NSF Project Reporting Format

This document has been developed to provide Principal Investigators (PIs), co-PIs, and research organizations with:

- a listing of the questions that will be asked in the new NSF project reporting format;
- assistance in planning for the submission of the report; and
- a tool to help PIs collaborate with other contributors in answering these questions, if needed.

The project reporting service on Research.gov and the associated help documentation provides more detailed instructions and contextual assistance.

Note: NSF project reports are not cumulative and should always be prepared for the specific project reporting period only.

PIs should NOT complete and upload this document to Research.gov in order to meet your reporting requirement. You are required to enter text in the text boxes available

Accomplishments

You have the option of selecting "nothing to report" in this section.

What are the major goals of the project?

The project team has been developing a digital library including many webpage archives and tweet archives related to disasters, in collaboration with the Internet Archive. The goals of the CTRnet project are to provide such archived data sets for analysis, including by researchers who are seeking deep insights about those events, and to support a range of services and infrastructure regarding those tragic events for the various stakeholders and the general public, allowing them to study and learn.

What was accomplished under these goals (you must provide information for at least one of the 4 categories below)?

Major Activities:

We have been taking steps toward these goals by archiving disaster information from both mainstream media webpages and social media (e.g., tweets) in collaboration with the Internet Archive. We also have been developing an analysis methodology and visualizations for archived social media data.

For large webpage collection data, we applied a Big Data processing tool called LucidWorks in collaboration with a company, LucidWorks, formerly called Lucid Imagination, Inc. The detail workflow of processing large data sets using this tool became a thesis topic of a graduate student, Kiran Chitturi, who has been working as a CTRnet team member. He worked to make accessible the data related to the Boston Marathon Bombing. His thesis is planned to be published Fall 2013.

Regarding our tweet data analysis, we had two approaches. We conducted studies applying sociology theories such as information diffusion and finding opinion leaders. Several survey studies for the use of information communication technologies were conducted in collaboration with our partners in Egypt and Tunisia as well.

The other approach was to apply text analysis approaches such as classification, data extraction, and visualization. To identify and visualize the four phases of emergency management possibly appearing in disaster tweets, we adopted machine-learning classification methods to categorize Hurricane Isaac tweets. The volume of tweets in each category was visualized, along with other information such as user locations and a mention network.

Specific Objectives:

One objective has been to archive all of the important disaster events occurring around the world. Generally we do this first using tweets, and then for webpages, adapting the Internet Archive tools to help.

Another objective has been to use LucidWorks to take the data that has been collected at the Internet Archive, copy that to our digital library at Virginia Tech, and research the best ways to make it accessible.

Significant Results:

We have helped expand awareness of and interest in what is coming to be called computing for disasters, or disaster informatics.

We have shown how to integrate the collection of tweets, the crawling of webpages, and the combination of digital library and archiving methods to address public and scholarly interests about important events.

We have amassed and preserved a very large collection of content about many of the most important natural or manmade disasters, especially since 2008.

Key outcomes or other achievements:

During the year, we published journal articles, papers in conference proceedings, and posters, as well as organized and presented our efforts in workshops and webinars. The publication details can be found in the 'Products' section.

In addition, we developed a prototype tweet analysis and visualization tool called PhaseVis, based on our analysis of our Hurricane Isaac tweet archive. The details of this study and tool were presented in the ISCRAM'13 conference. The prototype can be accessed at http://spare05.dlib.vt.edu/~ctrvis/phasevis/index_may.html.

As part of this project and for a dissertation study of a GRA, Seungwon Yang, a prototype of a topic identification system called Xpantrac has been built. Its user interface is being implemented, and will be available on the CTRnet website in this fall.

Over the years we have been working on tweet data processing such as extraction of URLs, hashtags, mentions, or cleaning of the tweet texts. In the process, we developed several utility scripts written in Python and its packages. We also developed scripts to extract tweet users' locations using their profiles. To apply the Latent Dirichlet Allocation (LDA) algorithm in our experiment, we used Python and its package called Gensim. All of these scripts can be downloaded from a public Github repository at: https://github.com/seungwonyang/public.

What opportunities for training and professional development has the project provided?

