



SKOS OF THE 1910 LIBRARY OF CONGRESS SUBJECT HEADING FOR THE TRANSFORMATION OF THE KEYWORDS TO CONTROLLED VOCABULARY OF THE NINETEENTH-CENTURY ENCYCLOPEDIA BRITANNICA

SONIA PASCUA, JANE GREENBERG, PETER LOGAN, AND JOAN BOONE

OUTLINE



Introduction



Background and
Motivation



Statement of the
Problem



Methodology



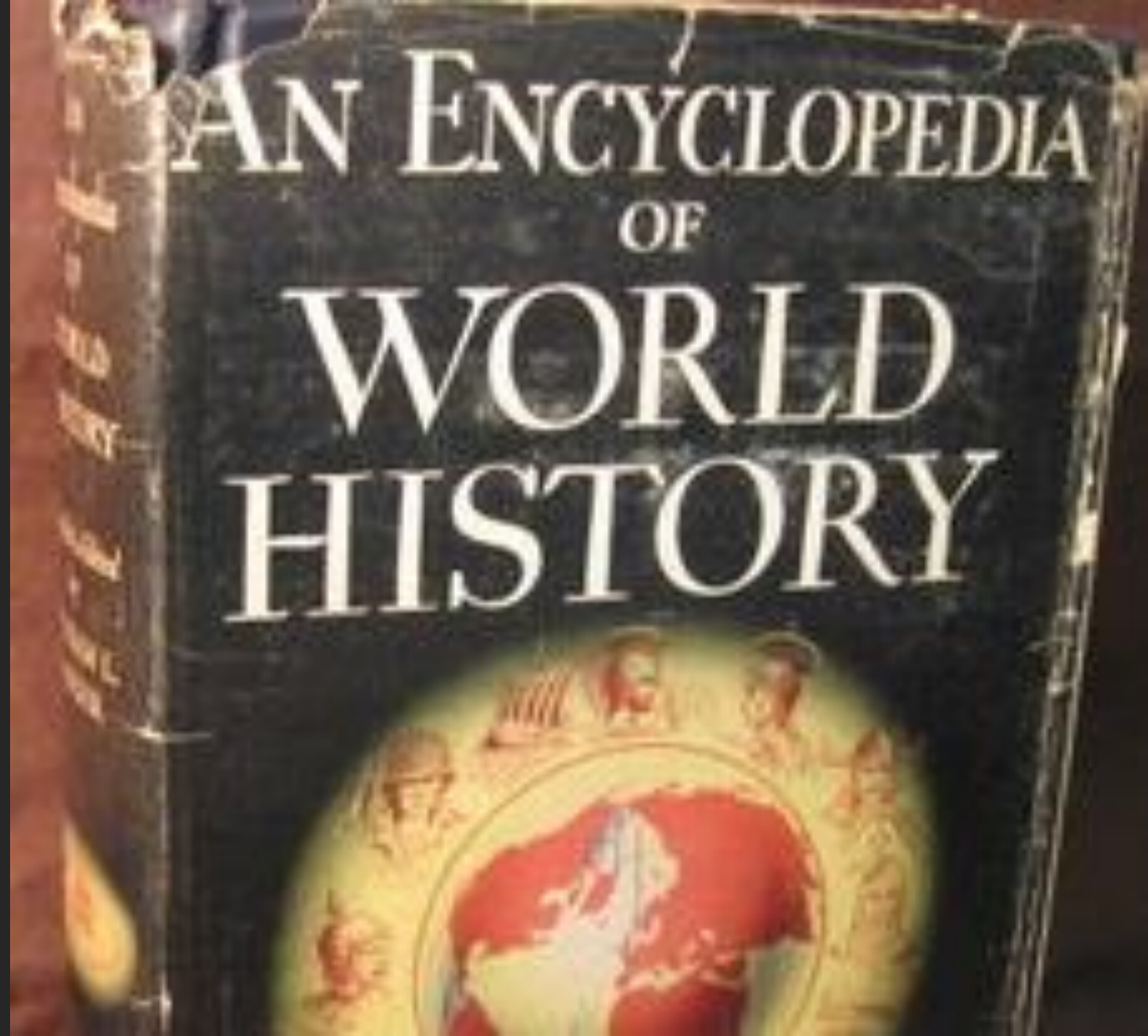
Results and
Findings



Conclusion

INTRODUCTION

“ENCYCLOPEDIA,” AS A
GENRE, IS A SIGNIFICANT
RESOURCE FOR
RESEARCHERS ACROSS
THE HUMANITIES, ARTS,
AND SCIENCES.



