Query Parameters for Harvesting Digital Video and Associated Contextual Information

Gary Marchionini, Chirag Shah, Christopher A. Lee, Robert Capra
School of Information & Library Science
University of North Carolina
Chapel Hill, NC 27599-3360
{march, chirag, callee, rcapra}@ils.unc.edu

ABSTRACT
Video is increasingly important to digital libraries and archives as both primary content and as context for the primary objects in collections. Services like YouTube not only offer large numbers of videos but also usage data such as comments and ratings that may help curators today make selections and aid future generations to interpret those selections. A query-based harvesting strategy is presented and results from daily harvests for six topics defined by 145 queries over a 20-month period are discussed with respect to, query specification parameters, topic, and contribution patterns. The limitations of the strategy and these data are considered and suggestions are offered for curators who wish to use query-based harvesting.

Categories and Subject Descriptors
H.3.7 Information Storage and Retrieval: Digital Libraries, H5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous.

General Terms
Management, Documentation, Design.

Keywords
Digital curation, harvesting, video mining.

1. INTRODUCTION
Curators in digital libraries, archives, and museums add value to their collections through selection, exhibition, and enhancement activities. The growing availability of digital video offers curators new opportunities to add value to collections in several ways. First, video may serve as primary content in a collection. Documentaries of people, events, or places that are core to the collection provide multisensory information to patrons. For example, a presidential library may include footage of speeches or important events in the life of the president. Secondly, video may serve as contextual information for other content. Video can provide background information that supports interpretation and understanding. For example, a digital library of data and papers by a scientist might also include video from news stories of the time that help a user understand why a discovery or idea was important at that point in history.

JCDL'09, June 15-19, 2009, Austin, Texas, USA.

It is especially important that such assets be gathered today because video has become an important and increasingly common mode of expression. Given the quantity and variety of digital video on the web today, techniques that help curators harvest useful content and context would be highly beneficial.

Harvesting may be done in a pre-coordinated manner through agreements to participate in consortia or adoption of open standards such as the OAI-PMH (http://www.openarchives.org/). Alternatively, systematic link following using open source (e.g., Heritrix http://crawler.archive.org/) or customized crawlers (e.g., [8,9] can be used to turn the classical pearl-growing search strategy into a harvesting strategy. For digital video circa 2009, which is heavily dominated by YouTube, we argue that repeated issuing of highly specified queries is an effective harvesting strategy. Such query-based harvesting depends on YouTube's API and ranking algorithms, and is thus less precisely controlled than pre-coordinated harvesting, however, it requires far less coordination among institutions and individuals. In this paper, our first goal is to illustrate the applications and limitations of query-based harvesting of digital video.

A second argument presented here is that harvesting digital video bits is not enough—curators must capture context that helps future users interpret the video. We suggest that the dynamic nature of WWW-based video content makes usage data one kind of contextual data that also can be harvested. Preservation scholars (e.g., 4,5) have long argued that access and use are crucial to digital preservation. Thus, investigating usage patterns as contextual information (CI) is a second goal of the work presented here.

More specifically, this paper describes our efforts to harvest video using a query-based technique for several kinds of topics and presents results based on 20-months of daily harvests of YouTube videos for these topics. Our focus in this paper is on key parameters related to queries, topics, and contributors and how these parameters affect harvesting outcomes. The paper is organized as follows: First the potentials and challenges of digital video harvesting and collection of contextual information are discussed. Next, a system developed to automate the process (ContextMiner) is briefly described. The results with respect to queries, topics, and contributors are summarized. Recommendations for curators who undertake such harvesting are then provided. Finally, we discuss future work.

2. CONTEXT AS AUGMENTATION
In addition to the traditional intellectual activities of selection and appraisal, curators of digital collections also decide how much and what types of CI to capture or create, in order to support future understanding and sensemaking. Librarians and archivists have traditionally provided CI by creating professionally crafted

catalog records, finding aids or other surrogates. However, we believe that computer-supported harvesting of CI from online sources will become increasingly important as (1) Web content plays a significant role either in phenomena that should be documented or as documentary evidence of those phenomena; and (2) the items themselves provide limited contextual information. The appropriate amount and types of CI to augment a collection and the strategies used to collect it will be highly dependent on collecting mission and available resources. One of the most scarce and valuable resources is the attention of human professionals. Professional judgment will continue to be as essential as ever, but human-generated CI is far too expensive to do the whole job of contextualizing collected materials. Thus, tools and strategies for convenient semi-automated harvesting of information from the Web will also be essential.

2.1 Collection Development, and Context

Two difficult curation decisions are (1) which and how much CI to incorporate into a collection, and (2) how then to create or capture this CI in a cost-effective way. Future users of today's digital materials may have full physical access to the bits of the target digital objects (TDOs)—i.e., objects that are the main focus of collecting activities--but they will not have access to the state of the world as it influenced or was influenced by the bits over time. A contextual entity (CE) is something in the world (object, agent, occurrence, purpose, time, place, form of expression, concept/abstraction, or relationship) that could be related to a TDO as part of its context. In order to ensure that information about a given CE (or context through which a digital object passes) is to persist over the long term, the information should be embedded in a resource that is ingested into an archive [6]. In this paper, we specifically discuss the harvesting of information from YouTube as a way to gather CI for possible ingest into repositories.

Full access to a past world state is impossible, but it is possible and often desirable to create or capture CI in order to promote the future understanding and meaningful use of materials in a collection. CI can be collected from the "primary use environment" of the TDOs, i.e. the environment in which a primary source is directly posted and intended to be used (e.g. from YouTube when one is building a collection of YouTube videos); or from "secondary use environments," i.e. any use environment outside the primary use environment. In other studies, we have examined the collection of CI from secondary use environments (blog postings that discuss the YouTube videos being collected) and the relevance of these harvests over time [2][3]. In this paper, we focus specifically on information harvested from YouTube itself.

Two important types of CI in harvested Web content are (1) the inherent hyperlinks that connect distinct digital objects and (2) traces created when digital objects are used (e.g. annotations, ratings, use statistics). Maslov et al. [7] provide a case study of how a digital library leveraged social networking and the WWW as platform to make an existing digital collection more accessible and useful. A foundational assumption for our work in the VidArch Project[1] is that hyperlinked materials and traces from users can be harvested by curators to augment their collections so that future generations can better access, use and understand them. Moreover, we believe that these and other usage patterns will themselves become important content in digital libraries.

2.2 Digital Video as Context

Digital video is increasingly common on the WWW, and YouTube is currently by far the largest public repository. According to Nielsen Online, YouTube alone served more than 5 billion video streams to more than 82 million unique visitors in the U.S. in October 2008, which represents the bulk of the almost 9 billion total video streamed in that period in the U.S. (http://www.nielsen.com/media/2008/pr_081217a_download.pdf).

