The International Journal of Digital Curation

Issue 1, Volume 4 | 2009

Mining Contextual Information for Ephemeral Digital Video Preservation

Chirag Shah

School of Information & Library Science (SILS)

University of North Carolina at Chapel Hill, USA

June 2009

Summary

For centuries the archival community has understood and practiced the art of adding contextual information while preserving an artifact. The question now is how these practices can be transferred to the digital domain. With the growing expansion of production and consumption of digital objects (documents, audio, video, etc.) it has become essential to identify and study issues related to their representation. A curator in the digital realm may be said to have the same responsibilities as one in a traditional archival domain. However, with the mass production and spread of digital objects, it may be difficult to do all the work manually. In the present article this problem is considered in the area of digital video preservation. We show how this problem can be formulated and propose a framework for capturing contextual information for ephemeral digital video preservation. This proposal is realized in a system called ContextMiner, which allows us to cater to a digital curator's needs with its four components: digital video curation, collection visualization, browsing interfaces, and video harvesting and monitoring. While the issues and systems described here are geared toward digital videos, they can easily be applied to other kinds of digital objects.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

"The impulses to record and to keep are a part of human nature; truth is embedded in the symbols and artifacts that we create and then keep by choice or by accident." (Conway, 1996)

For thousands of years, mankind recorded and preserved its social and cultural information and experiences in various forms. These records help us understand societies and cultures of different times and locations. Now that we live in the so-called "Information Age", it is natural to ask, of all the information that we encounter today, what is worth recording and preserving for future generations. More importantly, how can we capture and attach our social and cultural contexts to these information objects so that one can make sense out of them in a very different temporal and social context?

Our interest lies in studying various issues related to ephemeral digital video preservation. Digitized text has been around for a long time; digital images became common a decade back; and now digital videos are becoming a norm. Moreover, a large number of analog videos are being converted to digital format. Services such as YouTube[1] have made it very convenient and easy for almost everyone to host and share digital videos. As of August 2006 (Gomes, 2006), YouTube was hosting more than 6 million videos; the total time people spent watching these videos since its inception in February 2005 has been 9,305 years. There are many other venues where videos are kept with more moderation and control. Many such videos are either massively popular content that mainstream media organizations will preserve or which will be copied by very many people (the LOCKSS hypothesis (Reich & Rosenthal, 2001)). Many other videos will appear and disappear and receive scant attention. Our aim is to identify and preserve videos that include those outside the mainstream. While these videos do not represent the entire population, they do give a good idea of certain aspects of our culture, fashion, thoughts, and issues of the day.

However, future generations may not be able to make much sense of these videos merely by watching them. Collections such as Prelinger[2] and OpenVideo[3] contain some videos that are nearly 100 years old. Most of them are manually annotated so that one can understand them in the present context. With the rapid growth of digital videos, it is almost impossible to provide such a description of every video that we want to preserve. There is a need for identifying the factors that constitute the contextual information about the videos and capture it as automatically as possible. More practically, it would be useful to have a system that captures some contextual information for the digital video being preserved and presents it to the curator, who can then make an informed decision.

In this article we present some preliminary work on various issues of digital curation and, in particular, digital video preservation and building a system to implement it.

[1] YouTube - Broadcast Yourself http://www.youtube.com

[2] Internet Archive: Free Downloads: Prelinger Archives http://www.archive.org/details/prelinger

[3] The Open Video Project http://www.open-video.org

The Research Problem and Questions

A curator preserving a collection of videos for future use has to make various decisions about the content that is to be stored in the collection. These decisions are not easy, as the curator has to assure that the content is not only accessible, but that it should also make sense in the future. Therefore, there is a considerable need for capturing and storing contextual information along with the digital object being preserved. In the case of digital videos of an ephemeral nature, it is not always clear what may constitute the context, and how to capture it. If we watch these videos now, they may seem very understandable to us, but that is mostly due to the fact that we are aware of their spatio-temporal and other contexts. However, just as we may have difficulty comprehending an ephemeral video from the early 20th century without proper annotations, future generations may find it difficult to make sense of current videos. Metadata related to the video can help in describing it and assure retrieval of the video. However, being able to make sense of the video could require much richer and more highly contextualized information.

One can identify several important issues relating to this problem. These issues involve collection selection, metadata generation, storage, maintenance, and presentation. I shall limit our study focusing on the following research questions:

1. What is the difference between metadata and context for digital videos?
2. What constitutes the contextual information for ephemeral digital videos?
3. How should we capture this contextual information? Where should we look for it?
4. How should we incorporate the curator's knowledge about different sources and collections into the system?
5. How can we detect trends and patterns in a set of videos using the contextual information?
6. How should we present preserved information to users so they can make the most sense using the context?

Background

Preservation of digital objects falls under a broad topic of digital curation. In order to understand various aspects of digital curation and related works reported in the literature, we present an outline of digital curation in Figure 1. The review of literature is presented here keeping this outline in mind. As the figure shows, digital curation primarily involves selecting, preserving, and insuring access to a repository of digital information. The UK Digital Curation Centre (DCC) envisions digital curation as "communication across time" (Rusbridge et al., 2005) and recognizes the equivalence of preservation as "interoperability with the future". Buneman (2004) identified two cultures of digital curation:

- Preservers: that is, librarians, archivists and scientists
- Producers: that is, publishers of reference data and scientists with *complex* data.

He then argues that, even though both of these cultures have their differences, they have much in common. The producers should be worrying about preserving all their hard work and the preservers should be concerned about organizing, linking, and annotating their digital objects.

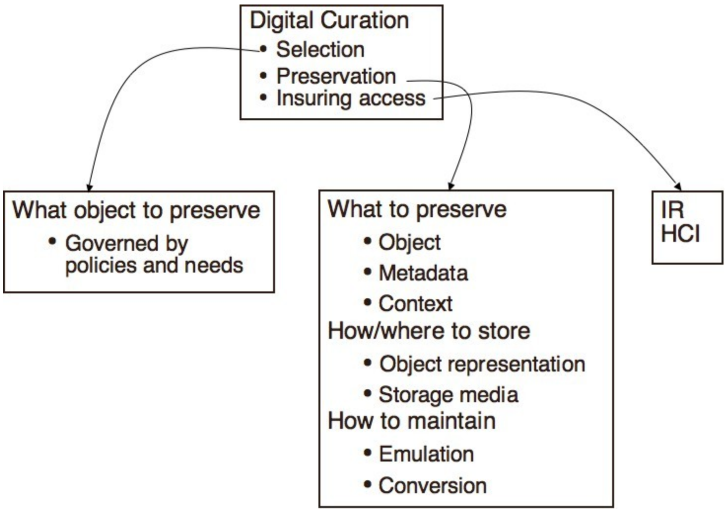


Figure 1. Various aspects of digital curation.

