Representing Aboutness: Automatically Indexing 19th-Century Encyclopedia Britannica Entries

Sam Grabus

Jane Greenberg

Peter Logan

Joan Boone

NEH grant #HAA-261228-18 & IMSL grant #RE- 70-17-0094-17

NATIONAL ENDOWMENT FOR THE Humanities

INSTITUTE of Museum and Library SERVICES

Aboutness:

Subjective & Objective

- **Subjective:** "Moby Dick is about more than just a whale" (Svenonius, 2000)

- **Objective:** a set of terms that can be agreed upon as useful for database Information Retrieval purposes. (Svenonius, 2000)

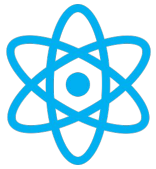
Still challenging!

Aboutness for Humanities Documents

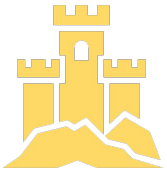
Scientific

- Linguistic Determinacy
- Domain-specific



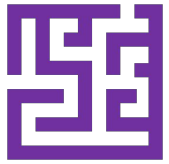




Humanities

- Linguistically Complex
- Multidisciplinary language
- Expressive





Controlled Vocabularies:

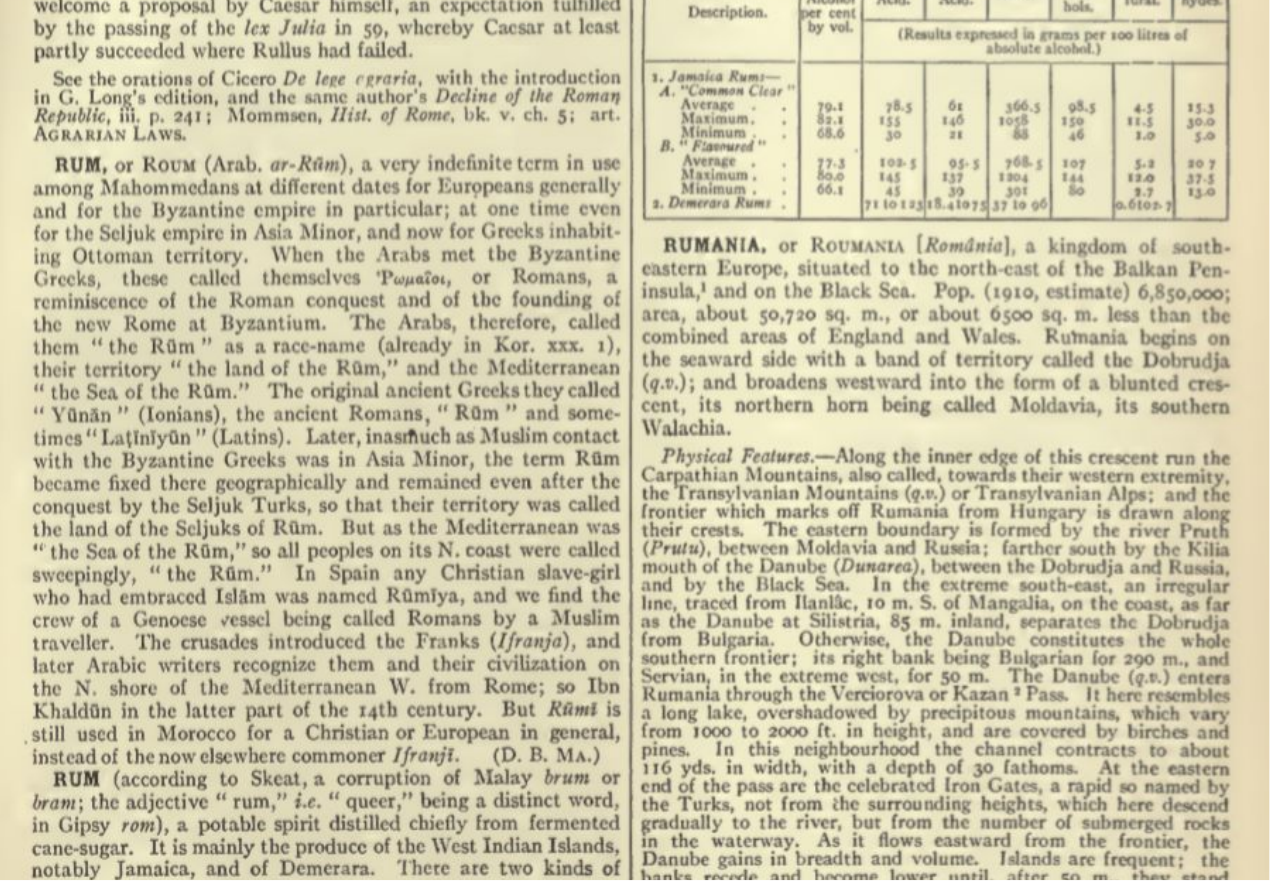
Helping to represent aboutness in a standardized way

- **Avoiding language idiosyncrasies:**
  - Regional word usage and spellings
  - Idioms
  - Abbreviations

  (Bair, S. & Carlson, S., 2008)

- **Text analysis and representation consistency**

  (Bueno-de-la-Fuente, G., Mateos R.D. & Greenberg, J., 2016)



welcome a proposal by Caesar himself, an expectation fulfilled by the passing of the *lex Julia* in 59, whereby Caesar at least partly succeeded where Rullus had failed.