As more events are documented with video and more people use video to express themselves, video increasingly acts as a primary material for repositories as well as CI for primary materials. Regardless of whether these videos are primary objects of collection or ancillary augmentations to help contextualize the primary objects, the video files alone are not likely to be sufficient for many future use cases. Without considerable contextual information, future users may be able to view the video but not understand or make sense of it. Although the declining costs of acquisition and storage devices support serious efforts to capture massive amounts of CI associated with everyday experiences, there will always be both practical and professional reasons for curators of digital collections to engage in selection.

There are many types of CI that can be captured at the time of creation. Within YouTube, this includes provenance metadata such as the name and profile of the account through which the video was posted, and the date it was posted. For digital video created in the near future, many other contextual elements are likely to be automatically captured at creation time. For example, geospatial, temporal, and other ambient physical states (e.g., temperature, radiation, chemical) will likely be automatically captured by video recording devices.

Many other types of CI take the form of system state at usage time. Examples include click streams, query logs, whether a user watched the entire video after selecting it, and what browser the user was using. The company or organization responsible for hosting the primary use environment– in this case, Google as the host of YouTube – usually has access to much more of this CI than outside parties. However, there are some forms of system state information that can be harvested by external parties, based on one-time or periodic snapshots. In YouTube, these include various forms of information that are dynamically presented on the page for a video at the time it is requested by a user agent, for example, comments and ratings. Additionally, temporally dependent information inherent to retrieval such as rank in the results list can also be harvested. We can envision a kind of usage timeline that begins with the video object's creation and extends over its usage lifetime – much like the Internet Archive's WayBack Machine interface, but incorporating usage and social dimensions beyond time and URL. This implies an archive that grows periodically as new usage event data is appended.

3. THE CONTEXTMINER APPROACH

To assist curators in harvesting videos and associated CI, we developed a strategy and set of tools. The strategy selected was query-based harvesting rather than link traversal crawling [11]. This strategy depends on APIs or search services provided by the primary use environment hosts. This is an acknowledged limitation on two fronts: APIs change or are restricted, and ranking strategies are typically proprietary. Regardless of these

[1] http://www.ils.unc.edu/vidarch/

limitations, this technique provides inexpensive sample streams of realistic and contemporaneously popular video related to the queries

ContextMiner has been in development since early 2007 and its architecture and technical details have been discussed in several venues (see project website for papers and demos). Here we focus on its use with YouTube campaigns run as part of our collaboration with the Library of Congress and the National Digital Information Infrastructure Preservation Program. ContextMiner lets a curator specify a set of queries for a topic (we refer to this collection of queries as a campaign) and a set of parameters for how often the queries are executed (e.g., daily or weekly), the number of results to harvest, and the primary use environment hosts (web sites) that should be queried (in this project, only YouTube data were harvested, however, the ContextMiner system also supports blog aggregator and Twitter, sources.

Using ContextMiner, we have issued 145 queries almost every day on YouTube since May 2007. Table 1 presents the exact queries run for each of the six topic campaigns. The specific query sets were developed based on group discussions. For the election topic, we selected all candidates listed in the Wikipedia article for U.S. Presidential election in May of 2007 and augmented the list of 52 names with six general queries such as 'election 2008', 'United States election 2008', and 'campaign 2008.' For the other campaigns, we aimed to select topics that would be of general interest and also might show large variations in activity. We also aimed to create query sets that were somewhat parsimonious with respect to alternative terminology and some that were more comprehensive. For multi-term queries, we used the phrases with and without quotations. Thus videos for a topic could appear in the top 100 results for either different queries (e.g., natural gas or green power) or for the same query expressed with the quotation syntax (e.g., Avian flu or "Avian flu").

Each day as ContextMiner runs each of these queries on YouTube, it extracts the top 100 results for each by relevance as given by YouTube. Thus, if there are 20 queries for a given topic, there will be up to 20*100 videos obtained, some of which overlap. From these videos, it stores only those that were not already collected**.** While storing a new video, ContextMiner also extracts available metadata related to the video, such as title, tags, author name, and genre. We also downloads the video in Flash format and later convert it to MPEG as a part of our agreement with the Library of Congress to provide all harvested video for historical purposes. Immediately after collecting new videos, ContextMiner revisits all the videos in the collection and records several indicators of social and temporal context. These indicators include the current total number of views, comments, and ratings.

Since all videos are "discovered" by issuing queries to YouTube, we also track the rank of the video in the list of results for that query. Each time the queries are run, we note the ranks of the videos and record them in the database. In order to support this ability to track the rank of a video across time, a video may be recorded multiple times in our database if it was "discovered" from different queries. For example, if video XYZ were discovered based on a query for Barack Obama and then at some later point discovered again from a query for Hillary Clinton, it would be entered in the database twice. This allows us to track its rank across time independently for each query. For example, it might be ranked #1 for the Barack Obama query but only ranked #8 for the Hillary Clinton query. We also store the unique YouTube video identifier with each record so that we can identify multiple instances of the same video. This allows us to compute

