ARTICLES

Evaluating the Impact of the Long-S upon 18th-Century Encyclopedia Britannica Automatic Subject Metadata Generation Results

Sam Grabus

ABSTRACT

This research compares automatic subject metadata generation when the pre-1800s Long-S character is corrected to a standard < s >. The test environment includes entries from the third edition of the Encyclopedia Britannica, *and the HIVE automatic subject indexing tool. A comparative study of metadata generated before and after correction of the Long-S demonstrated an average of 26.51 percent potentially relevant terms per entry omitted from results if the Long-S is not corrected. Results confirm that correcting the Long-S increases the availability of terms that can be used for creating quality metadata records. A relationship is also demonstrated between shorter entries and an increase in omitted terms when the Long-S is not corrected.*

INTRODUCTION

The creation of subject metadata for individual documents is long known to support standardized resource discovery and analysis by identifying and connecting resources with similar aboutness.[1] In order to address the challenges of scale, automatic or semi-automatic indexing is frequently employed for the generation of subject metadata, particularly for academic articles, where the abstract and title can be used as surrogates in place of indexing the full text. When automatically generating subject metadata for historical humanities full texts that do not have an abstract, anachronistic typographical challenges may arise. One key challenge is that presented by the historical "Long-S" < ſ >. In order to account for these idiosyncrasies, there is a need to understand the impact that they have upon the automatic subject indexing output. Addressing this challenge will help librarians and information professionals to determine whether or not they will need to correct the Long-S when automatically generating subject metadata for full-text pre-1800s documents.

The problem of the Long-S in Optical Character Recognition (OCR) for digital manuscript images has been discussed for decades.[2] Many scholars have researched methods for correcting the Long-S through the use of rule-based algorithms or dictionaries.[3] While the problem of the Long-S is well-known in the digital humanities community, automatic subject metadata generation for a large corpus of pre-1800s documents is rare, as is research about the application and evaluation of existing automatic subject metadata generation tools on 18th-century documents in real-world information environments. The impact of the Long-S upon automatic subject metadata generation results for pre-1800s texts has not been extensively explored. The research presented in this paper addresses this need. The paper reports results from basic statistical analysis and visualization using the Helping Interdisciplinary Vocabulary Engineering (HIVE) tool automatic

Sam Grabus (smg383@Drexel.edu) is an Information Science PhD Candidate at Drexel University's College of Computing and Informatics, and Research Assistant at Drexel's Metadata Research Center. This article is the 2020 winner of the LITA/Ex Libris Student Writing Award.

https://doi.org/10.6017/ital.v39i3.12235

subject indexing results, before and after the correction of the historical Long-S in the 3rd edition of the *Encyclopedia Britannica*. Background work was conducted over the Summer and Fall of 2019, and the research presented was conducted during Winter 2020. The work was motivated by current work on the "Developing the Data Set of Nineteenth-Century Knowledge" project, a National Endowment for the Humanities collaborative project between Temple University's Digital Scholarship Center and Drexel University's Metadata Research Center. The grant is part of a larger project, Temple University's "19th-Century Knowledge Project," which is digitizing four historical editions of the *Encyclopedia Britannica*.[4] The next section of this paper presents background covering the historical *Encyclopedia Britannica* data, the automatic subject metadata generation tool used for this project, a brief background of "the Long-S Problem," and the distribution of encyclopedia entry lengths in the 3rd edition. The background section will be followed by research objectives and method supporting the analysis. Next, the results are presented, demonstrating prevalence of terms omitted from the automatic subject metadata generation results if the Long-S is not corrected to a standard small < s > character, as well as the impact of encyclopedia entry length upon these results. The results are followed by a contextual discussion, and a conclusion that highlights key findings and identifies future research.

BACKGROUND

Indexing for the 19th-Century Knowledge Project

The 19th-Century Knowledge Project, an NEH-funded initiative at Temple University, is fully digitizing four historical editions of the *Encyclopedia Britannica* (the 3rd, 7th, 9th, and 11th). The long-term goal of the project is to analyze the evolving conceptualization of knowledge across the 19th century.[5] The 3rd edition of the *Encyclopedia Britannica* (1797) is the earliest edition being digitized for this project. The 3rd edition consists of 18 volumes, with a total of 14,579 pages, and individual entries ranging from four to over 150,000 words. For each individual entry, researchers at Temple have created individual TEI-XML files from the OCR output.

In order to enrich accessibility and analysis across this digital collection, The Knowledge Project will be adding controlled vocabulary subject headings into the TEI headers of each encyclopedia entry XML file. Considering the size of this corpus, both in terms of entry length and number of entries, automatic subject metadata generation will be required for the creation of this metadata. The Knowledge Project will employ controlled vocabularies to replace or complement naturally extracted keywords for this process. Using controlled vocabularies adheres to metadata semantic interoperability best practices, ensures representation consistency, and helps to bypass linguistic idiosyncrasies of these 18th and 19th Century primary source materials.[6] We selected two versions of the Library of Congress Subject Headings (LCSH) as the controlled vocabularies for this project. LCSH was selected due to its relational thesaurus structure, multidisciplinary nature, and continued prevalence in digital collections due to its expressiveness and status as the largest general indexing vocabulary.[7] In addition to the headings from the 2018 edition of LCSH, headings from the 1910 LCSH are also implemented in order to provide a more multi-faceted representation, using temporally-relevant terms that may have been removed from the contemporary LCSH.

