

Archival Discovery Through Network Graph Morphisms

Jeremiah Colonna-Romano; The University of Alabama Libraries, Hoole Special Collections Library; Tuscaloosa, Alabama, USA

Abstract

Digitized special collections archives are discoverable largely through full-text search via indexing [1] and categorical methods of librarianship expressed in finding aid arrangement, which privileges established terms and inherited description [2,3]. By contrast, this work is a computational arrangement method that compares clustered sentence embedding structures [4] to promote discovery when lexicon and semantic terms differ. We explore modeling texts as a graph of directed semantic displacements in a vector embedding space and conceptualize sections of these graphs as an extrinsically comparable morphism. We evaluate these graph structures as a semiotic medium by comparing alignments across graphs from various types of OCR texts, and language-like noise datasets. Findings suggest potential interoperative value in morphism aligned text. Work sees potential applications for interdisciplinary discovery and unsupervised archival arrangement.

Research Questions

1. When lexicon and semantic terms are not stable across materials text, how can a scholar discover and recognize content in archival collections without relying on the archive's expression of authoritative description?
2. If a document's network graph in vector embedding space encodes semantic features, can aligned sections between these graphs encode a comparable semantic morphism useful for discovery?
3. Given alignment feature comparisons of our semantic morphism, what thresholds produce matches with enough ambiguity to surface unintuitive adjacency? Does this threshold exceed feature alignment found when compared to tiered types of random text?

Motivation

When I am producing or supervising the production of archival metadata, I find myself regularly faced with the acceptably variable and structurally bound nature of the authority language we assign to archival documents we intend to describe in perpetuity. While subject and indexing system technology works easily with our interpretive assessment terms, the professionally defensible but arbitrary nature of these assessments cannot, due to their summary nature, hold the fullness of the entire language present in the archival document [1,5]. This work comes out of a desire to see if the complete text of a primary research material can be used to form a comparable computational identity that exposes more of the fullness of the materials text to discovery.

Rather than treating discovery methods derived from archival description as an authoritative statement on intent that locates a document's utility, this project treats discovery and arrangement as practices that produce meaning through tunable linguistic constraints.

Problem

Discovery of digitized library and archival collections is often done through full-text lexical index search, yet search privileges the terms, names, subjects, and phrases that reflect prior archival

description practices and can reinforce historical and systemic bias [2]. Researchers face a siloing problem: how can we recognize pertinent material when it is described in out-of-lexicon terms, and how can we search in other domains when we do not know the terminology?

Best practices for archival description rely on authoritative use of lexical and semantic terms [3]; in practice, discoverability is arbitrated through institutional policy priorities, asset management system technical restrictions, and individual expressions of archival best practices [5].

The technical problem addressed here is how to build a discovery system that does not require prior knowledge of authorized terms, accommodates the fuzzy expression of best practices, and still provides inspectable evidence for why items are suggested as related.

Approach

Bodies of archival documents text are converted into graphs of directed semantic displacements. Documents are segmented at the sentence level and embedded with an SBERT model; these sentence embeddings are then clustered by an agglomerative function [4,6]. Cluster centroids form the nodes of the document graph. The system computes a tensor of displacement vectors between nodes in a graph that captures intra-document semantic transitions.

We then estimate a principal component direction for each cluster [7], such that a generic comparable morphism is constructed by the location and connection of two clusters, their PC1 directions and the direction of the pairwise clusters displacement vector.

Semantic quality assessments and lexical composition metrics are generated from compared cluster text to describe observed values at different profiles. Aspects of archival item discovery through this approach are measured against a baseline of document embedding cosine similarity.

Fig. 1D and Fig. 1E show strong inter-document feature similarity between morphisms found on two different documents' network graph hulls. We measure these attributes to compare directional alignment of intra-document morphisms via cosine similarity and concordance of source and destination PC1 directions for potential as a surrogate for categorical language built directly from a document's text.

