Mining Art History: Bulk Converting Nonstandard PDFs to Text to Determine the Frequency of Citations and Key Terms in Humanities Articles

Amanda Wasielewski and Anna Dahlgren

Introduction

Text mining and other computational methods are not widely used in art historical scholarship.[1] One rationale for the lack of text mining in art history is that the field concerns itself with the study of visual and material objects rather than text-based ones. Although artists themselves have not produced huge volumes of text to study, art historians and critics *have* produced a large corpus of writing on art. This means that a computational methodology like text mining, which remains untested in art historiography, might provide insights into the state of the field of art history.[2] "Text mining" consists of extracting, sorting, and discovering patterns within a given set of text (typically a very large

[1] Matthew P. Long and Roger C. Schonfeld, "Preparing for the Future of Research Services for Art History: Recommendations from the Ithaka S+R Report," *Art Documentation: Journal of the Art Libraries Society of North America* 33, no. 2 (2014): 192–205.

[2] As far as we have been able to ascertain, computational methods have not yet been applied to art historiography. As Paul B. Jaskot notes, text mining has not been central to other types of art historical research either. See Paul B. Jaskot, "Digital Art History as the Social History of

How to cite this book chapter:
Wasielewski, Amanda, and Anna Dahlgren. "Mining Art History: Bulk Converting Nonstandard PDFs to Text to Determine the Frequency of Citations and Key Terms in Humanities Articles." In *Digital Human Sciences: New Objects—New Approaches*, edited by Sonya Petersson, 285–305. Stockholm: Stockholm University Press, 2021. DOI: https://doi.org/10.16993/bbk.l. License: CC-BY.

corpus) through automated/computational means. By processing bodies of text that are too large to sort through manually, such as a decade's worth of journal articles or a bibliography of books on a particular topic, text mining can uncover disciplinary trends and popular reference points unique to that dataset that provide a picture of the influences and biases contained therein.

In order to text mine an article in Adobe PDF format, it is necessary to export clean, plain text from the document. Most PDF articles today are already composed of searchable text (OCR) or have been created digitally and therefore have recognized text/characters. It is possible, therefore, to copy and paste out the text from articles manually when working with just one or a few articles. Manual extraction of text is, however, impractical for a large corpus. Text mining, like most computational methodologies, works best with the largest set of data possible. Indeed, large datasets *call for* the use of computational methodologies such as text mining because of how time-consuming it is to analyze large quantities of text manually.

The objective of this study is to develop a methodology for determining the most frequently appearing terms in a given set of articles, which is nearly impossible to do manually—even for one article. More specifically, the text mining method we outline addresses challenges that are particular to humanities journals such as those in art history, which do not adhere to a standardized format.[3] As noted, text mining is a new tool for art historians, and the majority of "digital art history" projects focus on image

Art: Towards the Disciplinary Relevance of Digital Methods," *Visual Resources* 35, no. 1–2 (April 3, 2019): 24.

[3] We are aware of that a large amount of scholarly writing in art history is published in monographs and edited volumes. These books are not easy to text mine at the present time, as many of them are either not available as e-books or not collected in a comprehensive way in any given online database. One reason for this is that the high cost of image permissions paired with centrality of images to the discipline makes mass digitization a difficult undertaking. See Long and Schonfeld, "Preparing for the Future," 203; Maureen Whalen, "What's Wrong with This Picture? An Examination of Art Historians' Attitudes About Electronic Publishing Opportunities and the Consequences of Their Continuing Love Affair with Print," *Art Documentation: Journal of the Art Libraries Society of North America* 28, no. 2 (Fall 2009): 13–22.

collections rather than text or historiography.[4] Since there is a shortage of research that addresses this particular area—that is, methodologies for text mining art history publications—the following chapter constitutes a unique contribution to the field of art history and has wider implications for humanities and social sciences digital scholarship in general.

Background

Journal databases for art history scholarship commonly store articles in PDF, which preserves the look and feel of printed journals in digital form. PDFs maintain the layout, pagination, and images for journal articles. The latter is particularly important for art history because we use images not merely as illustration but as essential components of our scholarly arguments.[5] Additionally, citation standards across disciplines demand page citations, which other online and text-based formats do not provide. For example, ebook formats are often problematic for citation because many of them are only available in file formats that are designed to vary the size of the text and the pagination according to the device and the readers' preference. Despite the advantages researchers gain by viewing journal articles in PDF format, the fact that articles may only be available in this format is a significant barrier for text mining.

Text mining in academic journals falls into four main categories: abstract mining, full text mining, metadata mining, and citation mining.[6] Text mining techniques for scientific articles in the

[4] See articles published in the *International Journal for Digital Art History*, on the website International Journal for Digital Art History, https://dahj.org.

[5] Keith Moxey, "Visual Studies and the Iconic Turn," *Journal of Visual Culture* 7, no. 2 (2008): 139.

