

Archive-it Partner Meeting,
Annapolis, Maryland
December 3, 2012

Real-time Archiving of Spontaneous Events (Use-Case : Hurricane Sandy)

Kiran Chitturi,
Digital Library Research Laboratory,
Virginia Tech.

Background

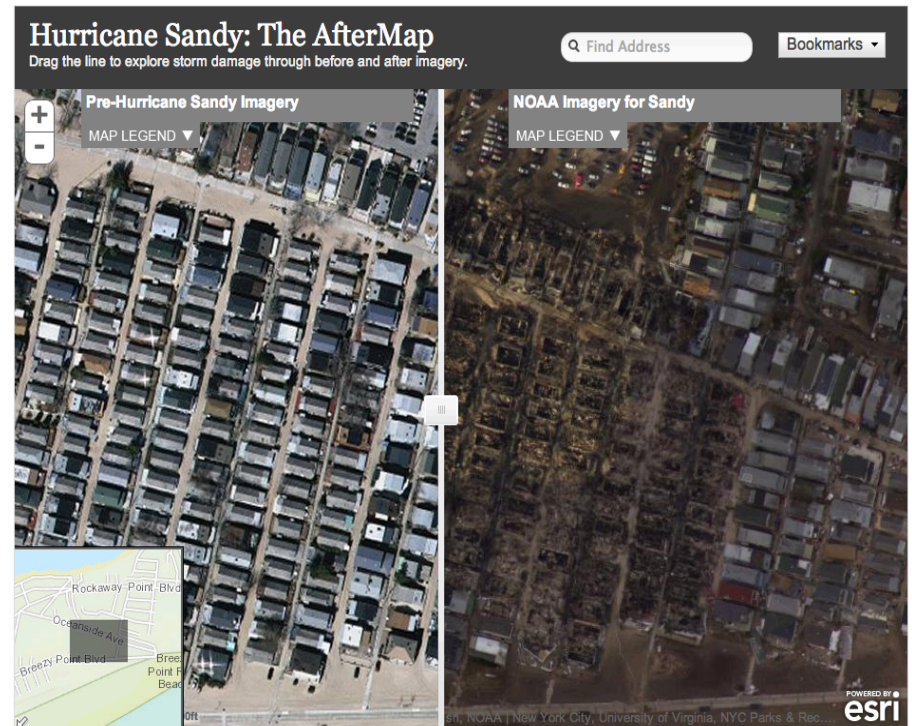
- Crisis, Tragedy and Recovery Network (CTRnet) [1]
 - Dr. Edward A. Fox (Principal Investigator)
 - 3+ year project started in August 2009 with grant from NSF
 - Researching the problems of integrating content, community, and services related to crises, tragedies, and recovery (CTR)
 - Building a domain specific distributed digital library, integrating information from many sources, with ontology support
- Partnership with Internet Archive for archiving data

Background

- Data collected using Archive-it
 - 8.8 TB + (43 archive collections)
- Collections
 - School Shootings
 - Earthquakes
 - Hurricanes
 - Floods
 - Plane Crashes

Hurricane Sandy [2]

- Devastating tropical cyclone
 - Formed - October 22, 2012
 - Dissipated - October 31, 2012
 - Damage - At least \$65.6 billion
 - Fatalities - 253 total
 - Spanning Seven Countries

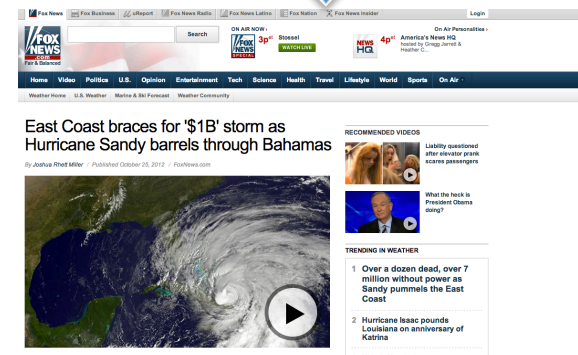


Archiving Hurricane Sandy - Tweet Collection

- Tweet collection based on hash tags and keywords
 - Example: #HurricaneSandy, #Sandy, Sandy prepare, Sandy Help, Sandy Shelter, Hurricane Sandy ...
- Our Sample
 - Collecting tweets since October 25, 2012
 - 8 million tweets
- Seed Extraction through tweets

@ [redacted] To friends and family up north, please take pre-caution and hope it all works out. This hurricane is shaping up to b... <http://t.co/HINSDbB4>

Thu, 25 Oct 2012 19:43:18 +0000 - tweet id 2.6155354316357E+17 - #7



The screenshot shows the Fox News website interface. The main headline reads "East Coast braces for '\$1B' storm as Hurricane Sandy barrels through Bahamas". Below the headline is a video player showing a satellite view of the hurricane. To the right of the video are sections for "RECOMMENDED VIDEOS" and "TRENDING IN WEATHER". The trending list includes: "1 Over a dozen dead, over 7 million without power as Sandy pummels the East Coast", "2 Hurricane lease pounds Louisiana on anniversary of Katrina", and "3 Miracle Monday!".