The tool applied for this process is HIVE, a vocabulary server and automatic indexing application.[8] HIVE allows the user to upload a digital text or URL, select one or more controlled vocabularies, and performs automatic subject indexing through the mapping of naturally extracted keywords to the available controlled vocabulary terms. HIVE was initially launched as an IMLS linked open

vocabulary and indexing demonstration project in 2009. Since that time, HIVE has been further developed, with the addition of more controlled vocabularies, user interface options, and the RAKE keyword extraction algorithm. The RAKE keyword extraction algorithm has been selected for this project after a comparison of topic relevance precision scores for three keyword extraction algorithms.[9]

The Long-S Problem

Early in our metadata generation efforts, we discovered that the 3rd edition of the *Encyclopedia Britannica* employs the historical Long-S. Originating in early Roman cursive script, the Long-S was used in typesetting up through the 18th century, both with and without a left crossbar. By the end of the 18th century, the Long-S fell out of use with printers.[10] As outlined by lexicographers of the 17th and 18th centuries, the rules for using the Long-S were frequently vague, complicated, inconsistent over time, and varied according to language (English, French, Spanish, or Italian).[11] These rules specified where in a word the Long-S should be used instead of a short < s >, whether it is capitalized, where it may be used in proximity to apostrophes, hyphens, and the letters < f >, < b >, < h >, and < k >; and whether it is used as part of a compound word or abbreviation.[12] This is further complicated by the inclusion of the half-crossbar, which occasionally results in two consequences: (a) The Long-S may be interpreted by OCR as an < f >, and < b > and < f > may be interpreted by OCR as a Long-S. Figure 1 shows an example from the 3rd edition entry on *Russia*, in which the original text specifies "of" (line 1 in top figure), yet the OCR output has interpreted the character as a Long-S. The Long-S may also occasionally be interpreted by the OCR as a lower-case < l >, such as the "univerlity of Dublin" in the 3rd edition entry on *Robinson (The most Rev Sir Richard)*.

These complications and inconsistencies are challenges when developing Python rules for correcting the Long-S in an automated way, and even preexisting scripts will need to be adapted for individual use with a particular corpus.

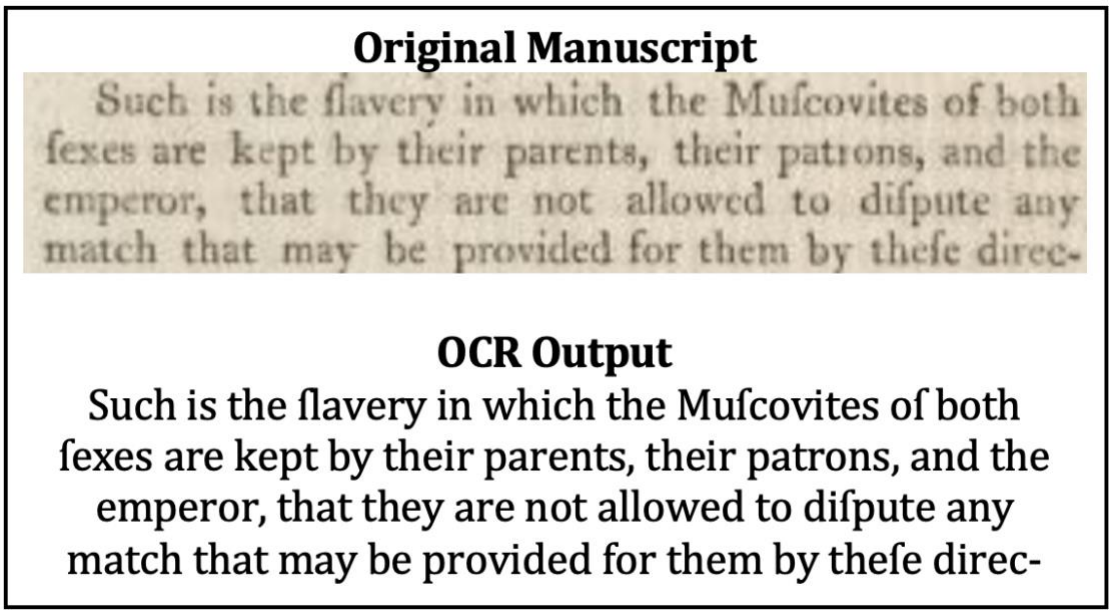

Figure 1. Example from the 3rd edition entry on *Russia*, comparing the original use of a letter < f > in "of" to the OCR output of the same passage, which mistakenly interprets the character as a Long-S.