One of the first significant issues these communities encounter is understanding and deciding what objects to preserve. Often this is addressed using policies and consideration of needs. For instance, Rauch, Pavuza, Strodl, and Rauber (2005) came up with nearly 350 criteria and policies for preserving audio and video records. Such criteria and policies are often influenced by the current infrastructure and assumptions about future use. Hedstrom (1998) recognized this issue and argued that digital preservation was primarily motivated by the needs and constraints of the repository, with little consideration for its usage as the scholarly source for the present and the future. Her work focused on storage media, migration, conversion, and overall management strategies for digital preservation. When it comes to making such decisions as to what to preserve and why, it becomes important to understand the role of preservation. Preservation is more than merely storage. As Conway (1990) clearly states, the purpose of preservation is to protect information of enduring value for access by present and future generations. As the pioneer of preservation microfilming, Binkley (1939) also said, “The objective of archival policy in a democratic country cannot be the mere saving of paper. It must be nothing less than the enriching of the complete historic consciousness of the people as a whole.”

In order to add such value to the preserved objects, collecting and attaching metadata is a common practice. Mao, Kim and Thoma (2004) at the National Library of Medicine (NLM) report a system that automatically generates descriptive metadata, which includes title, author, affiliation, and abstract from scanned medical journals. Kim and Ross (2006) at the Digital Curation Centre (DCC) show how to classify genres in order as a first step toward automating metadata extraction from documents. Metadata may be useful for organization and retrieval of the preserved data, but in order to make sense out of them, especially in a different time-frame, we need to add even richer information. In particular, we need to capture contextual information about the digital object being preserved Tibbo, Lee, Marchionini and Howard (2006) presented a framework for capturing such contextual information for digital videos that included elements such as related actors, events, objects, places and times. They also sought to identify those elements that are best documented today and secure contextualizing material for them while noting other elements that would be more deeply supported by material available in the future. Lee (2007) proposed an information model with eight classes of entities to capture contextual information: object, agent, event, function/purpose, time, place, form of expression, and concept/abstraction.

The next big issue that follows is where and how to store all the collected information (data, metadata, and context). This decision could be based on specific application or collection. For instance, Moore, Baru, Rajasekar, Ludaescher, Marciano, Wan et al. (2000b) talked about creation of a one-million message persistent E-mail collection concentrating on technologies such as XML, XSLT, and DTD. Gupta (2001) focused on representation for a Presidential Library. He separated the problems of creating an infrastructure-independent representation for the data objects in a Presidential Library and creating the same for the website that serves as the "annotation superstructure" of the data objects.

Since we are dealing with information in digital form, at the very basic level it is represented in bits. Then the question is how to add *meaning* to these data-bits. Moore (2001) demonstrated how to incorporate information and knowledge along with preservation data. He considers information as semantic tags that provide the meaning of the bits, and knowledge as structural relationships defined by a data model. Ludascher, Marciano and Moore (2001b) extended this idea to include a higher-level knowledge representation along with preservation data.

Since one of the primary goals of preservation is to retain the information for a long time, merely ensuring its storage is not enough; the preserved information needs to be constantly maintained. One of the biggest challenges in maintaining preservation information for a long period of time is the continuous change in storage and computing technologies. Dealing with this issue requires integration of expertise from different fields. Moore et al. (2000a) showed how the preserved information is maintained for a long time through the integration of archival storage technology from supercomputer centers, data grid technology from the computer science community, information models from the digital library community, and preservation models from the archive community. They present an approach for maintaining digital data for hundreds of years through the development of an environment that supports migration of collections onto new software systems. Ludascher et al. (2001a) proposed a migration approach for persistent archives based on XML infrastructure to prevent technological obsolescence over time and across platforms. Hunter and Choudhury (2004) presented a semi-automated digital preservation system based on semantic Web services. This system enables organizations to semiautomatically preserve their digital collections by dynamically discovering and invoking the most appropriate preservation service, as it is required.

If all the issues at different stages of preservation are addressed adequately, it typically ensures access to the preserved information. However, additional steps may need to be taken to present that information to the end-user in a meaningful way. In the archival domain this is typically done by creating a finding aid for the collection. Similar methods can be used for digital archives. However, the abundance, rapid accumulation, and dynamic nature of this information may make it almost impossible to manually create finding aids for them. Besides, we could use a host of tools to extract useful features from digital collections automatically. For instance, while preserving digital video, we could also extract and store some surrogates such as a storyboard, a picture frame, or fast-forwards. Presenting such surrogates along with the actual object can help the user understand that object better. Christel, Smith, Taylor and Winkler (1998) described the effects of different video skim techniques on comprehension, navigation, and user satisfaction. They found significant benefits for skims built from audio sequences meeting certain criteria. In the same spirit, Wildemuth et al. (2002) studied alternative surrogates for video objects in a digital library. They found that fast-forward surrogates attracted the most support from the users.

Based on the background studies, it became clear to us that:

- for effective preservation, storing contextual information along with the object is essential,
- the definition of this context and how much is enough vary with time and application,
- providing curators with the tools that help them make decisions about filtering the information, obtaining the context, and creating preservation policies

is the correct approach to dealing with digital preservation. In the next section we shall further explore this by proposing a way to mine contextual information, keeping the curator at the center.

Proposal for Mining Contextual Information

"A point of sharp contrast between the archaeologist and the looter is that the latter does not bother to record contextual information before removing an object." (Sharer & Ashmore, 2002)

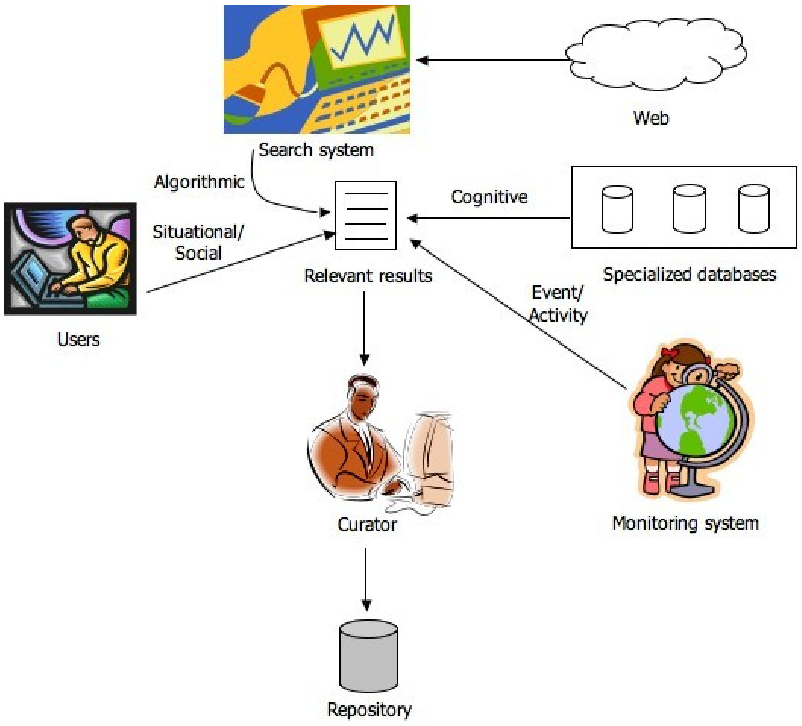


Figure 2. Digital Curation Architecture: Capturing context using various forms of relevances.