Archiving Hurricane Sandy - Seeds from Tweets

- Seeds – tweets statistics (Oct.25 – Nov.11)
 - 338,505 distinct seeds from 49,083 domains
 - Some top domains and seeds

Domain	Unique Seeds
Instagram.com	43098
Twitter.com	40863 (pictures)
Facebook.com	26176
youtube.com	12596
huffintonpost.com	4519
nytimes.com	3305

Archiving Hurricane Sandy - Seed Collection and Crawling

- Collecting seeds from public
 - Crawl type based on user choice
- Seeds extracted from tweets
 - Crawl type : single page
- Google news real-time coverage
 - Crawl type : News/RSS
 - Crawl Frequency : daily twice
- Popular news, business and weather portals
 - Crawl type : News/RSS
 - Crawl Frequency : daily once

Hurricane Sandy Archive-It Collection

This is a place to collect websites to archive as a part of a collection on Hurricane Sandy, in October/November 2012. Archived content will be available to the public at <http://archive-it.org/collections/3358>.

* Required

Please enter URL you would like to have archived

Let us know how much of this site you would like us to archive *

- The entire site
 Just one page

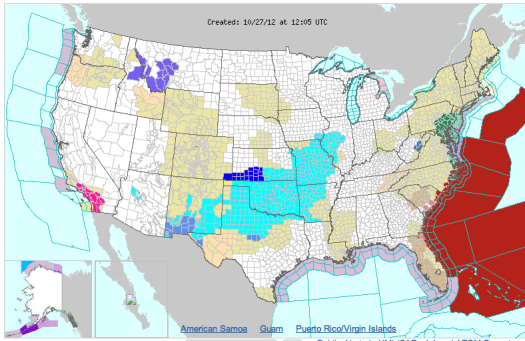
Powered by [Google Docs](#)

[Report Abuse](#) - [Terms of Service](#) - [Additional Terms](#)

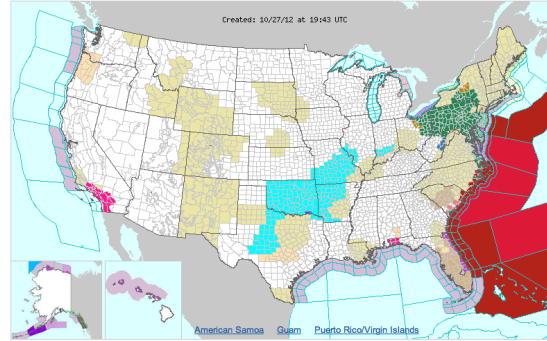
Archiving Hurricane Sandy - Seed Collection and Crawling (continued)

- Government, Emergency Organizations
 - Crawl type : whole website
 - Crawl Frequency : twice daily
- Facebook, twitter pages of local governments, and non-profit organizations
 - Crawl type : News/RSS
 - Crawl Frequency : daily once
- 83 Facebook pages about recovery of local communities (prepared by Steven Clift [3])
 - Crawl type : News/RSS
 - Crawl Frequency : weekly once

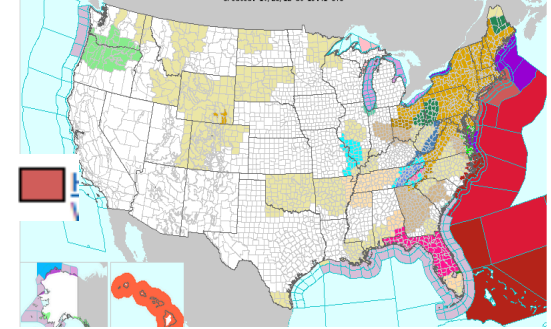
Archiving Hurricane Sandy – Weather maps from Wayback machine



