Subject Metadata Enrichment using Statistical Topic Models

David Newman
Department of Computer Science
University of California Irvine
Irvine, CA
newman@uci.edu

Kat Hagedorn
University of Michigan Libraries
University of Michigan
Ann Arbor, MI
khage@umich.edu

Chaitanya Chemudugunta
Padhraic Smyth
Department of Computer Science
University of California Irvine
{chandra,pjsmyth}@uci.edu

ABSTRACT

Creating a collection of metadata records from disparate and diverse sources often results in uneven, unreliable and variable quality subject metadata. Having uniform, consistent and enriched subject metadata allows users to more easily discover material, browse the collection, and limit keyword search results by subject. We demonstrate how statistical topic models are useful for subject metadata enrichment. We describe some of the challenges of metadata enrichment on a huge scale (10 million metadata records from 700 repositories in the OAIster Digital Library) when the metadata is highly heterogeneous (metadata about images and text, and both cultural heritage material and scientific literature). We show how to improve the quality of the enriched metadata, using both manual and statistical modeling techniques. Finally, we discuss some of the challenges of the production environment, and demonstrate the value of the enriched metadata in a prototype portal.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; I.7.4 [**Document and Text Processing**]: Electronic Publishing; I.5.3 [**Pattern Recognition**] Clustering

General Terms: Algorithms

Keywords: topic model, metadata enhancement, metadata enrichment, clustering, browsing, OAI, digital libraries

1. INTRODUCTION

Digital libraries grow continuously. As the number of resources in these libraries increases, enabling users to easily discover these resources becomes a fundamental issue for digital librarians. Only by sustaining and ensuring uniformly high quality access to increasingly large collections of resources can digital libraries fully unlock the value of their collections.

The Open Archives Initiative (OAI) has developed an interoperability standard for sharing digital content, allowing digital libraries to increase the size of their collections. Through OAI's Protocol for Metadata Harvesting (OAI-PMH), digital libraries can harvest (gather) metadata about digital content—for instance pertaining to some particular subject area—to create virtual collections.

One example of a virtual collection is the American West Project at the California Digital Library (CDL)[1], a prototype portal of cultural heritage material focused on the American West, gathered from a dozen different research institutions. The National Science Digital Library[2] (NSDL) is another virtual collection created from other digital library collections. Creating these collections by harvesting metadata from dozens of other institutions was, perhaps, the easy part of the process. The greater challenge has proven to be finding ways to enhance this metadata to allow users to go beyond simple keyword search.

Indeed, any digital library that aggregates content from various sources faces the challenge of enhancing heterogeneous metadata—whether through normalization, transformation and/or direct modification—because enhanced metadata provides more uniform access for end-user discovery. And user accessibility to the collection is the key feature of a successful digital library.

The world's largest collection of OAI metadata is OAIster (pronounced "oyster"), at the University of Michigan[3]. OAIster, a union catalog of digital resources, harvests from over seven hundred OAI repositories (i.e. data providers). Because OAIster's collection policy is to harvest *all* of the OAI repositories in the world, unlike CDL's American West portal or NSDL (which have a scope and theme), OAIster has, perhaps, the widest subject variety of any digital library. Thus, creating consistent enriched subject metadata is one of the biggest challenges of the OAIster collection.

Previous efforts on enriching subject metadata have focused on smaller collections of records that relate to one particular subject area or discipline—such as the American West collection. Larger scale subject metadata enrichment has been achieved in collections of (usually scientific) academic literature, where metadata records typically contain a highly descriptive abstract. In contrast, large-scale subject metadata enrichment of cultural heritage and mixed material has remained a challenge.

Statistical topic modeling, a recently developed machine learning technique (e.g. [6]), has great potential for subject metadata enrichment. Topic models simultaneously discover a set of topics

JCDL '07, June 17-22, 2007, Vancouver, BC, Canada.

[1] www.cdlib.org/inside/projects/amwest/

[2] www.nsdl.org

[3] www.oaister.org

or subjects covered by a collection of text documents (or in our case, metadata records), and determine the mix of topics associated with each document (or record). These topic models are gaining wide popularity because they produce easy-to-interpret topics and can quickly and effectively categorize the contents of large-scale collections.

In this paper, we present the first large-scale application of statistical topic models to subject metadata enrichment of highly heterogeneous metadata. We start by assessing the interpretability of topics, and show that more than one quarter of the learned topics are unusable as enhanced subject headings. We address this issue by deleting from the vocabulary words that do not contribute topically. Removing these words results in improved topics and improved subject metadata enhancement. We then propose a modified version of the topic model that automatically removes words that have less topical value. While the focus of this paper is on the back-end metadata enhancement, we finally demonstrate how enriched subject metadata—produced by our statistical topic models—enables higher quality searching in a prototype portal developed for the Digital Library Federation. This provides a model for other digital library projects and shows the value of topic modeling to subject metadata enhancement for virtual digital library collections.

2. SUBJECT METADATA ENRICHMENT

This section describes several approaches for automatically enriching subject metadata. A large and diverse collection of metadata records contains a varying amount and quality of subject information (and sometimes none). In practice, subject fields often contain a mix of controlled and uncontrolled text. Automatic enrichment aims to attach uniform and consistent subject headings to every record in the collection by using the existing descriptive text in each metadata record. These additional subject headings constitute the enriched subject metadata.

Rexa (rexa.info), a digital library of computer science research, makes extensive use of information extraction and topic modeling algorithms to create and enhance metadata. In Rexa, one can browse papers by topic. But for this type of scientific literature content, the rich descriptive text available in abstracts makes for relatively straightforward topic modeling and application of learned topic labels to papers.

Other researchers have investigated enriching subject metadata for cultural heritage material, but on a much smaller scale. Krowne and Halbert [8] presented an evaluation of subject metadata enrichment methods to support digital library browse interfaces, using metadata from AmericanSouth.org. They considered the case of creating digital library portals from content harvested by OAI-PMH, and enhancing the subject metadata of the resulting heterogeneous collection of metadata records. They chose a machine learning framework based on non-negative matrix factorization [9] to learn the topics that were ultimately used to drive subject browsing. While they reported success using this approach for subject metadata enrichment, only relatively small-scale tests were performed. This study also highlighted the widespread inconsistent use of Dublin Core fields, and the need for uniform subject metadata.