[6] There are many examples one could cite from each of these categories of research, e.g.: Sharon Block and David Newman, "What, Where, When, and Sometimes Why: Data Mining Two Decades of Women's History Abstracts," *Journal of Women's History* 23, no. 1 (March 10, 2011): 81–109; David Westergaard et al., "Text Mining of 15 Million Full-Text Scientific Articles," BioRxiv: The Preprint Server for Biology, published July 11, 2017, https://doi.org/10.1101/162099; Yang Yang et al., "The Researcher Social Network: A Social Network Based on Metadata of

STEM (science, technology, engineering, and mathematics) fields are well developed, but there is far less research that addresses text mining in humanities scholarship.[7] Journals and databases in the natural and social sciences have aided the process of text mining by providing inbuilt tools for analytics as well as structured text formats for articles. According to Giovanni Colavizza and Matteo Romanello,

> Citation mining is by now a "solved problem" with respect to STEM literature. Despite some drawbacks and margins for improvement, the crawlers behind mainstream indexes such as Google Scholar, Web of Science, Scopus, and Dimensions are largely capable of mining most citations accurately. The main problem which is left open is the skewness in literature coverage and mining

Scientific Publications," *Proceedings of WebSci '09: Society On-Line*, Athens, Greece, Mars 18–20, 2009, https://eprints.soton.ac.uk/267156; Iana Atanassova, Marc Bertin, and Philipp Mayr, "Mining Scientific Papers for Bibliometrics: A (Very) Brief Survey of Methods and Tools," paper presented at the 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 2015, arXiv: Preprint Archive:1505.01393, https://arxiv.org/abs/1505.01393v1.

[7] Aside from the many STEM field studies that have looked specifically at *academic literature* and/or *journal articles* in a particular discipline such as biomedical science, some of which are cited below, there have also been studies that looked at the literature of humanities disciplines (if not art history), which include: Block and Newman, "What, Where, When, and Sometimes Why"; David Mimno, "Computational Historiography: Data Mining in a Century of Classics Journals," *Journal on Computing and Cultural Heritage* 5, no. 1 (April 2012): 3:1–3:19; Andrew Goldstone and Ted Underwood, "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us," *New Literary History* 45, no. 3 (November 7, 2014): 359–84; Allen Beye Riddell, "How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models," *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, eds. Matt Erlin and Lynne Tatlock (Rochester, NY: Camden House, 2014), 91–114; Benjamin M. Miller, "The Making of Knowledge-Makers in Composition: A Distant Reading of Dissertations" (Graduate Center, CUNY, 2015); Elisabeth Günther and Emese Domahidi, "What Communication Scholars Write About: An Analysis of 80 Years of Research in High-Impact Journals," *International Journal of Communication* 11 (July 26, 2017): 21; Lino Wehrheim, "Economic History Goes Digital: Topic Modeling the Journal of Economic History," *Cliometrica* 13, no. 1 (January 1, 2019): 83–125.

> performance over different disciplines, with those within the humanities usually faring worse than most.[8]

STEM and social science researchers have addressed issues of text mining in journal articles using both journal indexes and PDF articles.[9] In one study of 15 million full-text articles in biomedical science, researchers found that, in addition to utilizing available structured text formats for articles, they *also* needed to address the issue of PDF to text/structured text conversion so that they could cover the vast array of articles included in their study.[10]

In the humanities, text/XML extraction from PDF is essential to forming a workable dataset. There are a number of tools that have been developed specifically to recognize the sections of a PDF journal article and convert it into structured format for text mining. While these tools were, for the most part, designed to parse natural and social science literature, many of them can also be adapted for use within humanities PDFs until more appropriate tools are developed. The techniques that have been created by

[8] Giovanni Colavizza and Matteo Romanello, "Citation Mining of Humanities Journals: The Progress to Date and the Challenges Ahead," *Journal of European Periodical Studies* 4, no. 1 (June 30, 2019): 37.

[9] For example, see: Chiquito J. Crasto et al., "Text Mining Neuroscience Journal Articles to Populate Neuroscience Databases," *Neuroinformatics* 1, no. 3 (September 1, 2003): 215–237; Stephan Dahl, "Current Themes in Social Marketing Research: Text-Mining the Past Five Years," *Social Marketing Quarterly* 16, no. 2 (June 1, 2010): 128–136; Allan Peter Davis et al., "A CTD–Pfizer Collaboration: Manual Curation of 88 000 Scientific Articles Text Mined for Drug–Disease and Drug–Phenotype Interactions," *Database: The Journal of Biological Databases and Curation*, 2013, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3842776; Tai-Quan Peng et al., "Mapping the Landscape of Internet Studies: Text Mining of Social Science Journal Articles 2000–2009," *New Media & Society* 15, no. 5 (August 1, 2013): 644–664; P. K. Jayasekara and Abu K. S., "Text Mining of Highly Cited Publications in Data Mining," in *2018 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS)* (Red Hook, NY: Curran Associates Inc., 2018), 128–130.

[10] Westergaard et al., "Text Mining of 15 Million Full-Text Scientific Articles"; Lindsay McKenzie, "Want to Analyze Millions of Scientific Papers All at Once? Here's the Best Way to Do It," Science AAAS website, July 21, 2017, https://www.sciencemag.org/news/2017/07/want-analyze-millions-scientific-papers-all-once-here-s-best-way-do-it.

researchers include PDFX,[11] GROBID,[12] ParsCit,[13] PDF-extract,[14] LA-PDFText,[15] Biblo (for citations, humanities-focused),[16] and CERMINE.[17] We chose to use CERMINE for the PDF extractions in this study. Our reasons for doing so are detailed in the methodology below.