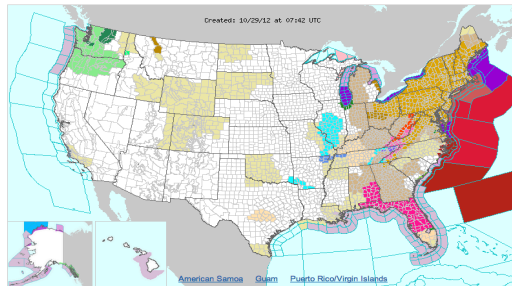
10/27/2012 – 12.05 UTC



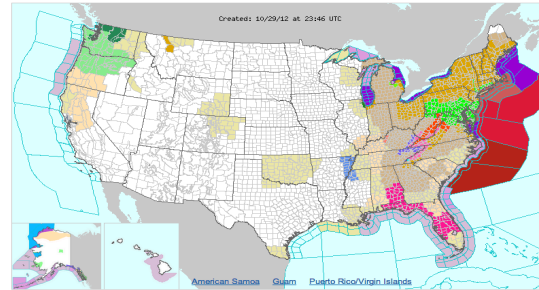
10/27/2012 – 19.43 UTC



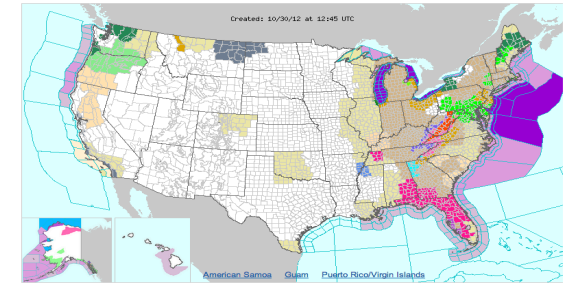
10/28/2012 – 19.41 UTC



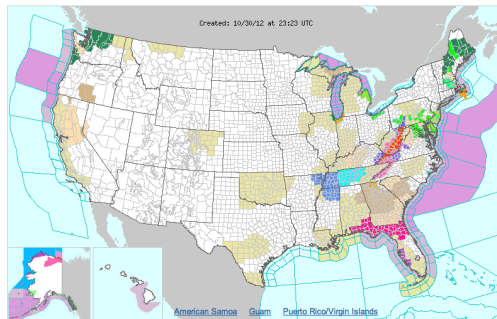
10/29/2012 – 07.42 UTC



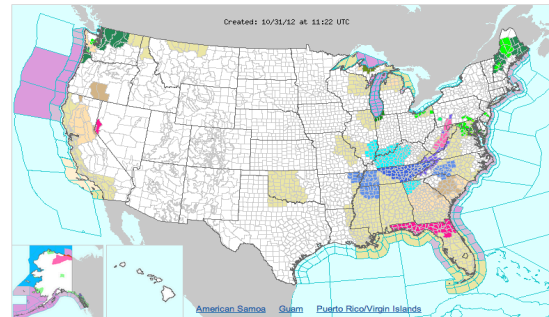
10/29/2012 – 23.46 UTC



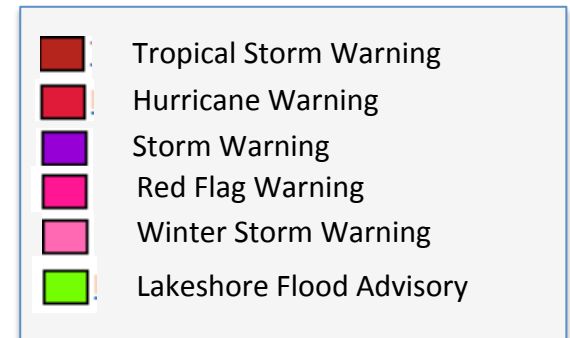
10/30/2012 – 12.45 UTC



10/30/2012 – 23.23 UTC



10/31/2012 – 11.22 UTC



Existing Problems : Real-time archiving

- Big collection sizes with frequent crawls
 - Distributed computing for crawling, filtering and analytics
- Seed collection after the event (recovery)
 - Crowd sourcing and identifying local sources
- Covering varied sources of information
 - Information Source graph (covering organizations and websites per geographic area)
- Disappearing webpages over time
 - Identifying and covering resources in time
- Spam
 - Classification and Filtering

Future Goals

- Use Lucid Works Big Data Software with the archive data
 - Integrated platform for solving information access problems with the tools for Search, Discovery, and Analysis of massive content sets
- Digital Library to provide access to the data
 - Features
 - Faceted Search
 - Image and Video extraction
 - Filtered, relevant documents per collection

Summary

- Real-time Archiving
 - Seed Collection
 - Big Data
 - Access

- Collaboration
 - Spotting Events
 - Access to data and software
 - IDEAL proposal

References and Resources

1. www.ctrnet.net
2. [www.en.wikipedia.org/wiki/Hurricane Sandy](http://www.en.wikipedia.org/wiki/Hurricane_Sandy)
3. <http://bitly.com/sandygroups>
4. <http://archive-it.org/collections/3358>
5. <http://archive-it.org/organizations/156>