Table 1. Queries for Each Topic Campaign

Election	**Energy**	**Epidemics**
election 2008	Bioenergy	epidemic
US election 2008	Biofuel	pandemic
United States election 2008	Biomass	SARS
presidential election 2008	Coal	Avian flu
campaign 2008	Crude Oil	"Avian flu"
decision 2008	"Crude Oil"	
Joe Biden	Electric Power	
Hillary Rodham Clinton	"Electric Power"	**Diabetes**
Christopher Dodd	Fossil Fuels	Diabetes
John Edwards	"Fossil Fuels"	Type 1diabetes
Mike Gravel	Fusion Energy	"Type 1diabetes"
Dennis Kucinich	"Fusion Energy"	Type 2 diabetes
Barack Obama	Gasification	"Type 2 diabetes"
Bill Richardson	Geothermal Energy	Gestational diabetes
Wesley Clark	"Geothermal Energy"	"Gestational diabetes"
Al Gore	Geothermal Power	Pre-diabetes
Tom Vilsack	"Geothermal Power"	Diabetes Cholesterol
Sam Brownback	Green Energy	"Diabetes Cholesterol"
John H. Cox	"Green Energy"	Diabetes Heart Disease
Jim Gilmore	Green Power	"Diabetes Heart Disease"
Rudy Giuliani	"Green Power"	Diabetes Stroke
Duncan Hunter	Hydrogen Energy	"Diabetes Stroke"
John McCain	"Hydrogen Energy"	
Ron Paul	Hydrogen Power	
Mitt Romney	"Hydrogen Power"	**Natural Disasters**
Tom Tancredo	Hydropower	natural disasters
Tommy Thompson	Natural Gas	"natural disasters"
Mike Huckabee	"Natural Gas"	avalanche
Newt Gingrich	Nuclear Energy	drought
Chuck Hagel	"Nuclear Energy"	earthquake
Fred Thompson	Nuclear Power	flood
James Gilchrist	"Nuclear Power"	hurricane
Dale Thompson	Renewable Energy Sources	typhoon
Elaine Brown	"Renewable Energy Source	tornado
Kat Swift	Solar Energy	tornadoes
Ralph Nader	"Solar Energy"	volcanoes
Cynthia McKinney	Sustainable Energy	cyclone
Rebecca Rotzler	"Sustainable Energy"	landslides
Mike Jingozian	Sustainable Power	storm damage
Bob Jackson	"Sustainable Power"	"storm damage"
Steve Kubby	Wind Energy	wildfires
George Phillies	"Wind Energy"	wind damage
Christine Smith	Energy Star	"wind damage"
Doug Stanhope	"Energy Star"	
Kent McManigal	Power Utilities	
Gene Chapman	"Power Utilities"	
Barry Hess	Energy Sciences	
Robert Milnes	"Energy Sciences"	
Wayne Allyn Root		
James H. Mccall		
Donald K. Allen	**Truth commission**	
Steve Adams	Truth commission	
David Koch	"Truth commission"	
John Taylor Bowles	Truth and reconciliation	
David J. Masters	"Truth and reconciliation"	
Michael Bloomberg		

the total number of **unique** videos[2] for each campaign. In this paper, we present different analyses, some using the total number of videos and some using the only the unique videos, depending on the goals of the analysis.

4. RESULTS

This procedure resulted in 52131 unique videos retrieved up to the time of this data analysis (December, 2008). These data were analyzed with a focus on the effects that query specification and contribution patterns have on query-based harvesting. The results are organized as follows. First, the data set is summarized and general observations about the collection are offered; next the effects of query sets are considered, and finally contribution patterns are presented and discussed.

4.1 The VidArch YouTube Data

Table 2 summarizes the number of videos (both total and unique) and number of contributors for each of the six topic campaigns over the 20-month period. The data show that the number of videos posted and the diversity of sources (contributors) that provide these videos both vary by topic. The highly popular and timely U.S. presidential election topic generated the largest number of contributors and largest number of unique videos in our collection, whereas the highly specific topic, 'truth commissions' generated the fewest videos and contributors. Very general topics such as energy and disasters, not surprisingly generated large numbers of videos and many contributors, while more specific topics like diabetes and epidemics generated fewer videos by fewer contributors. Clearly, the generality and the timeliness of the topic should be taken into consideration when setting harvesting parameters.

Table 2. Summary Data for 20-months

Topic	# Queries	# Videos	# Unique Videos	# Contributors
Elections	56	27355	24585	8190
Energy	48	39174	12339	5706
Epidemics	5	1936	1755	1039
Diabetes	14	11499	1919	868
Disasters	18	16456	10943	6584
Truth Comm	4	1720	590	275
Total	145	98140	52131	22662

About 20% of the videos that we have collected from YouTube have disappeared over the 20-month period, regardless of topic. The percentages of videos taken down by topic are 25% for election, 24% for energy, 19% for epidemics, 20% for diabetes, 23% for natural disasters, and 22% for truth commissions. Many of these videos were taken down within hours or days of their publishing and others at various times over the 20-month period. Therefore, it is possible that a typical web crawler would have missed such ephemeral content. Our approach, on the other hand, harvests YouTube every day in response to specific queries, and thus, there are good chances that we archived videos that other methods of crawling completely missed.

Note that our harvesting technique is dependent on YouTube APIs (which changed over time, causing us to miss a few days while the ContextMiner scripts were adjusted), and more importantly, the results depend on YouTube's ranking algorithms, which likely also change over time. We assume that our data in aggregate is nonetheless an accurate representation of a highly fluid digital video environment.

4.2 Queries and Content

One of our initial questions was how query formulation influences harvests. We first consider the terms in the tags in videos returned for the query sets for each of the topical campaigns. Table 3 shows the 20 most frequently occurring terms for the six campaigns with terms that do not occur in the queries highlighted. These data give us some encouragement that the videos retrieved are topically related to the queries. The total numbers of tags and the total numbers of unique tags for each set of videos are also given and demonstrate both the diversity of tags and typical long-tailed distributions of tag frequencies.

Table 3. Tag Distributions by Campaign.

Rank	Election		Energy		Epidemics	
1	2008	6374	energy	6267	sars	427
2	obama	5723	power	3271	flu	288
3	election	4783	green	2118	pandemic	262
4	president	4243	solar	1609	avian	231
5	john	3961	nuclear	1398	epidemic	220
6	barack	3755	oil	1278	sar	208
7	mccain	3395	gas	1234	bird	142
8	clinton	3205	renewable	1213	the	92
9	paul	3030	wind	1201	of	88
10	ron	2862	hydrogen	1086	influenza	79
11	presidential	2764	environment	1086	health	77
12	hillary	2686	alternative	1072	world	74
13	republican	2141	fuel	1071	music	71
14	iraq	2091	science	956	2008	69
15	news	1930	global	875	h5n1	65
16	politics	1856	water	781	virus	60
17	campaign	1793	the	776	disease	59
18	bush	1674	warming	748	nwo	58
19	debate	1668	free	669	s.a.r.s.	54
20	edwards	1638	electric	668	van	53
unique tags		19228		10347		5527
total tags		264811		135135		15922
	Diabetes		Disasters		Truth Comm.	
1	diabetes	1546	tornado	1242	truth	387
2	health	608	storm	1099	reconciliatior	171
3	type	407	volcano	852	commission	170
4	heart	361	the	846	halo	162
5	disease	306	hurricane	842	and	151
6	diet	291	flood	740	11-Sep	112
7	blood	276	earthquake	690	911	88
8	sugar	247	tornadoes	626	bush	82
9	insulin	219	cyclone	610	combat	70
10	weight	216	video	607	evolved	63
11	obesity	204	volcanoes	578	war	53
12	nutrition	201	damage	567	of	53
13	medicine	193	typhoon	554	conspiracy	49
14	cholesterol	171	wildfire	547	the	46
15	loss	168	weather	526	this	42
16	food	167	avalanche	485	providence	42
17	cancer	163	landslide	480	master	35
18	cure	156	music	471	chief	35
19	diabetic	133	wildfires	445	game	34
20	stroke	130	of	445	video	33
unique tags		4034		15698		1623
total tags		22635		95047		6700

Note that these data include stop words and terms that are not stemmed, however, tags that are not in the queries are mainly

[2] More precisely, we record the number of unique YouTube IDs. It is possible that a video is uploaded by different people and each version will have a unique YouTube ID.

related to the topic. The ratios of unique tags to total tags demonstrate the diversity of queries for the topics. The election and energy topics have only 7% and 8% unique tags respectively, whereas the other topics which are represented by fewer unique words show much larger overlaps in tags (35%, 18%, 17%, and 24% for epidemics, diabetes, disasters, and truth commissions respectively.

