Board 102: Crafting a Library on Belonging in Engineering: An Initial Review Using Textual Analysis

Ms. Denise Amanda Wetzel, Pennsylvania State University

Denise A. Wetzel (she/her) is a Science & Engineering Librarian at Pennsylvania State University Libraries. She is also the Patent and Trademark Resource Center Librarian for the University Park PTRC. She holds a MLIS from the University of Alabama and a Masters in Oceanography from Florida State University. Before joining PSU, Denise worked for Florida State University Libraries, Mississippi State University Libraries, and as a teacher.

Sara C. Kern, Pennsylvania State University

Sara Kern is an Engineering Librarian at Penn State University. She earned her MA in history from Penn State and her MSLIS at Syracuse University. Her research interests include inclusive library outreach and instruction.

Crafting a Library on Belonging in Engineering: An Initial Review using Textual Analysis

Introduction

What does it mean to belong in engineering? Who belongs in engineering? Where do libraries fit into this conversation? Many scholars have explored the concept of outreach, inclusion, accessibility, and belonging in STEM fields. This project examines a collection of these works, using a library of literature as a corpus for analysis with the intention of providing insight for future research on STEM belonging in libraries. The individual conclusions in these works are important, but examined together they may provide additional insight and understanding about creating spaces for belonging.

This work uses open-source tools both in the creation and analysis of the library; it is the authors hope that it makes both the project and the methodology more accessible for other scholars. While the library and analysis can stand alone and provide insight into themes and trends in belonging in STEM, this project is situated at the start of a larger project. The authors are interested in exploring concepts of belonging in the context of STEM libraries.

Researchers in the humanities and social sciences have increasingly used quantitative textual analysis to gain a high-level understanding of language and themes in texts. These studies are designed to investigate a corpus, or collections of words from selected texts, and use mathematical principles to look for high level connections and concepts not always apparent through individual reads [1], [2]. By moving beyond the standard literature review, the authors can examine the content of published material to explore themes and future research directions to understand belonging in STEM libraries. This methodology allowed the authors the opportunity to examine details in language and trends in texts in a corpus in a novel way that would be significantly more difficult and time consuming if read in a traditional way.

As the purpose of this study is preliminary research to inform future research projects, identifying themes and exploring high level connections between concepts is especially helpful to create a framework for future research. The authors briefly explored themes of belonging in libraries as well. Public libraries are often discussed as a “third place” - somewhere outside of home and work [3]. Researchers have recently turned this framework to academic libraries as well, in some cases using it to explore the “homeness” of a library space, a description that may tie in to understandings of belonging [4]. Because of this, the authors focused analysis on identifying research themes and insights that may be useful in the context of library inclusion efforts.

About the Corpus and Zotero

Developing the corpus for analysis involved a few different avenues. The authors employed a multi-database search approach, simultaneously using their institution's discovery layer, Educational Resources Information Center (ERIC) database, Web of Science, and Google Scholar to search for "belonging and STEM" which returned an overwhelming number of results that were not all applicable. The results were culled using "belonging AND STEM AND undergraduate student." After identifying articles central to the topic, the authors used citation chaining to narrow in on related research articles. This research strategy allows researchers the ability to review both cited works from an article and works citing the same article that may have a similar research topic. The authors also performed similar searches for "belonging AND STEM AND graduate student," "belonging AND engineering AND graduate student," "belonging AND women AND undergraduate student," and "belonging AND women AND graduate student." After gathering a much larger potential library of articles and book chapters, all articles were reviewed and narrowed down to the present library of 88 documents that fit into the analysis criteria. Knowing that this is a popular topic, citation alerts were also created for additional articles to be added for future research.

To keep track of their sources, the authors used Zotero, an open-source citation management tool with a paid option for additional storage. Zotero allowed the authors to organize PDFs before converting them to text files, take and share notes on specific resources, and create sub-collections on related topics for future research.

The authors chose to use Zotero because it allows researchers with free accounts to create researcher groups without restrictions on group size. Readers can access a public version of this library through a public Zotero group (https://www.zotero.org/groups/5364171/eld_belonging_shareable_library) and use it as a starting point for their own research on belonging in STEM. This library includes citations on belonging in STEM, which was used for this study, as well as a related sub-collection on outreach.