See the orations of Cicero *De lege agraria*, with the introduction in G. Long's edition, and the same author's *Decline of the Roman Republic*, iii. p. 241; Mommsen, *Hist. of Rome*, bk. v. ch. 5; art. AGRARIAN LAWS.

RUM, or ROUM (Arab. *ar-Rūm*), a very indefinite term in use among Mahommedans at different dates for Europeans generally and for the Byzantine empire in particular; at one time even for the Seljuk empire in Asia Minor, and now for Greeks inhabiting Ottoman territory. When the Arabs met the Byzantine Greeks, these called themselves Ῥωμαῖοι, or Romans, a reminiscence of the Roman conquest and of the founding of the new Rome at Byzantium. The Arabs, therefore, called them "the Rūm" as a race-name (already in Kor. xxx. 1), their territory "the land of the Rūm," and the Mediterranean "the Sea of the Rūm." The original ancient Greeks they called "Yūnān" (Ionians), the ancient Romans, "Rūm" and sometimes "Laṭīnīyūn" (Latins). Later, inasmuch as Muslim contact with the Byzantine Greeks was in Asia Minor, the term Rūm became fixed there geographically and remained even after the conquest by the Seljuk Turks, so that their territory was called the land of the Seljuks of Rūm. But as the Mediterranean was "the Sea of the Rūm," so all peoples on its N. coast were called sweepingly, "the Rūm." In Spain any Christian slave-girl who had embraced Islām was named Rūmīya, and we find the crew of a Genoese vessel being called Romans by a Muslim traveller. The crusades introduced the Franks (*Ifranja*), and later Arabic writers recognize them and their civilization on the N. shore of the Mediterranean W. from Rome; so Ibn Khaldūn in the latter part of the 14th century. But *Rūmī* is still used in Morocco for a Christian or European in general, instead of the now elsewhere commoner *Ifranjī*. (D. B. MA.)

RUM (according to Skeat, a corruption of Malay *brum* or *bram*; the adjective "rum," *i.e.* "queer," being a distinct word, in Gipsy *rom*), a potable spirit distilled chiefly from fermented cane-sugar. It is mainly the produce of the West Indian Islands, notably Jamaica, and of Demerara. There are two kinds of

Description.	Alcohol per cent by vol.	Acid.	Acid.	hols.	Total.	Ethers.
		(Results expressed in grams per 100 litres of absolute alcohol.)				
1. *Jamaica Rums*—						
A. "Common Clear"						
Average . .	79.1	78.5	61	366.5	98.5	4.5	15.3
Maximum . .	82.1	155	146	1058	150	11.5	30.0
Minimum . .	68.6	30	21	88	46	1.0	5.0
B. "Flavoured"							
Average . .	77.3	102.5	95.5	768.5	107	5.2	20.7
Maximum . .	80.0	145	137	1204	144	12.0	37.5
Minimum . .	66.1	45	39	301	80	2.7	13.0
2. *Demerara Rums* .		71 to 123	18.41075	37 to 96		0.6 to 2.7	

RUMANIA, or ROUMANIA [*România*], a kingdom of south-eastern Europe, situated to the north-east of the Balkan Peninsula,[1] and on the Black Sea. Pop. (1910, estimate) 6,850,000; area, about 50,720 sq. m., or about 6500 sq. m. less than the combined areas of England and Wales. Rumania begins on the seaward side with a band of territory called the Dobrudja (*q.v.*); and broadens westward into the form of a blunted crescent, its northern horn being called Moldavia, its southern Walachia.

