

# Metadata Enhancement Feasibility Study

Final Report

topicSeek and CDL  
5/23/2005

## Executive Summary

This feasibility study showed that topical clustering using the topic model<sup>1</sup> is useful for automated metadata enhancement. The topic model was run on 360,000 OAI-harvested objects related to the American West (AW). Three Dublin Core elements were extracted for clustering: Title, Subject and Description. We performed several iterations of the cycle: i) preprocess, ii) run model, iii) examine topics, and iv) modify preprocessing rules and expand stopword list. The final run of 400 topics produced more than 300 usable and interpretable topics, enhancing subject metadata for more than 80% of the 360,000 objects. We expect this proportion of usable topics and objects with enhanced metadata to increase as we further refine preprocessing rules and expand stopword lists. We also conclude that the OAI metadata is sufficiently rich to generate meaningful topics. Topic modeling was also useful in identifying non-American West objects. We conclude with an exploration of three options for how CDL can proceed using topic modeling.

We make a variety of recommendations. Highlighted below are the ones that focus on moving forward quickly with enriching topical metadata in American West Project digital objects. The complete set of recommendations follows Section 6 of this report.

- **Remove non-American West objects.** Out of 360,000 objects, there may be as many as 30,000 non-American West objects. Deleting these and rerunning the model would improve the topics.
- **Improve system for managing stopwords.** Removal of stopwords clearly affects the quality of the clustering. As the number of stopwords increased from the hundreds to the thousands, there was a greater need to be able to systematically manage this list. There are several categories of stopwords including names of people, geographic locations, form/genre/type terms, etc. We recommend that a separate list be created and managed for each category, and these lists be managed in a database.
- **Cluster the "production" American West Project collection using an expanded stopword list.** We saw for the EAD finding aid collection, a large improvement in topics by continually expanding the list of stopwords. This was not done to the same extent on the American West collection because of its larger size, and the time constraints of this feasibility study. We recommend a more extensive iterative analysis of the "production" American West Project collection (current plans call for this collection to be finalized by mid-September 2005, so that interface development can move forward) and a further expansion of the stopword list.
- **Try sub-clustering.** Several topical clusters emerged from this experiment around themes about *Buildings* (13 clusters), *Cities and towns* (9 clusters), *Dams* (7 clusters), *Japanese Americans -- Evacuation and relocation, 1942-1945* (7 clusters), *Logging* (5 clusters), and *United States -- Politics and government* (5 clusters). In these specific cases, re-running the topic model for the metadata records in these subsets of the American West Project collection would provide a more focused separation of sub-topics under these broader topical headings.
- **Begin development of a prototype classification utility to be used to automate the enhancement of American West metadata objects during the ingest process.** Since the American West Project is aiming to have a "production" version of its collection available for UI development by September 15, 2005, we recommend beginning the drafting of specifications for automating the hardcoding option (see Section 5 of this report).

---

<sup>1</sup> Topical clustering is the extraction and automatic summarization of topical information from large collections of text documents. There are several methods to compute topic clusters. The topic model and other methods are discussed in the article "Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper" by D.J. Newman and S. Block, forthcoming in *JASIST (Journal of the American Society of Information Science and Technology)*. A copy of an earlier draft of this article is available to CDL staff at [https://diva.cdlib.org/groups/tech\\_solutions/resources/topic\\_decomposition/NewmanBlockTopicDecomp.doc](https://diva.cdlib.org/groups/tech_solutions/resources/topic_decomposition/NewmanBlockTopicDecomp.doc)

## 1 Introduction

During this feasibility study we investigated the application of topical clustering to a collection of 360,000 American West-scoped harvested metadata objects. Topical clustering is the process of automatically determining the topics covered by a collection of text documents, without *a priori* topic definitions. Topics are automatically found by statistically determining words that tend to co-occur, and computing lists of the most likely words in each topic.

This report gives preliminary conclusions on the usefulness of topic modeling for automatically producing topical metadata, and makes recommendations on how CDL might proceed.

### 1.1 Objectives

The objectives of the study were to see if the topic model can:

- a) produce interpretable, sensible and useful topics, and
- b) produce useful and appropriate topical metadata (i.e. a topical decomposition) for every object in the collection.

### 1.2 Success Metrics

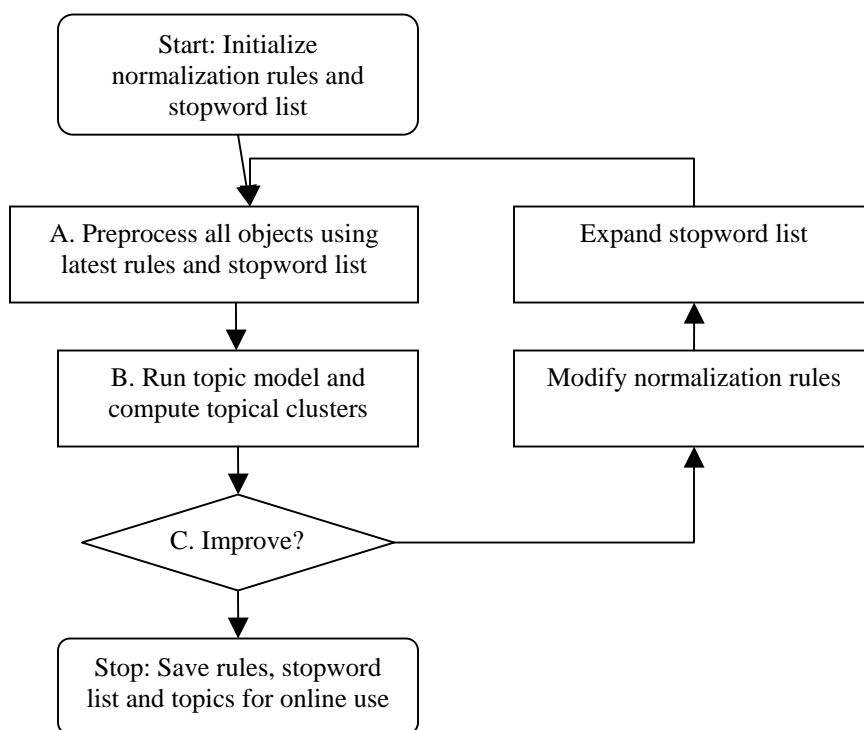
The success of the feasibility study will be based on the metrics in Table 1.

Metric	Comment
Number of usable topics	Topics are usable if they are coherent, interpretable and able to be named. They also need to be a valid subject area. The number and overall quality of usable topics can be increased by refining preprocessing rules and expanding the stopword list.
Number of objects with topic labels assigned (i.e. enhanced subject metadata)	If, for example, 1 in 4 topics are not usable, how many objects will not have any topics assigned?
Replicability / Reproducibility	Needs to give the same results when re-run.
Scalability	Needs to be able to run on large collections in reasonable time.
Extensibility	Needs to run on new (unseen) objects, e.g. objects from additional harvest sets.
Simplicity	Entire procedure needs to be simple to understand, with clear, straightforward processing rules.
Degree of automation	Manual processes should be minimized. System needs to generate report of what was done (e.g. report number of objects processed, number of enhanced objects).
Conformance with CDL practices	Any envisioned implementation of clustering must be able to comply with/conform to CDL's technical environment and practices.
Usefulness of information	Document vectors, a byproduct of topical decomposition, may be useful independent of the topical cluster metadata labels.

*Table 1.* Success metrics for metadata enhancement feasibility study.

## 2 Modeling Procedure

Topical clustering using the topic model is an iterative procedure. We start with some initial preprocessing and normalization rules and an initial generic stopwords list. We then run the topic model and examine the results. The preprocessing and normalization rules are then refined, and the stopwords list expanded to improve the topics. Figure 1 shows a diagram of this iterative procedure. The details of the preprocessing steps are given in Table 2. Table 3 and Table 4 show examples of the preprocessing of an OAI-harvested metadata object, showing how the original XML gets converted into a list of normalized extracted words, which are then converted to a vector representation.



*Figure 1.* Flowchart of modeling procedure showing how preprocessing rules and stopwords list are determined iteratively. This is the offline or batch version of the procedure. After rules, stopwords list and topics are saved, this object classification process can be used in an online and automated process.. Step A is performed with freely available Perl scripts developed by topicSeek specifically for CDL’s American West project. Step B is performed with topicSeek’s topic model, which is currently proprietary software. Step C is evaluated using the topic browser, which allows one to analyze, visualize and review the topic model results. This topic browser was developed by topicSeek using PHP and MySQL and is freely available to CDL. The topic browser has been installed on harvest-dev (at <http://harvest-dev.cdlib.org/aw/>), and can run on any Windows or Unix/Linux system that supports PHP and MySQL. Please see Appendix L: Description of Software for a full explanation of all software components.

Step	Description	Comment
Dublin Core (DC) element extraction	Selected DC elements are used for topical clustering. We investigated two combinations of DC elements: {Title, Subject, Description, Coverage}, and {Title, Subject, Description}. The second set was used for the final topical clustering runs. Given an object, all text is extracted for those DC elements. Duplicate information is not removed.	To increase separation of the four independently browsable metadata fields envisioned for the American West portal interface (Topic, Geographical location, Date, and Genre), we decided to omit dc:coverage (Geographical location) from the final topic model runs.
Normalization	The following steps are used to normalize word tokens <ul style="list-style-type: none"> <li>- convert to lowercase</li> <li>- replace all punctuation with a space</li> <li>- remove all single characters and all two-letter words</li> </ul>	No stemming has been used yet (normalizing to word-stems, e.g. combining 'houses' with 'house'). Stemming could be used – this is further discussed in Appendix H
Stopword removal	Frequently occurring words (e.g. the and for from) are removed. Additional words not related to topic/subject are removed. In particular, words relating to Coverage, Date, and Genre/Type are removed. The final stopword list consisted of <ul style="list-style-type: none"> <li>- 200+ standard stopwords</li> <li>- Words relating to html, file format, computer equipment</li> <li>- Words relating to Genre</li> <li>- The 50 U.S. states and their common abbreviations</li> <li>- All word tokens starting with a digit (e.g. all years)</li> <li>- Numbers, days of week, months of year</li> </ul>	The creation of a stopword list is an iterative process. Additional stopwords are added to improve topic clarity and coherence.
Collocation replacement	Frequently occurring word pairs or triples were identified and replaced as single word tokens (e.g. San Francisco becomes san_francisco).	This is useful because we can remove 'adobe_photoshop' and keep 'adobe'.
Vector representation	Once vocabulary is finalized, each word token in an object is replaced by its word ID within the context of the specific collection vocabulary. The final representation of each object is a list of word IDs and counts.	

*Table 2.* Preprocessing steps. These are the steps taken in the "Preprocess" block in Figure 1.