WHEN IN DOUBT—"LOOK IT UP" IN
The
Encyclopaedia Britannica



(New 11th Edition) issued 1910-11 by the
CAMBRIDGE UNIVERSITY PRESS (England)

The Sum of Human
Knowledge

29 volumes, 28,150 pages,
44,000,000 words of text.
*Printed on thin, but strong
opaque India paper, each
volume but one inch in
thickness.*

THE BOOK TO ASK QUESTIONS OF

FOR READING OR FOR STUDY

This Photo by Unknown author is licensed under [CC BY-SA](#).

INTRODUCTION

RESEARCHERS CAN USE
ENCYCLOPEDIAS PRODUCED
DURING EARLIER TIME-PERIODS
FOR STUDY BECAUSE THEY
PROVIDE ACCOUNTS NOT
ONLY OF HISTORY BUT ALSO
EVIDENCE OF HOW
KNOWLEDGE WAS PERCEIVED
AND ORGANIZED.

BACKGROUND OF THE STUDY

The 19th Century Knowledge Project is digitizing historical editions of the Encyclopedia Britannica (1797-1911)

This initiative will offer one of the most extensive, open, digital collections available today for studying the structure of 19th-century knowledge and its transformation

The metadata activities are being pursued through a collaboration with the Metadata Research Center (MRC) at Drexel University, supported by the U.S. National Endowment for the Humanities (NEH).

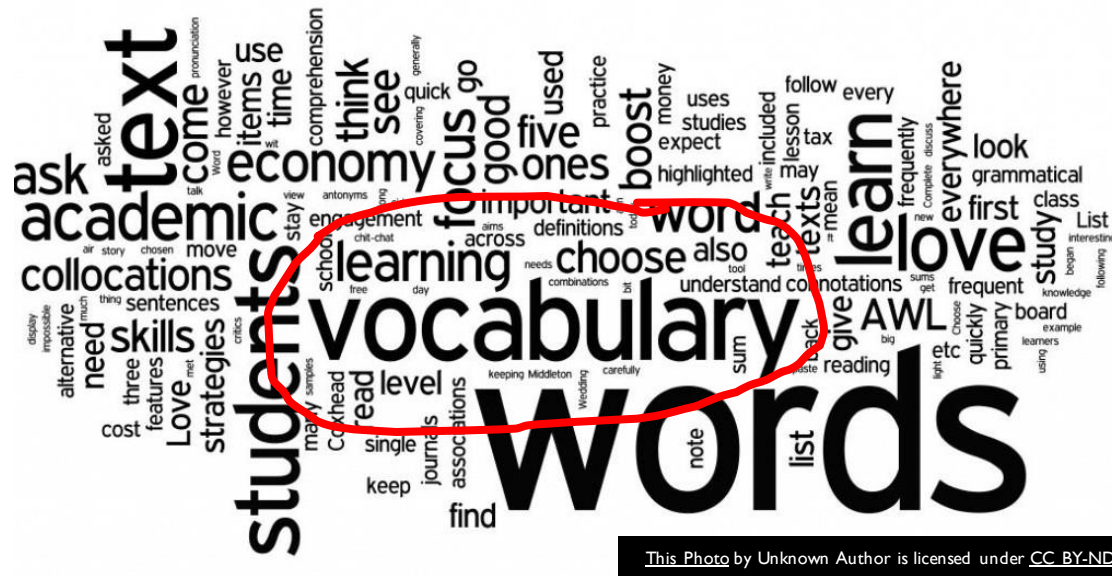


Digital Scholarship Center (DSC) of Temple University Libraries



MRC researchers are pursuing automatic methods of providing subject access for the Encyclopedia Britannica entries

Exploration using the 1910 version of LCSH in transforming the keywords of the 19th Century Encyclopedia Britannica to a controlled vocabulary



STATEMENT OF THE PROBLEM

- Vocabulary divide - terms found in the 19th C. KP are not represented in the current version of the LCSH (2016) due to deprecation
- LCSH updates represent new technical, scientific, and sociological ideas that are not recorded in the historical versions of the Encyclopedia.

