
Cal Lee
School of Information and Library Science
University of North Carolina, Chapel Hill

Sixth International Conference on Preservation of Digital Objects (iPRES)
October 6, 2009
San Francisco, CA
Remember the theme of this year’s conference?
Moving Into the Mainstream, Enabling Our Digital Future
So what is it that we want to “mainstream”?
Appropriate digital curation norms, practices, policies, systems
These goals can be supported and enacted through repositories that are trustworthy, loyal, helpful, friendly, courteous, kind, obedient, cheerful, thrifty, brave, clean and reverent*

*Also recommended for Boy Scouts
So what do we mean by “repository”? 
A combination of

- services
- resources (required to carry out those services and supported by the services)
- policies that determine how the services should be implemented

that are mixed and matched in various ways to manage and support access to digital collections over time
But online environments for managing digital collections are often not born as all-inclusive trustworthy repositories.
They’re often created to provide relatively focused sets of services (e.g. management and presentation of a specific type of digitized materials; author submission and annotation of pre-print articles; harvesting and dissemination of content from the Web)
For purposes of simplicity, services and resources are often co-located under the control of a single entity.
If an environment proves useful over time, those responsible for managing it begin to confront issues of interoperability, sustainability and scalability.
Move from developing and supporting specialized tools to developing and supporting a long-term digital repository.
A fundamental design question: how to “slide and dice” services, resources and policies: who will have responsibility, where they will reside, and how they will interact
Example of a specialized collecting building environment:

_ContextMiner_
ContextMiner
(http://www.contextminer.org)

• Web-based service for building collections, through execution & management of “campaigns” (i.e. sets of associated queries & parameters to harvest content over time)

• For campaign, user specifies how often to query, number of results to harvest, hosts to query

• Can collect information from various sources: blogs, Flickr, Twitter, YouTube, open Web

• Uses various site-specific APIs to collect data
Create a new campaign (step 1 of 2)
A campaign is like a project. Once you create a new campaign, you will be able to choose what objects you want to collect, from which sources, and how.

Enter a name for your campaign
(e.g. Swine flu, Economic recession)

Create
Create a new campaign (step 2 of 2)
Here you can choose what you want to collect for your campaign Economic recession.
Clicking on a source button will expand or collapse its options. Any query you add through "Common Queries" box will be added to Blogs, Twitter, YouTube, and Flickr.

Don't forget to click on "Start My Campaign" once you are done entering your queries and URLs, and configuring their options. Once your campaign is running, you can come back to this page any time to add more queries/URLs, or change any of the options.
Parameters for a Query within a Campaign

Add the **queries** that you want to monitor. ContextMiner will keep running them periodically on YouTube and collect the top results as per your criteria.

Current queries and their options:

<table>
<thead>
<tr>
<th>#</th>
<th>Query</th>
<th># of results</th>
<th>Criteria</th>
<th>What to collect</th>
<th>Frequency</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>stimulus package</td>
<td>100</td>
<td>Relevance, Published, View count, Rating</td>
<td>Results+Basic attributes, Results+Basic attributes +Periodic attributes</td>
<td>Weekly</td>
<td>Pause Delete</td>
</tr>
</tbody>
</table>
Campaign Management Features

• Changing campaign description, queries and some parameters
• Pausing, resuming, or deleting entire campaign or specific queries
• Adding new queries
• Applying judgments (relevant, non-relevant, neutral) to crawled items
• Deleting items that one doesn’t wish to retain
Three Different Campaigns for a Given User

<table>
<thead>
<tr>
<th>ID</th>
<th>Title</th>
<th>Date Created</th>
<th>Collection</th>
<th>Status</th>
<th>Last Export</th>
<th>Manage</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Elections 2008</td>
<td>2008-06-24</td>
<td>YouTube: 1537, In-links: 337756 Blogs: 38949 Tweets: 31616</td>
<td>Active</td>
<td>N/A</td>
<td>Description</td>
</tr>
<tr>
<td>516</td>
<td>Cancer</td>
<td>2009-06-16</td>
<td>Web: 500 YouTube: 532, In-links: 10757 Blogs: 707 Tweets: 3081 Flickr: 1678</td>
<td>Active</td>
<td>N/A</td>
<td>Campaign Options</td>
</tr>
<tr>
<td>523</td>
<td>Swine flu</td>
<td>2009-06-17</td>
<td>Web: 500 YouTube: 148, In-links: 3936 Blogs: 459 Tweets: 591 Flickr: 188</td>
<td>Active</td>
<td>N/A</td>
<td>Campaign Options</td>
</tr>
</tbody>
</table>
Items from YouTube within a Collecting Campaign