Despite the transition away from the Long-S towards the end of the 18th century, the 3rd edition of the *Encyclopedia Britannica* (published in 1797) implements the Long-S throughout, with approximately 100,594 instances of the Long-S in the OCR output. When performing metadata generation with the HIVE tool on the OCR output for an entry, the Long-S is most often interpreted by the automatic metadata generation tool as an < f >, which can result in (a) inaccurate keyword extraction (e.g., Russians→ Ruffians), and (b) when mapping extracted keywords to controlled vocabulary terms, essential topics could be unidentifiable, and HIVE will subsequently omit them from the results because they cannot be mapped to controlled vocabulary terms. Figure 2 provides a truncated view of Long-S words in the 3rd edition entry on *Rum*, which are subsequently removed from the pool of automatically extracted keywords when performing the automatic subject indexing sequence in HIVE. Using keyword extraction algorithms that are largely dependent upon term frequencies, automatic subject indexing for an entry on *Rum* may be substantially hindered when meaningful and frequently occurring words such as *sugar,* and *yeast* are removed.

"RUM a **ſpecies** of brandy or vinous **ſpirits**, **diſtilled** from **ſugar-canes**. Rum, according to Dr Shaw, differs from **ſimple ſugar-ſpirit**, in that it contains more of the natural flavour or **eſſential** oil of the **ſugar-cane** ; a great deal of raw juice and parts of the cane **itſelf** being often fermented in the liquor or **ſolution** of which the rum is prepared. The unctuous or oily flavour of rum is often **ſuppoſed** to proceed from the large quantity of fat **uſed** in boiling the **ſugar** ; which fat, indeed, if **coarſe**, will **uſually** give a **ſtinking** flavour to the **ſpirit** in our **diſtillations** of the **ſugar** liquor or waft, from our refining **ſugar-houſes** ;, but this is nothing of kin to the flavour of the rum, which is really the effect of the natural flavour of the cane. The method of making rum is this : When a **ſufficient ſtock** of the materials are got together, they add water to them, and ferment them in the common method, though the fermentation is always carried on very **ſlewly** at **firſt** ; **becauſe** at the beginning of the **ſeaſon** for making rum in the **iſlands,** they want **yeaſt** or some other ferment to make it work : but by degrees, after this, they procure a **ſufficient** quantity of the ferment, which **riſes** up as a head to the liquor in the operation ; and thus they are able afterwards to ferment and make their rum with a great deal of expedition, and in large quantities. When the **waſh** is fully fermented, or to a due degree of acidity, the **diſtillation** is carried on in the. common way, and the **ſpirit** is made up proof : though **ſometimes** it is reduced to a much

Figure 2. Examples of the Long-S in the 3rd edition Encyclopedia Britannica entry on *Rum*.

Using this example entry, the automatic subject indexing results were compared using Python, to determine which terms only appear when the Long-S has been corrected to the standard < s >. The comparison showed that 16 total terms no longer appeared in the results when the Long-S was not corrected to a standard < s >: ten terms using the 2018 LCSH, and six terms using the 1910 LCSH. These omitted results included the terms *sugar* and *yeast.*

The next section will discuss the encyclopedia entry word count for this corpus, and the possible impact that this may have upon automatic subject indexing between corrected and uncorrected Long-S instances.

Encyclopedia Entry Lengths

Consistent with other *Encyclopedia Britannica* editions in the 18th and 19th centuries, the encyclopedia entries in the 3rd edition vary substantially in length. A convenience sample of 3,849 3rd edition entries ranging in length from 2 to 202,848 words demonstrated an arithmetic mean of

826.60, and a median word count of 71. As shown in figure 3, this indicates a significant skew towards shorter entry lengths. For the vast majority of encyclopedia entries in this corpus, a low total word count may impact the degree of Long-S impact for automatic subject indexing results, given the importance of term availability and frequency for keyword extraction algorithms.