For a specific example of what these candidates for sublativ discourse look like in the archive, during development we found aligned morphisms that encoded potentially valuable unintuitive relatedness for research scholars in an issue of a 1960's student humor magazine and an issue of a 1990's LGBTQ newspaper. Both exhibited clustered language from first-person humorous anecdotal stories and clustered language expressing trepidation about U.S. relations with the U.S.S.R.

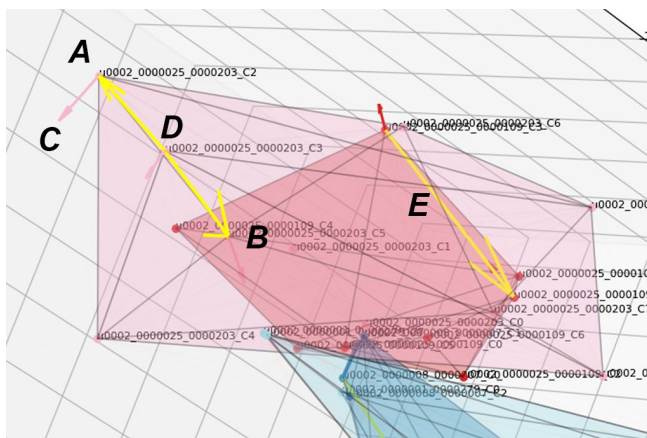


Figure 1. Projection of two different document graphs showing yellow aligned inter-document morphisms found on hulls shown in shades of pink. (A, B) Embedding cluster centroid nodes located at each end of the morphism. (C) Pink arrow indicates principal component direction of the sentence embedding cluster located at the end of the morphism. (D) Yellow arrow denotes displacement vector mapped between nodes. (E) Inter-document morphism located on dark pink hull of second document exhibits alignment between documents.

We anticipate interoperative domain value to providing researchers with discovery methods capable of identifying term-sentiment transitions like these.

Recent archival computing work has used OCR enrichment, handwritten text separation, named entity recognition, and semantic search to improve access to digitized collections [8-13]. Explainable document linkage and sentence embedding tracing of idea similarity further suggest that semantic comparison can surface relationships that lexical search alone may miss [14,15].

Method

Initial Tests: Alignment Only Comparison

Experimentally expressing the morphisms' comparison response was conducted on texts from seven different datasets with various compositional attributes characterized by the size and quality found in OCR generated from mixed manuscript collections. Data came from two groups: natural language documents from our special collections digital archive, and noise text with various language-like features. Initial response levels of a text's morphism features were graded on a two dimensional threshold scale of highest value between alignments across any inter-document morphisms present in the set tested. The two axes record data for displacement vector cosine similarity and cluster principal component concordance. First match between any two morphisms was recorded at its highest level as the system incrementally decreased the threshold for detecting an aligned match, leaving a plot of instances where every pair of morphisms match at the lowest percent of difference between them. Representative samples from the natural language text dataset had their morphism response tracked in a dataset space constructed with natural language texts and then similar assessments were made with the natural language text in a dataset space constructed from only noise texts. An inverse of this, starting with documents made of noise text, was also conducted. Samples exhibiting highest level threshold matches between morphisms were then evaluated for lexical overlap between their texts.

Because archival OCR is often affected by mixed text types, degradation, and historical orthography, testing the behavior of the

model against noisy and language-like text is relevant to real archival conditions [8-10,16].

Initial Tests: Observations

The highest alignment values recorded for natural language text subjects had an 0.82 cosine similarity score when evaluated in high lexical term overlap natural language datasets. These texts also saw representative matches in all quantiles across both axes of evaluation. Similarly, random character noise text saw equally high values within random character noise datasets with values at 0.81 cosine similarity.

