

Visualizing Digital Collections at Archive-It

Kalpesh Padia

Director: Michele C. Weigle
Committee: Michael L. Nelson
Ravi Mukkamala

Agenda

- + Introduction
- + Motivation
- + Related Work
- + Collection Retrieval and Processing
- + Visualizations
- + Case Studies
- + Future Work
- + Conclusion



INTRODUCTION AND MOTIVATION

Digital Archives

The screenshot displays the Library of Congress Digital Collections website. At the top, there is a navigation bar with the Library of Congress logo, links for 'ASK A LIBRARIAN', 'DIGITAL COLLECTIONS', and 'LIBRARY CATALOGS', a search box, and a 'GO' button. Below this is a breadcrumb trail: 'The Library of Congress > Digital Collections'. The main content area is titled 'Digital Collections & Services' with the subtitle 'Access to print, pictorial and audio-visual collections and other digital services'. A sidebar on the left lists 'DIGITAL COLLECTIONS' and 'More Resources' including 'Bibliographies and Guides', 'Finding Aids', 'Virtual Reference Shelf', 'Ask a Librarian', and 'Library Catalogs'. The main content features a 'Featured Digital Collections & Services' section with four items: 'American History & Culture' (a digital library of historical photos, documents, and videos), 'Historic Newspaper' (enhanced access to historical newspapers), 'International Collections' (materials and bilingual presentations from over 100 libraries), and 'Legislative Information' (THOMAS provides text of bills, Congress and more). Below these are 'Browse By Topic' and 'Browse By Subject' options. The right side of the page shows the 'Yale UNIVERSITY LIBRARY' logo and navigation tabs for 'Research Tools', 'Libraries & Collections', 'About the Library', and 'Library Services'. The 'Digital Collections' section includes a search box, a 'SEARCH' button, and a list of collections: 'Beinecke Rare Book and Manuscript Library Digital Images Online', 'Classics Department at Yale Digital Collection', 'Lewis Walpole Library Digital Collection', 'Manuscripts and Archives Digital Image Database', 'Visual Resources Collection at Yale', 'Yale Daily News Historical Archive', and 'ArtStor'. An 'In Focus' section highlights 'Notre Dame (Paris) : exterior sculpture : gargoyle' from the 'Visual Resources Collection'.

<http://www.loc.gov/index.html>

<http://digitalcollections.library.yale.edu/>

7/20/2012

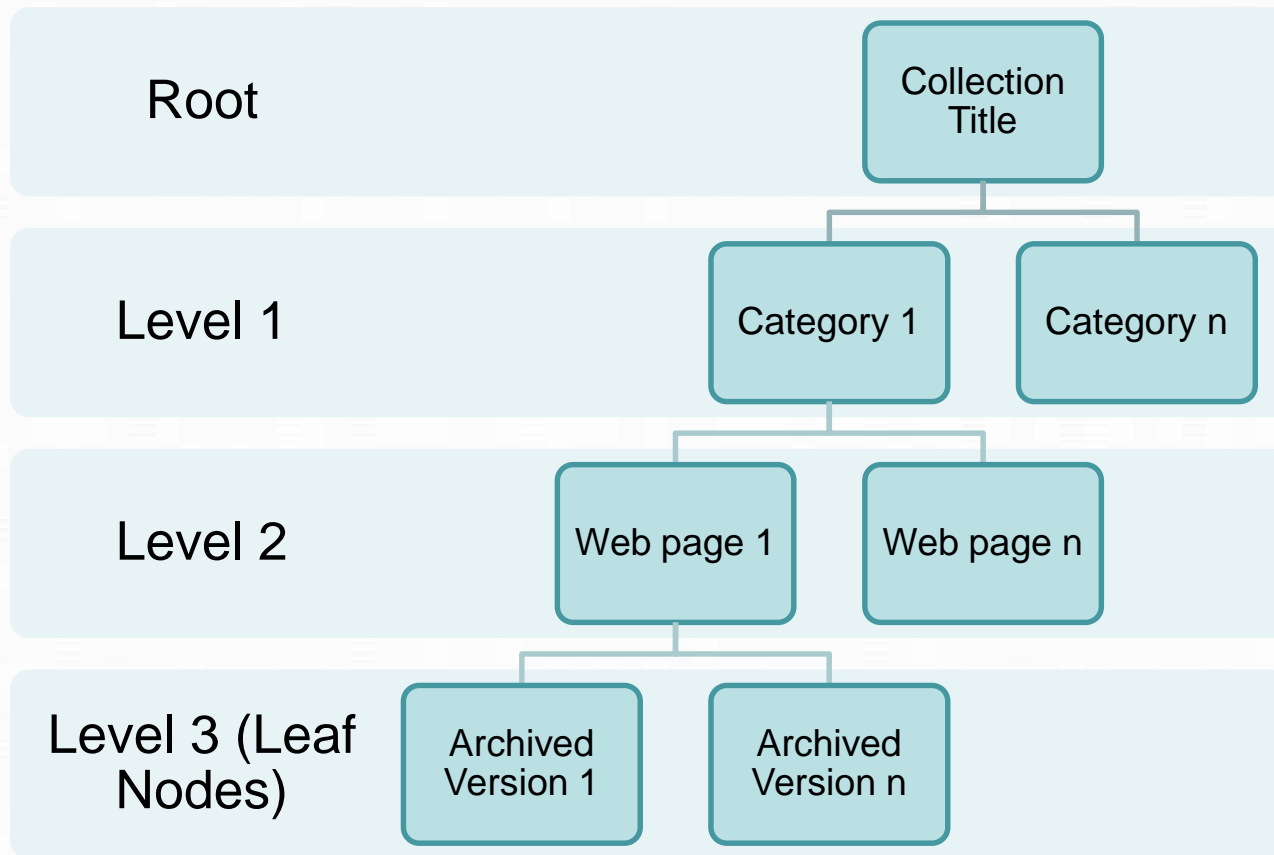
MS Thesis - August 2012

Archive-It

The screenshot shows the Archive-It website interface. At the top right, there are social media icons for Facebook and Twitter, and a 'Login' button. The main navigation bar includes the Archive-It logo, 'HOME', 'EXPLORE', 'LEARN MORE', and 'CONTACT US'. A secondary navigation bar features the text 'A web archiving service to harvest and preserve digital collections a service of the Internet Archive' and a classical building icon. Below this is an orange banner with a welcome message and webinar dates: 'Welcome to Archive-It! Attend a live informational webinar and demo to learn more about the service.' and 'Contact Us to sign up for an upcoming session: Jul 31 2012, 11:30 AM PDT Aug 14 2012, 11:30 AM PDT'. The 'Explore Collections' section has a search box and a 'Show All Collections' link. Three collection cards are displayed: 'Earthquake in Haiti' by Internet Archive Global Events, 'Maryland State Document Collection' by University of Maryland, and 'IT History Society' by IT History Society. Each card includes a representative image and a brief description of the collection's content.

<http://archive-it.org/>

Archive-It Collection Hierarchy



Exploring Archive-It Collections



The screenshot shows the Archive-It website interface. At the top, there is a navigation bar with links for HOME, EXPLORE, LEARN MORE, and CONTACT US. A social media icon for Facebook and a Login button are also present. Below the navigation bar, the breadcrumb trail reads: Explore >> Columbia University Libraries >> Human Rights. The main content area features a green header for the 'Human Rights' collection, collected by Columbia University Libraries. A red box highlights the collection title and the collector information. Below the header, there is a 'Narrow Your Results' section with filters for Group, Subject, and Creator. The 'Subject' filter is expanded, showing options like 'Human rights (396)', 'Human rights advocacy (191)', 'National human rights institutions (72)', 'Civil rights (46)', and 'Democracy (40)'. The search results section shows 'Page 1 of 6 (505 Total Results)' and lists two results. The first result is for the 'Anti Caste Discrimination Alliance ACDA' with a URL of <http://acda.co/>. The second result is for the 'Advocacy Forum--Nepal' with a URL of <http://advocacyforum.org/>.

<http://archive-it.org/collections/1068>

Exploring Archive-It Collections

ARCHIVE-IT HOME EXPLORE LEARN MORE CONTACT US A web archiving service to harvest and preserve digital collections a service of the Internet Archive

Explore >> Columbia University Libraries >> Human Rights

Center for Human Rights Documentation & Research

Human Rights
Collected by: Columbia University Libraries

Description: An initiative of CUL's Center for Human Rights Documentation and Research, the Human Rights Web Archive is an effort to preserve and ensure access to freely available human rights resources created mainly by non-governmental organizations, national human rights institutions, and individuals.
Subject: Society & Culture, Human rights, Non-governmental organizations, Human rights workers, National human rights institutions, Web archives
[More](#)

Narrow Your Results

Group

- Amnesty International sections (46)
- Blogs by individuals (6)
- National human rights institutions (71)
- Non-governmental organizations (396)
- Truth commissions, tribunals, and courts (12)

Subject

- Human rights (396)
- Human rights advocacy (191)
- National human rights institutions (72)
- Civil rights (46)
- Democracy (40)

Creator

- Asociacio'n Paz con Dignidad (3)
- Anti Caste Discrimination Alliance (2)
- Argentina. Defensor del Pueblo de la Nacio'n. (2)
- Asian Human Rights Commission (2)
- Beogradski centar za ljudska prava (2)

Language

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here

Sites Search Page Text

Page 1 of 6 (505 Total Results)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: Anti Caste Discrimination Alliance ACDA
URL: <http://acda.co/>
Description: UK-focused organization based in Derby, working to eliminate caste-based discrimination. In English. New site; see www.acdauk.org.uk/ for older site. Captured 4 times between Oct 13, 2011 and Jun 2, 2012
Subject: Discrimination, Caste, Caste-based discrimination, East Indians
[More](#)

Title: Advocacy Forum--Nepal
URL: <http://advocacyforum.org/>
Description: Nepal-focused organization based in Kathmandu. Includes reports. In English. Captured 14 times between Sep 3, 2010 and Jun 3, 2012
Videos: 4 Videos Captured
Subject: Human rights, Rule of law, Human rights advocacy
[More](#)

<http://archive-it.org/collections/1068>

Exploring Archive-It Collections

The screenshot shows the Archive-It website interface. At the top, there is a navigation bar with links for HOME, EXPLORE, LEARN MORE, and CONTACT US. Below this, a breadcrumb trail reads 'Explore >> Columbia University Libraries >> Human Rights'. The main content area features a green header for the 'Human Rights' collection, collected by Columbia University Libraries, with an archived date of May 2008. A description and subject tags are provided. Below the header, there is a 'Narrow Your Results' section with filters for Group, Subject, and Creator. A search bar is present with the text 'Enter search terms here'. A red box highlights a search result for 'Anti Caste Discrimination Alliance ACDA', showing its title, URL, description, and subject tags. Below it, another result for 'Advocacy Forum--Nepal' is visible.

<http://archive-it.org/collections/1068>

Exploring Archive-It Collections

Center for
Human Rights
Documentation
& Research

Human Rights Web Archive (Columbia University Libraries)



Enter Web Address: All

Searched for <http://acda.co/>

[Look up URL](#) in general Internet Archive web collection

4 Results [RSS](#) [Metadata](#)

[Proxy Mode Help](#)

Drag the cursor around the area you want to capture.

* denotes when page was updated

Found 4 Captures between Oct 13, 2011 - Jun 2, 2012

| | 2011 | | 2012 |
|--------------------------------|---------|-----------------------------|---------|
| | 2 pages | | 2 pages |
| Oct 13, 2011 * | | Mar 2, 2012 | |
| Dec 2, 2011 | | Jun 2, 2012 | |

[Home](#) | [Internet Archive](#)

http://wayback.archive-it.org/1068/*/http://acda.co/

Exploring Archive-It Collections

ARCHIVE-IT HOME EXPLORE LEARN MORE CONTACT US A web archiving service to harvest and preserve digital collections a service of the Internet Archive

Explore >> Columbia University Libraries >> Human Rights

Human Rights

Collected by: Columbia University Libraries

Archived since: May, 2008

Description: An initiative of CUL's Center for Human Rights Documentation and Research, the Human Rights Web Archive is an effort to preserve and ensure access to freely available human rights resources created mainly by non-governmental organizations, national human rights institutions, and individuals.

Subject: Society & Culture, Human rights, Non-governmental organizations, Human rights workers, National human rights institutions, Web archives

More ▾

Narrow Your Results

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here

Sites Search Page Text

Page 1 of 6 (505 Total Results)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: Anti Caste Discrimination Alliance ACDA
URL: <http://acda.co/>
Description: UK-focused organization based in Derby, working to eliminate caste-based discrimination. In English. New site; see www.acdauk.org.uk/ for older site. Captured 4 times between Oct 13, 2011 and Jun 2, 2012
Subject: Discrimination, Caste, Caste-based discrimination, East Indians

More ▾

Title: Advocacy Forum--Nepal
URL: <http://advocacyforum.org/>
Description: Nepal-focused organization based in Kathmandu. Includes reports. In English. Captured 14 times between Sep 3, 2010 and Jun 3, 2012
Videos: 4 Videos Captured
Subject: Human rights, Rule of law, Human rights advocacy

More ▾

Group

- Amnesty International sections (46)
- Blogs by individuals (6)
- National human rights institutions (71)
- Non-governmental organizations (396)
- Truth commissions, tribunals, and courts (12)

Subject

- Human rights (396)
- Human rights advocacy (191)
- National human rights institutions (72)
- Civil rights (46)
- Democracy (40)

More ▾

Creator

- Asociacio'n Paz con Dignidad (3)
- Anti Caste Discrimination Alliance (2)
- Argentina. Defensor del Pueblo de la Nacio'n. (2)
- Asian Human Rights Commission (2)
- Beogradski centar za ljudska prava (2)

More ▾

Language

<http://archive-it.org/collections/1068>

Exploring Archive-It Collections

The screenshot shows the Archive-It website interface. At the top, there is a navigation bar with links for HOME, EXPLORE, LEARN MORE, and CONTACT US. A social media icon for Facebook and a Login button are also present. Below the navigation bar, the breadcrumb trail reads: Explore >> Virginia Tech: Crisis, Tragedy, and Recovery Network >> Pakistan Floods (2011). The main content area features a collection card for 'Pakistan Floods (2011)', which includes a thumbnail image of people in a flooded area, a description of the event, and a 'More >' link. Below the collection card, there is a 'Narrow Your Results' section, which is highlighted with a red box. This section contains a search box and a 'Search' button. Below the search box, there are tabs for 'Sites' and 'Search Page Text'. The search results are displayed as a list of items, each with a title, URL, and capture date. The first result is titled 'Pakistan Flood Relief Fund - CIDA' and has a URL starting with 'http://aodi-cida.ca'. The second result is titled 'Pakistan Flood Relief Fund - CIDA' and has a URL starting with 'http://aodi-cida.ca'. The third result is titled 'Pakistan Flood Relief Fund - CIDA' and has a URL starting with 'http://aodi-cida.ca'. The fourth result is titled 'Pakistan Flood Relief Fund - CIDA' and has a URL starting with 'http://aodi-cida.ca'. The fifth result is titled 'Pakistan Flood Relief Fund - CIDA' and has a URL starting with 'http://aodi-cida.ca'.