An Introduction to Voyant Tools

Voyant Tools is an open-source, web-based application for performing text and data mining with code available through GitHub. Developed by researchers at McGill University and the University of Alberta, this dynamic tool supports scholarly reading and text interpretation. It is often used in the digital humanities, but it can be a useful resource for any STEM librarian looking to textually analyze a body of literature. A variety of tools built within the Voyant environment allow researchers to explore connections, create interactive visualizations, and easily export results for scholarly articles, websites, and more. While both authors have

experience in other statistical software, Voyant Tools was chosen because of the shallow learning curve creating a low barrier for use. It was in the hope that the authors may be able to recommend the tool for others with no experience with statistical software.

Figure 1: Screenshot of Voyant Tools landing page.

To get started using Voyant Tools, the researcher needs to create a corpus of texts to upload or use a preloaded corpus within the application. Voyant tools can accept files in a variety of formats, including .docx, .txt, .pdf, or even web page URLs. For this project, 88 files were converted from different formats into .txt files so that all files were similar. These files are available in the Zotero library linked above.

Once the files are added, Voyant tools will open with a default arrangement of tools across five panes. These tools are Cirrus, Reader, Trends, Summary, and Contexts. There are additional tools that can be used across the five panes, and everything is easily resized within the window itself to meet the needs of the researcher. A description of the five default tools plus other relevant tools used in the textual analysis performed for this paper can be found in Table 1 below.

Tool Name	Tool Description
Cirrus	A word cloud that allows librarians to hover over words to see word counts. Larger words in the visualization equal a higher prevalence within the corpus.
Reader	The tool provides easy access to the entire text of the corpus. Click

	on any word to highlight every occurrence of that word in the reader tool. It also shows frequency of terms near the bottom of the box.10
Trends	The tool shows the relative frequency of words throughout the corpus. It will automatically display the five most common words in the corpus, but more words may be added using the text box at the bottom of the pane.
Summary	This tool shows a summary of the number of words, number of unique words, longest and shortest documents, highest and lowest vocabulary density, average number of words per sentence, most frequent words, notable peaks in frequency, and distinctive words in the corpus chosen. Distinctive words tend to correlate to proper names.
Contexts	This tool shows each occurrence of a keyword with a bit of surrounding text for context. Both the left and right few words are shared and entries can be expanded to view longer contexts.
Document Terms	This tool is similar to the Trends tool but shows term frequencies using a table view for each document within a corpus. The table can also include the Z-score (a normalized value for the term's raw frequency compared to other term frequencies in the same document) and Significance (how important a term is in a document relative to the rest of the corpus) for terms.
Scatterplot	This visualization shows how words cluster in a corpus. Librarians can engage in principal component analysis (PCA) to reduce dimensions within a multidimensional space to see associations more clearly.

Table 1: Selected Voyant Tools tool descriptions that proved relevant to this textual analysis.

Initial textual analysis findings

This collection of documents focuses on belonging across a variety of STEM disciplines and is not limited to engineering. 41 documents focus on engineering, making it just under half of the entire corpus. The authors did not limit the documents by type other than the limitation of scholarly works, so there are journal articles, book chapters, and dissertations included in the corpus. Additionally, this analysis uses research on students in different stages of their educational careers, with an eye towards understanding challenges faced by historically marginalized people within the STEM community. An initial summary of the corpus reveals that the authors analyzed 88 documents with 1,123,499 total words [5]. This analysis also shows that the most frequent words present throughout the entire corpus are STEM, students, science, women, and belonging. STEM by itself was used 9,753 times across the 88 readings.

The creation and use of a word cloud using the Cirrus tool allows the authors another view on the most frequently occurring words in the corpus. [6] Figure 2 shows this word cloud, taken with an edited list of stop words removed. Voyant Tools defaults its list of stop words for corpus analysis, but researchers have the option to edit this list. The authors chose to edit the list since terms used in academic writings but that do not add meaning to the analysis could easily be removed. Examples of additional stop words added are, “e.g.,” “i.e.,” “doi,” etc. With Voyant Tools, the authors were able to use distance reading to identify word usage patterns.

