Orchestrating Cultural Heritage

Picca, Davide; Schnyder, Antonin; Eri, Kostina; Adamou, Alessandro; Rodighiero, Dario; Schnapp, Jeffrey T.

Published in:
Proceedings of the 34th ACM Conference on Hypertext and Social Media

DOI:
10.1145/3603163.3609066

Orchestrating Cultural Heritage: Exploring the Automated Analysis and Organization of Charles S. Peirce's PAP Manuscript

Davide Picca
davide.picca@unil.ch
Université de Lausanne
Lausanne, Vaud, Switzerland

Antonin Schnyder
antonin.schnyder@unil.ch
Université de Lausanne
Lausanne, Vaud, Switzerland

Eri Kostina
ekostina@bu.edu
Boston University
Boston, Massachusetts, United States

Alessandro Adamou
alessandro.adamou@biblhertz.it
Bibliotheca Hertziana - Max Planck
Institute for Art History
Rome, Italy

Dario Rodighiero
d.rodighiero@rug.nl
University of Groningen
Groningen, the Netherlands

Jeffrey Schnapp
jeffrey@metalab.harvard.edu
Harvard University
Cambridge, Massachusetts, United
States

ABSTRACT

This preliminary study introduces an innovative approach to the analysis and organization of cultural heritage materials, focusing on the archive of Charles S. Peirce. Given the diverse range of artifacts, objects, and documents comprising cultural heritage, it is essential to efficiently organize and provide access to these materials for the wider public. However, Peirce's manuscripts pose a particular challenge due to their extensive quantity, which makes comprehensive organization through manual classification practically impossible. In response to this challenge, our paper proposes a methodology for the automated analysis and organization of Peirce's manuscripts. We have specifically tested this approach on the renowned 115-page manuscript known as PAP. This study represents a significant step forward in establishing a research direction for the development of a larger project. By incorporating novel computational methods, this larger project has the potential to greatly enhance the field of cultural heritage organization.

CCS CONCEPTS

• **Applied computing** → **Online handwriting recognition**; **Document analysis**; *Hypertext / hypermedia creation*; *Annotation*; *Media arts*; Publishing; *Digital libraries and archives*; • **Human-centered computing** → Visual analytics; Visualization theory, concepts and paradigms; • **Information systems** → *Ontologies*; • **Computing methodologies** → *Cognitive science*; • **Social and professional topics** → **Historical people**.

KEYWORDS

Charles S. Peirce, data mining, document classification, data visualization, OCR, manuscripts, semantic analysis

ACM Reference Format:
Davide Picca, Antonin Schnyder, Eri Kostina, Alessandro Adamou, Dario Rodighiero, and Jeffrey Schnapp. 2023. Orchestrating Cultural Heritage: Exploring the Automated Analysis and Organization of Charles S. Peirce's PAP Manuscript . In *34th ACM Conference on Hypertext and Social Media (HT '23), September 4–8, 2023, Rome, Italy*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3603163.3609066

1 INTRODUCTION

Charles S. Peirce, a seminal figure in mathematics, logic, and philosophy, has profoundly impacted the intellectual evolution of humanity, particularly in the domain of scientific methodology [1]. However, the extensive intellectual heritage he left behind presents formidable challenges for organizing and categorizing it, due to constraints in accessibility and the task of organizing undated and revised texts.

Following Peirce's passing, Harvard University assumed stewardship of roughly 100,000 manuscript pages inherited from his widow [2]. Efforts to restructure and digitize Peirce's archive have proven challenging due to the sheer volume of his work and the ambiguity surrounding its chronological order.

This presentation introduces preliminary findings from an upcoming government-funded initiative, led in conjunction by the University of Lausanne, Harvard Houghton Library, Groningen University, and Bibliotheca Hertziana, to digitize and computationally reorganize Peirce's archive [6]. The study outlined here proposes a method for examining and classifying Peirce's manuscripts utilizing an automated system for more effective corpus organization.