<http://archive-it.org/collections/2836>

Drawbacks

- ✚ No visual feedback
- ✚ Discovering individual pages is difficult
- ✚ Optional metadata and categorization
- ✚ Collection structure known only to curator

Contribution

Interactive visualizations

- Treemap
- Time cloud
- Bubble chart
- Image plot
- Wordle
- Timeline

Temporal exploration of collections

Uncover collection structure



RELATED WORK

Microsoft Pivot

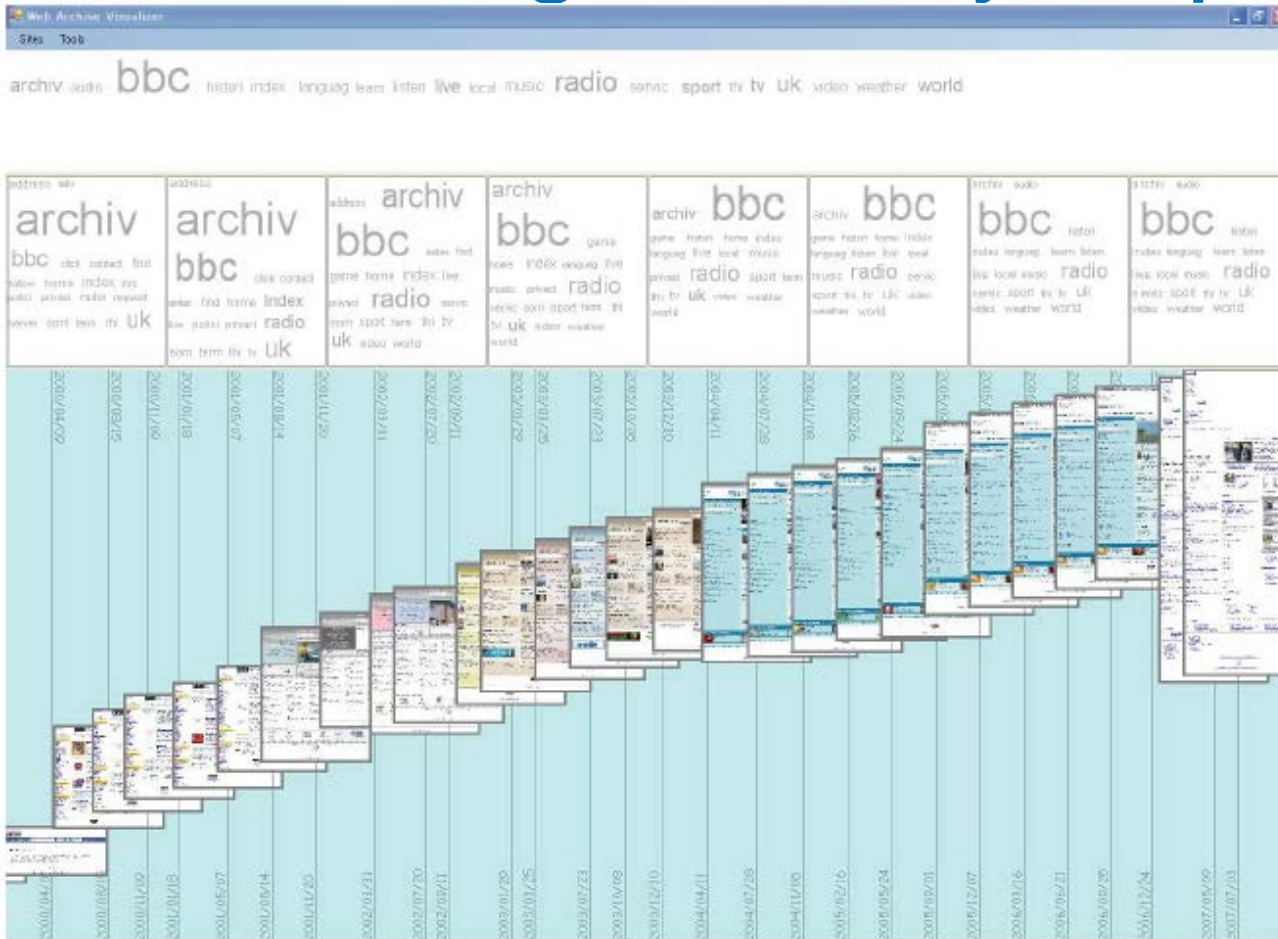
The screenshot displays the Microsoft Pivot viewer interface for Magic: The Gathering cards. The browser address bar shows the URL: <http://content.getpivot.com/Collections/MagicTheGathering/MagicTheGathering.xml>. The page title is "Magic: The Gathering (2007-2009) | Color: Blue".

On the left side, there is a "Filter by Keyword" panel with a "Clear All" button. Under the "Color" section, the "Blue" filter is selected, showing 446 items. Other filters include White (450), Black (446), Red (447), Green (448), Artifact (101), Multicolor (442), and Land (199). Below the color filters, there are sections for "Type", "Subtype", "Rarity", "Total Mana Cost", "Power", "Toughness", "Card Set", "Artist", and "Release Date", each with a dropdown menu.

The main area shows a grid of card sets, sorted by "Card Set". The sets are: Zendikar, Magic 2010, Alara Reborn, Conflux, Shards of Alara, Eventide, Shadowmoor, Morningtide, and Lorwyn. Each set is represented by a vertical column of card thumbnails.

<http://www.microsoft.com/silverlight/pivotviewer/>

Page History Explorer



A. Jatowt, Y. Kawai, and K. Tanaka, "Visualizing Historical Content of Web Pages," in *Proceedings of the 17th international conference on World Wide Web*, 2008.

3D Wall

Provided
by:

BRITISH
LIBRARY

Home
About
Search the archive
Browse the archive
Visualisation
Nominate a site
FAQ's
Technical information
Links to other archives
Archive statistics
UK Web Archive Blog
Contact

3D Wall for Special Collections

Blogs

Scroll through the wall of each Special Collection, click on any image to enlarge, then click on either the  symbol or the description/date to visit the archived website.



The 3D wall displays a grid of website thumbnails. Visible thumbnails include a red page with a bar chart, a page titled 'conservativehome', a page with 'Burrah!', and a page with 'cooliris'. Navigation arrows are visible on the left and right sides of the wall.

3D Wall Icons

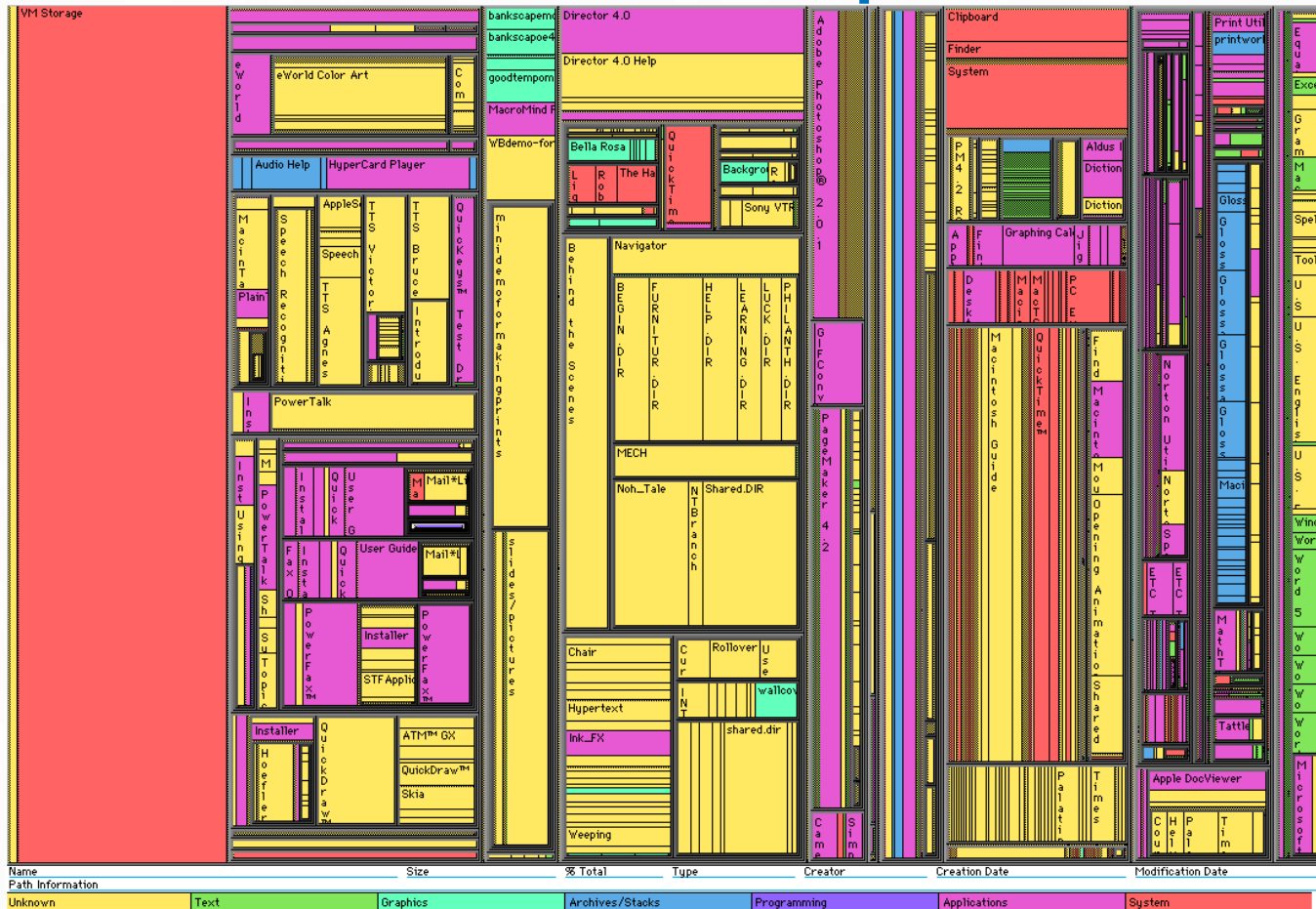
-  Open Page
-  Fullscreen
-  Scroll Wall
-  Start Slideshow
-  Close Slideshow
-  Prev/next
-  Close Page
-  Search

Use Tips

- You need to scroll the wall to locate the results of your search. Images which do not match your search string are greyed out.
- If nothing appears on the 3D Wall, you may have to wait a few seconds for the images to load. If that doesn't work, try

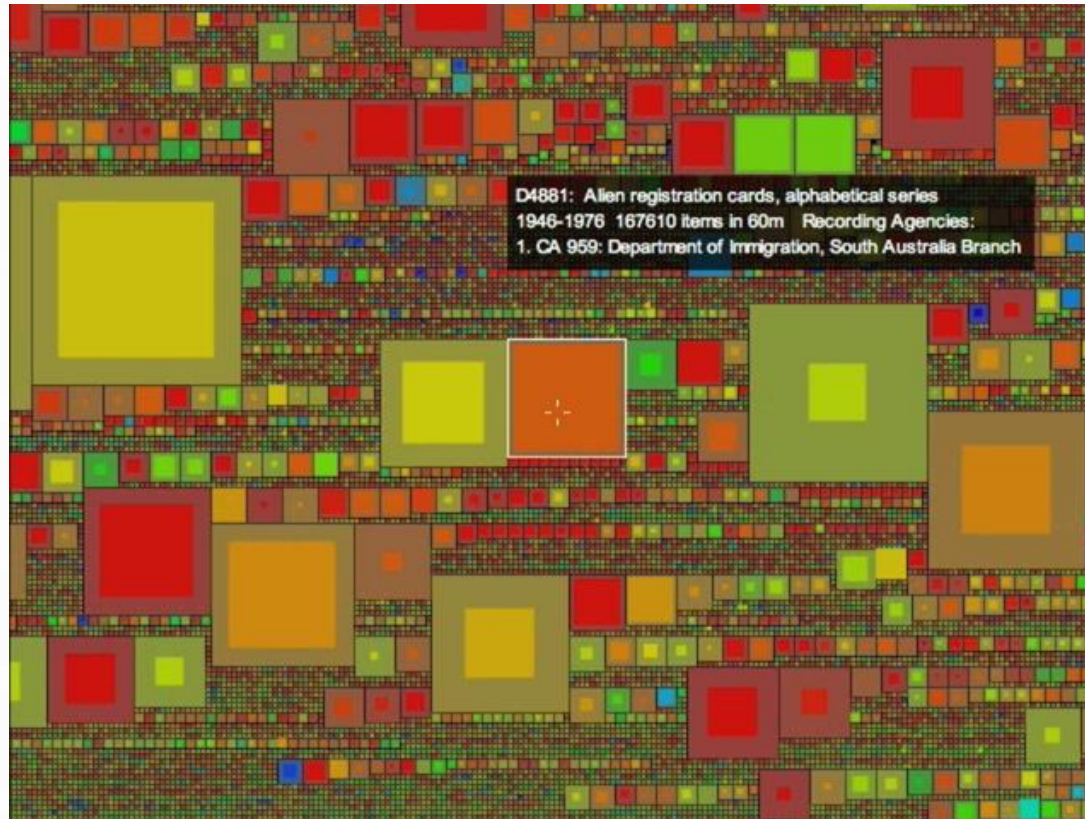
<http://www.webarchive.org.uk/ukwa/wall/Blogs>