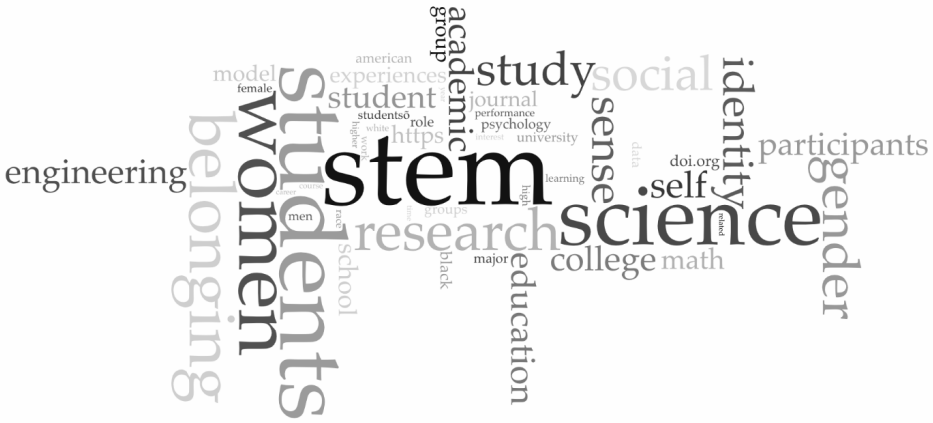

Figure 2: Word cloud of the 25 highest frequency terms in the corpus.

Considering the subject of this analysis, the most common words were unsurprising. The trends tool allowed the authors to move beyond commonly used words and better understand the relative frequencies of words within the corpus. Oftentimes the highest frequency, or raw frequency, words are apparent without analysis, but the authors found additional insight by exploring how often these words occur in relation to the total number of words. [7] In figure 3, a stacked column representation shows the four highest words across the corpus. Using this tool, readers can see that the highest occurrence of the word STEM occurs just 0.0300972 or just over 3% of the time in one document of the corpus.

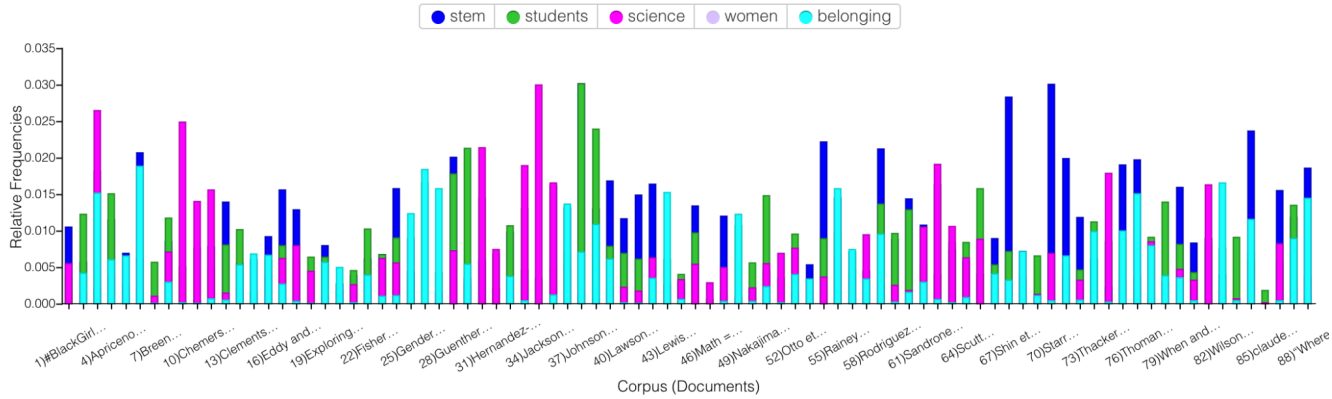

Figure 3: A stacked column representation of the relative frequency of the top 4 words across the corpus. The term “women” was removed from this analysis.

Voyant Tools also allows the researcher to identify frequently collocated terms. Collocated terms are two words that occur next to each other in the corpus. Frequently collocated terms can provide insight into the meaning and context of the individual terms in the pair. The authors used this list to identify terms that were of interest to their study and explore their frequency. A noteworthy example is shown in figure 4. It graphs occurrences of both “social belonging” (blue) and “STEM persistence” (green). Separately, the words “social belonging” are collocated 306 times in the corpus, where “STEM persistence” are collocated 278 times. These two collocated phrases were used in the same article in 13 of the 88 articles of the corpus [8]. To put this in context, figure 5 shows some of the most frequently collocated terms, such as, “science identity” (green), collocated 861 times, and “STEM belonging” (blue), collocated 823 times. While they are far more common across the corpus, they appear together in just 14 of 88 articles [9]. This project is intended to inform future research into belonging in STEM, so the authors focused on identifying terms collocated with “belonging” or “persistence.” The frequent occurrences of “social” with “belonging” and “persistence” indicates this is a theme in the published literature examined, indicating social belonging may be a fruitful area of investigation in the context of STEM libraries.