More recent work has addressed cultural heritage subject metadata enhancement on a somewhat larger scale than that undertaken by Krowne and Halbert. The California Digital Library created the American West collection, made up of 250,000 metadata records harvested from a dozen diverse repositories. They investigated tools and services to enrich metadata to support hierarchical faceted browse. In addition to normalizing date and location facets, they used topic modeling to enhance the subject metadata. The project was instrumental in highlighting issues surrounding access to heterogeneous metadata, particularly for cultural heritage material[4]. While some issues were highlighted (e.g., the problem of many metadata records containing boilerplate text from originating institutions), this collection was more homogeneous than OAIster and on a much smaller scale.

Thus, there is a growing recognition of the need for enhanced subject metadata in virtual digital collections. Researchers have successfully used topic modeling to do enhancement, but to date, have concentrated primarily on smaller or more homogeneous collections.

2.1 The Topic Model

In this paper we use the topic model for subject metadata enrichment of the OAIster collection. The topic model, a recently developed unsupervised machine learning technique, learns a set of topics that describe a collection of documents. In the topic model, documents are represented as mixtures of topics, and topics are probability distributions over words. Both the topic-word distributions and the assignment of words in documents to topics are learned in a completely unsupervised statistical manner.

Topic modeling evolved from earlier techniques such as Latent Semantic Analysis [4] and document clustering [5]. Both these methods can be used to extract semantic content from large document collections. But their use for subject metadata enhancement is limited. In Latent Semantic Analysis, topic "dimensions" are required to be orthogonal. This constraint produces topics that are more difficult to interpret, and harder to distinguish from one another. Document clustering suffers from a different problem. In document clustering, each document is forced to belong to a single topic cluster. This requirement is too limiting (in reality, records naturally have multiple, not single, subject headings), and produces lower quality topics. Using a collection of 80,000 18th-century newspaper articles, Newman and Block performed a detailed comparison of Latent Semantic Analysis, document clustering and probabilistic topic modeling to show some of these limitations [13]. Further comparisons of these three methods are discussed in [14].

The topic model is a recent extension of earlier work on statistical modeling of word count data/document collections, such as Probabilistic Latent Semantic Indexing [7] and Latent Dirichlet Allocation (LDA) [1]. Topic modeling uses efficient Gibbs sampling techniques to learn the topic-word and document-topic probability distributions for a collection of documents [6]. The topic model is now widely used for extracting semantic content from large document collections [2,10]

The topic model automatically enriches (or enhances) subject metadata as follows: First, the topic model is run on the descriptive text in the metadata collection, producing a set of

[4] www.cdlib.org/inside/projects/amwest/cdl_clusteringOAI_final.pdf

learned topics. These topics are interpreted and manually labeled with a topic label (subject heading). The topic model assigns one or more topic labels to each metadata record in the collection, creating the enhanced subject metadata. These topic labels can then be mapped into the search system's subject classification hierarchy to allow subject browse and limiting of search results by subject.

3. THE OAIster COLLECTION

University of Michigan's OAIster—a union catalog of digital resources—gets its collection by harvesting from over seven hundred OAI data providers. These data providers are required to expose their metadata in Simple Dublin Core format. While Dublin Core is a widely-adopted standard, the interpretation and population of the fifteen Dublin Core elements is ultimately up to the providers creating the metadata. Some institutions have resources to ensure high quality metadata. OAI records harvested from the Library of Congress repository have, not surprisingly, highly uniform Library of Congress Subject Headings in the Dublin Core Subject element. However, many institutions incorrectly or inconsistently use the Dublin Core fields. The purpose of the techniques we describe in this paper is to provide more adequate access to this unreliable content—specifically the Dublin Core subject field.

OAIster has built a unique collection of over ten million records. OAIster's reach often goes beyond that of major web search engines. For instance, English-language metadata from one data provider—Xiamen University Library—while publicly available for harvesting, cannot be crawled (and thus cannot be found) by search engines such as Google.

OAIster allows keyword and fielded search but because the collection is so large, searches can return thousands of results, with minimal limiting and sorting options. We would like to improve the search and discovery experience on OAIster by allowing users to restrict search results by subject. This functionality is only possible if we have reliable, consistent and appropriate subject metadata for each of the ten million records in OAIster. OAIster's collection has quadrupled in size in three years --- thus scalability and sustainability are a major focus in our evaluations. We carefully assess the amount of human labor (accompanying our automated topic modeling techniques) necessary to produce quality subject metadata enrichment.

Every month, the University of Michigan's Digital Library Production Service (DLPS) harvests—using OAI-PMH—the entire contents of each repository discovered by OAIster. For the results presented in this paper, we used the 9/2/2006 harvest that contained approximately nine million records from 668 repositories. The type of repository in large part determines what type of descriptive text is found in metadata records. For instance, repositories of scientific literature usually contain records with a title and abstract, while archives of images often only contain a short image caption.

The ten largest repositories (by size in MB) from our 9/2/2006 OAIster harvest are listed in Table 1. This list of ten further illustrates the variety of content found in metadata repositories. Five of the ten contain primarily scientific literature (CiteSeer, PubMed, CiteBase, arXiv, Institute of Physics). Pangaea contains records that tersely describe geoscientifc data sets, with minimal description. And four of the ten (Highwire, PictureAustralia, University of Michigan Digital Library, Capturing Electronic Publications) contain principally cultural heritage material.

Table 1. Ten largest repositories in OAIster. This list of ten repositories (out of 700) includes five scientific literature repositories, one data repository, and four cultural heritage repositories, and shows the diversity of OAIster material.