## 2.1 Preprocessing Examples

Harvested Object	Normalized Extracted Words	Document Vector Representation
<pre>&lt;dc oaiid="http://www.openarchives.org/OAI/2.0/oai_dc/" schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd" xmlns="http://purl.org/dc/elements/1.1/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xlink="http://www.w3.org/TR/xlink"&gt; &lt;title&gt;Left to right: Mrs. Nobu Kuniyoshi, Takako Tanioka, Bennie Kuniyoshi, Amy Tanioka, Mrs. Fude Tanioka, Marlene Tanioka, Marjorie Tanioka, Anna Tanioka, Mrs. Tanioka, Jimmy Tanioka, Mrs. Kajiro Tanioka, and George Kuniyoshi. The picture shows the Tanioka family at their home at Rt. 1, Box 685, Merced, Calif., to which they have recently returned. Mr. Tanioka returned to his home in April and the balance of the family arrived June 10, from the Granada Relocation Center. The Kuniyoshis, who were visiting the Taniokas, operate a farm about 1/2 mile from the Tanioka family, and they also returned early in June. In addition to the above members of the Tanioka family there is Charles, who is attending school at Boulder, Colorado. Other members of the Kuniyoshi family are Shinzen, head of the family; Yo, the eldest son; Denji, a son who is with the United States Army in France. Photographer: Iwasaki, Hikaru Merced, California. 6/29/45&lt;/title&gt; &lt;subject&gt;War Relocation Authority Photographs of Japanese-American Evacuation and Resettlement&lt;/subject&gt; &lt;subject&gt;Series 16: Resettlement&lt;/subject&gt; &lt;subject&gt;Group 9&lt;/subject&gt; &lt;publisher&gt;The Bancroft Library, University of California, Berkeley.&lt;/publisher&gt; &lt;type&gt;image&lt;/type&gt; &lt;identifier&gt;http://ark.cdlib.org/ark:/13030/tf6489p0f3&lt;/identifier&gt; &lt;relation&gt;http://oac.cdlib.org/findaid/ark:/13030/tf596nb4h0&lt;/relation&gt; &lt;relation&gt;ark:/13030/tf729p8s0&lt;/relation&gt; &lt;relation&gt;http://bancroft.berkeley.edu/&lt;/relation&gt; &lt;relation&gt;http://findaid.oac.cdlib.org/findaid/ark:/13030/tf596nb4h0&lt;/relation&gt; &lt;relation&gt;http://jarda.cdlib.org/&lt;/relation&gt; &lt;relation&gt;http://sunsite.berkeley.edu/CalHeritage/&lt;/relation&gt; &lt;/dc&gt;</pre>	<pre>war_relocation_authority photographs japanese_american evacuation resettlement series resettlement group left right mrs nobu kuniyoshi takako tanioka bennie kuniyoshi amy tanioka mrs fude tanioka marlene tanioka marjorie tanioka anna tanioka mrs tanioka jimmy tanioka mrs kajiro tanioka george kuniyoshi picture shows tanioka family their home box 685 merced calif which they have recently returned tanioka returned his home april balance family arrived june granada relocation_center kuniyoshis who were visiting taniokas operate farm about mile tanioka family they also returned early june addition above members tanioka family there charles who attending school boulder colorado other members kuniyoshi family are shinzen head family eldest son denji son who united_states_army france photographer iwasaki_hikaru merced california</pre>	<pre>253 1 907 1 1453 1 2030 1 3267 1 3320 1 4873 1 4873 1 8885 1 9950 6 9977 1 11455 1 12701 1 13256 2 14487 1 18116 1 18653 2 18712 2 18919 1 19549 4 22256 1 24056 1 24560 2 24669 3 25886 1 27359 2 28972 13 30656 1 31479 1</pre>

**Table 3.** Sample initial object, extracted words, and vector representation (file cdl-images-72025.xml). In this table stopwords are removed going from the normalized extracted words to the vector representation. Also note that all word order information is lost in the vector representation (referred to as the "bag of words" representation).

Harvested Object	Normalized Extracted Words	Document Vector Representation
<pre>&lt;dc oaiid="http://www.openarchives.org/OAI/2.0/oai_dc/" schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd" xmlns="http://purl.org/dc/elements/1.1/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xlink="http://www.w3.org/TR/xlink"&gt; &lt;title&gt;"Las Casitas." 7&lt;/title&gt; &lt;subject&gt;Views from a Trip to California.&lt;/subject&gt; &lt;subject&gt;Volume 1 (1905.06484:1-126)&lt;/subject&gt; &lt;subject&gt;Group 1&lt;/subject&gt; &lt;publisher&gt;The Bancroft Library, University of California, Berkeley.&lt;/publisher&gt; &lt;type&gt;image&lt;/type&gt; &lt;identifier&gt;http://ark.cdlib.org/ark:/13030/tf9f59p4p5&lt;/identifier&gt; &lt;relation&gt;http://oac.cdlib.org/findaid/ark:/13030/tf4c600848&lt;/relation&gt; &lt;relation&gt;ark:/13030/tf729p8s0&lt;/relation&gt; &lt;relation&gt;http://bancroft.berkeley.edu/&lt;/relation&gt; &lt;relation&gt;http://findaid.oac.cdlib.org/findaid/ark:/13030/tf4c600848&lt;/relation&gt; &lt;relation&gt;http://sunsite.berkeley.edu/CalHeritage/&lt;/relation&gt; &lt;/dc&gt;</pre>	<pre>las casitas trip</pre>	<pre>4492 15815 30164</pre>

**Table 4.** Sample initial object, extracted words, and vector representation (file cdl-images-125392.xml). Note that only three words remain in the final representation, and two of them, "las casitas," are related to geographical coverage (here, stopwords were removed going from the XML to the normalized extracted words). This is a good example of how some objects get classified based on relatively little information, in this case, almost solely on the word *trip*.

## 2.2 Collocations

We used collocation replacement during preprocessing to improve the clarity and readability of the topics. Collocations are adjacent work tokens that are split during normalization, but really belong together (e.g. 'san' followed by 'francisco' is replaced by 'san\_francisco'). Table 5 lists all the three-word collocations identified and replaced, while Table 6 shows selected frequently occurring two-word collocations identified and replaced. Note that some collocations were identified and then deleted because they were on the stopwords list. Collocation replacement allows us, for example, to delete 'adobe\_photoshop' but keep 'adobe', or to delete 'arkansas' but keep 'arkansas\_river' (a major Colorado river drainage).

<b>Word triple</b>	<b>Count</b>
grand_coulee_dam	45200
north_central_america	29400
indians_north_america	11500
war_relocation_authority	8400
united_states_army	960

*Table 5.* Number of occurrences of words-triples. These were the only word triples identified and replaced as a single token. 'north\_central\_america' is from the Thesaurus of Geographic Names (TGN) term 'North and Central America' and 'indians\_north\_america' is from the LCSH term 'Indians of North America'.

<b>Word pair</b>	<b>Count</b>
<i>washington_state</i>	153000
<i>united_states</i>	110000
san_francisco	30000
<i>dorothea_lange</i>	24000
columbia_river	12000
columbia_basin	10000
New_york	10000
native_americans	9000
santa_ana	8000
japanese_american	8000
national_park	5000

*Table 6.* Most frequently occurring word pairs. Word pairs in italics were included in the final stopword list.

### 3 Topic Model Runs

Table 7 lists the topic model runs performed. We initially investigated and compared the results of using the four DC elements {Title, Subject, Description, Coverage} versus the three DC elements {Title, Subject, Description}, referred to as "Run4" and "Run3" respectively. We focus the discussion on Run3, since that is what was finally used. We see that from Run3a thru Run3e, various modifications were made to the preprocessing rules, while the stopword list continually expanded.

Run	Description	Comment
Run4a	Extracted DC {Title, Subject, Description, Coverage} Standard stopwords removed	These two sets of runs using the four DC elements were discontinued in favor of the runs using the three DC elements.
Run4b	Run4a + extra stopwords removed	
Run3a	Extracted DC {Title, Subject, Description} Standard stopwords removed	Identified list of Format and Type words that were added to stopwords.
Run3b	Run3a + extra stopwords removed	Wanted better geographic separation, so removed U.S. state names and their common abbreviations.
Run3c	Run3b + state names removed	Wanted better chronologic separation, so removed all word tokens beginning with a digit (e.g. 1939).
Run3d	Run3c + dates removed	Wanted better readability of topics, so used collocation replacement to create common word pairs and triples.
Run3e	Run3d + collocations replaced	This was the final run for this feasibility study. This run was used for most of the analyses and data shown in this report.
EAD Runs <sup>2</sup>	Run1: Standard stopwords Run2: Run1 + extra stopwords Run3: Run2 + extra stopwords Run4: Run3 + extra stopwords	The limited size of the EAD collection (7000 docs) allowed us to iteratively refine the topic model results by aggressively expanding the stopword list. See Appendix K for further discussion of the EAD topic model runs

*Table 7.* A list of the topic model runs performed. Usually, the model was run for T = 50, 100, 200 and 400 topics.

#### 3.1 Results for Run3e

As described earlier, the modeling procedure is iterative, with each iteration further improving the results. The Run3e run was the final run, and used for the detailed analysis given in the remainder of this section. The run3e results can be browsed at <http://harvest-dev.cdlib.org/aw>. Table 8 gives some basic statistics for run3e, indicating that there were approximately 360,000 objects processed, with a collection vocabulary in excess of 30,000 words. The most frequently occurring words in run3e, Table 9, indicate a strong emphasis in the collection toward 'Seattle', the 'Grand Coulee Dam', and 'San Francisco'.

Number	Value
Number of objects	357,566
Size of vocabulary	32,652
Total number of words	7,419,183
Average number of words in object	21
Maximum number of words in object	864

*Table 8.* Basic statistics for run3e.

<sup>2</sup> A separate analysis of the clustering experiment on collection-level descriptive information extracted from the Online Archive of California's EAD-encoded finding aids will be submitted by Adrian Turner.

Word	Count
seattle	69000
history	66000
united	46000
grand_coulee_dam	45000
north_central_america	29000
buildings	29000
river	28000
construction	28000
facilities	27000
san_francisco	27000
county	26000
street	26000
building	24000
company	24000
government	24000
family	22000
portraits	22000
house	21000
water	20000
men	19000

Table 9. Most frequently occurring words in run3e.

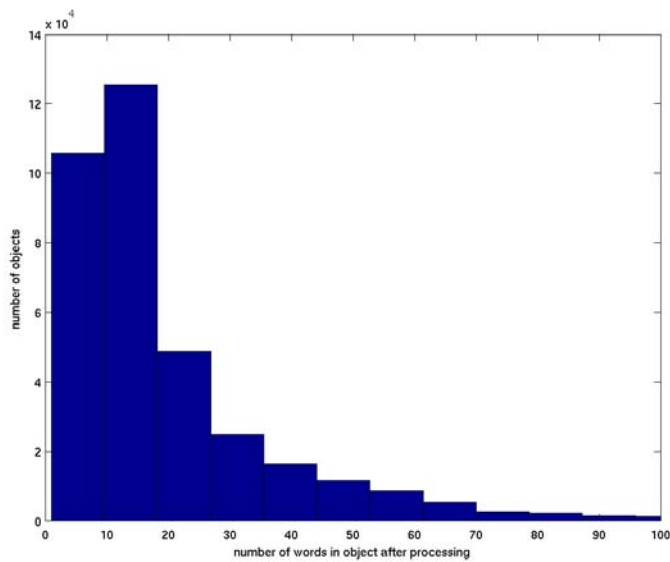


Figure 2. Distribution of number words per object after processing (for run3e). Note scale on y-axis: e.g. there are 100,000 objects that contain between 1 and 10 words. On average, there are 21 words per object after processing.