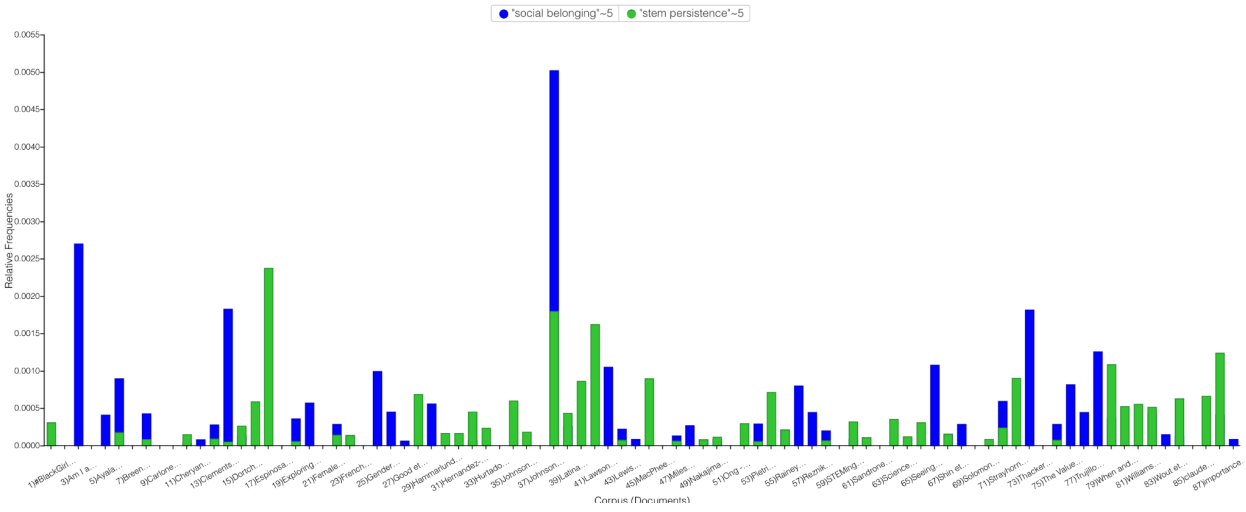

Figure 4: A stacked column representation of social belonging and STEM persistence.

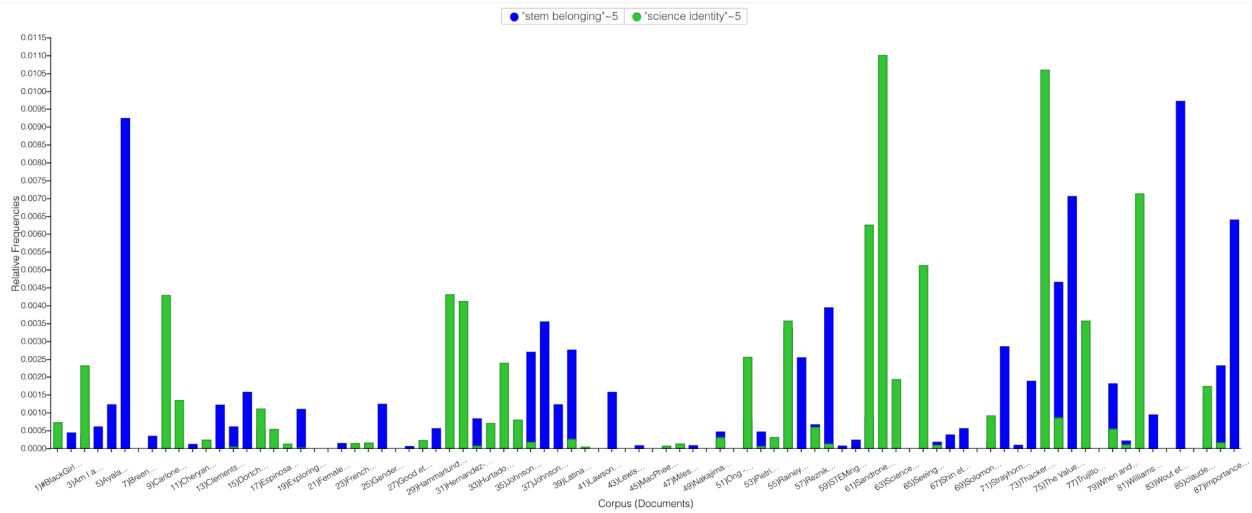

Figure 5: A stacked column representation of science identity and STEM belonging.