The presentation is organized as follows: Section 2 elaborates on the complexities of the proposed classification, acknowledging Richard Robin's catalog [11] and the Peirce Edition Project [9]. Section 3 details the dataset and pre-processing measures, while Section 4 discusses the methodology for information extraction and initial findings, before concluding with an outlook on the next steps.

2 PRESERVING CULTURAL HERITAGE

Cultural heritage encompasses a broad range of artifacts, objects, and documents that provide insights into the history, values, and cultural practices of society. Preserving cultural heritage is crucial for maintaining the memory of humanity's collective past and understanding its evolution. In recent years, digitizing cultural artifacts has become a popular approach to making them more

HT '23, September 4–8, 2023, Rome, Italy
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0232-7/23/09.
https://doi.org/10.1145/3603163.3609066

accessible to the broader public. However, organizing and categorizing digital collections of cultural heritage materials remains challenging.

Digitizing cultural heritage materials has several advantages, including the ability to make them more easily accessible to the public, to preserve fragile or deteriorating materials, and to facilitate research and study of the materials by scholars and researchers [10]. However, simply digitizing materials is not enough - the materials must also be organized to make them searchable and navigable.

One example of the challenges faced when organizing cultural heritage materials can be seen in Charles S. Peirce's manuscripts at Harvard Library. Peirce's manuscripts were left in disarray at the library, and scholars have attempted to classify the documents in various ways [2]. However, the quantity and nature of the manuscripts make it a daunting task. Houser and Kloesel [3] note that Peirce's papers were eventually classified by scientific discipline, but the organization did not consider the development of Peirce's thought. While the Hartshorne-Weiss volumes contributed significantly to categorizing Peirce's papers, they were semi-problematic. The Peirce Edition Project aimed to collate Peirce's writings chronologically, but a shortage of scholars and funding now threatens the project's survival.

Richard Shale Robin's catalog of Peirce's manuscripts remains a vital resource for researchers, although Robin acknowledged its imperfections and suggested automating Peirce's manuscript processing. Despite these challenges, preserving cultural heritage cannot be understated. Ongoing efforts to organize and digitize cultural heritage materials will help ensure these invaluable artifacts are available to future generations.

3 EMPIRICAL ANALYSIS

This presentation introduces the preliminary findings of an upcoming government-funded project, which aims to automate the organization of Peirce's corpus, ensuring consistency and accuracy in content categorization. This analysis aims to showcase an automated methodology that can compare parts of one or multiple documents, thereby creating a scalable model that can be applied to extensive collections, including the unpublished manuscripts of Peirce himself.

The case study centralizes on Peirce's renowned manuscript, PAP (Prolegomena for an Apology to Pragmatism), publicly available online through Houghton Library [2]. Identified as "Ma 293" and classified under the pragmatism category in Robin's catalog [11], PAP represents one of Peirce's most significant writings [16]. The current analysis examines all 115 manuscript pages of PAP, which include 56 pages of writing, 48 variant pages, and one blank page. These pages predominantly consist of textual content, with some inclusion of equations, graphs, and annotations. These statistics are summarized in Table 1.

The data obtained from this analysis was utilized for various purposes, as will be outlined in Section 4. High-level information about the entire text was acquired by merging all the pages into a single document. Depending on requirements, individual pages were examined separately or aggregated for comparison. For instance, the network illustrated in Section 4.2 was constructed from a knowledge graph extracted from all manuscript pages referred to as PAP, combined into a single file. Non-textual information was excluded from this analysis to focus entirely on text, as Peirce's diagrams warrant a project in their own right. All the subsequent initial studies were conducted without considering potential contributions from these graphical elements.

Table 1: Comparison of word count between sections

Section	Page Number	Word Count	Word Percent
First Draft	56	9,250	54.8%
Later Variants	48	7,627	45.2%
Blank Pages	1	0	0.00%

4 METHODOLOGY

The automated process discussed in this section follows a three-step procedure:

(1) Initial processing of manuscript pages, including layout recognition and text transcription.
(2) Semantic evaluation of pages and extraction of pertinent terms and connections.
(3) Grouping of pages based on similar lemma use and presenting the groups through a cartographic visualization.