While the majority of art history journals are not available in formats other than PDF, JSTOR—which contains a large number of humanities publications—has a service that provides "Data for Research" or DfR.[18] So studies that skip the complex step of PDF extraction are possible, given that JSTOR has the articles a researcher would like to study.[19] This is another area that can be developed across journal repositories to aid humanities scholars in understanding the literature in their disciplines.

Once the journal articles are in a format that is ready to process, the researcher must select the methodology for text mining.

[11] Alexandru Constantin, Steve Pettifer, and Andrei Voronkov, "PDFX: Fully-Automated PDF-to-XML Conversion of Scientific Literature," in *Proceedings of the 2013 ACM Symposium on Document Engineering*, DocEng '13 (New York: ACM, 2013), 177–180.

[12] Patrice Lopez, "GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications," in *International Conference on Theory and Practice of Digital Libraries* (Cham: Springer, 2009), 473–474.

[13] Isaac G. Councill, C. Lee Giles, and Min-Yen Kan, "ParsCit: An Open-Source CRF Reference String Parsing Package," in *LREC* 8 (2008): 661–667.

[14] The PDF-extract project has been "retired" and the creators now recommend using CERMINE. See "Pdfextract," CrossRef website, https://www.crossref.org/labs/pdfextract.

[15] Cartic Ramakrishnan et al., "Layout-Aware Text Extraction from Full-Text PDF of Scientific Articles," *Source Code for Biology and Medicine* 7 (May 28, 2012): 7.

[16] Young-Min Kim et al., "Automatic Annotation of Bibliographical References in Digital Humanities Books, Articles and Blogs," in *Proceedings of the 4th ACM Workshop on Online Books, Complementary Social Media and Crowdsourcing* (New York: ACM, 2011), 41–48.

[17] Dominika Tkaczyk et al., "CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature," *International Journal on Document Analysis and Recognition (IJDAR)* 18, no. 4 (December 1, 2015): 317–335.

[18] "JSTOR Data for Research," JSTOR website, https://www.jstor.org/dfr.

[19] See, e.g., Riddell, "How to Read 22,198 Journal Articles," 91–114.

For this study, we chose to use R (including the tm [text mining] package) because it offered flexibility and customizability that other methods did not offer. There are also text mining tools such as Voyant that do not require programming experience and can be operated from the browser window online, but these do not allow for the level of functionality that programmatic means such as R allow.[20] For example, one important step in text mining is cleaning the text of noise—punctuation, numbers, *et cetera*—which is easily and efficiently done in R. These cleaning steps can be customized according to the dataset at hand. It is also possible to analyze and extract particular segments of the data as well as create a wide variety of visualizations as well as export the data in a custom format.

Methodology

To prepare sample sets of articles to test, the PDFs must be downloaded from online scholarly databases. Attaining permission to bulk download PDFs can require lengthy permission negotiations between publishers, university libraries, and researchers. In assembling the corpuses for this study, therefore, 373 articles from the journal *Art History* and 215 from *Art Journal* from the last decade were assembled by manually downloading the articles from the Wiley and Taylor & Francis portals, respectively.[21] This is a relatively small set of data compared to some of the studies that have been done on academic articles previously, but nevertheless large enough to track significant trends within a journal.[22] It should be noted, however, that it would not be ideal to

[20] Voyant Tools website, "Voyant: See Through your Text," https://voyant-tools.org.

[21] While every page/section from these journals can be downloaded as PDFs, the text corpus we produced was limited to the editorial letters and articles in these journals. For *Art Journal*, six types of texts or PDFs were not included: the title page, information on the editorial board, table of contents, funding information, reviews and artist projects. For *Art History*, two types of texts were not included: abstracts/authors' biographies and reviews.

[22] See, e.g., Westergaard et al., "Text Mining of 15 Million Full-Text Scientific Articles"; McKenzie, "Want to Analyze Millions of Scientific Papers All at Once?"

manually download articles if the sample were much larger than this because it would be too time-consuming.

Once the articles have been downloaded, Adobe Acrobat has several export options for PDFs including plain text, rich text, and XML. Using the two corpuses of articles from *Art History* and *Art Journal*, it was clear that these are inadequate methods by which to export text from journal articles. Straight export in Acrobat preserves database headings, journal volume/issue headings, and other information that is repeated on every page of the article, which can skew text mining results. Furthermore, the XML is not exported with reliable individual tags to clearly denote the separate sections of the article. It is for this reason that a number of research groups have developed protocols, such as CERMINE and the others cited above, that are specifically designed to capture the data from scientific journal articles. In particular, these methods are useful for demarcating headings, body text, and bibliographic information, as well as recognizing columns and blocks of text.

Converting PDF to text and XML with CERMINE

In order to convert the sample of PDFs to text and XML, we used an open-source Java protocol called CERMINE (Content ExtRactor and MINEr).[23] Compared to other PDF to XML protocols, CERMINE is able to capture specific features of bibliographic information, including author, year, volume, *et cetera*.[24] Applying CERMINE to the two sample corpuses produced much more reliable and useful text/XML from PDFs than could be achieved by exporting full text from PDFs in Acrobat.[25] Since it is one of the newer protocols in this area, it has addressed some of

[23] Tkaczyk et al., “CERMINE.”