Our prototype tweet analysis and visualization tool, PhaseVis, started as a class project, where 3 graduate students from an Information Retrieval class, and 3 other graduate students, including Seungwon Yang (serving as GRA), collaborated together under the guidance of the CTRnet team. This effort was expanded to lead to a conference paper and related prototype development. Through this study, participating graduate students could experience every step of the process from Twitter data preprocessing, training set building for classification, understanding of three classification algorithms and tools, making decisions on which visualizations are appropriate, implementation of multiple visualizations, and collaboratively writing a conference paper. These experiences will be very helpful for the students involved, who may want to do any text analysis tasks in the future.

The project team had two undergraduate students collaborating with us. Jennifer Murphy, who had a sociology background, worked on detailed analysis of the webpages, which were embedded in disaster tweet messages collected under the supervision of the research assistant in the CTRnet team. Her study results were presented in a local peace promotion conference. The other student, Christopher Jones, has been learning about the workflow of web archiving as part of his research during summer and fall of 2013 in Virginia Tech's Scieneering program. For example, we had one-to-one teaching/learning regarding the details of tweet archive development, extraction of seed URLs from such tweets, and collecting webpages using the Internet Archive's Heritrix crawler. He plans to present his experience at the end of September 2013 for a meeting of the Scieneering program.

The project team collaborated with LucidWorks, on its Big Data software, also called LucidWorks. To help people learn about this sophisticated tool, which integrated several Apache projects such as Hadoop, HBase, Solr, Lucene, Mahout, Zookeeper, Pig, etc., several webinars were held online, and project members attended those webinars to assist.

In addition, in CS5604, Information Retrieval, in fall 2012, students prepared 6 educational modules to help those interested in learning about big data in general and LucidWorks in particular. See the list, with links to the resulting modules, in Table 3 on Wikiversity page http://en.wikiversity.org/wiki/Curriculum_on_Digital_Libraries. This further extends our prior work for expanding curricular resources related to digital libraries and other advanced information systems.

How have the results been disseminated to communities of interest?

Our tweet analysis and visualization study regarding PhaseVis tool was presented in the ISCRAM 2013 conference, which was held in Baden-Baden, Germany. Both researchers and the field practitioners in disaster area from all over the Europe and the US showed interest and provided feedback for our study during our presentation and the subsequent demonstration session. The presentation slides were posted to a public presentation-sharing site, SlideShare, for further dissemination to interested community members.

Other means of dissemination included publishing journal articles (see reference information under the 'Products' section), attending the Archive-It Partners Meeting, and organizing workshops for disaster informatics. Kiran Chitturi and Seungwon Yang attended the Partners Meeting, held by the Internet Archive in December 2012, to present our work regarding the use of Big Data analysis software for archived disaster webpages as well as tweet analysis and visualization studies. The PI of the CTRnet project, Edward A. Fox, organized a workshop 'Web Archiving and Digital Libraries (WADL'13)' following the JCDL 2013 conference. We had an opportunity to present our disaster webpage archiving and tweet studies.

What do you plan to do during the next reporting period to accomplish the goals?

This is the final report for the CTRnet project. However, Virginia Tech has just received an NSF grant IIS-1319578, "III: Small: Integrated Digital Event Archiving and Library (IDEAL)", that will build upon our CTRnet efforts.

NOTE: You may upload PDF files with images, tables, charts, or other graphics in support of the Accomplishments section. You may upload up to 4 PDF files with a maximum file size of 5 MB each.

Products

You have the option of selecting "nothing to report" in this section. There are no limitations to the number of entries you submit and you can also pull information directly from Thomson Search when using the online tool on Research.gov.

Within the Products section, you can list any products resulting from your project during the specified reporting period, such as:

Journals:

- Kavanaugh, A.L., Sheetz, S.D., Hassan, R., Yang, S., Elmongui, H.G., Fox, E.A., Magdy, M., and Shoemaker, D. J. Between a Rock and a Cell Phone: Communication and Information Use during the Egyptian Uprising. *International Journal of Information Systems for Crisis Response and Management, 5*(1): 1-21.

