fernandes de Santa Ma

BLOCK B
4. mes arcediano de paga san*s et Juan Gomes de Hornillos et Martin Gutierres de baluuscu
5. ra de la capilla de Santiago et Juan Fernandes de conparada et Ruy Sanches de
6. Mahamud et Pero Garcia de Villegos et Pero Peres de Laredo et Pero Ferrandes clérigo de
7. Santo Gil et Pedro Gomes de Lerma et Juan Alfonso clerigo de Santo Pedro et Juan

BLOCK C
8. Gutierres de Santdoval et Alfonso Dies clerigo de Santa Maria la Blanca et Die

go Munos Delgado y los quales dichos capellanes que presentes estavan dixeron que por
sy mesmos et por los otros capellanes del dicho numero que estavan absentes via*
asy como sy fuesen presentes et por virtud de una carta de licencia dada et otorga
da por Pero Garcia de Fuentes de Abat de Sant Arilla vicario general por el señor obispo

BLOCK D

de Burgos la qual paso por mi el dicho notario que davan a censo et por nombre de
censo et contrato menssual Audalla de Valladolid fijo del maestre Harache de Valladolid
vecino de Burgos que estava presente unas casas que los dichos capellanes tien en el
varrio que llaman de la Moreria de que son aladaños de la una parte casas de
maestre Mahomad de Vem*ibre et de la otra parte casas de Maestre Yuça de

BLOCK E

Toro et de Ruy Gomes de Castañeda et delante de la cal corriente et de parte
de tras casas Hali de Sant Roman las quales casas le dieron a censo por
precio et quantia de tres florines de oro del Cuño de Aragon que de et pague des
pague en cada año de censo et pension por las dichas casas a paga la
meytad por el dia de Navidat et la otra meytad por el dia de Sant Juan

BLOCK F

sopeña de tres maravedis por cada un dia de quantos dias pasare de los dichos pla
sos pasados en adelante que comenzarra la primera paga al dia de Nabidat
primero seguientes que sera en el año de treynta et tres el dicho Maho
mad obligo asy et a sus bienes et con tal condicion le dieron las di
chas casas si censo que si por ventura las quisiere vender a lo faga

BLOCK G

saber a los dichos capellanes por que sy ellos las quisieren que las ay*a por
sus d*s ante que otra persona alguna tanto por quanto como otro dellas
diere por compra et si las no quisiere comprar que los dichos capellanes
puedan aver et ayan el quanta cada vez que se vendieren et sy las ca

BLOCK H
32. sas fueren destrydas por qualquier caso fortuyto que el dicho Audalla seya
33. obligado a las fase et reparar sopena de ciento florines los dichos
34. capellanes obligaron los bienes del dicho numero de que las no* quita
35. por mas ni por menos sopena de ciento florines

Conclusion

Though the present project is still in its infancy, the pedagogy for teaching Spanish paleography online has developed over the years with the result that more documents are being transcribed. The creation of trained volunteers and new students enrolling in the MOOC will allow further documents to be transcribed in the years to come. The next phase is to mark the records up and prepare them for public consumption. In the meantime, we hope that the discussion here will help others to consider how to move forward with their own transcription projects – making them aware of both the challenges involved as well as the potential rewards for new scholarship and research opportunities.

The challenge for large-scale transcription and database projects is that the data needs to be shared to be useful. Current academic norms, however, do not encourage data sharing. If datasets (such as ours) can be merged with other datasets and shared with others scholars, though, we will see the development of more robust online collections of records that will allow other users to develop discrete questions or lines of inquiry not envisioned by the original constructors of the data.[52] This give-and-take, as well as the need to invite others via crowdsourcing to participate in scholarly endeavors, is a hallmark of the digital humanities. Yet, it is also difficult for scholars to relinquish control.[53] This project is currently collaborating with the *Global Middle Ages Project*, and hopefully, this participation will make our transcriptions and other findings available to scholars studying other aspects of the Middle Ages.[54]

Digital humanities research is project-based and collaborative. It requires us to rethink authorship because a digital project is not a single individual working alone. In the case of a crowdsourcing transcription project that issue is even more

[52] For a fuller discussion of data sharing and its challenges, see Ruth Mostern and Marieka Arksey, "Don't Just Build It, They Probably Won't Come: Data Sharing and the Social Life of Data in the Historical Quantitative Social Sciences," *International Journal of Humanities and Arts Computing* 10, no. 2 (2016): 205-224.

[53] Peter Robinson, "The Digital Revolution in Scholarly Editing" in *Ars Edendi Lecture Series*, vol. 4, eds. B. Crostini, G. Iversen, and B.M. Jensen (Stockholm: Stockholm University Press, 2016), 199-200 http://dx.doi.org/10.16993/baj.h (accessed September 12, 2018).

[54] The origins of the Global Middle Ages Project and the importance of collaboration among scholars are discussed in Geraldine Heng, "Romancing the portal: *MappaMundi* and the global middle ages" in *The Routledge Handbook of Digital Medieval Literature*, eds. Jen Boyle and Helen Burgess (London: Routledge, 2018), 31-46.

relevant – literally, thousands of people will contribute to the project.[55] Furthermore, by training citizen-scholars in paleography, the MOOC increases general interest in Spanish history by broadening the discussion beyond academic circles. Pedagogically, the MOOC teaches students the basic facts about the late Middle Ages and early modern period, familiarizes students with primary source material, and trains them in transcription and crowdsourcing methods.

[55] For more on the issue of authorship, see Brudrick et. al., *Digital_Humanities*, 83.