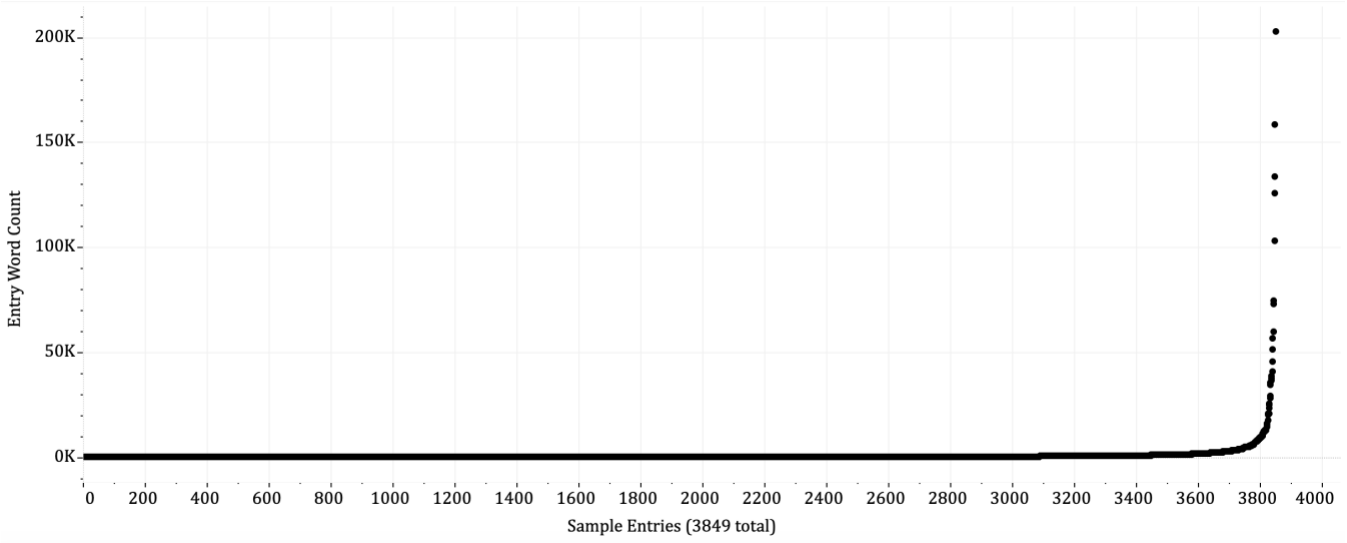

Figure 3. Scatterplot of word count for a convenience sample of 3,849 3rd Edition *Encyclopedia Britannica* entries.

Large-scale metadata generation requires time, labor, and resources, and it becomes more costly when accounting for the complications of correcting the Long-S for a particular corpus. Library and information professionals working with digital humanities resources will need to understand the impact of correcting or not corrected the Long-S in the corpus before designating resources and developing a protocol for generating the automatic or semi-automatic metadata for full-text resources. This includes understanding whether or not the length of each individual document will affect the degree of Long-S impact upon the results. This challenge, and issues reviewed above, are in the research presented below.

OBJECTIVES

The overriding goal of this work is to determine the prevalence of omitted terms in automatic subject indexing results when the Long-S is not corrected in the 3rd edition entries of the *Encyclopedia Britannica*.

Research questions:

1. What is the average number of terms that are omitted from automatic subject indexing results when the Long-S is not corrected to a standard < s >?
2. How does the encyclopedia entry length affect the number of terms that are omitted when the Long-S is not corrected to a standard < s >?

This analysis will approach these goals by performing a comparative analysis of automatic subject indexing results to determine the number of terms that are omitted from the results when the Long-S is not corrected to a standard letter < s >. Basic descriptive statistics are generated to determine central tendency. The quantity of terms omitted are then compared with encyclopedia

entry word counts. These objectives were shaped by collaboration between Drexel University's Metadata Research Center and Temple University's Digital Scholarship Center. The next section of this paper will report on methods and steps taken to address these objectives.

METHODS

We approached this research by performing a comparative analysis of subject metadata generated both before and after the correction of the historical Long-S in the 3rd edition of the *Encyclopedia Britannica*. The HIVE tool was used to automatically generate the subject metadata. Descriptive statistics were applied, and visualizations produced from the results were also examined to identify trends.

