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 1319578 - 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:	1319578
Project Title:	III: Small: Integrated Digital Event Archiving and Library (IDEAL)
PD/PI Name:	Edward A Fox, Principal Investigator Kristine Hanna, Co-Principal Investigator Andrea L Kavanaugh, Co-Principal Investigator Steven D Sheetz, Co-Principal Investigator Donald J Shoemaker, Co-Principal Investigator
Recipient Organization:	Virginia Polytechnic Institute and State University
Project/Grant Period:	09/01/2013 - 08/31/2016
Reporting Period:	09/01/2014 - 08/31/2015
Submitting Official (if other than PD\PI):	N/A
Submission Date:	N/A
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	N/A

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 IDEAL 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. 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 sub-event. 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, and visualization.

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

Major Activities:

1. We developed and applied software and tools for collecting, storing, organizing, indexing, and analyzing Web and tweet collections. Interfaces and visualization methods were developed for interacting with the collections.
2. We collected tweets and webpages about many events (community, disaster, and governmental). Web collections were processed using software including Python and Nutch. Tweet collections also were stored on local servers.
3. Collections were indexed and made available for searching, browsing, and other services.
4. Reports from projects related to IDEAL in three courses have been prepared: CS4984: Computational Linguistics - Fall 2014 - 30 students; CS5604: Information Storage and Retrieval - Spring 2015 - 23 students; CS4624: Multimedia, Hypertext, and Information Access - Spring 2015 - 60 students exposed, with 17 working on IDEAL-related term projects.

Specific Objectives:

1. Web and tweet collections were built using focused crawling, Nutch, and tweet archiving tools. Collections span many kinds of events (community, disasters, and governmental) and places around the world, expanding upon work in the previous year.
2. A prototype event detection system was developed. A new event website was prototyped as well. Thus, reporting and collecting software and interfaces were developed to aid event spotting and subsequent assembling of new event collections.
3. A more advanced system for storing, organizing, and analyzing event collections was prototyped in a graduate Information Retrieval class using both small (for testing) and big data (to ensure scalability) collections, processed using our Hadoop cluster.
4. A prototype system was developed for analyzing events, to produce meaningful summaries, also using small and big data collections and our Hadoop cluster.
5. Advisory (internal and external) meetings were held to aid project planning and dissemination.

Significant Results:

1. We have built tweet collections about roughly 200 events (disasters, crises, community, and political). A list of tweet collections can be found at http://hadoop.dlib.vt.edu:81/twitter/. For some of these we have derived web collections. This page lists roughly 700 tweet collections, with about 100M tweets, giving the keyword / hashtag information involved in the collecting, but other databases also describe additional aspects of our roughly 1B tweets archived.
2. Using an event focused crawler, we have built higher precision web collections about the Tunisia hotel attack, Same sex marriage, Charleston shooting, FIFA arrests, Nepal earthquake, Boat capsized in Libya, Typhoon Hagupit, AirAsia plane crash, Sydney Seige, and the Charlie Hebdo shooting.

3. Web collections built on the Internet Archive site, over 65 in number, are accessible, e.g., from https://archive-it.org/organizations/156. Web collections on our local systems were made accessible on a test basis through a SOLR interface for topic identification, classification, clustering, and summarization. All accessible Web collections are being made available through the project website http://www.eventsarchive.org/. A demo of an analyzer of collections to produce events models is at http://nick.dlib.vt.edu/EventModel/
4. A Hadoop cluster was constructed (from parts) and extended to 20 nodes by interested students to support several projects, software was set up, and students in classes as well as other volunteers have helped make it useful for IDEAL.

Key outcomes or Other achievements:

1. Student project reports have documented the learning and findings of the students already involved, and are available online for others to learn from too (see below under Other Publications).
2. Other publications and presentations have helped disseminate results and helped expand our collaboration with partners and stakeholders.