Identifying the difference between metadata and context is tricky. Anything that is called context could be declared as metadata and vice versa. For the purpose of this discussion, we would consider any static information about the object as its metadata and any (potentially) dynamic information about that object as its context. For instance, recording date, genre, and duration of a video are static and would be considered as metadata. Ratings and comments by the users watching this video could keep changing with the time, and therefore, they would be considered as the context.

To understand how different factors contribute to the context of a digital video, let us analyze a typical digital curation architecture. Figure 2 presents a schema where a curator of a digital library or archive is gathering contextual information about a digital video from different sources based on relevance. Depending on the way this relevance is defined or perceived, we could identify different kinds of contexts. This framework, based on relevance, is inspired by the works of Saracevic (1996, 1999), Borlund (2003) and Sabre (2004). There follows four different kinds of relevances that can be captured in the given architecture.

1. *Algorithmic relevance*: computer programs based on certain typical IR algorithms search the Web for contextualizing documents, news, images, or videos.
2. *Cognitive relevance*: curators use their background and knowledge to identify specialized databases and/or websites and the system searches these sources and returns candidate items.
3. *Situational/social relevance*: users annotate information based on their own contexts. Not everything annotated is reliable, but the annotations give the curators a good idea of how people relate information of various kinds.
4. *Event/activity relevance*: These explicit user tags can be augmented by a fourth kind of relevance that combines machine processes and user activity. A monitoring component is embedded in the system that can monitor certain sources for events and news and user community activities (e.g., search terms, click streams, links) and alert curators based on the preferences given.

In order to implement capturing the contexts defined above while following Lee's eight-entity model (Lee, 2007), I propose the following five kinds of contexts to be captured for digital video preservation. These contexts were originally presented by Shah and Marchionini (2007a).

1. *Spatial*. This includes location information. It could be where the video appeared, where it was shot, or what places it depicts.
2. *Temporal*. This is information about the time when the video was shot. It is important to note the difference between this time and the time period depicted in the video.

   The time in metadata form is about the time when the video was created. The time in the temporal context could be the time that the video is about. For instance, a video on the American Civil War (1861-1865) could be shot in the year 2007 and hence, year 2007 becomes time in metadata, whereas the temporal context of the video is the 19th century.
3. *Situational*. This context describes the situation in which the video is shot and the events being described in it. This will include background information and the history of the information in the video.

4. *Social*. This context reports people's comments, descriptions, tags, and ratings on the given video. For instance, in the case of YouTube, this could be the comments that people posted about a video.
5. *Cognitive.* Unlike the social context, the cognitive context is not left to the whole community; rather, it is filled in by the curator who can gather some knowledge about the video being preserved. For instance, he or she can list the related items based on personal knowledge or findings. He or she could also record other aspects of the video such as the novelty of the information, and opinions and sentiments expressed.

ContextMiner

To implement the concepts described above for capturing contextual information, we have devised a system called ContextMiner, a tool for collecting, maintaining, and providing preservation data, metadata, and context. Currently at its Alpha stage, ContextMiner provides a proof of concept and a promising direction for further development of the ideas presented here. In the current version, ContextMiner encapsulates the following four components. Their details are given in the sections that follow:

1. Digital video curation
2. Collection visualization
3. Browsing interfaces
4. Video harvesting and monitoring

Digital Video Curation

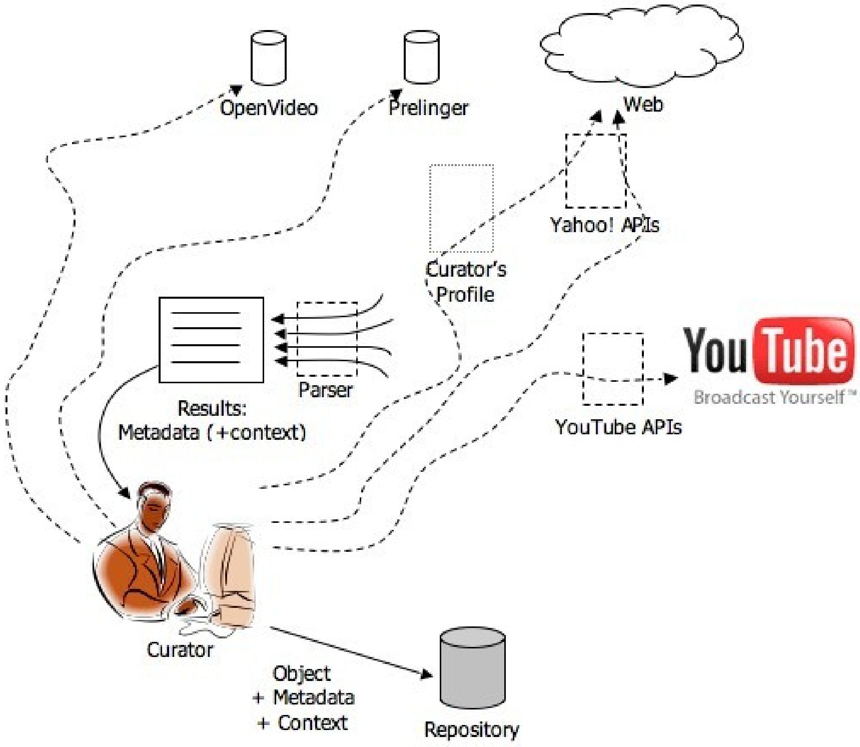


Figure 3. ContextMiner: digital video curation component.

This is the main component of ContextMiner. It helps the curator to find video information from various sources, collect metadata and contexts, edit and compile them, and finally store them in the repository. A typical flow of this process is illustrated in Figure 3, which was originally demonstrated by Shah and Marchionini, (2007b). However, since then, the interface has changed significantly based on the feedback we received and our own experience of working with this system. A brief description of the present interface is given below.

The interface in its initial condition is presented in Figure 4. The screen is divided into two panels. The panel on the right consists of a form with a number of fields that represent certain metadata or contextual information about the object being collected. In a full system, these fields would be mapped to the local content management system.