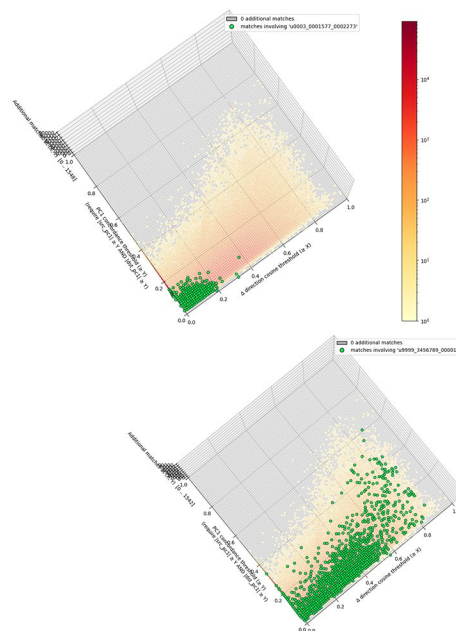


Figure 2. Natural language text morphism response to a random character noise dataset shown on top. Random character text morphism response to same noise dataset shown on bottom.

Inversely, Fig. 2 shows the response of natural language text morphisms in a noise character dataset with most of its matches falling under the first quantile on both axes and the highest value recorded only reaching 0.37 cosine similarity. Texts tracked did respond with higher threshold match values when the subject text was larger. Incremental increases in number of sentences and clusters in a document did show an increase in the number and higher alignment scores of matches present in the text.

When the natural language texts were arranged as an unsupervised finding aid on morphism likeness and inspected for content with other random, dirty OCR texts from our digital archive, lexical overlap did emerge at the end point groupings, not only in term and lemma but also in phrasing and text modality where the action of the text was working in the same way on the text's subjects. For instance, advertisement text was found grouped.

Primary Tests: Comparing Morphism, Text Composition, and Document Embeddings

The dataset used for these tests consists of items taken from the digital archive at the W. S. Hoole Special Collections Library of the University of Alabama, and from the OpenAlex repository of academic articles.

Text sourced from special collections was generated by OCR processes and contains random character noise, while the articles from OpenAlex do not.

Primary Tests: Dataset Overview

Items	800
Pages	25,525
Characters	65,172,062
Dictionary words	11,068,891
Tokens	17,314,410
Mean words/page	433
Mean tokens/page	678
Alphabet characters	73.35%
Numeric characters	1.84%
Punctuation	4.98%
Whitespace	19.48%

Collection domains of the text from special collections include humor magazines, newspapers, LGBTQ, literature review, and historical newspapers. The dataset pulled from OpenAlex was acquired by using the broad term “Study” with the intent of compiling texts from a range of fields.

Search Document Overview

Because this approach compares components of network graphs the initial search request is not queried with a word, keywords, or phrase but with a whole document. For the primary tests discussed here we centered observations on a single document from the dataset; document u0002_0000001_0000279. This document is an issue of a student publication from 1949, consisting of 28 pages, 75,725 characters, 13,887 dictionary words, 19,992 tokens, 496 mean words per page, 714 mean tokens per page, 73.69% alphabet characters, 0.78% numeric characters, 5.29% punctuation, and 19.86% whitespace characters. The landing page for the item in its current discovery system lists the subject headings: Universities and colleges--Humor--Periodicals; University of Alabama--Students--Periodicals.

Deconstructing the Search Document

The search documents body of text is tokenized at the sentence level. Each sentence segment is embedded with the SBERT model all-MiniLM-L6-v2, and those embeddings are associated with the segment they were generated from. This test uses agglomerative clustering which requires a predetermined number of cluster bins to group into before it can begin. Document network graphs used in this approach require at least three clusters to form a discernable set of network relationships, however distribution of the segments into the cluster bins can create a significant difference in cluster size. For example if a cluster C0 has 95% of the text of the document and a cluster C3 has 0.03% this detracts from the approaches supposition

that transition between clusters model an aspect of the documents discursive character. To mitigate an uneven set of cluster sizes, the document is clustered several times at a range of $k = \text{cluster_count}$. The different sets are evaluated for a best fit by weighted penalty for k values that produce clusters with a percentage of total document segments in the first and fifth quintile.

