

Mainstreaming Preservation through Slicing & Dicing of Digital Repositories: Investigating Alternative Service & Resource Options for ContextMiner Using Data Grid Technology

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.

Common Queries

Web

Blogs

Twitter

YouTube

Flickr

Optional Campaign Details

Start My Campaign

Parameters for a Query within a Campaign

YouTube

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.

Add

Current queries and their options:

#	Query	# of results	Criteria	What to collect	Frequency	Actions
1	stimulus package	<input style="width: 40px; text-align: center;" type="text" value="100"/> Update	<input checked="" type="radio"/> Relevance <input type="radio"/> Published <input type="radio"/> View count <input type="radio"/> Rating	<input checked="" type="radio"/> Results+Basic attributes <input type="radio"/> Results+Basic attributes +Periodic attributes <input type="checkbox"/> Collect inlinks	<input type="radio"/> Daily <input checked="" type="radio"/> Weekly <input type="radio"/> Monthly	Pause Delete

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


<input type="checkbox"/> Select: <input type="text"/> Action: <input type="text"/> <input type="button" value="Go"/>							
<input type="checkbox"/>	ID	Title	Date Created	Collection	Status	Last Export	Manage
<input type="checkbox"/>	2	Elections 2008	2008-06-24	YouTube: 1537, In-links: 337756 Blogs: 38949 Tweets: 31616	Active	N/A	Description Parameters Queries
<input type="checkbox"/>	516	Cancer	2009-06-16	Web: 500 YouTube: 532, In-links: 10757 Blogs: 707 Tweets: 3081 Flickr: 1678	Active	N/A	Campaign Options
<input type="checkbox"/>	523	Swine flu	2009-06-17	Web: 500 YouTube: 148, In-links: 3936 Blogs: 459 Tweets: 591 Flickr: 188	Active	N/A	Campaign Options

Items from YouTube within a Collecting Campaign

My Campaigns -> Cancer -> YouTube Campaign Options

Showing results of all the queries.

Select: [v] Action: [v] Go [First Page] [Prev] Showing page 1 of 22 pages [Next] [Last Page]

#	Title	Query	Category	Duration	Date
1	<p>Prostate Cancer Drug Improves Survival</p>  <p>Description: Dendreon said its prostate cancer treatment Provenge prolonged patient survival, providing fuel for a rally in its shares. Username: CapitalistPig1776 Category: News Duration: 4.85 min. Keywords: prostate, cancer, treatment, dendreon, Provenge, DNDN, FDA Full Record</p>	cancer survival	News	4.85 min.	2009-06-16
2	SUNDAY NEW YORK TIMES CHEMO BRAIN AND CANCER SURVIVAL	cancer survival	News	8.75 min.	2009-06-16
3	Treating Cancer - Dendreon's Provenge May Improve Survival Rate for Prostate Cancer	cancer survival	News	3.23 min.	2009-06-16
4	Tips For Cancer Survival	cancer survival	News	4.15 min.	2009-06-16

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](#)

Query:
cancer
survival

Category:
News

Duration:
4.85 min.

Crawl #	Crawl date	Views	Ratings	Avg Rating	Comments	Favorited
1	2009-06-16	2938	5	5	4	5
2	2009-06-17	2941	5	5	4	5
3	2009-06-18	2942	5	5	4	5
4	2009-06-19	2942	5	5	4	5
5	2009-06-20	2946	5	5	4	5
6	2009-06-21	2948	5	5	4	5

Items from Blogs within a Campaign

My Campaigns -> Cancer -> Blogs Campaign Options

Showing results of all the queries.

Select: Action: [First Page] [Prev] Showing page **1** of **29** pages [Next] [Last Page]

<input type="checkbox"/>	- +	#	Title	Query	Published	Collected
<input type="checkbox"/>	○ ○ ○	1	Cancer Research UK Researchers find clues to improve breast: Snippet: Checking lymph nodes during surgery and assessing the hormone status of tumours could help improve breast cancer survival in the UK according to research published today in Annals of Oncology Author: unknown, Site: cancerresearchuk Published: 2009-06-08, 00:05:00	cancer survival	2009-06-08	2009-06-16
<input type="checkbox"/>	○ ○ ○	2	Herceptin aids stomach cancer survival study HEALTH News	cancer survival	2009-06-01	2009-06-16
<input type="checkbox"/>	○ ○ ○	3	Campaigners blast Scots lung cancer survival rates Scotland	cancer survival	2009-06-08	2009-06-16
<input type="checkbox"/>	○ ○ ○	4	Predicting Breast Cancer Survival Rates Ivanhoe's Medical	cancer survival	2009-06-01	2009-06-16

Items from Flickr within a Campaign

[My Campaigns](#) -> [Cancer](#) -> [Flickr](#) [Campaign Options](#)

Showing results of all the queries.

Select: Action: [First Page] [Prev] Showing page 1 of 8 pages [Next](#) [Last Page](#)

<input type="checkbox"/>	- +	#	Picture	Author	Query	Collected
<input type="checkbox"/>	<input type="radio"/> <input type="radio"/> <input type="radio"/>	1	Cancer Survival Plaza Description: Tags: chicago, skydeck, willis, tower, millenium, park, chrome, jellybean, chinese, sculpture Taken: 2009-06-15, 19:45:43 Posted: 2009-06-15, 13:03:27	Henry Lopez Real name: Henry Lopez, Location: Chicago, United States	cancer survival	2009-06-16
<input type="checkbox"/>	<input type="radio"/> <input type="radio"/> <input type="radio"/>	2	IMG_0699 Description: 1000 paper birds Tags: paper, birds, cancer, survival, survivor, memorial, katrina Boemig Taken: 2009-04-25, 15:36:21 Posted: 2009-04-26, 13:20:40	katrina boemig Real name: N/A, Location: N/A	cancer survival	2009-06-16
<input type="checkbox"/>	<input type="radio"/> <input type="radio"/> <input type="radio"/>	3	Cancer Survival 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	TheEclecticArtisan Real name: Holly Smith, Location: Atlanta, USA	cancer survival	2009-06-16
<input type="checkbox"/>	<input type="radio"/> <input type="radio"/> <input type="radio"/>	4	Cancer survival Description: My BIL and nephew Love this old cocoa daisy kit Tags: Taken: 2009-03-08, 22:18:37 Posted: 2009-03-08, 23:18:37	Lan A Real name: Lan Amphone, Location: Andover	cancer survival	2009-06-16

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