Repository (type of metadata)	Description	Size in MB (no. records)
CiteSeer (science)	Scientific literature digital library	1106 (716772)
Highwire (cultural heritage)	Articles from 1000 journals	862 (995217)
PubMed (science)	National Library of Medicine digital archive	856 (715366)
PictureAustralia (cultural heritage)	Australiana images	758 (838983)
CiteBase (science)	Citation information for arXiv, etc.	600 (465428)
Pangaea (data)	Collection of geoscientific datafiles	582 (432507)
arXiv (science)	E-print archive of articles in physics and mathematics	477 (379344)
University of Michigan Digital Library (cultural heritage)	Digital collections at the University of Michigan	315 (308656)
Institute of Physics (science)	Journals from physics membership organization	266 (216498)
Capturing Electronic Publications (cultural heritage)	State documents from Illinois, Alaska, Arizona, Montana, etc.	235 (98626)

The metadata OAIster collects is in Simple Dublin Core format. In our subject metadata enrichment experiments, we used three of the fifteen Dublin Core elements: Title, Subject and Description. We determined (like Krowne and Halbert) that these three fields contained the bulk of the text relevant to determining the subject of a record. Words from the three fields were considered to be equally important because there was no way of knowing (in advance) from which field useful descriptive text might come. In theory, Dublin Core's Subject element should be the most relevant, but sometimes this field contains no text (and in that case we rely on text from the remaining two elements, Title and Description). Using the combined text from three Dublin Core elements reduces the problem of inconsistent use of individual elements. For example, some repositories routinely put content in Description that belongs in Subject. Since we combine the text from the three elements, this type of misuse does not affect our subject metadata enrichment.

3.1 Preprocessing

The input to the topic model is the so-called "bag-of-words representation" of a collection, in which every metadata record is represented by a sparse vector of word counts, i.e., a list of

{word-id, record-id, count} triples. We preprocessed the OAIster collection to produce the bag-of-words representation as follows: Starting with the 668 repositories in the 9/2/2006 harvest, we excluded 163 primarily non-English repositories, and 117 small repositories (containing fewer than 500 records), leaving 388 repositories. For each of these 388 repositories, the contents of Title, Subject and Description Dublin Core elements were tokenized (lowercase, punctuation removal, simple stemming), and bi-grams were identified using a t-test (e.g., "amino acid" was replaced by "amino_acid") as described in [12]. Identifying bi-grams is a straightforward preprocessing step that makes topics more interpretable. We augmented a standard English stopword list (*the, and*, *that, with*, etc.) with words that have little topic value, but occur frequently in metadata records. We found additional stopwords by scanning the list of frequently occurring words for words that did not relate directly to subject content, finding words such as: *volume, keyword, library, copyright,* etc.

The final bag-of-words representation of our processed OAIster collection included 7.5 million records, a 94,000 word vocabulary, and a total of 290 million word occurrences. This collection was too large to run on our system (the computation would have required 18GB of main memory (RAM), even using sparse data structures); consequently we used every third record producing a bag-of-words containing 2.5 million records and a total of 96 million word occurrences. Given the huge amount of data in this collection, it was reasonable to assume that using this smaller representation would still produce high-fidelity topics that represent the subjects spanned by all 7.5 million records.

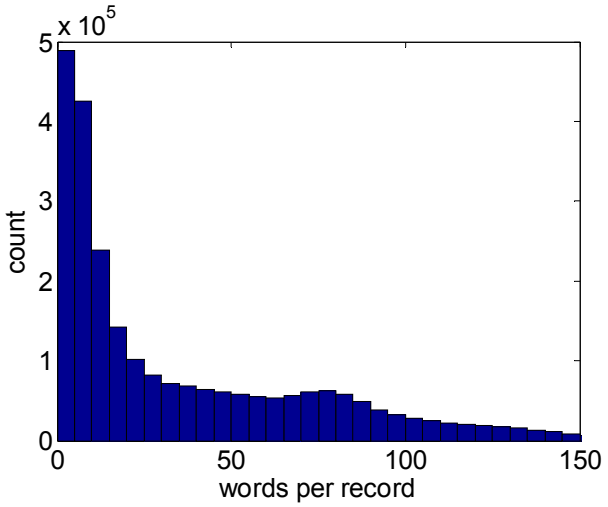

Figure 1. Histogram of words per processed record. On average, there are 38 words per processed record, but there are many records that contain fewer than 10 words.

The number of words in each record affects the topic model's ability to find interpretable topics, since the topic model works by finding patterns of co-occurring words. The distribution of words per processed record (Figure 1) shows that there were almost one million records with ten or fewer words. The mode around 75 words per record is from all the repositories containing metadata records with abstracts. On average, there were 38 words per processed record.

Once we had preprocessed the collection, we could begin to run experiments to evaluate the usefulness of the topic model for subject metadata enrichment.

4. EXPERIMENTS

We ran a total of three topic model computations, and evaluated their subject metadata enrichment performance. The first model (labeled the full-vocabulary model) was a baseline run, meaning that the preprocessing and topic model computation were performed in a standard way. The two subsequent models were designed to improve upon the performance of the full-vocabulary model. For all three topic models, we set the number of topics to be learned to 500. Based on past experience, we chose this number as a trade-off between avoiding topics that are too general to be meaningful and adding topics that are too specific to be of service to users.

4.1 Full-Vocabulary Topic Model

We ran the topic model on our collection of 2.5 million records (that has a vocabulary of 94,000 unique words) producing 500 topics. This computation required 6 GB of memory and 10 days of computing time on a 3GHz processor. Note that this is why we sampled one in three records—running the entire 7.5 million records would have required 18 GB of memory and 30 days of computing time. We note nonetheless, that the topic learning algorithm scales linearly in the total number of words.

4.1.1 Interpretation and Labeling of Topics

Four sample topics from this full-vocabulary model, selected from the 500 learned topics, are shown in Table 2. Each topic is shown as a list of words, in order of their likelihood (i.e., importance) in the topic. In each case, the list of words conveys a coherent theme or subject area, from "gene sequencing" through to "domestic architecture." While the full list of words defines a topic, a short topic label (i.e. subject heading) allows us to enhance searching in the production environment. Ideally a domain expert interprets the list of words to determine an appropriate topic label—in our case, our digital library colleagues labeled topics. The resulting topic labels/subject headings could then be made available to end-users to limit search results by subject.