### 3.2 Review of Topics

The primary success metric is the number of interpretable and usable topics. Measuring this number is a manual process. A cataloger or domain expert decides whether a topic is interpretable and usable by examining the list of

most frequently occurring words in the topic and the occurrences of those words in context in individual documents (in this case harvested metadata records). The primary factor in determining how time-intensive this examination becomes is the number of topics in the run (i.e., the process takes longer for a 400-topic run than it does for a 50-topic run). This examination by a cataloger or domain expert is also critical in refining the stopword list, which contributes to improving the quality of the topics in subsequent runs. Table 10 shows selected topics from the 400-topic run of run3e. The topical labels were supplied by CDL's Metadata Coordinator using the topical heading from LCSH that seemed to best capture the sense of each topical cluster. The process of supplying topical labels can be quite lengthy, involving exploration of the word clusters comprising a topic and the associated documents in order to clarify and validate the topic. In addition, the process requires some research in an appropriate topical thesaurus to determine the standardized vocabulary term that best captures the sense of the word cluster. In this particular experiment, the CDL Metadata Coordinator spent roughly 24 hours exploring the 400 word clusters in Run 3e (the final topic model run) and supplying LCSH subject headings to those topics deemed usable. Because of its size and complexity, LCSH may not be the best topical thesaurus to use for this kind of activity; however, it was chosen in this case because it is a familiar and frequently used standard in the digital library community. Other possibilities, such as the Dewey Decimal Classification (DDC) and the Thesaurus for Graphic Materials I: Subject Terms (TGM I) may be more appropriate for this work than LCSH.

Approximate Size (%)	Most frequently occurring words in topic	Topical Labels (LCSH)
1.0%	boat river trip boats lie testimony moab miles abstract mouth supplies san_juan trips greenriver green_river bluff party colorado_river oil canyon upstream motor	Rivers – Navigation
0.7%	evacuation resettlement relocation japanese_american war_relocation_authority homes places mrs iwasaki_hikaru continued relocation_center employed home farm relocated family	Japanese Americans -- Evactuation and relocation, 1942-1945
0.7%	columbia_basin irrigation columbia_river project history reclamation watershed valley basin league commission planning_commission development general spokane_chronicle surveys spokesman	Water reuse -- Pacific states
0.6%	grand_coulee_dam coulee dam history grand construction site spokane_chronicle design spokesman review columbia_river progress project spokane_press bureau_reclamation damsite	Dams -- Design and construction
0.5%	native_americans indians indian history yakima spokesman native_american review indians_north_america spokane_chronicle northwest_pacific tribe indian_reservation obituaries	Indians of North America
0.5%	water sand river channel bars rapids trouble bar deep rapid low places canyon current stream shallow waves stuck difficulty encountered rocks swift find run cataract bad change upstream	Rivers -- Navigation
0.5%	logging lumber lumber_company company loggers county logs mill industry coll crew kinsey grays_harbor business miles timber steam shingle equipment_supplies town donkeys log	Logging
0.4%	dress indians clothing man woman portraits beaded nez_perce wearing blanket wears named yakama note ceremonial pose shoulders necklaces decorated women wear blankets	Indians of North America -- Material culture
0.4%	mining gold mine mineral mines county industries miners perspectives resources nmhfm leadville climax showing lake dredge dressing gulch plants fairplay placer greenlee operation stillwater shaft	Mines and mineral resources
0.3%	grand_coulee_dam visitors history tourists coulee visit dam tourism officials review spokesman grand tourist tour spokane general geology spokane_chronicle trip visits project site visited inspection vista visitor	Dams -- Description and travel
0.3%	trees arboretum morton arb botanical_gardens flowers tree shrubs leaves limbs shadows maple cherry strybing autumn golden_gate malus flowering crab magnolia prunus oak blossoms willows	Trees

*Table 10.* Selected topics from run3e, T400. Supplied topical label in rightmost column provided by CDL's Metadata Coordinator. Note that if all 400 topics were equally important and evenly distributed, they would each represent 0.25%.

Topic cluster	brown cook san_sanfrancisco jesse law_enforcement scrapbooks documenting history ave sts cor feb bet east sept nov jany aug oct sic montgomery geary kearny dec alias folsom polk mcallister mar larkin howard turk rolph fulton ewing fillmore mission
Words to add to stopword list (ideally as a collocation)	Jesse Brown Cook Scrapbooks Documenting San Francisco History and Law Enforcement

*Table 11.* Example of opportunity to add words to stopword list. This topic may initially appear to be about law enforcement in San Francisco, but "Jesse Brown Cook Scrapbooks Documenting San Francisco History and Law Enforcement" is the name of the series (found in the DC:Subject element because of an awkward mapping by the OAI data providing institution). The result is that this topic will mostly contain objects from this series, and not much else. If the words "Jesse Brown Cook Scrapbooks Documenting San Francisco History and Law Enforcement" were removed as a unit, it would allow these objects to be described by broader topics.

Topic cluster	scott morris mabel turner leonard andrews hugh ruby sherman owen crawford peterson burns graves hazel willie irving reid dale cunningham lena esther clayton cora julian daisy vivian ada hawthorne parsons caldwell belmont kathleen cummings
Words to add to stopword list	All the above names!

*Table 12.* The topic of names. Here is another example of an opportunity to add words to stopword list. Topics of names are often seen during topic modeling. Here, the topic is not useful for subject metadata enhancement, and can either be ignored, or more preferably, removed by adding the names to the list of stopwords.

### 3.3 Mapping of Topic Clusters to 23 Broad Topical Categories

Earlier work at CDL on the American West project discussed the use of 23 Broad Topical Categories (BTCs) to categorize objects (see the report *American West Project High-level Topic Taxonomy*, 2005-01-14)<sup>3</sup>. These BTCs have been discussed and reviewed, and are will likely serve as the top-level of the hierarchical topical browse envisioned for the American West web portal.<sup>4</sup> Two key questions are: a) How well do the useful topics out of the topic model map to one or more of the BTCs?, and b) What is the best procedure to assign topics from the topic model to one or more BTCs? Table 13 shows mappings of a subset of the LCSH topical labels from the run3e 400-topic run to the 23 BTCs.

Mapping of the LCSH topical labels to the 23 BTCs was done by the CDL Metadata Coordinator and took approximately 6 hours. It is important to remember that this mapping is an unavoidably subjective process and represents a single person's viewpoint. One way of improving the reliability of the mapping process might be to invest further effort into expanding the scope notes for the 23 BTCs.

BTC	Name	List of T400 Topic Clusters	
1	Advertising & media	• Motion pictures	• Tourism
2	Agriculture and food production	• Archaeology • Industries	• Tourism • Water reuse -- Pacific states

<sup>3</sup> Available at: <[https://diva.cdlib.org/projects/american\\_west/metadata\\_coordinator/Content\\_analysis/AmWest\\_topics\\_firstlevel.pdf](https://diva.cdlib.org/projects/american_west/metadata_coordinator/Content_analysis/AmWest_topics_firstlevel.pdf)>

<sup>4</sup> At the time of the completion of this report, CDL's Assessment experts are engaged in an evaluation of the 23 BTC's via in-person and online survey instruments. Their report on these assessment activities will be available in June 2005 and will contribute to the final design of a topical browsing infrastructure for the American West web portal.

BTC	Name	List of T400 Topic Clusters	
3	American Indians	<ul style="list-style-type: none"> <li>• Archaeology</li> <li>• Children</li> <li>• Indian reservations</li> <li>• Indians of North America</li> </ul>	<ul style="list-style-type: none"> <li>• Indians of North America -- Government relations</li> <li>• Indians of North America -- Material culture</li> <li>• Indians of North America -- Social conditions</li> </ul>
4	Arts and architecture	<ul style="list-style-type: none"> <li>• Actors and actresses</li> <li>• Architecture</li> </ul>	<ul style="list-style-type: none"> <li>• Buildings</li> <li>• Theater</li> </ul>
5	Business and industry	<ul style="list-style-type: none"> <li>• Bridges -- Design and construction</li> <li>• Central business districts</li> <li>• Dams -- Design and construction</li> <li>• Finance</li> <li>• Globalization -- Political aspects</li> <li>• Industries</li> <li>• Logging</li> </ul>	<ul style="list-style-type: none"> <li>• Mines and mineral resources</li> <li>• Railroads</li> <li>• Stores, retail</li> <li>• Water resources development -- Finance</li> <li>• Water reuse -- Pacific states</li> <li>• Water-power</li> </ul>
6	Crime and violence	<ul style="list-style-type: none"> <li>• Indians of North America -- Social conditions</li> </ul>	<ul style="list-style-type: none"> <li>• Litigation</li> </ul>
7	Education	<ul style="list-style-type: none"> <li>• Children</li> <li>• Indians of North America -- Government relations</li> </ul>	<ul style="list-style-type: none"> <li>• Indians of North America -- Social conditions</li> <li>• Schools</li> </ul>
8	Environment and natural resources	<ul style="list-style-type: none"> <li>• Dams -- Description and travel</li> <li>• Dams -- Design and construction</li> <li>• Geology</li> <li>• Globalization -- Political aspects</li> <li>• Insects</li> <li>• Lakes</li> <li>• Logging</li> <li>• Mountains</li> </ul>	<ul style="list-style-type: none"> <li>• Rivers -- Description and travel</li> <li>• Rivers -- Navigation</li> <li>• Trees</li> <li>• Water resources development -- Finance</li> <li>• Water resources development -- Political aspects</li> <li>• Water reuse -- Pacific states</li> <li>• Water-power</li> </ul>
9	Ethnic groups	<ul style="list-style-type: none"> <li>• Children</li> </ul>	<ul style="list-style-type: none"> <li>• Japanese Americans -- Evacuation and relocation, 1942-1945</li> </ul>
10	Exploration and travel	<ul style="list-style-type: none"> <li>• Dams -- Description and travel</li> <li>• Exhibitions</li> <li>• Lakes</li> <li>• Mountains</li> </ul>	<ul style="list-style-type: none"> <li>• Railroads</li> <li>• Rivers -- Description and travel</li> <li>• Rivers -- Navigation</li> <li>• Tourism</li> </ul>
11	Gender	<ul style="list-style-type: none"> <li>• Clothing and dress</li> </ul>	
12	Government and law	<ul style="list-style-type: none"> <li>• Finance</li> <li>• Forts and fortifications</li> <li>• Indian reservations</li> <li>• Indians of North America -- Government relations</li> <li>• Litigation</li> <li>• Northwest, Pacific -- Politics and government</li> </ul>	<ul style="list-style-type: none"> <li>• Politics and government</li> <li>• United States -- Politics and government</li> <li>• Water resources development -- Finance</li> <li>• Water resources development -- Political aspects</li> </ul>
13	Migration		
14	Military and war	<ul style="list-style-type: none"> <li>• Exhibitions</li> <li>• Forts and fortifications</li> </ul>	<ul style="list-style-type: none"> <li>• Japanese Americans -- Evacuation and relocation, 1942-1945</li> </ul>
15	Political participation	<ul style="list-style-type: none"> <li>• Globalization -- Political aspects</li> <li>• Indian reservations</li> <li>• Litigation</li> <li>• Northwest, Pacific -- Politics and government</li> </ul>	<ul style="list-style-type: none"> <li>• Politics and government</li> <li>• United States -- Politics and government</li> <li>• Water resources development -- Political aspects</li> </ul>

