My Desktop
Prepare & Submit Proposals
Proposal Status
Proposal Functions
Awards & Reporting
Notifications & Requests
Project Reports
Submit Images/Videos
Award Functions
Manage Financials
Program Income Reporting
Grantee Cash Management Section Contacts
Administration
Lookup NSF ID

Preview of Award 1619028 - Annual Project Report

Cover |
Accomplishments |
Products |
Participants/Organizations |
Impacts |
Changes/Problems

Cover

Federal Agency and Organization Element to Which Report is Submitted:	4900
Federal Grant or Other Identifying Number Assigned by Agency:	1619028
Project Title:	III: Small: Collaborative Research: Global Event and Trend Archive Research (GETAR)
PD/PI Name:	Edward A Fox, Principal Investigator Andrea L Kavanaugh, Co-Principal Investigator Chandan K Reddy, Co-Principal Investigator Donald J Shoemaker, Co-Principal Investigator
Recipient Organization:	Virginia Polytechnic Institute and State University
Project/Grant Period:	01/01/2017 - 12/31/2019
Reporting Period:	01/01/2017 - 12/31/2017
Submitting Official (if other than PD\PI):	Edward A Fox Principal Investigator
Submission Date:	01/02/2018
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	Edward A Fox

Accomplishments

*** What are the major goals of the project?**

We will ingest tweets and Web-based content from social media and the general Web, including news and governmental information. In addition to archiving materials found, we will build an information system that includes related metadata and knowledge bases, consistent with the 5S (Societies, Scenarios, Spaces, Structures, Streams) framework, along with results from our intelligent focused crawler, to support comprehensive access to event related content. With the support of key partners, the GETAR team will undertake research, education, and dissemination efforts, to achieve three complementary objectives:
1. Collecting: We will spot, identify, and make sense of interesting events and trends. We also will accept specific or general requests about types of events. Given resource and sampling constraints, we will integrate methods to identify appropriate URLs as seeds, and specify when to start crawling and when to stop, with regard to each event or subevent. We will integrate focused crawling and filtering approaches in order to ingest content and generate new collections, with high precision and recall.
2. Archiving & Accessing: Permanent archiving, and access to those archives, will be ensured by our partner, Internet Archive (IA). Immediate access to ingested content will be facilitated through big data software built on top of our Hadoop cluster.
3. Analyzing & Visualizing: We will provide a wide range of integrated services beyond the usual (faceted) browsing and searching, including: classification, clustering, summarization, text mining, topic identification, trend analysis, and visualization.

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

Major Activities: The GETAR project, with more than 40 collaborators and 10 collaborating institutions, developed tweet and webpage collections, datasets, services, software, systems, and methods. The related activities include: collecting event-related content, software and system development and refinement, experimentation, evaluation, and working with diverse users (representing key stakeholder groups).

The Internet Archive (see also its separate report for IIS-1619371) has expanded its collections and technology support, as well as outreach activities. It hosts, preserves, and provides public access with attribution to web collections created by the project team through its public Wayback Machine interface and Archive-It service. The latter may be browsed by descriptive metadata and searched through Archive-It's newly deployed full-text Elasticsearch engine. New or updated Internet Archive and Archive-It API documentation and workshops provide project stakeholders with several means to query the data from and about these collections, and to derive datasets for further textual and visual analyses. Studies have proceeded of important events, including integration of survey and analysis approaches, and publishing findings (e.g., about communications, elections, and political events in Mexico and Tunisia).

Specific Objectives: Doctoral dissertation research led to improved methods for focused crawling and for assigning locations to tweets. More than 22 computers are connected, mostly in a Hadoop cluster. This network was constructed to support collection, processing, and access already to almost 2 billion tweets across over 1300 collections, along with millions of webpages, covering hundreds of important events. Regarding collections, prior collections were extended, new ones were launched as events occurred or requests were made by users, the event focused crawler was deployed, and diverse related curation efforts proceeded. Master's thesis and class (independent study and graduate courses) research led to improved tweet and webpage techniques for content cleaning and processing, information extraction, classification, clustering, topic analysis, sentiment analysis, indexing, searching, browsing, and visualization.

Significant Results:

Advances have been made in big data handling, computational linguistics, digital libraries, information retrieval, information visualization, machine learning, and Web archiving. These have been integrated into a large system built around a Hadoop cluster, that works with growing numbers of expanding collections of tweets and webpages, supplemented by cleaning, information extraction, and adding value through advanced analysis.

