View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Carolina Digital Repository

Lesley A. Skalla. Automatic subject indexing of dryad repository datasets: performance evaluation of HIVE and SmartHIVE. A Master's Paper for the M.S. in L.S. degree. April, 2011. 46 pages. Advisor: Jane Greenberg

HIVE is an automatic metadata application being considered for the Dryad data repository. The goals of this study were to 1) determine if HIVE performance can be improved by subject training and 2) determine HIVE's effectiveness in automatically generating controlled vocabulary terms that can be used to represent scientific data sets. HIVE performance was evaluated by 1) matching of HIVE-generated keywords with manually generated index terms (precision and recall) and 2) human evaluation of keyword relevancy. Part 1 of the study found that subject training HIVE had no effect on the number of "correct" keywords, precision, or recall. Part 2 found that despite large inter-evaluator inconsistency, the trend was for domain experts to assign a statistically similar number of keywords as "relevant" to both the data (x=4.0) and the articles (x=3.3) suggesting that HIVE may be useful in creating subject metadata for data sets from the full-text article.

Headings:

Automatic Indexing

Datasets

Metadata

Subject Cataloging

AUTOMATIC SUBJECT INDEXING OF DRYAD REPOSITORY DATASETS: PERFORMANCE EVALUATION OF HIVE AND SMARTHIVE

by
Lesley A. Skalla

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science.

Chapel Hill, North Carolina

April 2011

Approved by

_____

Jane Greenberg

Table of Contents

PREFACE

The master's paper research reported on in this document was conducted during the spring 2011 semester. A flaw in the document training for Kea++ was detected on April 3, a day before the final copy was due for submission, and as final edits were being made. It was determined that the training of HIVE via the 50 documents, as reported on in this paper, does not appear to have been fully processed. Given practical research constraints, the overall purpose of the master's paper, and the due date of April 4, my master's paper is being submitted with this noted flaw. This decision is supported by my advisor, Professor Jane Greenberg, noting that I fulfilled the requirements of the master's paper. It is my hope that I will be able to re-run the evaluation with the correct training, to more accurately assess the difference between regular HIVE and Smart HIVE. New results will be appended to this master's paper, or uploaded with the offprint for this work in the UNC repository.

INTRODUCTION

The digital age is having a profound effect on scientific research. Networked, collaborative, data-driven science has led to new ways of doing and communicating science (Wright, Sumner, Moore, & Koch, 2007). Data used to be collected, analyzed, published in research articles and then essentially forgotten. Data now takes a more central role as a valuable, citable, scholarly work whose life can be extended by re-use (Davis & Vickery, 2007). By encouraging the preservation and sharing of data sets, data can be available for verification, further analysis, or completely new research

opportunities within and across disciplines. Data that is archived in an appropriate public repository or data center will be preserved in a usable form for future use.

An excellent example of a public, international data repository is Dryad. Dryad was specifically developed for the preservation, discovery, and sharing of data underlying published research articles in the field of evolutionary biology and ecology (http://datadryad.org/) and has recently broadened to include data in the basic and applied biosciences. The repository was launched as the result of collaboration between the National Evolutionary Synthesis Center (NESCent) and the University of North Carolina Metadata Research Center, in coordination with a core group of societies and journals including *The American Naturalist, Evolution, the Journal of Evolutionary Biology, Molecular Ecology, and Heredity.* Starting in January 2011, authors publishing in these journals are required to deposit their data in an appropriate public repository such as Dryad upon publication of the research article (Rieseberg, Vines, & Kane, 2010; Whitlock, McPeek, Rausher, Rieseberg, & Moore, 2010). An important goal for Dryad is that the data submission process be easy, user-friendly, and place minimal burden on the author. To support this, Dryad partner journals provide Dryad with bibliographic information for each article (Vision, 2010). This bibliographic data in turn provides source metadata for each Dryad data package which can include one or more data files (Greenberg, 2009).

Another goal of Dryad is to allow end-users to perform specific searches of data by not only publication, but also by taxon, geography, geological age, biological concept, etc. Currently, the author-assigned keywords for the published article are used as subject metadata to describe the content of the data files. The Dryad curator will often enrich the

metadata by adding taxon name keywords. To further enhance data discoverability and reusability, keywords from controlled vocabularies will be added to the metadata records (Vision, 2010). Unfortunately, there is no one controlled vocabulary that will cover the wide range of interdisciplinary topics that represent research content in Dryad. In addition, controlled vocabularies have serious cost, interoperability, and usability constraints (HIVE, 2008). This in turn was the motivation for the creation of HIVE, or Helping Interdisciplinary Vocabulary Engineering.

HIVE is an automatic metadata generation approach that uses the Simple Knowledge Organization System (SKOS) to dynamically integrate discipline-specific controlled vocabularies. Dryad is considering utilizing HIVE to help support both authors and the repository curator in assigning subject metadata to the datasets. Candidate keywords extracted from document abstracts and potentially, the full-text of a document will be mapped to existing SKOS-encoded, discipline-specific controlled vocabularies resulting in a list of potential keywords that the author and/or Dryad curator can use to better represent the data set(s). In addition, HIVE can use machine learning techniques to be "trained" to better understand concept relationships. The integration of machine learning into HIVE is referred to as SmartHIVE.

Two early studies have provided some initial insight into HIVE system. Sherman (2010) demonstrated that the machine learning underlying HIVE produced higher quality results than basic term matching techniques used in NCBO's Bioportal and thus may potentially improve the quality of automatic indexing by HIVE. In addition, the standalone HIVE vocabulary server has been evaluated by a pilot usability testing study. Huang (2010) investigated what information professionals and scientists thought about

the pilot HIVE. Evaluative studies assessing the performance of HIVE within Dryad in automatically indexing data sets have yet to be conducted. Therefore, the research problem I will be addressing in this study is: To what extent can HIVE be used to generate controlled vocabulary terms that effectively describe the content of Dryad datasets?

To begin to approach this research problem, this study seeks to provide insight into three research questions:

Question one: How effective is HIVE for automatically generating controlled vocabulary terms from full-text articles?

Question two: How effective are HIVE-generated controlled vocabulary terms in describing the content of Dryad dataset(s)?

Question three: To what extent will training HIVE in a specific sub-discipline of evolutionary biology/ecology increase its ability to generate controlled vocabulary terms for that specific topic?

The literature review that follows intends to review current knowledge of the automatic indexing techniques utilized by HIVE for use in the Dryad digital data repository and is organized as follows. First, I will begin with a discussion about the metadata behind scientific datasets such as those in Dryad with an emphasis on subject descriptors. After reviewing the differences between keywords and subject descriptors, the main body of the review focuses on the automatic keyword generation techniques (KEA and KEA++) that power HIVE. Of particular importance for this study will be the ability of KEA++ to allow for sub-domain specific improvement in performance through training. Dryad's need for multiple vocabularies is then discussed. The review ends by

determining how the literature has assessed the quality of automatically-generated keywords.

LITERATURE REVIEW

Science Repository Datasets, Metadata, and Indexing

Data stored in scientific repositories can only be utilized if users can find and retrieve it. Quality descriptive metadata are essential for this process with subject metadata being particularly important for discovery and access. The gold standard for subject indexing is generally observed as the manual assignment of terms from a controlled vocabulary by an information professional. However, human generation of metadata is costly in terms of both money and time (Greenberg et. al., 2002). Data repositories generally require the author to submit supplemental information about the datasets (metadata) upon deposition thus contributing to metadata creation. Because all datasets submitted to Dryad are linked to published journal articles, the bibliographic information provided by the journal is automatically captured and utilized as source metadata for each Dryad data package (Greenberg, 2009). This process of automatic metadata creation supports Dryad's goal to make the data submission process as easy, user-friendly, and as burden-free as possible (http://datadryad.org/factSheet). So for example, author metadata is automatically applied to the author field for the data object metadata records. Likewise, the keywords assigned by the author for the article are used as the subject terms in Dryads metadata scheme.