Physical Features.—Along the inner edge of this crescent run the Carpathian Mountains, also called, towards their western extremity, the Transylvanian Mountains (*q.v.*) or Transylvanian Alps; and the frontier which marks off Rumania from Hungary is drawn along their crests. The eastern boundary is formed by the river Pruth (*Prutu*), between Moldavia and Russia; farther south by the Kilia mouth of the Danube (*Dunarea*), between the Dobrudja and Russia, and by the Black Sea. In the extreme south-east, an irregular line, traced from Ilanlâc, 10 m. S. of Mangalia, on the coast, as far as the Danube at Silistria, 85 m. inland, separates the Dobrudja from Bulgaria. Otherwise, the Danube constitutes the whole southern frontier; its right bank being Bulgarian for 290 m., and Servian, in the extreme west, for 50 m. The Danube (*q.v.*) enters Rumania through the Verciorova or Kazan[2] Pass. It here resembles a long lake, overshadowed by precipitous mountains, which vary from 1000 to 2000 ft. in height, and are covered by birches and pines. In this neighbourhood the channel contracts to about 116 yds. in width, with a depth of 30 fathoms. At the eastern end of the pass are the celebrated Iron Gates, a rapid so named by the Turks, not from the surrounding heights, which here descend gradually to the river, but from the number of submerged rocks in the waterway. As it flows eastward from the frontier, the Danube gains in breadth and volume. Islands are frequent; the banks recede and become lower until, after 50 m., they stand


```
http://www.tei-c.org/ns/1.0" xml:id="eb11-23-r05-0825-02" facs="encyclopaediabri23chisrich_0882.jp2" type="entry">

<p><label>RUM</label>
(according to Skeat, a corruption of Malay <hi rend="italic">brum</hi> or <hi rend="italic">branr,</hi> the adjective
“rum,” <hi rend="italic">i.e.</hi> “queer,” being a distinct word, in Gipsy <hi rend="italic">ram),</hi> a potable
spirit distilled chiefly from fermented cane-sugar. It is mainly the produce of the West Indian Islands, notably
Jamaica, and of Demerara. !’here are two kinds of Jamaica rum, namely, “common” or “clean” rum, and , “flavoured” or
“German” rum. The latter is used almost entirely for purposes of blending with lighter types of spirit. Compared with
other potable spirits such as whisky and brandy, the Jamaica rums are distinguished by their very high proportion of
secondary products, particularly of the compound esters. Among the latter butyric “ether” (ethyl butyrate)
predominates. The Demerara rums are of a lighter character. Rum has a deep brown colour imparted by caramel or by
storage in sherry· casks, or, most generally, by both. “Tafia” is an inferior quality of rum produced in the French
colonies. “Negro” rum, which is the lowest quality of all, and into the wash for which the <hi rend="italic">debris</hi
> of the sugar-cane enters, is consumed locally by the coloured workers. The spirit prepared from beet-sugar molasses
cannot be regarded as rum, for, unless it is highly rectified, it possesses a disagreeable\&gt;dour’and taste.
Fictitious rum is, however, sometimes prepared from highly rectified beet spirit and rum “essence”-a mixture of
artificial esters (ethyl butyrate, &amp;c.L birch bark oil and so on. Highly rectified</p><p>beet spirit is also
occasionally used for blending with genuine rum, particularly with the “flavoured” or “German” rum. The latter name
originated in the fact that this kind of rum was exported very largely to Gcrmany for the purpose of blending. The
general composition of various kinds of rum is manifest from the annexed table. The consumption of rum in the United
Kingdom has fallen off considerably of late years, concurrently with the general tendency of the public towards
lighter and “drier” alcoholic beverages (see Spirits).</p><table> <row> <cell cols="8" rows="1"><p><hi rend="smallcaps"
>Composition of Different Varieties of Rum (Analyses by W. Collingwood Williams; cf.</hi> <hi rend="italic">J. Soc.
Chem. Ind.</hi></p><p>1907, P. 498.)</p></cell> </row> <row> <cell cols="1" rows="2"><p>Description.</p></cell> <cell
cols="1" rows="2"><p>Alcohol per cent by vol.</p></cell> <cell cols="1" rows="1"><p>Total</p><p>Acid.</p></cell> <cell
cols="1" rows="1"><p>Volaine</p><p>Add.</p></cell> <cell cols="1" rows="1"><p>Esters.</p></cell> <cell cols="1" rows="1
"><p>Higher</p><p>Alco</p><p>hols.</p></cell> <cell cols="1" rows="1"><p>Fur</p><p>fural.</p></cell> <cell cols="1"
rows="1"><p>Alde</p><p>hydes.</p></cell> </row> <row> <cell cols="6" rows="1"><p>(Results expressed in grams per ioo
litres of absolute alcohol.)</p></cell> </row> <row> <cell cols="1" rows="1"><p>1. <hi rend="italic">Jamaica Rums—</hi>
</p><p><hi rend="italic">A.</hi> <hi rend="italic">"Common Clear”</hi> Average . Maximum. Minimum .</p><p><hi rend="
italic">B. “Flavoured”</hi> Average . Maximum . Minimum .</p><p>2. <hi rend="italic">Demerara Rums .</hi></p></cell> <
cell cols="1" rows="1"><p>79.1</p><p>52.1 68.6</p><p>773</p><p>80.0</p><p>66.1</p></cell> <cell cols="1" rows="1"><p>
78.5</p><p>155</p><p/><p>102.5</p><p>145</p><p/><p>71 to 123</p></cell> <cell cols="1" rows="1"> <p/><p>146</p><p/><p>
95*5</p><p>137</p><p>«</p><p>18.41075</p></cell> <cell cols="1" rows="1"><p>366.5</p><p>Il</p><p>768. S</p><p>1204</p><
p>391</p><p>37 to 96</p></cell> <cell cols="1" rows="1"><p>98.5</p><p>150</p><p/><p>107</p><p>144</p><p/> </cell> <cell
 cols="1" rows="1"><p>45</p><p>11.5</p><p>Ï.O</p><p>5-2</p><p>12.0</p><p>2.7</p><p>0.6 to 2. 7</p></cell> <cell cols="1
" rows="1"><p>15.3</p><p>30.0</p><p>5.0</p><p>20 7 375 130</p></cell> </row> <row> <cell cols="8" rows="1"/> </row> </
table>
</div></body></text></TEI>
```