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

1. Research has involved IDEAL staff and volunteer students regarding an event focused crawler prototype, tweet location disambiguation, and our new Hadoop cluster.
2. Three courses (2 undergraduate and 1 graduate, see above) included class presentations and term projects related to the IDEAL project. Two students in CS4994, undergraduate research, focused on helping with IDEAL.
3. Megan Eyler, an undergraduate student in Sociology, volunteered and worked with co-PI Dr. Donald Shoemaker.
4. Ph.D. student Yue Sun volunteered and worked with co-PI Dr. Andrea Kavanaugh.
5. Visiting scholar Dr. Sultan Al-Daihani, an associate professor at Kuwait University on sabbatical, collaborated with the project team during many meetings, discussions, and research explorations.
6. M.S. student Ziqian Song worked with co-PI Dr. Andrea Kavanaugh.
7. Two project GRAs, Mohamed Farag and Sunshin Lee, made substantial progress on their doctoral dissertations, expected to be completed in 2016, that are closely related to this project.

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

1. The IDEAL project has provided publicly available information and pointers for previous and current Web collections, and summaries for tweet collections.
2. The IDEAL project team presented our event focused crawler work at the Internet Archive-it 2014 annual meeting.
3. The IDEAL project team presented our Web archive content analysis work at the IIPC General Assembly 2015 annual meeting.
4. A short paper and poster were presented at the JCDL 2015 conference, covering big data text summarization and tweet location disambiguation. A general tutorial on digital libraries also was presented there by the PI, that included discussion of IDEAL as a case study.
5. We organized and ran the Web Archiving and Digital Library (WADL 2015) workshop at JCDL 2015.
6. We presented our research on intelligent event focused crawling (and have an extended abstract in the preliminary proceedings) at WADL 2015.
7. Undergraduate student Megan Eyler, working with co-PI Dr. Donald Shoemaker, presented a poster at a student research symposium run by Virginia Tech's Center for Peace Studies and Violoence Prevention, one of our local partners, see http://www.sociology.vt.edu/cpsvp/.

8. We refined our 2014 digital government paper and submitted it for journal publication.
9. At THATCamp (The Humanities and Technology Camp) 2015 Virginia, Blacksburg, VA, April 10-11, 2015, http://virginia2015.thatcamp.org/, the PI led a pre-conference workshop, aided by many on the project team, entitled Event Crawling, Archiving, and Exploring. Related discussions continued throughout THATCamp.
10. Co-PI Dr. Andrea Kavanaugh attended the June 15-16 workshop at Bentley University, helping connect our project with those studying data related to police and government activities.

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

1. More tweets will be collected covering all relevant kinds of events, using DMI-TCAT, which will supplement our prior tweet collecting with yourtwapperkeeper, since more metadata is available and since additional approaches to tweet collection are supported. After subsequent processing, data from these tweets will aid our collecting of webpages about those events.
2. More detailed and tailored analyses of our data will be prepared, along with development of better tools to aid in those analyses.
3. We will integrate multiple prototypes developed over the last two years into a publicly accessible system for Integrated Event Archiving and Digital Library (IDEAL). In addition to searching and browsing, it will support tailored analysis and visualization, for both tweet and webpage collections. It will add value by leveraging our work on classification, clustering, topic modeling, social network analysis, and named entity recognition.
4. We will extend the many contacts made over the past year, including with various Virginia Tech groups in the Library or in Digital Humanities (e.g., with the Center for Peace Studies and Violoence Prevention), as well as at other sites, to support the diverse current and prospective stakeholders who can benefit from IDEAL.
5. Working with the Internet Archive and others, we will further advance technology transfer of our research findings.
6. The project GRAs, Mohamed Farag and Sunshin Lee, will work to complete their IDEAL-related dissertations.