Many topics (including the topics shown in Table 2) clearly define subject areas, and the corresponding topic labels can be used as subject headings during metadata enrichment. However, some learned topics are less interpretable and thus less useful. The three topics shown in Table 3 are unusable for subject metadata enrichment. The first topic is unusable because the "size" concept conveyed by the words is not sensible as a subject heading. The second topic is in Spanish (and in our case not usable; while the topic model does work in other languages, for this study, we limited ourselves to English language topics). Finally, the third topic shows some possibility, but is ultimately not usable. Some of the words are thematically about street scenes, but the topic is polluted by specific words such as Santa Ana and Orange (two cities in California).

Table 2. Sample topics from full-vocabulary topic model. The list of words in each topic relate to a particular subject area. A human interprets each list of words to determine an appropriate topic label.

Words in topic	Topic label
gene sequence genes sequences cdna region amino_acid clones encoding cloned coding dna genomic cloning clone	gene sequencing
social cultural political culture conflict identity society economic context gender contemporary politic world examines tradition sociology institution ethic discourse	cultural identity
general_relativity gravity gravitational solution black_hole tensor einstein horizon spacetime equation field metric vacuum scalar matter energy relativity	relativity
house garden houses dwelling housing homes terrace estate home building architecture residence homestead residences road cottage domestic fences lawn historic	domestic architecture

Table 3. Example topics from full-vocabulary topic model that are unusable as subject headings.

Words in topic	Usefulness
large small size larger smaller sizes scale sized largest	Reasonable but unusable
foi para pacientes por foram dos doen resultados grupo das tratamento entre	Topic about patient treatment, in Spanish
building street visible santa_ana view avenue public_library front orange corner	Not usable: mix of concept words and specific geographic location words

From the 500 topics produced by the full-vocabulary topic model, 352 topics were interpreted to be about a particular subject area and given topic labels. These topics were then marked as usable for subject metadata enhancement. The remaining 148 topics were deemed unusable (for subject metadata enhancement). We thus set out to increase the number of usable topics, because this directly affects the number of metadata records enhanced. Increasing the yield and quality of usable topics was the focus of the subsequent topic models, and is described in Sections 4.2 and 4.3.

4.1.2 Assigning Topics to a Record

Recall that the topic model simultaneously learns a set of topics *and* assigns a topic label to every word in every record. By counting the number of times each topic label occurs in a record, we can list, in order of proportion, the topics assigned to that record. An example of topics assigned to a record is shown in Table 4. The table shows part of a metadata record from Australian National University's DSpace repository. This record describes an article about the theory of choice and voting. Almost half of the words in the record (after preprocessing) are assigned to the topics of "game theory," "argument," "criteria" and "voting." Note that unusable topics (described above) are never assigned to records, even if they are high probability (i.e. proportion), so some records may be left with fewer than four assigned topics.

Table 4. Topics assigned to a selected metadata record. The topic model determines that this record is mostly about the topics: game theory, argument, criteria and voting.

Metadata record	Topic labels (% words assigned)
Aggregating sets of judgments: two impossibility results compared. (C. List and P. Pettit) May's celebrated theorem (1952) shows that, if a group of individuals wants to make a choice between two alternatives (say x and y), then majority voting is the unique decision procedure satisfying a set of attractive minimal conditions ...	game theory (21%) argument (12%) criteria (7%) voting (6%)

After assigning these topics to this record, an end-user can find it by using any of the four topic labels (subject headings) in a search system. An end-user can also browse a subject area and view all records assigned to a particular topic. Table 5 shows the ten most relevant records in the "game theory" topic. The power of topic modeling is that it allows users to access records across the institutional boundaries of individual repositories; in Table 5 the top ten records come from five different repositories.

Table 5. Top ten records in the "game theory" topic. These ten records come from five different repositories.

game theory	game games equilibrium preferences player cooperative preference equilibria cooperation collective utility individual choice bargaining coalition nash strategy
Top 10 records	1. Fundamental Components of the Gameplay Experience: Analysing Immersion 2. The Ethics of Computer Game Design 3. Backward Induction and Common Knowledge 4. Designing Puzzles for Collaborative Gaming Experience 5. Aggregating sets of judgments: two impossibility results compared 6. Games for Modal and Temporal Logics 7. Configuring the player - subversive behaviour in Project Entropia 8. From Mass Audience to Massive Multiplayer: How Multiplayer Games Create New Media Politics 9. Bargaining with incomplete information; Handbook of Game Theory with Economic Applications 10. Testable Restrictions of General Equilibrium Theory in Exchange Economies with Externalities

4.1.3 Qualitative Analysis of Assigned Topics

We qualitatively assessed how well the records were described by the topics assigned to them. We looked at 75 randomly chosen records, and answered three questions for each record (Table 6).

1. What fraction of the four topics assigned to a record were appropriate? Both full and partial descriptions of the record by the topic were accepted. For example, a record about medieval architecture that had an architecture topic assigned to it was accepted.
2. Were the records science-based or humanities-based? Previous experimentation for the Metadata Enhancement OAI Workshop[5] showed that science records were described better overall by topics than humanities records.

[5] www.metascholar.org/events/2006/meow/

3. Did short records being described by minimal metadata (e.g., a title, a note, a URL) receive fewer appropriate topic labels? With insufficient metadata, the topic model may have had difficulty in assigning appropriate topics.

Table 6. Analysis of topic assignments to 75 randomly selected records for full-vocabulary topic model.