The GETAR project has developed novel methodology and workflows, tailored to addressing the challenging problem of working with events. At a high level, for collection building, is a workflow to collect tweets about each event or event class, extract URLs, use the URLs present therein as seeds to our event focused crawler, and add resulting webpages to our Web collection. The event focused crawler workflow uses the extracted URLs as seeds to construct an event model that guides the selection and focused crawling for webpages. Key new methods were developed to analyze and accordingly add value (and metadata) to the collected content. Regarding our processing of tweets, a new framework was devised to streamline a variety of tweet analysis and transformation workflows. Regarding building tweet classifiers for the hundreds of events studied, a learning optimizer method was devised employing iterative processing with minimal human effort to yield high quality classification of tweets into collections for particular real world events. Regarding the problem that few tweets have associated latitude and longitude values, a methodology was devised for associating locations with tweets based on location indicative words.

Key outcomes or Other achievements:

Collection building and analysis (of both tweets and webpages) has improved through advances in classification, big data workflows, focused crawling (to identify webpages focused on an event of interest), inferring the location of tweets from their text when GPS data is unavailable, topic analysis, and natural language processing (including Arabic). Insights gained have been shared regarding juvenile delinquency, school shootings, and the use of information during conflicts, crises, elections, and uprisings. Collections are available to support other research and exploration regarding important events since 2007 such as the above, as well as attacks, bombings, celebrations, climate change, collapses, community activities, crashes, disease outbreaks, earthquakes, eclipses, environmental disruptions, erosion, explosions, fires, floods, hurricanes, innovations, judicial decisions, pollution, power outages, protests, revolutions, shootings, sports, storms, summits, tornadoes, transportation failures, tsunamis, typhoons, and veteran activities.

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

In the Fall 2016 class CS5604 (Information Retrieval, IR), the class-wide term project, carried out by students working in six teams (each uploading deliverables into the local institutional repository), was in support of GETAR. Through project based learning they applied IR theory and methods, using our Hadoop cluster, to ingest, analyze, index, and visualize event-related tweets and webpages. In Spring 2017, two teams in CS6604 (Digital Libraries) worked on projects related to GETAR, also uploading deliverables (e.g., reports, presentations, data, code). Andrej Galad completed his MS Independent Study, while Matthew Bock and Saurabh Chakravarty completed their MS theses. In addition to the earlier doctoral dissertations of Seungwon Yang and Tarek Kanan, two more dissertations were completed in this period, by Mohamed Magdy Gharib Farag and Sunshin Lee.

*** How have the results been disseminated to communities of interest?**

Dissemination has been through the reported publications and presentations. Further dissemination was through the project website (http://eventsarchive.org) and the website connected to the tweet collections and descriptions (http://hadoop.dlib.vt.edu/). In addition, we led the 2017 Web Archiving and Digital Libraries (WADL) workshops, with related proceedings.

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

We have updated our cluster software late in 2017, and are proceeding with collection and data analysis activities, using software and other results from CS5604 projects in fall 2017. In spring 2018 there should be some term projects in CS4624 (Multimedia, Hypertext, and Information Access), as well as several volunteer student efforts, to further extend our efforts. One effort will be to consolidate and extend results from our work with data related to the August 2017 Solar Eclipse, using tweets collected since summer 2017. Another effort will be to finalize the MS thesis work of Abigail Bartolome, who has been working with GETAR since its inception. Then in fall 2018 the CS5604 class will again focus on team term projects related to GETAR.

Products

Books

Book Chapters

Inventions

Journals or Juried Conference Papers

Ahuja, Aman, Wei Wei, Wei Lu, Kathleen M. Carley and Chandan K. Reddy (2017). A Probabilistic Geographical Aspect-Opinion Model for Geo-tagged Microblogs. *In Proceedings of the IEEE International Conference on Data Mining (ICDM), New Orleans, LA, November 2017*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1109/ICDM.2017.82

Castro, Eduardo P. S.; Saurabh Chakravarty; Eric Williamson; Denilson Alves Pereira; Edward A. Fox (2017). Classifying Short Unstructured Data Using the Apache Spark Platform. *Proceedings Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on, 19-23 June 2017, Toronto, ON, Canada, IEEE*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1109/JCDL.2017.7991567

Dave, Vachik S., Mohammad Al Hasan, and Chandan K. Reddy (2017). How Fast Will You Get a Response? Predicting Interval Time for Reciprocal Link Creation. *In Proceedings of Eleventh International AAAI Conference on Web and Social Media (ICWSM), Montréal, Canada, May 2017*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15676

Fox, Edward A. (2017). Introduction to digital libraries. *Proceedings Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on, 19-23 June 2017, Toronto, ON, Canada, IEEE*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1109/JCDL.2017.7991620