BTC	Name	List of T400 Topic Clusters	
16	Popular and domestic culture	<ul style="list-style-type: none"> <li>Actors and actresses</li> <li>Archaeology</li> <li>Architecture</li> <li>Buildings</li> <li>Church buildings</li> <li>Clothing and dress</li> <li>Indians of North America -- Material culture</li> </ul>	<ul style="list-style-type: none"> <li>Indians of North America -- Social conditions</li> <li>Motion pictures</li> <li>Photography of families</li> <li>Schools</li> <li>Stores, Retail</li> <li>Theater</li> <li>Tourism</li> <li>Universities and colleges</li> </ul>
17	Recreation and leisure	<ul style="list-style-type: none"> <li>Actors and actresses</li> <li>Children</li> <li>Dams -- Description and travel</li> <li>Exhibitions</li> <li>Lakes</li> </ul>	<ul style="list-style-type: none"> <li>Motion pictures</li> <li>Mountains</li> <li>Rivers -- Description and travel</li> <li>Schools</li> <li>Theater</li> </ul>
18	Religion and philosophy	<ul style="list-style-type: none"> <li>Archaeology</li> </ul>	<ul style="list-style-type: none"> <li>Church buildings</li> </ul>
19	Science and medicine	<ul style="list-style-type: none"> <li>Geology</li> </ul>	<ul style="list-style-type: none"> <li>Insects</li> </ul>
20	Social and family life	<ul style="list-style-type: none"> <li>Buildings</li> <li>Children</li> <li>Church buildings</li> <li>Clothing and dress</li> <li>Indian reservations</li> </ul>	<ul style="list-style-type: none"> <li>Indians of North America -- Material culture</li> <li>Indians of North America -- Social conditions</li> <li>Japanese Americans -- Evacuation and relocation, 1942-1945</li> <li>Photography of families</li> <li>Schools</li> </ul>
21	Transportation	<ul style="list-style-type: none"> <li>Bridges -- Design and construction</li> <li>Exhibitions</li> <li>Railroads</li> </ul>	<ul style="list-style-type: none"> <li>Rivers -- Description and travel</li> <li>Rivers -- Navigation</li> </ul>
22	Urban life	<ul style="list-style-type: none"> <li>Central business districts</li> <li>Exhibitions</li> <li>Industries</li> </ul>	<ul style="list-style-type: none"> <li>Railroads</li> <li>Stores, Retail</li> </ul>
23	Work and labor	<ul style="list-style-type: none"> <li>Archaeology</li> <li>Bridges -- Design and construction</li> <li>Dams -- Design and construction</li> <li>Indians of North America -- Government relations</li> </ul>	<ul style="list-style-type: none"> <li>Industries</li> <li>Logging</li> <li>Mines and mineral resources</li> </ul>

*Table 13.* Classification of T400 topical labels (LCSH) to BTCs. This is only a partial mapping of the topical labels, representing approximately 20% of the topics.

### 3.4 Comparison 100-Topic run and 400-Topic Run

Table 14 shows how the topic model produces topics of different specificities, using as an example topics relating to the Grand Coulee Dam. In the T=100 row, we see two topics relating to the Grand Coulee Dam; the first about construction and design, and the second about irrigation and water reclamation. In the T=400 row we see eight topics relating to the Grand Coulee Dam ranging from construction, contracts and bids, financing, employment, and tourism. To help differentiate the topics, the words Grand, Coulee and Dam were removed from the table for display purposes. These words are not on the stopword list.

T	Topics							
100	<b>construction</b> history <b>design</b> concrete company mwak contracts spokesman review columbia_river bids site spokane_chronicle work atkinson_kier mason_walsh engineers excavation progress cofferdams mason_city pouring				columbia_basin <b>irrigation</b> history project columbia_river <b>reclamation</b> commission sullivan james basin watershed spokane valley secretary league review spokane_chronicle appropriations spokesman chamber commerce			
400	history <b>construction</b> site spokane_chronicle design spokesman review columbia_river progress project spokane_press bureau_reclamation site almira spokane elmore build power mason_city huge completion complete	construction columbia_river history design <b>coffers</b> coffer excavation <b>diversion</b> work progress review spokesman columbia river piling channel steel huge engineering east bedrock mighty pumping channels dynamite shore	<b>company</b> mwak history construction atkinson_kier mason_walsh mason_city design spokesman review general contractors spokane_chronicle work site job builders atkinson bureau_reclamation pomeroy contract schedule ahead	Construction <b>contracts bids</b> history design mason letting contract silas spokane_chronicle bid review spokesman low slocum company job contractors work harvey opened awarded bureau_reclamation site spokane completion	<b>president</b> history roosevelt franklin_delano <b>appropriations</b> roosevelt_franklin <b>financing</b> project finance interior department appropriation budget united <b>fun</b> ds review reclamation spokesman spokane_chronicle	<b>employment</b> construction construction_workers <b>wages</b> history work design men statistics <b>payroll</b> labor review bureau_reclamation jobs spokesman payrolls site figures mwak job total employed service reemployment beery	<b>concrete</b> construction history pouring s editorials feature <b>design</b> articles hoover yards poured commentary work manley foundation cubic mixing plant progress placing forms spillway west_side pour record statistics yard	<b>visitors</b> history <b>tourists</b> visit <b>tourism</b> officials review spokesman tourist tour spokane general geology spokane_chronicle trip visits project site visited inspection vista visitor inspect party sight caravan

Table 14. Hierarchy of topics relating to the Grand Coulee Dam. The words 'grand' 'coulee' and 'dam' were removed just in this table for display purposes. Words in bold differentiate the topics within a row.

### 3.5 Sample Topic Decompositions/Metadata Enhancement

After the topic model is run and topical labels are supplied for the usable topics, objects can have their subject metadata enhanced with these topics. Table 15 shows this topic decomposition or metadata enhancement for a randomly selected object (wsu-clipping-12499.xml). Here, the model calculates that this object can be 79% explained by the topics [Water reuse -- Pacific states], [Water-power] and [Water resources development -- Finance]. This object would then be enhanced with additional subject metadata listing these three additional subject headings. Another example is given in Appendix J, showing the proportion of the object cdl-images-83026.xml that is explained by the top topics.

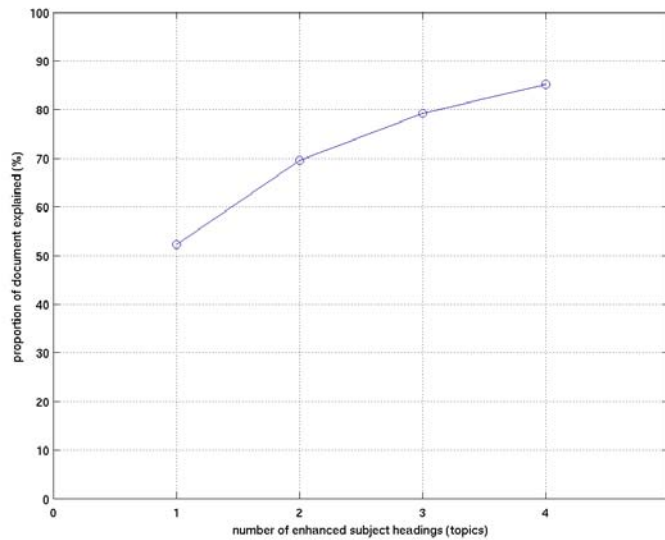
wsu-clipping-12499.xml		
<title>State history. Columbia Basin. Columbia Basin Planning Commission. 1936-02-19. p. 2.</title> <title>Oregonian ; 1936-02-19</title> <title>Planning commission outlines scheme in report on Columbia basin</title> <description>Planning commission outlines scheme in report on Columbia basin. - Supreme Court decision on TVA opens way for announcement of proposal meant to aid development / by Lee Bostwick. Creation by congress of a Pacific northwest power agency as a public corporation to develop the transmission and sale of power from Bonneville and Grand Coulee dams was recommended yesterday to President Roosevelt by the Pacific northwest regional planning commission in its Columbia basin report.</description> <subject>State history ; Columbia Basin ; reclamation ; irrigation ; Columbia Basin Commission ; Columbia Basin Planning Commission ; President Franklin Delano Roosevelt ;</subject> <date>1935</date> <subject>Columbia River Watershed Irrigation--Washington (State)--Columbia River Valley Columbia Basin Commission (Wash.) Pacific Northwest Regional Planning Commission Roosevelt, Franklin D. (Franklin Delano), 1882-1945.</subject>		
Percent	Topic Words	Supplied Topical Label
41%	columbia_basin irrigation columbia_river project history reclamation watershed valley basin league commission planning_commission development general spokane_chronicle surveys spokesman review gill roy institute districts spokane land committee	Water reuse -- Pacific states
21%	power bonneville_dam columbia_river northwest ross electric_power bonneville history hydro hydroelectric plants rates dam administrator navigation administration oregonian project federal politics mcnary authority government distribution policy	Water-power
17%	grand_coulee_dam president history roosevelt franklin_delano appropriations coulee roosevelt_franklin financing project dam finance grand interior department appropriation budget united funds review reclamation spokesman spokane_chronicle	Water resources development -- Finance

Table 15. Sample topic decomposition of object wsu-clipping-12499.xml . The three topics together reflect the contents of the object. Some DC elements were omitted for clarity. Full text from this object is given in Appendix I.

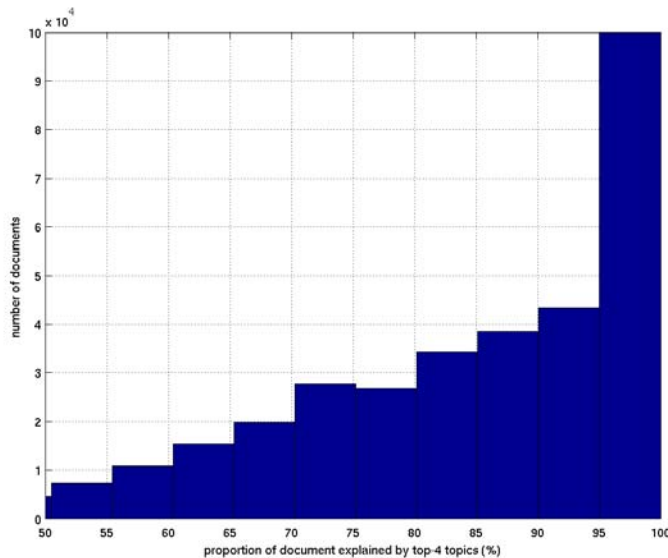
### 3.6 Proportion of Object Explained by Topics

On average, what proportion of a metadata object is explained by the top-1 (2, 3 and 4) topics? The previous section shows one example of this, but averaging over all 360,000 objects, what should we expect? Figure 3 shows that, for run3e, the top topic explains 50% of a metadata object, on average. Furthermore, the top-2 topics explain 70%, the top-3 topics explain 80% and the top-4 topics explain 85% of a metadata object, on average. This information is helpful to plan for the number of additional subject headings that will be used to enhance an object.