4.1 Initial Processing

The first phase of our approach involves transcribing manuscript pages fetched from HOLLIS, the repository of Harvard University that holds Peirce's archive [17]. TIFF images of the PAP document were locally downloaded using the International Image Interoperability Framework (IIIF), a protocol that delivers high-quality images and their corresponding metadata online for cultural heritage institutions [15]. The 115-page images were then uploaded to the Transkribus platform [5] for layout detection, Optical Character Recognition (OCR), and Handwritten Text Recognition (HTR). To minimize decoding errors, a manual revision followed this automated process. Given the focus on the handwritten text, figures, and graphs were replaced with a placeholder "Fig. x", and unrecognized words were noted with "`||||unidentified-word||||`". No similarities were detected among these unidentified words. The transcription was then exported to XML format for text analysis using the spaCy [18] and REBEL [4] libraries.

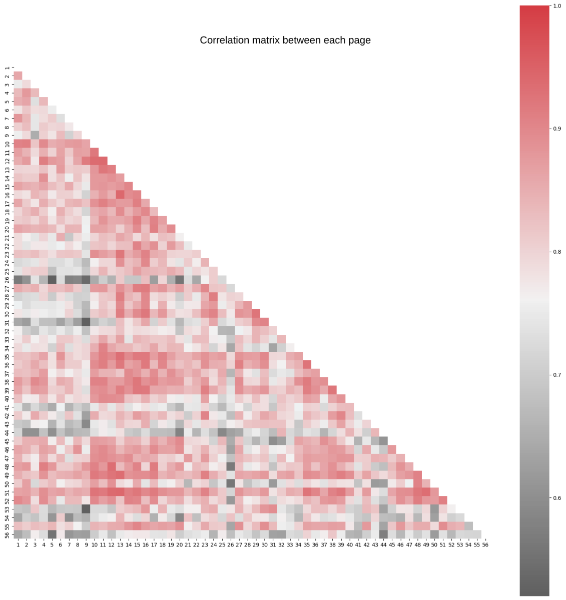

Figure 1: Correlation matrix showing the semantic proximity of pages to each other. Variants have been aggregated.

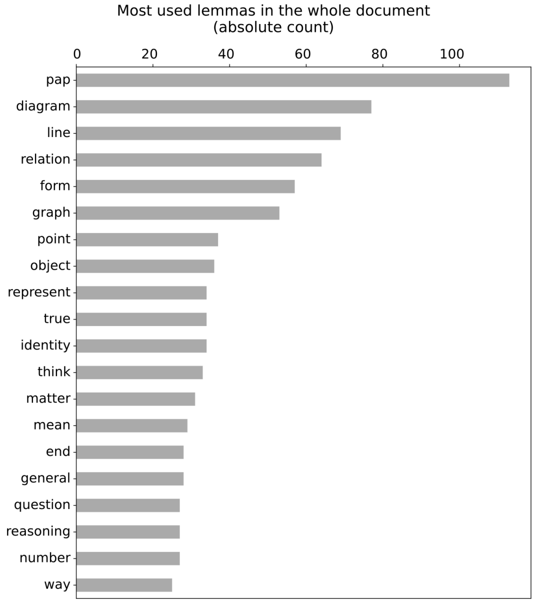

Figure 2: Bar chart showing the most frequent lemmas in the document.

4.2 Text Analysis

Several computational techniques were employed to analyze the manuscript, taking into account Peirce's unique writing style. The first technique involved identifying capitalized words, which Peirce used for emphasis, such as "Diagram", "Induction", or "Line of Identity". The average ratio of capitalized words per page was 9.3, with a standard deviation of 8.7. Semantic proximity between manuscript pages was estimated using vector representations of words computed by spaCy. This approach created a square correlation matrix that displayed values between -1 and +1, indicating the proximity between each pair of documents (See Figure 1). The latter variant pages were grouped and linked with the corresponding first draft pages, leading to 56 documents.

