

Internet Archive Partner Meeting; Washington, D.C.; 4 Nov. 2009

Crisis, Tragedy and Recovery Network (CTRnet)

A Global Human Network and Distributed Digital Library

Venkat Srinivasan, Steve Sheetz, Edward A. Fox

Virginia Tech, Blacksburg, VA

{svenkat,sheetz,fox}@vt.edu



Funded by

NSF Award # 0916733



Crisis, Tragedy, and Recovery Network

<http://www.ctrnet.net>

CTRnet Team

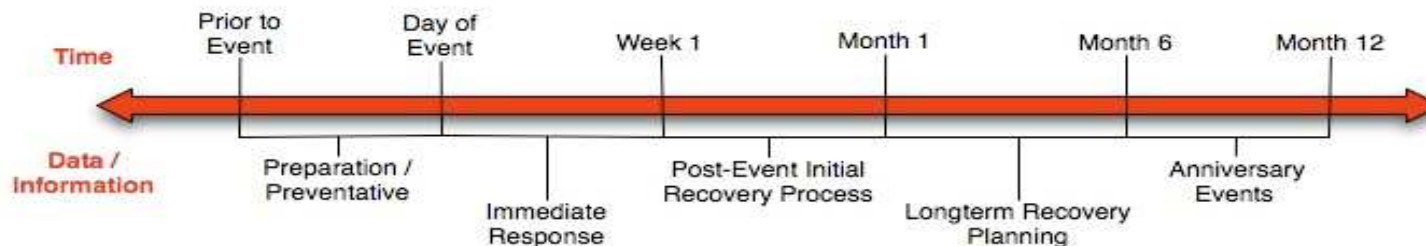
- ▶ **Faculty:** Ed Fox, Andrea Kavanaugh, Naren Ramakrishnan, Steve Sheetz, Don Shoemaker
- ▶ **Students:** Venkat Srinivasan, Bidisha Dewanjee, Amine Chigani
- ▶ **Collaborators:** Internet Archive, Monterrey Tech, Iowa State, and many more
- ▶ **Sponsors :** NSF via grant number IIS-0916733
- ▶ **Student team in Fox's Fall 2009 VT class CS5604:** Tram Bethea, Yipan Deng, Seth Fox, Min Li, and Chao Peng

Outline

- ▶ Introduction
- ▶ Vision
- ▶ Challenges
- ▶ Methods
- ▶ Experiments
- ▶ Results
- ▶ Outreach
- ▶ Conclusions

Introduction

- ▶ Human tragedies that result from natural and man-made events affect communities significantly.
- ▶ Series of information needs have to be addressed, during or after the tragic event.
 - ▶ During the event – there is a need to communicate, get accurate information, etc.
 - ▶ After the event – there is need to analyze, recover, prevent future instances, etc.



Introduction (Cont'd.)

- ▶ Not enough reliable information available always – in terms of quality and quantity