Products

Books

Book Chapters

Conference Papers and Presentations

Tarek Kanan, Xuan Zhang, Mohammed Magdy, and Edward Fox (2015). *Big Data Text Summarization for Events: a Problem Based Learning Course*. Joint Conference on Digital Libraries (JCDL 2015). Knoxville, TN. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Mohamed Farag and Edward Fox (2015). *Building and Archiving Event Web Collections: A focused crawler approach*. Web Archiving and Digital Libraries Workshop (WADL 2015) - Joint Conference on Digital Libraries (JCDL). Knoxville, TN. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Tarek Kanan, Souleiman Ayoub, Eyad Saif, Ghassan Kanaan, Prashant Chandrasekar, Edward Fox (2016). *Extracting Named Entities Using Named Entity Recognizer and Generating Topics Using Latent Dirichlet Allocation Algorithm for Arabic News Articles*. International Computer Sciences and Informatics Conference (ICSIC 2016). Amman, Jordan. Status = AWAITING_PUBLICATION; Acknowledgement of Federal Support = Yes

Sunshin Lee, Mohammed Farag, Tarek Kanan, and Edward A. Fox (2015). *Read between the lines: A Machine Learning Approach for Disambiguating the Geo-location of Tweets*. Joint Conference on Digital Libraries (JCDL 2015). Knoxville, TN. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Megan Eyler and Donald Shoemaker (2015). *Standardizing Information of Collections on School Shootings: Contextualizing Digitalized Data*. Cultivating Peace: A Student Research Symposium on Violence Prevention. Virginia Tech, Blacksburg, VA. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Richard Gruss, Tarek Kanan, Xuan Zhang, Mohamed Farag, Mary C. English, and Edward A. Fox (2015). *Teaching Big Data Through Project-based Learning in Computational Linguistics and Information Retrieval*. Consortium for Computing in Small Colleges: Southeastern Conference. Roanoke College, Salem, VA. Status = AWAITING_PUBLICATION; Acknowledgement of Federal Support = Yes

Mohamed Magdy Farag and Edward A. Fox (2015). *Web Archive Content Analysis: Disaster Events Case Study*. International Internet Preservation Consortium (IIPC) 2015 General Assembly (GA2015). Stanford, CA. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Edward A. Fox and Zhiwu Xie (2015). *Web Archiving and Digital Libraries (WADL)*. Joint Conference on Digital Libraries (JCDL 2015). Knoxville, TN. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Inventions

Journals

Jonathan P. Leidig and Edward A. Fox (2014). Intelligent Digital Libraries and Tailored Services. *Journal of Intelligent Information Systems*. 43 (3), 463-480. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1007/s10844-014-0342-3

Tarek Kanan and Edward A. Fox (2015). Automated Arabic Text Classification with P-Stemmer, Machine Learning, and a Tailored News Article Taxonomy. *Journal of the Association for Information Science and Technology (JASIST)*. 66 . Status = AWAITING_PUBLICATION; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Licenses

Other Products

Other Publications

Ayoub, Souleiman; Freeman, Julia (2015). *Arabic News Article Summarization*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/52339. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Cui, Xuewen; Tao, Rongrong; Zhang, Ruide (2015). *Classification Team Project for IDEAL*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/52253. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Bialousz, Kenneth; Kokal, Kevin; Orleans-Pobee, Kwamina; Wakeley, Chris (2015). *Computational Linguistic Analysis of Earthquake Collections*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/51132. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Crowder, Nicholas; Nguyen, David; Hsu, Andy; Mecklenburg, Will; Morris, Jeff (2015). *Computational Linguistics Hurricane Group*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/51136. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Thumma, Sujit Reddy; Kalidas, Rubasri; Torkey, Hanaa (2015). *Document Clustering for IDEAL*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/52341. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Antol, Stanislaw; Ayoub, Souleiman; Folgar, Carlos; Smith, Steve (2015). *Exploring the Blacksburg Community*