Treemap



Johnson and Shneiderman, "Space-Filling Approach to the Visualization of Hierarchical Information Structures" in proceedings of the 2nd conference on Visualization '91

Series Browser



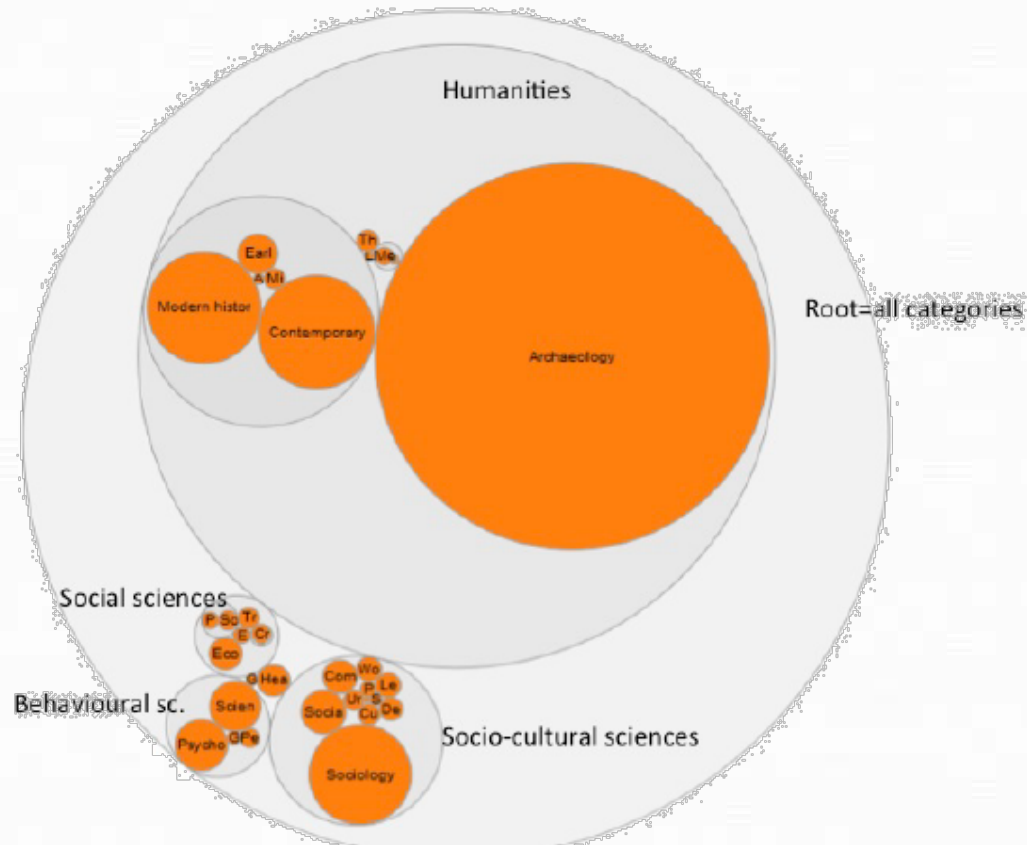
M. Whitelaw, "Visualising Archival Collections: The Visible Archive Project," in *Archives and Manuscripts*, vol. 37, Issue 2, 2009.

A1 Explorer



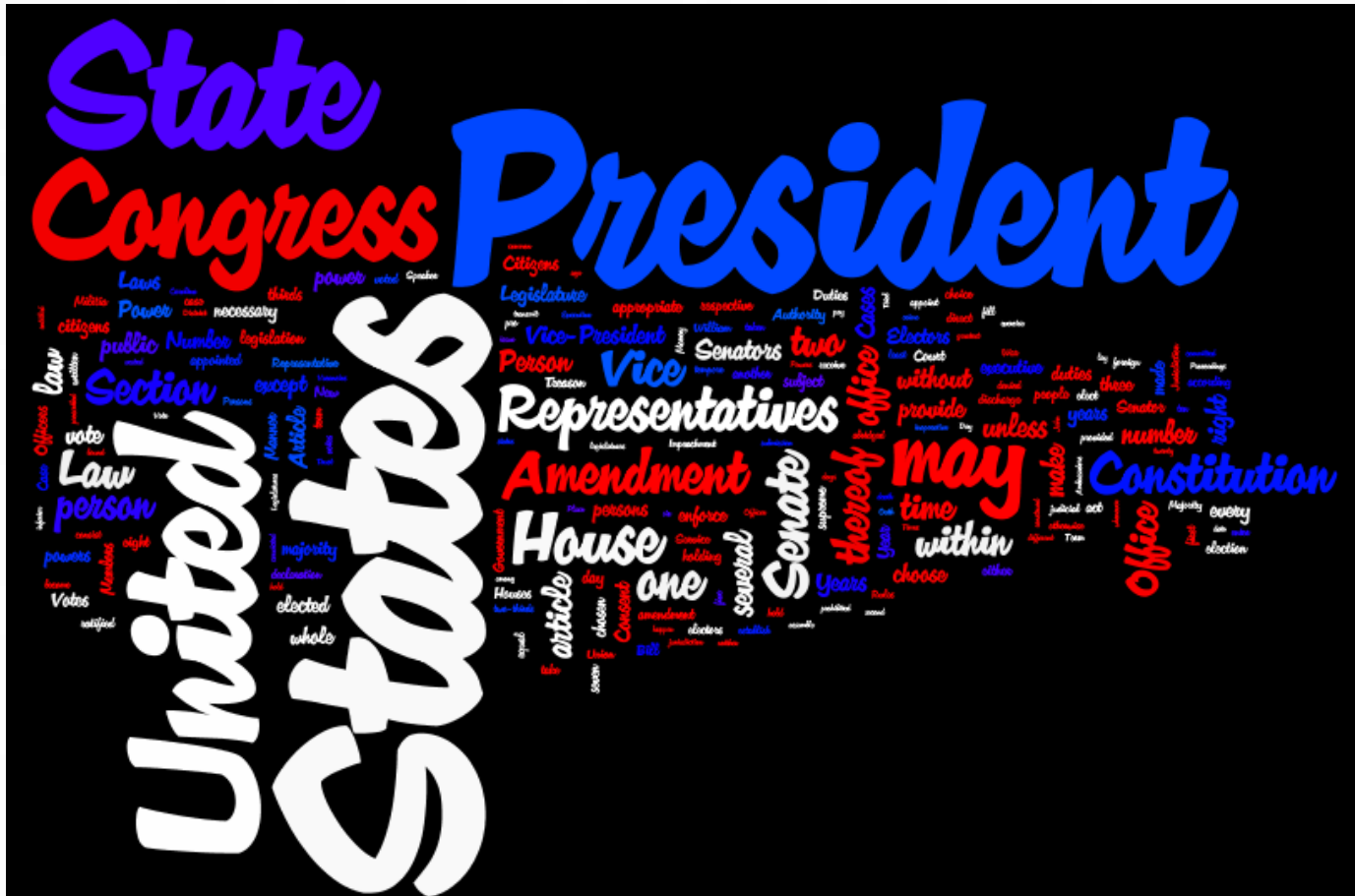
M. Whitelaw, "Visualising Archival Collections: The Visible Archive Project," in *Archives and Manuscripts*, vol. 37, Issue 2, 2009.

EASY



Scharnhorst et.al. "Looking at a digital research data archive Visual interfaces to EASY," in CORR, 2012, <http://arxiv.org/abs/1204.3200>

Wordle



. Jonathan Feinberg, <http://wordle.net/> , Dogear



DATA RETRIEVAL AND PROCESSING

11 Collections, 2K+ Web pages, 70K+ Mementos

| ID | Collection Name | Time Span | Groups | URI Domains | # of Webpages |
|------|---|-----------|--------|-------------|---------------|
| 11 | South Dakota Government | 3 Days | 1 | 50 | 88 |
| 12 | State Minnesota Sites | 3 Weeks | 1 | 6 | 6 |
| 13 | Ari Salomon Archive | 1 Day | 1 | 1 | 1 |
| 194 | NC State Government Web Site Archive | 14 Years | 1 | 451 | 609 |
| 499 | Archive Montana: Preserving State Agency Websites | 15 Years | 36 | 132 | 144 |
| 667 | The New York Greens | 1 Day | 1 | 1 | 1 |
| 677 | Actors Equity Association | 1 Day | 1 | 1 | 1 |
| 1068 | Human Rights | 3 Years | 5 | 341 | 365 |
| 1621 | Chile Earthquake | 1 Day | 1 | 13 | 19 |
| 2323 | Jasmine Revolution - Tunisia 2011 | 5 Months | 4 | 147 | 223 |
| 2836 | Pakistan Floods (2011) | 20 Days | 1 | 253 | 623 |

Data Retrieval & Processing


Retrieval:

- Screen scrape
- Copy collection hierarchy
- Store page content


Processing:

- Calculate TF and TF-IDF
- Generate screenshots
- Generate wordles
- Rule-based categorization
- Construct JSON


No categorization

HOME | EXPLORE | LEARN MORE | CONTACT US

A web archiving service
to harvest and preserve digital collections
a service of the Internet Archive



Explore >> Virginia Tech: Crisis, Tragedy, and Recovery Network >> Pakistan floods (2011)



Pakistan floods (2011)

Collected by: [Virginia Tech: Crisis, Tragedy, and Recovery Network](#)

Archived since: Sep, 2011

Description: The monsoon rains in Sindh province (Sept 2011), in southeastern Pakistan, have submerged more than 4.5 million acres of farming land, damaging an estimated 80% of cash crops. At least 5.5 million people have so far been affected, with more than 200 dead.

Subject: [Floods](#), [Pakistan](#), [Spontaneous Events](#)

[More](#) ▼

Narrow Your Results

There are no further ways to narrow your results.

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Sites **Search Page Text**

Page 1 of 7 (655 Total Results)

Sort By: [Title \(A-Z\)](#) | [Title \(Z-A\)](#) | [URL \(A-Z\)](#) | [URL \(Z-A\)](#)

URL: <http://abnews.go.com/MillionMomsChallenge/slideshow/flooding-ravages-pakistan-unhcr-relief-14565409>
captured on Sep 28, 2011

URL: <http://acdi-cida.gc.ca/acdi-cida/ACDI-CIDA.nsf/eng/ANN-820133234-NKW/>
Title: Pakistan Flood Relief Fund - CIDA
Captured on Sep 16, 2011

URL: <http://adf.ly/2roFB>
captured on Sep 28, 2011

URL: <http://adf.ly/2roFC>
captured on Sep 28, 2011

URL: <http://adf.ly/2roOP>
captured on Sep 28, 2011

Improper Categorization

The screenshot shows the Internet Archive interface for the collection 'Jasmine Revolution - Tunisia 2011'. The page includes a navigation menu, a search bar, and a list of search results. A red box highlights a result with the URL <http://www.youtube.com/watch?v=ob5Vs9BkN3M>, which is categorized as 'Social Media Sites'.

ARCHIVE-IT HOME EXPLORE LEARN MORE
CONTACT US

A web archiving service to harvest and preserve digital collections
a service of the Internet Archive

Explore >> Internet Archive Global Events >> Jasmine Revolution - Tunisia 2011

INTERNET ARCHIVE

Jasmine Revolution - Tunisia 2011
Collected by: [Internet Archive Global Events](#)
Archived since: Jan, 2011
Description: This collection consists of websites documenting the revolution in Tunisia in 2011. Our partners at Library of Congress and Bibliothèque Nationale de France have contributed websites for this collection, and the sites are primarily in French and Arabic with some in English.

Narrow Your Results

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Group
Blogs and Other Sites (95)
News Sites and Articles (12)
Social Media Sites (53)

Enter search terms here

Sites Search Page Text

Page 1 of 3 (235 Total Results)

Sort By: [Title \(A-Z\)](#) | [Title \(Z-A\)](#) | [URL \(A-Z\)](#) | [URL \(Z-A\)](#)

URL: <http://www.youtube.com/watch?v=pGMqdOgERBw>
Captured 2 times between Feb 16, 2011 and Feb 28, 2011
Videos: 2 Videos Captured

Group: Social Media Sites

URL: <http://www.youtube.com/watch?v=omWjoDISOLE&feature=related>
Captured on Feb 28, 2011
Videos: 5 Videos Captured

URL: <http://www.youtube.com/watch?v=ob5Vs9BkN3M>
Captured 5 times between Mar 18, 2011 and Oct 22, 2011
Videos: 22 Videos Captured

Rule based categorization



Pakistan floods (2011)
Collected by: Virginia Tech: Crisis, Tragedy, and Recovery Network
Archived since: Sep, 2011
Description: The monsoon rains in Sindh province (Sept 2011), in southeastern Pakistan, have submerged more than 4.5 million acres of farming land, damaging an estimated 80% of cash crops. At least 5.5 million people have so far been affected, with more than 200 dead.
Subject: [Floods](#), [Pakistan](#), [Spontaneous Events](#)
[More ▾](#)

Narrow Your Results

There are no further ways to narrow your results.

Page 1 of 7 (655 Total Results)

[Next Page ▶](#)

Sort By: [Title \(A-Z\)](#) | [Title \(Z-A\)](#) | [URL \(A-Z\)](#) | [URL \(Z-A\)](#)

URL: <http://abcnews.go.com/MillionMomsChallenge/slideshow/flooding-ravages-pakistan-unhcr-relief-14565409>
Captured on Sep 28, 2011

URL: http://fewww.wordpress.com/2011/09/22/pakistan-flood-death-toll-tops-1000/?utm_source=twitterfeed&utm_medium=twitter
Captured on Sep 28, 2011

URL: <http://floodmaps.lums.edu.pk/>
Title: Pakistan FloodMAPS
Captured on Sep 16, 2011

URL: http://hithot.cc/en_UK/Pakistan/

URL: <http://www.facebook.com/147433618618300/posts/148775358484126/>
Captured on Sep 28, 2011

URL: http://www.youtube.com/watch?v=egZFQ_kucTE
Captured on Sep 28, 2011

News Web pages

Blogs

Social Media

Videos

Special URI and TLD based categorization



Pakistan floods (2011)
Collected by: Virginia Tech: Crisis, Tragedy, and Recovery Network
Archived since: Sep, 2011
Description: The monsoon rains in Sindh province (Sept 2011), in southeastern Pakistan, have submerged more than 4.5 million acres of farming land, damaging an estimated 80% of cash crops. At least 5.5 million people have so far been affected, with more than 200 dead.
Subject: [Floods](#), [Pakistan](#), [Spontaneous Events](#)
[More ▾](#)

Narrow Your Results

There are no further ways to narrow your results.