Books:

- Rao Shen, Marcos Andre Goncalves, and Edward A. Fox. Key Issues Regarding Digital Libraries: Evaluation and Integration. Morgan & Claypool Publishers, San Francisco, Feb. 2013, 110 pages, ISBN paperback 9781608459124, ebook 9781608459131, DOI 10.2200/S00474ED1V01Y201301ICR026, http://www.morganclaypool.com/doi/abs/10.2200/S00474ED1V01Y201301ICR026

Book Chapters:

- Edward A. Fox, Monika Akbar, Sherif Hanie El Meligy Abdelhamid, Noha Ibrahim Elsherbiny, Mohamed Magdy Gharib Farag, Fang Jin, Jonathan P. Leidig, Sai Tulasi Neppali. Digital Libraries. In Computing Handbook Vol. 2 (Information Systems and Information Technology), Section 3, Ch. 18, ed. by Heikki Topi, Chapman & Hall/CRC Press, Taylor and Francis Group, in press for 2014, http://www.crcpress.com/product/isbn/9781439898444

Thesis/Dissertations:

- Thesis by Kiran Chitturi (expected Fall 2013):

  - Use of Big Data software tool for processing WARC files

- Dissertation by Seungwon Yang (expected October 2013)
  - Finding topic tags using expansion-extraction approach

Conference Papers and Presentations:

- Edward A. Fox, Seungwon Yang, and the CTRnet team (2013). Crisis Tragedy, and Recovery Network Digital Library (CTRnet) + Web Archiving in Qatar and VT. Workshop presentation at Web Archiving and Digital Libraries (WADL 2013), a JCDL workshop, July 25-26, 2013, Indianapolis, Indiana, USA

- Lin Tzy Li, Otavio A.B. Penatti, Edward A. Fox, and Ricardo da S. Torres. 2013. Domain-specific image geocoding: a case study on Virginia Tech building photos. In Proceedings of the 13th ACM/IEEE-CS joint conference on digital libraries (JCDL '13). ACM, New York, NY, USA, 363-366. DOI=10.1145/2467696.2467727 http://doi.acm.org/10.1145/2467696.2467727

- Kiran Chitturi, Edward A. Fox and CTRnet team. (2013). Virtual Workshop on Big Data and Emergency Informatics. Jan. 10, 2013.

- Yang, S., Chung, H., Lin, X., Lee, S., Chen, L., Wood, A., Kavanaugh, A. L., Sheetz, S. D., Shoemaker, D. J., and Fox, E. A. (2013). PhaseVis: What, When, Where, and Who in Visualizing the Four Phases of Emergency Management through the Lens of Social Media. *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2013)*, Baden-Baden, Germany, May 2013.

- Kiran Chitturi and Seungwon Yang. (2012). Real-time Archiving of Spontaneous Events (use case: Hurricane Sandy). Archive-It Partners Meeting, Dec. 3, 2012. Annapolis, MD.

- Steven D. Sheetz and Edward A. Fox (2012). Webinar on Emergency Informatics and Digital Libraries. July 24, 2012.

Other Publications:

- Jennifer Murphy, Donald J. Shoemaker, Seungwon Yang, Andrea Kavanaugh, Steven D. Sheetz, and Edward A. Fox. 2012. Examining the Content of Online Resources Embedded in Tweets: Hurricane Isaac (8/23-9/26/2012). Cultivating Peace: A Student Research Symposium for Violence Prevention. Poster, Nov. 16-18, 2012. Blacksburg, VA. USA.

Technologies or Techniques:

Advances in working with tweets, web archiving crawlers, digital libraries, and big data systems.

Patents:

Inventions:

Licenses:

Websites:

http://www.ctrnet.net (with many additions since prior years, including the final report just added from March 2011 of a related project: http://www.ctrnet.net/sites/default/files/CCSR White Paper Report VT IBM Kavanaugh Natsev_0.pdf)

Other Products:

- Prototype visualization tool, PhaseVis:
http://spare05.dlib.vt.edu/~ctrvis/phasevis/index_may.html

- Tweet processing scripts at Github:
https://github.com/seungwonyang/public

- Webpage Archives
  - Blasts in Boston Marathon
  - Boko Haram Attack
  - Global Emergency Overview Site
  - Texas Fertilizer Plant Explosion
  - Atlanta School Shooting 2013

- Tweet Archives (since June/July 2013, with number of tweets shown)
  - Iran election: 71,609
  - midwest storm: 1,122
  - NSA leak: 72,766
  - #rouhani: 18,489
  - flood: 975,971
  - earthquake: 692,905
  - tornado: 656,878
  - hurricane: 806,983
  - train explosion: 13,152
  - pollution: 113,301
  - diabetes: 649,239
  - obesity: 247,071
  - cancer: 2,809,407
  - autism: 136,076
  - suicide: 911,262
  - mental illness: 67,895
  - bullying: 555,612
  - #R4BIA: 150,390
  - #EgyAntiCoup: 8,231
  - #Rabaa: 10,560
  - #AntiCoup: 34,845
  - Georgia school shooting: 11,469
  - Antoinette Tuff: 35,119
  - Fukushima: 68,574
  - Nuclear plant: 12,300
  -

NOTE: You may upload PDF files with images, tables, charts, or other graphics in support of the Products section. You may upload up to 4 PDF files with a maximum file size of 5 MB each.