Whereas automatic metadata propagation is obviously useful and saves time, how do we know that a metadata record for a published research article can legitimately serve as a metadata source for the data object(s)? Though this question has not been empirically

addressed, Greenberg (2009) effectively makes the case that a metadata record for a published research article can serve as a source of metadata for data objects represented in the article based on the logic that published research is in effect an artifact generated by the data.

Does this logic hold true for author-assigned keywords which are used as subject descriptors for Dryad datasets? Is the content of data packages within Dryad accurately described by subject keywords assigned by the authors to represent the intellectual content of the research article? Based on the premise argued by Greenberg that published research (the output) is closely linked to its underlying data (the input), one can agree to this logic. One might consider however, the possibility of improving the metadata record by adding additional keywords from controlled vocabularies.

Keywords versus Subject Descriptors

Keywords (often called keyphrases) are used to provide a concise description of an information resource such as a document, image, or in Dryad's case, a dataset. Whereas keywords are often free-text, natural-language terms drawn from the document and assigned by the author, descriptors are terms that are drawn from a controlled vocabulary and typically assigned by an indexer. The debate concerning the effectiveness and usefulness of both controlled and natural indexing languages began in the 1960's with the Cranfield experiments evaluating 1) the performance of various indexing languages in retrieval and 2) the effectiveness of vocabulary control (Cleverdon, 1968; Cleverdon & Mills, 1963; and Cleverdon, Mills, & Keen, 1966). These studies suggested that under certain circumstances, natural language systems can perform as well as or

better than controlled vocabulary systems. Over time, many follow-up studies corroborated these views (Rowley, 1994) though in the 1970's, Carrow and Nugent (1977) first proposed the idea of complementarity; that the two approaches were complementary and that optimal performance would be obtained by using both methods.

More recently, studies in the literature have compared documents assigned descriptors by indexers with documents assigned keywords by authors. One study by Gil-Leiva and Alonso-Arroya (2007) found that nearly 25% of keywords matched exactly with descriptors and an additional 21% of keywords either partially matched or were a variant form of the descriptor. The authors suggested that both human indexers and automatic indexing programs could be guided by author-assigned keywords. On the contrary, the fact that approximately 54% of the keywords did not match the descriptors suggested that keywords might be used as additional points of entry for users. A similar finding was reported by Strader (2009) who looked at the overlap between author-assigned keywords and cataloger-assigned Library of Congress Headings (LCSH) for a set of electronic theses and dissertations in an online catalog. The author goes on to say that the use of LCSH complements the use of author-assigned keywords by offering unique terms for discoverability.

Automatic Indexing Techniques

Likewise, Dryad plans to enhance the discoverability of its datasets by enriching the author-assigned keywords currently used as dataset subject descriptors with controlled vocabulary terms (Vision, 2010). In general, the manual creation of subject metadata is cost-prohibitive and time consuming: to minimize a metadata bottleneck as

described by Liddy et al. (2002), Dryad's metadata plan emphasizes the use of automated techniques as much as possible. Automated indexing techniques have improved markedly over the years. The concept of automatic indexing goes back many decades. In fact, Lancaster (2003) discusses the first examples of automatic indexing based on word frequency published in the 1950's (Baxendale, 1958; Luhn, 1957).

Automatic Keyphrase Extraction versus Keyphrase Assignment

There are two classic approaches to automatic indexing as explained in Lancaster (2003). Automatic keyphrase extraction is based on extraction indexing where words or phrases in the text are extracted and used to represent the aboutness of the text. Extracted phrases are then chosen based on statistical algorithms that analyze the properties of the candidate's keyphrases such as frequency of occurrence and length. Keyphrase extraction can be criticized because it often generates terms that are non-sensical or inappropriate.

In contrast, keyphrase assignment (also known index term assignment) is based on the same premise as human indexing: terms are selected from a controlled vocabulary. While this approach provides more consistency than keyphrase extraction, it also requires a large set of training documents that have been manually indexed in order to provide positive and negative examples (Medelyan & Witten, 2008). This type of inductive learning scheme is known as machine-learning and is used to build rules that can predict the classification of new documents. The literature is replete with various algorithms that incorporate various statistical and machine learning techniques to both term assignment and keyphrase extraction approaches. Because Dryad's automatic metadata extractor

utilizes the KEA++ algorithm, this literature review will focus on the main studies describing its use and performance.

KEA

The original KEA algorithm was published by Frank, Paynter, Witten, Gutwin, and Nevill-Manning in 1999. The KEA algorithm automatically extracts keyphrases from text by choosing candidate keyphrases using lexical methods, determining values for each candidate's features, and then predicting which candidates are "good" key phrases using a machine-learning algorithm (Witten, Paynter, & Frank, 1999). Training involves a set of training documents that must include the author's keyphrases or manually-assigned descriptors so numerical values can be assigned to calculated features that are either positive ("is a keyphrase") or negative ("is not a keyphrase"). This model built during training is then applied to new "test" documents.

Witten et al. (1999) assessed the quality of KEA-extracted keyphrases and found that one to two of KEA-assigned keyphrases matched the five author-assigned keyphrases. A later study by Jones and Paynter (2002) further evaluated KEA using human phrase assessment. Twenty–eight subjects were given a technical document to read and asked to rate the suitability of approximately 60 candidate phrases to represent the document. Overall, subjects found that the majority (80%) of KEA-generated keyphrases were viewed positively and considered relevant to the document.

KEA++

Medelyan and Witten (2006) then improved upon the original KEA algorithm by creating KEA++ which enhanced the keyphrase extraction process by using a controlled

vocabulary. KEA++ can utilize any controlled vocabulary (or thesaurus) that is encoded in the Simple Knowledge Organization System (SKOS). SKOS is a semantic web language used for representing and applying controlled vocabularies such as thesauri, classifications schemes, subject headings, or taxonomies (Miles & Perez-Aguera, 2007). Candidates are identified and then matched against terms in the vocabulary. Terms that are considered non-descriptors are replaced with the corresponding descriptor thus allowing terms that are not in the document to become generated keyphrases. The final set of keyphrases are determined from this candidate list using machine-learning based on four term attributes including TFxIDF, position of the first appearance, length, and node degree (Medelyan & Witten, 2006). The machine-learning model used to predict the best keyphrases is the same as used in the original KEA algorithm.

Medelyan and Witten (2008) evaluated KEA++ extensively on agricultural documents from the United Nations Food and Agriculture Organization which are manually indexed with terms from the Agrovoc thesaurus. The study used 780 randomly-selected, full-text documents as both the training and evaluation corpus. A separate corpus of 30 different documents was manually indexed by six professional indexers at FAO so that the consistency of KEA++ could be compared with the consistency of human indexers. Overall, the authors found that KEA++ (automatic controlled-vocabulary indexing) outperformed KEA (automatic free-text indexing) and that most KEA++-assigned terms either matched or were similar to terms assigned by human indexers. KEA++ indexing was approximately 30% consistent with human indexers compared with human indexers who were 39% consistent with each other. KEA++ did

have problems assigning some keyphrases that were identified as relevant by human indexers but not by KEA++; these were terms that did not appear (or appear often) in the text.

Subject Area of Training Documents

There are two different aspects of KEA and KEA++ training that need to be addressed. The first is in reference to the subject area of the training documents. Frank et al. (1999) studied the extent to which models formed by KEA "transfer" from one subject domain to another by training KEA on one collection of journals articles and then testing a different subject collection. The authors concluded there was a "trend" towards improved results when testing and training documents were from the same population (i.e., same domain); however, the differences were not statistically significant. This result lead the authors to exploit the domain-specific information by including a new attribute into the machine learning algorithm: the keyphrase frequency. This new attribute keeps track of the number of times an author-assigned keyphrase occurs in the training set. Results of the studies showed that utilizing domain-specific information increased the number of correctly extracted keyphrases. In addition, performance improved as more documents were included in the training set. The authors suggest that the quality of KEA-generated keyphrases can be improved when domain-specific information is utilized.