Analyzing and Visualizing (Tweet) Archives

Presenter: Seungwon Yang

12/3/2012

Internet Archive Partners Meeting

Overview

- Description
 - Four Phase Model of Emergency Management
- Data Flow Diagram
 - Selection/cleaning → classification → visualization
- PhaseVis Components
 - What, When, Where, and Who
- A Sample Use Case

Tweet Analysis & Visualization Study

- Title:

PhaseVis: What, When, Where, and Who in Visualizing the Four Phases of Emergency Management Through the Lens of Social Media

- Conference Submitted:

The 10th International ISCRAM Conference'13

(Information Systems for Crisis Response and Management)

- Prototype Link:

- <http://spare05.dlib.vt.edu/~ctrvis/phasevis/>

Four Phases of Emergency Management

Phases	Description
Mitigation	Activities providing a critical foundation in the effort to reduce or eliminate the risks to persons or property, and to lessen or avoid the impact of a disaster
Preparedness	Range of deliberate, critical tasks and activities necessary to build, sustain, and improve the operational capability to prevent , protect against, respond to, and recover from emergency and disasters.
Response	Immediate and ongoing actions following an emergency or disaster , which are designed to provide emergency assistance for casualties such as search/rescue, emergency shelter, medical care, mass feeding.
Recovery	The development, coordination, and execution of service and site restoration plans. The reconstruction of government operations and services.