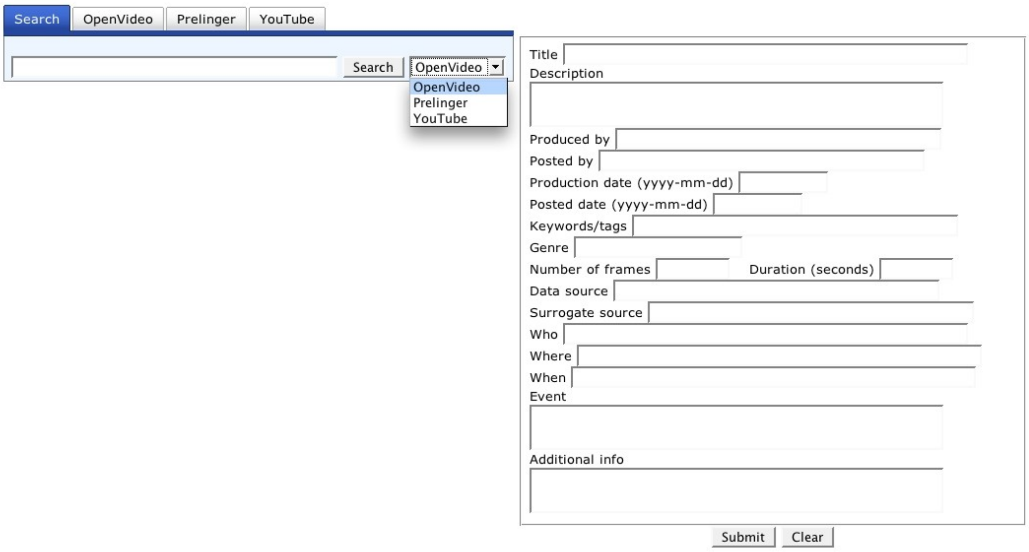
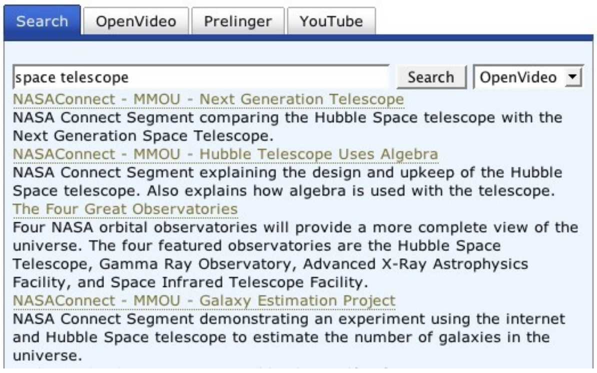


Figure 4. The curator interface.

The panel on the left has four tabs: Search, OpenVideo, Prelinger, and YouTube. One can toggle between these tabs without reloading the whole page. All the tabs, other than the "Search" tab contain a form similar to the one displayed on the right panel. The "Search" tabs include a search box where the curator can enter a query. Along with this query, the curator can specify in which source to execute this query. This can be indicated by selecting the source from the dropdown box with three options: OpenVideo, Prelinger, and YouTube. A sample result list for query "space telescope" in OpenVideo is shown in Figure 5 (note that only the result panel is shown to give more readable detail). When the curator clicks on a result, the system fetches as much information about that object as possible from the given source. At this point the lef-thandside panel automatically switches to the corresponding source tab and fills in the fetched information in various fields on that tab. Figure 6 shows the result of clicking on a search result in the "Search" tab that came from YouTube. As we can see, the left-hand side panel has most of the fields from the right-hand side. Each field on the left is followed by an arrow button. Clicking on this button transfers the content of that field to the corresponding field on the right. Thus, if the curator is satisfied with the automatically extracted information, he or she can collect them in just a few clicks.

In summary, the interface on one side provides curators a workspace, where they can search on and extract data from different sources, compare and compile them, and transfer them to the form on the other side to store in the collection.



space telescope

NASAConnect - MMOU - Next Generation Telescope
NASA Connect Segment comparing the Hubble Space telescope with the Next Generation Space Telescope.

NASAConnect - MMOU - Hubble Telescope Uses Algebra
NASA Connect Segment explaining the design and upkeep of the Hubble Space telescope. Also explains how algebra is used with the telescope.

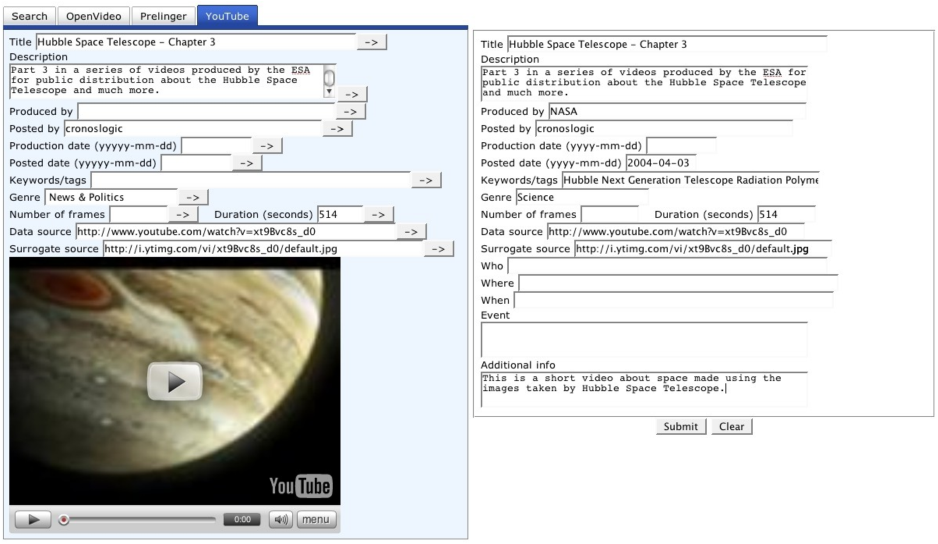
The Four Great Observatories
Four NASA orbital observatories will provide a more complete view of the universe. The four featured observatories are the Hubble Space Telescope, Gamma Ray Observatory, Advanced X-Ray Astrophysics Facility, and Space Infrared Telescope Facility.

NASAConnect - MMOU - Galaxy Estimation Project
NASA Connect Segment demonstrating an experiment using the internet and Hubble Space telescope to estimate the number of galaxies in the universe.

Figure 5. Searching in OpenVideo.