Figure 4 takes a slightly different view, and shows the distribution of proportion of a metadata object explained by top-4 topics. On average, 85% of an object is explained by the top-4 topics. Approximately 100,000 of the 360,000 objects have close to 100% of the object explained by the top-4 topics.



*Figure 3.* Proportion of metadata object explained by top-1, 2, 3 and 4 topics. On average, 50% of an object is explained by the top topic, and 85% of an object is explained by the top-4 topics.



*Figure 4.* Distribution of proportion of document (object) explained by top-4 topics. On average, 85% of an object is explained by the top-4 topics. Approximately 100,000 of the 360,000 objects have close to 100% of the object explained by the top-4 topics.

### 3.7 Yield of Enhanced Objects

Two key success metrics are the number of usable topics, and the number of objects enhanced with these usable topics (see Table 1). Preliminary estimates of these numbers were computed as follows. The 400 topics from run3e were scanned for their potential to be usable topics. Usable requires that the topic is both relatively coherent and interpretable, and that it is useful as a subject category (e.g. a topic of people's names is coherent and interpretable, but not useful as a subject category). An initial scan of the 400 topics indicated approximately 310 topics (78%) could potentially be labeled and useable. Given these 310 usable topics, how many of the 360,000 objects get enhanced with one or more usable topics? One decision that will affect the number of enhanced objects is setting

the minimum proportion of an object that must be explained by, for example, the top four topics. To estimate this number, we required that at least 40% of an object be explained by the top four topics. This value of 40% is called the aggregate threshold. Also, we require that topics individually explain at least 5% of an object. This value of 5% is called the individual topic threshold. Using these thresholds, and with 90 out of 400 topics not used, we estimated that approximately 288,000 of 360,000 (80%) objects would have their metadata enhanced with at least one usable topic. These numbers are shown in Table 16.

This value of 80% would decrease if there were fewer usable topics, or if we increased the aggregate threshold of the proportion of an object that must be explained by the top topics. This value of 80% would increase if there were more usable topics. We believe that it is possible to increase the number of usable topics by further expanding stopword lists, which might improve previously unusable topics and would increase the proportion of objects that get enhanced. We estimate that it would take one person less than 24 work-hours to create a reasonably complete collection of stopwords specific to the American West collection. This estimate is based on time spent documenting stopwords for the American West collection, and for the much-smaller EAD collection<sup>5</sup>. This estimate might increase slightly if the stopwords were being managed in a database and had to be broken down into categories.

We recommended that the quality and usefulness of topical assignments be reviewed when the proportion of the object explained by the top four topics is marginal (close to the aggregate threshold, e.g. in the 40%-50% range). This review may guide one in deciding where to set this aggregate threshold (currently at 40%). This review is a manual process of examining randomly selected objects that are deemed marginal.

Quantity	Value	Percent
Total number of topics	400	
Number of labeled and useable topics	310	78%
Number of unlabeled topics	90	22%
Total number of objects	360,000	
Number of objects enhanced	288,000	80%
Number of objects not enhanced	72,000	20%

*Table 16.* Yield of enhanced objects. Note that with 22% of topics were unlabelled (i.e. not used), 20% of objects end up not enhanced with topic assignments. Some portion of the 30,000 un-enhanced objects will be non-AW objects. This proportion of un-enhanced objects will decrease as a) the non-AW objects are removed from the collection, and b) stopword lists are expanded.

---

<sup>5</sup> Preliminary work on a 50-topic run on the smaller EAD collection indicate that it is possible, through iterative refinement of stopword lists, to achieve nearly 100% usable topics.

#### 4 Identification of non-American West Objects

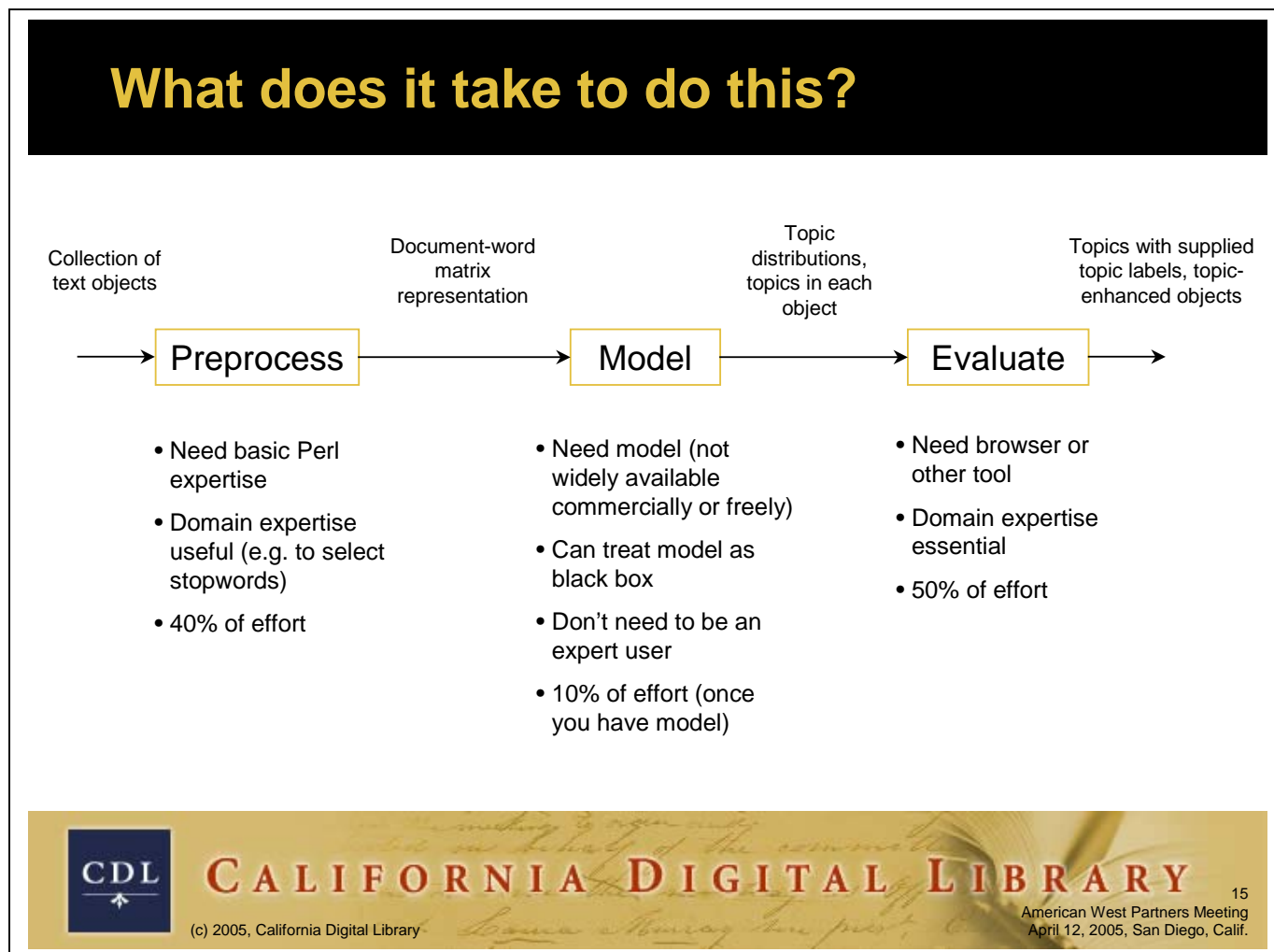
We wanted to discover rules for identifying non-AW objects that would not have the side-effect of excluding some valid AW objects. For example, identifying Egyptian Pyramid-related objects by searching for "pyramid" would exclude many valid AW objects mentioning Pyramid Lake in Nevada. Examination of the topics from the topic model revealed clues for finding groups of non-American West objects. For a topic that looked unrelated to AW, we examined randomly selected objects from the top objects in that topic. This examination usually led to the discovery of a simple rule (or pattern) for identifying the group of non-AW objects. For example, looking at the objects in the topic (egypt world male female pyramids pyramid africa) led to the discovery of a group of 6500 objects from cdl-images containing <subject>Dorothea Lange Collection</subject> followed by <subject>Around the World (1958-1963)</subject>. Using this approach we identified a possible 30,000 non-AW objects out of the 360,000 harvested objects. We also note here that, in future, removal of these objects would improve the quality of topics, increase the number of usable topics, and increase the proportion of objects that get metadata enhancement.

Instution-set	Containing	Estimated number of objects
cdl-images-*	<subject>Dorothea Lange Collection</subject> <subject>Around the World (1958-1963)</subject>	6500
cdl-images-*	<subject>Dorothea Lange Collection</subject> <subject>Photographic Essays (1953-1959)</subject> <subject>Ireland</subject>	2400
cdl-images-*	<subject>Keystone-Mast Collection,</subject> <subject>Stereographic Photoprints by Geographical Location</subject> <subject>Europe</subject> OR Asia OR South America OR Africa	2300 (Europe) 1400 (Asia) 900 (Sth Am) 400 (Africa)
cdl-images-*	<subject>Keystone-Mast Collection,</subject> <subject>Stereographic Photoprints by Geographical Location</subject> <subject>North and Central America</subject> <subject>Canada</subject>	2700 (some may be western Canada)
cdl-images-*	<subject>Keystone-Mast Collection,</subject> <subject>Stereographic Photoprints by Geographical Location</subject> <subject>North and Central America</subject> <subject>United States</subject> <subject>Illinois</subject> OR New York OR District of Columbia OR Florida OR Pennsylvania	2900 (IL) 2500 (NY) 2200 (DC) 1600 (FL) 1000 (PA)
uw-*	<coverage>*France*</coverage> OR <subject>*(France)*</subject>	3000
uw-*	<coverage>*Egypt*</coverage> OR <subject>*(Egypt)*</subject>	800
iu-cushman-*	<coverage>Chicago, Illinois, United States (Cook county)</coverage>	1300
byu-jackson-*	<subject>*Russia (Federation);*</subject>	100
uw-* (and *-*)	<language>Thai</language>	900
<b>Approximate Total (PRELIMINARY ESTIMATE)</b>		<b>30,000</b>

*Table 17.* Rules for identifying potential non-American West objects. These rules were discovered by examining the topic clusters computed by the topic model. There may be as many as 30,000 non-AW objects out of a total of 360,000 harvested objects.

## 5 Options for CDL to Use Topic Modeling for American West Metadata Objects

We suggest three possible approaches for CDL to use Topic Modeling. These approaches are not necessarily mutually exclusive, and can be used in combination with one another. The first, ONE-TIME HARDCODE is a mainly offline/batch approach where the hard-coding of topics into American West objects is performed once. The second, AUTOMATE, is an extension of the first, and envisions an online/in-house approach where CDL has the information and programs necessary to perform topical classification using topic vectors and labels generated by the topic modeling process. The third, VIRTUAL, is a slightly different approach where the topics assignments are not pre-computed or saved, but computed on the fly during browsing using topic vectors and labels generated by the topic modeling process. The following sections outline a possible workflow for each of for these three approaches, and Section 5.4 lists advantages and disadvantages to each approach. Figure 5 provides a graphic overview of the basic process of topic modeling to derive topic vectors and associated topic labels, regardless of how those labels are used to enhance metadata or facilitate access through a browse interface.