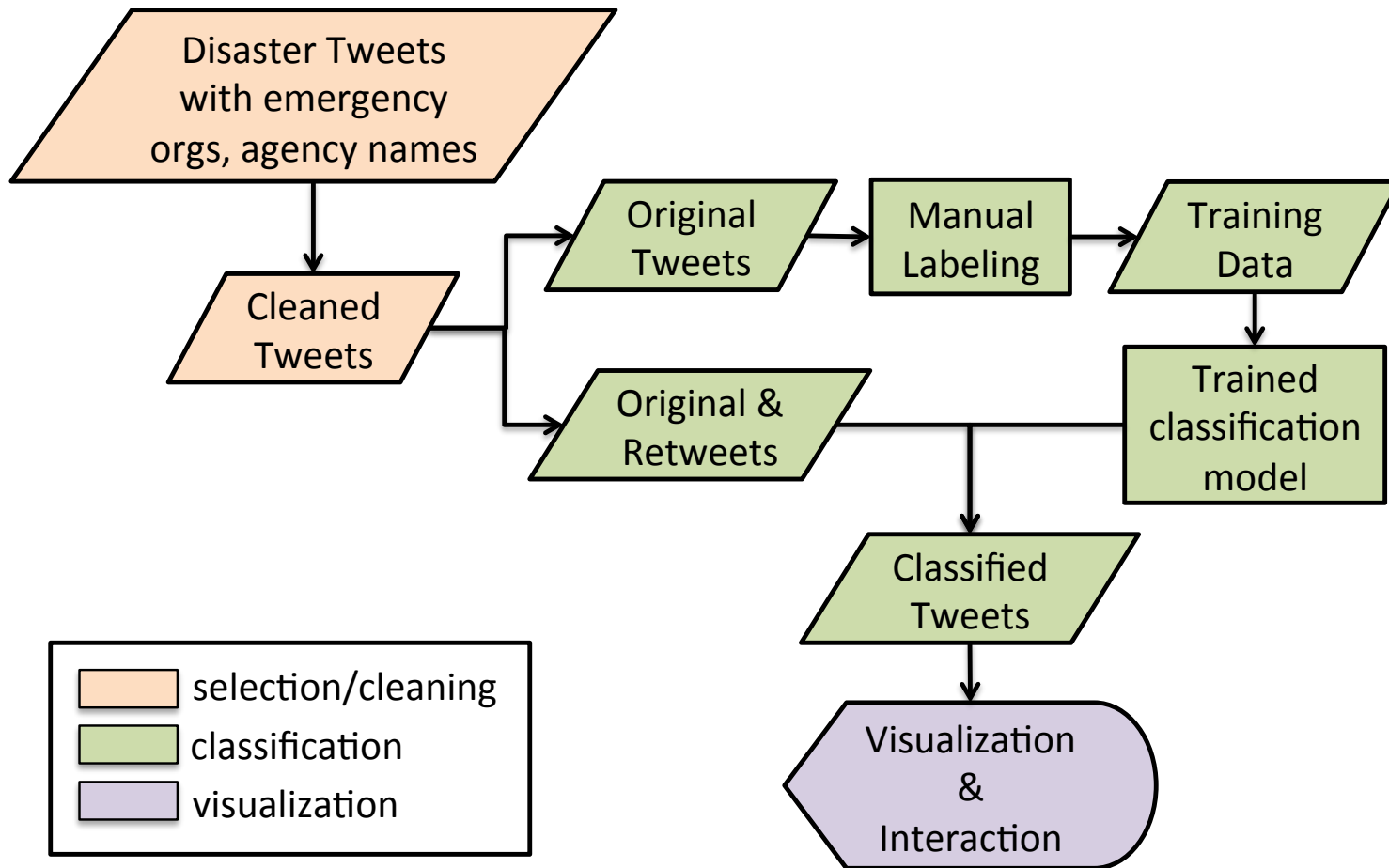
Four Phases of Emergency Management

- Tweet Examples -

Phase (Category)	Example Tweet
1: Response	Red Cross provided 20,000+ overnight shelter stays and served 400,000+ meals/snacks to help those impacted by #Hurricane #Isaac. THANK YOU!!
2: Recovery	Hurricane #Isaac relief efforts in the Gulf continue. Pls help! Txt REDCROSS to 90999, to give \$10.
3: Mitigation	FEMA mitigation to offer rebuilding tips in St. John Parish at Home Depot in LaPlace, La. http://t.co/8UFJFEz5 #Isaac
4: Preparedness	RT @jeffiel: Seriously, everybody should have this #RedCross app on their phone. What to do in #emergencies http://t.co/MtrQcOo1 #isaac