[24] Tkaczyk et al., “CERMINE,” 319.

[25] It should be noted that one study, by the researchers who created CERMINE, found that GROBID outperformed CERMINE “out of the box” for citations. Nevertheless, we found CERMINE more straightforward to use as non-experts in computer science. See Dominika Tkaczyk et al., “Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers,” in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL ’18 (New York: ACM, 2018), 99–108.

the issues and omissions in the design of other similar protocols.[26] However, CERMINE works best for articles that have a reference list at the end of the article rather than just a "Notes" section (either endnotes or footnotes), as is common in art history. When faced with notes that contain more than just citation information, that is, commentary or added information rather than simply bibliographic sources, the XML tags for article title, authors' given names and surname, year, volume, issue, *et cetera*, are sometimes garbled. Nevertheless, CERMINE was able to reliably capture the surnames of authors from the endnotes of 373 articles in *Art History* when they contained references rather than commentary.

Since notes in *Art History* appear at the end of the article, CERMINE was effective at demarcating the bibliography section of the article accurately. *Art Journal*, however, provided a number of issues that had to do with the layout of the journal. In *Art Journal*, notes appear along the left side of the page next to where they are cited. In Figure 1, which was taken as a sample from the summer 2005 issue of *Art Journal*, the notes are highlighted in dark gray for both the first page and an interior page of an article. This format has been used by *Art Journal* since summer 1998, so a corpus of articles from the last 10 years all have this format. Since all the notes are along the margin on the left side of the page, we were able to solve this formatting issue by batch cropping our corpus of 215 documents from *Art Journal* in Acrobat to create a corpus of PDFs that only contain the left margin of the page. This corpus of left margins could then be run through CERMINE to export plain text and XML from the PDFs. Likewise, the reverse operation was executed in which just the notes were cropped off the page to leave the main body text of the article. In cases such as this, where formatting is unusual but consistent throughout, this work-around produces satisfactory results.

If a larger sample of the journal from the 1960s to the present had been used, however, we would have had to contend with a number of formatting changes that would have complicated the operation. Figure 2 shows the format of the journal from autumn 1965, for example, with the notes highlighted in gray for the first page and several interior pages. Here, notes appear under only the

[26] Tkaczyk et al., "CERMINE," 318–319.

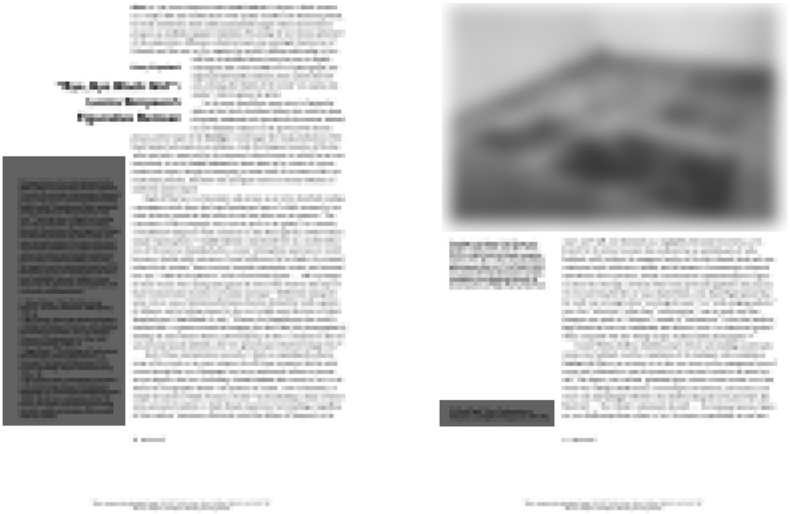


Figure 1: A sample page from the summer 2005 issue of *Art Journal* (footnotes in dark gray). Copyright: Authors. License: CC BY 4.0.

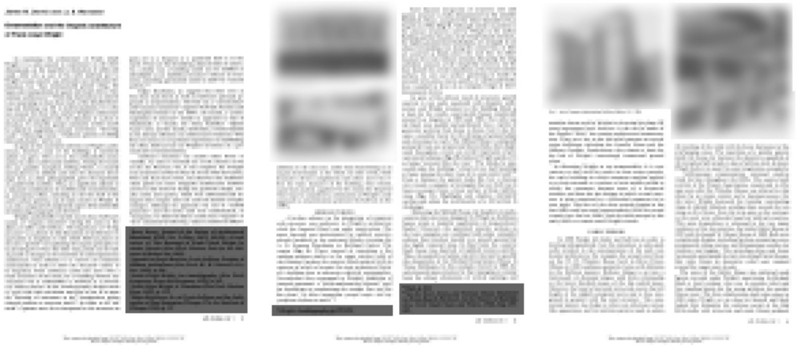


Figure 2: A sample page from the autumn 1965 issue of *Art Journal* (footnotes in dark gray). Copyright: Authors. License: CC BY 4.0.