Page 1 of 7 (655 Total Results) [Next Page ▶](#)

Sort By: [Title \(A-Z\)](#) | [Title \(Z-A\)](#) | [URL \(A-Z\)](#) | [URL \(Z-A\)](#)

URL: <http://abcnews.go.com/MillionMomsChallenge/slideshow/flooding-ravages-pakistan-unhcr-relief-14565409>
Captured on Sep 28, 2011

URL: http://fewww.wordpress.com/2011/09/22/pakistan-flood-death-toll-tops-1000/?utm_source=twitterfeed&utm_medium=twitter
Captured on Sep 28, 2011

URL: <http://floodmaps.lums.edu.pk/>
Title: Pakistan FloodMAPS
Captured on Sep 16, 2011

URL: http://hithot.cc/en_UK/Pakistan/

URL: <http://www.facebook.com/147433618618300/posts/148775358484126/>
Captured on Sep 28, 2011

URL: http://www.youtube.com/watch?v=egZFQ_kucTE
Captured on Sep 28, 2011

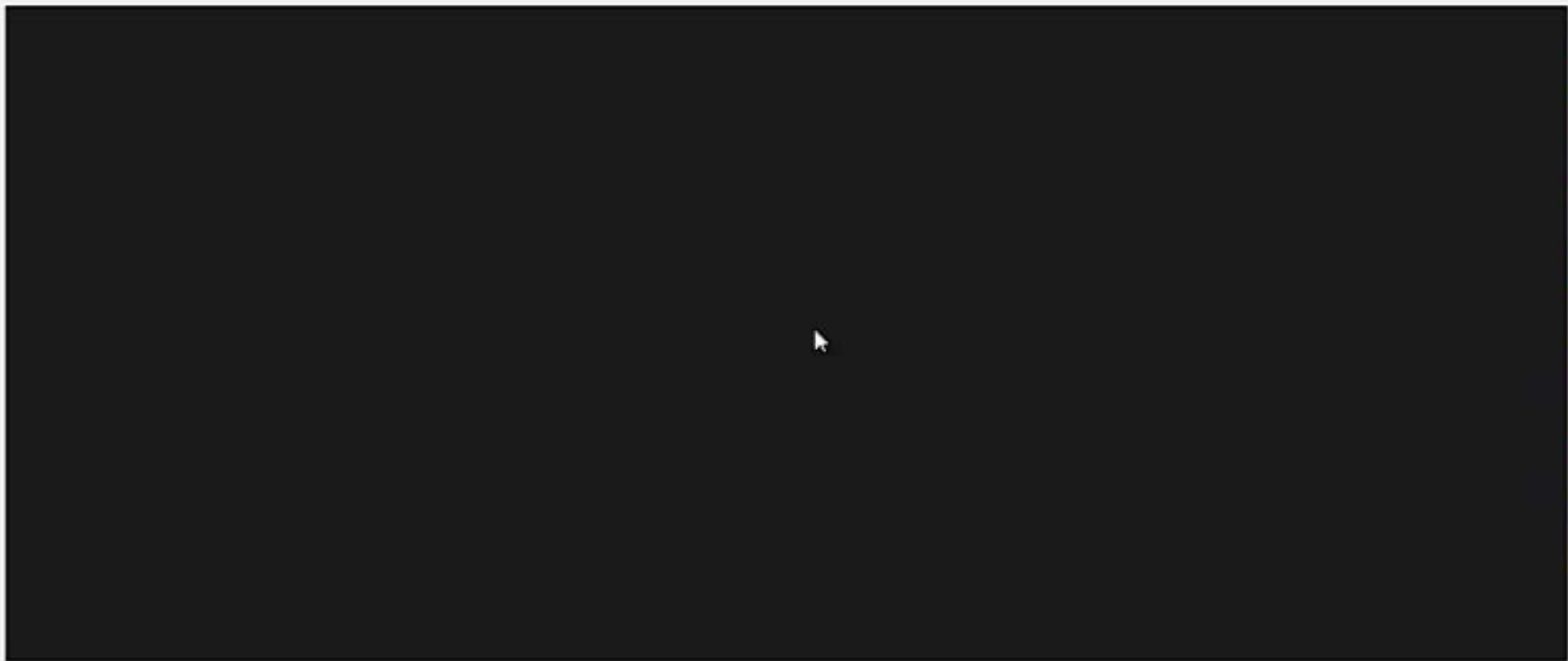
➔ Pakistani news web pages



VISUALIZATIONS

Treemap

Collection ID:



Days old

- 1 day
- < 1 week
- 1 - 2 weeks
- 2 - 4 weeks
- 1 - 2 months
- 2 - 4 months
- 4 - 8 months
- 8 - 12 months
- > 1 year