Question	Analysis
Number of appropriate topics	56% of records had 2 or more appropriate topic labels 25% of records had 1 appropriate topic label 19% of records had no topic labels assigned
Humanities vs. science	47% of humanities records had appropriate topic labels 83% of science records had appropriate topic labels
Short records (< 10 words)	4 short records had 3 or 4 appropriate topics assigned 11 short records had 1 or 2 appropriate topics assigned

As a point of clarification, "25% of records had 1 appropriate topic label" means that the other (up to three) topic labels in those records were not appropriate, and "no topic labels assigned" results from the highest probability topics being unusable topics, and therefore not assigned. In summary, the majority of the time, the description of records by topic assignments was accurate. On average, a smaller percentage of humanities records had appropriate topic labels than science records, likely due to the fact that humanities records have less metadata. Short records also suffered the same fate of generally having fewer appropriate topic labels.

4.2 Reduced-Vocabulary Topic Model

For the full-vocabulary topic model, we performed relatively basic preprocessing. Despite the many idiosyncrasies of particular repositories, we did no repository-specific preprocessing. The full-vocabulary model yielded 352 usable and labeled topics out of the 500 topics learned by the topic model. A straightforward way to improve the yield and quality of usable topics is to remove from the vocabulary words that do not significantly contribute to the topics. Words contribute little either because they are topically too broad, or topically too specific. Examples of topically broad words include *january*, *february* (months of the year) and *result, paper, study* (words often used in research articles). Examples of topically specific words include *santa_ana* (city in California) and *ladies_repository* (the name of a historic periodical).

We deleted such unwanted words from the full vocabulary using two methods. The first method eliminated unwanted words by reviewing topics. We built a browser-based tool that allowed reviewers to examine topics, and click on unwanted words that were degrading the topic. For example, in the third topic of Table 3, we marked *santa_ana* for deletion because it is a specific geographic location. The second method eliminated unwanted words by reviewing high-frequency words in each of the 388 repositories. Through this review we eliminated word tokens such as *ladies_repository* (the name of a periodical) and *repec* (the name of a repository). Finally, we removed words overlooked in our initial stopword list, such as *jpg* and some non-English words (which were neatly clustered together into topics).

Using primarily manual techniques, we deleted a total of 12,000 topically uninteresting words from our initial vocabulary of 94,000 words, leaving a revised vocabulary of 82,000 words. While this reduction in vocabulary caused the average number of words per record to decrease from 38 to 26, our hope was that the better quality of the words would produce more usable topics, better quality topics, and better topic assignments to individual records.

We assessed the effect of the reduced vocabulary by finding matching topics between the full-vocabulary model and the reduced-vocabulary model, and evaluating improvement in topic interpretability. An illustration of this is given in Table 7. This topic—about family photographs—becomes more refined, interpretable and usable. In the full-vocabulary model, we see words such as *george_edward* and *anderson_photograph* degrading the topic. These specific words degrade the topic because they don't help define the subject area described by the entire list of topic words. In the reduced-vocabulary model, these unwanted words have been omitted from the vocabulary, resulting in a much clearer topic about family photographs, clothing and dress.

Table 7. Comparison of single topic between full-vocabulary model and reduced-vocabulary model. The reduced-vocabulary model version of the topic is clearer and more interpretable.

Model	Words in topic
Full-vocabulary	family_photograph mss jpg george_edward anderson_photograph plate_negative women_portrait gelatin_dry photograph_portrait south_africa studio_portrait children_portrait hair standing sitting portrait underwood portrait_portrait front infant_portrait
Reduced-vocabulary	family_photograph wearing woman hair dress clothing shoulder baby suit dressed chair clothing_dress wear hand tie shirt jacket costume boy ribbon collar dark lap bow white full_face beard young_woman leaning striped outdoor

The reduced-vocabulary topic model increased the number of usable topics from 352 to 412. Furthermore, the topics themselves were generally clearer and more easily interpreted. Our manual reduction in vocabulary led to a marked improvement. But a major concern with any manual process is scalability, and this is addressed by a new model we developed (explained in the next section).

An important point is that users still can search for these specific words that have been deleted. We emphasize that these words are removed *only* for topic modeling purposes, and will not show up in the topic representations. But these words still exist in the metadata record, and can therefore be found using keyword search.

4.3 Background-Words Topic Model

We introduce an extension of the topic model called the background-words topic model. The idea behind this model came from the manual process undertaken in the reduced-vocabulary model, in which we identified and deleted some high-frequency words from each repository. We developed an algorithm that modeled a collection of *sub-collections*. In this model, a word in a record is either generated from a shared topic across sub-collections, or a background topic (distribution) from that record's sub-collection. Words in a background distribution tend to occur in records in the sub-collection, and tend *not* to occur in records

outside the sub-collection. The output of the background-words model were shared topics (topics shared by all sub-collections) and background distributions (specific to each sub-collection). Our expectation was that these shared topics would be more interpretable (and useful as subject headings).

The background-words model is a variation on the special words model developed and presented by Chemudugunta, et al. [3], in which the model learns words special to one record, words to be used in shared topics, and words generic to the entire collection. The background-words model has a pair of parameters that sets a prior probability that a word belongs to a background distribution. The first parameter sets the expected fraction of words that belong to a background distribution, and the second parameter sets the strength of this prior belief. For our purposes, background words were usually words related to the originating institution of the metadata record (such as *York gift*), archival terms (such as *bound volumes* or *artifact material*), or other source-type descriptors (such as *jpg* or *pixel*; or a place of publication) that don't relate directly to the subject matter.

To see how well the background-words model replicated our manual process of deleting words that were specifically related to individual repositories, we ran the full 94,000-word vocabulary bag-of-words data set. The results are shown in Tables 8 and 9. Table 8 shows selected background distributions for three repositories. Most of the words in these background distributions are, indeed, very specific to their respective repository, and have little topical value for shared topics. Note that some words (e.g., *civil_war* from Library of Congress) should belong in shared topics (if other repositories have material on the Civil War). As in the topic model, words (such as *civil_war*) can belong in multiple topics, or in the case of the background-words model, both shared topics and background distributions.

Table 8. Background words found by the background-words model for three repositories. Generally, these words have lower topical value. The grey font color shows words that also occur in shared topics.

