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 Jefferson J Bailey, 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 - 08/31/2017
Reporting Period:	09/01/2015 - 08/31/2016
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. The Digital Library Research Laboratory (DLRL) Hadoop cluster was enhanced and utilized to support many of these efforts.

2.We collected tweets and webpages about hundreds of events (community, disaster, and governmental). Web collections were processed using software including Python and Nutch. Tweet collections were stored on local servers. Our focused crawler research led to new software for leveraging tweets to efficiently find additional relevant webpages.

3. Collections were indexed and made available for searching, browsing, and other services.

4. We led publication from the Web Archiving and Digital Libraries workshop in 2015 for an IEEE TCDL Bulletin issue, and organized and ran WADL 2016, which will lead to another Bulletin issue. We led work on a special issue on this topic of the International Journal on Digital Libraries, to appear within the next year.

5. In addition to a broad range of other publications and presentations, reports from projects related to IDEAL in two courses have been prepared: CS5604: Information Storage and Retrieval - Spring 2016 - 20 students, all working with IDEAL data, software, and systems; CS4624: Multimedia, Hypertext, and Information Access - Spring 2016 - 56 students exposed, with 7 working on IDEAL-related term projects.

6. We disseminated results at the 2015 Virginia Science Festival, also collecting data to better understand public interest in important events, see http://wordpress.cs.vt.edu/csblog/2015/10/05/cs-department-represented-at-the-virginia-science-festival/

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 years.

2. 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. This year we used Spark to extend our prior work with MapReduce, so as to allow fast continuous updating.

3. Two closely related dissertations by project research assistants were pursued.

4. Advisory meeting guidance aided project planning and dissemination.

Significant Results:

1. We have built tweet collections for more than 500 events (disasters, crises, communities, and political activities). A list of tweet collections and analysis services for the collections can be found at http://hadoop.dlib.vt.edu. For some of these we have derived web collections. The Collection (Archiving) DB lists roughly 900 tweet collections, with over 1.2B tweets, giving the keyword / hashtag information involved in the collecting. The GETAR-related collection lists 315 collections including global warming, climate change, and energy issues, with over 60M tweets. We also have 1% sampling of tweets, with over 65M tweets.

2. Using an event focused crawler, we have built higher precision web collections about many events. Evaluation studies have validated the effectiveness of the methods using recent events: Orlando shooting, Ecuador earthquake, Panama papers, California shooting, Brussels attack, Paris attack, and Oregon 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. Our Hadoop cluster was expanded to now have more than 150TB of storage, as well as better support for searching. Cloudera Hadoop software was updated to a newer version (5.6.0) to support multiple projects, and students in classes as well as other volunteers have both learned and helped make it useful for IDEAL.

Key outcomes or Other achievements:

1. The connection of classes with our cluster and the IDEAL project led to PI Fox receiving a 2016 XCaliber Award "for making extraordinary contributions to technology enriched active learning", see https://vtnews.vt.edu/articles/2016/04/fs-edwardfox.html

2. 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).

3. Other publications and presentations have helped disseminate results and helped expand our collaboration with partners and stakeholders.

4. We initiated an exploratory collection and analysis of tweets about Saudi youth employment search in collaboration with VT Economics faculty member Djavad Salehi-Isfahani, CS Ph.D. student Liuqing Li, and CS undergraduate Azizabdul Alayadi. This collection includes all the tweets of the 5000 most active followers of the Saudi Human Resources and Development Foundation (HRDF) Twitter account.

*** 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 Hadoop cluster.

2. Two courses (1 undergraduate and 1 graduate, see above) included class presentations and term projects related to the IDEAL project. Two students completed independent studies (one graduate and one undergraduate), focused on helping with IDEAL.

3. Megan Eyler, an undergraduate student in Sociology, volunteered and worked with co-PI Dr. Donald Shoemaker, as well as students involved in the project.

4. Ph.D. student Yue Sun volunteered and worked with co-PI Dr. Andrea Kavanaugh and co-PI Steven Sheetz on survey data analysis related to political crises in Mexico and Tunisia.