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Title</th>
<th>Query</th>
<th>Category</th>
<th>Duration</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Prostate Cancer Drug Improves Survival</td>
<td>cancer survival</td>
<td>News</td>
<td>4.85 min.</td>
<td>2009-06-16</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SUNDAY NEW YORK TIMES CHEMO BRAIN AND CANCER SURVIVAL</td>
<td>cancer survival</td>
<td>News</td>
<td>8.75 min.</td>
<td>2009-06-16</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Treating Cancer - Dendreon’s Provenge May Improve Survival Rate for Prostate Cancer</td>
<td>cancer survival</td>
<td>News</td>
<td>3.23 min.</td>
<td>2009-06-16</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Tips For Cancer Survival</td>
<td>cancer survival</td>
<td>News</td>
<td>4.13 min.</td>
<td>2009-06-16</td>
</tr>
</tbody>
</table>
Detailed Metadata for a Video from YouTube

My Campaigns -> Cancer -> Prostate Cancer Drug Improves Survival

Prostate Cancer Drug Improves Survival [YouTube Video]

Description: Dendreon said its prostate cancer treatment Provenge prolonged patient survival, providing fuel for a rally in its shares.

Username: CapitalistPig1776

Keywords: prostate, cancer, treatment, dendreon, Provenge, DNDN, FDA

In-links to this item: 22

<table>
<thead>
<tr>
<th>Crawl #</th>
<th>Crawl date</th>
<th>Views</th>
<th>Ratings</th>
<th>Avg Rating</th>
<th>Comments</th>
<th>Favorited</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2009-06-16</td>
<td>2938</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>2009-06-17</td>
<td>2941</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>2009-06-18</td>
<td>2942</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>2009-06-19</td>
<td>2942</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>2009-06-20</td>
<td>2946</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>2009-06-21</td>
<td>2948</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
## Items from Blogs within a Campaign

<table>
<thead>
<tr>
<th>#</th>
<th>Title</th>
<th>Query</th>
<th>Published</th>
<th>Collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cancer Research UK Researchers find clues to improve breast cancer survival</td>
<td>cancer survival</td>
<td>2009-06-08</td>
<td>2009-06-16</td>
</tr>
<tr>
<td>2</td>
<td>Herceptin aids stomach cancer survival study</td>
<td>HEALTH News</td>
<td>cancer survival</td>
<td>2009-06-01</td>
</tr>
<tr>
<td>3</td>
<td>Campaigners blast Scots lung cancer survival rates</td>
<td>Scotland</td>
<td>cancer survival</td>
<td>2009-06-08</td>
</tr>
<tr>
<td>4</td>
<td>Predicting Breast Cancer Survival Rates</td>
<td>Ivanhoe's Medical</td>
<td>cancer survival</td>
<td>2009-06-01</td>
</tr>
</tbody>
</table>
## Items from Flickr within a Campaign

### My Campaigns -> Cancer -> Flickr

**Search**

Showing results of all the queries.

<table>
<thead>
<tr>
<th>#</th>
<th>Picture</th>
<th>Author</th>
<th>Query</th>
<th>Collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="image" alt="Cancer Survival Plaza" /></td>
<td><strong>Henry Lopez</strong>&lt;br&gt;Real name: Henry Lopez,&lt;br&gt;Location: Chicago, United States</td>
<td>cancer survival</td>
<td>2009-06-16</td>
</tr>
<tr>
<td>2</td>
<td><img src="image" alt="IMG_0699" /></td>
<td><strong>Katrina Boemig</strong>&lt;br&gt;Real name: N/A, Location: N/A</td>
<td>cancer survival</td>
<td>2009-06-16</td>
</tr>
<tr>
<td>3</td>
<td><img src="image" alt="Cancer Survival" /></td>
<td><strong>TheEclecticArtisan</strong>&lt;br&gt;Real name: Holly Smith,&lt;br&gt;Location: Atlanta, USA</td>
<td>cancer survival</td>
<td>2009-06-16</td>
</tr>
<tr>
<td>4</td>
<td><img src="image" alt="Cancer survival" /></td>
<td><strong>Lan A</strong>&lt;br&gt;Real name: Lan Amphone,&lt;br&gt;Location: Andover</td>
<td>cancer survival</td>
<td>2009-06-16</td>
</tr>
</tbody>
</table>