4.2.1. Overlaps

As Table 2 shows, there was considerable overlap for videos retrieved across the different queries within given topics except for the election and epidemics topics, which both had only about 10% more total videos than unique videos. The epidemic topic campaign only used five queries, only one of which (Avian flu) had two terms and thus there was only a single duplicate query with and without quotes. Because the election topic did not use any duplicate queries with quotation marks, there were many fewer overlaps overall. An analysis of the overlaps among all pairs of queries in the election set showed considerable overlap between pairs of the general queries (e.g., US election 2008 and United States election 2008 yielded 577 overlaps and US election 2008 and presidential election 2008 yielded 455 overlaps). Among the candidate names, the queries for Ron Paul and Kat Swift yielded 171 overlaps, which was the highest number among different candidates. Not surprisingly, using variant syntax to gain more coverage for a specific query will yield more redundancy in results, which has implications for harvesting, particularly from services that limit or charge fees for access. Given the data and the amount of noise in the technique (uncontrolled API and ranking) it seems reasonable to use only one syntactic variation for phrases (e.g., only use the phrase in quotes).

4.2.2. Query Effects Over Time: Churn

Another way to look at the effects of queries on harvesting results is to consider the number of unique videos yielded over time. If a query retrieves 100 unique videos each day in the top 100 results, then the topic activity is very high and changing rapidly (high churn), whereas if the same 100 videos are in the top 100 each day, then the topic is much less dynamic. To represent this churn[3] in our data, we compute an idealized statistic by dividing the number of unique videos found over the 20-months for each topic by the maximum number of videos that could have possibly been retrieved if each query variation yielded 100 different videos. Thus for the 56 election queries, we assume a denominator of 56queries*100results*600days=3,360,000, thus yielding a churn index of 24585/3360000=0.0073. Because the data show that considerable redundancy is added by the additional queries with multiple terms, we use the number of unique queries for each topic campaign for the other 5 topics in computing the churn index. For example, the harvesting for the topic diabetes used 14 query variants (diabetes, type 1 diabetes, "type 1 diabetes," type 2 diabetes, "type 2 diabetes," gestational diabetes, "gestational diabetes," pre-diabetes, diabetes cholesterol, "diabetes cholesterol", diabetes heart disease, "diabetes heart disease," diabetes stroke, and "diabetes stroke") but for computing the churn, we only use the 8 query variants without quotation marks. Table 4 displays the churn index for the six topics and the number of videos per unique query.

Table 4. Topic-Query Churn Effects

Topic	Unique Queries	Churn Index	Videos/Query
Elections	56	0.0070	439.0
Energy	27	0.0043	457.0
Epidemics	4	0.0059	438.8
Diabetes	8	0.0023	239.9
Disasters	15	0.0101	729.5
Truth Comm	2	0.0025	295.0

The results of our queries related to disasters show considerable variability over time. The results of our queries related to Diabetes and Truth Commissions seem more stable as the top videos tend to remain consistent at almost three times the rate of the results from our disaster queries. Curators are advised to adjust their harvesting parameter settings for the churn in their topic of interest. For example, it may be acceptable to harvest less often than daily for a low churn topic; or if the objective is to collect more comprehensively, to set a higher result set threshold.

4.3 Queries Over Time and Ranking Effects

One harvesting parameter that will have strong influence on the total number of items harvested is the threshold for the number of results that are collected for each crawl. We set a threshold of 100 in our data collection, but because we recorded the ranks of videos that appeared in the top 100 results for each query for each crawl, we were able to estimate the effects of harvesting at thresholds lower than 100. Note that we do not get a rank for every video each day, because items can fall out of the top 100. Thus we have rank values for each video at least once and often many times, thus allowing a reasonable estimation over the large number of days and videos harvested. We consider rankings over time in three ways: for two specific videos, for specific election 2008 queries, and for overall topic campaigns.

It is important to note that due to the way the rank data was collected, it contains some gaps and anomalies. We describe these limitations in the section below. Given these limitations, we present the rank data here as a way to describe the challenges in collecting such data and to make general observations about *apparent* trends.

4.3.1. Rank Variability and Noisy Data

We collect rank information for each <query, video> pair in the database. For example, if we have a video that was discovered through a query for Barack Obama, we continue to track its rank in the results list for the Barack Obama query over time. If we also discovered the same video through a query for Hillary Clinton, we also keep track of its rank in the results list for Hillary Clinton.

In order to collect the top 100 videos for a given query on a given crawl, we issue two YouTube API calls: one that returns the first 50 results and one that returns the second 50. These results are then combined into a "top 100" list of videos for that query for that crawl and is stored in temporary storage. After all the queries have been run on a given day, a second process compares the day's query results to the full list of all videos that we have

[3] Note that customer churn is more complex than the simple change in rank that we use here, taking into consideration customer loyalty and human decision processes over time. See [10] for an example of these more comprehensive treatments of churn.

"discovered" from this or prior crawls for the given queries. For each <query, video> pair, the process looks in the day's top 100 list for that query and records the rank position of the video. If the video does not appear in the day's top 100 list, it is assigned a rank of 0. Thus, for many of the videos in our collection, we do not know its absolute rank for every crawl. For example, if we have collected 2500 videos for the query "Barack Obama", on any given query instance we will only have rank data for 100 of them (the top 100).

Gaps / No rank data – In our data set, we have observed strange situations in which a video will go from being highly ranked to having a zero rank and then resume its highly ranked position. An understandable situation would be for a video to gradually become less popular (rank gradually moves toward 100), then fall out of the top 100 (rank = 0), and then re-enter the top 100 at a low rank (e.g. rank = 95). But what we have occasionally observed are videos that suddenly drop from rank 5 to rank 0 (i.e., not in top 100) and then go back to rank 5. Our suspicion is that we stopped getting data about this video for the period of time when it jumped to zero. It could be that our crawlers were blocked or experienced network problems on those days, or perhaps our scripts experienced problems. Regardless of the cause, these jumps to zero represent one source of noise in our data.