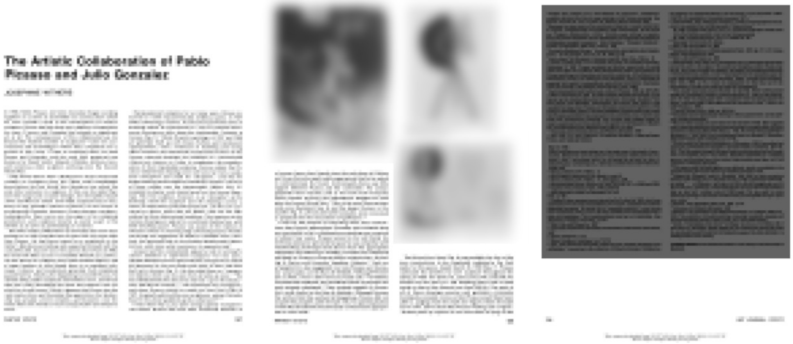
second column alone, only the first column, under both columns, or under neither column. It would be very difficult to reliably crop just the notes from these pages without including any of the body text of the article. Although CERMINE *does* recognize references that are interspersed in the text, it places them in situ in the XML (i.e., among the <p> tags), so references may be garbled in exporting the information rather than cleanly demarcated, as they are when they appear at the end of the article. These differences in the layout mean that, even though the text mining itself can be automated, every volume that goes into the dataset has to be analyzed separately (based on the particularities of its layout) in order to get appropriate and reliable text extractions.

Moving through the years that the journal has been published, one would have had to contend with a number of additional changes in format. An example from winter 1975–76 (Figure 3) shows a two-column format with notes at the end. CERMINE is relatively good at handling columns in articles and has checks in place for determining the layout of the columns in the article.[27] This is useful to note, as *Art Journal* had a three-column format in an example from spring 1985 (Figure 4), and notes were placed at the end of the article. By winter 1995, the journal had returned to a two-column format (Figure 5). In 1998, as noted, the journal transitioned to its current format, which places the references at the left margin of the page. It is, therefore, important to bear in mind that the unique style formats of humanities journals vary not only in comparison to each other but also in comparison to their earlier or later manifestations.

Analyzing the plain text

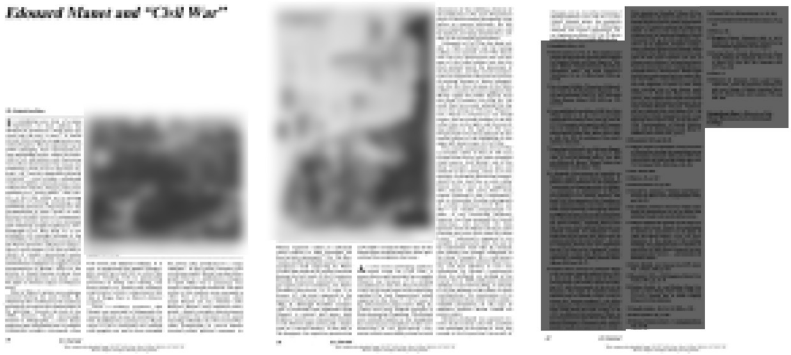
In the given sample of art history journal articles, we used CERMINE to export both plain text and XML. Although the plain text files created by CERMINE exclude many (if not all) of the abovementioned auxiliary PDF headers that were left in the text files exported from Adobe Acrobat, they still contain a lot of text, which creates unwanted noise in the final text mining operations. For example, copyright information for images and figure tags may be undesirable in a corpus collected for text mining since particular image collections and repositories that provide art images for publication will repeatedly occur. In the case of *Art History*, the headings for the publication itself were successfully removed, but, owing to the unusual format of *Art Journal*, the plain text files created by CERMINE for this journal (before page cropping) still contain unwanted headings such as the issue (e.g., “Winter 2011”). Given these limitations, extracting particular

[27] Tkaczyk et al., “CERMINE,” 320–321.



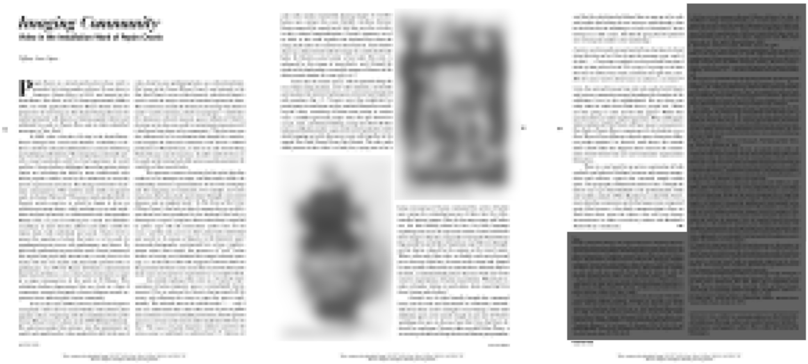
The Artistic Collaboration of Pablo Picasso and Julio Gonzalez

Figure 3: A sample page from the winter 1975–76 issue of *Art Journal* (endnotes in dark gray). Copyright: Authors. License: CC BY 4.0.



Edouard Manet and "Civil War"

Figure 4: A sample page from the spring 1985 issue of *Art Journal* (endnotes in dark gray). Copyright: Authors. License: CC BY 4.0.



Imaging Community

Figure 5: A sample page from the winter 1995 issue of *Art Journal* (endnotes in dark gray). Copyright: Authors. License: CC BY 4.0.

sections from XML files, used alone or in combination with each other, provides more flexibility and clarity in the data.