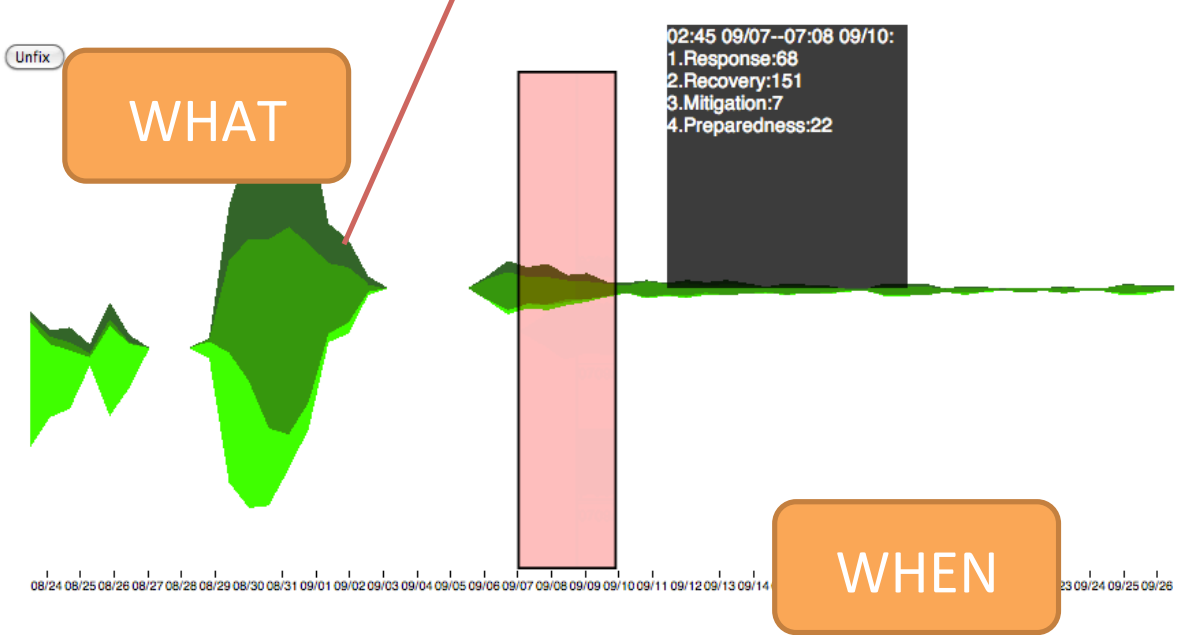
Data Flow Diagram

- Selection, classification, & visualization -

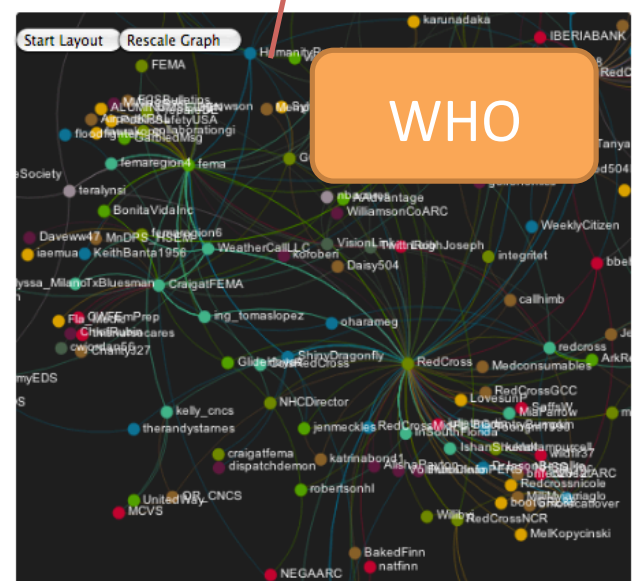


What, When, Where, and Who in PhaseVis

(a) Phase View



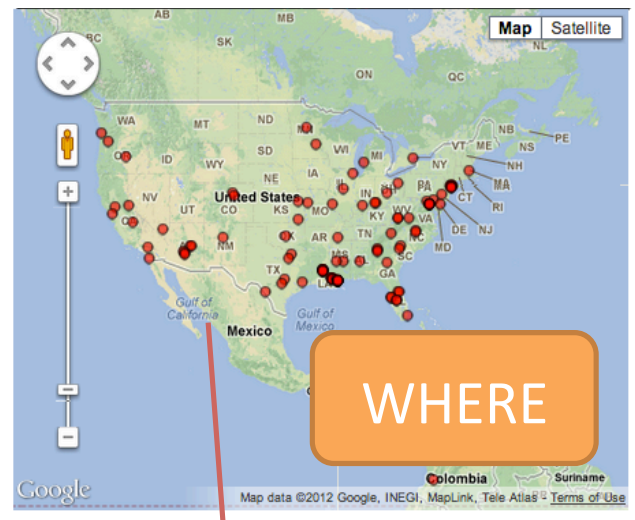
(c) Social Network View



Tweets in Selected Region

No.	Is_R	Text	Class	Date	Timestamp
1323061		RT @RedCrossSELA: Please RT! We NEED #volunteers! Help us distribute supplies in communities hit hard by #Is	1	Sat, 08 Sep 2	1347144
1323260		How many of Ya'll applied to FREEMA for #Isaac? FREEMA AIN'T GIVIN UP SHIT! No Mo Free from FEMA. They se	1	Sun, 09 Sep 2	1347149
1324240		Can you give back to #RedCross volunteers? Contact me if you can offer discounted meals or attractions as a tha	1	Sun, 09 Sep 2	1347163
1324661		RT @RedCrossSELA: Please RT! We NEED #volunteers! Help us distribute supplies in communities hit hard by #Is	1	Sun, 09 Sep 2	1347170
1324871		RT @RedCrossSELA: Please RT! We NEED #volunteers! Help us distribute supplies in communities hit hard by #Is	1	Sun, 09 Sep 2	1347172
1327361		RT @RedCrossSELA: Please RT! We NEED #volunteers! Help us distribute supplies in communities hit hard by #Is	1	Sun, 09 Sep 2	1347205
1326741		RT @jenmeckles: Just saw about a dozen @RedCross relief trucks in southern Mississippi heading to the Gulf #Is	1	Sun, 09 Sep 2	1347215
1325860		#Isaac has left more than #flood waters & destruction behind. #RedCross relief efforts continue in the Gulf Coast	1	Sun, 09 Sep 2	1347226
1325871		#Isaac has left more than #flood waters & destruction behind. #RedCross relief efforts contin	1	Sun, 09 Sep 2	1347226
1325061		#Isaac has left more than #flood waters & destruction behind. #RedCross relief efforts contin	1	Mon, 10 Sep 2	1347242
1330541		#Isaac has left more than #flood waters & destruction behind. #RedCross relief efforts contin	1	Mon, 10 Sep 2	1347244
1330661		#Isaac has left more than #flood waters & destruction behind. #RedCross relief efforts contin	1	Mon, 10 Sep 2	1347245
1330751		#Isaac has left more than 5,200 people stayed in #RedCross shelters. Support them by textin	1	Mon, 10 Sep 2	1347247