A mean value cluster embedding is calculated from each clusters segment embeddings to serve as the cluster centroid. In turn a mean value document embedding is calculated from the full set of centroid embeddings and assigned at the document level.

Displacement vectors are calculated for every directed pair of cluster centroids (C0-C1, C1-C0, ...) producing a complete graph of the documents embedding clusters connected by its displacement vector edges. The first principal component of each cluster embedding is computed through singular value decomposition and stored along with the cluster identity. The search document is now represented as a set of semantically aligned edges and principal components encoding the geometric transitions of language held in its clustered vector embeddings.

Network Graph Morphisms and Word2vec Analogy

Within the word2vec analogy hypothesis, the important claim is not that words become points in a vector space, but that relations between meanings can emerge from vector displacements. Mikolov et al. show that sufficiently large continuous word representations can encode syntactic and semantic regularities such that a relation like man to woman or country to capital is recoverable through vector arithmetic; their noteworthy example is that King – Man + Woman lands near Queen [18]. The displacement vector component of this approach draws from the word2vec analogy when it maps the displacement between clusters of document segments. Each document is treated not only as a point or even as a set of points, but as a directed system of semantic regions.

If cluster centroids are represented as C0 and C1, then the morphism edge as C1 minus C0 is interpreted as a local semantic transformation from one region of the document to another. Comparing displacement vectors across documents asks whether two items contain similarly oriented internal semantic movements, rather than asking whether two whole documents are close in embedding space. This approach extends the vector displacement intuition from word analogies [18], [19] into encoding cluster offsets within documents language so that locally analogous morphology in text embeddings networks can be compared by cosine similarity across the embedding space.

The PC1 alignment term adds a second constraint to the analogy. An aligned displacement vector cosine similarity value indicates that a pair of source to destination cluster transitions point in a similar direction, but it does not establish that the source or destination regions themselves have comparable internal semantic structure. By estimating the principal component direction of the segment embeddings inside a cluster, the approach asks whether the local semantic variance of the source or destination cluster is also concordant across documents. By adding the principal component of the cluster as a second alignment filter, a strong morphism match also asks how globally aligned the local cluster semantic transition is. This is useful because word2vec analogies have been criticized as depending partly on pairwise similarity and neighborhood effects rather than pure relational structure [20]. The displacement plus PC1 approach becomes diagnostic rather than a claim that document morphisms are exact analogies. It extends the word2vec hypothesis to a linkage of alignments across and between network graphs attempting to infer semantic relevance from a progression of alignments rather than proximity.

Graphing Alignments with Lexical Overlap Coefficient

The unpredictability of document text size across special collections OCR transcripts creates a need to assess the lexical overlap of compared clusters and the ratio of lexical containment. We calculate lexical overlap by treating each cluster's segment text as a token field and compare the vocabulary shared between two matched destination clusters. For a morphism match A: Csrc, Cdst and B: Csrc, Cdst, each destination cluster is reduced to a set of unique terms: $Uniq_A_Cdst$ and $Uniq_B_Cdst$. The lexical overlap coefficient is calculated as $|Uniq_A_Cdst \cap Uniq_B_Cdst| / \min(|A_Cdst|, |B_Cdst|)$. This differs from Jaccard similarity as the overlap coefficient is more sensitive to representing the smaller cluster's percent of overlap. If a smaller cluster's vocabulary is mostly contained inside a larger cluster, the overlap coefficient can be high even when Jaccard is modest. That makes it useful for cluster comparisons where one destination cluster has less tokens or contains a narrow lexical region and the other contains a larger or broader lexical field.