Anomalous rank data – Because we request the top 100 videos for each query, the recorded ranks should be integers from 1 to 100. However, during roughly the first year of our data collection, we occasionally recorded some ranks greater than 100. Overall, from May 2007 to May 2008, approximately 1.5% of the ranks we recorded were greater than 100. In May 2008, we made changes to our crawlers to update them to work with different YouTube APIs. After this change, all recorded rank data falls below 100. Because of this, there is additional noise present in the rank data prior to May 2008 that does not appear to be present in the later data.

Black box ranking algorithm – In addition to the issues mentioned above, we do not fully understand the YouTube ranking algorithm and have observed some ranking anomalies (e.g., it is impossible to say whether a specific video ranking changes because of usage popularity or the distributional matches of a days' queries with the static metadata in the video).

To help understand these aspects of the rank data, we present examples of ranking data for two specific <query, video> pairs. The first video has been highly ranked for the "Barack Obama" query for almost the entire 20-months of tracking. Figure 1 shows a plot of the rank of this video over time. For the majority of the time, this video has been in the top 10 for the query "Barack Obama". At several points it fell below the top 10, with some spikes down to the 60s and one spike to over 100. At several points, the rank jumped from around 5 to 0 for a period of time but then resumed at 5.

The second example is a highly-ranked video for the Truth Commissions set of queries (Figure 2). This video has also remained in the top 10 for most of the time it has been tracked and has had less volatility than the previous example. Again, there was a period when we lost tracking data on this <query, video> pair, but the overall trend shows a consistently ranked video.

These two cases demonstrate that although there is considerable noise in these data, general trends over time may be discerned.

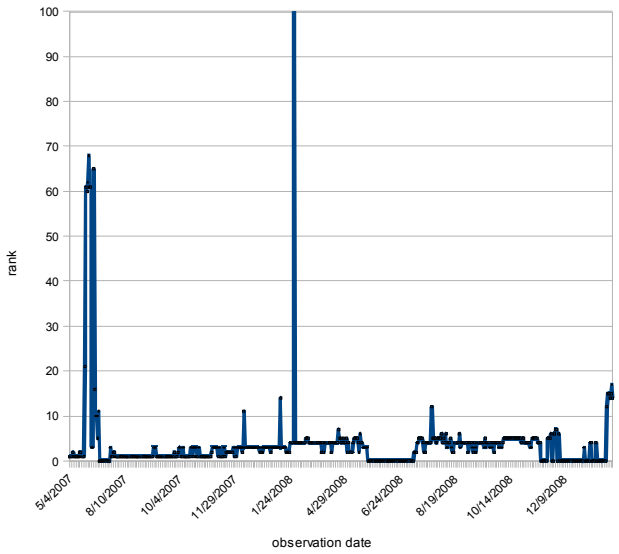

Figure 1. Ranks over Time for One Highly Ranked Video for the "Barack Obama" Query.

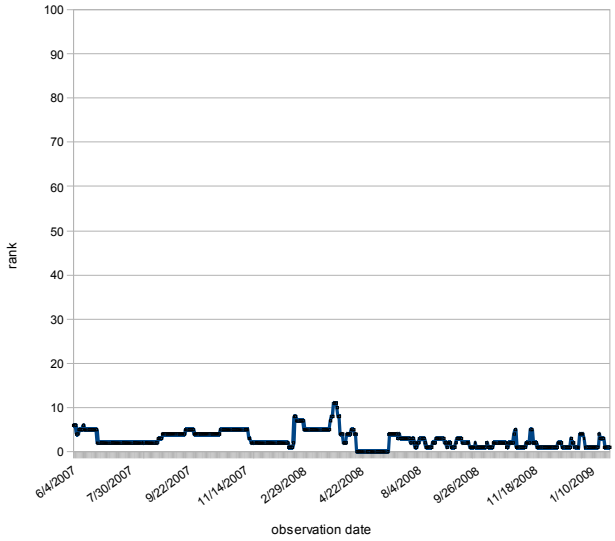

Figure 2. Ranks Over Time for One Highly-Ranked Video for the Truth Commissions Campaign

Curators may way to collect consistently highly-ranked videos and may also want to invest more manual effort in examining the comments, ratings, and related information for those videos, or collect further contextual information related to them (e.g., blog pages that link to the videos).

4.3.2. Rank Threshold Effects for Specific Queries.

Figures 3-5 show the number of unique videos retrieved by three different queries in the 2008 Presidential Election topic campaign. Note that the previous section describes important limitations of the rank data set that are important to understand when interpreting these graphs.

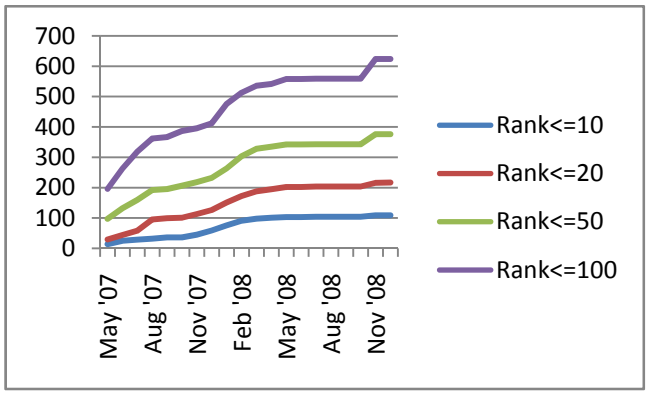

Figure 3. Unique Videos Ranked for 'Election 2008' query over 20-months for top 10, 20, 50, and 100 thresholds.

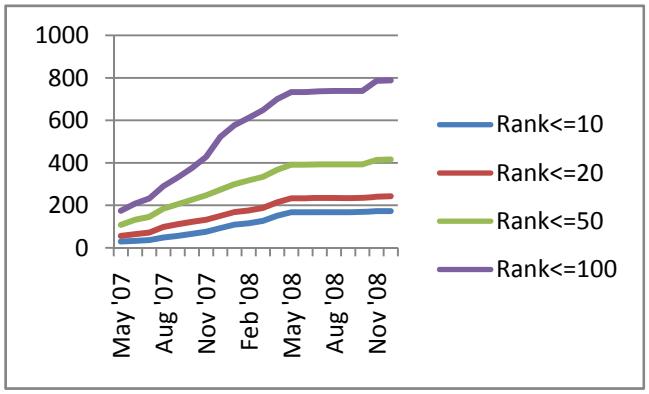

Figure 4. Unique Videos Ranked for 'Barak Obama' query over 20-months for top 10, 20, 50, and 100 thresholds.