Second, the number of training sets required to produce quality keyphrases is of interest to catalogers wanting to utilize KEA++ as their automated indexing approach. Witten et al. (1999) found that performance of the original KEA (keyphrase extraction

with no controlled vocabulary) did not improve after the training set reached 50 documents in contrast to the performance of KEA++ in which precision and recall increased when the training set increased from 50 to 100 documents and then leveled off (Medelyan & Witten, 2008). Interestingly, when domain-specific information was incorporated into the machine-learning model as described by Witten et al. (1999) marked improvement was observed when the training corpus was increased from 100 to 1000 documents.

Dryad's Controlled Vocabulary Requirements

The previous studies provide evidence to support the idea that thesaurus-based automatic keyphrase indexing (KEA++) is superior in performance to free text keyphrase indexing (KEA). But what about documents whose discipline either does not have a controlled vocabulary, or is multi-disciplinary in nature and therefore would benefit from more than one controlled vocabulary? This is the difficulty encountered by the creators of the Dryad repository. As previously mentioned, Dryad initially contained the underlying datasets of research published in the area of evolutionary biology and ecology [note: Dryad recently expanded to accept data from all the basic and applied biosciences]. This is a highly interdisciplinary field that integrates a vast range of different scientific fields including ecology, developmental biology, genetics, molecular biology, systematics, and paleontology.

A team study reported by Greenberg (2009) determined which controlled vocabulary systems would best fit the needs of Dryad. The study evaluated a sample of approximately 600 author-assigned keywords obtained from 104 articles in Dryad's

partner journals. Keywords were assigned to nine different facets including *topic*, *research method*, *geographic location*, *taxon*, *personal name*, *agency name*, *anatomical aspect*, *discipline*, and *habitat*. Terms in each facet were then mapped to terms in appropriate controlled vocabularies and ontological sources (e.g., Education Resources Information Center (ERIC) Thesaurus, National Biological Information Infrastructure's Biocomplexity Thesaurus (NBII Thesaurus), Medical Subject Headings (MeSH), Library of Congress Subject Headings (LCSH), Getty Thesaurus of Geographic Names (TGN), Gene Ontology (GO), Integrated Taxonomic Information System (ITIS) are examples). Matches were categorized as either "exact" meaning that the keyword exactly matched to either the preferred or non-preferred term OR as a "partial and non-match" meaning the author-assigned keyword either partially matched the preferred or non-preferred term or did not match any term. Overall, 22% of the keywords mapped exactly to the NBII Thesaurus; 23% mapped to MeSH; and 33% mapped to LCSH. These results indicate that a single controlled vocabulary will not sufficiently represent the range of concepts present in a Dryad dataset and that instead multiple vocabularies could provide better representation.

Evaluating Automatically-generated Keyphrase Quality

The final section of the literature review will focus on the various techniques used to evaluate performance of automatic indexing algorithms. Two basic methods for evaluating automatically generated keyphrases are apparent in the literature. The first approach utilizes precision and recall to determine how well either extracted or generated keyphrases match a set of "relevant" phrases. The second approach uses human

evaluation to rate extracted or generated keyphrases. Both methods have their advantages and disadvantages.

The majority of studies reviewed evaluated the quality of keyphrases generated by automatic indexing algorithms by calculating precision and recall (e.g., Frank et al., 1999; Jones & Paynter, 2002; Medelyan & Witten, 2008; Witten et al., 1999). This approach compares the algorithm-generated phrases with a set of "relevant" phrases (e.g., author-assigned keywords). The precision of a set of automatically-generated phrases is usually defined as the proportion of the set that match the author-assigned keywords whereas recall is the proportion of the total number of author-assigned keywords that appear in the set of automatically-generated keywords (Jones & Paynter, 2002). A potential problem with this method arises if the documents being tested do not have author-assigned keywords. In addition, Jones and Paynter (2002) questioned whether author-assigned keyphrases were an acceptable standard against which to measure performance. Their results suggest that authors do provide quality keyphrases and that keyphrases are an acceptable standard in which to compare automatically-generated descriptors against.

Many studies have also determined phrase quality using human assessment (Barker & Cornacchia, 2000; Jones & Paynter, 2002; Tolle & Chen, 2000). Usually, subjects are given a document to read and a phrase list comprised of automatically generated keyphrases. Subjects then rate the relevance of individual phrases (or phrase sets) to the document. In addition, Jones and Paynter (2002) demonstrated how precision and recall metrics can be applied to subjective evaluation. Based on a method used by Tolle and Chen (2000), Jones and Paynter used subject precision (SP) and subject recall

(SR) to determine the proportion of extracted phrases chosen as relevant by a rater. In this way, one does not have to have the set of "relevant" keyword phrases identified beforehand. Instead, relevance is a subjective factor determined by the users in the experiment.

Of course, the main disadvantage of using humans to evaluate keyphrases is the subjectivity of the assessment; this can lead to inconsistent scores between raters. On the other hand, gaining people's actual opinion of generated-keywords may provide a more true-to-life assessment of keyword quality. It is important to note that it may be possible to reduce inter-assessor inconsistency by using human raters whose domain knowledge matches that of the documents being tested (Jones & Paynter, 2002; Tolle & Chen, 2000).

Summary

Data sets deposited in digital data repositories such as Dryad require rich metadata to ensure data accessibility and usability. Neither data creators nor data curators have the time to manually assign subject metadata to datasets and therefore the utilization of automatic metadata generation techniques such as automatic keyphrase indexing is required. The literature reveals a long history of automatic keyphrase indexing techniques that are continually improving. KEA++'s new approach to thesaurus-based indexing using machine learning combines the best attributes of both keyphrase indexing and keyphrase assignment techniques. HIVE uses KEA++ and thus can utilize any thesaurus that is SKOS-encoded. This is most beneficial to the Dryad repository which due to its interdisciplinary nature requires multiple controlled vocabularies to sufficiently represent those concepts found in Dryad datasets. Therefore,

HIVE is potentially a very useful automatic metadata generation tool to the Dryad data repository but one that needs to be tested. In addition, the literature reveals conflicting evidence to whether or not domain-specific training sets will improve performance of KEA++. Thus, the purpose of this research paper is to begin to answer the following questions:

Question one: How effective is HIVE for automatically generating controlled vocabulary terms from full-text articles?

Question two: How effective are HIVE-generated controlled vocabulary terms in describing the content of Dryad dataset(s)?

Question three: To what extent will training HIVE in a specific sub-discipline of evolutionary biology/ecology increase its ability to generate controlled vocabulary terms for that specific topic?

METHODS

A quasi-experiment, supported by human evaluation, was used to answer the questions posited above. This study is divided into two parts. The first part seeks to determine if HIVE can be "trained" in a specific sub-domain of ecology/evolutionary biology using machine-learning techniques. Performance of this trained HIVE (or SmartHIVE) will be based on the modified recall and precision metrics previously described in the literature review (Frank et al., 1999; Jones & Paynter, 2002; Medelyan & Witten, 2008; Witten et al., 1999). The second part of the study seeks to determine how domain experts in ecology and/or evolutionary biology rate the ability of HIVE-generated subject descriptors to describe both the journal article and its underlying data. Ideally, both HIVE and SmartHIVE would have been evaluated by domain experts; however, in

order to assess evaluator's opinions of the relevancy of HIVE-generated keywords to both the article and the underlying data set within the small scale of a master's paper, I chose to look at only the SmartHIVE-generated keywords.