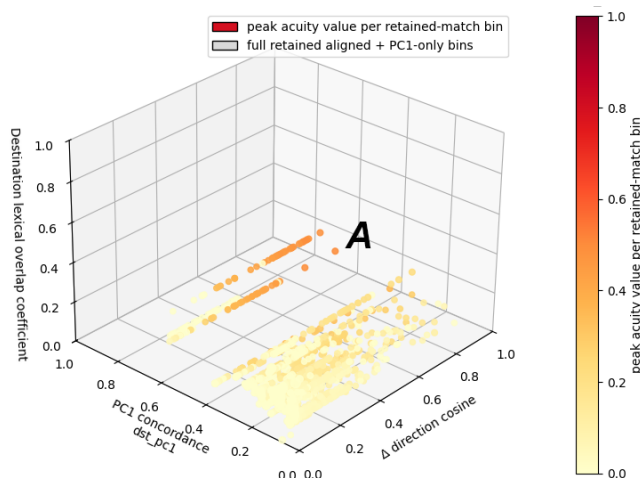


Figure 3. Three axis graph of morphism matches between document u0002_0000001_0000279 and u0002_0000001_0000222. (A) Point shows match at displacement vector cosine similarity 0.84, pc1 concordance 0.57, and lexical overlap coefficient 0.20.

Mapping matches across a three-axis graph of displacement-vector alignment, PC1 concordance, and lexical overlap shows how morphism matches distribute across geometric and lexical evidence.

A point or threshold bin high on the displacement axis means two Csrc, Cdst transformations point in a similar embedding space direction. A high PC1 value means the compared endpoint clusters have similar internal semantic orientation. A high lexical overlap value means the matched clusters share substantial vocabulary.

The most obvious matches are high on all three axes: geometrically aligned, locally concordant, and lexically similar. Fig. 3A shows a match at a high, medium, low axis profile suggesting there may be interest in this relationship. The more revealing region for this method may be high displacement alignment and high PC1 concordance but low lexical overlap. Those cases suggest that the morphism comparison is identifying a similar semantic or discursive transformation even though the cluster texts are not simply matching through shared vocabulary. Conversely, high lexical overlap with weak displacement or PC1 alignment suggests topical or vocabulary similarity without a strongly shared morphism structure.

Primary Tests: Observations

The following is a selection of matches between the search document u0002_0000001_0000279 and other documents in the test dataset that exhibit a range of match value profiles and example observations of text patterns and features found between those clusters.

Morphism match u2_1_279; C4,C1 to u2_1_222; C3,C0

Disp. vec. cos sim.	0.8485912084579468
PC1 concordance	0.5794119238853455
Doc. emb. cos sim.	0.2094784900287649
Lexical overlap coef.	0.2094017094017094

This match between two different issues of the same magazine presents as a genre to form transition rather than a document level lexical duplicate. In u0002_0000001_0000279, the source cluster C4 is dominated by comic phrases, dialogue, jokes, punchlines, campus romance, and narrative gag sequences such as a staged shipwreck story, repeated "you crazy fool" dialogue, and many brief joke blocks. The destination cluster C1 shifts into a more magazine editorial miscellany register: letters to the editor, masthead staff lists, publication information, campus news parody, institutional notices, ads, service copy, longer topical entries and registration jokes.

The u0002_0000001_0000222 edge appears to make a comparable move: C3 is also a joke, poem dialogue cluster, with comic verse, mock courtship, short one liners, dialect jokes, and gag exchanges while C0 shifts toward a mixed magazine page field containing jokes, poetry, ads, fashion copy, campus notices, and page layout fragments. The high displacement vector cosine value of 0.849, makes sense because both edges seem to move from a dense humorous dialogic cluster toward a broader periodical miscellany cluster. The moderate PC1 concordance, 0.579, suggests the destination clusters are internally organized in somewhat similar ways but not identically: both mix page furniture, ads, campus context, and humor, but their internal contents are heterogeneous. The low document embedding cosine of 0.209, says the two whole items are not globally close, and the low lexical overlap coefficient of 0.209, confirms that the match is not driven by shared wording. This is an example of the morphism method finding a parallel internal movement across two different issue's page environments from joke, narrative material into the broader magazine apparatus of campus humor, ads, editorial matter, and layout miscellany.