The most frequent lemmas were also identified to better understand the manuscript's meaning. After removing less significant words, lemmas were extracted and sorted in decreasing order of occurrence. This information was presented as a bar chart (Figure 2). Finally, a network analysis was performed using REBEL, a method presented by Cabot and Navigli [4]. A total of 950 triples were extracted, and 43 relations were identified. These include relations that carry a methodological or scholarly significance (e.g., "discoverer or inventor", "practiced by", "studied by" or "used by") as well as those with causal (e.g., "has cause", "has effect") or topological meanings (e.g., "part of", "location" and "country"): aligning them with standard ontologies is part of future work. These relations were presented separately in interactive graphs using Streamlit[1].

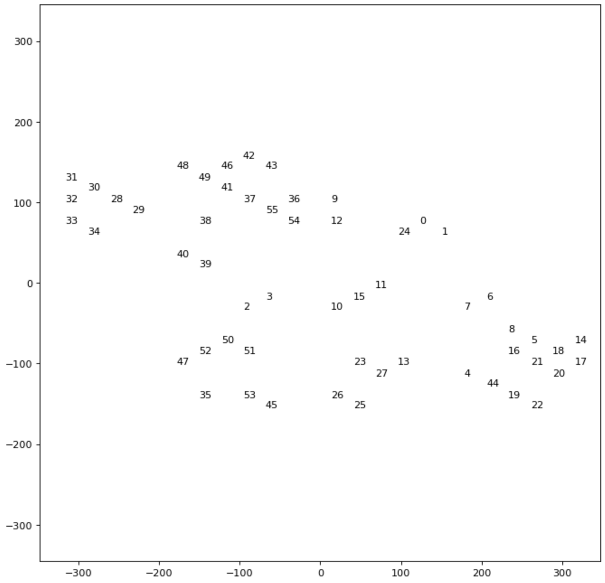

Figure 3: The UMAP algorithm maps PAP manuscripts according to similarity.

4.3 Clustering

The final stage of the text analysis employs lemmas to chart manuscript pages onto a two-dimensional plane through cartographic data visualization. First, the contents of all the pages were reduced to a frequency table of lemmas, which was subsequently modified utilizing the TF-IDF metric to highlight the similarity between

[1] The application is freely available at https://pap-viz.streamlit.app/

pages [14]. Next, the multi-dimensional frequency of lemmas that describe each page is reduced to two dimensions on the Cartesian plane through UMAP dimensionality reduction [7]. The UMAP algorithm provides a spatial distribution of manuscript pages based on similarity, creating page clusters with similar usage of lemmas as illustrated in Figure 3.

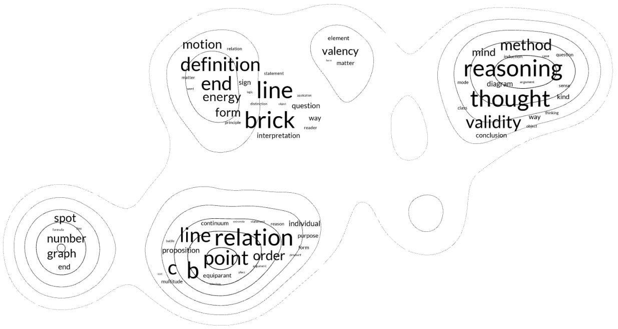

Figure 4: This web-based application allows the user to see the UMAP embedding with new graphical elements that resemble a visual classification of the PAP manuscript pages.

Advanced data visualization techniques can be employed to customize this embedding [8, 12, 13]. Furthermore, word clouds display the most frequently used words in each cluster. This potential visual classification provides an alternative interpretation of the manuscript pages, in addition to close reading and typical text-analysis visuals. This two-dimensional classification of PAP writings can be extended to more extensive manuscript collections like the Peirce archive.

5 CONCLUSION

The preliminary findings from this ongoing project have demonstrated the potential of an automated approach to categorize and analyze Peirce's unpublished manuscripts. By implementing computational text transcription and analysis techniques, we have successfully identified semantic connections between different sections of the PAP document. The resulting cartographic visualization provides an intuitive representation of the manuscript's structure and the relationships between its various parts.

