#437

The HIVE Impact: Contributing to Consistency via Automatic Indexing

Hollie White[1,2], Craig Willis[2], Jane Greenberg[2]

[1] J. Michael Goodson Law Library, Duke University
[2] Metadata Research Center, School of Information and Library Science, University of North Carolina

Metadata Research Center <MRC>
UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

Summary:

The rapid and continuous growth of interdisciplinary digital collections is straining manual indexing processes, requiring more expedient means of indexing, often using multiple controlled vocabularies. Recent advances in automatic indexing are promising, but automatic indexing alone is insufficient. While automatic indexing improves efficiency, in practice indexing quality remains a central concern.

Indexing consistency and quality

- The degree to which two or more indexers describe the same resource using the same terms
- Related to statistical concepts of inter-judge or inter-rater agreement
- Typically used to evaluate consistency between human indexers
- Commonly used measures include Rolling's and Hooper's
- Recently, inter-indexer consistency measures have been used to evaluate automatic indexing techniques.
- ***Inter-indexer consistency alone is not a measure of quality.***

Can automatic indexing techniques, such as those supported by HIVE, be used to improve manual inter-indexer consistency?

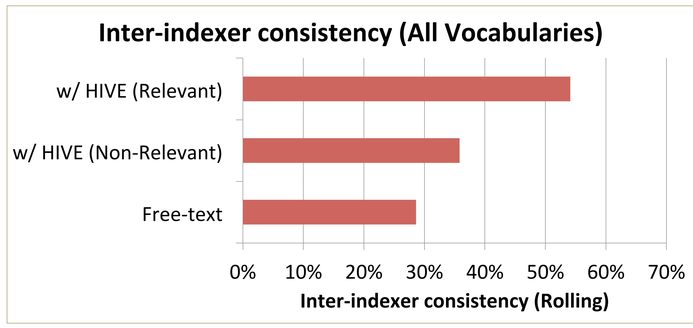
Methods:

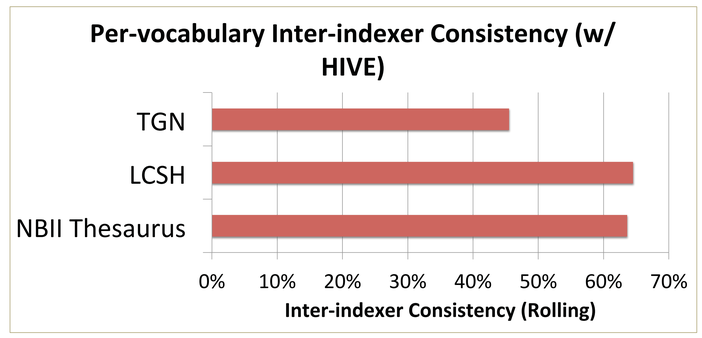
This study uses a within-subjects design using task scenario-based questionnaires to collect indexing terms for scientific recordsfrom the Dryad data repository based on two scenarios:

1. Assign free-text keywords without an indexing aid
2. Assign controlled terms from multiple vocabularies with an indexing aid. Identify terms from the automatic indexing aid regarded as relevant and not relevant to the document.

Participants:

- 31 participants recruited from HIVE workshops in January, March, and May 2011.
- Librarians, technologies, programmers, and other information professionals

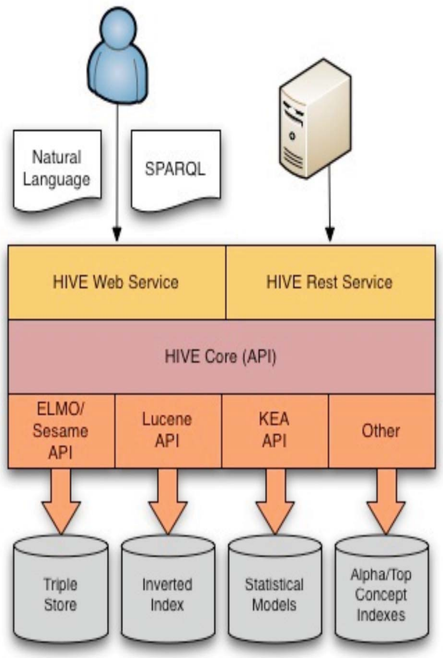




HIVE
Helping Interdisciplinary Vocabulary Engineering
http://hive.nescent.org

Helping Interdisciplinary Vocabulary Engineering (HIVE):

- IMLS-funded demonstration system
- Supports integration of multiple SKOS-based vocabularies



- Two primary functions:
  - 1) manual searching
  - 2) automatic indexing
- Open-source technologies (Lucene, Sesame)
- KEA++ algorithm for automatic indexing

KEA keyphrase extraction algorithm · openRDF.org Sesame · Lucene · Semantic Web

Partner vocabularies:

- The Getty Thesaurus of Geographic Names (TGN)
- Library of Congress Subject Headings (LCSH)
- NBII Biocomplexity Thesaurus

- Results suggest that HIVE is a successful indexing aid and that automatic indexing techniques may be used to improve indexing consistency.
- Inter-indexer consistency alone is not a measure of quality and needs to be considered in the context of other known measures for indexing, such as specificity and exhaustivity.
- Future work will focus on studying efficiency, cost, and quality.
- Information about the impact of automatic indexing techniques on manual indexing can be used to inform new workflows and procedures.
- Machine-aided approaches to indexing can reduce indexer burden while maintaining acceptable levels of quality as compared to fully automatic approaches.
- To have tools that contribute to consistency across different skill/experience levels can help the act of indexing in general.

HIVE is supported by IMLS grant LG-07-08-0120-08: INSTITUTE of Museum and Library SERVICES

In collaboration with: DRYAD · NESCent · NSF

With acknowledgements to Ryan Scherle, José Ramón Pérez Agüera, and Lina Huang for their work on the HIVE vocabulary server.

References:

Anderson, J.D., and Perez-Carballo, J. (2001a). "The nature of indexing: how humans and machines analyze messages and text for retrieval. Part I: Research, and the nature of human indexing." *Information Processing and Management* 37(2000), 231-254.
Anderson, J.D., and Perez-Carballo, J. (2001b). "The nature of indexing: how humans and machines analyze messages and text for retrieval. Part II: Machine indexing, and the allocation of human versus machine effort." *Information Processing and Management* 37(2001), 255-277.
Greenberg, J., Losee, R., Pérez Agüera, J.R., Scherle, R., White, H., and Willis, C. (2011). "HIVE: Helping Interdisciplinary Vocabulary Engineering." *Bulletin of the American Society for Information Science and Technology*, 37 (4): 23-26. Lancaster, F.W. (2003). *Indexing and Abstracting in Theory and Practice*. London: Facet.
Mann, T. (1997). "'Cataloging Must Change!' and Indexer Consistency Studies: Misreading the Evidence at Our Peril." *Cataloging and Classification Quarterly 23(3): 3-45.*
Medelyan, O. and Witten, I.H. (2008). "Domain independent automatic keyphrase indexing with small training sets." *Journal of American Society for Information Science and Technology*, 59(7)
Rolling, L. (1981). "Indexing Consistency, Quality, and Efficiency." *Information Processing and Management* 17: 69-76