Fox, Edward A.; Zhiwu Xie; Martin Klein (2017). Web Archiving and Digital Libraries (WADL). *Proceedings Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on, 19-23 June 2017, Toronto, ON, Canada, IEEE*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1109/JCDL.2017.7991625

Kim, Hannah, Jaegul Choo, Changhyun Lee, Hanseung Lee, Chandan K. Reddy, and Haesun Park (2017). PIVE: Per-Iteration Visualization Environment for Real-time Interactions with Dimension Reduction and Clustering. *In Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI), San Francisco, CA, February 2017*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER: https://aaai.org /ocs/index.php/AAAI/AAAI17/paper/download/14381/13873

Rakesh, Vineeth, Niranjan Jadhav, Alexander Kotov, and Chandan K. Reddy (2017). Probabilistic Social Sequential Model for Tour Recommendation. *In Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM), Cambridge, UK, February 2017*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1145/3018661.3018711

Suh, Sangho, Jaegul Choo, Joonseok Lee, and Chandan K. Reddy (2017). Local Topic Discovery via Boosted Ensemble of Nonnegative Matrix Factorization. *In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Melbourne, Australia, August 2017*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.24963/ijcai.2017/699

Vinzamuri, Bhanukiran, Yan Li, and Chandan K. Reddy (2017). Pre-Processing Censored Survival Data Using Inverse Covariance Matrix Based Calibration. *IEEE Transactions on Knowledge and Data Engineering (TKDE), pp.2111-2124, October 2017*. 29 (10), 2111. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1109/TKDE.2017.2719028

Licenses

Other Conference Presentations / Papers

Shoemaker, Donald J., Jason Callahan, Liuqing Li, Edward Fox, and Islam Harb (2017). *Analysis of U.S. School Shootings Using Big Data*. Annual Meeting of the American Society of Criminology. Philadelphia, Pennsylvania. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Other Products

Other Publications

Galad, Andrej (2016). *ArchiveSpark - MS Independent Study Final Submission*. Virginia Tech, 2016-12-13, http://hdl.handle.net/10919/77457. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Kohler, Rachel; Tasooji, Reza; Sullivan, Patrick (2016). *CS 5604 Information Storage and Retrieval Front-End Team Fall 2016 Final Report*. Virginia Tech, 2016-12-08, http://hdl.handle.net/10919/73711. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Williamson, Eric R.; Chakravarty, Saurabh (2016). *CS5604 Fall 2016 Classification Team Final Report*. Virginia Tech, 2016-12-08, http://hdl.handle.net/10919/73713. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Li, Liuqing; Pillai, Anusha; Wang, Ye; Tian, Ke (2016). *CS5604 Fall 2016 Solr Team Project Report*. Virginia Tech, 2016-12-07, http://hdl.handle.net/10919/73710. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Wagner, Mitchell J.; Abidi, Faiz; Fan, Shuangfei (2016). *CS5604: Information and Storage Retrieval Fall 2016 - CMT (Collection Management Tweets)*. Virginia Tech, 2016-12-08, http://hdl.handle.net/10919/73739. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Chon, Jieun; Wang, Haitao; Bian, Yali; Niu, Shuo (2017). *CS5604: Information and Storage Retrieval Fall 2017 - FE (Front-End Team)*. Virginia Tech, 2017-12-24, http://hdl.handle.net/10919/81423. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Li, Liuqing; Harb, Islam; Galad, Andrej (2017). *CS6604 Spring 2017 Global Events Team Project*. Virginia Tech, 2017-05-03, http://hdl.handle.net/10919/77867. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Bartolome, Abigail; Islam, MD; Vundekode, Soumya (2016). *Clustering and Topic Analysis in CS 5604 Information Retrieval Fall 2016*. Virginia Tech, 2016-12-08, http://hdl.handle.net/10919/73712. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Eagan, Mackenzie; Liang, Xiao; Michael, Louis; Patil, Supritha (2017). *Collection Management Webpages*. CS5604 Information Retrieval, Fall 2017, Virginia Polytechnic Institute and State University, 2017-12-25, http://hdl.handle.net/10919/81428. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Dao, Tung; Wakeley, Christopher; Weigang, Liu (2017). *Collection Management Webpages - Fall 2016 CS5604*. Virginia Tech, 2017-03-23, http://hdl.handle.net/10919/76675. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Fox, Edward A., Sunshin Lee, Mohamad Farag, Andrea Kavanaugh, Donald Shoemaker, Jefferson Bailey, Steve Sheetz, Liuqing Li, Islam Harb, Seungwon Yang (2017). *GETAR Overview*. Presentation for Discovery Analytics Center, Virginia Tech, 2 February 2017, http://fox.cs.vt.edu/talks/2017/20170202GETAR-DAC.pptx. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Manchester, Emma; Srinivasan, Ravi; Crenshaw, Sean; Masterson, Alec; Grinnan, Harrison (2017). *Global Event Crawler and Seed Generator for GETAR*. CS4624: Multimedia, Hypertext, and Information Access, Spring 2017, Virginia Tech, 2017-04-28, http://hdl.handle.net/10919/776202017. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Fox, Edward A. (2017). *Information Research*. Presentation for STEP 2017 Faculty Research Seminar, 7/7/2017, Virginia Tech, http://fox.cs.vt.edu/talks/2017/20170707STEPfacultySeminar.pptx. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Fox, Edward A. (2016). *Information Research*. Presentation for ENGR 1014: Engineering Research Seminar, 2 September 2016, Virginia Tech, http://fox.cs.vt.edu/talks/2016/20160902ENGR1014Fox.pptx. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Fox, Edward A. (2017). *Introduction to Digital Libraries*. Tutorial presentation at JCDL 2017 (Toronto – 19 June 2017), http://fox.cs.vt.edu/talks/2017/20170613JCDL2017FoxTutorialSlides.pptx. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Bartolome, Abigail; Bock, Matthew; Vinayagam, Radha Krishnan; Krishnamurthy, Rahul (2017). *Sentiment and Topic Analysis*. CS6604 Digital Libraries, Spring 2017, Virginia Tech, 2017-05-03, http://hdl.handle.net/10919/77883. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Renugopal, Jishnu; Zargarpur, Mattin; Zhao, Haiyu; Richardson, Christian; Zhang, Kevin; Schmidt, Will (2017). *VR4GETAR*. CS4624: Multimedia, Hypertext, and Information Access, Spring 2017, Virginia Tech, 2017-04-28, http://hdl.handle.net/10919/77616. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Fox, Edward A. (2017). *Web Archiving Research Supported by Text and Data Mining*. Presentation for Text and Data Mining Discussion Forum, Wednesday, 4/12/2017, Virginia Tech, Newman Library, http://fox.cs.vt.edu/talks/2017/20170412TDMforumFox.pptx. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Patents