Digital Scans in PDF

To

Over 100,000 Individual TEI-XML entries

19th Century Knowledge Project

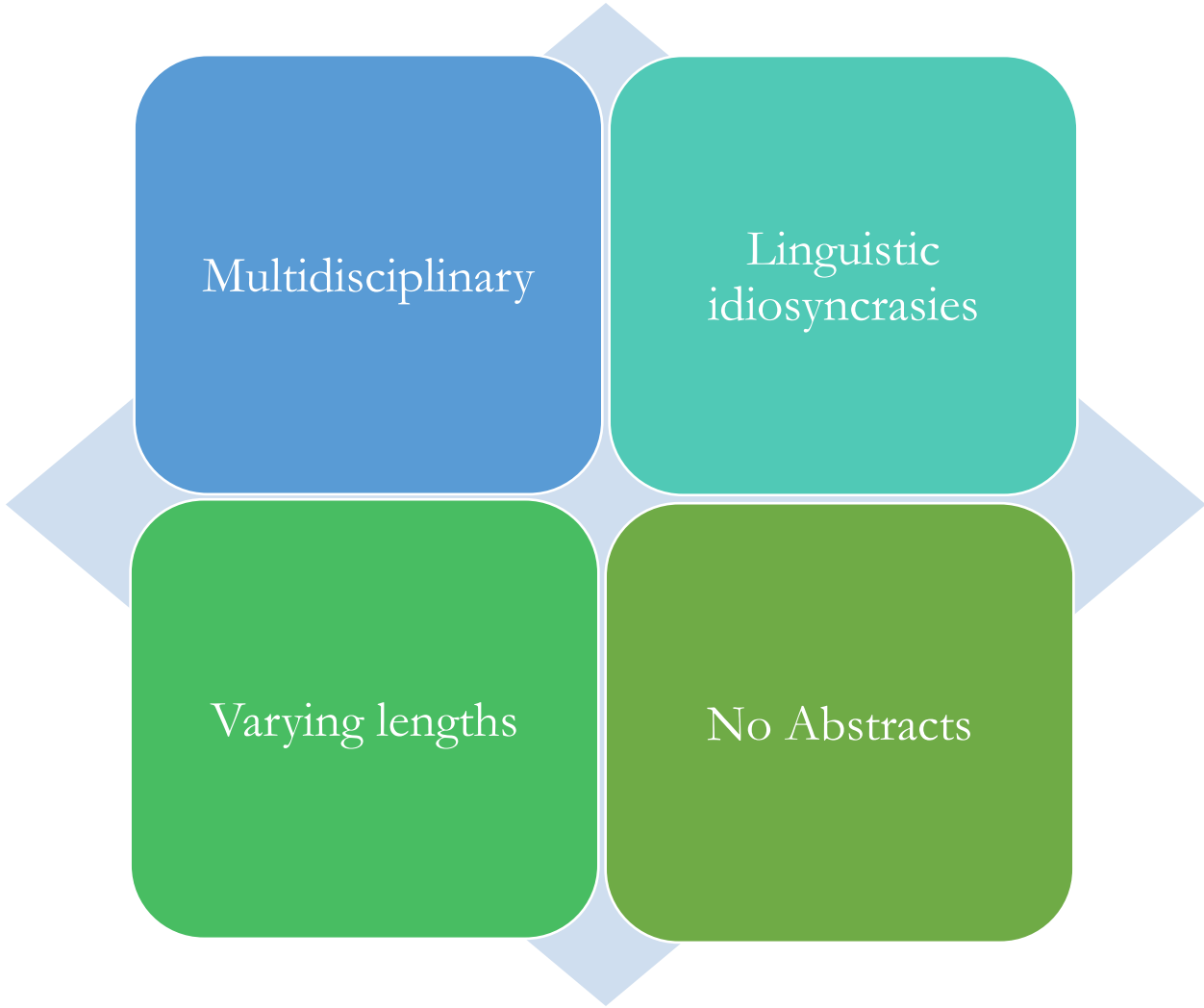
- 3rd ed., 18 vols., 1797
- 7th ed., 21 vols., 1842
- 9th ed., 25 vols., 1889
- 11th ed., 29 vols., 1911

Goals of the 19th-Century Knowledge Project

- **Long term question:** How does the specification of concepts change over time across four 19th-Century *Encyclopedia Britannicas* (1797-1911)?

- **Short term goal:** Automated descriptive subject metadata creation for integration into the individual encyclopedia entry TEI-XML headings