Time Cloud

01-05-1997 02-10-1997



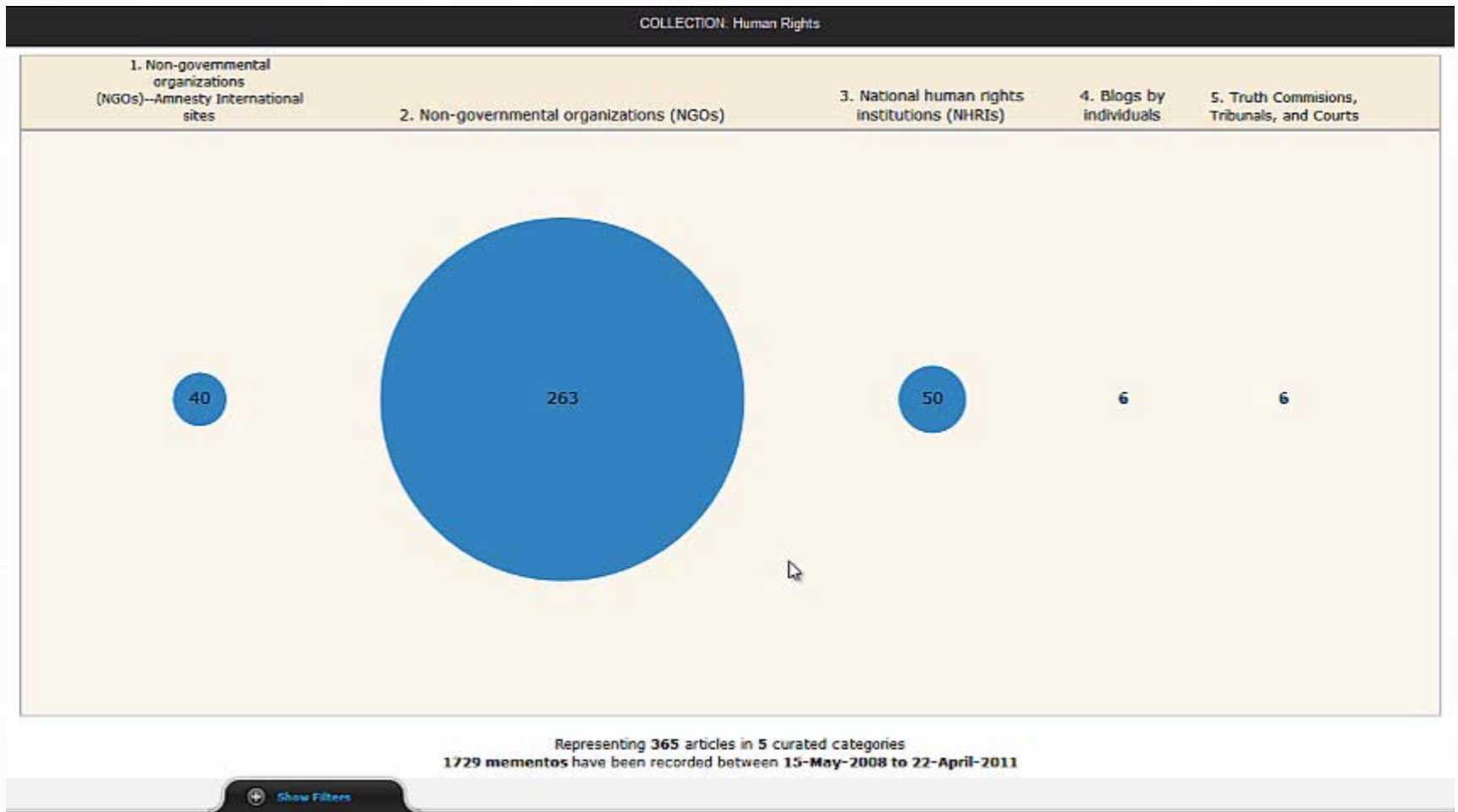
TF



TF-IDF



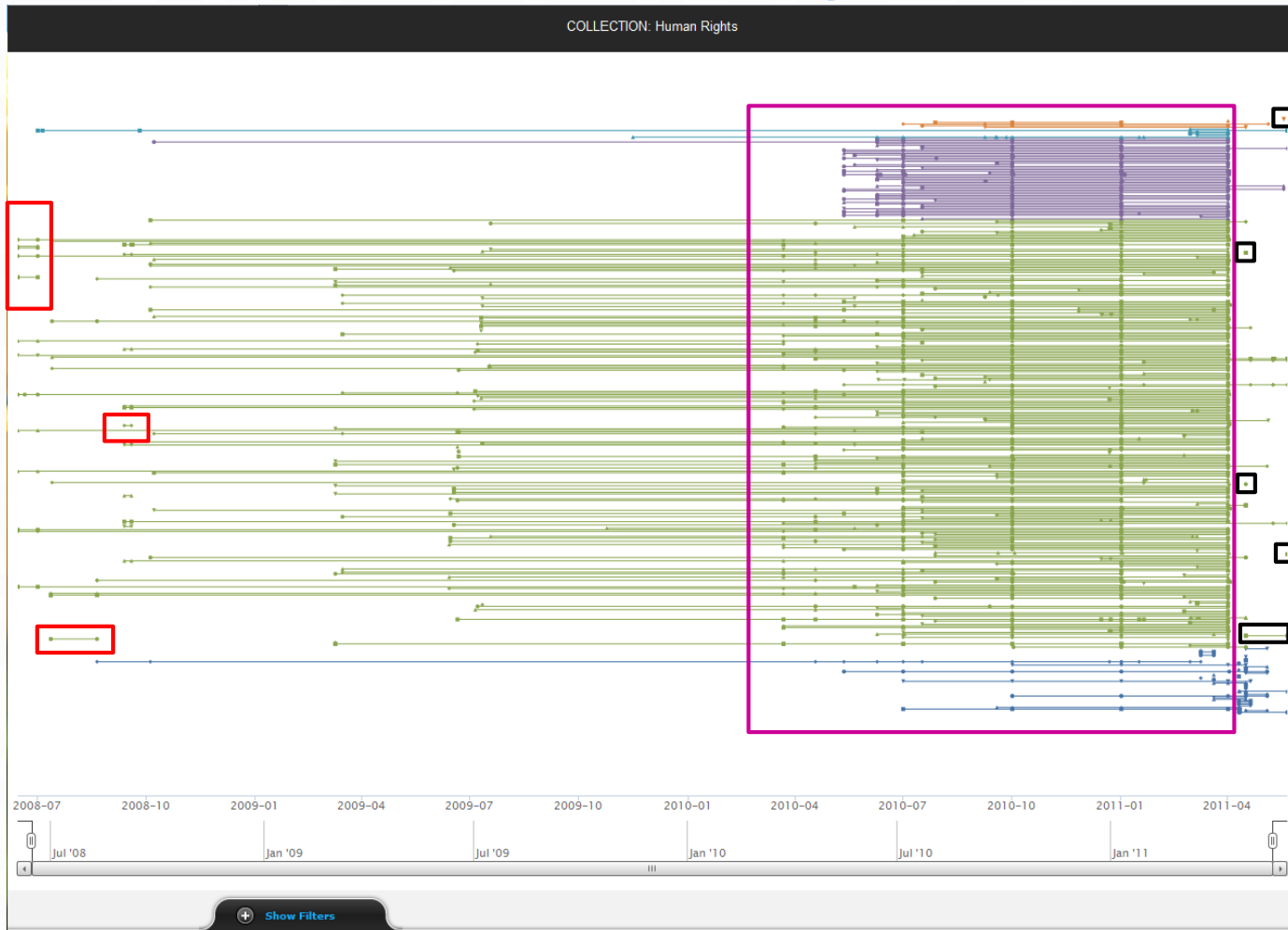
Bubble Chart, Image Plot & Timeline



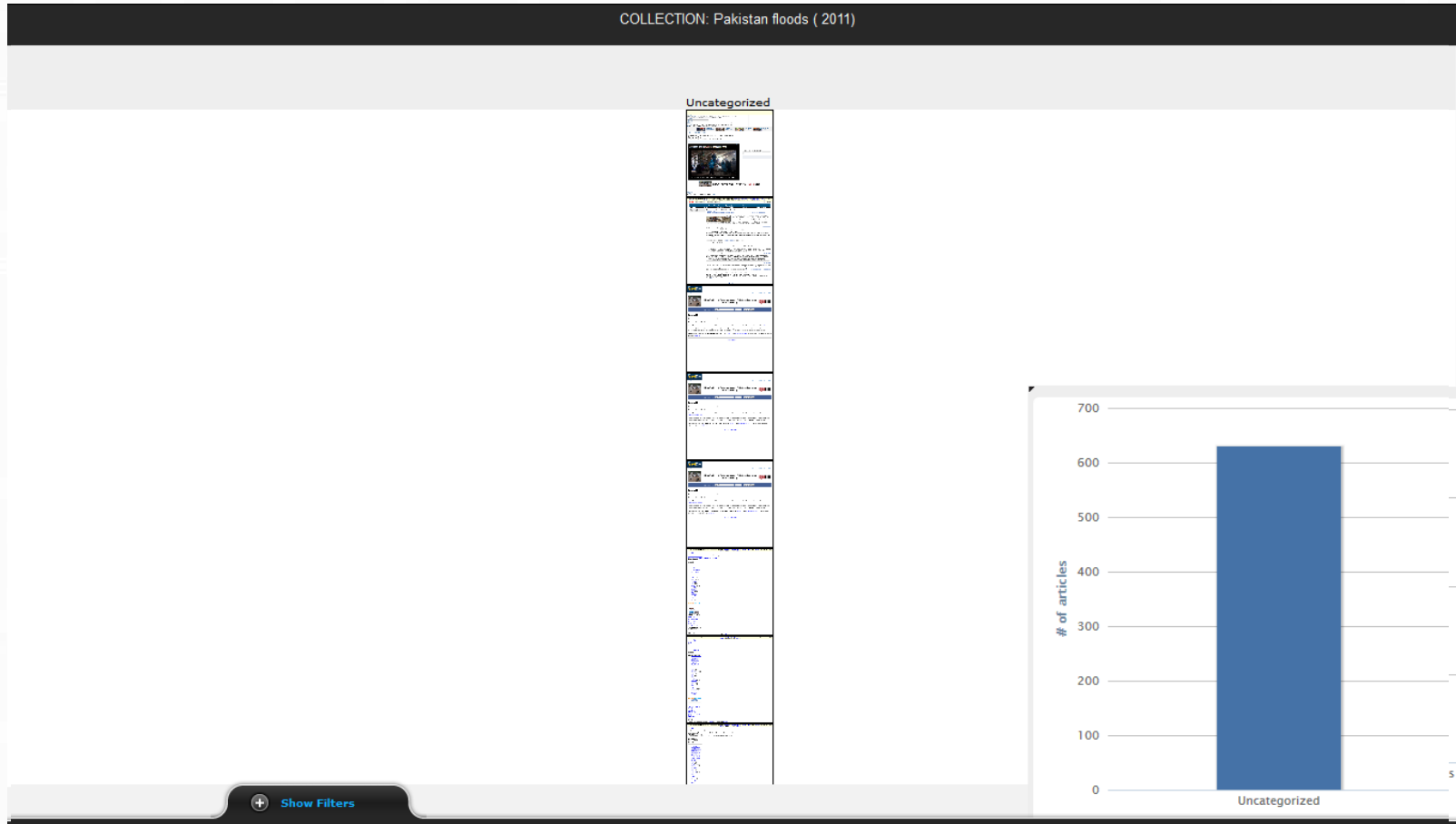


CASE STUDIES

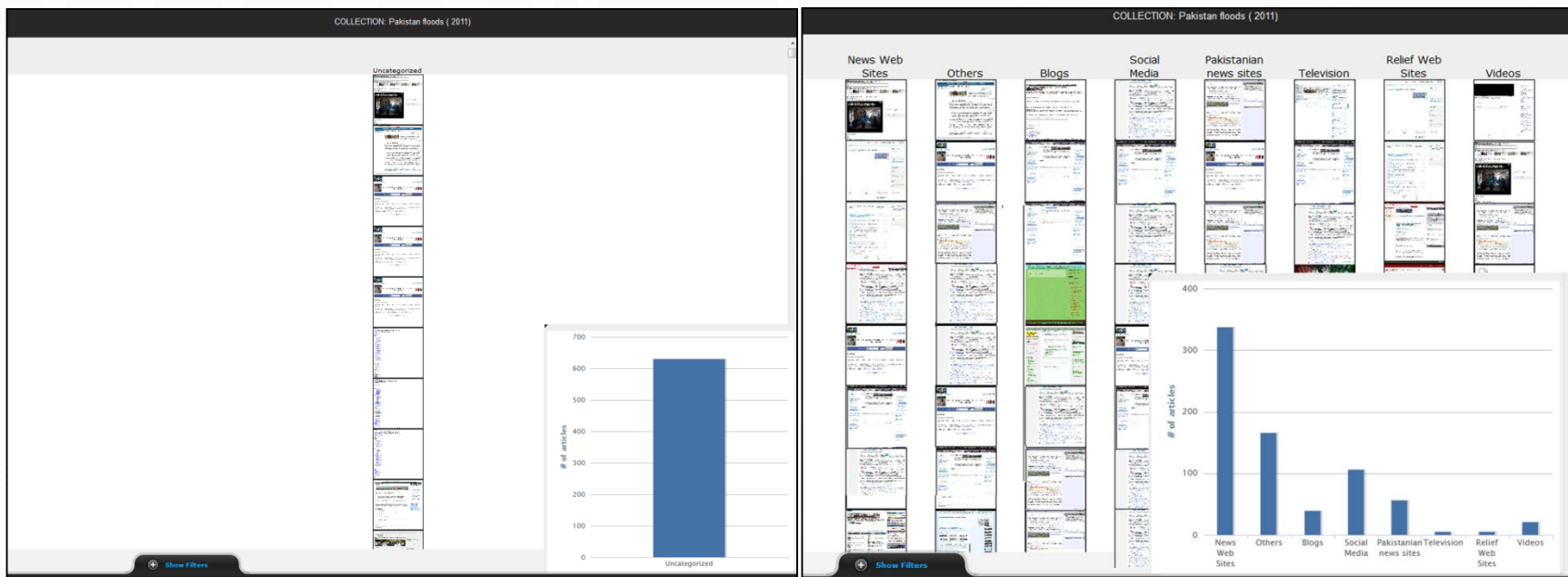
1. Collection Building and Growth



2. Re-Categorization (Pakistan Flood: no categorization)



2. Re-Categorization (Pakistan Flood: after categorization)



3. Collection Synopsis



The screenshot displays the Archive-It website interface. At the top right, there are social media icons for Facebook and Twitter, and a 'Login' button. The main navigation bar includes 'HOME', 'EXPLORE', 'LEARN MORE', and 'CONTACT US'. A tagline reads: 'A web archiving service to harvest and preserve digital collections a service of the Internet Archive'. The breadcrumb trail is 'Explore >> Virginia Tech: Crisis, Tragedy, and Recovery Network >> Pakistan floods (2011)'. The main content area features a photo of people wading through floodwaters, with the title 'Pakistan floods (2011)' and the collector 'Virginia Tech: Crisis, Tragedy, and Recovery Network'. It also lists the archive date as 'Sep, 2011', a description of the monsoon rains in Sindh province, and subject tags: 'Spontaneous Events, Floods, Pakistan'. A 'More' dropdown menu is visible. Below the synopsis, the text 'Narrow Your Results' is followed by instructions on how to filter the collection.

ARCHIVE-IT

HOME EXPLORE LEARN MORE CONTACT US

A web archiving service
to harvest and preserve digital collections
a service of the Internet Archive

Explore >> Virginia Tech: Crisis, Tragedy, and Recovery Network >> Pakistan floods (2011)

Pakistan floods (2011)
Collected by: [Virginia Tech: Crisis, Tragedy, and Recovery Network](#)
Archived since: Sep, 2011
Description: The monsoon rains in Sindh province (Sept 2011), in southeastern Pakistan, have submerged more than 4.5 million acres of farming land, damaging an estimated 80% of cash crops. At least 5.5 million people have so far been affected, with more than 200 dead.
Subject: Spontaneous Events, Floods, Pakistan
[More](#) ▼

Narrow Your Results

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

3. Collection Synopsis

The screenshot shows a web browser window displaying the Internet Archive website. The address bar shows the URL archive-it.org/collections/2836. The page title is "Archive-It - Pakistan floods (2011)".

The main content area features a collection synopsis for "Pakistan floods (2011)". The synopsis includes a small image of people wading through floodwaters. The text reads: "Pakistan floods (2011) Collected by: Virginia Tech: Crisis, Tragedy, and Recovery Network Archived since: Sep, 2011 Description: The monsoon rains in South province (Sep 2011), in southeastern Pakistan, have submerged more than 4.5 million acres of farming land, damaging an estimated 80% of cash crops. At least 5.5 million people have so far been affected, with more than 200 dead. Subject: Spontaneous Events - Floods, Pakistan More ▾".

Below the synopsis is a "Narrow Your Results" section with a search bar and a "Search" button. The search bar contains the text "Enter search terms here".

The search results section shows "Page 1 of 7 (655 Total Results)". The results are sorted by "Title (A-Z)". The first result is: "URL: <http://abcnews.go.com/MillionMomsChallenge/slideshow/flooding-ravages-pakistan-unhcr-relief-14565409> Captured once on Sep 26, 2011".

The second result is: "Title: Pakistan Flood Relief Fund - CIDA URL: <http://acdi-cida.gc.ca/acdi-cida/ACDI-CIDA.nsf/eng/ANN-820133234-NKW/> Captured once on Sep 16, 2011".

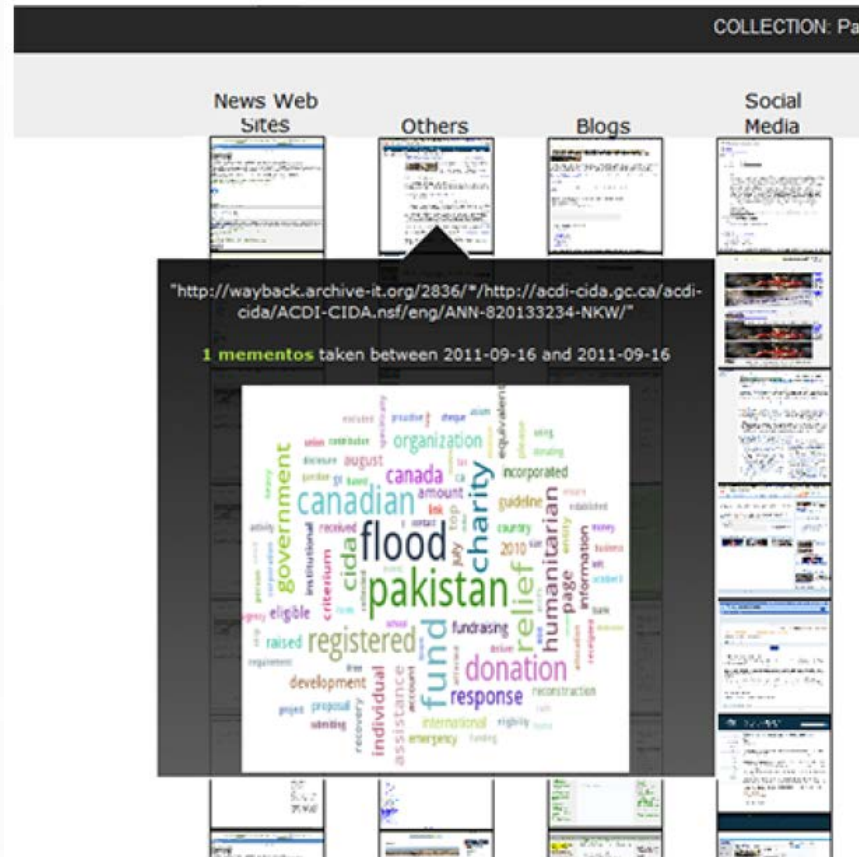
The third result is: "URL: <http://adf.ly/2roFB> Captured once on Sep 28, 2011".

The fourth result is: "URL: <http://adf.ly/2roFC> Captured once on Sep 28, 2011".

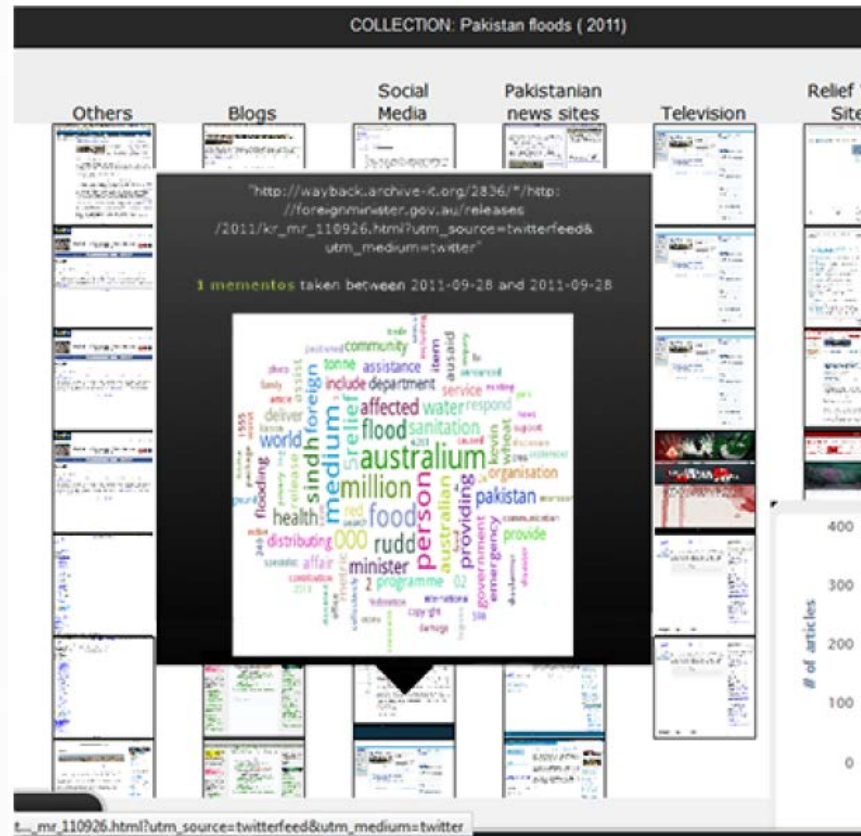
The fifth result is: "URL: <http://adf.ly/2roOP> Captured once on Sep 28, 2011".

The sixth result is: "URL: <http://aidnews.org/pakistan-another-year-another-flood/> Captured once on Sep 28, 2011".