*Figure 5.* Flowchart showing a possible process for automatically enhancing subject metadata for newly harvested objects. Online process. Step A is currently handled by Perl scripts. Step B could be done by a small Java, C/C++, or Perl program. Steps A and B, which are currently performed using small Perl scripts and/or C++ programs, could be implement in Java.. Step C is CDL dependent. Please see Appendix L: Description of Software for a full explanation of all software components.

**5.1 ONE-TIME HARDCODE: Hard-Coding of Topics into American West Metadata Records**

ONE-TIME HARDCODE consists of refining the topic modeling until the topics are deemed satisfactory, then performing a one-time hard-coding of topics into the American West metadata records. The workflow tasks are shown in Table 18 and illustrated in Figure 6. Steps 1-7 have been completed during this feasibility study (possibly more iterations of steps 2-4 will be required).

Step	Workflow Task	Who
1	Harvest American West objects	CDL
2	Preprocess and remove stopwords	Vendor
3	Run topic model	Vendor
4	Examine topics, modify preprocessing rules and expand stopword list	CDL (Analysis), Vendor
5	Repeat steps 2-4 until topics are satisfactory	Vendor, CDL (Analysis)
6	Give useable topics a topical label corresponding to a topical thesaurus like LCSH	CDL
7	Compute topics associated with each object	Vendor
8	Hard-code topics into the METS record for each object from vendor-supplied data	CDL

*Table 18.* Possible workflow for ONE-TIME HARDCODE.

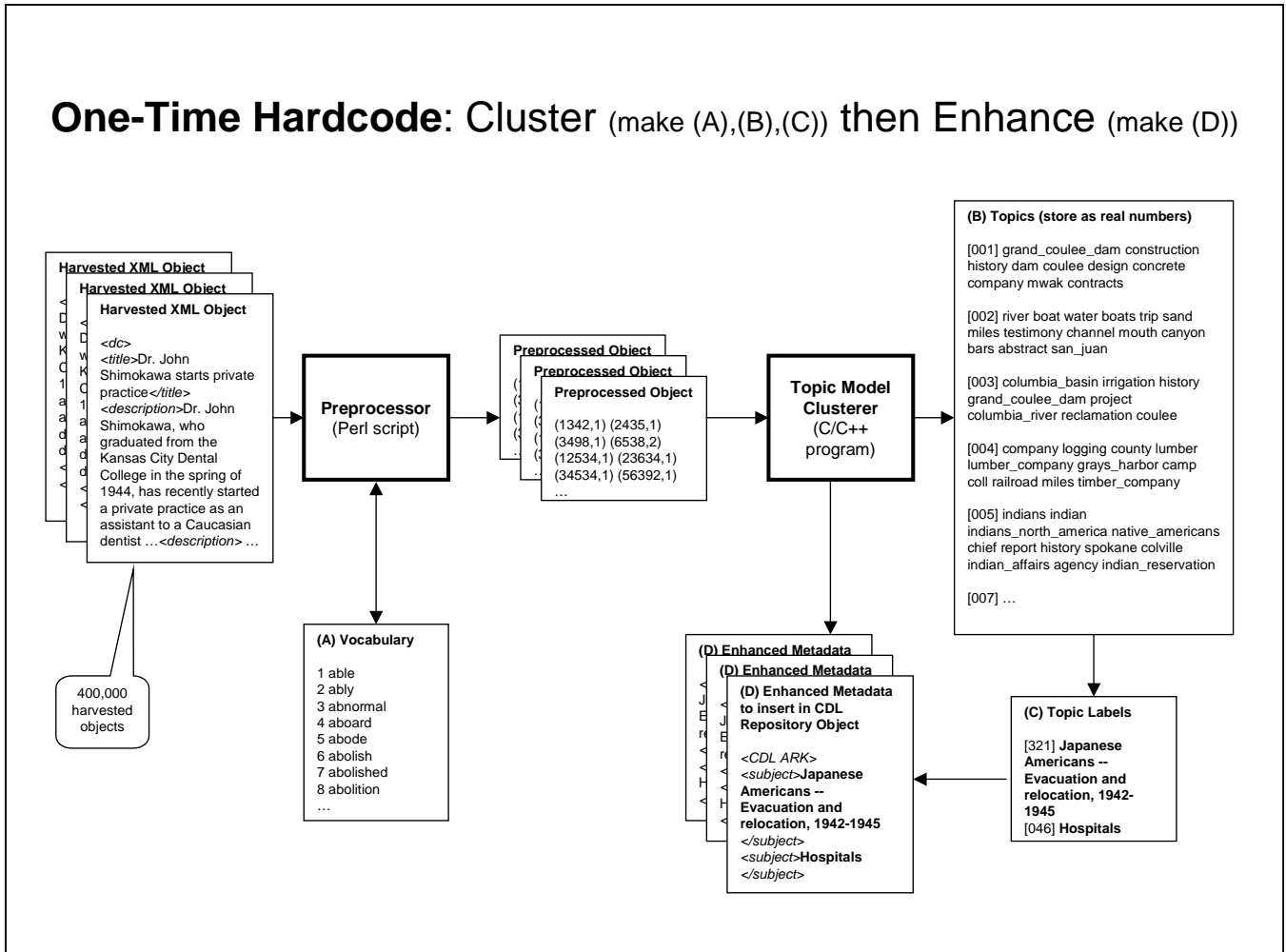


Figure 6. Process overview for ONE-TIME HARDCODE.

## 5.2 AUTOMATE: Process to Automatically Classify American West Metadata Records

AUTOMATE consists of completing steps 1-6 of ONE-TIME HARDCODE and then providing CDL with all the data and programs necessary to perform in-house metadata enhancement. The workflow tasks are shown in Table 19 and illustrated in Figure 7. These steps are currently performed using small Perl scripts and/or C++ programs. It would be possible to implement these scripts and programs in Java.

Step	Workflow Task	Who
1	Perform Steps 1-6 of ONE-TIME HARDCODE	Vendor, CDL
2	Provide CDL with all preprocessing rules, stopword list, topic distributions and names, vocabulary, scripts and programs	Vendor → CDL
3	(Re-) Harvest American West objects (may include additional sets)	CDL
4	Exclude non-AW objects using rules (Classify-AW)	CDL
5	Process and vectorize all objects using rules, stopword list and collection vocabulary (Perl)	CDL
6	Classify each object using the saved topics (Classify-Topic) (Perl and/or C++)	CDL
7	Hard-code topics into each object	CDL

*Table 19.* Possible workflow for AUTOMATE.

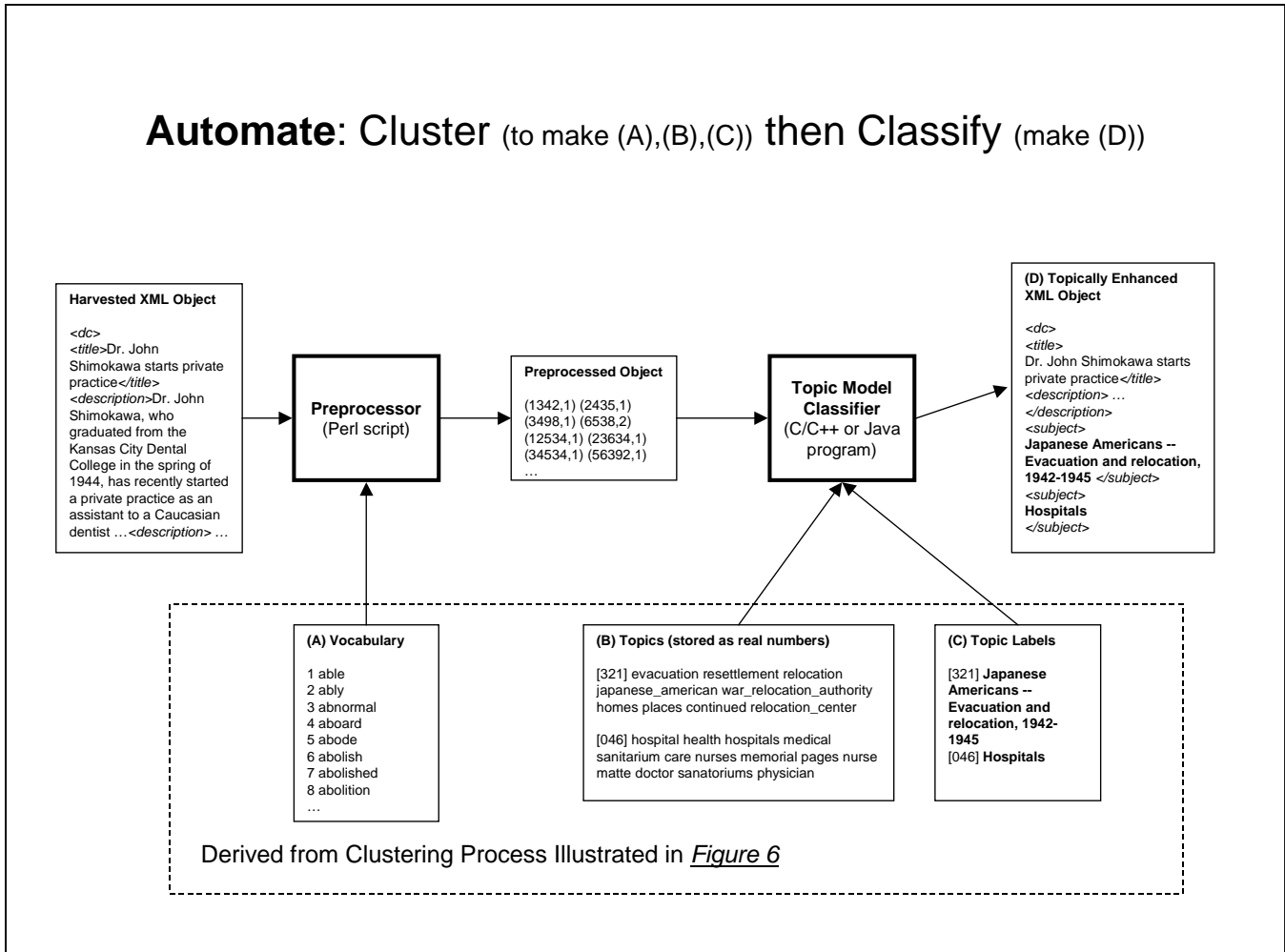


Figure 7. Process overview for AUTOMATE.

**5.3 VIRTUAL: Use Topic Vectors in Browsing Interface**

VIRTUAL consists of completing steps 1-6 of ONE-TIME HARDCODE and then providing CDL with all the data and programs necessary to perform in-house metadata enhancement. The workflow tasks are shown in Table 20. In this approach the topics are not hard-coded into the American West metadata records. Topical associations are computed on the fly during a browse session.