Events Collection. Technical Report, Virginia Tech, http://hdl.handle.net/10919/51135. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Pumma, Sarunya; Liu, Xiaoyang (2015). *Extracting Topics from Tweets and Webpages for IDEAL*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/52343. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Chandrasekaran, Arjun; Sharma, Saurav; Sulucz, Peter; Tran, Jonathan (2015). *Generating an Intelligent Human-Readable Summary of a Shooting Event from a Large Collection of Webpages*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/51137. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Cadena, Jose; Chen, Mengsu; Wen, Chengyuan (2015). *Hadoop Project for IDEAL*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/52342. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Du, Qianzhou; Zhang, Xuan (2015). *Named Entity Recognition for IDEAL*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/52254. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Acanfora, Joseph; Evangelista, Marc; Keimig, David; Su, Myron (2015). *Natural Language Processing: Generating a Summary of Flood Disasters*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/51134. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Cummins, John; Ciambrone, Andrew; Haile, Beakal; Quinn, Liu (2015). *New Event Detection*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/53836. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Crabb, Briana; Aconfora, Joe; Morris, Jeff (2015). *News Event Website*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/52336. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Gruss, Richard; Morgado, Daniel; Craun, Nate; Shea-Blymyer, Colin (2015). *OutbreakSum: Automatic Summarization of Texts Relating to Disease Outbreaks*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/51133. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Wang, Xiangwen; Chandrasekar, Prashant (2015). *Reducing Noise for IDEAL*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/52340. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Harb, Islam; Jin, Yilong; Cedeno, Vanessa; Mallampati, Sai Ravi Kiran; Bulusu, Bhaskara Srinivasa Bharadwaj (2015). *Social Network Project for IDEAL*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/52264. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Gruss, Richard; Choudhury, Ananya; Komawar, Nikhil (2015). *Solr Team Project Report*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/52265. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Plahn, Jordan; Zamani, Michael; Lee, Hayden; Trujillo, Michael (2015). *Summarizing Fire Events with Natural Language Processing*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/51131. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Leskanic, Tyler; Kays, Kevin; Maier, Emily; Cannon, Seth (2015). *Tracking FEMA*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/52871. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Conley, Chris; Druckenbrod, Alex; Meyer, Karl; Muggleworth, Samuel; Fox, Edward A. (2015). *Tweets Metadata*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/52363. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Xuan, Zhang; Wei, Huang; Ji, Wang; Tianyu, Geng (2014). *Unsupervised Event Extraction from News and Twitter*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/47954. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Patents

Technologies or Techniques

Thesis/Dissertations

Tarek Ghaze Kanan. *Arabic News Text Classification and Summarization: A Case of the Electronic Library Institute SeerQ (ELISQ)*. (2015). Virginia Polytechnic Institute and State University. Acknowledgement of Federal Support = Yes

Seungwon Yang. *Automatic Identification of Topic Tags from Texts Based on Expansion-Extraction Approach, http://hdl.handle.net/10919/25111*. (2014). Virginia Polytechnic Institute and State University. Acknowledgement of Federal Support = Yes

Websites

DL-VT416: A Digital Library Testbed for Research Related to 4/16/2007 at Virginia Tech
http://dl-vt-416.org/

Our current website, for IDEAL, also can be reached with this URL, since the IDEAL site also includes results from NSF IIS-0736055: SGER: DL-VT416: A Digital Library Testbed for Research Related to 4/16/2007 at Virginia Tech, $199,993+REU Supplement, PI Edward A. Fox, Co-PIs: Christopher L. North, Donald J. Shoemaker, Naren Ramakrishnan, Weiguo Fan, August 15, 2007 - July 31, 2008. That prequel project led first to the CTRnet and then the IDEAL project, so all three projects now share the same website. To ensure continuity, this URL, as well as http://www.ctrnet.net/, all resolve to eventsarchive.org

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Fox, Edward	PD/PI	1
Hanna, Kristine	Co PD/PI	0
Kavanaugh, Andrea	Co PD/PI	1
Sheetz, Steven	Co PD/PI	1
Shoemaker, Donald	Co PD/PI	1
Sandoval-Almazan, Rodrigo	Faculty	0
Skandrani, Hamida	Faculty	0

Name	Most Senior Project Role	Nearest Person Month Worked
Xie, Zhiwu	Faculty	0
Al-Daihani, Sultan	Postdoctoral (scholar, fellow or other postdoctoral position)	0
Yang, Senugwon	Postdoctoral (scholar, fellow or other postdoctoral position)	0
Mansour, Riham	Other Professional	0
Farag, Mohamed	Graduate Student (research assistant)	7
Lee, Sunshin	Graduate Student (research assistant)	7
Ayoub, Souleiman	Undergraduate Student	1
Cummins, John	Undergraduate Student	0
Eyler, Megan	Undergraduate Student	0
Sebastian, Joseph	Undergraduate Student	0
Kanan, Tarek	Other	1
Sun, Yue	Other	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, supervising and participating in all key project activities

Funding Support: This project

International Collaboration: No
International Travel: No

Kristine Hanna
Email: kristine@archive.org
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 0

Contribution to the Project: Kristine Hanna served as our contact at Internet Archive. Paperwork is being submitted to change the Internet Archive Co-PI to Jefferson Bailey, who will be the new contact replacing Kristine Hanna.