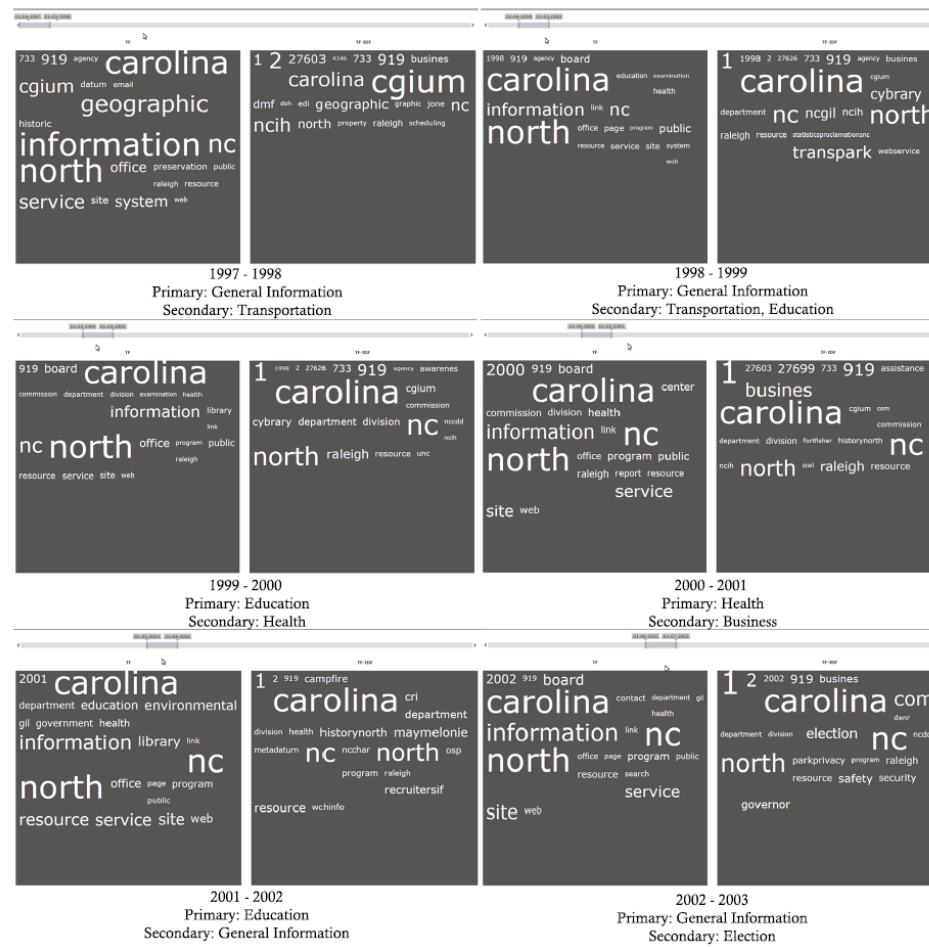
3. Collection Synopsis



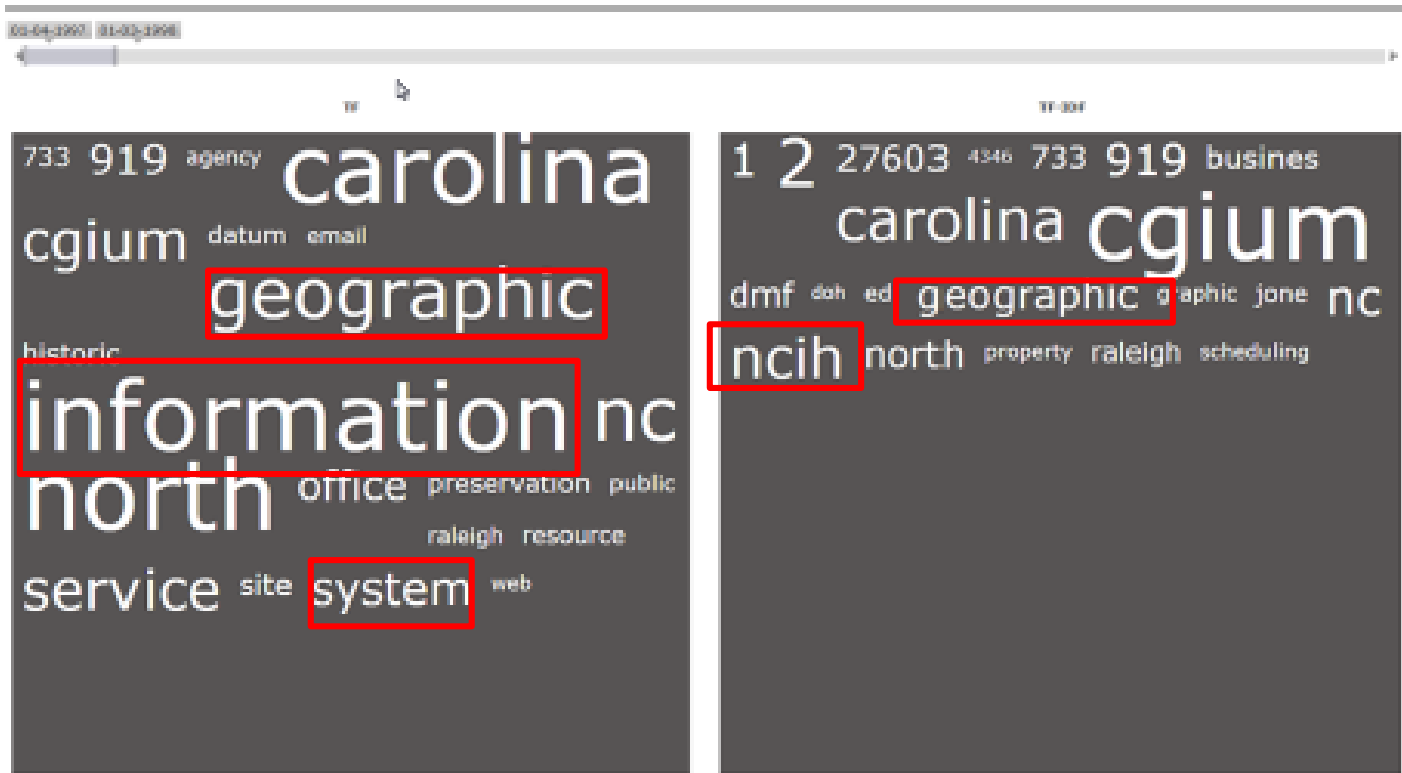
3. Collection Synopsis



4. Theme Tracking



4. Theme Tracking

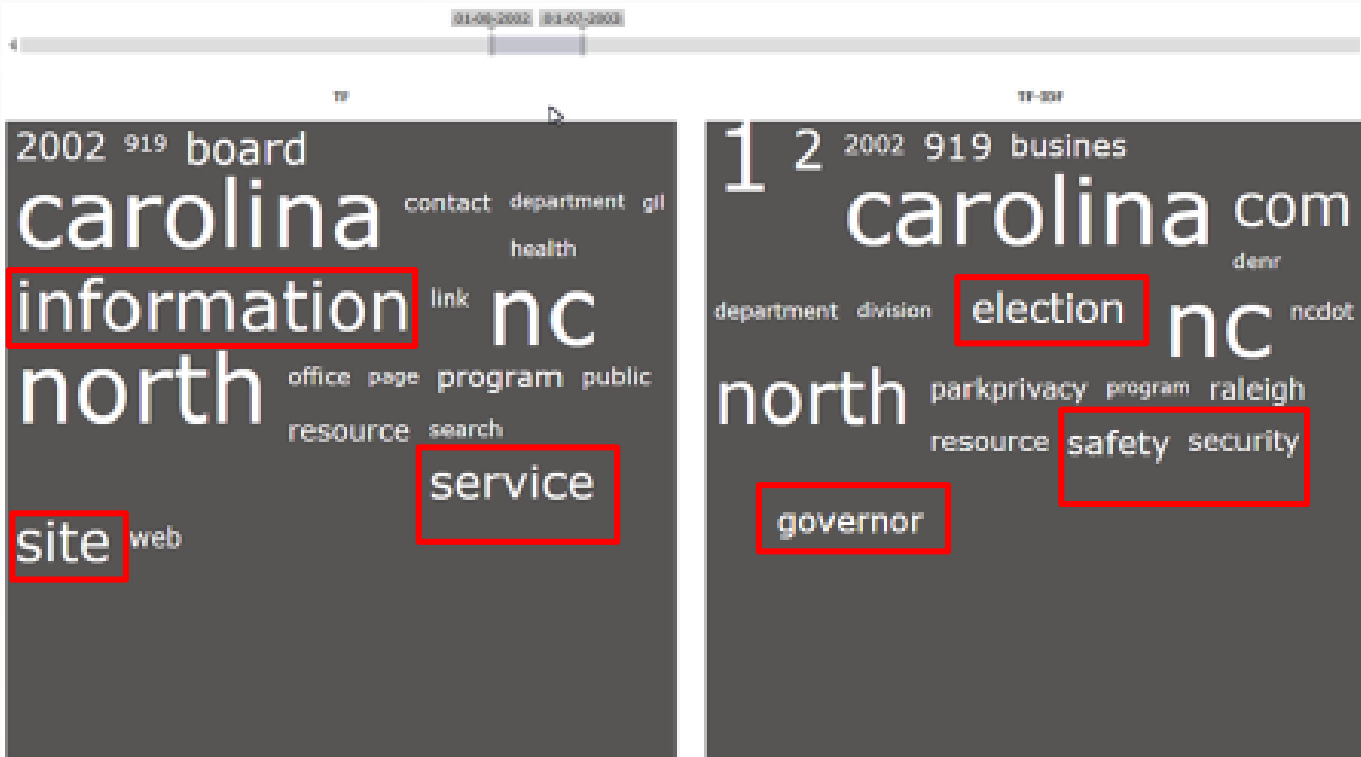


1997 - 1998

Primary: General Information

Secondary: Transportation

4. Theme Tracking

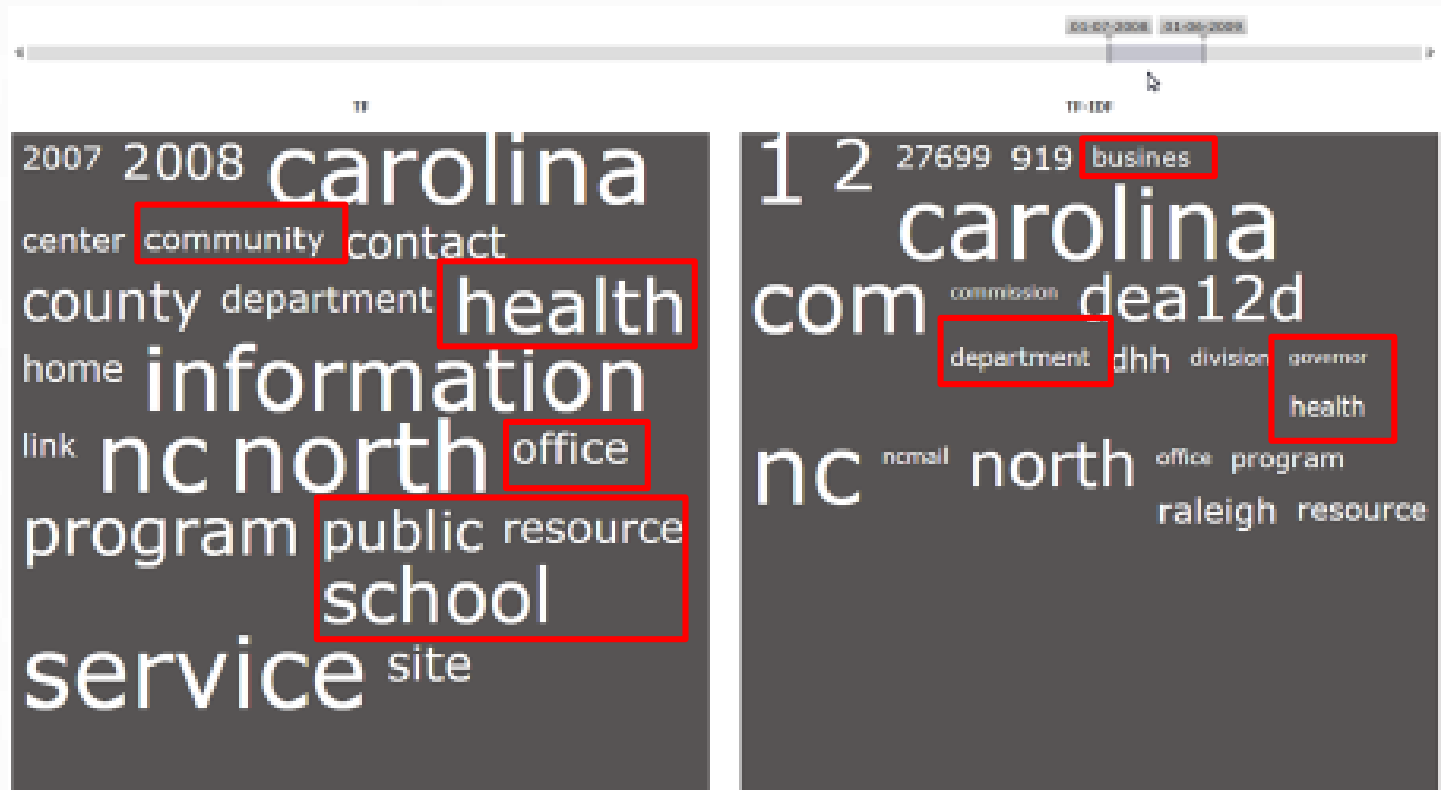


2002 - 2003

Primary: General Information

Secondary: Election

4. Theme Tracking



2008 - 2009

Primary: Education, Health
Secondary: Government, Business

Informal User Evaluation

✚ Alex Thurman, Columbia University Libraries

✚ Feedback on

- ease of browsing and obtaining information
- *user-friendliness of the interface*
- *whether they prefer textual or graphical interface*
- *most effective visualization*
- *effectiveness of the rule-based categorization in exploring archives*

Feedback

✚ Effective visualizations:

- Treemap – color coding useful for identifying newer additions
- Image plot – screenshots with mouse-over wordles allow for good navigation
- Timeline – useful for visualizing development of groups in collection

✚ Suggestions

- Broader timescale for treemaps
- Include stop words from other languages



FUTURE WORK AND CONCLUSION

Future Work

- ✚ N-Gram wordles
- ✚ Term expansion
- ✚ Krovetz stemmer (dictionary based stemmer)
- ✚ Integration with Archive-It
- ✚ Detailed user evaluation
- ✚ Implementation for other archives