RESEARCH DESIGN

Conversion of MeSH to SKOS

The Medical Subject Headings (MeSH) thesaurus is a controlled vocabulary used for the indexing of journals in MEDLINE and is maintained by the National Library of Medicine. The MeSH thesaurus contains over 26,000 subject descriptors. An XML version of the MeSH vocabulary was converted to SKOS RDF/XML for indexing in HIVE because an official MeSH SKOS is not currently available. For more information on the conversion of the MeSH vocabulary to SKOS, please refer to the HIVE wiki (http://code.google.com/p/hive-mrc/wiki/MeshToSKOS).

HIVE's Keyphrase Indexing Algorithm

As previously mentioned, the HIVE Automatic Concept Indexer relies on SKOS-encoded vocabularies and KEA++ machine learning to automatically generate subject metadata. KEA++ works in two stages: *candidate identification* and *keyphrase selection* (Medelyan & Witten, 2008). In the first stage, candidate terms and phrases are extracted from the full-text document and matched against terms in the SKOS-encoded vocabularies. Candidate terms that are non-descriptors in the controlled vocabulary are replaced by their corresponding descriptors. The second phase uses a model to identify the most significant terms based on certain features of the terms. This model has to be

first learned by KEA++ through training data which can be defined as manually indexed full-text documents. For each training document, four different features are calculated:

- **Term frequency X inverse document frequency weight (TFxIDF)** which assesses how specific a phrase is to a document by comparing the frequency of the phrase in the document with the frequency of that phrase in the training set (Salton & MacGill, 1983).
- **Position of the first occurrence of a keyphrase** determines the proportion of the document that precedes the phrase's first appearance; terms that have very low or very high values are more likely to be index terms because they are positioned in the beginning (i.e., title, abstract, or introduction) or end of an article (i.e., conclusion).
- **Length of candidate phrase (# of words)** allows the algorithm to choose best phrase length.
- **Node degree** reflects how many links exist between the candidate keyphrase and terms in the thesaurus, between the keyphrases and other candidate phrases, or as a ratio of the two.

Using the manually assigned index terms as a positive example, each candidate phrase is identified as an index term or not an index term. Based on the calculated feature values for each of these positive and negative examples, KEA++ creates a model that can then be applied to candidate keyphrases extracted from new documents. Candidate phrases from new documents are identified, feature values calculated, and overall probability for being an index term determined.

Training Documents

The first goal of this study was to determine if HIVE's automatic indexing ability could be improved by training in a specific sub-domain using the machine learning methods just described. The basic HIVE indexing algorithm used in this study is one based on KEA++ which is "pre-trained" on the AGROVOC training set as described in Medelyan and Whitten, 2008. As previously mentioned, to build a model for HIVE using KEA++, documents in the training set must have been manually indexed, preferentially with the controlled vocabularies that will be utilized during automatic indexing (NBII, MeSH, and LCSH). In this study, I chose to train HIVE using articles indexed with MeSH because the majority of data archived in Dryad were published in journals indexed by MEDLINE thus providing a convenient source of training documents.

To build a sub-domain-specific model for HIVE, a training set of 50 articles indexed in MEDLINE related to reproduction within the field of ecology and evolutionary biology were selected. Articles were identified in PubMed using the following MeSH [mh] descriptors as search terms: reproduction, "mating preference, general", "sexual behavior, animal", "maternal behavior", "paternal behavior", courtship, fertilization, and pregnancy in conjunction with Boolean AND (biological evolution [mh] OR ecology [mh]). To build the keyphrase extraction model, KEA was provided with training documents (full-text documents in a .txt format) accompanied by the "key" files (text files with containing manually assigned MeSH terms in which subheadings and the main descriptor indicators [*] were removed). The Apache Tika toolkit was used to convert PDFs to text. For detailed information relating to the training of KEA++, please refer to the HIVE wiki (http://code.google.com/p/hive-mrc/wiki/TrainingKEA). The

"training" of KEA was performed by Craig Willis, research assistant for the Metadata Research Center.

Test Documents

To test the ability of the HIVE Automatic Concept Indexer to create subject metadata for the Dryad data repository, a test set of eighteen domain-specific articles retrieved from Dryad relating to reproduction were purposely selected. Articles were obtained by searching the Dryad repository using the following search string: reproduction OR reproductive OR breeding OR mate OR mating OR courtship OR pregnancy OR fertility OR fecundity OR monogamy OR polyandry OR hermaphrodit*. Articles and datasets were limited to studies using animals. In addition, articles must have been indexed in MEDLINE.

A control set of eight domain-neutral articles was also created in order to determine if improvement in automatic subject indexing was limited to the domain in which HIVE was trained. To obtain a random selection of domain-neutral test articles, data packages in Dryad (as identified by their unique DOIs) were consecutively labeled 1- 367. A random sequence of numbers (from 1-367) was generated using a random number generator at http://www.random.org/ and the first eight Dryad data sets that were indexed in MEDLINE were selected.

Automatic Subject Metadata Generation

Selected documents were converted from PDF to text using the Apache Tika toolkit. The eight domain-neutral and eighteen domain-specific documents were run

through both the standard HIVE (untrained) and "SmartHIVE" (trained). Generated output was obtained for evaluation.

Evaluation of HIVE versus SmartHIVE

As previously discussed in the literature review, the majority of studies evaluating the quality of automatically generated keyphrases calculated a version of precision and recall. This approach compares automatically generated keyphrases with a set of "relevant" keyphrases, often author-assigned keywords. In this study, the MeSH descriptors assigned by the professional indexers at the National Library of Medicine are the "relevant" keywords. To determine precision and recall, the number of "correct" or identically matching MeSH terms between the automatically generated HIVE subject descriptors and those assigned the article in MEDLINE was calculated.

- **Precision** was determined using the formula: # of matching MeSH terms (i.e., # of "correct" HIVE generated terms)/ # of generated HIVE terms.
- **Recall** was determined using the formula: # of matching MeSH terms (i.e., # of "correct" HIVE generated terms)/the # of manually assigned MeSH terms.

Excel was used to determine the mean number of matching (or "correct") SmartHIVE-generated terms, precision, and recall. The standard error of the mean (SEM) was calculated by dividing the standard deviation by the square root of the sample size (n). Means were compared using an unpaired T test with Excel.

Domain Expert Evaluation of SmartHIVE-generated Subject Descriptors

The quality of SmartHIVE-generated subject descriptors was further evaluated using human assessment. Six evaluators were recruited from the biology departments at

the University of North Carolina at Chapel Hill and Duke University in Durham, NC and NESCent (National Evolutionary Synthesis Center), a nonprofit science center in Durham, NC dedicated to cross-disciplinary research in evolution. The evaluators were graduate students in the fields of ecology and evolutionary biology (with a special emphasis on reproductive biology) because inter-evaluator inconsistency has been reported to be reduced by using human raters whose domain knowledge matches that of the documents being tested (Jones & Paynter, 2002; Tolle & Chen, 2000). Evaluators were each provided a $100.00 honorarium upon completion of the work.

Human Assessment Procedure

The human evaluation portion of this research study is based on the work of Jones and Paynter (2002) and Tolle and Chen (2000). Each of the 6 participants evaluated the automatically-generated subject keywords from nine different Dryad data packages which were randomly assigned. After students agreed to participate as evaluators, they were sent an email that contained a brief introduction, nine PDFs, and an Excel spreadsheet containing the study description and rationale, detailed instructions for completing the evaluations, and nine worksheets, one for each article/data package. Each worksheet contained the article title, the name of the file containing the article PDF, the link to the Dryad data package, and two identical lists of the twenty SmartHIVE-generated keywords. Each evaluator was instructed to read/scan the article and the dataset(s) and then rate each of the twenty SmartHIVE-generated keywords for their relevancy to both the article AND the data set(s). They could choose between the following choices: relevant, partially-relevant, and not relevant. Space was also

provided for evaluators to write in comments on each keyword. Evaluators were provided with the following definitions to guide the rating process:

- **Relevant phrases** best "represent" the topic covered in document or data set.
- **Partially-relevant phrases** somewhat "represent" the topic covered in the document or data set and can be considered related to the topic.
- Phrases that are **not relevant** are *not* considered "representative" of the document or data set topic.