Collection Visualization

This component of the ContextMiner is developed to help the curator visualize a collection in different ways. A user of an IR system is often left in the dark with a query box. It may be useful in many cases to present some kind of overview representation of the underlying collection to the user. Obviously, constructing and providing a representation of a massive collection such as the Web is very difficult and probably not very useful. However, such functionality could be feasible and useful for a small or specialized collection. The other issue that keeps the user in the dark is the presentation of the retrieved information. A typical search engine presents the retrieved results ordered by their relevance. While working with specialized collections in which most of the documents are about a specific domain or topic, relevance is not enough to order the retrieved set. A user may also want to know how much the retrieved documents differ from each other. In other words, in addition to relevance, it becomes useful to display any novelty among those documents.



Title Hubble Space Telescope – Chapter 3
Description
Part 3 in a series of videos produced by the ESA for public distribution about the Hubble Space Telescope and much more.
Produced by
Posted by cronoslogic
Production date (yyyyy-mm-dd)
Posted date (yyyyy-mm-dd)
Keywords/tags
Genre News & Politics
Number of frames Duration (seconds) 514
Data source http://www.youtube.com/watch?v=xt9Bvc8s_d0
Surrogate source http://i.ytimg.com/vi/xt9Bvc8s_d0/default.jpg

Title Hubble Space Telescope – Chapter 3
Description
Part 3 in a series of videos produced by the ESA for public distribution about the Hubble Space Telescope and much more.
Produced by NASA
Posted by cronoslogic
Production date (yyyy-mm-dd)
Posted date (yyyy-mm-dd) 2004-04-03
Keywords/tags Hubble Next Generation Telescope Radiation Polyme
Genre Science
Number of frames Duration (seconds) 514
Data source http://www.youtube.com/watch?v=xt9Bvc8s_d0
Surrogate source http://i.ytimg.com/vi/xt9Bvc8s_d0/default.jpg
Who
Where
When
Event
Additional info
This is a short video about space made using the images taken by Hubble Space Telescope.

Figure 6. Result of clicking on a YouTube search result. The form on the right shows some of the fields filled in using the metadata automatically extracted on the left.

ContextMiner's collection visualization component facilitates this using three sub-components, which are illustrated in Figure 7 and described below:

1. *Search*: The system provides a simple searching interface for doing full-text search and retrieval. Using Indri[4] , ContextMiner indexes text, HTML, XML, and PDF documents available in the collection.
2. *Browse*: ContextMiner's visualization system prepares two kinds of hyperlinked term clouds:
   (a) Term-collection cloud: based on the occurrences of terms in the entire collection.
   (b) Term-document cloud: based on the number of documents in which a term occurs.
   The user can browse through the clickable term clouds and find associated documents.
3. *Discover*: This system not only retrieves relevant information from the indexed collection, but can also evaluate novelty across documents. This can help the user to discover not only the relevant, but also the novel information. At present novelty for document d_j with respect to d_i is implemented as

$$Novelty(d_i,d_j)=1 - \text{Fraction of words overlap between } d_i \text{ and } d_j \quad (1)$$

The system evaluates novelty among the top 10 retrieved documents and presents these pairwise relationships in a matrix format. In addition to the numbers, the system also uses color-coding to indicate the degree of novelty.

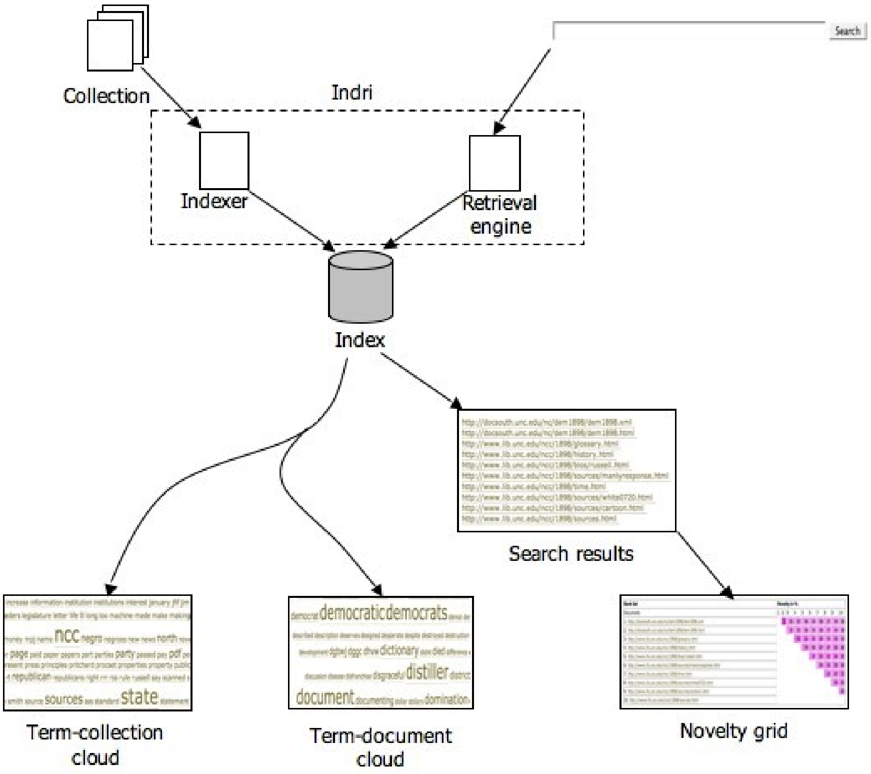


Figure 7. ContextMiner: collection visualization component.

[4] INDRI - Language modeling meets inference networks http://www.lemurproject.org/indri/

Currently this component is working on The North Carolina Election of 1898 Collection[5] available from UNC Chapel Hill Library. The collection has nearly 500 historical documents containing about 8 million terms. We have also developed this component as a stand-alone tool called DiscoverInfo (Shah & Marchionini, 2007c) that allows one to point to any URL on the Web for the purposes of crawling that website, indexing it, and finally preparing the interfaces described above for browsing.[6]

Browsing Interfaces

Curation is not only about preserving the information or objects, but also making sure that they are retrievable and comprehensible. Therefore, ContextMiner also includes tools to present the preserved information from the repository. At present, there are three ways of presenting this information.

Typical Digital Library Table.

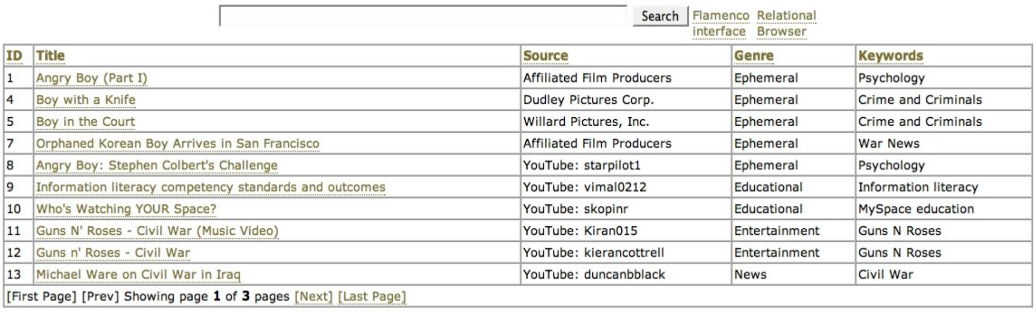
By default, this interface is given for browsing. It presents a few fields such as title, source, genre, and keywords in a table form. The user can search in the entire repository and retrieve the results in table form (Figure 8). The results are presented with paging, displaying 10 results on each page. The user can also sort the results by any of the fields. Clicking on any of the items brings up the full information about that item.