Technologies or Techniques

Thesis/Dissertations

Bock, Matthew. *A Framework for Hadoop Based Digital Libraries of Tweets*. (2017). MS Thesis, Virginia Tech, 2017-07-17, http://hdl.handle.net/. Acknowledgement of Federal Support = Yes

Chakravarty, Saurabh. *A Large Collection Learning Optimizer Framework*. (2017). MS Thesis, Virginia Tech, 2017-06-30, http://hdl.handle.net/. Acknowledgement of Federal Support = Yes

Lee, Sunshin. *Geo-Locating Tweets with Latent Location Information*. (2017). Doctoral Dissertation, Virginia Tech, 2017-02-13, http://hdl. Acknowledgement of Federal Support = Yes

Farag, Mohamed Magdy Gharib. *Intelligent Event Focused Crawling*. (2016). Doctoral Dissertation, Virginia Tech, 2016-09-23, http://hdl. Acknowledgement of Federal Support = Yes

Websites

Events Archiving
http://eventsarchive.org

Website about event and trend archiving research, covering work on a series of related NSF-supported projects: Global Event and Trend Archive Research (GETAR), IIS-1619028 and 1619371; Integrated Digital Event Archiving and Library (IDEAL), IIS-1319578; Crisis, Tragedy, and Recovery Network, CTRnet, IIS-0916733; and DL-VT416: A Digital Library Testbed for Research Related to 4/16/2007 at Virginia Tech, IIS-0736055

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Fox, Edward	PD/PI	1
Kavanaugh, Andrea	Co PD/PI	1
Reddy, Chandan	Co PD/PI	1
Shoemaker, Donald	Co PD/PI	1
Bailey, Jefferson	Co-Investigator	1

Name	Most Senior Project Role	Nearest Person Month Worked
Agozino, Onwubiko	Faculty	0
Angermeier, Paul	Faculty	0
Farag, Mohamed	Faculty	0
Horning, Mike	Faculty	0
Jelesko, John	Faculty	0
Kanan, Tarek	Faculty	0
Krometis, Leigh	Faculty	0
Lee, Sunshin	Faculty	1
Murray-Tuite, Pamela	Faculty	0
North, Chris	Faculty	0
Pereira, Denilson	Faculty	0
Salehi-Isfahani, Djavad	Faculty	0
Sandoval-Almazan, Rodrigo	Faculty	0
Sheetz, Steven	Faculty	0
Skandrani, Hamida	Faculty	0
Smith, Eric	Faculty	0
Tedesco, John	Faculty	0
Wimberley, Dale	Faculty	0
Xie, Zhiwu	Faculty	0
Yang, Seungwon	Faculty	0
Coleman, Shane	Staff Scientist (doctoral level)	0