Conclusion

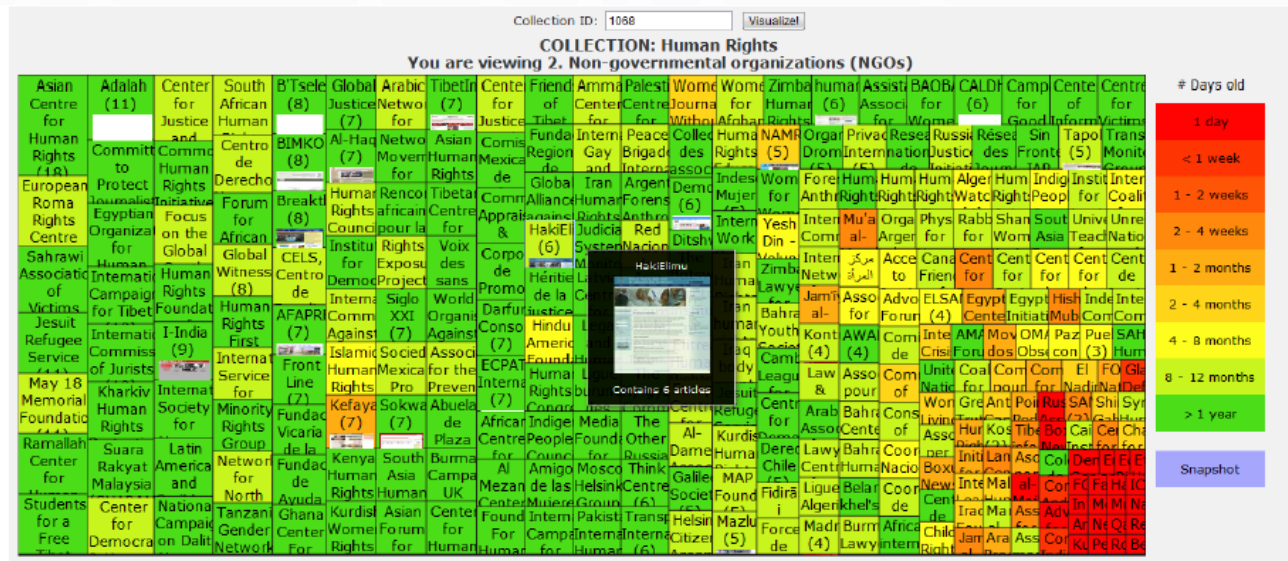
 ***Identified metrics for collections***

Conclusion

Identified metrics for collections

Visualizations

- Treemap



Conclusion

✚ Identified metrics for collections

✚ Visualizations

- Treemap
- ***Time cloud***

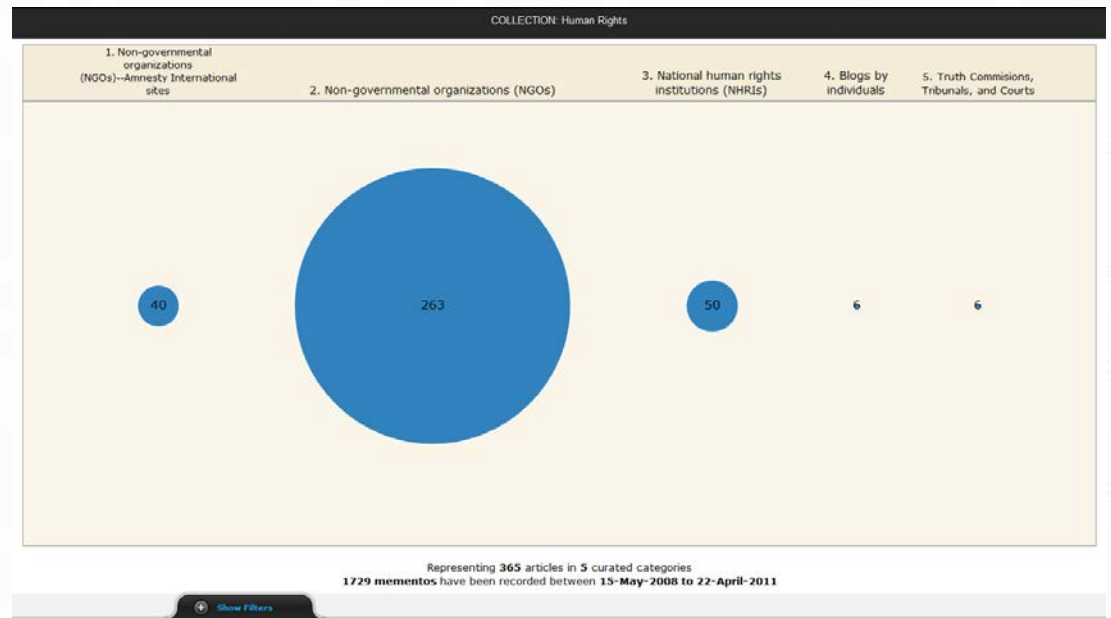


Conclusion

✚ Identified metrics for collections

✚ Visualizations

- Treemap
- Time cloud
- ***Bubble chart***

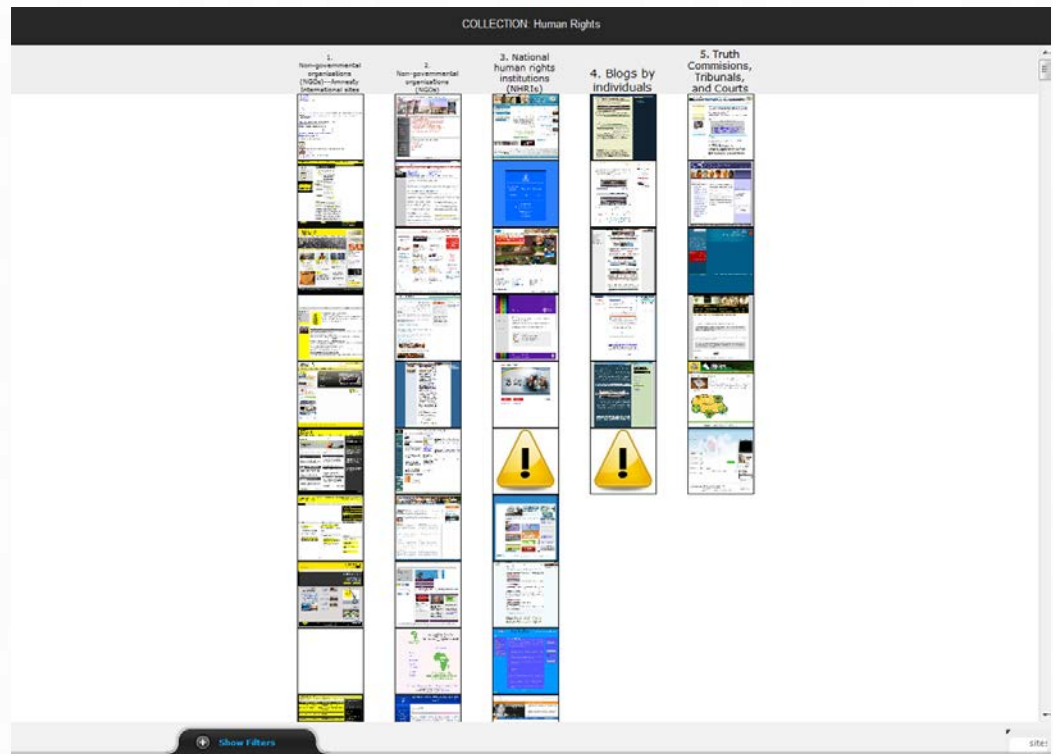


Conclusion

Identified metrics for collections

Visualizations

- Treemap
- Time cloud
- Bubble chart
- ***Image plot***



Conclusion

✚ Identified metrics for collections

✚ Visualizations

- Treemap
- Time cloud
- Bubble chart
- Image plot
- **Wordle**

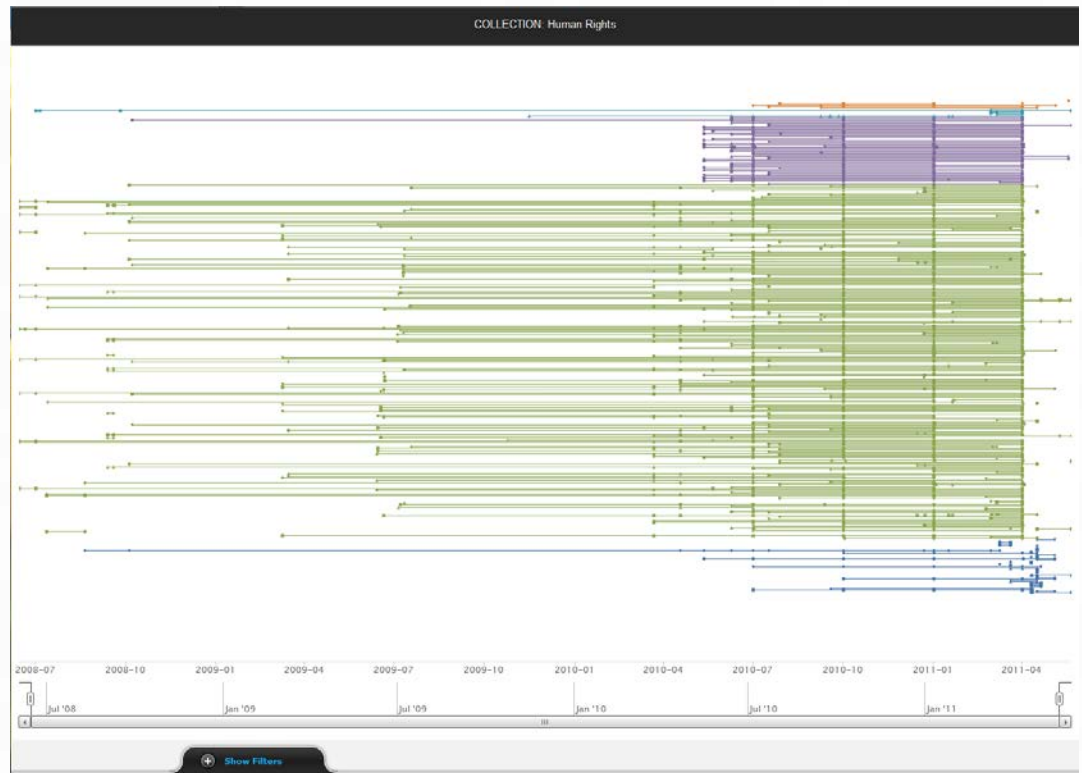


Conclusion

✚ Identified metrics for collections

✚ Visualizations

- Treemap
- Time cloud
- Bubble chart
- Image plot
- Wordle
- ***Timeline***

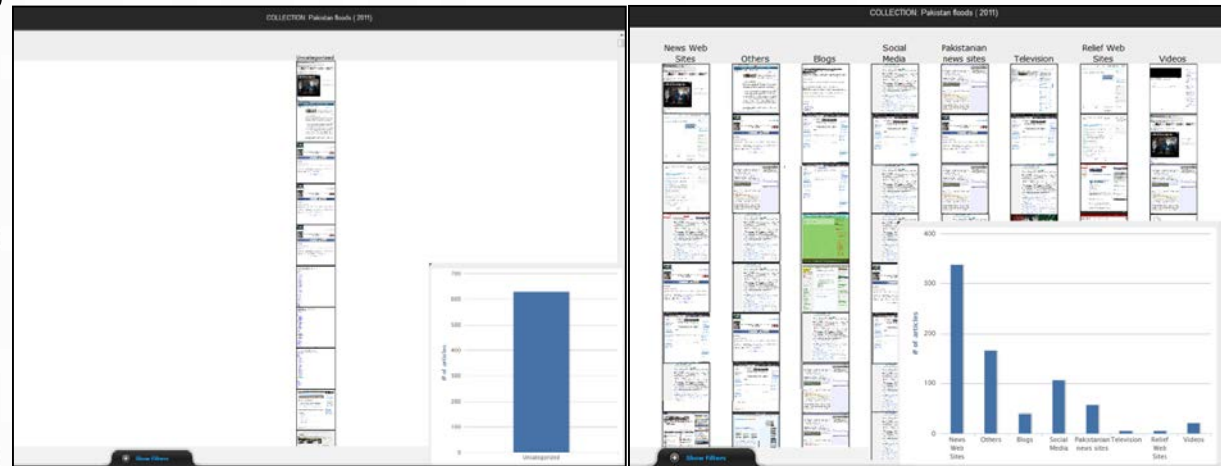


Conclusion

✚ Identified metrics for collections

✚ Visualizations

- Treemap
- Time cloud
- Bubble chart
- Image plot
- Wordle
- Timeline



✚ ***Rule – based categorization***



BACKUP

Time Span

Center for
Human Rights
Documentation
& Research

Human Rights Web Archive (Columbia University
Libraries)



Enter Web Address: All

Searched for <http://amigosdemujeres.org/> 10 Results [RSS](#) [Metadata](#)
[Look up URL](#) in general Internet Archive web collection [Proxy Mode Help](#)

* denotes when page was updated

| Found 10 Captures between May 14, 2009 - Mar 2, 2012 | | | |
|--|---|--|-------------------------------|
| 2009 | 2010 | 2011 | 2012 |
| 1 page | 4 pages | 4 pages | 1 page |
| May 14, 2009 * | Mar 19, 2010 * Jun 2, 2010 * Sep 2, 2010 * Dec 2, 2010 * | Mar 2, 2011 * Jun 2, 2011 * Sep 2, 2011 * Dec 2, 2011 * | Mar 2, 2012 * |

[Home](#) | [Internet Archive](#)

http://wayback.archive-it.org/1068/*/http://amigosdemujeres.org/

| | | |
|-----------|--------|--------------------|
| Time span | Small | 1 Day - 2 Weeks |
| | Medium | 2 Weeks - 4 Months |
| | Large | > 4 Months |

Groups

ARCHIVE-IT HOME EXPLORE LEARN MORE CONTACT US

A web archiving service to harvest and preserve digital collections a service of the Internet Archive

Explore >> Columbia University Libraries >> Human Rights

Human Rights
Collected by: [Columbia University Libraries](#)

Archived since: May, 2008

Description: An initiative of CUL's Center for Human Rights Documentation and Research, the Human Rights Web Archive is an effort to preserve and ensure access to freely available human rights resources created mainly by non-governmental organizations, national human rights institutions, and individuals.