5. Visiting scholar Dr. Denilson Alves Pereira, a professor at Departamento de Ciência da Computação da Universidade Federal de Lavras, Brasil, collaborated with the project team during many meetings, discussions, and research explorations. He also helped to classify tweets' locations using reverse geo-coded tweets with his many-class classifier.

6. M.S. student Ziqian Song worked with co-PI Dr. Andrea Kavanaugh on the analysis of Twitter data related to community and government in the New River Valley region surrounding Virginia Tech.

7. Two project GRAs, Mohamed Farag and Sunshin Lee, made substantial progress on their doctoral dissertations, that are closely related to this project. One was completed in August 2016 and the other should be finalized late in 2016.

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

1. All told, there are 33 products in our list this year, including some works listed last year for which the status has progressed.

2. The IDEAL project has provided publicly available information and pointers for previous and current Web collections, and summaries for tweet collections.

3. A poster was presented at the JCDL 2016 conference, covering our evaluation of traditional approaches to Web archiving, showing the need for our methods. A general tutorial on digital libraries also was presented there by the PI, that included discussion of IDEAL as a case study.

4. We organized and ran the Web Archiving and Digital Library (WADL 2016) workshop at JCDL 2016.

5. We presented our research on intelligent event focused crawling (and have an extended abstract in the preliminary proceedings) at WADL 2016. We also had a poster there about our cluster processing.

6. We have 3 journal articles and one conference publication related to our analyses about important world events. One book, on juvenile justice, was published, including a discussion of our project on page 6.

*** 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. In addition, we plan to collaborate with GWU and deploy their Social Feed Manager tool to distribute the work on collecting tweets and related social media content. After subsequent processing, data from these efforts will aid our collecting of webpages about those events, leveraging our work on event focused crawling.

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 three 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. This will connect with the NSF-funded Global Event and Trend Archive Research (GETAR) project, soon to start, that will involve some 30 other interested investigators.

5. Working with the Internet Archive and others, we will further advance technology transfer of our research findings.

6. Project GRA Sunshin Lee will soon complete his IDEAL-related dissertation.

7. We will enhance our support for searching and browsing by improving the hardware and software related to Solr in our cluster. We also will connect with the Blacklight family of interfaces to improve usability.

Products

Books

Donald J. Shoemaker and Timothy W. Wolfe (2016). *Handbook of Juvenile Justice. 2nd ed.* ABC-CLIO. Santa Barbara, CA. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; ISBN: 978-1-4408-4074-6

Book Chapters

Inventions

Journals or Juried Conference Papers

Andrea Kavanaugh, Steven D. Sheetz, Hamida Skandrani, John C. Tedesco, Yue Sun, and Edward A. Fox (2016). The Use and Impact of Social Media during the 2011 Tunisian Revolution. *17th International Digital Government Research Conference (dg.o 2016), Fudan University, China, June 8-10, 2016, Yushim Kim and Monica Liu (Eds.)*. 20. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: http://dx.doi.org/10.1145/2912160.2912175

Edward A. Fox (2016). Introduction to Digital Libraries. *ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2016, http://www.jcdl2016.org/), Rutgers Univ., Newark, NJ*. 283. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: http://dx.doi.org/10.1145/2910896.2925429

Edward A. Fox, Zhiwu Xie, and Martin Klein (2016). WADL 2016: Third International Workshop on Web Archiving and Digital Libraries. *ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2016, http://www.jcdl2016.org/), Rutgers Univ., Newark, NJ*. 293. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: http://dx.doi.org/10.1145/2910896.2926735

Kavanaugh, A., Sheetz, S. Skandrani, H., Sun, Y., Tedesco, J. and Fox, E. (2016). The Use and Impact of Social Media during the 2011 Tunisian Revolution. *Information Polity*. . Status = ACCEPTED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Kavanaugh, A., Sheetz, S., Sandoval-Almazan, R., Tedesco, J., and Fox, E. (2016). Media Use during Conflicts: Information Seeking and Political Efficacy during the 2012 Mexican Elections. *Government Information Quarterly*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: http://dx.doi.org/10.1016/j.giq.2016.01.004