The set of SmartHIVE-generated descriptors for each of the eighteen domain specific article/data packages was evaluated by three different people.

Evaluation of SmartHIVE-generated Keywords by Domain Experts

This study sought to determine how scientists rate the relevance of HIVE-generated subject keyphrases to both the original full-text article and to its underlying data sets. The set of eighteen domain-specific articles was run through SmartHIVE and the output collected. Evaluators rated each of the twenty keywords as relevant, partially-relevant, and non-relevant to both the full-text article and to its associated data set(s). The number of relevant, partially-relevant, and non-relevant keywords was tabulated and the means for each paper (n=3) calculated. The mean number ±standard error of the mean (standard deviation/square root of n) of relevant, partially-relevant, and non-relevant keywords for both the article and the data set(s) were calculated using Excel. Means were compared using an unpaired T test in Excel. In addition, the mean number of manually-assigned MeSH terms was identified for each category.

RESULTS

Matching, Precision, and Recall

Both the domain-specific (n=18) and domain-neutral (n=8) test sets were converted to text and run through both the minimally trained HIVE and the sub-domain trained SmartHIVE algorithms. Generated-output was limited to twenty keywords. This evaluation was designed to address the question of whether training HIVE in a sub-domain improves its ability to assign subject descriptors to articles in that specific domain. The first measurement of keyword quality was the extent to which HIVE produced the same subject descriptors as manually assigned by indexers at the National Library of Medicine (or number "correct"). For each article, HIVE-generated keywords were compared with manually assigned MeSH index terms. Table 1 summarizes the results and includes the number of HIVE keywords that match exactly manually assigned MeSH index terms. Precision and recall are also included because previous evaluations of KEA, KEA++, and other keyphrase extraction tools have utilized these measures to determine performance.

Overall, performance results were similar for keywords generated from domain-specific articles and domain-neutral articles run through both minimally-trained HIVE and sub-domain trained SmartHIVE (Table 1). The mean number of correctly identified MeSH terms produced by HIVE and SmartHIVE was not statistically different for domain-specific test documents (3.6±.40 vs. 3.0±.40, respectively; p=.33) or domain-neutral test documents (3.4±1.3 vs. 3.3±1.2, respectively; p=.84). Likewise, Table 1 shows that there was no statistical difference between mean precision and recall for

keywords generated by HIVE or SmartHIVE for domain-specific documents or domain-neutral documents.

Table 1: Performance of HIVE versus SmartHIVE

	# Exactly Matching MeSH Terms ±SEM	Precision (%)	Recall (%)
Domain specific test set: n=18			
HIVE (untrained)	3.6±.40	17.8	31.2
SmartHIVE (domain-trained)	3.0±.40	15.0	25.8
p=	.33	.33	.18
Domain neutral test set: n=8			
HIVE (untrained)	3.4±.45	16.9	28.6
SmartHIVE (trained)	3.3±.41	16.3	27.9
p=	.84	.84	.86

Table 1 strictly reports whether or not HIVE-generated terms match exactly manually assigned MeSH index terms. In some instances, there were HIVE-generated terms that were very close to those assigned by indexers but were not exact. For example, the HIVE-generated keyword “Photoreceptor Cells” is very close to the assigned “Photoreceptor Cells, invertebrate”. Another common example is that the indexers assigned the descriptor “Animals” and HIVE would assign the more specific “Insects”. To account for these semantically related terms, Table 2 reports the number and percentages of exact matches, terms that are non-matching but similar within the immediate hierarchy including broader, narrower, or related, and terms that are unrelated.

The percentages of exact matches between HIVE and SmartHIVE were similar for both domain-specific and domain-neutral test sets (18% vs. 15%) and (17% and

16%), respectively. Overall, HIVE and SmartHIVE do a rather poor job selecting keywords that correctly match the terms assigned by indexers. When terms that are similar but not perfectly matched are factored in, the percentage of agreement between HIVE and human indexers increased to 20%-22% of generated terms. Still, between 78%-80% of HIVE-generated terms are of no relation to MeSH terms manually assigned by humans.

Table 2: Distribution of exact matches, similar, and unrelated terms.

	Total terms	**Exact match**	**Similar within the immediate hierarchy (broader, narrower, related)**	**Unrelated**
Domain specific test set:				
HIVE (untrained)	360	64 (18%)	15 (4%)	281 (78%)
SmartHIVE (domain-trained)	360	54 (15%)	17 (5%)	289 (80%)
Domain neutral test set:				
HIVE (untrained)	160	27 (17%)	7 (4%)	126 (79%)
SmartHIVE (trained)	160	26 (16%)	10 (6%)	124 (78%)

Subject Evaluation of SmartHIVE

The eighteen domain-specific articles were run through SmartHIVE (HIVE trained on articles indexed with MeSH that were related to reproduction) and output collected. Evaluators rated nine sets of SmartHIVE-generated keywords (twenty keywords/set) resulting in each of the eighteen articles being evaluated three times. The mean number of relevant, partially-relevant, and non-relevant keywords was calculated for each set of SmartHIVE-generated keywords. Figure 1 shows the mean number of "relevant" keywords for each of the eighteen articles and underlying datasets. It is

obvious that there was wide variation in the ability of SmartHIVE to generate keywords viewed as relevant by human evaluators. The mean number of "relevant" keywords per article ranged from 1.3 – 7.7 and 0.3 – 6.7 per data set. In general, the trend was for evaluators to assign more keywords as "relevant" in reference to the article than for data sets however this was not always the case.

Figure 1. Number of relevant keywords for articles vs. data set(s).

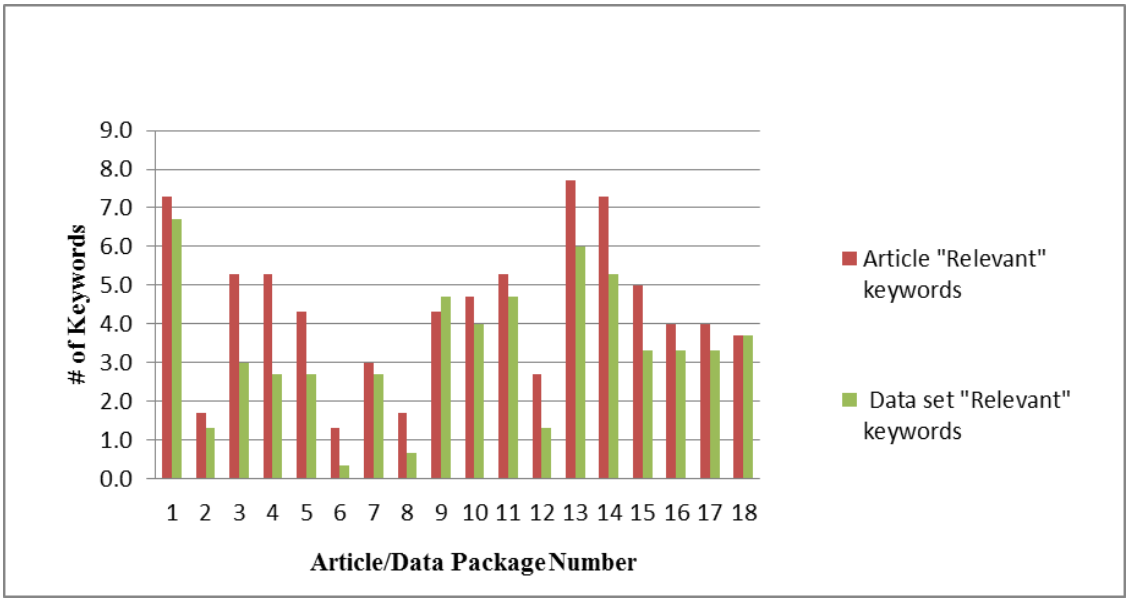

Figure 2 shows the number of "partially-relevant" keywords for articles versus data sets as assessed by study participants. A similar amount of variation was observed for keywords described as "partially-relevant" by evaluators: the mean number of keywords rated as "partially-relevant" in regards to their ability to describe the article ranged from 1.0-8.7 and 0.3-6.7 for data sets. As observed with keywords rated as "relevant", there were more keywords rated as "partially-relevant" as applied to the article than as applied to the data sets.