Morphism match u2_1_279; C4,C6 to u2_125_21; C0,C4

Disp. vec. cos sim.	0.7186671495437622
PC1 concordance	0.9088007211685181
Doc. emb. cos sim.	0.8758722340719971
Lexical overlap coef.	0.28291316526610644

This match involves the search document magazine and an issue of a quarterly literature review publication. It is driven by a shared periodical miscellany, comic literary movement, but with

different local vocabularies. In u0002_0000001_0000279, C4 is a compact humor cluster: short jokes, comic dialogues, fraternity coed punchlines, and repeated gag phrases. Its destination C6 broadens to a mixed humor magazine field: comic narrative, record reviews, cover girl fragments, ads, and a longer underage drinking routine. The matched edge in u0002_0000125_0000021, C0 contains extended fiction around characters, soup, snow, drunken driving, and a gas station rescue scene, while C4 also begins in the character narrative text but opens outward into a broader literary field, including poetry, dramatic fragments, and issue miscellany.

The high displacement vector cosine of 0.719, suggests that both edges make an internal movement from a comic scene cluster toward an editorial magazine content cluster. The very high PC1 concordance of 0.909, indicates that the destination clusters have strongly similar internal organization: both are mixed with dialogue, narrative, short form pieces, page fragments, and periodical structure. Unlike the previous example, the document embedding cosine is also high at 0.876, so the items are already close under a global embedding baseline; this match is less about discovering a hidden cross document relation and more about showing that the morphism edge is consistent with the broader document level similarity. The lexical overlap coefficient of 0.283, is moderate. There is some shared periodical and narrative vocabulary, but the match is not just duplicate wording. It appears to align a similar internal magazine shift from compact comic narrative material into broader miscellany and issue structure across two globally similar periodical objects.

Discussion

Bias in creating archival metadata using best practices with language from inherited precedent and SBERT embeddings systemic bias operate at different layers of description. Metadata bias is an assigned language effect. Titles, subject headings, genre terms, and collection descriptions reflect cataloging conventions, institutional priorities, controlled vocabularies, and the professional mandate to normalize messy objects into stable access assets. That can make unlike materials appear alike because they share a sanctioned descriptive term, or make related materials disappear from one another because the assigned language differs. SBERT bias is instead an embedding space effect.

The model positions text according to statistical associations, dominant language patterns, genre regularities, and training corpus material. The morphism alignment approach partly mitigates both effects because it does not rely only on item level metadata or a single document embedding location; it compares internal movements between clusters, then checks displacement vector alignment, PC1 concordance, semantic quality, lexical overlap, and document-level embedding proximity. In the examples we reviewed, some matches had low document embedding similarity and low lexical overlap but still showed meaningful morphism alignment, such as a movement from organized editorial community discourse into mixed periodical page miscellany across very different publications, suggesting that the method can surface parallel document functions not captured by metadata labels or document embedding proximity. At the same time, morphism categorization can exacerbate bias if treated as a neutral truth. Considering repeated layout forms, OCR artifacts, mastheads, advertisements, jokes, sponsor lists, or culturally loaded model associations can create categories simply because they are geometrically regular. The campus humor magazine examples show how strongly repeated joke, advertisement, staff, and page structures can form meaningful but form driven transitions. So the method's

advantage is not that it removes bias, but that it makes categorization more inspectable. A match can be interpreted as lexical, structural, semantic, document level, or morphism specific depending on profile values. For the activity of archival discovery document embeddings compare documents as points. This network graph morphism approach of comparison treats documents as a system of directed aligned internal semantic transformations. Where document embeddings answer the question; What documents is this document near? Morphism comparison asks; What internal semantic transitions does this document share with other documents despite how near they may be?