Conceptual Model for the Encyclopedia Britannica Controlled Vocabulary



Automatic
Indexing



Subject
Cataloging



Articles of 19th
Century Encyclopedia
Britannica



Encyclopedia
Britannica
Controlled
Vocabulary

SKOS of the 1910 LCHS for the Transformation of Keywords
to Controlled Vocabulary of the 19th Century Encyclopedia
Britannica

Abacus.

QA135

Abandoned children. *See* Children—*Charities, protection, etc.*;
Foundlings; Orphans and orphan-asylums.

Abattoirs. *See* Slaughtering and slaughter-houses.

Abbeys.

NA4800-6113 (Architecture)

See also Cathedrals; Convents; Monasteries.

Abbots.

Abbreviations.

Z111 (Paleography)

See also Cipher; Shorthand.

Abbreviations, English, [French, Hebrew, etc.]

RESULTS AND FINDINGS

DOCX AND TEI FORMAT OF
THE 1910 LIBRARY OF
CONGRESS SUBJECT
HEADINGS

```

<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#"
>

  <skos:Concept rdf:about="http://mysite.com/#Abacus"><prefLabel>Abacus</prefLabel>
  <skos:notes>      QA135
  </skos:notes>
  </skos:Concept>

  <skos:Concept rdf:about="http://mysite.com/#Abandoned children"><prefLabel>Abandoned children</prefLabel>
  <USE>Abandoned children.</USE>
  </skos:Concept>

  <skos:Concept rdf:about="http://mysite.com/#Abattoirs"><prefLabel>Abattoirs</prefLabel>
  <USE>Abattoirs.</USE>
  </skos:Concept>

  <skos:Concept rdf:about="http://mysite.com/#Abbeys"><prefLabel>Abbeys</prefLabel>
  <skos:notes>      NA4800-6113 (Architecture)
  </skos:notes>
  <RT>Cathedrals ; Convents ; Monasteries.</RT>
  </skos:Concept>

  <skos:Concept rdf:about="http://mysite.com/#Abbots"><prefLabel>Abbots</prefLabel>
  </skos:Concept>

  <skos:Concept rdf:about="http://mysite.com/#Abbreviations"><prefLabel>Abbreviations</prefLabel>
  <skos:notes>      Z111 (Paleography)
  </skos:notes>
  <RT>Cipher ; Shorthand.</RT>
  </skos:Concept>

  <skos:Concept rdf:about="http://mysite.com/#Abbreviations, English, [French, Hebrew, etc]><prefLabel>
Abbreviations, English, [French, Hebrew, etc</prefLabel>
  </skos:Concept>

  <skos:Concept rdf:about="http://mysite.com/#Abdomen"><prefLabel>Abdomen</prefLabel>
  <skos:notes>      QM543 (Regional anatomy)
  </skos:notes>
  <RT>Groin; Intestines; Kidneys; Liver; Peritoneum; Stomach.</RT>
  </skos:Concept>

  <skos:Concept rdf:about="http://mysite.com/#Abdomen-Diseases"><prefLabel>Abdomen-Diseases</prefLabel>
  <skos:notes>      RC941
  </skos:notes>
  </skos:Concept>