Figure 2. Number of partially-relevant keywords for articles vs. data set(s).

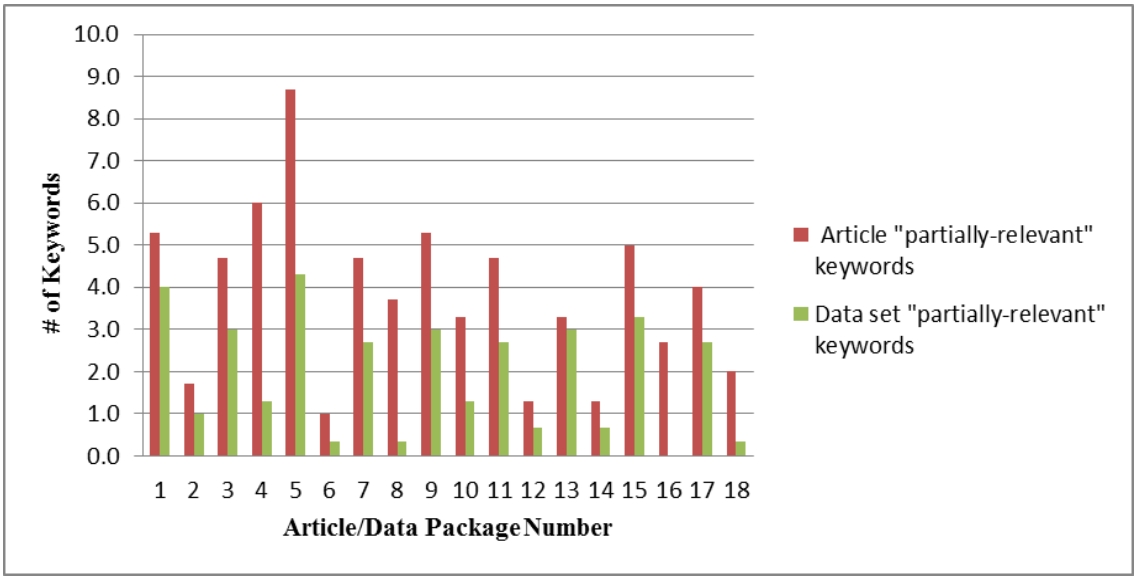

Table 3 presents the mean number of "relevant" and "partially-relevant" ratings assigned by evaluators to article and to data set(s). Evaluators rated an average of 4.0±.45 keywords/document as relevant compared with 3.3±.41 keywords/data package and did not differ statistically (p=.25). Evaluators assigned significantly (p=.002) fewer keywords as "partially-relevant" for data packages than compared with articles (1.9±.33 vs. 3.8±.47, respectively). Accordingly, significantly (p=.0008) more terms were considered "non-relevant" for data packages than compared with articles (14.7±1.0 vs. 11.2±.68). If one considers the ratings "relevant" and "partially-relevant" as useful keywords, then the combined subject precision is 39% for articles and 26% for data. Subject precision was calculated by dividing the # of keywords designated as "relevant" by the subject by the total # of keywords generated.

Table 3. Evaluator's ratings of SmartHIVE-generated keywords: relevance to article and to data set(s).

	Mean # of relevant keywords ±SEM	Mean # of relevant keywords that were also manually-assigned	Mean # of partially-relevant keywords ±SEM	Mean # of partially-relevant keywords that were also manually-assigned	Mean # of non-relevant keywords ±SEM	Mean # of non-relevant keywords that were also manually-assigned
Relevant to Article	4.0±.45	1.8	3.8±.47	0.7	11.2±.68	0.5
Relevant to Data set(s)	3.3±.41	1.5	1.9±.33	0.5	14.7±1.0	0.8

It is also of interest to determine how many of the keywords deemed "relevant" or "partially-relevant" by our evaluators were also assigned MeSH terms in MEDLINE. Table 3 shows that more keywords in the "relevant" category were manually-assigned MeSH keywords than in either the "partially-relevant" or "non-relevant" categories. Specifically, for every 4 article-relevant keywords, 1.8 were also manually-assigned MeSH keywords. Likewise, for every 3.3 data-relevant keywords, 1.5 were manually-assigned keywords. When counting the number of keywords that were also manually assigned MeSH descriptors, I only included exact matches, not semantically similar keywords such as broader, narrower, or related terms.

There are some instances when a manually-assigned MeSH term was viewed as "non-relevant" by the evaluators. This appeared to occur more often when applying keywords to data sets (occurred 21/54 times) compared with articles (occurred 6/54 times). This may be due to the fact that only certain data sets are deposited in Dryad possibly representing only specific aspects of the article's findings. Therefore, general

keywords that may apply to the article in general would not apply to a specific data set. In addition, some MeSH descriptors are just too broad to apply to datasets (e.g., Disorders of Sex Development).

DISCUSSION

The two main goals of this study were to determine:

1. To what extent will training HIVE in a specific sub-discipline of evolutionary biology/ecology increase its ability to generate controlled vocabulary terms (specifically, MeSH) for that specific topic?
2. HIVE's effectiveness in automatically generating controlled vocabulary terms (specifically, MeSH) from full-text articles that can be used to describe the content of the article and that of its underlying data set(s) archived in Dryad.

The results of the first part of this small study indicate that training HIVE in a specific sub-domain of ecology/evolutionary biology was unsuccessful in improving the automatic generation of keywords for articles in that sub-domain. Specifically, articles related to the topic of reproduction that were automatically indexed using the HIVE algorithm trained specifically on reproductive-related articles (SmartHIVE) performed no better than articles automatically indexed using the minimally-trained HIVE based on matching, precision, and recall. Indeed, though not statistically significant, the trend was for HIVE to produce slightly more "correct" keywords than SmartHIVE. It is interesting to note the differences in keywords assigned by HIVE and SmartHIVE. Table 4 demonstrates the differences in keyword production in one example document. The most notable difference is that the HIVE algorithm produced five "correct" keywords but only two remained in the keyword list generated by SmartHIVE. In addition, note the various

keywords unique to each algorithm. Table 4 also provides an example of how training HIVE resulted in less total number of "correct" keywords but also produced a unique keyword that was not assigned as an index term by MEDLINE that evaluators deemed as "relevant" to the data set. There appears to be no definitive positive or negative effect of training HIVE; for some articles indexing improved with SmartHIVE and for others, performance declined.