Repository	Background words
University of Michigan	moa view son europe detail artifact_function mich vol building house architecture artifact_material art periodical_devoted ladies_repository literature_art ann_arbor
Library of Congress	piano_music civil_war home york bound_volumes washington negative_sleeve york_york gift_episcopal view song_piano district_columbia portfolio_folder printed_ephemera detroit_publishing
University of Chicago	university_chicago building_ground archives_photographic chicago_illinois department_botany american_environmental record lantern_slides county exterior_view files building_sites river_valley interior_view construction_sequence images_pixel photographic_print

The background-words model clearly improved the interpretability of many topics. One such example is the full-vocabulary model's topic about music and song (Table 9). In the full-vocabulary model, words specific to particular repositories (*moa periodical_devoted ladies_repository literature_art nla_mus*) and geographic locations (*music_australian south_america peru*) are in the topic. The version of this topic from the background-words model is cleaner, more interpretable, and more usable across a wide range of records.

Table 9. Comparison of "music" topic between full-vocabulary model and background-words model. The shared topic from the background-words model is more interpretable.

Model	Words in topic
Full-vocabulary	music moa periodical_devoted ladies_repository literature_art song musical mus gov_nla nla_mus music_australian cover instrument piano musician south_america voice_piano drum peru song_piano
Background-words model	music poster musical dance song theatre instrument actor concert entertainment theater piano sound festival theatrical musician drama performances opera art

The background-words model did provide advantages over the full-vocabulary model, and is promising as an automated and scalable alternative to the labor-intensive effort required for the reduced-vocabulary model. By comparing multiple aspects of the three models we can analyze the advantages of a fully automated versus a manually-reduced vocabulary metadata enhancement system.

4.4 Comparison of the Three Models

Table 10 shows that the number of usable topics increased from 70% (352 out of 500 topics) for the full-vocabulary model to 83% for the reduced-vocabulary model. The background-words model resulted in an in-between yield (76%) of usable topics. Beyond the results in this table, the clarity and interpretability of topics from the reduced-vocabulary and background-words models were uniformly better than those from the full-vocabulary model. We were not surprised to see the percent of usable background-words topics fall between the other two models. Recall that in the reduced-vocabulary model words were removed either because they were polluting otherwise clean topics, or because they were words specific to a repository and otherwise topically uninteresting. The background-words model addresses only the second set of words that are removed—those specific to a repository and topically uninteresting—and therefore results in half the improvement.

We saw similar differences in the three models for the percent of records enhanced. Only usable topics were kept for subject metadata enrichment, and topics that accounted for fewer than 5% of the words in a record were suppressed. This threshholding reduced the rate of false-positive labeling in which a record received a topic label, but the topic was not particularly relevant to that record. The proportion of records receiving at least one topic label increased from 92% for the full-vocabulary model to 99% for the reduced-vocabulary model. Again, the background-words model (with 96%) fell in between the full-vocabulary and reduced-vocabulary models.

We also compared the average coverage provided by the top four topic labels (i.e., the percentage of words in a record that received one of its top four topic labels). A higher percentage indicates that the topics are more efficiently describing records (if every word in a record is assigned to the top four topics, then the coverage is

100%). The coverage in the full-vocabulary model was 55%. This coverage increased to 68% for the reduced-vocabulary model, and again, the background-words model fell in between at 62%.

Table 10. Comparison of the three topic models. The Reduced-vocabulary model and Background-words model outperform the Full-vocabulary model.

	Full-vocabulary	Reduced-vocabulary	Background-words model
% usable topics	70%	83%	76%
% records enhanced	92%	99%	96%
Average coverage by top 4 topics	55%	68%	62%

4.4.1 Comparison of Topics

Examining some individual topics across the three models shows how the various models performed. All three models learned a topic about navy ships and vessels (Table 11). In the full-vocabulary model, the topic is polluted by geographic-specific words (*darwin)*, type words (*side_view*) and repository-specific words (*evan_antoni*). The reduced-vocabulary model omits these unwanted words from the topic (since they have been deleted from the vocabulary). The background-words model considers these words background words specific to a collection, and, as in the reduced-vocabulary model, they don't appear as highly relevant words in the topic. This is a good example of the background-words model doing exactly what we expected it to do. We point out that *all* these words still exist in the metadata record, so no searchable information has been lost.

Table 11. Comparison of topic about navy ships/vessels. Words common to all three models appear in grey, highlighting less relevant words.

Model	Words in topic
Full-vocabulary	darwin hmas naval navy world_war ship ran port sea naval_historical cruiser side_view vessel note hm aerial destroyer gun submarine deck starboard sailor bow crew aboard harbour fleet photograph_picture evan_antoni
Reduced-vocabulary	ship hmas navy naval vessel world_war ran naval_historical sea cruiser port deck gun destroyer submarine sailor fleet merchant starboard officer aboard inch gulf bow brisbane warship aerial patrol
Background-words model	ship port hmas vessel gun navy ran naval_historical naval note view destroyer australian cruiser world_war aerial starboard submarine

While some topics improved from the full-vocabulary model to the two subsequent models, others topics were fairly static over the three models. An example of such a topic (about antigens) is shown in Table 12. Because there were few repository-specific words in these topics, and all the words contributed to the conveyed subject area, neither the reduced-vocabulary model nor the background-words model greatly affected what was already a fairly coherent topic.

Table 12. Comparison of topic about antigens.

Model	Words in topic
Full-vocabulary	antigen anti antibody antibodies serum sera igg immunoglobulin rabbit complement mab elisa igm reactivity human monoclonal_antibodies
Reduced-vocabulary	anti antigen antibodies antibody serum sera igg immunoglobulin human mab complement elisa igm monoclonal_antibodies assay rabbit
Background-words model	antigen anti antibodies antibody serum sera igg immunoglobulin mab complement elisa igm rabbit monoclonal_antibodies

5. PROTOTYPE PORTAL

Our final step in this research was to deploy our topic modeling metadata enrichment techniques on a prototype search service that uses the assigned topic labels as subject headings. While the emphasis of this work is on the (back-end) metadata enrichment, we briefly demonstrate how the enriched metadata is used in a live portal.