Funding Support: A sub-award to Internet Archive from this project provided related assistance.

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 leading community and digital government aspects of the project, as well as surveys related to elections and other political events.

Funding Support: This project

International Collaboration: No
International Travel: No

Steven D Sheetz
Email: sheetz@vt.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Co-PI helping with ontology and database work, including regarding tweets. Assisting with surveys and analysis of survey data, and preparation of related publications.

Funding Support: This project

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 taking the lead in work related to Sociology and related disciplines. Serving as main liaison to the social science and humanities groups interested in our data and in support for access, analysis, and reporting. Aiding especially regarding shootings.

Funding Support: This project

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: Professor in Autonomous U. of Mexico State, Administration & Accounting Faculty, Toluca, Mexico. Collaborating regarding surveys, analyses, and publications connected with events in Mexico.

Funding Support: N/A

International Collaboration: Yes, Mexico
International Travel: No

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

Contribution to the Project: Professor at Université de la Manouba, Département de Gestion, Tunisia, in Marketing. Collaborating regarding surveys, analyses, and publications connected with events in Tunisia.

Funding Support: N/A

International Collaboration: Yes, Tunisia
International Travel: No

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

Contribution to the Project: Ph.D. member of Library faculty, collaborating on related Web archiving projects, and serving as co-chair of WADL 2015.

Funding Support: N/A

International Collaboration: No
International Travel: No

Sultan M. Al-Daihani
Email: s.aldaihani@ku.edu.kw
Most Senior Project Role: Postdoctoral (scholar, fellow or other postdoctoral position)
Nearest Person Month Worked: 0

Contribution to the Project: Visiting scholar working in the Digital Library Research Laboratory at Virginia Tech for his sabbatical, attending project meetings and working with graduate students on issues related to Arabic and/or Library/Information Science

---

Funding Support: N/A

International Collaboration: Yes, Kuwait
International Travel: No

Senugwon Yang
Email: seungwon@vt.edu
Most Senior Project Role: Postdoctoral (scholar, fellow or other postdoctoral position)
Nearest Person Month Worked: 0

Contribution to the Project: Helped developed topic extraction prototype, Xpantrac, which is the subject of his Virginia Tech dissertation completed before this year of funding, after which he served as a postdoc at GMU, though still helping with IDEAL. Working with the PI, he submitted a journal paper on XPANTRAC for publication. He also supervised the Tracking FEMA project in Spring 2015 in CS4624.

Funding Support: N/A

International Collaboration: No
International Travel: No

Riham Hassan Abdel-Moneim Mansour
Email: rihamma@microsoft.com
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 0

Contribution to the Project: Helped in research, design, and analysis of the survey about social media use during the Egyptian uprising. Helped supervise the work of Mohamed Farag. Earlier, at the Arab Academy of Science and Technology in Cairo, and having visiting Virginia Tech a few years ago and worked with the NSF funded project team, she recruited another Egyptian collaborator (Prof. Hicham Elmongui, of the University of Alexandria).

Funding Support: N/A

International Collaboration: Yes, Egypt
International Travel: No

Mohamed Magdy Farag
Email: mmagdy@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 7

Contribution to the Project: Developed prototypes that demonstrate research goals and ideas. Helped guide class projects related to IDEAL during the fall and spring. Helped with webpage collections and their processing. Building upon his doctoral proposal, has been carrying out the planned research related to IDEAL. Has made progress especially with regard to event modeling and focused crawling.