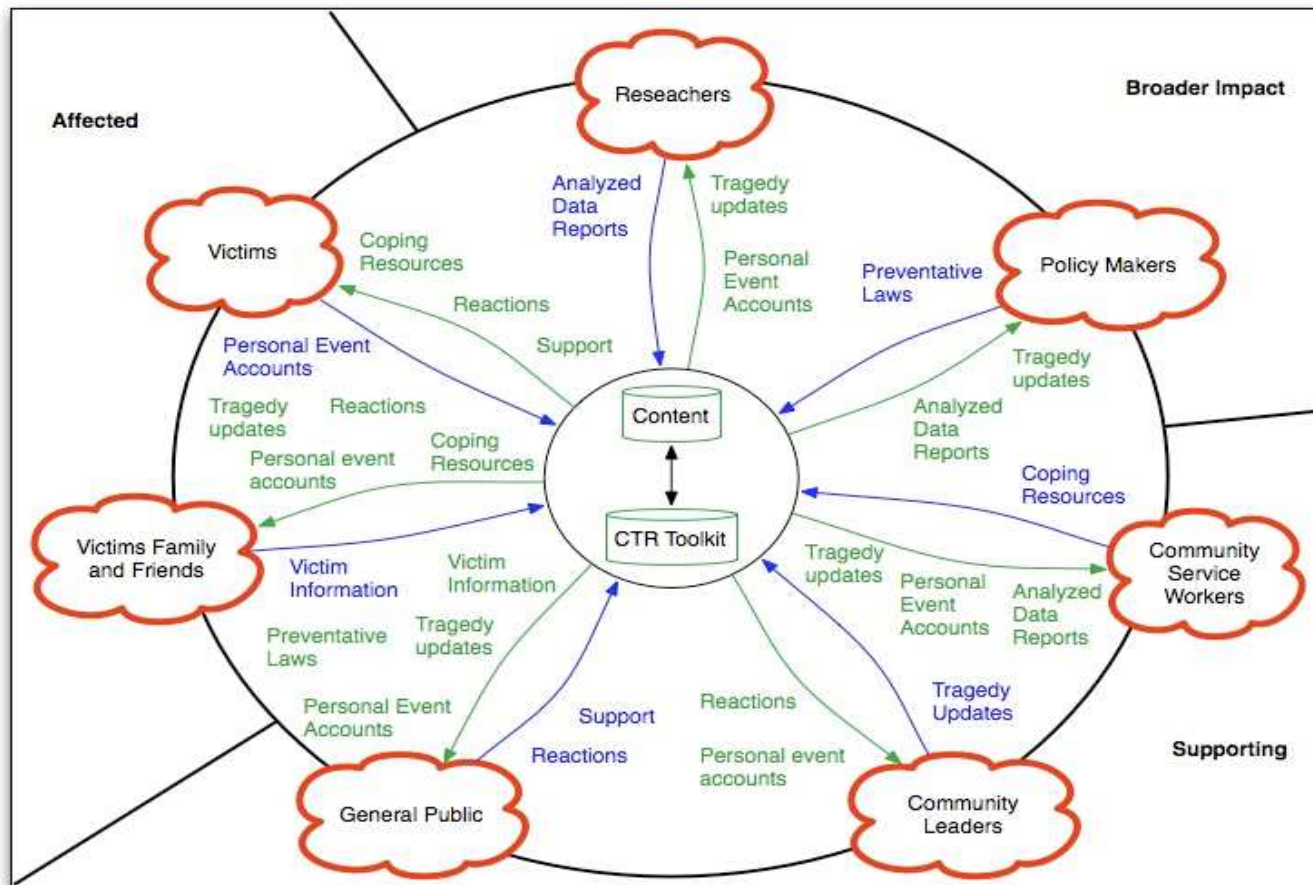
Year	Event	IA Collection #	Wikipedia suffix, other URLs
2004	Asian Tsunami	2004	Indian_Ocean_earthquake tsunami.archive.org
2007	Burmese Uprising	937	2007_Burmese_anti-government_protests
2007	California Wildfires	877	California_wildfires_of_October_2007
2008	Georgia and Russia Conflict	1120	2008_Georgia-Russia_crisis
2005	Hurricane Katrina	174	Hurricane_katrina, www.hurricanearchive.org, hellicane.blogspot.com, everythingandnothing.typepad.com katrinapoetry
2008	Iowa Flood	1092	Iowa_flood_of_2008
1998	Matthew Shepard murder	1075	Matthew_Shepard
2008	N. Illinois U. Shooting	970	Northern_Illinois_University_shooting
2008	Tibet protests	1044	Tibet_protests
2007	VT April 16 Shooting	694	Virginia_Tech_massacre, www.april16archive.org
2008	Zimbabwean crisis	1048	Zimbabwe, www.crisiszimbabwe.org

CTR related collections at
Internet Archive

Vision

- ▶ Supporting Crisis, Tragedy and Recovery (CTR) specific information needs
- ▶ Supporting diverse groups of stakeholders
 - ▶ Victims, family/friends, researchers, policy makers, general public
- ▶ Intelligent information integration
 - ▶ Not just archiving, but also analyzing information
- ▶ Rich *Digital Library* services
- ▶ Building a global collaborating community
- ▶ **Current focus is on school shootings** (but will be expanded later).

Vision (Contd.)