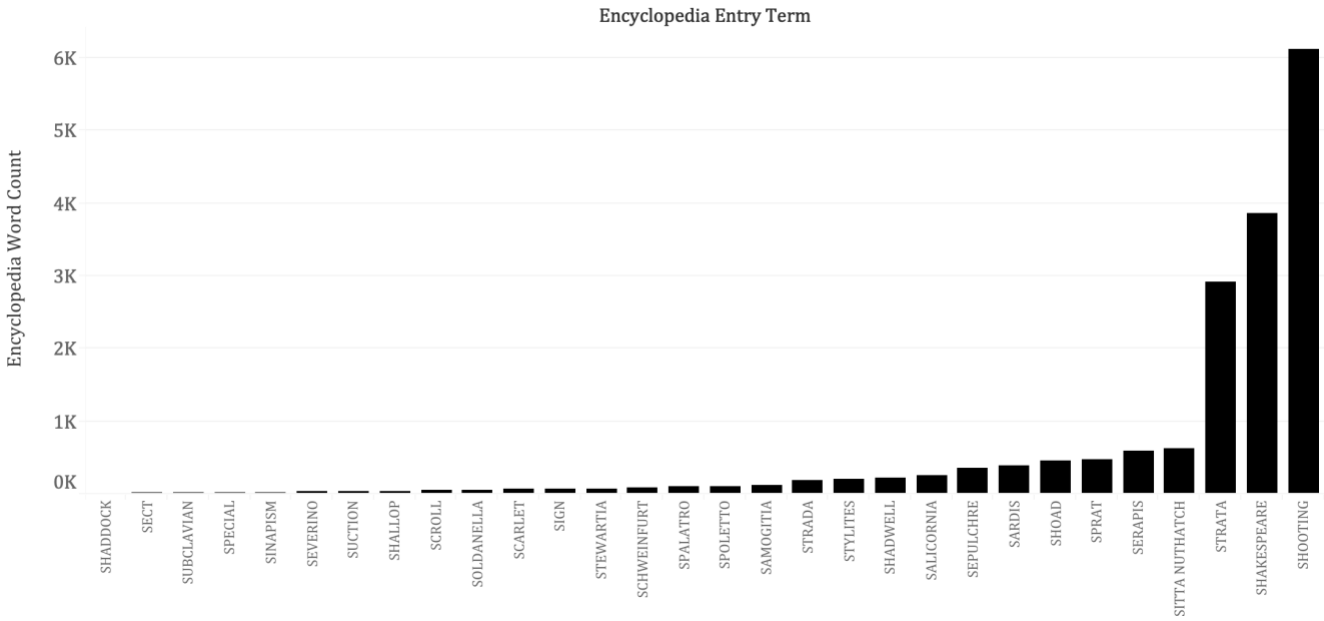

Figure 4. The 30 Encyclopedia Britannica 3rd edition *Encyclopedia Britannica* entries randomly selected for this study, sorted in ascending order by their word counts.

The protocol for performing this research involved the following steps:

1. Compile a sample for testing:
   1.1. A random sample of 30 encyclopedia entries was identified from a convenience sample of entries that comprise the letter S volumes of the 3rd edition. The entries range, in length, from 6 to 6,114 words. The median word count for entries in this sample is 99 words.
   1.2. The sample of terms selected for this study and their respective word counts are visualized in figure 4.
   1.3. For each entry, the Long-S terms in the original XML file were extracted to a list.
2. Perform automatic subject indexing sequence upon entries to generate lists of terms:
   2.1. Using the 2018 and 1910 versions of the LCSH.
   2.2. With fixed maximum subject heading results set to 40: 20 maximum terms returned with the 2018 LCSH, and 20 maximum terms returned with the 1910 LCSH.
   2.3. Before Long-S correction and after Long-S correction, using the Oxygen XML Editor TEI to TXT transformation.

3. Perform outer join on Python Data Frames, between terms generated when the Long-S has been corrected vs. terms generated when the Long-S has not been corrected. The resulting left outer join list displays terms that are omitted from the automatic indexing results if the Long-S is not corrected to a standard small < s >. The quantity of terms omitted are recorded for comparison.
4. Analysis: Descriptive statistics were generated to determine central tendency for the number and percentage of words omitted when the Long-S is not corrected. The quantity of terms omitted are also visualized in a continuous scatterplot with the corresponding word counts, to demonstrate that the quantity of terms omitted when the Long-S is not corrected seems to relate to the length of the document being automatically classified.

RESULTS

The results report the prevalence of omitted terms when the Long-S is not corrected to a standard < s >, as well as a visualization of the number of terms omitted as they relate to the encyclopedia entry length. For each of the 30 sample entries automatically indexed with HIVE, a fixed maximum number of 40 entries were returned: a maximum of 20 terms using the 2018 LCSH, and a maximum of 20 terms using the 1910 LCSH. As seen in table 1, central tendency is measured using the arithmetic mean and median, along with the standard deviation and range. The average number of terms omitted from an entry's results is 6.73, and the average percentage of terms omitted from an entry's results is 26.51 percent, with the 2018 and 1910 editions of LCSH performing at similar rates. The full results are displayed in appendix A.

Table 1. Measures of centrality, standard deviation, range, and percentage for quantity of terms omitted when the Long-S is not corrected to a standard < s >, rounded to the hundredth. For each entry, a maximum of 40 terms were returned: 20 using 2018 LCSH and 20 using 1910 LCSH. The total results returned varies according to the entry length. These totals are reported in appendix B. (N= 30 entries.)

	Both Vocabularies	2018 LCSH	1910 LCSH
Average, Terms Omitted	6.73	3.67	3.07
Median, Terms Omitted	5	3	2
Standard Deviation	6.53	3.84	3.17
Range, Terms Omitted	0-24	0-13	0-11
Average Percentage, Omitted Terms	26.51%	27.51%	24.28%
Median Percentage, Omitted Terms	22.36%	20.00%	19.09%

For each entry in the sample, the results in appendix A display the total words omitted when the Long-S is not corrected, the number of 2018 LCSH terms omitted, the number of 1910 LCSH terms omitted, and the encyclopedia entry word count. Figure 5 visualizes the total number of terms omitted for each entry when the Long-S is not corrected, demonstrating an increase in terms omitted for entries with lower word counts. These results are broken down by vocabulary used in figure 6, demonstrating that both vocabularies used to generate these results indicate a significant increase in omitted terms for shorter entries.