Thoughts on Archival Applications

In an archival setting, this approach is best understood as an interpretive discovery apparatus, not a content management system, or a digital assets manager. At archive scale, the practical challenges are substantial. OCR quality varies; sentence segmentation can turn structure like: tables, advertisements, citations, mastheads, and page components into misleading semantic signals. Cluster counts strongly affect the number of morphism edges and full pair wise comparison can become computationally time consuming unless the analysis is scoped to search documents, thresholds, or top candidate sets. Even when the system performs well, the output is not self-explanatory. High alignment values indicate that two edge movements are worth inspecting, but they do not determine the meaning of the match, and this is by design so researchers are given the opportunity to observe and arbitrate a texts utility.

The examples we reviewed show that some matches reflect shared topical content, some reflect periodical structure, some reflect genre movement, and some reflect low lexical overlap parallels in document function. That means the method requires close engagement by librarians or patrons with cluster texts, lexical overlap reports, quality diagnostics, and document embedding baselines. Its value is not that it eliminates curatorial judgment, but that it can direct expert attention toward otherwise obscured relationships. Additional application research should investigate means of arrangement by induction from morphism signatures. Recurrent patterns of internal semantic features to document form geometry was seen in the primary tests output data. By aggregating characteristic morphism shapes across many known examples of common material forms, this method could define signatures and score unknown items against those signatures. This would support unsupervised arrangement by identifying how an item behaves as a document and compares among known recurring textual regions. Such a system could provide explainable evidence for archival processing series arrangement and be used to provide access in instances where collection size or archival staff resources make it infeasible to process.

Conclusions

This system is not designed to recover a latent intent in archival documents; it is designed to negotiate the conditions where meaning becomes searchable.

By framing discovery as a post-structural mode of language generation [17], the system's abstractions are understood as situated productions that arise from contingent, affine transformation operations enacted by a researchers' selected language domain. From this perspective, ambiguity in a texts relevance is not a failure state it is a resource that can be tuned and inspected. Threshold alignment controls become a user interface component as they attempt to make the black box of computational access more legible [14]: the researcher can inspect and decide what level of overlap and alignment is creatively valuable for their question and through what

lens of language they want to situate their organizing view of collection materials. This discovery environment shifts computational access methods from opaque retrieval to term agnostic discovery that supports an unmade archive model open to their question.

Additional work exploring how clustering methods affect the grouping of sentence embeddings is crucial to all downstream imports of morphism comparisons.

Acknowledgements

Funding for this research was generously provided by The University of Alabama Office of Research and Economic Development 2025 FUSE Grant (Grant no. RG1410A). The author gratefully acknowledges Brycen Shiver graduate student from the Department of Mechanical Engineering and Ryan Adolfs graduate student from the Department of Computer Science for their role in developing the archival computational data pipeline supporting this project, and Parisa Akhtari Zavareh graduate student from the Department of Mechanical Engineering for their roll in research methods and data analysis.