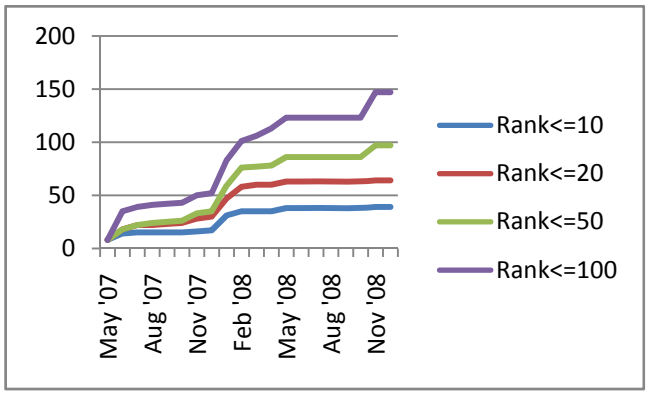

Figure 5. Unique Videos Ranked for 'Kat Smith' query over 20-months for top 10, 20, 50, and 100 thresholds.

These trends show that raising the result set threshold for a given query eventually results in diminishing returns. For the general election query, harvesting the top 20 rather than 10 exactly doubles the number of videos; whereas a threshold of 100 yields 4.7 times as many videos. In the case of the specific but very popular query "Barack Obama", there is a 41% increase for a threshold of 20 rather than 10; and 3.6 times as many videos for a setting of 100. For a specific and less popular topic (Kat Smith), there is a 64% increase from 10 to 20 and 2.8 times as many videos for 100.

These charts also show different kinds of inflections in cumulative growth over time. The two queries with large numbers of results are more continuous than the Kat Smith query that shows significant jumps just before the primary elections began in early 2008 and again just before the election. Thus, curators are advised to match specific queries with different relevance ranking thresholds, in order to reduce system and human effort.

4.3.2. Rank Threshold Effects for Topics

Figures 6-10 show curves for five topics. In these graphs, the aggregate of all queries provide a view of the effects of threshold settings for multi-query topics.

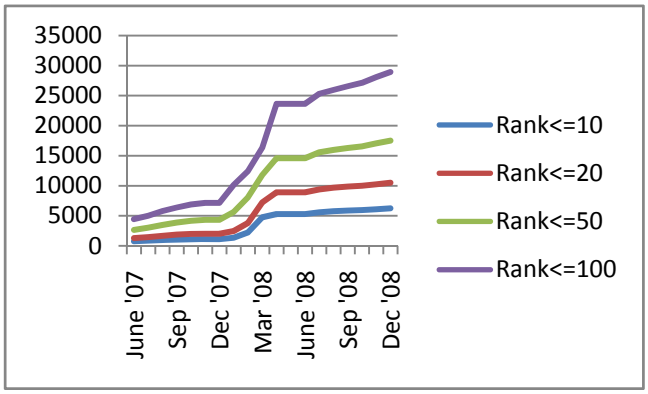

Figure 6. Videos Ranked for all energy queries over 20-months for top 10, 20, 50, and 100 thresholds.

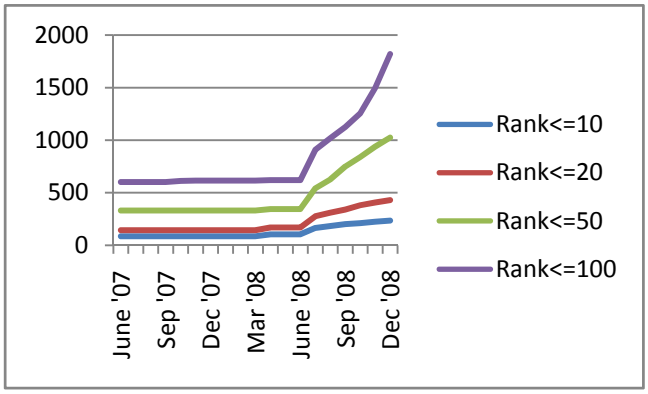

Figure 7. Videos Ranked for all epidemic queries over 20-months for top 10, 20, 50, and 100 thresholds.

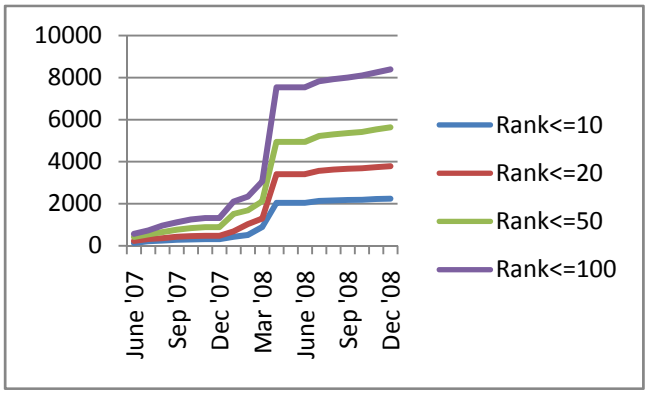

Figure 8. Videos Ranked for all diabetes queries over 20-months for top 10, 20, 50, and 100 thresholds.

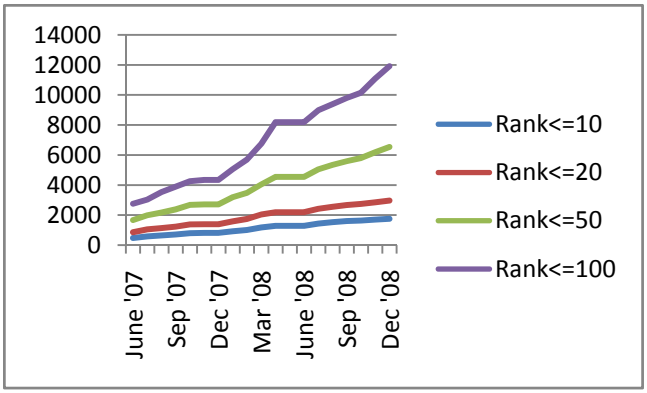

Figure 9. Videos Ranked for all natural disaster queries over 20-months for top 10, 20, 50, and 100 thresholds.