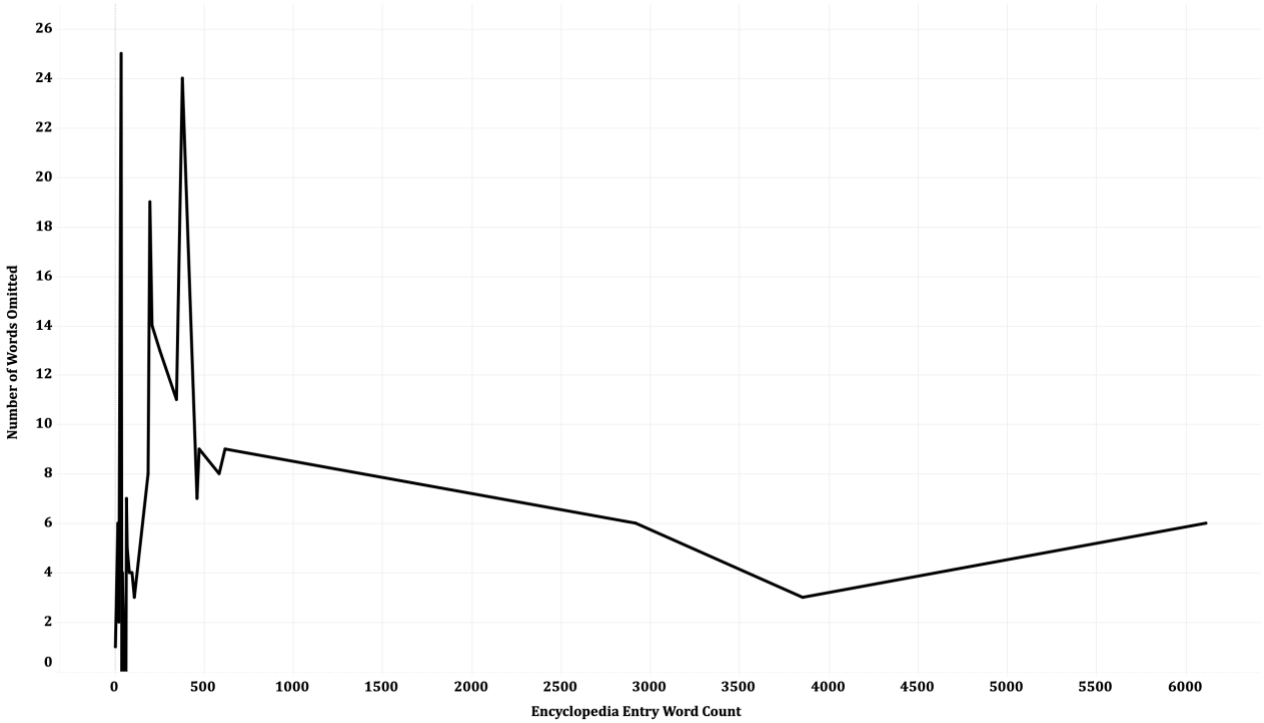

Figure 5. Number of automatic subject indexing terms that are omitted when the Long-S is not corrected to a standard < s > as compared by encyclopedia entry word count.