Extracting XML tags with Python

Certain XML tags in the documents created by CERMINE were of particular interest for text mining the journals in this study. Pulling the <p> (paragraph) tags generates a text document that contains only the body text of a given article minus headings, bibliography, and any other information in the article. This can be used to look at term frequency and create word clouds that capture the main themes of the journal. Additionally, pulling all the text contained in the subchildren of the <back> tag creates a text output of the entire bibliography of the article. This is useful in determining the most frequently cited authors. The <surname> tag was also pulled from the XML in order to compare frequently cited authors to the full bibliography text. While all of the surnames were not accurately captured in the <surname> tag, as mentioned, pulling just this tag creates a very clean list of authors to work with and provides a good basis of comparison to the results of the full bibliography, which need to be cleansed of nonauthor names to a much larger extent before text mining.

In order to pull XML tags, we created two different scripts using Python that could pull the three tags listed above: <p>, <back> (plus children/subchildren), and <surname>. One of the scripts takes the given corpus of journal articles and compiles the selected tag from all those articles into one single plain text file. The other script creates a separate text file with the data from the given tag for each of the journal articles. These were later used to determine overall frequency of terms or authors from a given corpus versus frequency of term or author across the documents. For example, Author X might be a very important reference for three articles out of 10. If these three articles reference her 10 times each or a total of 30 times, she would be placed high on the list of frequently cited scholars in the corpus of 10 documents. If Author Y, however, is mentioned in all 10 of the articles but only one time for each, her frequency of total citations would be lower than Author X (a total of 10). Given this, frequency alone is not the only indicator of influence or popularity. One could argue that Author X is more influential, as she is mentioned in all the articles, even if she is not mentioned *as much*. Therefore, this methodology looks at *both* total frequency and frequency across the documents. So, Author X would be counted for three documents and Author Y for 10, putting Author Y higher on the list of most-cited authors across all the articles.

Text mining with R

Once the desired data from particular XML tags have been pulled from the files, preparations for text mining can begin. For this purpose, we used the tm (text mining) package in R. The first step after loading the given corpus of text files was to clean the data. Some of the text exported from the PDF did not come out as clean as we would have liked. For example, some words contained spaces between vowels, such as the word “figure,” which sometimes appeared in the XML as “fi gure.” In most cases, this kind of export error was corrected *ad hoc* with find/replace within the text. Some words were also mashed together without spaces in the export. These would later be combined with instances of the words that came out cleanly in the frequency lists when they were found in the resultant CSV file.

In order to clean the text, we used several different cleaning protocols in R/tm that are designed to remove unwanted characters and words. Often, within a given corpus, there are combinations of words that have to be excluded from the bibliography, as they could skew the results. For example, “Journal of the Warburg and Courtauld Institutes” and “Warburg Institute” were removed from bibliographic references in this study, as they would have distorted the frequency for citations of the art historian Aby Warburg, after whom the institute takes its name. Another facet of the language found in art history journals is that there are plenty of non-ASCII characters from non-English languages, such as accented characters. In order to simplify the export of these characters in the given data, we converted accented characters to ASCII equivalents. So, for example, ö became o. After attending to these special cases, which depend on the contents of the corpus at hand, we followed a number of common steps for text cleaning. These included: removing punctuation and numbers, making text lowercase, removing stopwords, removing whitespace, and stemming the text in order to combine words with similar roots such as "paintings" and "painting."

In the case of looking at bibliography text, choice of stopwords became an issue. Since we were only interested primarily in the names found in the references, the standard English stopword list was not comprehensive enough. Stopwords are the words that are common to a given language but not interesting or important for text mining, for example "the," "and," or "for." R has its own stopwords function and lists where the language (in this case, English) can be specified to remove these commonly used but unimportant words from the frequency output. To analyze the bibliography, however, we created our own custom stopwords lists of the top 20,000 most frequently used words in the English language.[28] Applying these lists to the text gave a much cleaner picture of the authors cited. After cleaning the text, the next step was to create a document term matrix (dtm) from the corpus and export a CSV spreadsheet of the terms in decreasing order of frequency. This produces the total frequency for the single combined text file of the given corpus of documents.

In order to find the spread of a given term or author across the documents, however, a slightly different method was used. For this, another corpus was created from a directory containing all the individual text files of each exported tag and then cleaned using the same steps as above. Once the text was cleaned, it was exported as a term document matrix (tdm) and written to a CSV file. Each column was then counted if there was a value > 0 (indicating the author or term had been cited at least once) and those totals were used to create a list of the top-cited authors or terms by number of documents in which they appear.

Findings and Discussion

This study found that CERMINE can be used on art history journal articles to export body text and bibliographic information in a parseable XML format, which can then be used to perform text

[28] This list was compiled from preexisting word lists that are readily available online. The inbuilt stopwords exclusion function in R, however, has a much smaller list of words such as those listed above.

mining in R (tm package) to determine the frequency of citations and terms for a given set of journal articles. While the <surname> tags exported in CERMINE were less reliable than extraction of either the full bibliography or the <p> tags, it still provided a useful, clean set of data to work with in combination with the other extractions. The text mining protocol in R is well established and can be used as a flexible means for a variety of text mining functions beyond simple frequency of terms. Before data can be put into this tried and tested text mining method, the input data needs to be relatively clean. This study shows that there is a need for a protocol like CERMINE that is specifically designed for footnotes and endnotes that include both bibliographic information and commentary for the study of art history and other humanities journal articles. The highly varied formats of such journals, however, make the task very difficult. Text mining of frequently cited authors and terms in journal articles can then be used to provide an overall picture of the themes and biases of a particular journal over a given time period.