```

SKOS OF THE 1910 LIBRARY OF CONGRESS SUBJECT HEADINGS

RDF / XML

MACHINE-READABLE FORMAT



Helping Interdisciplinary
Vocabulary Engineering



Vocabularies Search Index

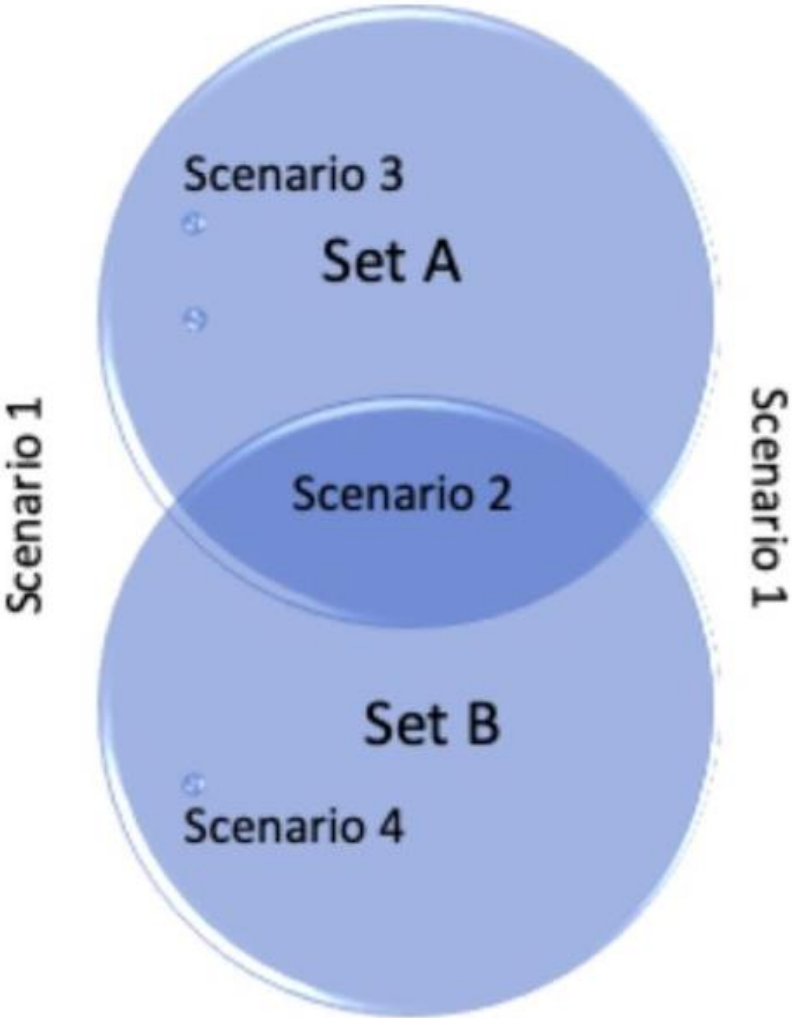
Vocabulary	Short name	Concepts	Last Updated
Asthma Ontology	Asthma	289	03/02/2016
Cardiology	Cardiology	155	07/25/2018
Combined Consumer Health Vocabulary	CCHV	210153	08/08/2019
Diabetes Mellitus Diagnostic Ontology	Diabetes	6439	12/20/2015
Ephraim Chambers Cyclopaedia	Chambers	57	05/05/2019
Food and Agriculture Organization	AGROVOC	35542	08/17/2018
Gastroenterology	Gastroenterology	112	07/25/2018
Library of Congress Subject Heading 1910	LCSH1910	23707	08/29/2019
Library of Congress Subject Headings	LCSH	421572	07/26/2018
Medical Subject Headings	MeSH	377824	08/25/2018
Metals	Metals	44	01/01/2016
OCHV	OCHV	87879	07/09/2019
Oncology	Oncology	132	07/25/2018
Pediatrics	Pediatrics	450	06/18/2018
Radiation Oncology	ROO	1183	07/07/2015
Radiology Lexicon	RADLEX	45471	11/16/2016
Respiratory	Respiratory	142	07/25/2018
Smart Appliances REference ontology	SAREF	112	02/10/2015
US Geological Survey	USGS	968	01/01/2016
Unified Astronomy Thesaurus	UAT	1843	01/31/2017

1910 LIBRARY OF
CONGRESS
SUBJECT
HEADING IN HIVE

DATA SETS SPECIFICATIONS:

SET A: THE 19TH CENTURY
ENCYCLOPEDIA BRITANNICA ARTICLES
ARE INDEXED USING THE 1910 LCHS

SET B: THE 19TH CENTURY
ENCYCLOPEDIA BRITANNICA ARTICLES
ARE INDEXED USING 2016 LCHS



U
S
E
C
A
S
E
S

Scenario 1: Union of Set A and Set B ($A \cup B$) gives all the keywords found in the NCEB using both versions of LCHS. These keywords are all used in the transformation of the keywords to the controlled vocabulary of NCEB.