Name	Most Senior Project Role	Nearest Person Month Worked
Klein, Martin	Staff Scientist (doctoral level)	0
Mather, Paul	Staff Scientist (doctoral level)	0
Sforza, Peter	Staff Scientist (doctoral level)	0
Ahuja, Aman	Graduate Student (research assistant)	0
Bartolome, Abigail	Graduate Student (research assistant)	0
Bock, Matthew	Graduate Student (research assistant)	1
Callahan, Jason	Graduate Student (research assistant)	0
Chakravarty, Saurabh	Graduate Student (research assistant)	1
Galad, Andrej	Graduate Student (research assistant)	1
Harb, Islam	Graduate Student (research assistant)	1
Ireland, Leanna	Graduate Student (research assistant)	0
Li, Liuqing	Graduate Student (research assistant)	5
Niu, Shuo	Graduate Student (research assistant)	0
Patil, Supritha	Graduate Student (research assistant)	0
Song, Ziqian	Graduate Student (research assistant)	0
Wang, Ji	Graduate Student (research assistant)	0
Wang, Xinyue	Graduate Student (research assistant)	0
Zhang, Xuan	Graduate Student (research assistant)	0
Todorov, Teddy	Undergraduate Student	0

Full details of individuals who have worked on the project:

Edward A Fox

Email: fox@vt.edu
Most Senior Project Role: PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: PI/PD in charge of this grant, supervising graduate assistants, teaching students completing class projects, supervising theses and dissertations

Funding Support: NSF IIS-1619028

International Collaboration: No
International Travel: No

Andrea L Kavanaugh
Email: kavan@vt.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Co-PI working on collaboration, data collection, data analysis, publication, and supervision of students

Funding Support: NSF IIS-1619028

International Collaboration: No
International Travel: No

Chandan K Reddy
Email: reddy@cs.vt.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Co-PI working on algorithm and software development, collaboration, data collection, data analysis, publication, and supervision of students

Funding Support: NSF IIS-1619028

International Collaboration: No
International Travel: No

Donald J Shoemaker
Email: shoemake@vt.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Co-PI working on collaboration, data collection, data analysis, publication, and supervision of students

Funding Support: NSF IIS-1619028

International Collaboration: No
International Travel: No

Jefferson Bailey
Email: jefferson@archive.org
Most Senior Project Role: Co-Investigator
Nearest Person Month Worked: 1

Contribution to the Project: Internet Archive is a collaborative partner, also receiving funds on this project from NSF, through IIS-1619371. We use their equipment and services and data, and collaborate on research.

Funding Support: This project, i.e., IIS-1619371

International Collaboration: No
International Travel: No

Onwubiko Agozino
Email: agozino@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Paul Angermeier
Email: biota@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Mohamed Farag
Email: mohamedmagdy@gmail.com
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaborating regarding the software he developed as part of his earlier doctoral work supported by a prior NSF project, IDEAL, related to this research, as well as this project.

Funding Support: Local support

International Collaboration: Yes, Egypt
International Travel: No

Mike Horning
Email: mhorning@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

John Jelesko
Email: jelesko@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Tarek Kanan
Email: tarek.kanan@gmail.com
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaborate regarding NLP and handling of Arabic texts, extending his doctoral work completed earlier at VT

Funding Support: Local support

International Collaboration: Yes, Jordan
International Travel: No

Leigh Anne Krometis
Email: lehenry@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Sunshin Lee
Email: slee116@radford.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 1

Contribution to the Project: Worked as GRA on this project, then as postdoc, now as faculty at Radford, collaborating to extend his doctoral research.

Funding Support: Local support

International Collaboration: No
International Travel: No

Pamela Murray-Tuite
Email: pmmurra@clemson.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Formerly a VT faculty member, now on faculty at Clemson, she is PI on another NSF project in which PI Fox serves at co-PI, and is helping with curation and analysis of data related to disasters that effect both transportation and power systems.