Challenges of Automatically Indexing this Corpus



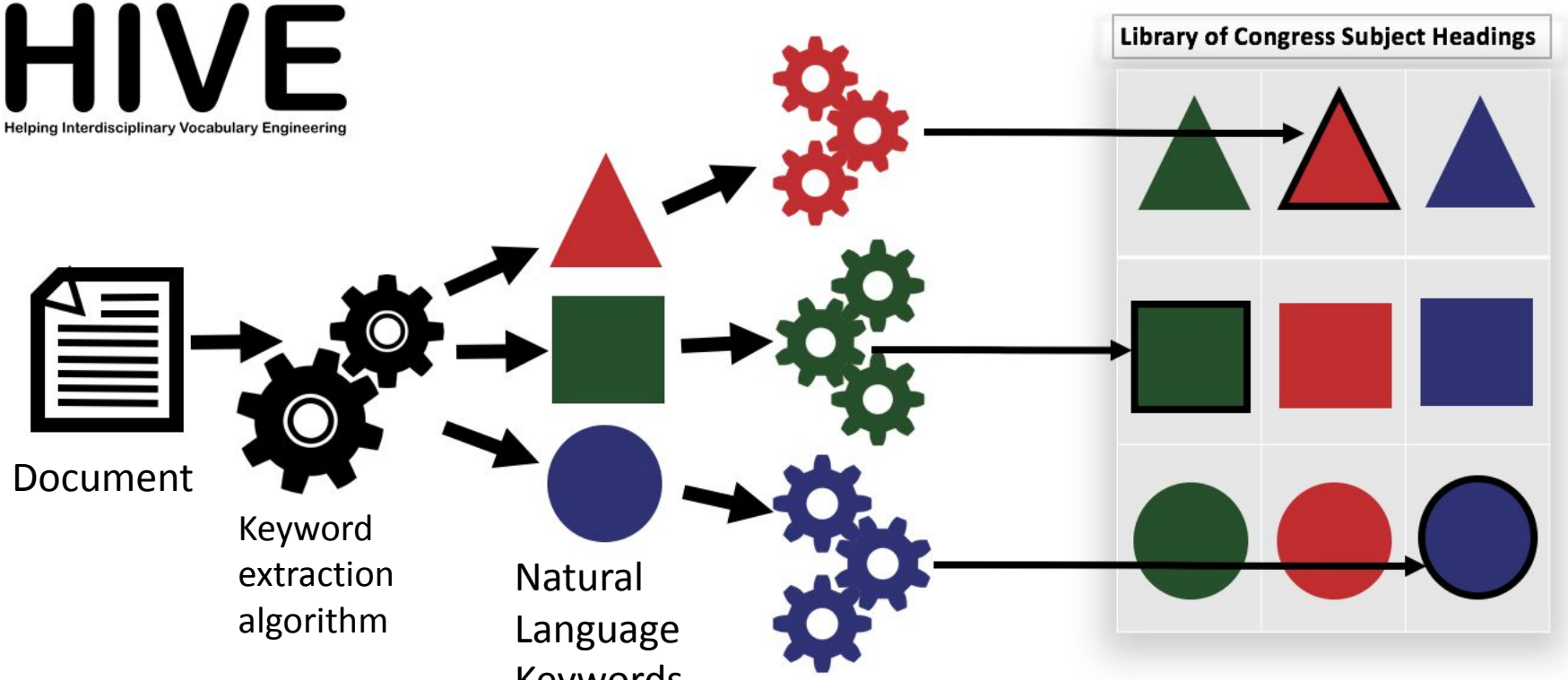
Our Goals

- Automated keyword extraction
- Transformation of keywords into controlled vocabulary terms
- Possibility of indexing with multiple controlled vocabularies

- **End Result: Large-scale automatic subject metadata generation with Controlled Vocabulary terms**