The University of Michigan Libraries Digital Library Production Service (DLPS)[6] specializes in creating digital library services for over 250 collections, including OAIster. The DLF Portal[7] (a sub-set of OAIster) was developed by DLPS for use during the DLF/Institute for Museum and Library Studies (IMLS) grant period to both provide access to all OAI records available at DLF institutions and to test new functionality developed during the grant period. When new functionality was available, we presented and discussed it with our grant colleagues, our scholar board, and other interested parties. As part of this grant we integrated the topic model output into the DLF Portal.

5.1 Subject Hierarchy

From practical experience within DLPS, we knew that navigating through 500 topics is a burden to place on the end-user during search. Consequently, we needed to map the topic labels to a more manageable subject hierarchy.

We mapped our learned topics (topic labels) into a classification system designed at the University of Michigan Libraries called High-level Browse. This classification system, which was designed for access to Electronic Journals[8], consists of 8 top-level categories, and nearly 100 second-level sub-categories. The predefined sub-categories were then manually linked to the automatically-learned topics to provide a 3-level hierarchy as depicted in Figure 2. The topics serve as automatically-created subject headings for the full collection of 2.5 million records, since each record has a topic distribution assigned to it during the Gibbs sampling process. In this manner the topic model provides the "semantic glue" to link high level subject hierarchy with individual records.

[6] www.umdl.umich.edu

[7] www.hti.umich.edu/i/imls/

[8] www.lib.umich.edu/ejournals/

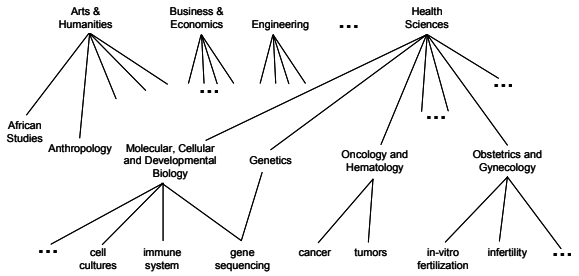

Figure 2. Subject Hierarchy. Topics learned by the topic model are located (lowest level) under the High-level Browse Categories (top-level) and Sub-categories (mid-level) already in use at the University of Michigan Libraries.

5.2 Incorporating the Enhanced Metadata

The software used by DLPS is our own in-house Digital Library eXtension Service (DLXS), www.dlxs.org. For the prototype we modified this software to incorporate the enhanced metadata.

1. Each metadata record was assigned (up to) four topics using a newly created tool. This tool output one file per repository (for each of the 61 in the DLF Portal) that listed record numbers and the topics assigned to them.
2. The tool that transforms OAIster metadata from Simple Dublin Core to our native DLXS Bibliographic Class was modified so that it could ingest the file from the first step, and output a transformed metadata record.

The transformed metadata record now contained three new fields: topic label, High-level Browse Sub-category, and High-level Browse Category. Our final step was to re-engineer the DLXS software to recognize these new fields in the search interface, and re-design this interface to accommodate the changes. Searching using assigned topics incurs no additional cost at time of search – the topics are just additional metadata for each record. The primary computational cost occurs offline when learning the topic model.

5.3 Search Interface

Our search interface (Figure 3) provides two drop-down menus where an end-user can "Select a Topic" that corresponds to the High-level Browse Categories, and, if desired, a "Sub-Topic" that corresponds to the High-level Browse Sub-categories. The system uses these search limiters in conjunction with the word or word(s) entered in the top search boxes to find results.

As a result of including the topic labels in the records, a search by an end-user on a topic label will retrieve all the records containing that label. For instance, if an end-user searches on "gene sequencing," all records containing that label will be retrieved. We ultimately plan to use a three-level hierarchy of subject headings in the search interface, which would provide a more granular way of retrieving relevant records (users would be able to select a topic label via a drop-down menu).

The search results (Figure 4) were designed to show the end-user which topic they had searched within (the white highlighted row in the left column under "Results by Topic"), and additionally that they could revise their search (either expand or restrict) by choosing a different topic.

In Figure 4, the end-user retrieved 239 records within the "Engineering" topic as a result of the search—the records are listed on the right. If the end-user revised the search by choosing the "Science" topic in the left column, the user would see the 347 records found within that topic for the search query "conductor*".

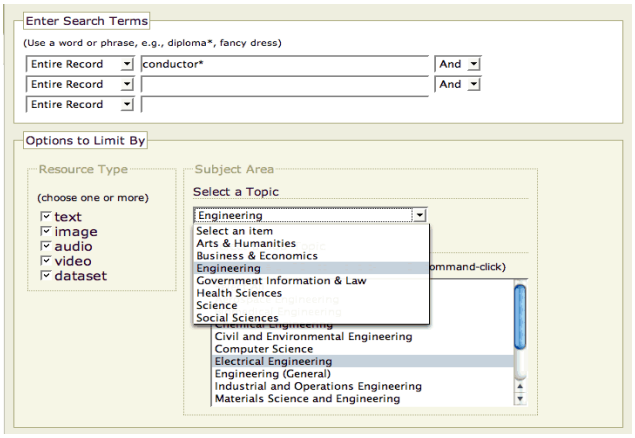

Figure 3. Search interface on DLF Portal. The topic model creates the enhanced subject metadata that allows searches to be limited by the selected subject area.

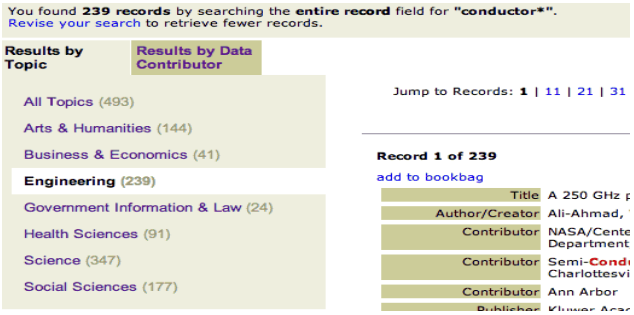

Figure 4. Search result on DLF Portal. The ability to revise search results by topic is enabled by the enhanced subject metadata.