The overall goal of dealing with a large corpus of text is to automate as many steps as possible, so the steps of this methodology that had to be conducted manually point toward areas for further development when it comes to text mining humanities PDFs.[29] The first such area is in downloading the original corpus of PDFs. This is easily solved for a larger corpus, as some journal databases are now willing to allow researchers access to articles for text mining studies.[30] Given some of the inadequacies in the quality of text exported and the manual text cleaning that was necessary to correct errant spaces and words that were mashed together in the converted text produced by CERMINE, this study would have certainly benefited from delivery of text in plain text format. When PDF text—even OCR or born-digital articles—are captured from the document, the quirks of spacing do not always translate, as anyone who has ever tried to copy/paste text from a

[29] There are existing out-of-the-box software packages to do text analysis like NVivo, but acquiring the actual *text* of journal articles in a usable and machine-readable form remains the main barrier here.

[30] This varies by database and country where the researcher is based and is by no means a resolved issue for research such as this.

PDF will have learned. If journal articles were available as structured text or even just plain text, these issues would be irrelevant.

There were also issues that arose within the final frequency lists of terms and authors. One of the major issues encountered was that, in focusing on surnames, the output of common surnames such as "Smith" and "Jones" were conflated with one another. In further studies, a means by which specific Smiths are associated with their given name or specific references would help to solve this issue. In the literature, an author named Smith might be referred to by surname several times, with their given name only mentioned in the first citation. This makes connecting all the *same* Smiths more difficult. Another major issue in the text cleaning process was that stemming was not always very successful. The particular jargon of art history, that is, the nominalization and adjectivization of certain words and authors, is not accounted for in standard stemming protocols. For example, the name of the art historian Aby Warburg is sometimes adjectivized to the descriptor "Warburgian" or the verb "perform" may be nominalized as "performativity," following scholar Judith Butler.[31] In other words, one of complications in dealing with text mining of art history scholarship is the complexity of the language used and its deviation from standard English. The nominalized or adjectival forms of certain terms are, however, rather easy to locate through searching for the root in the CSV output files.

Despite these issues, the methodology described in this study produced results that were rich with insight and often defied expectations or preconceptions we had about particular journals or the concerns of the field in general. The frequency lists of terms and authors can be used in the future to compare and analyze differences and similarities across sample sets.[32] Possible avenues of further research using this methodology include comparisons of the content in a selection of different journals or analysis of the content in one particular journal over time.

[31] Judith Butler, *Bodies That Matter: On the Discursive Limits of "Sex"* (New York: Routledge, 1993); Judith Butler, *Gender Trouble: Feminism and the Subversion of Identity* (New York: Routledge, 1999).

[32] For this aim we will make these datasets open access on our research project's website metadataculture.se in the future.

Acknowledgments

The research for this chapter was conducted within the project *Sharing the Visual Heritage* (metadataculture.se) at Stockholm University, funded by the Swedish Research Council.

References

Atanassova, Iana, Marc Bertin, and Philipp Mayr. "Mining Scientific Papers for Bibliometrics: A (Very) Brief Survey of Methods and Tools." Paper presented at the 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 2015. arXiv: Preprint Archive:1505.01393. https://arxiv.org/abs/1505.01393v1.

Block, Sharon, and David Newman. "What, Where, When, and Sometimes Why: Data Mining Two Decades of Women's History Abstracts." *Journal of Women's History* 23, no. 1 (March 10, 2011): 81–109.

Butler, Judith. *Bodies That Matter: On the Discursive Limits of "Sex."* New York: Routledge, 1993.

Butler, Judith. *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge, 1999.

Colavizza, Giovanni, and Matteo Romanello. "Citation Mining of Humanities Journals: The Progress to Date and the Challenges Ahead." *Journal of European Periodical Studies* 4, no. 1 (June 30, 2019): 36–53.

Constantin, Alexandru, Steve Pettifer, and Andrei Voronkov. "PDFX: Fully-Automated PDF-to-XML Conversion of Scientific Literature." In *Proceedings of the 2013 ACM Symposium on Document Engineering*, 177–180. DocEng '13. New York: ACM, 2013.

Councill, Isaac G., C. Lee Giles, and Min-Yen Kan. "ParsCit: An Open-Source CRF Reference String Parsing Package." *LREC* 8 (2008): 661–667.

Crasto, Chiquito J., Luis N. Marenco, Michele Migliore, Buqing Mao, Prakash M. Nadkarni, Perry Miller, and Gordon M. Shepherd. "Text Mining Neuroscience Journal Articles to Populate Neuroscience Databases." *Neuroinformatics* 1, no. 3 (September 1, 2003): 215–237.

Dahl, Stephan. "Current Themes in Social Marketing Research: Text-Mining the Past Five Years." *Social Marketing Quarterly* 16, no. 2 (June 1, 2010): 128–136.