Faceted Search Interface with Flamenco.

We have also implemented the browsing interface using Flamenco[7], which provides a nice way to create faceted search interfaces. We used source, keywords, and genre as the facets for generating this interface. Figure 9 shows this interface. The details of this implementation is beyond the scope of this article and the reader is referred to the Flamenco website[7] for further information.

Faceted Search Interface with Relational Browser.

Relational Browser (RB)[8] is another tool for creating faceted search interfaces. Once again, we used the same three facets and implemented the RB interface with them over our repository. A screenshot of this interface is given in Figure 10.



ID	Title	Source	Genre	Keywords
1	Angry Boy (Part I)	Affiliated Film Producers	Ephemeral	Psychology
4	Boy with a Knife	Dudley Pictures Corp.	Ephemeral	Crime and Criminals
5	Boy in the Court	Willard Pictures, Inc.	Ephemeral	Crime and Criminals
7	Orphaned Korean Boy Arrives in San Francisco	Affiliated Film Producers	Ephemeral	War News
8	Angry Boy: Stephen Colbert's Challenge	YouTube: starpilot1	Ephemeral	Psychology
9	Information literacy competency standards and outcomes	YouTube: vimal0212	Educational	Information literacy
10	Who's Watching YOUR Space?	YouTube: skopinr	Educational	MySpace education
11	Guns N' Roses - Civil War (Music Video)	YouTube: Kiran015	Entertainment	Guns N Roses
12	Guns n' Roses - Civil War	YouTube: kierancottrell	Entertainment	Guns N Roses
13	Michael Ware on Civil War in Iraq	YouTube: duncanbblack	News	Civil War

[First Page] [Prev] Showing page **1** of **3** pages [Next] [Last Page]

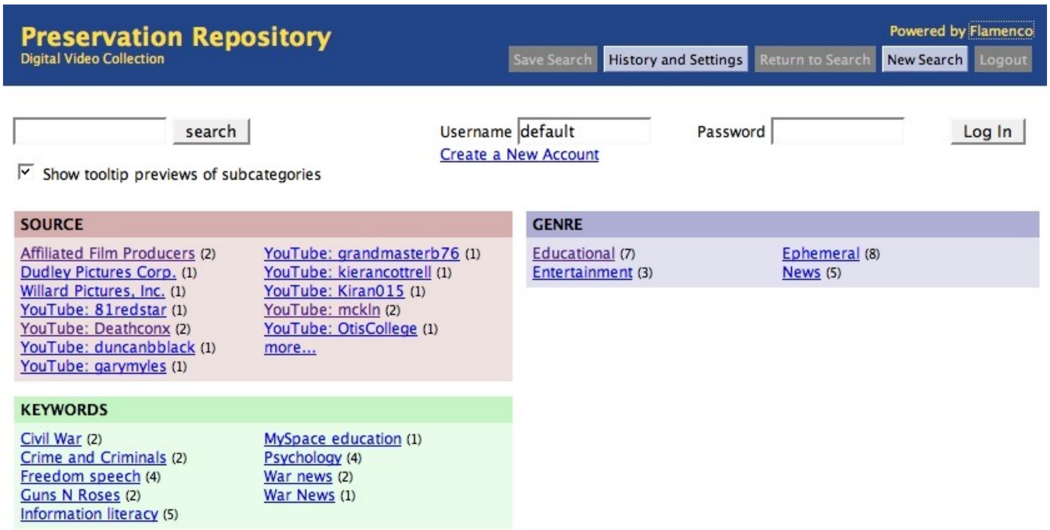
Figure 8. ContextMiner: typical digital library interface to the repository.

[5] The North Carolina Election of 1898 http://www.lib.unc.edu/ncc/1898/

[6] Another spin-off from this is a toolkit called DITookkit, which allows anyone to build such an interface for almost any collection that consists indexable documents. This toolkit is available for free download from http://idl.ils.unc.edu/~chirag/DIToolkit/

[7] The Flamenco Search Interface Project http://flamenco.berkeley.edu/

[8] Interaction Design Laboratory Presents RAVE http://idl.ils.unc.edu/rave/



Username default

Create a New Account

Password

☑ Show tooltip previews of subcategories

SOURCE

Affiliated Film Producers (2)
Dudley Pictures Corp. (1)
Willard Pictures, Inc. (1)
YouTube: 81redstar (1)
YouTube: Deathconx (2)
YouTube: duncanbblack (1)
YouTube: garymyles (1)
YouTube: grandmasterb76 (1)
YouTube: kierancottrell (1)
YouTube: Kiran015 (1)
YouTube: mckln (2)
YouTube: OtisCollege (1)
more...

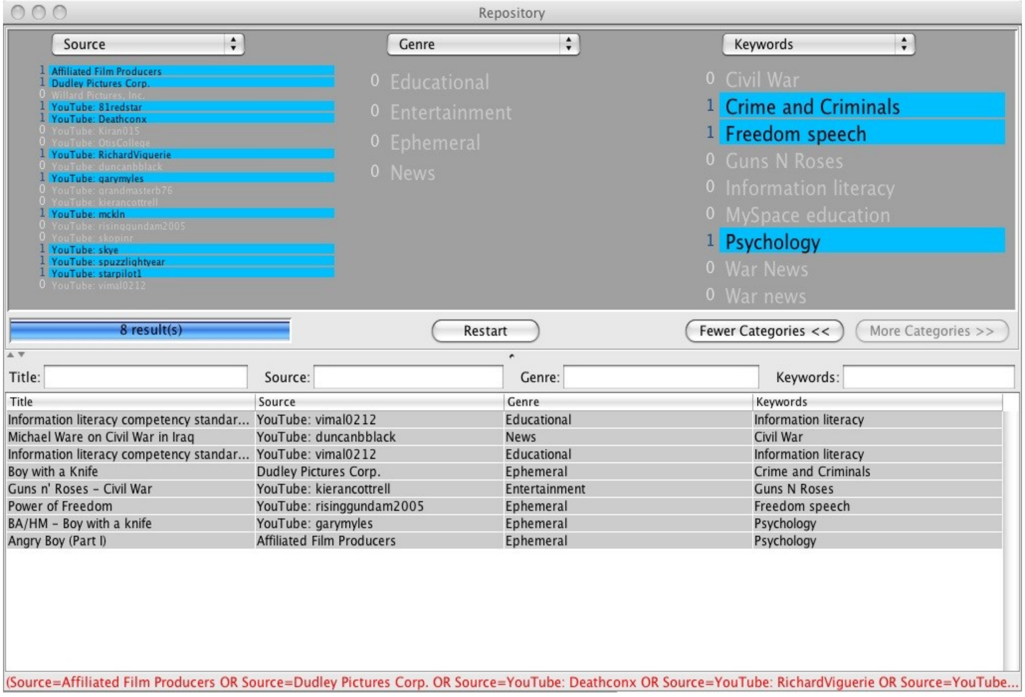
GENRE

Educational (7)
Entertainment (3)
Ephemeral (8)
News (5)