(b) Tweet View



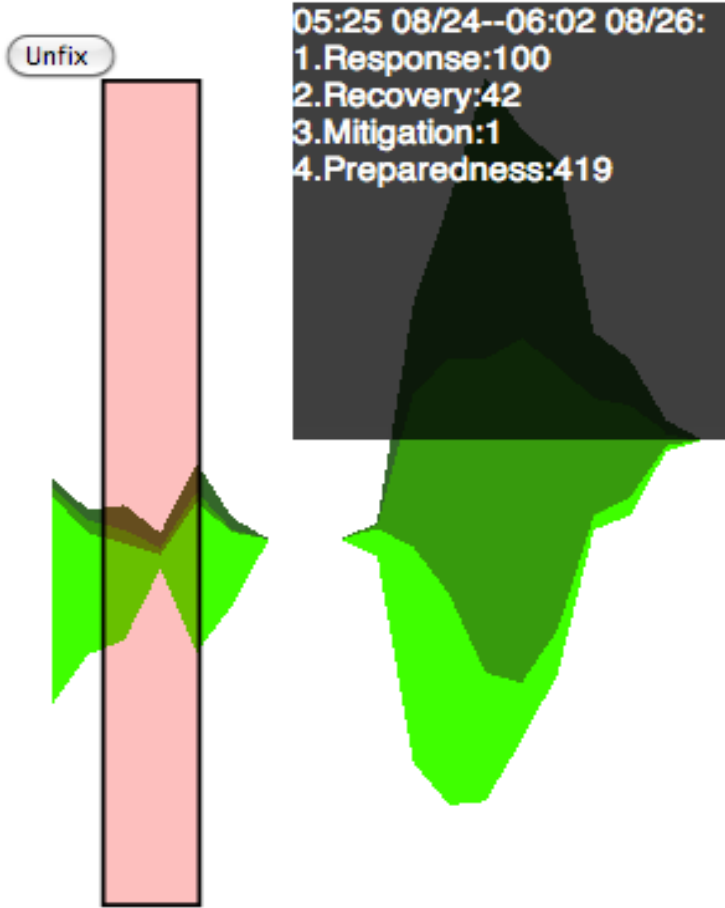
(d) Map View

A Sample Use Case (1/2)

(A)

8/24 - 8/26

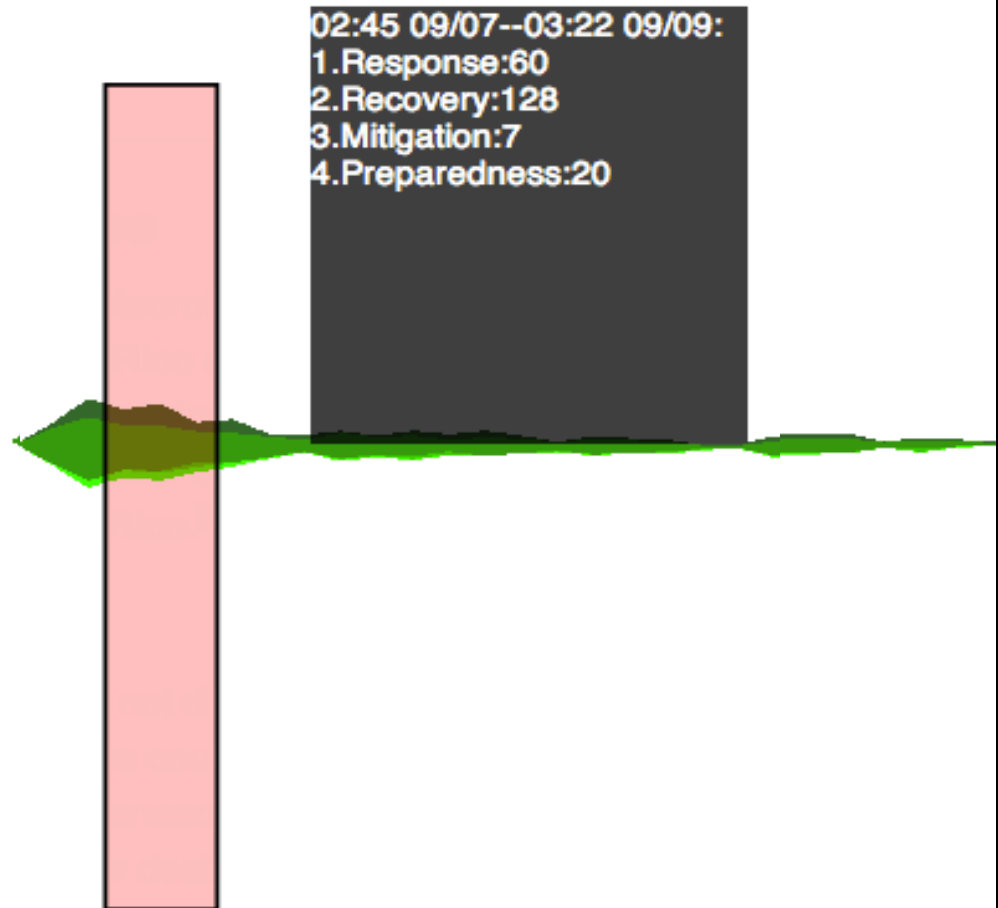
Preparedness is dominant



(B)