Table 4. An example of keyword sets assigned an article run through HIVE and SmartHIVE

HIVE Keywords	SmartHIVE Keywords	Assigned MeSH Index Terms
Seasons	Salmonidae	MH - Animals
Sexuality	Individuation	MH - Bayes Theorem
Trout	Sex Characteristics	MH - Biological Evolution
Salmon	Mental Competency	MH - Body Size
Ecology	Oncorhynchus kisutch	MH - Female
Probability	Crassostrea	MH - Genotype
Reproduction	Trout	MH - Male
Genotype	Salmon	MH - Models, Statistical
Rivers	Seasons	MH - Norway
Viverridae	Elastomers	MH - Pedigree
Parenting	Sharks	MH - Reproduction/*genetics
Ficus	Pedigree	MH - Sequence Analysis, DNA
Salmo salar	Ambystoma	MH - *Sexual Behavior, Animal
Salmonidae	Pliability	MH - Trout/*genetics
Body Size	Salmo salar	
Individuation	Cesarean Section	
Population	Normal Distribution	
Breeding	Microsatellite Repeats$	
Pedigree	Principle-Based Ethics	
Egg Shell	Rivers	

*Terms shaded in green were also manually assigned in MEDLINE
* The term shaded in yellow is a narrower MeSH term for Animal.
$ Term was not assigned manually but was selected as a "relevant" keyword to represent the data by 2/3 evaluators.
*(Serbezov, D. Bernatchex, L, Olsen, E.M. & Vollestad, L.A. (2010). Mating patterns and determinants of individual reproductive success in brown trout (*Salmo trutta*) revealed by parentage analysis of an entire stream living population. Molecular Ecology. 19. 3193-3205.)

One odd outcome of training HIVE in the sub-domain of reproduction was that articles indexed with SmartHIVE no longer were assigned the MeSH term “reproduction” or “breeding”. These two terms were both produced in 9/18 keyword lists when run through HIVE but never appeared as a keyword from SmartHIVE. Similarly, the term “mating preference, animal” was never assigned by either HIVE or SmartHIVE despite that heading being prevalent in the training set used to create SmartHIVE. Likewise, the keywords “sexuality” or “sex characteristics” were produced as keywords but never the term “sexual behavior, animal”, which was commonly assigned manually as an index term in MEDLINE. The reasons for these occurrences are unknown and will require further exploration.

Human evaluation of SmartHIVE-generated keywords was conducted to provide more “ecological validity” to the performance indicators of precision and recall. I wanted to know not only if HIVE could generate the same keywords as those assigned by the indexers of NLM, but also to determine what scientists thought of those assigned keywords. The results of the second part of this study suggest that the term sets produced by HIVE are moderately useful in describing the subject content of both the article and its underlying data set(s). On average, evaluators rated 4.0/20 keywords as relevant/document and 3.3/20 keywords as relevant/data package. There was however, great variation in the evaluator’s scoring of keywords within the same document/data package. However, there was no trend observed for any one specific evaluator (i.e., no one evaluator consistently chose more or less terms as relevant). This made it impossible to remove any one evaluator’s scores in order to improve consistency. Evaluator 5 tended

to choose less terms as relevant, especially in respect to relevancy to the data packages, however, it was not consistent enough to remove that evaluator from the study.

An important goal of this study was to determine if the HIVE-generated keywords assigned to an article were also considered useful in describing the article's data thus providing a beneficial tool to assist in the indexing of archived datasets. Overall, it appeared that MeSH index terms automatically generated by SmartHIVE that were considered by the evaluators as relevant to the article were not statistically different from that of the data package (4.0 relevant terms to 3.3 relevant terms, respectively). Coincidentally, there was statistically more "non-relevant" index terms assigned to the data than compared with those assigned the article. Still, these results suggest that keywords assigned to an article can potentially be used as index terms for data sets and is rather encouraging considering that HIVE is basing its term selection on the full-text article and not the actual data set(s). Additionally, these findings support Greenberg (2009) who suggested that a metadata record for a published research article can serve as a source of metadata for data objects represented in the article based on the logic that published research is in effect an artifact generated by the data.

On the other hand, one must consider that the inability of SmartHIVE to consistently generate high quality keywords (based both on % matching with manually assigned MeSH terms and evaluator opinion) is disappointing. For example, 8/18 keyword sets contained two or less matches with manually-assigned MeSH terms. What are some possible explanations for these poorly indexed documents? First of all, in general, terms assigned from the MeSH thesauri are very complex and difficult to match. Second, the MeSH model does not map cleanly to SKOS resulting in a loss of

information. Finally, past studies of KEA have mentioned possible problems with the conversion of the document from PDF to plain text. For example, errors in conversion could possibly result in erroneous keywords. Careful examination of the converted text may reveal possible explanations for non-matching and non-relevant keywords.

An interesting finding was that HIVE-generated index terms did not have to actually be a "correct" match with a manually-assigned term in order to be rated positively by evaluators. Specifically, slightly less than half of keywords considered relevant to both the article and the data package by evaluators (45% and 46%, respectively) were also designated MeSH index terms for MEDLINE as well. One observed reason for this is that sometimes evaluators chose keywords that appeared "relevant" to the study or data but were not actually the "proper" use of the subject descriptor. For example, SmartHIVE-generated the term "color" but the human indexer used the term "pigmentation" resulting in a SmartHIVE-generated term that is close but not technically correct. However, the term "color" may work just as well as a keyword as the term "pigmentation".

LIMITATIONS

There are two main limitations present in this study. The first limitation to consider is that this study only looked at MeSH, one of the several thesauri available in HIVE. One of the main advantages of utilizing HIVE is that it dynamically integrates multiple discipline-specific controlled vocabularies so that it overcomes the difficulty in providing access to multiple vocabularies for metadata descriptions for interdisciplinary resources. For the purposes of this study, it was necessary to choose one controlled vocabulary in order to train HIVE in a specific subject. Training documents need to be

indexed with the controlled vocabulary that will be used to produce keywords. Most articles whose data are archived in Dryad are indexed using MeSH descriptors in MEDLINE and thus provided a convenient source of training information. Thus, only looking at the ability of HIVE to assign keywords from the MeSH thesauri while suitable for a research study, does not utilize the full power of HIVE's multiple vocabularies.

This is especially important when considering the interdisciplinary nature of the articles indexed in this study. Greenberg (2009) showed that a single controlled vocabulary does not sufficiently represent the wide range of concepts present in a Dryad dataset. Specifically, out of 600 author-assigned keywords from a set of 104 articles in Dryad partner journals, only 23% of keywords mapped to MeSH. Another difficulty with only using MeSH to index datasets that focus on ecology and evolutionary biology is that MeSH is a biomedical thesauri and there are concepts in Dryad datasets that are not present in MeSH. For example, the subject of one of the test articles was migration. "Seasonal migration" is an author-assigned keyword but "migration" is not a MeSH term. Future studies should focus on user's ratings of keywords assigned from multiple vocabularies.

The second major limitation to this study was the subjective nature of the evaluations. It is apparent from the wide variations in subject assigned relevancy ratings that the evaluation of keywords for relevancy is an extremely subjective process. The author of this study was aware of the potential difficulties involved with human evaluators and attempted to minimize inter-evaluator inconsistency by utilizing subject evaluators who were knowledgeable in the domain of ecology and evolutionary biology. In fact, the author went as far as to recruit evaluators whose research focus was related to

reproduction. Unfortunately, efforts in this regard were not successful and there existed high inter-evaluator inconsistency.

CONCLUSION

The study presented in this paper addressed two major questions. One, can HIVE be subject-trained (SmartHIVE) in order to improve its automatic indexing performance and two, are the keywords produced by SmartHIVE effective in describing the content of both the journal article, and its underlying data? To address the first question, HIVE was first trained in the sub-discipline of reproduction with a set of 50 training documents—articles with the topic of reproduction and ecology/evolutionary biology *and* indexed by MEDLINE with MeSH subject descriptors. Eighteen documents in the sub-domain of reproduction and whose data were archived in Dryad were used as a test set that was run through both the minimally-trained HIVE algorithm and the subject-trained SmartHIVE. Overall, the results of Part One of the study suggest that subject training HIVE did not improve its ability to assign subject descriptors to articles in that subject domain. To address the second question, keyword sets generated from the eighteen subject-specific test sets run through SmartHIVE were evaluated by human evaluators knowledgeable in the domain of ecology and evolutionary biology for relevancy to both the article and to the underlying data. Results of Part Two showed that 39% of keywords were rated positively by human evaluators (either rated “relevant” or “partially-relevant”) in reference to the article whereas 26% of keywords were rated positively in describing the data.