Funding Support: NSF

International Collaboration: No

International Travel: No

Sunshin Lee
Email: sslee777@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 7

Contribution to the Project: Developed prototypes that demonstrate research ideas and goals of the project. Led work on tweet collection. Guided undergraduate research students and multiple undergraduate class projects. Led the planning, ordering, construction, software setup, and operation of the Hadoop cluster. Building upon his doctoral proposal, has been carrying out planned research related to IDEAL. Has made progress on inferring locations implicit in tweet content.

Funding Support: NSF

International Collaboration: No
International Travel: No

Souleiman Ayoub
Email: siayoub@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: In addition to working on summarizing our community event collection in CS4984, Computational Linguistics, in fall 2014, in 2015 he worked, in CS4624 and in two undergraduate research courses, on Arabic natural language processing, helping also with resulting publications.

Funding Support: N/A

International Collaboration: No
International Travel: No

John Alex Cummins
Email: jcvt@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 0

Contribution to the Project: Volunteered help with the Hadoop cluster, tweet collection, and related management of data. Developed a template to help summarize fire events.

Funding Support: N/A

International Collaboration: No
International Travel: No

Megan Eyler
Email: emegan@vt.edu

Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 0

Contribution to the Project: As an undergraduate student in Sociology, volunteered and worked with co-PI Dr. Donald Shoemaker. She assisted especially with regard to school shooting events, presenting a poster at a local symposium.

Funding Support: N/A

International Collaboration: No
International Travel: No

Joseph Braeden Sebastian
Email: jbraeden@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 0

Contribution to the Project: Volunteered to upload web and tweet archives to Hadoop cluster. Developed template to describe flood events.

Funding Support: N/A

International Collaboration: No
International Travel: No

Tarek G. Kanan
Email: tarekk@vt.edu
Most Senior Project Role: Other
Nearest Person Month Worked: 1

Contribution to the Project: Doctoral student, helping prepare for the fall Computational Linguistics course that will use IDEAL collections, and assisting with search and Arabic research, using SOLR and other tools

Funding Support: Qatar

International Collaboration: Yes, Qatar
International Travel: No

Yue Sun
Email: syue88@vt.edu
Most Senior Project Role: Other
Nearest Person Month Worked: 0

Contribution to the Project: Doctoral student, helping analyze survey data

Funding Support: N/A

International Collaboration: No
International Travel: No

What other organizations have been involved as partners?

Name	Type of Partner Organization	Location
Alexandria University	Academic Institution	Egypt
Autonomous University of the State of Mexico, Toluca	Academic Institution	Toluca, Mexico
High Institute of Management of Tunis	Academic Institution	Tunis, Tunisia
Internet Archive	Other Nonprofits	San Francisco, CA, USA
University of the Philippines, Diliman	Academic Institution	Philippines
Université de la Manouba	Academic Institution	Manouba, Tunisia

Full details of organizations that have been involved as partners:

Alexandria University

Organization Type: Academic Institution
Organization Location: Egypt

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: 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.

Autonomous University of the State of Mexico, Toluca

Organization Type: Academic Institution
Organization Location: Toluca, Mexico

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: 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

High Institute of Management of Tunis

Organization Type: Academic Institution
Organization Location: Tunis, Tunisia

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: collaborated on survey research on social media and information/communication technology use in Tunisia during and since the political crisis of the revolution.

Internet Archive

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

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

More Detail on Partner and Contribution: The project team is using IA's Archive-It service, specifically the Heritrix crawler and the Wayback machine, for webpage archiving tasks.

University of the Philippines, Diliman

Organization Type: Academic Institution
Organization Location: Philippines

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Collaboration as discussed in the writeup on Dr. Shoemaker's work, including his 3 month visit to the Philippines in the spring.

Université de la Manouba

Organization Type: Academic Institution
Organization Location: Manouba, Tunisia

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Foreign partner helping with data collection, analysis, and publication related to events in Tunisia.

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?