Beyond using topics to simply limit keyword search results (i.e. only displaying results that match the keyword search and topic label), there are approaches that probabilistically combine general topics and specific keywords, as described by [3]. Furthermore, the topic model allows additional features such as annotating individual words in records by topic, and supporting search of similar records given a selected record (i.e. "Show me similar records").

6. DISCUSSION

"Better search comes from better metadata" remains a widely-held belief in the digital library community. But metadata often comes as-is, and many institutions don't have the resources to improve metadata, either through manual intervention or retrospective cleanup. Automated metadata enhancement techniques, such as topic modeling to enhance subject metadata, offer scalable solutions that start to address the difficulty of enriching metadata.

Our efforts to enhance subject metadata for a wide range of OAI records revealed many issues. While our topic modeling approach is statistical, and can handle some degree of noise, we found that

improved preprocessing of metadata records produced better results. Creating individual preprocessing rules for each repository in the collection is not a scalable solution for OAIster, or any other large metadata collection. However, it is possible we could improve results by creating individual preprocessing rules for larger repositories, or for sets of similar repositories (e.g., DSpace), or specifically for repositories that primarily contain humanities records.

The best way to measure quality is through user testing—in the results presented we have not yet performed such tests. Nonetheless, our human review of a limited number of records did demonstrate that the additional subject headings assigned by our topic models were generally appropriate and useful, and that some manual editing of stopword lists increased the quality of the topics assigned. Our new background-words model gave improved results, but it had the handicap of working with the original vocabulary that was cluttered with many topically uninformative words. While it may have been interesting to run the background-words topic model with the reduced vocabulary, this would not have been a fair comparison because the reduced vocabulary had *already* deleted repository-specific words.

We suggest that the overall best result will come from using the background-words model, along with human tailoring of the vocabulary. This type of hybrid approach, where the model takes over some of the labor-intensive work, ultimately provides a more scalable solution.

7. CONCLUSION

Generating uniform subject metadata for millions of records from hundreds of repositories is difficult. However we have shown that automated methods, such as the topic model, augmented with human review and intervention, can produce good results. This paper presents an approach in which humans intervene during the process when the added value is high and the human labor cost is bounded, thus offering a scalable approach.

We show that removing topically low-value words from the vocabulary improves the interpretability of appropriate topics, increases the yield of usable topics, and improves the assignment of topics to metadata records. While this produces a good result, the manual process of removing words is not scalable to increasingly large collections.

To address this scalability problem, we developed and present a new model, called the background-words topic model, specifically designed for collections of repositories. The background-words topic model identifies words that are likely to be specific to a repository and have little meaning or relevance outside that repository. The background-words topic model computes shared topics (relevant to all records across the entire collection) and a background distribution specific to each repository in the collection. We found these shared topics to be better than the topics from the standard topic model. Using the background-words model may ultimately result in less human labor required to get high quality and high yield subject metadata enhancement. Furthermore, this type of automated modeling approach is highly scalable.

Finally, we demonstrate the practical usefulness of the topic modeling approach by performing metadata enhancement in a production environment, and deploying a live search portal on a subset of OAIster.

8. ACKNOWLEDGMENTS

We thank the Andrew W. Mellon Foundation, the Institute for Museum and Library Studies (grant no. LG-02-04-0006-04), the Digital Library Federation, and the entire Digital Library Production Service and Scholarly Publishing Office departments at the U of Michigan Libraries for their support. In particular we would like to thank Perry Willett, Suzanne Chapman, Phil Farber, Christina Powell and Kevin Hawkins. Authors DN, CC and PS were supported in part by the National Science Foundation under grants IIS-0083489 and SCI-0225642.

9. REFERENCES

[1] Blei, D., Ng, A., Jordan, M. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993-1022, 2003.

[2] Buntine, W., Lofstrom, J., Perki, J., Perttu, S., Poroshin, V., Silander, A Scalable Topic-Based Open Source Search Engine. *In IEEE/WIC/ACM International Conference on Web Intelligence*, 228-234, 2004.

[3] Chemudugunta, C., Smyth, P., Steyvers, M., Modeling general and specific aspects of documents with a probabilistic topic model. *In NIPS'06, Advances in Neural Information Processing Systems 19*,. 2006.

[4] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., Harshman, R. A. Indexing by latent semantic analysis. *JASIS*, 41(6):391-407, 1990.

[5] Dhillon, I.S., Modha, D.S., Concept decompositions for large sparse text data using clustering. *Machine Learning.* 42:143-175, 2001.

[6] Griffiths, T., Steyvers, M., Finding Scientific Topics. *PNAS,* 101(suppl. 1):5228-5235. 2004.

[7] Hoffman, T. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42, 177-196, 2001.

[8] Krowne, A., Halbert, M. An initial evaluation of automated organization for digital library browsing. *Joint Conference on Digital Libraries*. pp246-255. June 7-11, 2005

[9] Lee, D., Seung, H. S., Learning the parts of objects by non-negative matrix factorization. *Nature*, v.401, 788-791, 1999.

[10] Li, W., McCallum, A. Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations. *In ICML'06*. 2006

[11] Mann, G. S., Mimno, D., McCallum, A. Bibliometric impact measures leveraging topic analysis. *Joint Conference on Digital Libraries.* pp65-74. June 11-15,. 2006.

[12] Manning, C., Schutze, H. Foundations of Statistical Natural Language Processing, *MIT Press*. Cambridge, MA: 1999.

[13] Newman, D., Block, S. Probabilistic Topic Decomposition of and Eighteenth Century Newspaper. *JASIST*, 57(6):753-767, 2006.

[14] Newman, D., Chemudugunta, C., Smyth, P., Steyvers, M. Analyzing Entities and Topics in News Articles Using Statistical Topic Models. *In LNCS -- IEEE Conference on Intelligence and Security Informatics.* pp93-104. San Diego, 2006.