Step	Workflow Task	Who
1	Perform Steps 1-6 of ONE-TIME HARDCODE	Vendor, CDL
2	Save preprocessing rules, stopword list, topic distributions and names, vocabulary	Vendor → CDL
3	(Re-) Harvest American West objects (may include additional sets)	CDL
4	Identify and exclude non-AW objects using saved rules (Classify-AW)	CDL
5	Process and vectorize all objects using saved rules, stopword list and vocabulary	CDL
6	Use predetermined topic vectors to query vectorized documents on the fly during browsing	CDL

*Table 20.* Possible workflow for VIRTUAL.

## 5.4 Comparison of Approaches

Approach	Advantages	Disadvantages
ONE-TIME HARDCODE	<ul style="list-style-type: none"> <li>• Should provide current AW collection with 80% of objects enhanced with topics</li> <li>• Minimum work for CDL</li> </ul>	<ul style="list-style-type: none"> <li>• Metadata enhancement will be a one-time</li> <li>• Not readily extensible to future American West collection additions, or to other future CDL collections without contracting for vendor services</li> </ul>
AUTOMATE	<ul style="list-style-type: none"> <li>• Allows for maintenance of dynamic (growing) American West collection</li> <li>• Allows for classification of previously-unseen objects (e.g. coming from additional harvested sets)</li> <li>• Can reuse entire infrastructure for other collections/projects</li> <li>• Topical classification may improve as partners improve their mapping of OAI-exposed metadata from their native data management systems</li> <li>• Classification software (currently in Perl) could be easily implemented in Java</li> </ul>	<ul style="list-style-type: none"> <li>• More infrastructure for CDL to support</li> <li>• Potential issues in conforming to CDL environment and practices (e.g. porting or translating scripts and programs)</li> <li>• May require future human mediation to identify non-AW objects from newly harvest sets</li> </ul>
VIRTUAL	<ul style="list-style-type: none"> <li>• Document vectors representing metadata records can be searched on the fly using predetermined topic vectors during browsing</li> <li>• Provides a possible strategy for integration with licensed content</li> <li>• Vectors will be available to other systems (e.g. keyword search)</li> <li>• Allows for shifting topic definitions</li> <li>• Avoids need to hardcode topics into metadata records to drive browsing infrastructure</li> </ul>	<ul style="list-style-type: none"> <li>• Potentially inefficient to not pre-compute and save topics associated with each object (if topics are fixed)</li> <li>• More computational load on browser</li> </ul>

*Table 21.* Comparison of approaches.

## 6 Options for CDL to Use Topic Modeling on Other Collections

The previous section described options for CDL to use topic modeling specifically for American West metadata enhancement. Here we discuss how CDL could use topic modeling on other collections. One key fact is that topic modeling is only a semi-automated process. While the metadata enhancement process can be designed to be relatively automated, the topic modeling that occurs beforehand must include significant human input to produce meaningful and usable topics. This iterative process, illustrated in Figure 1, requires that humans manually create and refine preprocessing and normalization rules and stopwords lists. This manual effort needs to be done for every new collection modeled.

There is, however, some knowledge gained during this process (e.g. having a systematic means of maintaining stopwords lists, perhaps assisted by some tool). And while some of the software components have pieces specific to a given collection, the majority of the software components are general and could be re-used for different collections. Below is a list of what software could potentially be re-used on other collections

Software Component/ Data	Specific to AmWest?	Comments
Topic Model	No	Topic Model uses word vectors produced by preprocessing scripts. Topic Model code is not specific to American West. Note: Topic Model is currently proprietary topicSeek code, and only available through topicSeek. (See Appendix L for more details)
Topic Browser	No	Topic Browser PHP scripts are general. Data in MySQL database is specific to American West, but can easily be replaced by some other collection.
Preprocessing and Processing Scripts	Yes	Preprocessing and Processing Scripts are relatively specific to American West. These scripts are, however, fairly simple and could be easily modified for other collections.
Stopword Lists	Yes	Stopword list is currently highly specific to American West. If stopwords list were managed in a more modular fashion (e.g., in a database), as recommended in this report, not all segments of the stopwords list would be specific to any given project.
Classifier	No	Classifier software can be used with any set of topic vectors and labels derived from a topic modeling process..
Metadata Enhancer	No	Implemented by CDL.

*Table 22.* Description of how software components and data can be re-used for other collections. Please see Appendix L: Description of Software for a full explanation of all software components.

## Recommendations

- **Remove non-American West objects.** Out of 360,000 objects, there may be as many as 30,000 non-American West objects. Deleting these and rerunning the model would improve the topics.
- **Improve system for managing stopwords.** Removal of stopwords clearly affects the quality of the clustering. As the number of stopwords increased from the hundreds to the thousands, there was a greater need to be able to systematically manage this list. There are several categories of stopwords including names of people, geographic locations, form/genre/type terms, etc. We recommend that a separate list be created and managed for each category, and these lists be managed in a database.
- **Expand stopwords.** We saw for the EAD finding aid collection, a large improvement in topics by continually expanding the list of stopwords. This was not done to the same extent on the American West collection because of its larger size, and the time constraints of this feasibility study. We recommend a more extensive iterative analysis of the "final" American West Project collection and a further expansion of the stopword list.
- **Use stemming.** Usage of the Porter stemming algorithm will improve the topics by combining words that come from the same word root (<http://www.tartarus.org/~martin/PorterStemmer/>).
- **Devise strategy for dealing with non-enhanced objects.** After producing a high-quality list of topics, there will be some number of objects that do not receive any topical metadata enhancement. There may be additional techniques for classifying these objects that "fall through the cracks" into one or more of the usable topics.
- **Try sub-clustering.** Several topical clusters emerged from this experiment around themes about *Buildings* (13 clusters), *Cities and towns* (9 clusters), *Dams* (7 clusters), *Japanese Americans -- Evacuation and relocation, 1942-1945* (7 clusters), *Logging* (5 clusters), and *United States -- Politics and government* (5 clusters). In these specific cases, re-running the topic model for the metadata records in these subsets of the American West Project collection would provide a more focused separation of sub-topics under these broader topical headings.
- **Begin development of a prototype classification utility to be used to automate the enhancement of American West metadata objects during the ingest process.** Since the American West Project is aiming to have a "production" version of its collection available for UI development by September 15, 2005, we recommend beginning the drafting of specifications for automating the hardcoding option (see Section 5 of this report).
- **Evaluate other options for CDL to use topic modeling.** All three options ONE-TIME HARDCODE, AUTOMATE and VECTOR, described in Section 5, should be discussed and evaluated.
- **Review aggregate threshold for proportion of object explained by topics.** We suggested that the topics assigned to each object should together explain more than 40% of that object. We recommend that this aggregate threshold be evaluated by manually examining marginal cases (e.g. when only 40-50% of an object is explained by the topics). The individual topic threshold of 5% should also be reviewed.

## Appendices

### A. Harvest Statistics

CDL used the OAI-harvesting protocol to harvest American West-scoped sets of metadata objects from a list of partner institutions. Sets were selected based on their potential relevance to the American West. Table A.1 gives some basic harvest information, and Table A.2 lists the top-10 institution-sets based on object count. Note that about 70% of the collection comes from CDL, UW and WSU.

Information	Value
Scoping	American West
Harvest date	2/7/2005
Number of objects	364,603
Format	xml containing Dublin Core

*Table A.1.* Harvest statistics.

Institution-set	Object count	Percent of total
California Digital Library - Images	150548	42%
University of Washington - All	75920	21%
Washington State University - Clippings	23191	6%
Indiana Historical Society	16988	5%
Indiana University - Cushman	14424	4%
University of Texas, Austin - Runyon Photography Collection	8085	2%
Colorado Digitization Program - Florissant Fossil Beds Nat'l Monument	7051	2%
Colorado Digitization Program - Fort Lewis College	4816	1%
Library of Congress - Panoramic Photographs	4213	1%
Western Waters Digital Library	2863	1%

*Table A.2.* Top-10 institution-sets by object count.

### B. Coverage

Table B shows the estimate of the number of objects that mention a U.S. state. The state name could occur in any of the four DC elements: Coverage, Title, Subject and Description. State names that occurred more than once in an object were only counted once. We searched for the U.S. state names as well as the Associated Press common abbreviations. This preliminary estimate indicates that 30% of the objects relate to Washington State, and 20% of the objects relate to California.

State	Object count	Percent of total
Alabama	500	-
Alaska	9600	3%
Arizona	6800	2%
Arkansas	1100	-
California	60000	20%
Colorado	24200	8%
Connecticut	600	-
Delaware	1000	-
Florida	2900	1%
Georgia	600	-
Hawaii	600	-

State	Object count	Percent of total
Idaho	3600	1%
Illinois	8500	2%
Indiana	17000	5%
Iowa	800	-
Kansas	600	-
Kentucky	200	-
Louisiana	1400	-
Maine	400	-
Maryland	400	-
Massachusetts	2300	-
Michigan	1200	-
Minnesota	800	-
Mississippi	2400	-
Missouri	1300	-
Montana	2000	-
Nebraska	600	-
Nevada	4900	1%
New_Hampshire	300	-
New_Jersey	800	-
New_Mexico	3800	1%
New_York	6600	2%
Carolinas	400	-
Dakotas	900	-
Ohio	1200	-
Oklahoma	600	-
Oregon	6100	2%
Pennsylvania	2200	-
Rhode_Island	100	-
Tennessee	400	-
Texas	10000	3%
Utah	7900	2%
Vermont	200	-
Virginias	1800	-
Washington	89900	30%
Wisconsin	3500	1%
Wyoming	1100	-

*Table B.* Number of occurrences of state name in dc:title, dc:subject, dc:description, and dc:coverage. Does not count multiple occurrences of state names in object. Total number of state name occurrences is approximately 297,000 out of 358,000 objects. Percentages less than 1% are represented as '-'. Based on single-word counts, so a) cannot separately count e.g. NC and SC, and b) cannot disambiguate Mexico the country, versus the state of New Mexico.

### C. Date

Table C shows the number of objects that relate to a particular decade, estimated from data in the dc:date element. Counts from 1990s and 2000s are likely to be digitization dates. Approximately 110,000 out of 360,000 (30%) of objects had no dc:date element. The other fourteen DC elements were not to estimate the following object counts, even though date information sometimes occurs in these fields.

Year	Object count
2000s	52200
1990s	17400
1980s	6800
1970s	8300
1960s	18800
1950s	30600
1940s	28200
1930s	46600
1920s	38000
1910s	50400
1900s	66300
1890s	30800
1880s	10200
1870s	6700
1860s	6200
1850s	2900
1840s	200
1800s	500

*Table C.* Number of occurrences of a 4-digit year in dc:date, grouped by decade. Total number of 4-digit years found in dc:date is approximately 420,000. Counts from 1990s and 2000s are likely to be digitization dates.

**D. Language**

Language	Object count
English	59000
Thai	860
German	170
French	140

*Table D.* Number of occurrences of words in dc:language. Total number is approximately 60,000. In most cases, objects identified as non-English are likely to be non-AW.

**E. Type**

Type	Object count
image	290000
photograph	59000
text	45000
clipping	23000

*Table E.* Number of occurrences of words in dc:type (top-4 listed).

**F. Other DC Elements**

Counts of word frequencies in all 15 DC elements are given in [http://www.topicseek.com/cdl/freq\\_\\*.txt](http://www.topicseek.com/cdl/freq_*.txt), where \* denotes any of the 15 DC elements.