9/7 - 9/9

Recovery is dominant

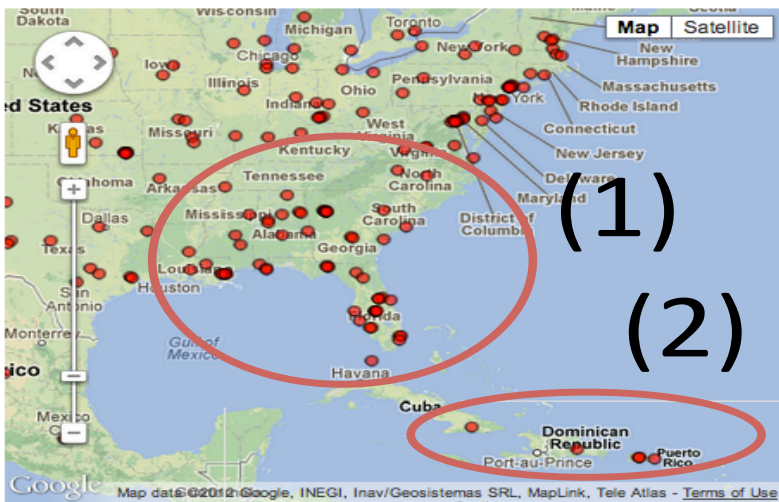
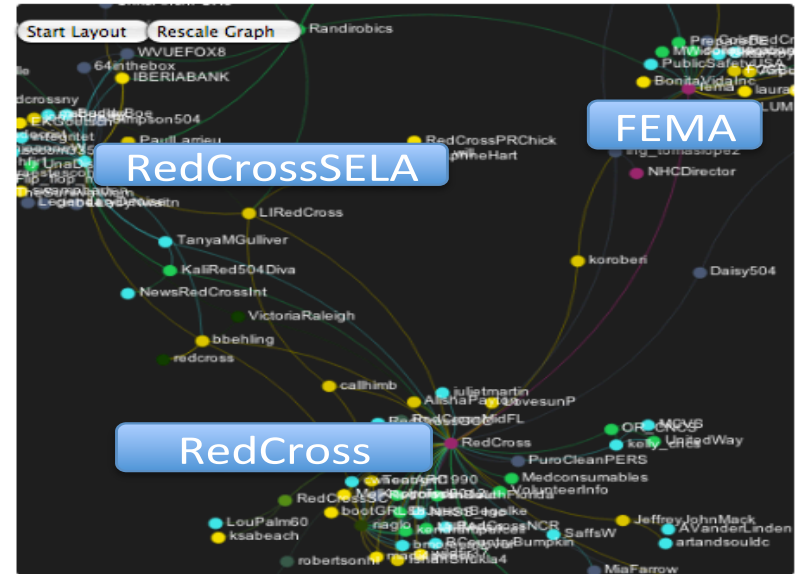
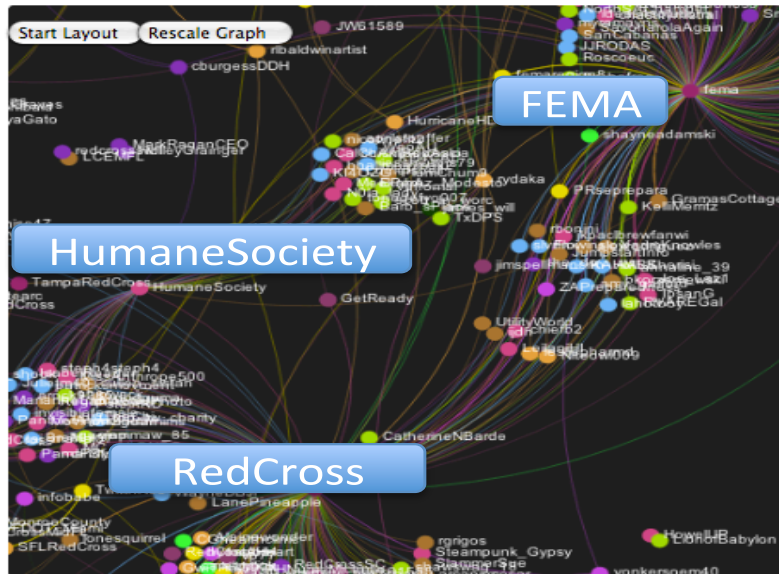


08/24 08/25 08/26 08/27 08/28 08/29 08/30 08/31 09/01 09/02 09/03 09/04 09/05 09/06 09/07 09/08 09/09 09/10 09/11 09/12 09/13 09/14 09/15 09/16 09/17 09/18 09/19 09/20 09/21

A Sample Use Case (2/2)

(A)

(B)



Acknowledgments

- Members of the **Digital Library Research Laboratory**, Virginia Tech
- We thank **NSF** for supporting the CTRnet project (IIS-0916733), of which this work is part. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Thank you!
Questions?