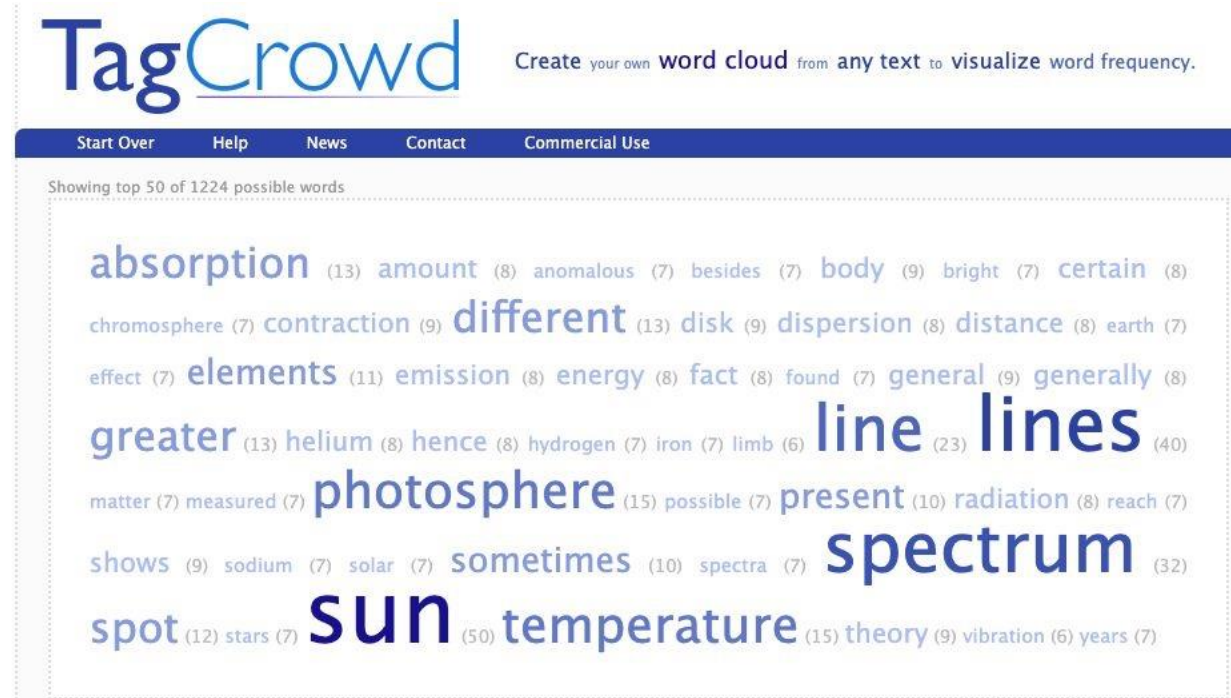
Scenario 2: Intersection of Set A and Set B ($A \cap B$) gives the keywords found which are common only to both versions of LCHS. This data represents the keywords that are still being used from the 19th century until 2016.

Scenario 3: Difference of Set A and Set B ($A - B$) gives the keywords found in the 1910 LCHS and could be deprecated already in the 2016 LCHS version or changed through time and has new terminology that replaces it.

Scenario 4: Difference of Set A and Set B ($B - A$) gives the keywords found in the 2016 LCHS and but could be non-existent yet in the 1910 LCHS version. This could also be the new keywords used in the later century.

SUBJECT ANALYSIS

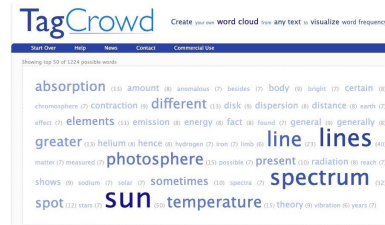
I. The text format of the article sun from the 19th Century Encyclopedia Britannica was subjected to text analysis using TagCrowd.



SUBJECT ANALYSIS

1. The text format of the article sun from the 19th Century Encyclopedia Britannica was subjected to text analysis using TagCrowd.

2. To apply the “Aboutness” approach to subject analysis, the following words were determined as the keywords. 16 out of 51 descriptors were selected as the keywords. They were spotted to reflect both the richness of the topic being discussed in the article and the expressiveness they contain in relation to the word “SUN”.