KEYWORDS

Civil War (2)
Crime and Criminals (2)
Freedom speech (4)
Guns N Roses (2)
Information literacy (5)
MySpace education (1)
Psychology (4)
War news (2)
War News (1)

Figure 9. ContextMiner: faceted search interface to the repository with Flamenco.



1 Affiliated Film Producers
1 Dudley Pictures Corp.
0 Willard Pictures, Inc.
1 YouTube: 81redstar
1 YouTube: Deathconx
0 YouTube: Kiran015
0 YouTube: OtisCollege
1 YouTube: RichardViguerie
0 YouTube: duncanbblack
1 YouTube: garymyles
0 YouTube: grandmasterb76
0 YouTube: kierancottrell
1 YouTube: mckln
0 YouTube: risinggundam2005
0 YouTube: skopinr
1 YouTube: skye
1 YouTube: spuzzlightyear
1 YouTube: starpilot1
0 YouTube: vimal0212

0 Educational
0 Entertainment
0 Ephemeral
0 News

0 Civil War
1 Crime and Criminals
1 Freedom speech
0 Guns N Roses
0 Information literacy
0 MySpace education
1 Psychology
0 War News
0 War news

Title: Source: Genre: Keywords:

Title	Source	Genre	Keywords
Information literacy competency standar...	YouTube: vimal0212	Educational	Information literacy
Michael Ware on Civil War in Iraq	YouTube: duncanbblack	News	Civil War
Information literacy competency standar...	YouTube: vimal0212	Educational	Information literacy
Boy with a Knife	Dudley Pictures Corp.	Ephemeral	Crime and Criminals
Guns n' Roses – Civil War	YouTube: kierancottrell	Entertainment	Guns N Roses
Power of Freedom	YouTube: risinggundam2005	Ephemeral	Freedom speech
BA/HM – Boy with a knife	YouTube: garymyles	Ephemeral	Psychology
Angry Boy (Part I)	Affiliated Film Producers	Ephemeral	Psychology

(Source=Affiliated Film Producers OR Source=Dudley Pictures Corp. OR Source=YouTube: Deathconx OR Source=YouTube: RichardViguerie OR Source=YouTube...

Figure 10. ContextMiner: faceted search interface to the repository with Relational Browser (RB).

Video Harvesting and Monitoring

One of our objectives in this project is to understand the trends or patterns in a set of videos. More specifically, we are interested in identifying the usefulness and impact of social context on the significance of digital videos. In order to study this, ContextMiner includes a video-harvesting and monitoring component. The schema for this component is given in Figure 11. There follows a brief description of its workflow.

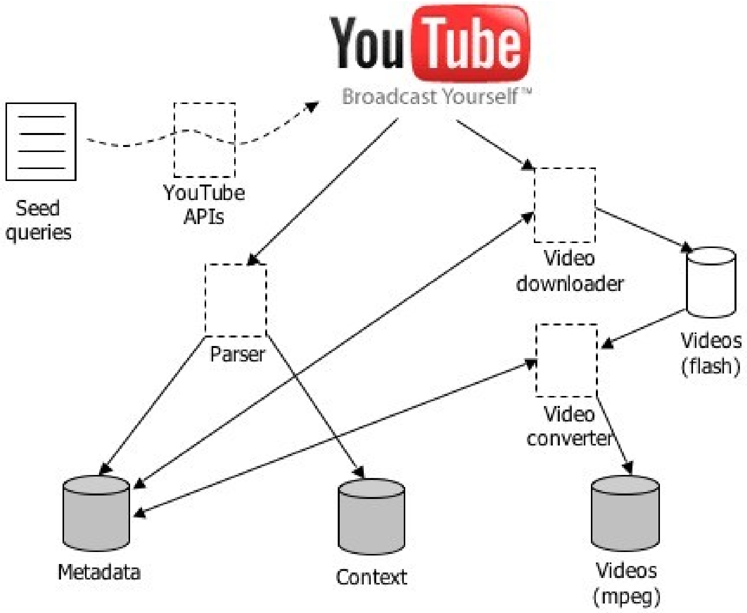


Figure 11. ContextMiner: YouTube video harvesting and monitoring component.

1. A set of seed queries are given by the curator.
2. ContextMiner uses these queries to go out and search on YouTube using YouTube APIs
3. A set of metadata is extracted from the top 100 results returned from YouTube
4. The video downloader component checks the metadata table to see which videos are not downloaded and grabs those videos in flash format.
5. The video converter component checks which videos are downloaded and not converted, and converts them into mpeg format.
6. The context-capturing component goes out to YouTube and captures various kinds of contextual information about the video items for which metadata are already collected. Some of these contextual items include number of views, number of comments (including all the text comments), ratings, number of honors, and number of times favorited. Each time a context is captured, a time-stamp is recorded.

As we can see, there are four major parts to this component:

- metadata collection,
- context collection,
- video downloader, and
- video converter.

Each of these parts can be run independently and they all will check the overlapping functions with other parts to guarantee consistency and integrity of the whole component.

Conclusion

Preserving ephemeral digital videos is not only a task that is intellectually interesting, but also a process that is culturally and socially important. These videos which achieve a certain importance, popularity, or attention albeit for a limited period of time can tell us a lot about the cultural and social values, opinions, sentiments, and community dynamics of that time. However, simply storing the videos is not sufficient. If we want to ensure that the future generations can access and make sense of these videos, we need to add value to them. This goes beyond storing some metadata. In this article I have argued that, in order to make sense out of the videos being preserved, we need to store additional contextual information. This problem of capturing context brings up a host of research questions including defining context, capturing and validating it, and presenting it.