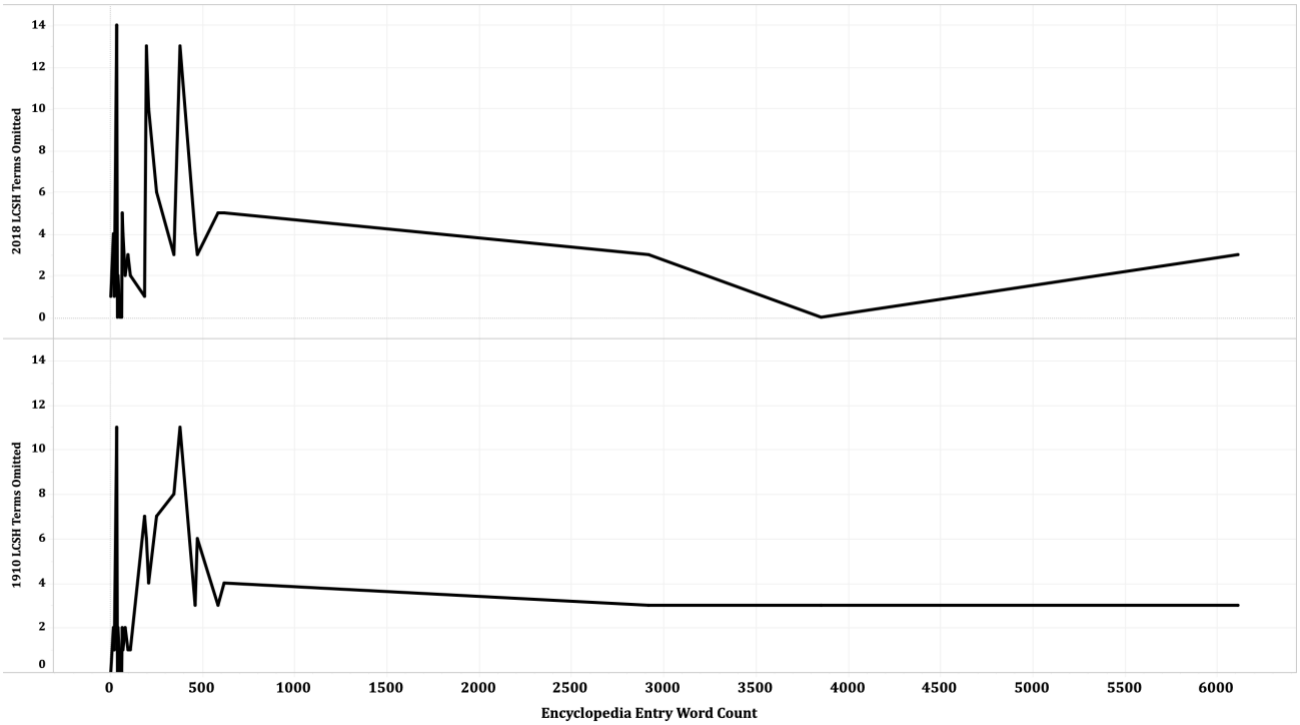

Figure 6. Number of automatic subject indexing terms that are omitted when the Long-S is not corrected to a standard < s > as compared by encyclopedia entry word count, separated by controlled vocabulary version.

DISCUSSION

The analysis above presents measures of centrality for quantity of terms omitted if the Long-S is not corrected to a standard < s > prior to automatic subject indexing using HIVE, as well as a visualization to represent the relationship between encyclopedia entry word count and number of terms omitted. Although researchers have identified challenges with the Long-S and have focused a great deal on the technologies and methods used to correct it, there is still limited work on looking at the results of not correcting the Long-S character when performing an automatic subject indexing sequence.

This research demonstrated an average of 6.73 potentially relevant terms omitted from automatic indexing results when the Long-S is not corrected, accounting for an average of 26.51 percent of the total results, with an approximately equal distribution of omitted terms across the two controlled vocabulary versions used. When the quantity of terms omitted is visualized using a continuous scatterplot, the results also demonstrated a significant increase in omitted terms for shorter entries, with longer entries less affected. These results reflect the impact of term frequency and total word count in keyword extraction and automatic subject indexing, with longer documents having a greater pool of total terms from which to identify key terms.

Considering the complexities and similarities of the typographical characters in the original manuscript, the OCR output process for this corpus occasionally mistakes the letters < s >, < f >, < r >, and < l >. As a result, an occasional Long-S word in this study did not originally contain an < s > (e.g., *sor* instead of *for*). Correction of these Long-S OCR errors requires the development of a dictionary-based script. An additional complication of this research is that the corrected OCR output for the encyclopedia entries still contains a few errors not related to the Long-S, which will prevent the mapping of the term to any controlled vocabulary term (e.g., in the entry on *Sepulchre,* the OCR output for the term *Palestine* was *Palestinc*).

These results are specific to this particular corpus of 3rd edition *Encyclopedia Britannica* entries, but it is very likely that testing another set of pre-1800s documents containing the Long-S would also illustrate that for best results with any algorithm or tool, the Long-S needs to be corrected. The results are also specific to the two versions of the LCSH used, both the 1910 LCSH and the 2018 LCSH, which are available in the HIVE tool. The 1910 version is key for the time period being studied, and the 2018, more contemporary to today, has supported additional analysis on the impact of the Long-S. Both of these vocabularies are important to the larger 19th-Century Knowledge Project. It should be noted that while the LCSH is updated weekly, we were limited to what is available via the HIVE tool, and any discrepancies that may be found with the 2020 LCSH will very likely have a minimal effect upon metadata generation results. It should be noted that the 2020 LCSH will be incorporated into HIVE soon and can be explored in future research.

CONCLUSION AND NEXT STEPS

The objective of this research was to determine the impact of correcting the Long-S in pre-1800s documents when performing an automatic metadata generation sequence using keyword extraction and controlled vocabulary mapping. This was accomplished by performing an automatic subject indexing sequence using the HIVE tool, followed by a basic statistical analysis to determine the quantity of terms omitted from the results when the Long-S is not corrected to a standard < s >. The number of omitted terms was also compared with the encyclopedia entry word count and visualized to demonstrate a significant increase in omitted terms for shorter

encyclopedia entries. The study was conclusive in confirming that the correction of the Long-S is a critical part of our workflow.

The significance of this research is that it demonstrates the necessity of correcting the Long-S prior to performing an automatic subject indexing on historical documents. Beyond the correction of the Long-S, the larger next steps for this project are to continue to explore automatic metadata generation for this corpus. These next steps include the comparison of results using contemporary vs. historical vocabularies and streamlining a protocol for bulk classification procedures and integration of terms into the TEI-XML headers. The research presented here can inform other digital humanities and even science-oriented projects, where researchers may not be aware of the impact of the Long-S on automatic metadata generation not only for subjects, but also named entities, particularly when automatic approaches with controlled vocabularies are desired.

ACKNOWLEDGEMENTS

The author thanks Dr. Jane Greenberg and Dr. Peter Logan for their guidance. The author acknowledges the support of the NEH grant #HAA-261228-18.