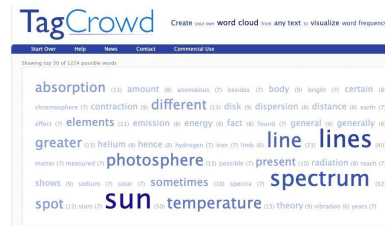
Article (19th Century Encyclopedia Britannica)	Tag Crowd	Word Frequency	Aboutness
Sun	Sun	50	Sun
	Lines	40	Lines
	Spectrum	32	Spectrum
	Line	23	
	Photosphere	15	Photosphere
	Temperature	15	Temperature
	Absorption	13	Absorption
	Different	13	
	Greater	13	
	Spot	12	Spot
	Elements	11	Elements
	Present	10	
	Sometimes	10	
	Body	9	Body
	Contraction	9	Contraction
	Disk	9	Disk
	General	9	
	Shows	9	
	Theory	9	
	Amount	8	
	Certain	8	
	Dispersion	8	Dispersion
	Distance	8	Distance
	Emission	8	Emission
	Energy	8	Energy

SUBJECT ANALYSIS

1. The text format of the article sun from the NCEB was subjected to text analysis using TagCrowd.

2. To apply the “Aboutness” approach to subject analysis, the following words were determined as the keywords. 16 out of 51 descriptors were selected as the keywords. They were spotted to reflect both the richness of the topic being discussed in the article and the expressiveness they contain in relation to the word “SUN”.

3. For the purpose of simulation, manual indexing using 1910 LCSH (OCR version) was performed capturing the structure, subdivisions and subject headings for the index of the article “SUN”



Article (19th Century Encyclopedia Britannica)	Tag Crowd	Word Frequency	Aboutness
Sun	Sun	50	Sun
	Lines	40	Lines
	Spectrum	32	Spectrum
	Line	23	
	Photosphere	15	Photosphere
	Temperature	15	Temperature
	Absorption	13	Absorption
	Different	13	
	Greater	13	
	Spot	12	Spot
	Elements	11	Elements
	Present	10	
	Sometimes	10	
	Body	9	Body
	Contraction	9	Contraction
	Disk	9	Disk
	General	9	
	Shows	9	
	Theory	9	
	Amount	8	
	Certain	8	
	Dispersion	8	Dispersion
	Distance	8	Distance
	Emission	8	Emission
	Energy	8	Energy

Absorption (Physiology)

QP88

See also Digestion ; Fat ; Lymphatics ; Osmosis ; Skin.

Absorption, Atmospheric. *See* Solar radiation.

Absorption of light.

QC437

Absorption spectra.

QC437

See also Heat—Radiation and absorption; Spectrum analysis.

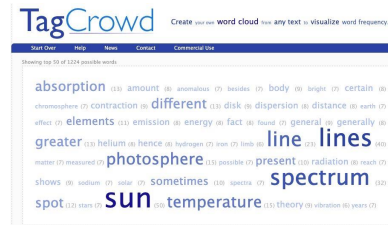
SUBJECT ANALYSIS

1. The text format of the article sun from the NCEB was subjected to text analysis using TagCrowd.

2. To apply the “Aboutness” approach to subject analysis, the following words were determined as the keywords. 16 out of 51 descriptors were selected as the keywords. They were spotted to reflect both the richness of the topic being discussed in the article and the expressiveness they contain in relation to the word “SUN”.

3. For the purpose of simulation, manual indexing using 1910 LCSH (OCR version) was performed capturing the structure, subdivisions and subject headings for the index of the article “SUN”

4. The same article was uploaded to HIVE for automatic indexing, generating the 2016 LCSH version of the keyword “Absorption”



Article (19th Century Encyclopedia Britannica)	Tag Crowd	Word Frequency	Aboutness
Sun	Sun	50	Sun
	Lines	40	Lines
	Spectrum	32	Spectrum
	Line	23	
	Photosphere	15	Photosphere
	Temperature	15	Temperature
	Absorption	13	Absorption
	Different	13	
	Greater	13	
	Spot	12	Spot
	Elements	11	Elements
	Present	10	
	Sometimes	10	
	Body	9	Body
	Contraction	9	Contraction
	Disk	9	Disk
	General	9	
	Shows	9	
	Theory	9	
	Amount	8	
	Certain	8	
	Dispersion	8	Dispersion
	Distance	8	Distance
	Emission	8	Emission
	Energy	8	Energy

Absorption (Physiology)

QP88

See also Digestion ; Fat ; Lymphatics ; Osmosis ; Skin.

Absorption, Atmospheric. *See* Solar radiation.

Absorption of light.

QC437

Absorption spectra.

QC437

See also Heat—Radiation and absorption ; Spectrum analysis.