Inspired by Saracevic's types of relevance (Saracevic, 1999) and Lee's eight-entity information model (Lee, 2007), we proposed to capture five different kinds of contexts for a digital video. This proposal is implemented by a system called ContextMiner, which includes componets to search for contextual information, add value to a digital object being preserved, visualize connections among digital objects and their contextual information, and harvest data and attributes, as well as monitor the objects of interest over a period of time.

While discussion here was focused on digital videos, the framework proposed here could be mapped to any kind of digital object. The ContextMiner system presented here serves as a proof of concept for this framework. I hope in the future to present an evaluation of ContextMiner, along with various analyses of data and processes associated with it.

Acknowledgment

The author is indebted to his advisor Gary Marchionini for his constant encouragement and guidance; and Barbara Wildemuth, as well as other members of the VidArch team - Rob Capra, Paul Jones, Sarah Jordan, Cal Lee, Terrell Russell, Laura Sheble, Yaxiao Song, and Helen Tibbo - for their feedback on this work. The work reported here is supported by NSF grant # IIS 0455970.

References

Binkley, R. C. (1939). Strategic objectives in archival policy. *American Archivist*, *2, (July),* pp. 162–168.

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science*, *54*(10), pp. 913–925.

Buneman, P. (2004). The two cultures of digital curation. In *The 16th International Conference on Scientific and Statistical Database Management (SSDBM'04), IEEE Computer Society*, pp. 7.

Christel, M. G., Smith, M. A., Taylor, C. R., & Winkler, D. B. (1998). Evolving video skims into useful multimedia abstractions. In *Proceedings of SIGCHwe Conference on Human Factors in Computing Systems*, pp. 171–178, New York, NY. ACM Press/Addison-Wesley.

Conway, P. (1990). Archival preservation in a nationwide context. *American Archivist*, *53*(2), pp. 204–222.

Conway, P. (1996). *Preservation in the digital world*. Technical Report, Yale University Library.

Gomes, L. (2006, August 30). Will all of us get our 15 minutes on a YouTube video? In *Wall Street Journal*. Retrieved June 25, 2009, from http://online.wsj.com/public/article/SB115689298168048904-5wWyrSwyn6RfVfz9NwLk774VUWc_20070829.html

Gupta, A. (2001). *Preserving presidential library websites*. Technical Report SDSC TR-2001-3, San Diego Supercomputer Center.

Hedstrom, M. (1998). Digital preservation: A time bomb for digital libraries. *Computers and the Humanities*, *31*, pp. 189–202.

Hunter, J., & Choudhury, S. (2004). A semi-automated digital preservation system based on semantic web services. In *Proceedings of ACM IEEE Joint Conference on Digital Libraries*, pp. 269– 278.

Kim, Y., & Ross, S. (2006). Genre classification in automated ingest and appraisal metadata. In *Proceedings of the Tenth European Conference on Research and Advances in Technology for Digital Libraries*, pp. 63–74, Alicante, Spain. Springer-Verlag.

Lee, C. A. (2007). *From simply finding to making sense of digital objects: Toward an information model for contextual information*. Technical Report, SILS, UNC Chapel Hill.

Ludascher, B., Marciano, R., & Moore, R. (2001a). Preservation of digital data with self-validating, self-instantiating knowledge-based archives. In *ACM SIGMOD*, *30*, pp. 54 – 63.

Ludascher, B., Marciano, R., & Moore, R. (2001b). Towards self-validating knowledge-based archives. In *Eleventh International Workshop on Research Issues in Data Engineering on Document Management for Data Intensive Business and Scientific Applications*.

Mao, S., Kim, J. W., & Thoma, G. R. (2004). A dynamic feature generation system for automated metadata extraction in preservation of digital materials. In *Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04)*. IEEE Computer Society.

Moore, R. (2001). *The preservation of data, information, and knowledge*. Technical Report, San Diego Supercomputer Center.

Moore, R., Baru, C., Rajasekar, A., Ludaescher, B., Marciano, R., Wan, M. et al. (2000a, March). Collection-based persistent digital archives - part 1. *D-Lib Magazine, 6*(3). Retrieved June 25, 2009, from http://www.dlib.org/dlib/march00/moore/03moore-pt1.html

Moore, R., Baru, C., Rajasekar, A., Ludaescher, B., Marciano, R., Wan, M. et al. (2000b, April). Collection-based persistent digital archives - part 2. *D-Lib Magazine, 6*(4). Retrieved June 25, 2009, from http://www.dlib.org/dlib/april00/moore/04moore-pt2.html

Rauch, C., Pavuza, F., Strodl, S., & Rauber, A. (2005). Evaluating preservation strategies for audio and video files. In *DELOS Workshop on Digital Repositories: Interoperability and Common Services*, *52*, pp. 172– 180, Heraklion, Greece.

Reich, V., & Rosenthal, D. S. H. (2001, June). LOCKSS: A permanent web publishing and access system. *DLib Magazine*, *7*(6). Retrieved June 25, 2009, from http://www.dlib.org/dlib/june01/reich/06reich.html

Rusbridge, C., Burnhill, P., Ross, S., Buneman, P., Giaretta, D., Lyon, L. et al. (2005). The digital curation centre: a vision for digital curation. In *Local to Global Data Interoperability - Challenges and Technologies*, pp. 31–41.

Sabre, J. M. (2004). *"Relevance" in information retrieval*. Technical Report, Penn State University.

Saracevic, T. (1996). Relevance reconsidered. In *Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2)*, pp. 201–218.

Saracevic, T. (1999). Information science. *Journal of the American Society for Information Science*, *50*(12), pp. 1051–1063.

Shah, C., & Marchionini, G. (2007a). Capturing relevant information for digital curation. In *IEEE ACM Joint Conference on Digital Libraries (JCDL)*, pp. 496.

Shah, C., & Marchionini, G. (2007b). ContextMiner: A tool for digital library curators. In *IEEE ACM Joint Conference on Digital Libraries (JCDL)*, pp. 514.

Shah, C., & Marchionini, G. (2007c). DiscoverInfo: A tool for discovering information with relevance and novelty. In *ACM SIGIR*, pp. 902.

Sharer, R. J., & Ashmore, W. (2002). *Archaeology: Discovering our past*. McGraw-Hill.

Tibbo, H. R., Lee, C. A., Marchionini, G., & Howard, D. (2006). VidArch: Preserving meaning of digital video over time through creating and capture of contextual documentation. *Proceedings of Archiving*, pp. 210-215.

Wildemuth, B. M., Marchionini, G., Wilkens, T., Yang, M., Geisler, G., Fowler, B. et al. (2002). Alternative surrogates for video objects in a digital library: User's perspective on their relative usability. In M. Agosti & C. Thanos (Eds.), *Proceedings of the 6th European Conference on Research and Advances in Technology for Digital Libraries*, pp. 493–507, Berlin: Springer-Verlag.