This preliminary examination of the various DC elements indicated the most promising elements to provide useful information relating to subject. We decided to initially investigate and compare the results of using the four DC elements {Title, Subject, Description, Coverage} versus the three DC elements {Title, Subject, Description}.

### G. Additional Harvest Statistics

DC element	Word count
contributor	369000
coverage	539000
creator	555000
date	469000
description	5845000
Format	1493000
identifier	3447000
language	60000
publisher	1653000
relation	5003000
rights	3044000
source	1829000
subject	4817000
title	2808000
type	589000
TOTAL	32520000

*Table G.* Total number of words in each DC element. Note: run3 contains 12.3 million words in total, and run4 contains 12.8 million words in total. The addition of the "Coverage" DC field (going from run3 to run4) only increased the total number of words by 4%.

### H. Stemming

Stemming could be applied to reduce the vocabulary and improve the topic clusters. Standard stemming algorithms (e.g. the Porter stemming algorithm, available at <http://www.tartarus.org/~martin/PorterStemmer/>) use a list of grammar rules to identify words that have the same root. Table H shows a selection of the stemmed run3e vocabulary.

List of words that map to same word root
ARRESTED ARREST ARRESTS
ARRIERE
ARRIGO
ARRIVED ARRIVAL ARRIVALS ARRIVE ARRIVES ARRIVING
ARRONDISSEMENT
ARROW ARROWS
ARROWHEAD ARROWHEADS
ARROYO
ARSENAL
ARSENIC
ARSON
ART ARTE ARTS
ARTANDREVOLUTION
ARTEMISIA
ARTERY
ARTESIA
ARTESIAN
ARTHUR
ARTICHOKE
ARTICLE ARTICLES
ARTIFACTS ARTIFACT
ARTIFICIAL ARTIFICIALLY
ARTILLERY
ARTIODACTYLA
ARTIS
ARTISANS
ARTIST ARTISTIC ARTISTS

Table H. Sample of stemmed vocabulary generated using the Porter stemming algorithm.

I. Object wsu-clipping-12499.xml

<pre> &lt;dc oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" dc="http://purl.org/dc/elements/1.1/" schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd" xsi="http://www.w3.org/2001/XMLSchema-instance"&gt; &lt;title&gt;State history. Columbia Basin. Columbia Basin Planning Commission. 1936-02-19. p. 2.&lt;/title&gt; &lt;title&gt;Oregonian ; 1936-02-19&lt;/title&gt; &lt;title&gt;Planning commission outlines scheme in report on Columbia basin&lt;/title&gt; &lt;description&gt;Planning commission outlines scheme in report on Columbia basin. - Supreme Court decision on TVA opens way for announcement of proposal meant to aid development / by Lee Bostwick.&lt;br&gt;Creation by congress of a Pacific northwest power agency as a public corporation to develop the transmission and sale of power from Bonneville and Grand Coulee dams was recommended yesterday to President Roosevelt by the Pacific northwest regional planning commission in its Columbia basin report.&lt;/description&gt; &lt;subject&gt;State history ; Columbia Basin ; reclamation ; irrigation ; Columbia Basin Commission ; Columbia Basin Planning Commission ; President Franklin Delano Roosevelt ;&lt;/subject&gt; &lt;date&gt;1935&lt;/date&gt; &lt;identifier&gt;sh76-434-2&lt;/identifier&gt; &lt;subject&gt;Columbia River Watershed&lt;br&gt;Irrigation--Washington (State)--Columbia River Valley&lt;br&gt;Columbia Basin Commission (Wash.)&lt;br&gt;Pacific Northwest Regional Planning Commission&lt;br&gt;Roosevelt, Franklin D. (Franklin Delano), 1882-1945.&lt;/subject&gt; &lt;type&gt;Clippings&lt;br&gt;Text&lt;/type&gt; &lt;format /&gt; &lt;format&gt;Image/JPEG&lt;/format&gt; &lt;source&gt;State history box 76&lt;/source&gt; &lt;language&gt;English&lt;/language&gt; &lt;relation /&gt; &lt;rights /&gt; &lt;creator /&gt; &lt;date /&gt; &lt;identifier&gt;http://content.wsulibs.wsu.edu/clipping/image/14137.jpg&lt;/identifier&gt; &lt;/dc&gt; </pre>
--

Table I. Object wsu-clipping-12499.xml.

**J. Proportion of Object Explained by Topics**

```

<title>In the foreground is a field of tomatoes and directly behind that is a field of beans, two of the many truck
crops raised under the direction of Yoshimi Shibata by the Midwestern Farm Company, owned by three resettlers,
which raised 100 acres of truck crops on three pieces of farmland near Bartlett, Lombard, and Melrose Park, Illinois.
The land was leased from the three farm owners on a share-rental basis. Crops included tomatoes, melons, carrots,
onions, beans and pickles. The pickles were the most successful crop and onions proved the least financially
successful. On the whole, Mr. Shibata and the 12 regular men who worked with him throughout the season were
pleased with the venture. Sometimes as many as 20 extra workers were hired at the peak of the harvest. Mr. Shibata
was a greenhouse man and a farmer at Mt. Eden, California, prior to evacuation and came to Chicago from the Tule
Lake Relocation Center. In the background may be seen the buildings on the farm at Melrose Park where all three
resettlers reside. Photographer: Iwasaki, Hikaru Melrose Park, Illinois. 9/?/44</title>
<subject>War Relocation Authority Photographs of Japanese-American Evacuation and Resettlement</subject>
<subject>Series 13: Relocation (continued)</subject>
<subject>Group 40</subject>
<publisher>The Bancroft Library. University of California, Berkeley.</publisher>
<type>image</type>
    
```

*Table J.1.* Object cdl-images-83026.xml.

Percent of object explained by topic	Supplied topical label	Words in topic
36%	Japanese Americans -- Evacuation and relocation, 1942-1945	evacuation resettlement relocation japanese_american war_relocation_authority homes places
31%	Agriculture	agricultural farms farm agriculture farming harvesting farmers laborers plantations croplands machinery
10%	Land	land lands acres public sale sales timber property clearing acre ellensburg commissioner leases
77%	TOTAL	

UI. Percent of cdl-images-83026.xml explained by top topics. Here we see that, in total, the top three topics explain 77% of this object.

**K. Impact of Stopword List Expansion Using EAD Topics (50-Topic Run)**

#	EAD Run 1
1	male female child ireland adult egypt san francisco dixon tree building interior korea house ravine field steep car window chair man road marin automobile cable
2	committee minutes commission meetings legislation governor federal administration bill development releases staff bureau planning control agency legislative agencies proposed thereunder services activities members organization personnel
3	ship crowd shown mayor president standing lieutenant stand men class san building captain speaks francisco hotel sergeant private stands harry car wife man sitting group
4	trees mountains automobiles forests facilities stores clothing shops dress lakes men signs streets ponds sports snow parks roads recreation national children resorts women hotels animals
5	found early order period time organized group divided interest large collected individual bulk important activities represented collections documentation chronological alphabetical dates created topics major provide
6	water river district dam irrigation valley san power contents construction creek engineering supply prepared maps development proposed reservoir reclamation land engineer basin canal joaquin flood
7	richard paul cast walter oranges joseph herbert michael peter lee miller wilson fred ruth donald unknown arthur helen roy williams margaret bill harry moore edward

#	EAD Run 1 (continued)
8	national president committee political campaign club law issues service housing health member labor council college government chairman policy international speech business democratic party act executive
9	haynes march smith april anderson february edward johnson frederick clark adams francis louis league edwin robinson joseph jones stevenson thompson young samuel theodore ernest brown
10	lord sir french prince king russian von german senator miss edward refers gen pasha kaiser witte colonel baron tsar dillon typescripts fox sage signor cuttings

Table K.1. Top ten topics for EAD, run 1.

#	EAD Run 4
1	tree building house ravine car chair window road man automobile scene cable wall market stone asia woman pedestrian store war sign sidewalk table horse fence
2	trees mountains automobiles forests facilities stores shops clothing dress signs lakes men streets ponds sports snow parks roads recreation children resorts hotels animals furniture shrubs
3	ship president men attorney car market neighborhood signs man table conference war building labor children background judge police speaking hands chief hotel sit women committee
4	committee commission governor legislation bureau federal act legislative bills agricultural resolutions agriculture chief appointments executive government hearings law newspaper labor regulations legislature safety tax insurance
5	administrative development planning administration organization staff personnel proposals training meeting study operations grant budget youth management service funding policies procedures conferences policy assistance regional community
6	water district irrigation committee development association valley river reclamation commission power engineers land municipal conservation conference engineering meeting construction engineer flood districts basin specifications contract
7	time place due interesting contract country details called past twenty composed earlier code death deal put close support ten word importance finally gave entire total
8	committee national labor council union president housing political workers organization conference campaign international executive community chairman strike association election movement congress farm committees convention policy
9	chicano literature college conference academic graduate mexican community professor student development association curriculum faculty teaching national educational dealing committee students language president literary hispanic future
10	railroad construction gas engineer railway electric valuation financial highway power telephone operation companies exhibit road lines property district oil commission utilities equipment maintenance survey traffic

Table K.2. Top ten topics for EAD, run 4. Comparing this with Table J.1 we see that, after 3 iterations of expanding the stopword list, we end up with more coherent and usable topics.

## L. Description of Software

Component	Availability	Language(s)	Description
Topic Model	Not yet available. Currently proprietary to topicSeek.	C/C++	(Step B in Figure 1) The Topic Model is the core software component that computes the topics in a collection of text objects, and assigns topics to each object. Currently, the Topic Model is only available by retaining topicSeek. Licensing agreements (where the Topic Model is run independently of topicSeek) may be possible in the future.
Topic Browser	Freely available to CDL	PHP and MySQL. (& Apache) Uses Swish-e (* this is installed at CDL)	(Step C in Figure 1) The topic browser has been implemented as a web-based tool for analysis and visualization. Simple PHP scripts are used to extract pre-computed topical information that is stored in a MySQL database. The topic browser can operate on any Windows or Unix/Linux system that supports PHP and MySQL. The topic browser has been tested on Windows and Unix/Linux systems running the Apache webserver. Also, Swish-e was used for indexing (freely available at <a href="http://swish-e.org/">http://swish-e.org/</a> ). This code is freely available to CDL, and currently operating on harvest-dev at <a href="http://harvest-dev.cdlib.org/aw/">http://harvest-dev.cdlib.org/aw/</a> .
Preprocessing and Processing Scripts	Freely available to CDL	Perl	(Figures 6 and 7) The Perl scripts that have been specifically developed for CDL's American West project are freely available to CDL.
Classifier	Available to CDL in next phase	Perl or C/C++	(Figure 7) If CDL wanted to pursue Option 5.2 (AUTOMATE), topicSeek could develop and deliver Perl or C/C++ software to do object classification.
Metadata Enhancer	To be implemented by CDL		(Figures 6 and 7) Once topics have been assigned to each object (either using ONE-TIME HARDCODE or AUTOMATE) it is up to CDL to enhance the metadata objects.

*Table L.* Description of all software used in metadata enhancement feasibility study.