The following steps of this project will involve refining the methodology and extending the analysis to other manuscripts in Peirce's corpus. In addition, future research could incorporate the examination of graphical content, such as diagrams and equations, to provide a more comprehensive understanding of Peirce's work.

REFERENCES

[1] Robert Burch. 2001. Charles Sanders Peirce. In *The Stanford Encyclopedia of Philosophy* (summer 2022 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2022/entries/peirce/

[2] Harvard Library. 2022. Charles S. Peirce Papers | Harvard Library. https://library.harvard.edu/collections/charles-s-peirce-papers

[3] Nathan Houser. 1989. Nathan Houser, "The Fortunes and Misfortunes of the Peirce Papers" at Arisbe: The Peirce Gateway. https://arisbe.sitehost.iu.edu/menu/library/aboutcsp/houser/fortunes.htm

[4] Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation Extraction By End-to-end Language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 2370–2381. https://aclanthology.org/2021.findings-emnlp.204

[5] Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 04. IEEE, Kyoto, Japan, 19–24. https://doi.org/10.1109/ICDAR.2017.307

[6] Mary Keeler. 2020. Pragmatically Improving Access to Peirce's Archive. *Chinese Semiotic Studies* 16, 1 (Feb. 2020), 167–187. https://doi.org/10.1515/css-2020-0009

[7] Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *arXiv.org* stat.ML (Feb. 2018), 63 pages. https://arxiv.org/pdf/1802.03426.pdf

[8] Chloe Ye Eun Moon and Dario Rodighiero. 2020. Mapping as a contemporary instrument for orientation in conferences. In *Proceedings of the IX Annual Conference of the Association for Humanities and Digital Culture (AIUCD). The inevitable turning point: challenges and perspectives for Humanistic Informatics*, Carlo Passarotti, Eleonora Maria Gabriella Litta Modignani Picozzi, Greta Franzini, and Cristina Marras (Eds.). Associazione per l'Informatica Umanistica e la Cultura Digitale, Milan, Italy, 156–162. https://doi.org/10.5281/zenodo.3611340

[9] Peirce Edition Project. 2021. Peirce Edition Project. https://peirce.iupui.edu/

[10] Davide Picca, Alessandro Adamou, Yumeng Hou, Mattia Egloff, and Sarah Kenderdine. 2022. Knowledge organization of the Hong Kong Martial Arts Living Archive to capture and preserve intangible cultural heritage. In *Digital Humanities 2022 - Conference Abstracts, 25-29 July 2022, Tokyo, Japan*. University of Tokyo, online, 329–332. https://dh2022.dhii.asia/abstracts/files/PICCA_Davide_Knowledge_organization_of_the_Hong_Kong_Martial.html

[11] Richard S. Robin. 1967. Robin Catalog. https://peirce.sitehost.iu.edu/robin/rcatalog.htm

[12] Dario Rodighiero. 2021. *Mapping affinities: democratizing data visualization* (open-access English ed.). Métis Presses, Geneva. https://doi.org/10.37866/0563-99-9

[13] Dario Rodighiero and Alberto Romele. 2022. Reading network diagrams by using contour lines and word clouds. In *Proceeding of Graphs and Networks in the Humanities*. Royal Netherlands Academy of Arts and Sciences, Amsterdam, 7 pages. https://doi.org/10.31235/osf.io/3g7bt

[14] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24, 5 (Jan. 1988), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

[15] Stuart Snydman, Robert Sanderson, and Tom Cramer. 2015. The International Image Interoperability Framework (IIIF): A. In *Proceedings of IS&T Archiving Conference*. Los Angeles, CA, 16–21.

[16] Frederik Stjernfelt. 2007. *Diagrammatology: an investigation on the borderlines of phenomenology, ontology, and semiotics*. Number 336 in Synthese library. Springer, Dordrecht. OCLC: 255632682.

[17] Harvard University. 2023. Charles S. Peirce papers. https://hollisarchives.lib.harvard.edu/repositories/24/resources/6437 Accessed: 2023-03-08.

[18] Yuli Vasiliev. 2020. *Natural language processing with Python and spaCy: a practical introduction*. No Starch Press, San Francisco.