Subject: [Society & Culture](#), [Human rights](#), [Non-governmental organizations](#), [Human rights workers](#), [National human rights institutions](#), [Web archives](#)
[More ▼](#)

Narrow Your Results

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Group

- Amnesty International sections (48)
- Blogs by individuals (6)
- National human rights institutions (71)
- Non-governmental organizations (356)
- Truth commissions, tribunals, and courts (12)

Subject

- Human rights (398)

Enter search terms here

Sites Search Page Text

Page 1 of 5 (499 Total Results)

Sort By: [Title \(A-Z\)](#) | [Title \(Z-A\)](#) | [URL \(A-Z\)](#) | [URL \(Z-A\)](#)

| | | |
|--------|--------|-------|
| Groups | Small | 1 |
| | Medium | 2 - 5 |
| | Large | > 5 |

<http://www.archive-it.org/collections/1068>

URI Domains

ARCHIVE-IT HOME EXPLORE LEARN MORE CONTACT US A web archiving service to harvest and preserve digital collections a service of the Internet Archive

Explore >> Virginia Tech: Crisis, Tragedy, and Recovery Network >> Pakistan floods (2011)

Pakistan floods (2011)
Collected by: Virginia Tech: Crisis, Tragedy, and Recovery Network
Archived since: Sep, 2011
Description: The monsoon rains in Sindh province (Sept 2011), in southeastern Pakistan, have submerged more than 4.5 million acres of farming land, damaging an estimated 80% of cash crops. At least 5.5 million people have so far been affected, with more than 200 dead.
Subject: Spontaneous Events, Floods, Pakistan
More ▾

Narrow Your Results

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

There are no further ways to narrow your results.

Enter search terms here

Sites Search Page Text

Page 1 of 7 (655 Total Results)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

URI: <http://abcnews.go.com/MillionMomsChallenge/slideshow/flooding-ravages-pakistan-unhcr-relief-14565409>
Captured once on Sep 28, 2011

URI: <http://beenasarwar.wordpress.com/2011/09/23/general-observations-about-pakistan-floods/>
Captured once on Sep 28, 2011

URI: <http://beenasarwar.wordpress.com/flood-relief-info/>
Captured once on Sep 28, 2011

Title: Pakistan floods
URI: http://care.ca/main/index.php?pakistan_flood
Captured once on Sep 16, 2011
More ▾

URI: <http://chimalaya.org/2011/09/21/un-appeals-for-pakistan-flood-aid/>
Captured once on Sep 28, 2011

URI: http://chimalaya.org/2011/09/21/un-appeals-for-pakistan-flood-aid/?utm_medium=referral&utm_source=pulseneews

| | | |
|-------------|--------|---------|
| URI Domains | Small | 1 - 10 |
| | Medium | 11 - 20 |
| | Large | > 20 |

<http://www.archive-it.org/collections/2836>

Number of Web Pages



Pakistan floods (2011)
Collected by: [Virginia Tech Crisis, Tragedy, and Recovery Network](#)
Archived since: Sep, 2011
Description: The monsoon rains in Sindh province (Sept 2011), in southeastern Pakistan, have submerged more than 4.5 million acres of farming land, damaging an estimated 80% of cash crops. At least 3.5 million people have so far been affected, with more than 200 dead.
Subject: [Spontaneous Events](#), [Floods](#), [Pakistan](#)
[More ▾](#)

Narrow Your Results

There are no further ways to narrow your results.

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Sites Search Page Text

Page 1 of 7 (655 Total Results)

Sort By: [Title \(A-Z\)](#) | [Title \(Z-A\)](#) | [URL \(A-Z\)](#) | [URL \(Z-A\)](#)

URI: <http://abcnews.go.com/MillionMomsChallenge/slideshow/flooding-ravages-pakistan-unhcr-relief-14565409>
Captured once on Sep 28, 2011

Title: **Pakistan Flood Relief Fund - CIDA**
URI: <http://acdi-cida.gc.ca/acdi-cida/ACDI-CIDA.nsf/eng/ANN-820133234-NKW/>
Captured once on Sep 16, 2011

URI: <http://adf.ly/2roFB>
Captured once on Sep 28, 2011

URI: <http://adf.ly/2roFC>
Captured once on Sep 28, 2011

URI: <http://adf.ly/2roOP>
Captured once on Sep 28, 2011

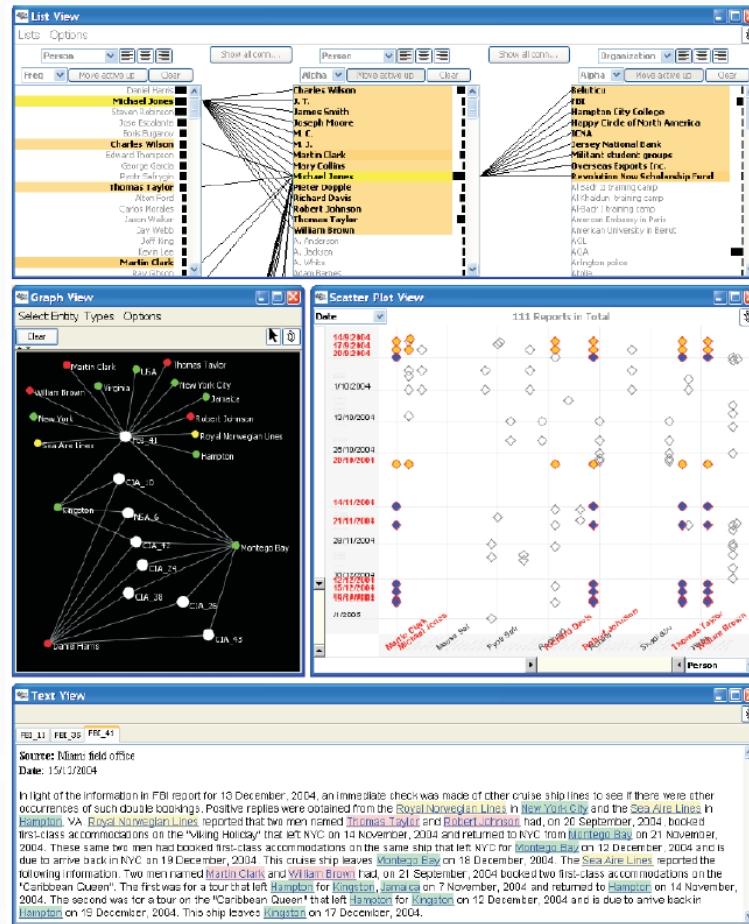
URI: <http://aidnews.org/pakistan-another-year-another-flood/>
Captured once on Sep 28, 2011

URI: <http://aidresources.org/pakistan-another-year-another-flood/>
Captured once on Sep 28, 2011

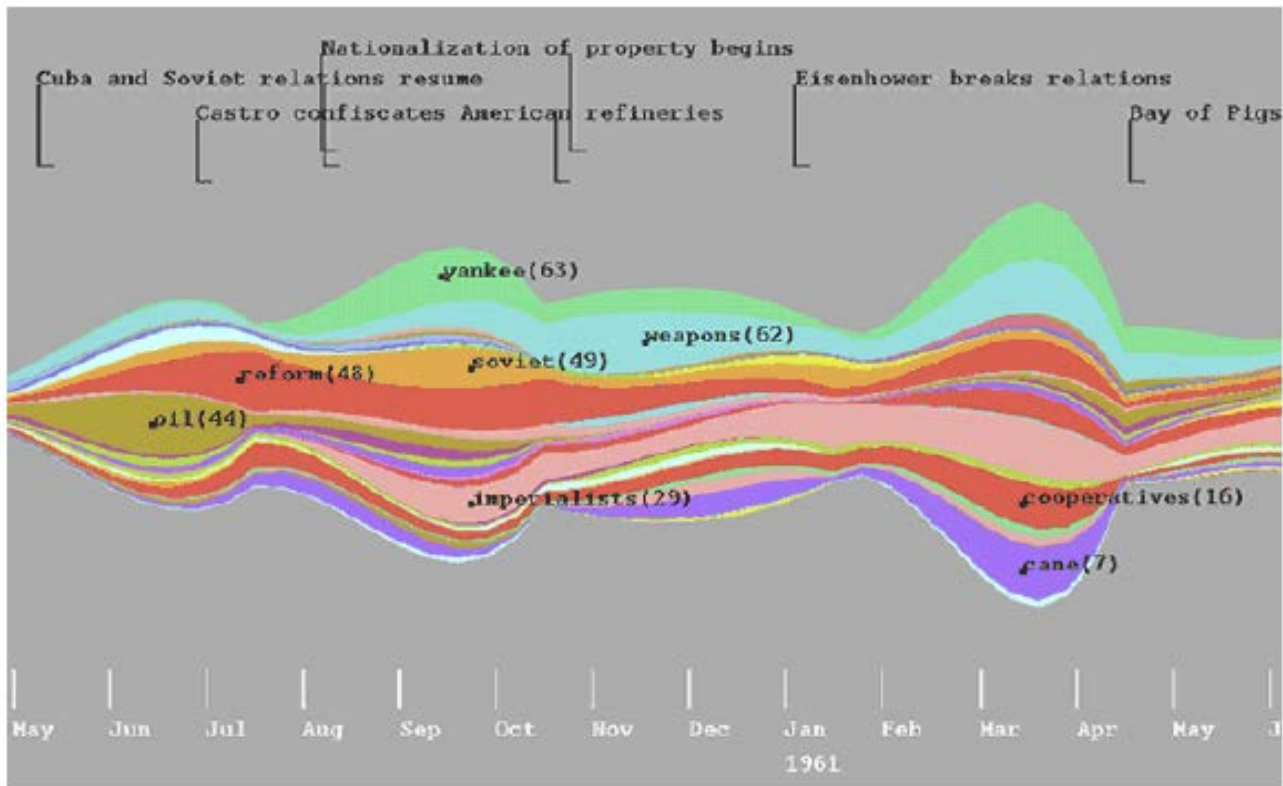
| | | |
|----------------|--------|---------|
| # of Web Pages | Small | 1 - 10 |
| | Medium | 11 - 99 |
| | Large | > 99 |

<http://www.archive-it.org/collections/2836>

Jigsaw



Themeriver



Wei et.al. in SIGKDD, 2010

Time Cloud



Bubble Chart

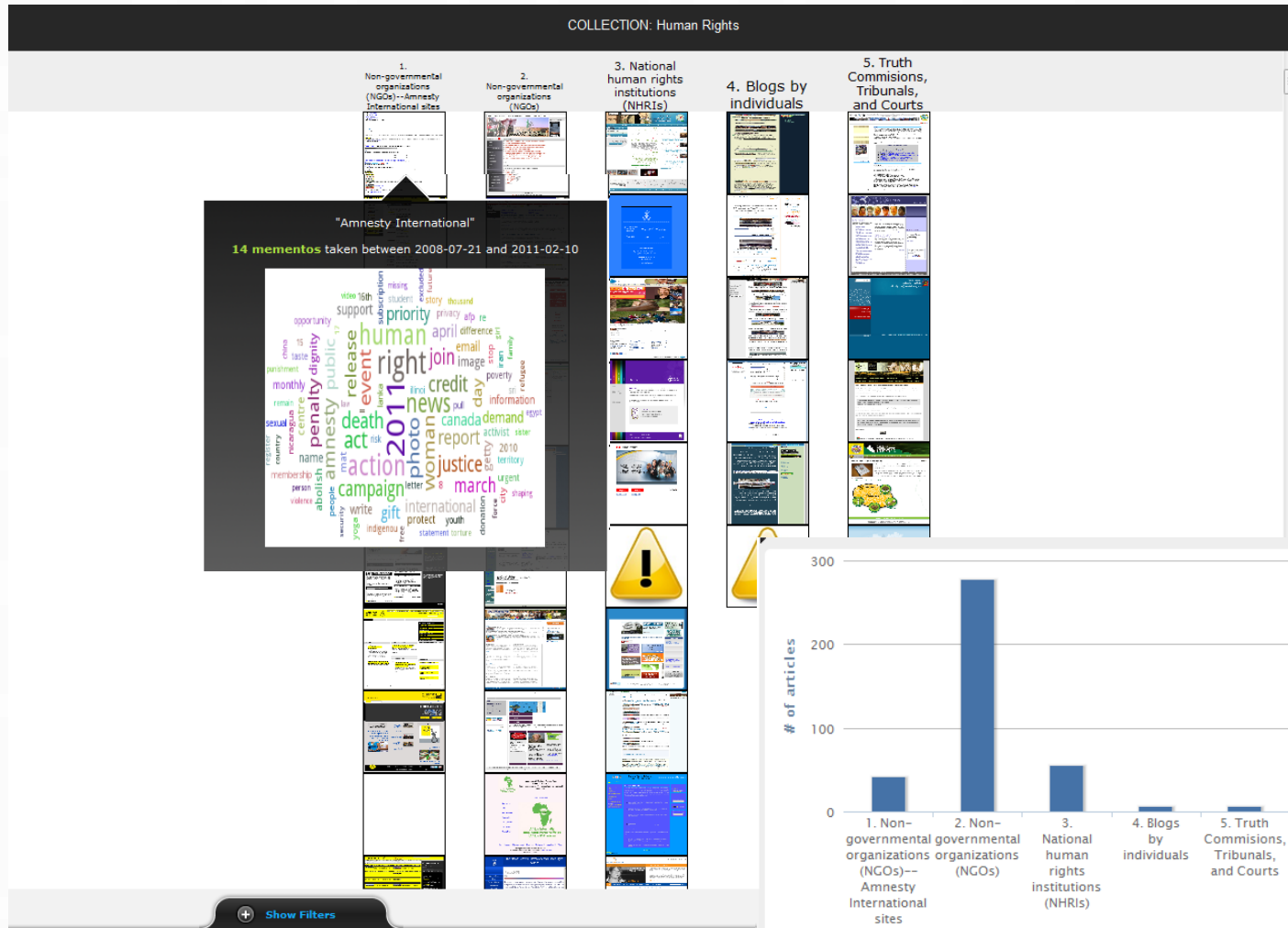
COLLECTION: Human Rights



Representing 365 articles in 5 curated categories
1729 mementos have been recorded between 15-May-2008 to 22-April-2011

+ Show Filters

Image Plot with Wordle



Timeline