Davis, Allan Peter, Thomas C. Wiegers, Phoebe M. Roberts, Benjamin L. King, Jean M. Lay, Kelley Lennon-Hopkins, Daniela Sciaky, et al. "A CTD–Pfizer Collaboration: Manual Curation of 88 000 Scientific Articles Text Mined for Drug–Disease and Drug–Phenotype Interactions." *Database: The Journal of Biological Databases and Curation* (2013). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3842776.

Goldstone, Andrew, and Ted Underwood. "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." *New Literary History* 45, no. 3 (November 7, 2014): 359–384.

Günther, Elisabeth, and Emese Domahidi. "What Communication Scholars Write About: An Analysis of 80 Years of Research in High-Impact Journals." *International Journal of Communication* 11 (July 26, 2017): 21.

International Journal for Digital Art History website. https://dahj.org.

Jaskot, Paul B. "Digital Art History as the Social History of Art: Towards the Disciplinary Relevance of Digital Methods." *Visual Resources* 35, no. 1–2 (April 3, 2019): 21–33.

Jayasekara, P. K., and Abu K. S. "Text Mining of Highly Cited Publications in Data Mining." In *2018 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS)*, 128–130. Red Hook, NY: Curran Associates inc., 2018.

JSTOR website. "JSTOR Data for Research." https://www.jstor.org/dfr.

Kim, Young-Min, Patrice Bellot, Elodie Faath, and Marin Dacos. "Automatic Annotation of Bibliographical References in Digital Humanities Books, Articles and Blogs." In *Proceedings of the 4th ACM Workshop on Online Books, Complementary Social Media and Crowdsourcing*, 41–48. New York: ACM, 2011.

Long, Matthew P., and Roger C. Schonfeld. "Preparing for the Future of Research Services for Art History: Recommendations from the Ithaka S+R Report." *Art Documentation: Journal of the Art Libraries Society of North America* 33, no. 2 (2014): 192–205.

Lopez, Patrice. "GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications." In *International Conference on Theory and Practice of Digital Libraries*, 473–474. Cham: Springer, 2009.

McKenzie, Lindsay. "Want to Analyze Millions of Scientific Papers All at Once? Here's the Best Way to Do It." Science AAAS website, July 21, 2017. https://www.sciencemag.org/news/2017/07/want-analyze-millions-scientific-papers-all-once-here-s-best-way-do-it.

Miller, Benjamin M. "The Making of Knowledge-Makers in Composition: A Distant Reading of Dissertations." Graduate Center, CUNY, 2015.

Mimno, David. "Computational Historiography: Data Mining in a Century of Classics Journals." *Journal on Computing and Culteral Heritage* 5, no. 1 (April 2012): 3:1–3:19.

Moxey, Keith. "Visual Studies and the Iconic Turn." *Journal of Visual Culture* 7, no. 2 (2008): 131–146.

CrossRef website. "Pdfextract." https://www.crossref.org/labs/pdfextract.

Peng, Tai-Quan, Lun Zhang, Zhi-Jin Zhong, and Jonathan J. H. Zhu. "Mapping the Landscape of Internet Studies: Text Mining of Social Science Journal Articles 2000–2009." *New Media & Society* 15, no. 5 (August 1, 2013): 644–664.

Ramakrishnan, Cartic, Abhishek Patnia, Eduard Hovy, and Gully Burns. "Layout-Aware Text Extraction from Full-Text PDF of Scientific Articles." *Source Code for Biology and Medicine* 7 (May 28, 2012): 7.

Riddell, Allen Beye. "How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models." In *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, edited by Matt Erlin and Lynne Tatlock, 91–114. Rochester, NY: Camden House, 2014.

Tkaczyk, Dominika, Andrew Collins, Paraic Sheridan, and Joeran Beel. "Machine Learning vs. Rules and Out-of-the-Box vs.

Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers." In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, 99–108. JCDL '18. New York: ACM, 2018.

Tkaczyk, Dominika, Paweł Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Łukasz Bolikowski. "CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature." *International Journal on Document Analysis and Recognition (IJDAR)* 18, no. 4 (December 1, 2015): 317–335.

Voyant Tools website. "Voyant: See Through your Text." https://voyant-tools.org.

Wehrheim, Lino. "Economic History Goes Digital: Topic Modeling the Journal of Economic History." *Cliometrica* 13, no. 1 (January 1, 2019): 83–125.

Westergaard, David, Hans-Henrik Stærfeldt, Christian Tønsberg, Lars Juhl Jensen, and Søren Brunak. "Text Mining of 15 Million Full-Text Scientific Articles." BioRxiv: The Preprint Server for Biology, July 11, 2017. https://doi.org/10.1101/162099.

Whalen, Maureen. "What's Wrong with This Picture? An Examination of Art Historians' Attitudes About Electronic Publishing Opportunities and the Consequences of Their Continuing Love Affair with Print." *Art Documentation: Journal of the Art Libraries Society of North America* 28, no. 2 (Fall 2009): 13–22.

Yang, Yang, Ching Man Au Yeung, Mark J. Weal, and Hugh Davis. "The Researcher Social Network: A Social Network Based on Metadata of Scientific Publications." *Proceedings of WebSci '09: Society On-Line*, Athens, Greece. Mars 18–20, 2009. https://eprints.soton.ac.uk/267156.