Conclusions and Future Research

The authors were able to build a library related to belonging in STEM that they and others can use to inform future research. Using those documents, the authors used Voyant Tools to complete high-level textual analysis with the goal of identifying noteworthy trends to inform future research projects. The main takeaway of this paper was to analyze these trends using a novel tool so that readers could see its functionality while sharing the authors' own processes. Additionally, this analysis revealed connections in the literature between identity, social experiences, persistence, and belonging. This is a preliminary study; the authors intend to examine the trends listed above more deeply both in the literature and through additional textual analysis.

Through the exploration of collocated terms in figure 4, the authors sought to further contextualize the meaning and place of students' social life and its relationship to STEM belonging. If, as many of these studies suggest, social belonging is closely tied to persistence and general belonging in STEM, and if academic libraries can serve as third places, which can support social belonging, then it may be helpful to better understand the role STEM libraries can play in supporting this aspect of belonging.

Further investigation into the corpus gathered and used for this paper's analysis will hopefully reveal trends not readily apparent in this beginning review. The authors used Voyant Tools for their initial analysis because it is open source and is accessible to novice textual analysis researchers doing most analysis. However, Voyant Tools does not allow users to re-order the corpus once it is entered. As these works were not entered in chronological order, the authors were unable to look at trends over time with this corpus.

Additionally, even after cleaning the data of non-significant words, the most used words, like “STEM” and “belonging”, were so frequently used as to make it difficult to gain insight beyond these topics being related in the context of this corpus. The authors found some interesting relationships that can inform future projects, but they also plan to re-investigate this corpus with a more robust analytical tool to search for additional insights.

Future research may make use of more robust analytical tools to delve further into this corpus and examine trends over time. That work will inform research to better understand how belonging, STEM, and libraries are linked. The authors also plan to explore a variety of disciplines within the STEM fields to better understand the inclusion needs of students across curriculum.

Citations

[1] M. Gavin, *Literary Mathematics: Quantitative Theory for Textual Studies,* 1st ed. Stanford University Press, 2022. [E-book] Available: DOI: 10.1515/9781503633919.

[2] B. McGillivray and G. M. Tóth, *Applying Language Technology in Humanities Research: Design, Application, and the Underlying Logic*, 1st ed. Palgrave Macmillan, 2020. [E-book] Available: DOI: 10.1007/978-3-030-46493-6.

[3] R. Oldenburg, *The Great Good Place: Cafes, Coffee Shops, Community Centers, Beauty Parlors, General Stores, Bars, Hangouts and How They Get You Through the Day*, 1st ed. New York: Paragon House, 1989.

[4] P. Mehta and A. Cox, "At Home in the Academic Library? A Study of Student Feelings of "Homeness", *The New Review of Academic Librarianship*, vol. 27-1, pp. 4-37, 2021. https://doi.org/10.21900/j.jloe.v3.956.

[5] S. Sinclair and G. Rockwell, "Summary," Voyant Tools. https://voyant-tools.org/?corpus=d7e1d044cff8d7577774d1a8cbbfdc74&view=Summary (accessed Jan. 9, 2024).

[6] S. Sinclair and G. Rockwell, "Cirrus," Voyant Tools. https://voyant-tools.org/?stopList=keywords-2514b81bce59676898909ccd4fbd51c3&corpus=3356a1bcf9f2a854475535eb8f9718ac&view=Cirrus. (accessed Jan. 20, 2024).

[7] S. Sinclair and G. Rockwell, "Trends," Voyant Tools. https://voyant-tools.org/?stopList=keywords-2514b81bce59676898909ccd4fbd51c3&query=stem&query=students&query=science&query=women&query=belonging&mode=&chartType=stacked&corpus=3356a1bcf9f2a854475535eb8f9718ac&view=Trends. (accessed Jan. 20, 2024).

[8] S. Sinclair and G. Rockwell, "Trends," Voyant Tools. https://voyant-tools.org/?view=Trends&query=%22stem%20persistence%22~5&query=%22social%20belonging%22~5&chartType=stacked&corpus=d7e1d044cff8d7577774d1a8cbbfdc74
(accessed Feb. 6, 2024).

[9] S. Sinclair and G. Rockwell, "Trends," Voyant Tools. https://voyant-tools.org/?view=Trends&query=%22stem%20belonging%22~5&query=%22science%20identity

%22~5&chartType=stacked&corpus=d7e1d044cff8d7577774d1a8cbbfdc74 (accessed Feb. 6, 2024).