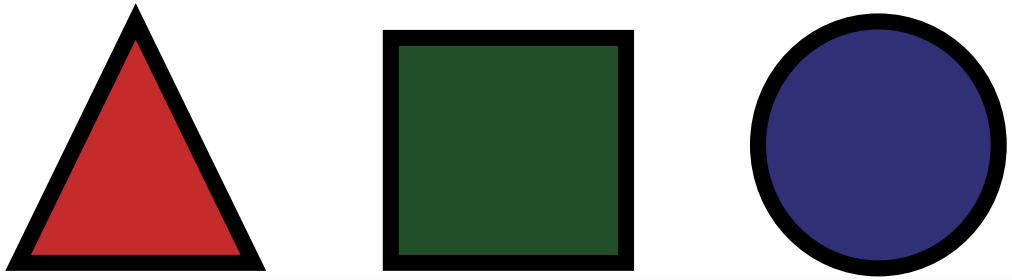
HIVE

Helping Interdisciplinary Vocabulary Engineering

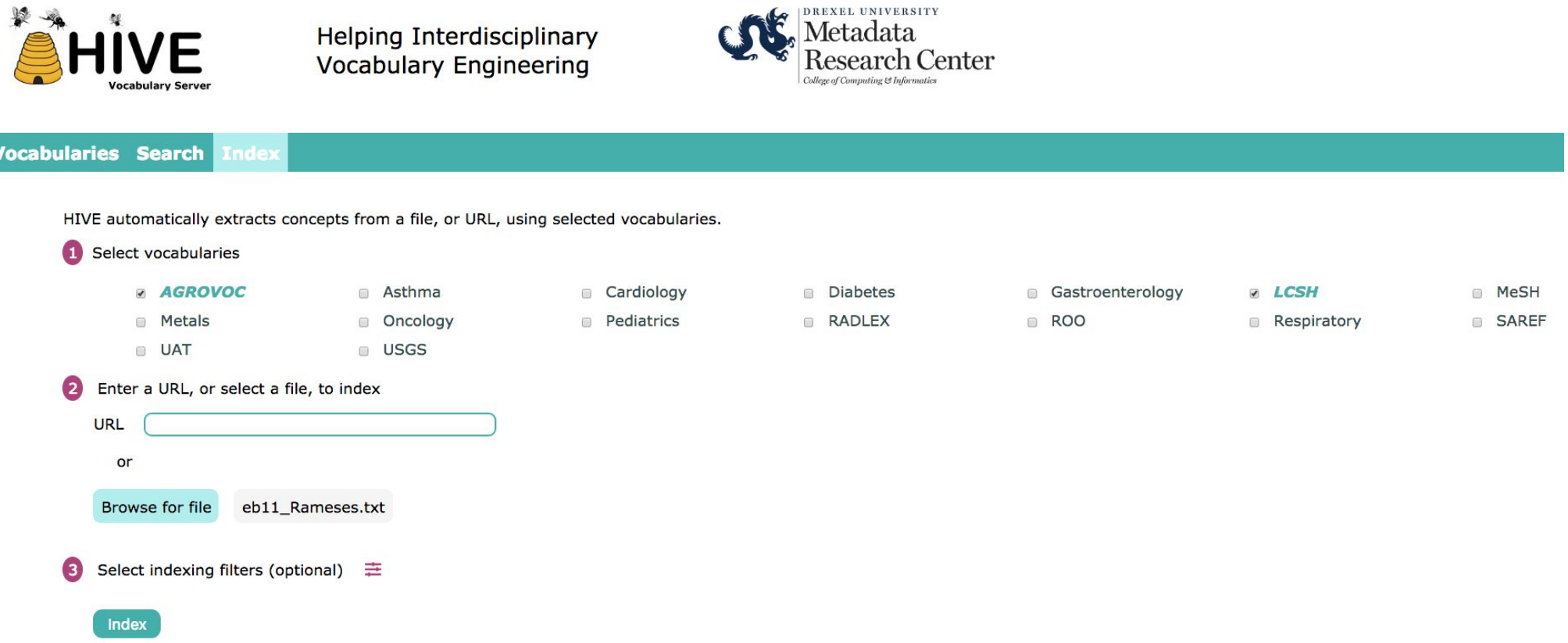


2nd Algorithm maps keywords to controlled vocabulary terms

HIVE Automatically Indexed Document Subject Headings:



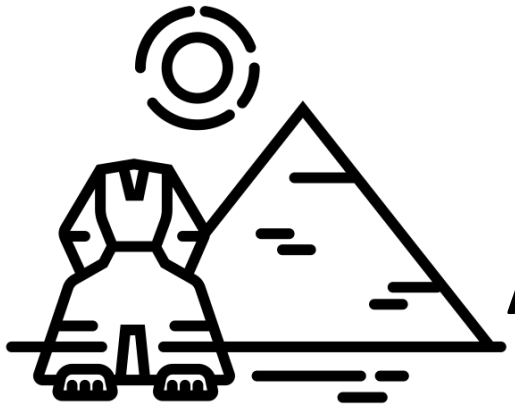
HIVE Indexing Example: 11th edition entry on Rameses (the city in Egypt)



http://hive2.cci.drexel.edu:8080/indexer

HIVE Indexing Example: 11th edition entry on Rameses

(the city in Egypt)



Automatically-generated LCSH and Agrovoc results

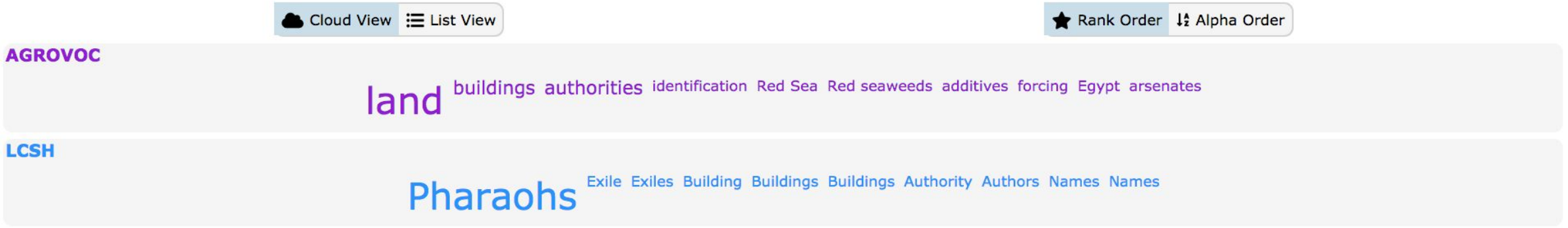


Image: Egypt by Laymik from the Noun Project

Three Keyword Extraction Algorithms

Which to use?