Challenges

- ▶ Many CTR events to crawl
- ▶ Sometimes too little data is available.
- ▶ A vast amount of data needs to be crawled.
 - ▶ Information integration is a key aspect of the project.
 - ▶ Must crawl different kinds of data
 - ▶ **“Scoping” crawls for each event independently is not practical.**
- ▶ Considerable amount of non-relevant data arrives in each category.
 - ▶ Makes data analysis very hard
- ▶ Likely to run out of budget very quickly
- ▶ **Some general principles, best practices needed**

Methods Outline

- ▶ Internet Archive's Archive-It tool to collect data
- ▶ Data in 6 major categories (where available)
 - ▶ News
 - ▶ Web 2.0 (blogs, social networks, forums, etc.)
 - ▶ Images
 - ▶ Videos
 - ▶ Academic publications (books, journal articles, reports, etc.)
 - ▶ Organizations (Red Cross, UN, etc.)

Experiments

- ▶ Crawl using a local Heritrix set-up
 - ▶ Try to get a lower bound on what to expect (worst case scenario)
 - ▶ See how much of data is likely to be relevant
 - ▶ Learn how to select seeds
 - ▶ Learn some general principles (using the above) that can be used for crawling CTR events
- ▶ Crawled for recent typhoon in Asia-Pacific (Ketsana)
- ▶ Also tried to learn using our April 16 shootings collection (at IA)

Results

- ▶ Some results from pilot experiments
- ▶ Very low precision
 - ▶ For April 16 data, we crawled ~4million URLs. Only ~100,000 of them contain the word “shootings”
 - ▶ Many crawled results are non-relevant.
 - ▶ For Ketsana, we crawled ~600,000 URLs. Only ~10,000 had the word “Ketsana” in them.
- ▶ **Crawl broad and not deep**
 - ▶ Crawl with more seeds
 - ▶ **Need a way to set crawl depth with Archive-It.**

CTRnet Outreach

CTRnet website (<http://www.ctrnet.net>)

The screenshot shows the CTRnet website interface. At the top, there is a navigation menu with buttons for HOME, BACKGROUND, RESOURCES, CONNECT, FAQs, and ABOUT. A search bar is located to the right of the menu. Below the menu, the page title is 'Home » Resources'. On the left side, there is a 'LOGIN' section with a key icon, a username field (containing 'svenkat'), a password field (masked with dots), and a 'Log in' button. Below the login form are links for 'Create new account' and 'Request new password'. To the right of the login form is a 'LATEST' section with a pencil icon, listing recent news items such as 'First CTRnet Local Advisory Board Meeting' and 'CTRnet Team Wins Award at OutreachNOW Conference'. The main content area is titled 'Resources' and shows a timestamp 'Thu, 09/03/2009 - 18:34 - svenkat'. It contains two main sections: 'CTR related digital collections' and 'Websites'. The 'CTR related digital collections' section lists seven items, including 'April 16 Collection at Internet Archive', 'April 16 Digital Library', 'April 16 Archive', 'Northern Illinois Shooting Digital Collection at Internet Archive', 'September 11 Digital Archive', 'Hurricane Digital Memory Bank', and 'California Wildfire Collection at Internet Archive'. The 'Websites' section lists five items, including 'Virginia Tech Campus Shooting in the Yahoo Directory', 'Archive for the "campus shooting" category from the Crime Report', 'Timeline of worldwide school shooting', 'Recent Worldwide school shooting Map', and 'List of Deadliest campus shootings in United States'.

- ▶ Contribute resources, seeds for events
- ▶ Browse collections
- ▶ Join forums
- ▶ Build network

Crisis, Tragedy, and Recovery Network



Conclusions

- ▶ Strong case for adding **focused crawling** feature to Archive-It
 - ▶ Determine on the fly if the webpage is relevant or not
 - ▶ If not, then do not follow links from there
 - ▶ Substantial savings in resources
 - ▶ 1 news story = 47 URLs (from our experiments)
- ▶ Especially useful when building multiple collections
- ▶ Also, the idea of crawling, then filtering to get rid of non-relevant documents, and using the result instead of the full crawl, might aid users.
- ▶ **Virginia Tech and some partners are interested in working on this for possible integration with Archive-It .**