Participants

There are no limits on the number of participants you list for this section; however, you must list participants who have worked one person month or more for the project reporting period. You have the option of selecting "nothing to report" in this section. For Research Experience for Undergraduates (REU) sites and supplements, specific questions will be listed in this section. The online service will also ask for additional information on participants such as:

- What individuals have worked on the project?
- What organizations have been involved as partners?
- Have other collaborators or contacts been involved?

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Edward A. Fox	Principal Investigator (PI)	
Donald J. Shoemaker, Steven D. Sheetz, Andrea L. Kavanaugh, Naren Ramakrishnan	Co-PI	
Venkat Srinivasan, Seungwon Yang	GRA	
Tram Bethea, Yinlin Chen, Amine Chigani, Kiran Chitturi, Bidisha Dewanjee, Yipan Deng, Noha ElSherbiny, Seth Fox, S. M. Shamimul Hasan, Nadia P. Kozievitch, Sunshin Lee, Lin Tzy Li, Min Li, Mohamed Magdy, Chao Peng, Chet Rosson, Travis F. Whalen	Graduate collaborators	
Bernadel Benoit, Jason Browning, Mario Calixte, Sherley Codio, Rachel Coston, Jennifer Francois, Jason Heim, Robert Leith, Fabrice Marcelin, Ashley Phelps, Jason Smith, Justin Tillar, Keith Wooldridge, Jennifer Murphy, Christopher Jones, Tanmana Sarkar	Undergraduate collaborators	
Paul Mather (Library), John Tedesco (Communications)	Other collaborators	

What other organizations have been involved as partners?

The online service will also ask you for additional information such as:

- Type of Partner Organization: NGO
- Name: Internet Archive (IA)

- Location: San Francisco, CA, USA
- Partner's contribution to the project: The project team is using IA's Archive-It service, specifically the Heritrix crawler and the Wayback machine, for webpage archiving tasks.

- Type of Partner Organization: Research Lab of a company
- Name: IBM Watson Research Laboratory
- Location: NY, USA
- Partner's contribution to the project: collaborated with two researchers at IBM Watson for social media studies (e.g., Arlington study) – see http://www.ctrnet.net/sites/default/files/CCSR White Paper Report VT IBM Kavanaugh Natsev_0.pdf

- Type of Partner Organization: Government organization
- Name: Arlington County
- Location: Arlington, VA, USA
- Partner's contribution to the project: collaborated with the CTRnet team in social media study for government officials and civic organizations in Arlington area) – see http://www.ctrnet.net/sites/default/files/CCSR White Paper Report VT IBM Kavanaugh Natsev_0.pdf

- Type of Partner Organization: University
- Name: Arab Academy of Science and Technology
- Location: Cairo, Egypt
- Partner's contribution to the project: collaborated on survey research on social media and information and communication technology use in Egypt during and since the political crisis of the revolution in Egypt.

- Type of Partner Organization: University
- Name: Texas A&M University
- Location: TX, USA
- Partner's contribution to the project: collaborated with regard to helping develop broad interest in Computing for Disasters

- Type of Partner Organization: University
- Name: State Autonomous University of Mexico, Toluca
- Location: Toluca, Mexico
- Partner's contribution to the project: collaborated on survey research on social media and information and communication technology use in Mexico during and since the political turmoil surrounding Presidential and Parliamentary elections in Mexico in July 2012.

- Type of Partner Organization: University
- Name: High Institute of Management of Tunis

- Location: Tunis, Tunisia
- Partner's contribution to the project: collaborated on survey research on social media and information/communication technology use in Tunisia during and since the political crisis of the revolution.

- Type of Partner Organization: University
- Name: University of Alexandria
- Location: Alexandria, Egypt
- Partner's contribution to the project: collaborated on survey research on social media and information and communication technology use in Egypt during and since the political crisis surrounding the revolution in Egypt.