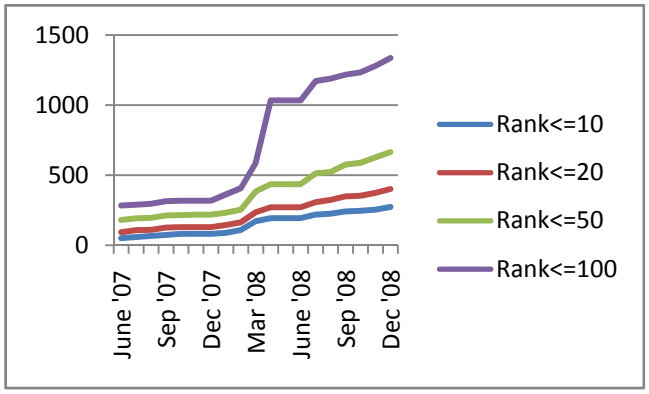

Figure 10. Videos Ranked for all truth commission query over 20-months for top 10, 20, 50, and 100 thresholds.

These charts show that for highly general topics with large numbers of videos retrieved, the differences between threshold values is fairly constant over time. However, for very specific topics with lower numbers of videos retrieved, the differences between low and high thresholds can be quite large. As with churn, topic characteristics should influence curator decisions about thresholds.

These different topical graphs show unique cumulative inflection points. For example, Figure 8 shows a large jump in the April 2008 rankings for the diabetes campaign. Looking more closely at the data, we see that the largest increases for the topic came from the 'gestational diabetes' queries and are likely a result of the news on March 28 that Angelina Jolie was reported to have gestational diabetes. It is interesting that this effect was much larger than the increased activity generated by the two major diabetes studies released on mid February 2008. Clearly, harvesting campaigns will be strongly affected by popular events and curators would be well served by making adjustments in their strategies in response to such events. Early in our project we hoped we might be able not only to detect such patterns post hoc but also to predict some (e.g., activity after presidential debates or primaries that might lead to significant campaign changes). However, the data is quite noisy and more sophisticated filtering strategies will be needed to detect other than quite gross changes.

4.4 Contributor Patterns and Stop Source Lists

Curators who aim to harvest video content from environments such as YouTube should also consider the sources of content. An experienced curator will often know the most likely creators or distributors of related materials and can make targeted requests to those sources. When harvesting from more general sources such as YouTube, curators must develop more nuanced filtering and targeting strategies to save time and effort when reviewing the harvested materials for potential ingest into an archive. Thus, another thread of work in this project was to investigate contributor patterns.

In addition to targeting specific sources for specific queries (e.g., the Library of Congress strategy of harvesting 2008 presidential election videos from the candidates' official campaign offices), we expect that curators will want to develop 'stop source lists' (akin to stop word lists for text retrieval) that limit the videos that are harvested from specific primary use environments. These lists can be established either *a priori* based on knowledge about the topic and key sources, or *a posteriori* as data collection progresses. *A priori*, it is less clear what decisions should be made about inclusion or exclusion. For a collecting campaign related to epidemics, it is perhaps obvious to exclude videos from a music group named 'Epidemic" or the song titled "Epidemic" by the group Slayer. However, decisions about popular news or topically pertinent sources are more dependent on collecting goals and available resources. For example, when building collections related to energy policy, one repository might already have a mechanism in place for obtaining materials from the U.S. Department of Energy, and might, therefore, wish to filter any DOE videos out of their collections from YouTube; whereas another repository without an established pipeline for receiving DOE materials might wish to directly target DOE videos for collection from YouTube. Filters and targets should often be adjusted, based on data from the results of queries over time. In this section, we present data about the contributors of the videos we harvested.

There were more than 20000 contributors for the more than 50000 unique videos in our collection[4]. Not surprisingly, these contribution patterns show long-tail distribution patterns overall with a very small number of contributors (sources) providing multiple videos and most contributors providing a single video. For example, of the 5706 contributors in the energy campaign data, 2502 (44%) contributed exactly one video.

To further investigate the data we conducted a search in YouTube for each unique contributor to determine how many videos this source provided overall. These 20000 contributors contributed more than 2.7 million videos to YouTube. This indicates that on average, a contributor in our list posted 135 videos to YouTube. Given this base, we compare patterns within different topics. An average contributor from the election topic campaign posted only 94 videos on YouTube. Thus, these contributors' activity on other YouTube topics was low compared to the average contributor in our entire data set. However, these contributors had the highest number of topic-related videos per contributor (more than 3 each). This shows that active contributors to the election topic were not

[4] Although the contributors in each topic are unique in this table, some contributors appear in more than one topic so there are somewhat fewer than 22662 unique contributors.

particularly active contributors to YouTube in general. We can call these contributors "election mavens".

Table 5 displays the number of contributors for each topic and the number of videos we harvested for those topics. Column three (local) reports the number of contributors in our collection. Column four presents the number of videos contributed to YouTube by all of these contributors. Column five provides the range of number of videos contributed by all contributors for each topic. Thus, for the election topic, most contributors posted one video, but one contributor posted 645 videos (BarackOmabamdotcom). For the energy topic, most contributors posted one video, but one contributor posted 1099 videos (etvmedia). The sixth column shows the range of number of videos contributed to YouTube by all contributors who also provided videos that appear in our database. For the epidemics topic, CBS provided two videos that were harvested by our queries and 13811 total videos to YouTube. The final column presents the most frequent YouTube poster among our contributors and the number of those videos we harvested for each topic. TourFactory is a service that does video house tours for realtors and homeowners and has posted almost 50,000 home tours to YouTube. On January 27, 2008, one of their videos appeared in the result list for the query "David J. Masters" (a possible U.S. presidential candidate). We believe the video appeared for this query because it had 'David' as one of the tags, and 'master bedroom' in its description. This video[5] was later removed. This vignette shows the importance of using and continually monitoring heuristic rules to filter harvesting. The case of Expertvilliage is more predictable, because it is a major contributor of how-to videos to YouTube on a variety of topics (including e.g. diabetes).

Clearly, some of these sources could be stop listed or targeted *a priori* (e.g., specific news channels). However, other sources can be quite surprising and curators will want to use regular monitoring after harvesting begins to cull or target specific contributors. As we gain more experience with managing large collections of harvested data, alerts and filters can be added to help curators as they monitor incoming data streams.