Funding Support: Local support

International Collaboration: No
International Travel: No

Chris North
Email: north@cs.vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on software, collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Denilson Pereira
Email: denilsonpereira@dcc.ufla.br
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaboration on publications and related research connected with analysis of tweets, classification, disambiguation, and text analysis

Funding Support: Local support

International Collaboration: Yes, Brazil
International Travel: No

Djavad Salehi-Isfahani
Email: salehi@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Rodrigo Sandoval-Almazan
Email: rsandovuaem@gmail.com
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaborating regarding publication, analysis, and collection/curation of data related to events in Mexico.

Funding Support: Local support

International Collaboration: No
International Travel: No

Steven D. Sheetz
Email: sheetz@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaborator on IDEAL, prior related project, as co-PI.

Funding Support: Local support

International Collaboration: No
International Travel: No

Hamida Skandrani
Email: hamida.skandrani@gmail.com
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaborated on publications and studies related to events in Tunisia and the region.

Funding Support: Local support

International Collaboration: Yes, Tunisia
International Travel: No

Eric Smith
Email: epsmith@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

John Tedesco
Email: tedesco@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Dale Wimberley
Email: wimberly@vt.edu
Most Senior Project Role: Faculty

Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Zhiwu Xie
Email: zhiwuxie@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Library researcher collaborating on web archiving.

Funding Support: Local support

International Collaboration: No
International Travel: No

Seungwon Yang
Email: seungwonyang@lsu.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaborate on collecting and analyzing tweets related to events, especially related to the Gulf region.

Funding Support: Local support

International Collaboration: No
International Travel: No

Shane Coleman
Email: shanec4@vt.edu
Most Senior Project Role: Staff Scientist (doctoral level)
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Martin Klein
Email: martinklein0815@gmail.com
Most Senior Project Role: Staff Scientist (doctoral level)
Nearest Person Month Worked: 0

Contribution to the Project: Collaborate on collecting data (e.g., tweets) and undertaking related analysis.

Funding Support: Local support

International Collaboration: No
International Travel: No

Paul Mather
Email: pmather@vt.edu
Most Senior Project Role: Staff Scientist (doctoral level)
Nearest Person Month Worked: 0

Contribution to the Project: Collaborate on managing our equipment and software, connecting with Library efforts.

Funding Support: Local support

International Collaboration: No
International Travel: No

Peter Sforza
Email: psforza@vt.edu
Most Senior Project Role: Staff Scientist (doctoral level)
Nearest Person Month Worked: 0

Contribution to the Project: Director of local center on GIS collaborating on spatial location.

Funding Support: Local support

International Collaboration: Yes, Mexico
International Travel: No

Aman Ahuja
Email: aahuja@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Graduate student collaborating on project research.

Funding Support: Local support

International Collaboration: No

International Travel: No

Abigail Bartolome
Email: abijbart@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Graduae student working on thesis in collaboration with project.

Funding Support: Local support

International Collaboration: No
International Travel: No

Matthew Bock
Email: mattb93@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 1

Contribution to the Project: Graduate student completing thesis in support of this project.

Funding Support: Local support

International Collaboration: No
International Travel: No

Jason Callahan
Email: jcallaha@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Graduate student collaborating on project research.

Funding Support: Local support

International Collaboration: No
International Travel: No

Saurabh Chakravarty
Email: saurabc@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 1

Contribution to the Project: Graduate student completing thesis in support of this project.

Funding Support: Local support

International Collaboration: No
International Travel: No

Andrej Galad
Email: agalad@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 1

Contribution to the Project: Graduate student completing independent study project in support of this project.

Funding Support: Local support

International Collaboration: No
International Travel: No

Islam Harb
Email: iharb@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 1

Contribution to the Project: GRA supported by this project, working on all aspects.

Funding Support: This project

International Collaboration: No
International Travel: No

Leanna Ireland
Email: lireland@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Graduate student collaborating on project research.

Funding Support: Local support

International Collaboration: No
International Travel: No

Liuqing Li
Email: liuqing@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 5

Contribution to the Project: GRA supported by this project, working on all aspects.

Funding Support: This project

International Collaboration: No
International Travel: No

Shuo Niu
Email: shuoniu@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Graduate student collaborating on project research.

Funding Support: Local support

International Collaboration: No
International Travel: No

Supritha Patil
Email: patil93@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Graduate student collaborating on project research.

Funding Support: Local support

International Collaboration: No
International Travel: No

Ziqian Song
Email: ziqian@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Graduate student collaborating on project research.

Funding Support: Local support

International Collaboration: No
International Travel: No

Ji Wang
Email: wji@cs.vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Graduate student collaborating on project research.

Funding Support: Local support

International Collaboration: No
International Travel: No

Xinyue Wang
Email: xw0078@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Graduate student collaborating on project research.