- Type of Partner Organization: Software company
- Name: LucidWorks
- Location: San Francisco, CA, USA
- Partner's contribution to the project: provided a Big Data analysis tool called, LucidWorks Big Data, as well as necessary support to install and use the tool on a cluster computer. Held workshops regarding LucidWorks.

Have other collaborators or contacts been involved? No

Impacts

You have the option of selecting "nothing to report" in this section.

What is the impact on the development of the principal discipline(s) of the project?

We have been refining our archiving process, which starts with the identification of events, and continues with tweet collection development and crawling of the actual webpages. Twitter data analysis and Big Data processing were also the techniques that we adopted and applied during the project. Our experiences were shared with other researchers through workshops and webinars, and it contributes to the archiving community and computer science field in positive ways.

What is the impact on other disciplines?

Traditionally, the practice of web archiving in the Library and Information Science (LIS) field required much human intervention, especially for the selection of seed URLs. However, in the case of large-scale disasters, the traditional LIS approach was almost incapable of identifying important seed URLs due to the volume of the webpages being generated. Our web archiving approach, which is based on crowd-sourced URLs using tweets (thus involving less human intervention), may

provide insights to the archiving personnel in the LIS field, as a more efficient and effective web archiving method with broader coverage of events. In addition, our project contributes to research in the social sciences, as indicated in the survey studies in Egypt, Mexico and Tunisia, discussed earlier. Specifically, the survey research contributes to social network analysis, political participation research, and knowledge about communication behavior and effects.

What is the impact on the development of human resources?

We had multiple collaborators, who were graduate students and faculty members on campus and in other institutions who kindly helped with this project. We believe that all those people have learned about the general procedure of web archiving and tweet data archiving and analysis as well as details of technology used. Such experiences would be a direct and indirect help for their studies.

What is the impact on physical resources that form infrastructure?

We used local funds to add 6 drives (3TB each) to a local server so we could manage the growing amount of CTRnet data for our digital library.

We have been discussing with the campus high performance computing (HPC) groups about the needs for managing locally the CTRnet and other similar types of data, leading to plans for some of the campus HPC systems to support this genre of Big Data requirements.

We have been discussing with the Department of Computer Science technical staff about these needs, and they now plan on acquiring, in support of such research, a cluster to use for research that would combine the Hadoop software and a number of 'fat nodes" with storage.

What is the impact on institutional resources that form infrastructure?

The campus and department groups focused on infrastructure for the institution are now planning on extending support for the type of research carried out in CTRnet.

What is the impact on information resources that form infrastructure?

We could develop many more disaster webpage archives, as well as more tweet archives, including not only disaster events but also health problems. Our archives are hosted by the Internet Archive and accessible through the Wayback Machine. Such webpages will provide an open and broader access to researchers in the disaster field and the general public.

What is the impact on technology transfer?

Our interactions with LucidWorks in several webinars and news coverage, as well as the WADL 2013 workshop held in conjunction with JCDL 2013, has spread awareness about this project and

the related resources and technologies developed. For example, Judith Lamont, on March 31, 2013, in her article “Search: power tools that leverage corporate knowledge”, appearing in the April 2013 issue of KMWorld (vol. 22, issue 4), discusses “Understanding disasters” and “Exploring, analyzing information”, referring to our project.

Kiran Chitturi, who worked on our project, now is employed by LucidWorks, another key part of our technology transfer results.

What is the impact on society beyond science and technology?

As mentioned above, our archived disaster webpages are hosted at the Internet Archive and accessible with the Wayback Machine. Due to its openness and potential for broader impact, people who are not in science and technology fields can freely access the archived webpages. Based on such information access, those users may have insights and detailed understanding related to their interests in disaster events.

Changes / Problems

If not previously reported in writing to the agency through other mechanisms, provide the following additional information or state, "Nothing to Report", if applicable.

Changes in approach and reason for change:

Nothing to Report

Actual or Anticipated problems or delays and actions or plans to resolve them:

Nothing to Report

Changes that have a significant impact on expenditures:

Nothing to Report

Significant changes in use or care of human subjects:

Nothing to Report

Significant changes in use or care of vertebrate animals:

Nothing to Report

Significant changes in use or care of biohazards:

Nothing to Report

Special Requirements

This report section is only available when Special Requirements are specifically noted in the solicitation and approved by the Office of Management and Budget.

NOTE: You may upload PDF files in support of the Special Requirements section. You may upload PDF files with a maximum file size of 10 MB each. There is no limit to the number of files uploaded.