**Description**: My husband has just been diagnosed with terminal metastatic pancreatic cancer that has invaded his spine, liver, abdomen, and lymphatic system. The doctors have sent him home to die, but that’s not my truth. I believe, in every cell of my body, that he can be healed. We are throwing every compatible holistic therapy at his cancer, and I believe that we can be triumphant. I made this graphic to remind us of what it will take to survive and overcome this cancer. I’ll be placing them all over the house.

**Tags**: cancer, survivor, terminal, illness, hope, inspiration, inspirational, colorful, Holly Smith, Michael Smith, cancer survivor, overcoming cancer, holistic, naturopathy, cesium, rubidium, laetrile

**Taken**: 2009-04-05, 14:29:35  **Posted**: 2009-04-05, 14:29:35
Uses of *ContextMiner*

- By VidArch project at UNC: 145 queries to YouTube every day -> 98,000+ unique videos (May 2007 - Dec 2008)
- In July 2008, public beta released
  - Nearly 300 users
  - 600+ campaigns
Growth Pains of ContextMiner

• Current implementation (single MySQL database & associated code) not scalable or sustainable

• Challenges and opportunities:
  – Storage
  – Collaboration among users
  – Secure data sharing
  – Passive users
  – Preservation & sustainability
Integrated Rule-Oriented Data System (iRODS)

- Data grid technology supports various forms of “slicing and dicing” of resources, e.g.
  - Storage in multiple locations
  - Many different storage technologies
  - Resolution of resource identifiers across diverse systems
- The “Rule-Oriented” part
  - Policies -> Rules -> Micro-services
  - Apply rules through various software environments (e.g. DSpace, Fedora)
  - Document and prove that particular policies were defined and correctly applied at given point in time
ContextMiner Meets iRODS – Sliding and Dicing Options
Transfer of Data - Considerations

- Carried out only once or periodically, based on trigger events or pre-defined schedules.
- Scope of transfer – videos, static metadata, dynamic metadata.
- Undifferentiated bitstream or broken into discrete data elements within iRODS as AVUs (attribute value units).
- Transfer mechanism – XML export/import or Rule-oriented Database Access (RDA) system.
- iRODS mediation of storage of data and metadata in different places (e.g. the collecting institution, consortial data center, private-sector storage provider).
Transfer for Features and Functions

• iRODS execution of web harvesting after collecting campaign has been initiated

• iRODS implementing user account actions based on customized policies (e.g. disabling crawls after given period of inactivity)
Future Directions & Implications

• Further investigation of alternative arrangements based on: efficiencies of resource use; management of dependencies across entities; business model most appropriate to participating organizations

• Generalizing findings by elaborating reference rule sets for use by organizations undergoing similar transitions in their collecting environments - Distributed Custodial Archival Preservation Environments (DCAPE) project

• Incorporating hooks from user interfaces of repository and collection management environments into iRODS
Examples of Potential iRODS Hooks in ContextMiner Interface

• “replicate my campaign data X times in Y locations”
• “verify the integrity of my campaign data by running a checksum every X days”
• ”notify me through email if my campaigns are about to be disabled”
• “pause my campaign if it grows beyond X bytes”
• “every X hours, harvest the blog pages identified in my campaign using wget and store the videos in the following Y locations”
In short:

Mainstreaming digital curation by pushing elements of sustainability into the places where people already work and live
# Acknowledgements

## Financial Support

- National Science Foundation - VidArch Project (#IIS 0455970 DigArch Program)
- National Historical Publications & Records Commission – DCAPE Project (NAR08-RE-10010-08)

## Research Collaborators

- Chien-Yi Hou
- Richard Marciano
- Chirag Shah