APPENDIX A

Entry Term	Total Words Omitted	2018 LCSH Terms Omitted	1910 LCSH Terms Omitted	Encyclopedia Entry Word Count
SARDIS	24	13	11	381
SUCTION	24	13	11	38
STYLITES, PILLAR SAINTS	19	13	6	199
SHADWELL	14	10	4	211
SALICORNIA	13	6	7	254
SEPULCHRE	11	3	8	348
SITTA NUTHATCH	9	5	4	620
SPRAT	9	3	6	475
SERAPIS	8	5	3	587
STRADA	8	1	7	189
SHOAD	7	4	3	463
SIGN	7	5	2	68
SHOOTING	6	3	3	6114
STRATA	6	3	3	2920
STEWARTIA	5	4	1	72
SUBCLAVIAN	5	3	2	20
SCHWEINFURT	4	2	2	84
SCROLL	4	2	2	45
SPALATRO	4	3	1	99
SPECIAL	4	3	1	24
SAMOGITIA	3	2	1	112
SHAKESPEARE	3	0	3	3855
SINAPISM	2	1	1	25
SECT	1	1	0	20
SEVERINO	1	1	0	38
SHADDOCK	1	1	0	6
SCARLET	0	0	0	65
SHALLOP, SHALLOOP	0	0	0	42
SOLDANELLA	0	0	0	56
SPOLETTO	0	0	0	99

APPENDIX B

*N = 30 entries	Average Terms Returned	Median Terms Returned
Corrected	24.77 / 40 possible	28 / 40 possible
Uncorrected	26.47 / 40 possible	29 / 40 possible
2018 LCSH Corrected	14.10 / 20 possible	19 / 20 possible
2018 LCSH Uncorrected	13.47 / 20 possible	18.5 / 20 possible
1910 LCSH Corrected	11.27 / 20 possible	11 / 20 possible
1910 LCSH Uncorrected	10.13 / 20 possible	9 / 20 possible

ENDNOTES

[1] Liz Woolcott, "Understanding Metadata: What is Metadata, and What is it For?," Routledge (November 17, 2017), https://doi.org/10.1080/01639374.2017.1358232; Koraljka Golub et al., "A framework for evaluating automatic indexing or classification in the context of retrieval," *Journal of the Association for Information Science and Technology* 67, no. 1 (2016), https://doi.org/10.1002/asi.23600; Lynne C. Howarth, "Metadata and Bibliographic Control: Soul-Mates or Two Solitudes?," *Cataloging & Classification Quarterly* 40, no. 3-4 (2005), https://doi.org/10.1300/J104v40n03_03.

[2] A. Belaid et al., "Automatic indexing and reformulation of ancient dictionaries" (paper presented at the First International Workshop on Document Image Analysis for Libraries, Palo Alto, CA, 2004), https://doi.org/10.1109/DIAL.2004.1263264.

[3] Beatrice Alex et al., "Digitised Historical Text: Does it have to be mediOCRe" (paper presented at the KONVENS 2012 (LThist 2012 workshop), Vienna, September 21, 2012); Ted Underwood, "A half-decent OCR normalizer for English texts after 1700," *The Stone and the Shell*, December 10, 2013, https://tedunderwood.com/2013/12/10/a-half-decent-ocr-normalizer-for-english-texts-after-1700/.

[4] "Nineteenth-century knowledge project," (GitHub Repository), 2020, https://tu-plogan.github.io/.

[5] "Nineteenth-century Knowledge Project."

[6] Marcia Lei Zeng and Lois Mai Chan, "Metadata Interoperability and Standardization - A Study of Methodology, Part II," *D-Lib Magazine* 12, no. 6 (2006); G. Bueno-de-la-Fuente, D. Rodríguez Mateos, and J. Greenberg, "Chapter 10 - Automatic Text Indexing with SKOS Vocabularies in HIVE" (Elsevier Ltd, 2016); Sheila Bair and Sharon Carlson, "Where Keywords Fail: Using Metadata to Facilitate Digital Humanities Scholarship," *Journal of Library Metadata* 8, no. 3 (2008), https://doi.org/10.1080/19386380802398503.

[7] John Walsh, "The use of Library of Congress Subject Headings in digital collections," *Library Review* 60, no. 4 (2011), https://doi.org/10.1108/00242531111127875.

[8] Jane Greenberg et al., "HIVE: Helping interdisciplinary vocabulary engineering," *Bulletin of the American Society for Information Science and Technology* 37, no. 4 (2011), https://doi.org/10.1002/bult.2011.1720370407.

[9] Sam Grabus et al., "Representing Aboutness: Automatically Indexing 19th- Century Encyclopedia Britannica Entries," NASKO 7 (2019), pp. 138-48, https://doi.org/10.7152/nasko.v7i1.15635.

[10] Karen Attar, "S and Long S," in *Oxford Companion to the Book*, eds. Michael Felix Suarez and H. R. II Woudhuysen (Oxford: Oxford University Press, 2010); Ingrid Tieken-Boon van Ostade, "Spelling systems," in *An Introduction to Late Modern English* (Edinburgh University Press, 2009).

[11] Andrew West, "The Rules for Long-S," *TUGboat* 32, no. 1 (2011).

[12] Attar, "S and Long S."