Funding Support: Local support

International Collaboration: No
International Travel: No

Xuan Zhang
Email: xuancs@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Graduate student collaborating on project research.

Funding Support: Local support

International Collaboration: No
International Travel: No

Teddy Todorov
Email: ttodorov@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 0

Contribution to the Project: Volunteer helping with software to identify important events.

Funding Support: Local support

International Collaboration: No
International Travel: No

What other organizations have been involved as partners?

Name	Type of Partner	Location

Organization		
Al Zaytonah University of Jordan	Academic Institution	Jordan
Arab Academy for Science and Technology	Academic Institution	Alexandria, Egypt
University of Tunis - Manouba Campus	Academic Institution	Tunisia
Clemson University	Academic Institution	Clemson, SC
George Washington University	Academic Institution	Washington, D.C.
Internet Archive	Other Nonprofits	San Francisco, CA
Los Alamos National Laboratory	State or Local Government	Los Alamos, New Mexico
Louisiana State University	Academic Institution	Baton Rouge, LA
Radford University	Academic Institution	Radford, VA
Universidad Autónoma del Estado de México (UAEM)	Academic Institution	Mexico
Universidade Federal de Lavras (UFLA)	Academic Institution	Lavras, MG, Brasil

Full details of organizations that have been involved as partners:

Al Zaytonah University of Jordan

Organization Type: Academic Institution
Organization Location: Jordan

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Tarek Kanan continues to collaborate regarding NLP and handling of Arabic texts, extending his doctoral work completed earlier at VT.

Arab Academy for Science and Technology

Organization Type: Academic Institution
Organization Location: Alexandria, Egypt

Partner's Contribution to the Project:

Collaborative Research

More Detail on Partner and Contribution: Mohamed Farag, on the faculty, is collaborating regarding the software he developed as part of his earlier doctoral work supported by a prior NSF project, IDEAL, related to this research, as well as this project.

Clemson University

Organization Type: Academic Institution
Organization Location: Clemson, SC

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Dr. Pamela Murray-Tuite, formerly a VT faculty member, now on faculty at Clemson, is PI on another NSF project in which PI Fox serves at co-PI, and is helping with curation and analysis of data related to disasters that effect both transportation and power systems.

George Washington University

Organization Type: Academic Institution
Organization Location: Washington, D.C.

Partner's Contribution to the Project:
In-Kind Support

More Detail on Partner and Contribution: We use the Social Feed Manager software from GWU Libraries, which they continue to support and enhance through collaboration.

Internet Archive

Organization Type: Other Nonprofits
Organization Location: San Francisco, CA

Partner's Contribution to the Project:
Facilities
Collaborative Research

More Detail on Partner and Contribution: Internet Archive is a collaborative partner, also receiving funds on this project from NSF, through IIS-1619371. We use their equipment and services and data, and collaborate on research. Jefferson Bailey is co-PI on GETAR.

Los Alamos National Laboratory

Organization Type: State or Local Government
Organization Location: Los Alamos, New Mexico

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: We collaborate on collecting data (e.g., tweets) and undertaking related analysis. This work is led by Dr. Martin Klein.

Louisiana State University

Organization Type: Academic Institution
Organization Location: Baton Rouge, LA

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: We collaborate on collecting and analyzing tweets related to events, especially related to the Gulf region. This is led by Dr. Seungwon Yang, whose doctoral work was supported in part by prior NSF-funded related projects at VT.

Radford University

Organization Type: Academic Institution
Organization Location: Radford, VA

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Dr. Sunshin Lee, whose Ph.D. work was supported by this project before he started as a faculty member at Radford, continues to collaborate with our research. He served also as volunteer postdoc aiding GETAR before going to Radford.

Universidad Autónoma del Estado de México (UAEM)

Organization Type: Academic Institution
Organization Location: Mexico

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Dr. Sandoval Almazan is collaborating regarding publication, analysis, and collection/curation of data related to events in Mexico.

Universidade Federal de Lavras (UFLA)

Organization Type: Academic Institution
Organization Location: Lavras, MG, Brasil

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Dr. Pereira is continuing collaboration on publications and related research connected with analysis of tweets, classification, disambiguation, and text analysis.

University of Tunis - Manouba Campus

Organization Type: Academic Institution
Organization Location: Tunisia

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Hamida Skandrani has collaborated on publications and studies related to events in Tunisia and the region.