Helping Interdisciplinary
Vocabulary Engineering



DREXEL UNIVERSITY
Metadata
Research Center
College of Computing & Information

Vocabularies Search Index

HIVE automatically extracts concepts from a file, or URL, using selected vocabularies.

1 Select vocabularies

- ☐ AGROVOC
- ☐ Diabetes
- ☐ Metals
- ☐ ROO
- ☐ USGS
- ☐ Asthma
- ☐ Gastroenterology
- ☐ Oncology
- ☐ Respiratory
- ☐ Cardiology
- ☒ LCSH
- ☐ Pediatrics
- ☐ SAREF
- ☐ Chambers
- ☐ MeSH
- ☐ RADLEX
- ☐ UAT

2 Enter a URL, or select a file, to index Indexing...

URL

or

3 Select indexing filters (optional)

LCSH

Sun

Temperature Temperature Contracts Contraction Sodium Founding Bodies Helium Case

Cases Cases Case Iron Hydrogenation Hydrogen Measurement Measurement Distances Calcium

Leveling Levellers Atmosphere Atmospherics Absorption Stars Red Dues Regionalism Regions Waves Light

Lighting Lighting Titanium Losses Heat Heating Heating Sources Sources Scandium Vibration Vibrators Vibration

Inference Magnesium Chromium Oxygen Oxygenators

 Cloud View

 Rank Order

 List View

 Alpha Order

LCSH

Sun Temperature

Temperature Contracts

Contraction Sodium Founding

Bodies Helium Case Cases

Cases Case Iron

Hydrogenation Hydrogen

Measurement Measurement

Distances Calcium Leveling

Levellers Atmosphere Atmospherics

Absorption Stars Red Dues

Regionalism Regions Waves Light

Lighting Lighting Titanium Losses

List

JSON-LD

SKOS RDF/XML

Dublin Core

XML

Preferred label Absorption

URI <http://id.loc.gov/authorities/subjects/sh85000245>

Alternate label Sorption;

Notes label Not provided

Broader

Chemistry, Physical and theoretical
Packed towers

Narrower

Gases--Absorption and adsorption
Light absorption
Photoabsorption
Sorbents

Related No related concepts

RESULTS FROM THE
2016 LIBRARY OF CONGRESS SUBJECT
HEADING IN HIVE

Article (19th Century Encyclopedia Britannica)	Tag Crowd	Word Frequency	Aboutness	1910 LCHS Manual Subject Indexing	Output Controlled Vocabulary	LCHS in Hive	Scenario
Sun	Sun	50	Sun			Preferred label: Absorption URI http://id.loc.gov/authorities/subjects/sh85000245 Alternate label Sorption Notes label Not provided Broader Chemistry, Physical and theoretical Packed towers Narrower Gases--Absorption and adsorption Light absorption Photoabsorption Sorbents Related No related concepts	
	Absorption	13	Absorption	*Absorption (Physiology) *Absorption, Atmospheric See Solar -- Radiation Absorption of light *Absorption spectra See also Heat -- Radiation a	Absorption of light		Scenario 2

WORD ANALYSIS MATRIX

Table above is a word analysis matrix for the descriptor “Absorption” that shows the result of the subject analysis conducted in the article “SUN”. The simulation of the word – Absorption fell in scenario 2 of the use cases. This means that the word “Absorption” intersects both data sets, thus it exists from 1910 till 2016.

Article (19th Century Encyclopedia Britannica)	Tag Crowd	Word Frequency	Aboutness	1910 LCSH Manual Subject Indexing	Output Controlled Vocabulary	LCSH in Hive	Scenario
Sun	Sun	50	Sun				
	Absorption	13	Absorption	*Absorption (Physiology) *Absorption, Atmospheric See Solar -- Radiation Absorption of light *Absorption spectra See also Heat -- Radiation a	Absorption of light	Preferred label: Absorption URI http://id.loc.gov/authorities/subjects/sh85000245 Alternate label Sorption Notes label Not provided Broader Chemistry, Physical and theoretical Packed towers Narrower Gases--Absorption and adsorption Light absorption Photoabsorption Sorbents Related No related concepts	Scenario 2