	KEA++	MAUI	RAKE
Features	• TFxIDF • First occurrence • Keyphraseness (keyphrase frequency) (Medelyan and Whitten 2008)	• TF-IDF • First occurrence • Keyphraseness • Length • Node degree • Wikipedia-based keyphraseness • Spread • Semantic relatedness • Inverse Wikipedia linkage (Medelyan, Perrone, and Witten 2010)	• Word frequency • Word degree • Ratio of degree to frequency • Co-occurrences • Stop words • Adjustable parameters: -Minimum characters per word -Maximum words per phrase -Minimum word frequency (Rose et al. 2010)
Machine-Learning	Yes	Yes	No

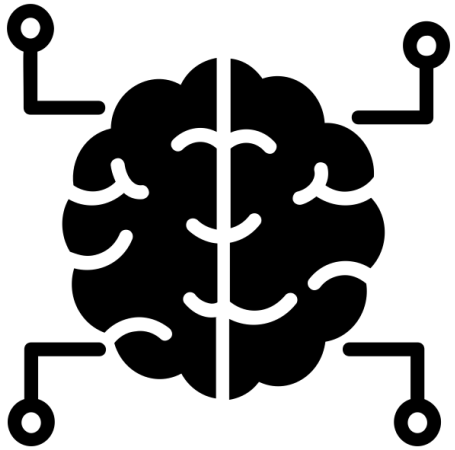


Image: Neural interface by ProSymbols from the Noun Project

Topic Relevance Evaluation

Two Questions:

- For each article in the preliminary sample, **what proportion of the 10 subject headings returned are relevant?**

- Since HIVE ranks the results according to relevancy, **what proportion of HIVE's *highest-ranked* results for each algorithm are relevant results?**

Steps

1. Compiling a sample for testing
2. Uploading entries to HIVE
3. Improving HIVE's interface
4. Re-testing and optimization
5. Topic Relevance Analysis

- Evaluator reads entries
- Ranks the relevance of the HIVE subject heading results

Relevance Measures

- Relevant (R)
- Partially-Relevant (PR)
- Non Relevant (NR)

Why Partial Relevance?

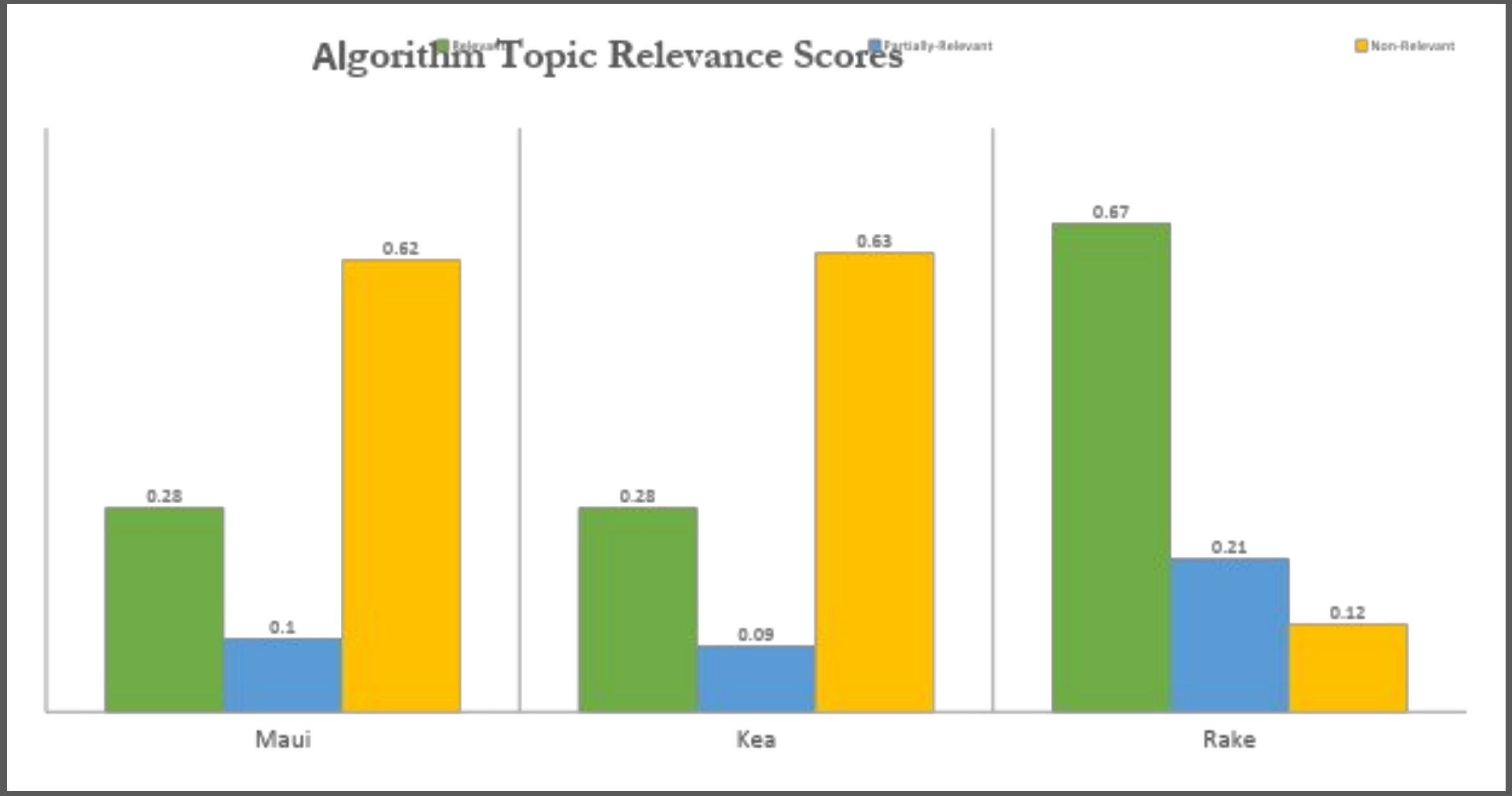
- **User uncertainty** regarding the information object's degree of relevance
- Degree of relevance in **relation to an information goal**