References

- [1] J. Schaffner, "The Metadata is the Interface: Better Description for Better Discovery of Archives and Special Collections, Synthesized from User Studies," OCLC Research, Dublin, OH, 2009. <https://doi.org/10.25333/dp1k-3348>
- [2] J. M. Schwartz and T. Cook, "Archives, records, and power: The making of modern memory," *Archival Science*, vol. 2, pp. 1-19, 2002. <https://doi.org/10.1007/BF02435628>
- [3] Society of American Archivists, "Describing Archives: A Content Standard," version 2022.0.3.2, Chicago, IL, 2022. Available: <https://saa-ts-dacs.github.io/>
- [4] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982-3992. Available: <https://aclanthology.org/D19-1410/>
- [5] G. Wiedeman, "Designing Digital Discovery and Access Systems for Archival Description," *Code4Lib Journal*, no. 55, Jan. 2023. Available: <https://journal.code4lib.org/articles/16963>
- [6] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *WIREs Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86-97, 2012. <https://doi.org/10.1002/widm.53>
- [7] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2065, Art. 20150202, 2016. <https://doi.org/10.1098/rsta.2015.0202>
- [8] A. Jääskeläinen and L. Uosukainen, "OCR: Unleash the hidden information," *Proc. IS&T Archiving Conf.*, 2018, pp. 83-87. <https://doi.org/10.2352/issn.2168-3204.2018.1.0.19>
- [9] M. Vafaie, O. Bruns, N. Pilz, J. Waitelonis, and H. Sack, "Handwritten and Printed Text Identification in Historical Archival Documents," *Proc. IS&T Archiving Conf.*, 2022, pp. 15-20. <https://doi.org/10.2352/issn.2168-3204.2022.19.1.4>
- [10] M. Vafaie, J. Waitelonis, and H. Sack, "Improvements in Handwritten and Printed Text Separation in Historical Archival Documents," *Proc. IS&T Archiving Conf.*, 2023, pp. 36-41. <https://doi.org/10.2352/issn.2168-3204.2023.20.1.7>
- [11] A. Rattinger, G. Alliata, K. Benzi, and S. Kenderdine, "AI-Driven Metadata Extraction and Semantic Search for Audiovisual Archives," *Proc. IS&T Archiving Conf.*, 2025, pp. 112-116. <https://doi.org/10.2352/issn.2168-3204.2025.22.1.21>
- [12] V. Poso, M. Lipsanen, I. Toivanen, and T. Välsälö, "Making Sense of Bureaucratic Documents - Named Entity Recognition for State Authority Archives," *Proc. IS&T Archiving Conf.*, 2024, pp. 6-10. <https://doi.org/10.2352/issn.2168-3204.2024.21.1.2>
- [13] D. Spiliotopoulos, E. Tzoannos, C. Cabulea, and D. Frey, "Digital Archives: Semantic Search and Retrieval," in A. Holzinger and G. Pasi, eds., *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, LNCS 7947, Springer, 2013, pp. 173-182. https://doi.org/10.1007/978-3-642-39146-0_16
- [14] E. Antoine, H. J. Kang, I. Rousseau, G. Azémard, F. Bechet, and G. Damnat, "Exploring Social Sciences Archives with Explainable Document Linkage through Question Generation," in *Proc. 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 2023, pp. 141-151. Available: <https://aclanthology.org/2023.latechclfl-1.16/>
- [15] L. Li, "Tracing the Genealogies of Ideas with Sentence Embeddings," in *Proc. 4th Int. Conf. on Natural Language Processing for Digital Humanities*, 2024, pp. 9-16. <https://doi.org/10.18653/v1/2024.nlp4dh-1.2>
- [16] A. Michail, C. Raclé, J. Opitz, and S. Clematide, "Adapting Multilingual Embedding Models to Historical Luxembourgish," in *Proc. 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, 2025, pp. 291-298. Available: <https://aclanthology.org/2025.latechclfl-1.26/>
- [17] T. Cook, "Fashionable Nonsense or Professional Rebirth: Postmodernism and the Practice of Archives," *Archivaria*, no. 51, pp. 14-35, 2001. Available: <https://archivaria.ca/index.php/archivaria/article/view/12792>
- [18] Mikolov, T., Yih, W., & Zweig, G. "Linguistic Regularities in Continuous Space Word Representations." *NAACL-HLT*, 2013. Available: <https://aclanthology.org/anthology-files/pdf/N/N13/N13-1090.pdf>
- [19] Mikolov, T., Chen, K., Corrado, G., & Dean, J. "Efficient Estimation of Word Representations in Vector Space." *arXiv*, 2013. Available: <https://doi.org/10.48550/arXiv.1301.3781>
- [20] Levy, O., & Goldberg, Y. "Linguistic Regularities in Sparse and Explicit Word Representations." *CoNLL*, 2014. Available: <https://doi.org/10.3115/v1/W14-1618>

Author Biography

Jeremiah Colonna-Romano received his BFA in Fine Arts from Maryland Institute College of Art (2000), and his MFA in Painting from Indiana University (2007). Currently he is the Digital Initiatives Librarian at the University of Alabama Libraries' W. S. Hoole Special Collections Library.