What other collaborators or contacts have been involved?
Nothing to report

Impacts

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

The success of repeated offerings of CS5604, Information Retrieval (IR), being run using the pedagogial method of problem/project based learning, with the problem of how to develop an advanced information retrieval system in support of GETAR, should help others teaching IR to improve the learning in their classes by similar connection to research.
The event focused crawler extends the scope of Web crawling to situations where webpages are sought about an event, rather than about a topic or organization or website.
The methodology of using location indicative words to infer location for tweets that lack latitude and longitude values should expand the utility of tweet collections and related social network studies, by enabling analyses and visualizations that involve locations or geospatial reasoning.
The integration of processing of tweets and webpages, all related to important events, in one system with linked workflows, should broaden the scope of studies that largely just use only one of these two sources for digital library and Web archiving research.

What is the impact on other disciplines?

Tweet and webpage collections are of interest to many disciplines studying recent history and current events, including history, sociology, political science, economics, environmental science, linguistics, communications, government, etc. As a result of this project, scores of Virginia Tech scholars, from a variety of departments as well as University Libraries, have expressed interest in our methods and activities, and a number have worked with us on focused studies. We have collected information and shared that with them, as well as helped with related analysis. This shows how broad impact is likely to spread to a number of other disciplines.

What is the impact on the development of human resources?

GETAR has led to 2 dissertations, 2 theses, and 13 student reports across 4 offerings in 3 different courses. The application of problem/project based learning has been very popular with students, who are highly motivated, and apply their skills in other courses as well as internships and work after graduation. Students working on the project are now in faculty positions at Louisiana State University and Radford University as well as universities in Egypt and Jordan. A number of those involved are woman, and a number come from underrepresented groups.

The project has involved more than 40 collaborators and 10 collaborating institutions. People in diverse fields have been exposed to advanced data analytics and visualization, enhancing their appreciation of science and understanding about working with data.

What is the impact on physical resources that form infrastructure?

Project success led in 2017 to support in the form of two high-end desktop computers, each with 128GB RAM, paid through the State Council of Higher Education for Virginia (SCHEV), being added to the equipment supporting GETAR. With those additions, and ongoing maintenance of hardware and software, our complex of machines for collecting tweets, using a Hadoop cluster for large-scale processing, and using other computers to support searching and visualization, has led to a powerful integrated infrastructure to support our research, related education, and support for students, faculty, and staff at Virginia Tech, as well as beyond.

What is the impact on institutional resources that form infrastructure?

In addition to stimulating support from the Deparment of Computer Science for our infrastructure, University Libraries has built a very similar infrastructure, and the campus IT groups have launched several clusters to support other similar types of investigations.

What is the impact on information resources that form infrastructure?

Aided in part by the Web Archiving and Digital Libraries workshops, and other dissemination of project activities and accomplishments, other teams involved in Web archiving have engaged in related studies and efforts to devise software and methods, as well as build collections. There is a growing movement for collecting and archiving tweets and/or webpages, and to broaden the support for working with those archives. The enormous collection of over 300 billion webpages at the Internet Archive, as well as other archives, has stimulated broad interest in these information resources. Our methods to add value through analysis, and to support event-oriented studies and access, shows promise to expand the utility of the expanding information resources.

What is the impact on technology transfer?

The Internet Archive is a partner, working with the GETAR team, and has access to our technology, software, and data. Its actions broadly influence the rest of the worldwide Web archiving community.

66 total web collections representing 15 TB of data among more than 250 million unique web objects were created with the Internet Archive’s Archive-It service. They are available to be seen through its public Wayback Machine interface (archive.org/web) and by collection at: https://archive-it.org/organizations/156. The former provides attribution to the project team for collecting each web page in its collection. The latter enables stakeholders to browse specific collections by descriptive metadata and by full-text with Archive-It’s newly deployed Elasticsearch engine.

Internet Archive staff have updated or written new documentation for stakeholders and the general public to query and use data from and about these collections and their contents through its general and collection-specific Wayback index (CDX) APIs, OpenSearch API, and “WASAPI” web archive data transfer API. Improvements to derivative dataset generation and analysis processes that the project team may use to mine and/or visualize these archives were likewise documented and formed the basis of workshops in the United States and abroad to train further librarians, archivists, and researchers to use similar resources and tools.

What is the impact on society beyond science and technology?

The collections developed can be used by any interested groups. As our software and systems mature, open access to suitable portions of our collections will be provided to the public.

Changes/Problems

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

We have saved some funding assigned to students, that will be spent in the remainder of the project: The half-time GRA who worked on the project in Spring 2017 left with short notice. His replacement will start 1/1/2018. The full-time GRA who has worked since the project inception was offered a fall 2017 half-time teaching assistantship in CS5604, which allowed him to guide 36 students whose work was benefiting GETAR; he is resuming his full-time GRA for GETAR as of 1/1/2018.

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.