(Hjørland and Christensen 2002)

Topic relevance precision scores for three algorithms tested in HIVE

Precision score percentages are out of the 100 total subject headings for each algorithm

	Maui	Kea++	RAKE
Relevant	28%	28%	67%
Partially Relevant	10%	9%	21%
Non-Relevant	62%	63%	12%
HIVE top-ranked as true relevant	0%	100%	100%



TOPIC RELEVANCE PRECISION

KEYWORD EXTRACTION ALGORITHM

Types of Indexing Errors

INDEXING ERROR	EXAMPLE
1. Too many or too few results	• HIVE Version(s): 1 • Algorithm(s): Maui • Entry: 3rd edition, "Rhetoric" • Results: Zero results
2. Inappropriate levels of granularity (too broad)	• HIVE Version(s): Both • Algorithm(s): All • Entry: 11th edition, “Rameses” (the city)* • Broad Result: “Names”
3. The absence of essential subjects	• HIVE Version(s): 1 • Algorithm(s): Maui & Kea++ • Entry: 9th edition, “Rice” • Missing Result: “Rice”
4. Presence of obviously incorrect subject headings	• HIVE Version(s): 2 • Algorithm(s): Kea • Entry: 11th edition, “Rose” (the flower)* • LCSH Result: Heterosexual teachers
5. Different semantic meanings of a word	• HIVE Version(s): 1 • Algorithm(s): Maui • Entry: 11 edition, "Rum” (the liquor)* • Results: "Rummy (Game),” “Spirits (Islam)”
6. Time-Inappropriate Subject Heading	• HIVE Version(s): 2 • Algorithm: RAKE • Entry: 11th edition, “Rifle” • Result: “ZSU-23-4 (Antiaircraft gun)”

Table 3: Examples of common indexing errors as found in the research on this corpus
** Italics indicate comments added by researchers for clarity*

Identified by Lancaster 2003 & Golub et al. 2016

Additional errors identified working with this corpus

Next Steps:

"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**

- Comparative topic relevance testing for before and after correction of the historical Long S in the 3rd edition.
  - The **"long S"** used in the 3rd edition, which is indexed by HIVE as an f
  - *Example from the 3rd edition entry on Rum*
- Comparative topic relevance testing to refine RAKE's minimum word frequency parameter to accommodate for entries of varying lengths
- Integrating historical controlled vocabularies into HIVE
  - Can we avoid time-inappropriate subject headings and common homonyms?

References

- Bair, Sheila and Carlson, Sharon. 2008. “Where Keywords Fail : Using Metadata to Facilitate Digital Humanities Scholarship Where Keywords Fail : Using Metadata to Facilitate Digital Humanities Scholarship.” *Journal of Library Metadata*, *8*, no. 3: 249-262. https://doi.org/10.1080/19386380802398503
- Bueno-de-la-Fuente, Gema, Mateos R., David and Greenberg, Jane. 2016. "Chapter 10 - Automatic Text Indexing with SKOS Vocabularies in HIVE." In , 231-245: Elsevier Ltd.
- Golub, Koraljka, Soergel, Dagobert, Buchanan, George, Tudhope, Douglas, Lykke, Marianne and Hiom, Debra. 2016. “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: 3-16. https://doi.org/10.1002/asi
- Hjørland, Birger and Frank Sejer Christensen. 2002. "Work Tasks and socio-cognitive Relevance: A Specific Example." Journal of the American Society for Information Science and Technology 53, no. 11: 960-965.
- Lancaster, F. Wilfrid. 2003. Indexing and Abstracting in Theory and Practice. Third ed. London: Facet Pub.
- Medelyan, Olena, Vye Perrone, and Ian Witten. 2010. "Subject Metadata Support Powered by Maui."ACM. doi:10.1145/1816123.1816204.
- Medelyan, Olena and Ian H. Witten. 2008. "Domain-independent Automatic Keyphrase Indexing with Small Training Sets." Journal of the American Society for Information Science and Technology 59, no. 7: 1026-1040.
- Moens, Marie-Francine. 2000. Automatic Indexing and Abstracting of Document Texts. 1. Aufl. ed. Vol. 6.;6;. Boston: Kluwer Academic Publishers. doi:10.1007/b116177.
- Rose, Stuart, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. "Automatic Keyword Extraction from Individual Documents." In , 1-20. Chichester, UK: John Wiley & Sons, Ltd.
- Svenonius, Elaine. 2000. The Intellectual Foundation of Information Organization. Cambridge, Mass: MIT Press.

Questions?

Thank you!