The finding that training HIVE in a specific sub-discipline was ineffective in increasing the number of correctly-indexed keywords was surprising. Previous studies by

Frank et. al. (1999) showed that the quality of Kea-generated keyphrases was improved when domain-specific information was utilized. It is possible that HIVE's training set needs to be increased from 50 papers to 100 papers because Medelyan and Witten (1990) demonstrated that the performance of KEA++ (precision and recall) improved when the training set was increased from 50 to 100 documents. The specific content of the training documents should also be evaluated more closely. Perhaps the content of the training papers needs to be more directly related to the content of the documents in the test set. It might be useful to systematically compare the article content of the test set in relation to training set in order to identify any relationships to SmartHIVE's indexing performance. Results of the human relevance assessment indicated that MeSH keywords generated by HIVE were almost as effective in describing the content of the data set(s) as the article itself. This encouraging result suggests that HIVE could be useful for assisting both scientists and Dryad's curator in creating useful subject metadata for datasets in Dryad from the full-text journal article. Future work must focus on improving the quality of HIVE's output by both optimizing its machine learning algorithms and its ability to be subject-trained. The results of this future research will not only apply to the specific needs of the Dryad data repository but to all repositories that contain interdisciplinary collections and could benefit from the use of an interdisciplinary vocabulary system such as HIVE.

REFERENCES

Barker, K. K. & Cornacchia, N. (2000). Using noun phrase heads to extract document keyphrases. *Advances in artificial intelligence : 13th biennial conference of the Canadian Society for Computational Studies of Intelligence, AI 2000, Montréal, Québec, Canada, May 14-17, 2000 :* Springer: Berlin Heidelberg, 40-52.

Baxendale, P.B. (1958). Machine-made index for technical literature-an experiment. *IBM Journal of Research & Development, 2.* 354-361.

Carrow, D., & Nugent, J. (1981). Comparison of free text and index search abilities in an operating information system. In: *Information management in the 1980's Proceedings of the American Society for Information Science 40th Annual Meeting, September 26 – October 1, 1977.* White Plains, NY: Knowledge Industry Publications. 131-138.

Cleverdon, C.W. (1968). Effect of relevance assessments on comparative performance of index languages. Cranfield, U.K.: Cranfield College of Aeronautics.

Cleverson, C.W., & Mills, J. (1963). The testing of index language devices. *Aslib Proceedings, 15(4),* 106-130.

Cleverdon, C.W., Mills, J., & Keen, E.M. (1966). *Factors determining the performance of indexing systems, Vol. I -Design.* Cranfield, England: Aslib Cranfield Research Project.

Coyle, K. (2008). Machine indexing. *The Journal of Academic Librarianship, 34*(6), 530-531.

Davis, H.M. & Vickery, J. N. (2007). Datasets, a shift in the currency of scholarly communication: implications for library collections and acquisitions. *Serials Review, 33,* 26-32. Doi:10.1016/j.serrev.2006.11.004

Frank, E. Paynter, G.W., Witten, I.H., Gutwin, C., & Nevill-Manning, C.G. (1999). Domain-specific keyphrase extraction. *Proceedings of 16th International Joint Conference on Artificial Intelligence IJCAI-99,* 668-673.

Gil-Leiva, I., & Alonso-Arroyo, A. (2007). Keywords given by authors of scientific articles in database descriptors. *Journal of the American Society for Information Science and Technology, 58*(8), 1175-1187.

Greenberg, J. (2009). Theoretical considerations of lifecycle modeling: An analysis of the dryad repository demonstrating automatic metadata propagation, inheritance, and value system adoption. *Cataloging & Classification Quarterly, 47*(3-4), 380-402.

Greenberg, J., Pattuelli, M.C., Parsia, B., & Robertson, W.D. (2002). Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization. *Journal of Digital Information, 2*(2).

Greenberg, J., White, H. C., Carrier, S., & Scherle, R. (2009). A metadata best practice for a scientific data repository. *Journal of Library Metadata, 9*(3-4), 194-212.

Huang, L. (2010). Usability testing of HIVE: a system for dynamic access to multiple controlled vocabularies for automatic metadata generation. (Unpublished master's paper). The University of North Carolina, Chapel Hill, NC.

Jones, S. S., & Paynter, G.W. (2002). Automatic extraction of document keyphrases for use in digital libraries: Evaluation and applications. *Journal of the American Society for Information Science and Technology*, *53*(8), 653-677.

Lancaster, F.W. (2003). Automatic Indexing & Abstracting. In *Indexing and abstracting in theory & practice* (pp. 282-336). Champaign, Il: University of Illinois.

Liddy, E.D., Allen, E., Harwell, S., Corieri, S., Yilmazel, O., Ozgencil, N.E., Diekma, A., McCracken, N.J., Silverstein, J., & Sutton, S.A. (2002). Automatic metadata generation & evaluation. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, August 11-15, 2002, Tampere, Finland New York: ACM Press, 401-402.

Luhn, H.P. (1957). A statistical approach to mechanized encoding & searching of literary information. *IBM Journal of Research & Development, 1*. 309-317.

Medelyan, O., & Witten, I. H. (2006). Thesaurus based automatic keyphrase indexing. *Digital libraries, 2006. JCDL '06. Proceedings of the 6th ACM/IEEE-CS joint conference on Digital Libraries.* New York, New York, USA: ACM Press, 296-297.

Medelyan, O., & Witten, I. H. (2008). Domain-independent automatic keyphrase indexing with small training sets. *Journal of the American Society for Information Science and Technology, 59*(7), 1026-1040.

Miles, A., & Perez-Aguera, J. (2007). SKOS: Simple knowledge organization for the web. *Cataloging & Classification Quarterly, 43*(3/4)

National Evolutionary Synthesis Center. (2007, May 16-17th) "*Data Preservation, Sharing, and Discovery: Challenges for Small Science in the Digital Era*". A report of a workshop held May 16 -17th 2007 at NESCent, Durham, NC; Retrieved from https://www.nescent.org/wg/dryad/images/b/b0/Workshop2007FinalReport.pdf

Rieseberg, Vines, & Kane, (2010). Editorial and retrospective 2010. *Molecular Ecology, 19*, 1–22. doi: 10.1111/j.1365-294X.2009.04450.x

Rowley, J. (1994). The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. *Journal of Information Science, 20(2),* 108-119.

Salton, G. & MacGill, M.J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Sherman, J.K. (2010). Automatic metadata generation: a comparison of two annotators. (Unpublished master's paper). University of North Carolina, Chapel Hill, NC.

Strader, C. R. (2009). Author-assigned keywords versus library of congress subject headings: Implications for the cataloging of electronic theses and dissertations. *Library Resources & Technical Services, 53*(4), 243-250.

Tolle, K.M. & Chen, H. (2000). Comparing noun phrasing techniques for use with medical digital library tools. *Journal of the American Society for Information Science. 51*(4), 352-370.

Whitlock, M. C., McPeek, M. A., Rausher, M. D., Rieseberg, L., & Moore, A. J. (2010). Data archiving. *American Naturalist. 175*, 145-146.

Witten, I.H., Paynter, G.W., & Frank, E. (1999). KEA: practical automatic keyphrase extraction. *International Conference on Digital Libraries. Proceedings of the 4th ACM Conference on Digital Libraries*. ACM Press, 254-255.

Wright, M., Sumner, T., Moore, R., & Koch, T. (2007). Connecting digital libraries to eScience: the future of scientific scholarship. *International Journal on Digital Libraries, 7*, 1-4. DOI 10.1007/s00799-007-0030-9

Vision, J. (2010). Open data and the social contract of scientific publishing. *BioScience, 60*, 330-33.