Cloud View
List View

Rank Order
Alpha Order

LCSH

Sun Temperature

Temperature Contracts

Contraction Sodium Founding

Bodies Helium Case Cases

Cases Case Iron

Hydrogenation Hydrogen

Measurement Measurement

Distances Calcium Leveling

Levellers Atmosphere Atmospheric

Absorption Stars Red Dues

Regionalism Regions Waves Light

Lighting Lighting Titanium Losses

Heat Heating Heating Sources

Sources Scandium Vibration

Vibrators Vibration Inference

Magnesium Chromium Oxygen

Oxygenators

List
JSON-LD
SKOS RDF/XML
Dublin Core
XML

Preferred label Absorption

URI <http://id.loc.gov/authorities/subjects/sh85000245>

Alternate label Sorption;

Notes label Not provided

Broader

Chemistry, Physical and theoretical
Packed towers

Narrower

Gases--Absorption and adsorption
Light absorption
Photoabsorption
Sorbents

Related No related concepts

WORD ANALYSIS MATRIX

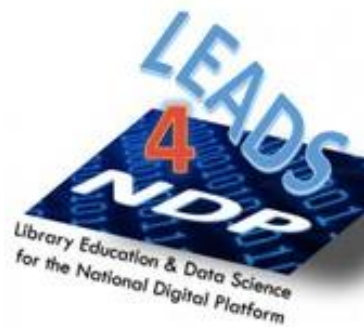
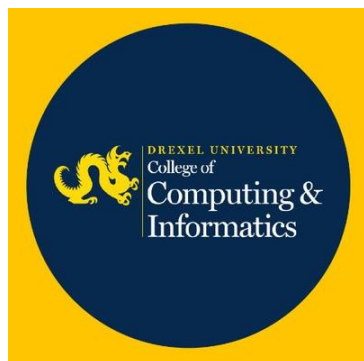
Table above is a word analysis matrix for the descriptor “Absorption” that shows the result of the subject analysis conducted in the article “SUN”. The simulation of the word – Absorption fell in scenario 2 of the use cases. This means that the word “Absorption” intersects both data sets, thus it exists from 1910 till 2016.

CONCLUSIONS

The use of both versions of the LCSH, 1910 and 2016, as a controlled vocabulary to extract keywords from the articles of 19TH Century Encyclopedia Britannica purposively rendered a wider consideration in the establishment of the 19TH Century Encyclopedia Britannica controlled vocabulary. Use cases showed the inclusive generation of keywords producing substantial vocabularies.

Scenarios put forward the impositions of different inferences of the data.

- Scenario 1 was the integration of all descriptors that matched to the vocabulary of the 2 versions of LCSH.
- Scenario 2 captured all the descriptors that commonly matched with both versions of LCSH. This meant that these words were used in the 19th century until 2016.
- Scenario 3 represented the words that were not existent anymore in the present day since it's only found as a vocabulary in the 1910 LCSH.
- On the other hand, scenario 4 is telling us that there were new terms formulated by men in the next centuries that came. Thus using the old LCSH or the appropriate version of LCSH would yield higher percent of accuracy in terms of completeness of 19th Century Encyclopedia Britannica keywords to be transformed to a controlled vocabulary. The goal is to take simulations of all scenarios to be able to justify the hypothesis and to conduct manual indexing using the current LCSH (2016) which serve another use case.



REFERENCES

- Encyclopedia. (2019, June 14). Retrieved June 16, 2019, from <https://en.wikipedia.org/wiki/Encyclopedia> HIVE2. (n.d.). Retrieved June 16, 2019, from <http://hive2.cci.drexel.edu:8080/> Greenberg, J., Losee, R., Pérez Agüera, J. R., Scherle, R., White, H., & Willis, C. (2011). HIVE:
- Helping Interdisciplinary Vocabulary Engineering. Bulletin of the American Society for Information Science and Technology, 37(4): 23-26.
- Library of Congress Subject Headings. (2019, January 27). Retrieved June 16, 2019, from https://en.wikipedia.org/wiki/Library_of_Congress_Subject_Headings
- Library of Congress Subject Headings. (n.d.). Retrieved June 16, 2019, from <https://www.loc.gov/aba/publications/FreeLCSH/freelcsh.html>
- Logan, P.M. Nineteenth-Century Knowledge Project. Retrieved June 16, 2019, from <https://tu-plogan.github.io/>
- Logan, P. M., Greenberg, J., & Grabus, S. (In Press). Knowledge Representation: Old, New, and Automated Indexing. In Proceedings of Digital Humanities Conference 2019, Utrecht, The Netherlands.