1. The IDEAL team extended the Hadoop cluster that is used for storing, extracting, analyzing, and visualizing

Web and tweet collections. The Hadoop cluster will help speed up processing and analyzing stored collections. This should be an exemplar demonstration of how low cost equipment can support this type of research, and should guide others with similar interests or with related big data applications.
2. A novel technique, event archive analysis, for collecting, storing, and analyzing webpages about a specific event was developed. The developed technique will help automate the process of collecting, storing, analyzing, and summarizing Web collections as well as preparing high quality archives. This technique is tightly coupled with our advanced focused crawler, since both include models of events, as well as methods to build and utilize those models.
3. We have advanced the teaching of information retrieval and computational linguistics by developing a new problem/project-based learning approach closely tied to the IDEAL project, that engages students in each of those courses in research as well as in learning and applying state-of-the-art techniques and technologies in the big data area.
4. We developed an improved technique for location disambiguation in tweets that makes use of tailored knowledge bases we constructed, as well as natural language processing, machine learning, and GIS-related methods.

What is the impact on other disciplines?

1. We have supported related work in library and information science, digital humanities, social sciences, and additional fields (see above), engaging many who are interested in working with our collections. Others are likely to use our collections and services as those expand, and as we further disseminate results and extend our collaboration.
2. Our project activities and findings are being included in a new sociology book (Donald J. Shoemaker and Timothy W. Wolfe, Juvenile Justice: A Reference Handbook, second edition, in press with ABC-CLIO), as well as a related course: Sociology 4424 – Juvenile Delinquency.

What is the impact on the development of human resources?

1. At Virginia Tech, many students learned about and from this project. Students in the spring 2015 offerings of CS4624 (Multimedia, Hypertext, and Information Access) and CS5604 (Information Storage and Retrieval) carried out term projects in groups related to event detection, reporting, developing tweet metadata standards, and prototyping a novel framework for a search engine integrating multiple sources of information. Students in the Fall 2014 offering of CS4984 (Computational Linguistics) also carried out projects in groups to produce meaningful English summaries of small and big-sized tweet and web collections. Multiple student reports were prepared and made globally accessible through VTechWorks, each listed under Other Publications. Each submission includes final presentation slides, final report, and all other appropriate deliverables. Students learned not only about related topics in computational linguistics and connected big data issues, but also how to prepare high quality content and upload it to VTechWorks for sharing with others who are interested in learning about these matters.
2. Also at Virginia Tech, three other students carried out undergraduate research over the last year. Joseph Sebastian and Alex Cummins worked on preparing a big data framework for ingesting, storing, and organizing tweet collection into our Hadoop cluster, as well as on collection summary templates. Another student, Souleiman Ayoub, worked on preparing meaningful summaries from news articles.

What is the impact on physical resources that form infrastructure?

1. At Virginia Tech, a 21 node Hadoop cluster was set up to aid our project. In the fall of 2014 a smaller version of it was used in the Computational Linguistics course. It then was expanded and used in 2015 in the Information Storage and Retrieval course as well as the Multimedia, Hypertext, and Information Access course.
2. Other computers, servers, and virtual machines have been deployed to help with the research and services provided through the IDEAL project

What is the impact on institutional resources that form infrastructure?

The IDEAL project supported and aided a variety of Digital Humanities efforts on campus, especially in the College of Liberal Arts and Human Studies. This led to multiple collaborative projects. Also the IDEAL project helped support activities of Virginia Tech's Center for Peace Studies and Violoence Prevention.

What is the impact on information resources that form infrastructure?

There is ongoing work to gather both tweets and webpages to be assembled into collections that will be analyzed, summarized, and made accessible. There are already more than 600 tweet collections related to over 200 events, over 65 webpage collections at the Internet Archive, and over 11 TB of webpages being categorized according to event type and event instance.

What is the impact on technology transfer?

1. Educational modules were developed that facilitate learning about subjects related to the IDEAL project, regarding big data, computational linguistics, information retrieval, and machine learinng.
2. The IDEAL project started collaboration with Altiscale and Cloudera regarding support for distributed processing using Hadoop clusters.

What is the impact on society beyond science and technology?

Public users can access web collections through the IDEAL project website and the IA website. Services are provided that help the public search and browse information about events in our collections. Stakeholder groups also can benefit, that have interest in crises, disasters, tragedies, recovery, and other events (related to communities and governments), by requesting that we start collecting tweets and webpages about events they identify; we also collaborate with such groups to aid in their research.

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
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.