Table 5. Contributor Patterns

Topic	# Contributor	# Videos Local	# Videos YT	Local Range	YT range	Most Frequent
Elections	8190	23524	771022	1--645	1--49034	TourFactory (1 local)
Energy	5706	35002	776997	1--1099	1--137969	expertvillage (19 local)
Epidemics	1039	1698	122697	1--50	1--13811	CBS (2 local)
Diabetes	868	7367	266402	1--580	1--137969	expertvillage (535 local)
Disasters	6584	13550	814888	1--195	1--137969	expertvillage (2 local)
Truth Comm	275	923	40592	1--48	1--6815	AlJazeeraEnglish (28 local)

5. RECOMMENDATIONS FOR CURATORS

Curators should carefully consider queries, sources, crawling parameters and how they relate to the intended collecting goals. Ongoing monitoring and assessment of data is important to make adjustments and gauge the costs of assessing potential videos for inclusion in the collection. Decisions about how much contextual information to harvest and how much to assume will be preserved elsewhere is also an important issue. We summarize the recommendations that come from our data as follows:

- Take the generality and the timeliness of the topic into consideration when setting harvesting parameters. This applies to query formulation as well as setting harvesting parameters such as frequency of crawls and result threshold settings.
- If variant syntax is used to gain more coverage for a specific query, it will yield more redundancy in results, which has implications for harvesting, particularly from services that limit or charge fees for access. Given the noise in query-based harvesting, it is advisable to focus on synonymic variations rather than syntactic variations in queries.
- Adjust harvesting parameter settings based on the churn in the topic of interest. It may be acceptable to harvest less often than daily for a low churn topic; or if the objective is to collect more comprehensively, to set a higher result set threshold.
- Extract consistently highly ranked videos and then invest more manual effort in examining the comments, ratings, and other related information for those videos.
- Adapt filters and targets on an ongoing basis by using and continually monitoring heuristic filtering rules.
- When attempting to build collections about popular events, make adjustments in the harvesting strategy over time in response to unanticipated related events.

These recommendations must be considered with respect to the time, resources, and effort available to monitor the harvesting and examine the results. Gathering thousands of videos without the time to review their relevance to the collection is likely to be a waste of resources for most collecting institutions. Note that although we have presented YouTube data, many of the harvesting strategies apply for other video sources or web-based materials.

6. CONCLUSIONS AND FUTURE WORK

We have developed ContextMiner APIs to share our data under a Creative Commons License with the research community, so that such rich data collected over nearly two years can be further analyzed and used[6]. We have also made several of our tools available freely with open source to promote research and education on this topic. These include TubeKit[7] and the public beta of ContextMiner.[8]

TubeKit is a query-based YouTube crawling toolkit that allows one to build one's own harvester, such as those described in this paper, that can harvest YouTube based on a set of seed queries and collect a number of attributes. TubeKit assists in several

[5] It was available at http://www.youtube.com/watch?v=j-y8o1KLmX4 with title "4700 Charmwood" and genre "Travel & Events".

[6] Note that videos that have been removed from YouTube will not be shared openly, however, they will be provided to the Library of Congress as part of the historical record.

[7] http://www.tubekit.org/

[8] http://www.contextminer.org/

phases of this process starting with database creation to finally providing access to the collected data with browsing and searching interfaces. Now in its third release of public beta, TubeKit is used by several research groups and organizations, not only for collecting valuable data from YouTube, but also in making sense of it.

The ContextMiner public beta extends our ContextMiner framework to include blog and Twitter sources and in-link information. It has been available for general use since July 2008.

At the time of writing this, there are nearly 200 accounts set up to use ContextMiner. Users have created more than 300 campaigns with more than 600 queries, and have collected millions of objects (YouTube videos, blogs) and related contextual information. ContextMiner is also in use by several members of the National Digital Information Infrastructure Preservation Program (NDIIPP)[9] and can be used by teachers or others who wish to harvest content on specific topics. Further development is underway to provide access to more sources and tools for information exploration. ContextMiner is available as open source code or as a web-based service.

More theoretically, we will continue to collect data with an eye toward understanding the applications and limitations of query-based harvesting and what curators can do to maximize their time as they assess the harvested streams. We hope to gather impressions of curators and teachers who are using ContextMiner to harvest content to understand how they are using the harvested streams.

More generally, although we are not able to predict events based on usage patterns in the data collected in this project, we look forward to examining patterns in the comments and related videos in the future. At present, YouTube comments offer more noise than substance; however, as considered commentary becomes more common, text analysis can be applied to better classify videos and support inferences about trends. Furthermore, as video comments become more commonplace, video similarities based on color, optical flow, sound intensity, and other media features can be leveraged for similar purposes.

7. ACKNOWLEDGEMENTS

This work has been supported by a grant from the National Science Foundation (IIS 0455970) and a contract from the Library of Congress as part of the National Digital Information Preservation Program.

8. REFERENCES

[1] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2009). *Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation.* http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf

[2] Capra, R., Lee, C., Marchionini, G., Russell, T., Shah, C., & Stutzman, F. (2008). Selection and Context Scoping for Digital Video Collections: An Investigation of YouTube and Blogs. *Proceedings of ACM/IEEE JCDL 2008* (Pittsburgh, PA, June 16-20, 2008), 211-220.

[3] Clemens, R., Capra, R., Lee, C., and Sheble, L. (2009). Contextual Information from Blogs in Video Digital Curation. Proceedings of Society of American Archivists 2008 Research Forum.

[4] Conway, P. (2000). Overview: Rational for digitization and preservation. In *Handbook for digital projects: A management tool for preservation and access.* Northeast Document Conservation Center, Andover, MA. http://www.nedcc.org/digital/dman.pdf.

[5] Lavoie, B. & Dempsey, L. (2004). Thirteen ways of looking at...digital preservation. *D-Lib Magazine,* 10(7/8). http://www.dlib.org/dlib/july04/lavoie/07lavoie.html.

[6] Lee, C. 2007. "Taking Context Seriously: A Framework for Contextual Information in Digital Collections." UNC SILS TR-2007-04.

[7] Maslov, A., Mikeal, A., & Leggett, J. (2009). Cooperation or Control? Web 2.0 and the Digital Library. *Journal of Digital Information,* 10(1), https://journals.tdl.org/jodi/issue/view/65.

[8] Najork, M. Wiener, J. (2001). Breadth-First Crawling Yields High-Quality Pages. In: Proceedings of the 10th International Conference on the World Wide Web (Hong Kong, May 01 - 05, 2001). WWW '01, 114-118. ACM Press, New York, NY.

[9] Pant, G., & Srinivasan, P. (2005). Learning to Crawl: Comparing Classification Schemes. *ACM Trans. Inf. Syst*. 23, 430-462.

[10] Rosset, S., Neumann, E., Eick, U., Vatnik, N., and Idan, Y. 2002. Customer lifetime value modeling and its use for customer retention planning. In Proceedings of the Eighth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Edmonton, Alberta, Canada, July 23 - 26, 2002). KDD '02. ACM, New York, NY, 332-340. DOI= http://doi.acm.org/10.1145/775047.775097

[11] Shah, C., and Marchionini, G. 2007. Preserving 2008 US Presidential Election Videos. In the *Proceedings of International Web Archiving Workshop (IWAW) 2007.*

[9] http://www.digitalpreservation.gov/