Mohamed Farag and Edward A. Fox (2016). Which webpage should we crawl first? Social media-based webpage source importance guidance. *WADL 2016: Third International Workshop on Web Archiving and Digital Libraries, Rutgers Univ., Newark, NJ*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER:

Mohamed Farag, Pranav Nakate and Edward A. Fox (2016). Big Data Processing of School Shooting Archives. *ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2016, http://www.jcdl2016.org/), Rutgers Univ., Newark, NJ, June 19-23, 2016*. 271. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: http://dx.doi.org/10.1145/2910896.2925466

Mohamed Magdy, Sunshin Lee, Edward Fox (2016). Focused Crawling for Events. *International Journal on Digital Libraries (IJDL)*. . Status = UNDER_REVIEW; Acknowledgment of Federal Support = Yes ; Peer Reviewed = 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. *Journal of Computing Sciences in Colleges*. 31 (2), 260. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; ISSN: 1937-4771

Tarek Kanan and Edward A. Fox (2016). 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 = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: http://dx.doi.org/0.1002/asi.23609

Licenses

Other Conference Presentations / Papers

Edward A. Fox, Mohamed Farag, Sunshin Lee, Xuan Zhang, Richard Gruss (2016). *Conversation: Problem/project-based Learning with Big Data*. 2016 Conference on Higher Education Pedagogy. Virginia Tech, Blacksburg, VA. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Edward A. Fox and the IDEAL team (2015). *Demonstration: IDEAL (Integrated Digital Event Archive & Library)*. Virginia Science Festival, Sept. 26, 2015. Virginia Tech, Blacksburg, VA. 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 = PUBLISHED; Acknowledgement of Federal Support = Yes

Sunshin Lee, Mohamed Magdy, Edward A. Fox (2015). *IDEAL: Integrated Digital Event Archiving and Library*. Center for Human Computer Interaction (CHCI) 20-Year Celebration Conference. Virginia Tech, Blacksburg, VA. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Sunshin Lee and Edward A. Fox (2016). *Poster: Archiving and Analyzing Tweets and Webpages with the DLRL Hadoop Cluster*. WADL 2016: Third International Workshop on Web Archiving and Digital Libraries, June 22-23, 2016. Rutgers Univ., Newark, NJ. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Sunshin Lee, Mohamed Magdy, Richard Gruss, Tarek Kanan, Xuan Zhang, and Edward A. Fox (2016). *Poster: Enhanced problem-based learning connecting big data research with classes*. HPC Day 2016, by Virginia Tech's Advanced Research Computing. 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*. 29th Annual Consortium for Computing in Small Colleges: Southeastern Conference (CCSC:SE). Roanoke College, Salem, VA. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Kavanaugh, A., Sheetz, S., Skandrani, H., Tedesco, J., and Fox, E (2016). *The Impact of Information Sources on Political Information Efficacy in Tunisia: A Case Study of the 2014 Elections*. 50th Annual Meeting of the Middle East Studies Association (MESA 2016), November 17-20, 2016. Boston, MA. Status = ACCEPTED; Acknowledgement of Federal Support = Yes

Other Products

Other Publications

Pranav Nakate (2016). *Big Data Processing of School Shooting Archives*. Independent study for MS, with report at "http://eventsarchive.org/sites/default/files/PranavNakate20160111Big Data Processing of School Shooting Archives - Report.docx" and slides at "http://eventsarchive.org/sites/default/files/PranavNakate20160111Big Data Processing of School Shooting Archives - Presentation.pptx". Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Masiane, Moeti; Warren, Lawrence (2016). *CS5604 Front-End User Interface Team*. Virginia Tech, CS5604 Team Term Project Report website, Blacksburg, VA, http://hdl.handle.net/10919/70935. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Vishwasrao, Saket; Thorve, Swapna; Tang, Lijie (2016). *CS5604: Clustering and Social Networks for IDEAL*. Virginia Tech, CS5604 Team Term Project Report website, Blacksburg, VA, http://hdl.handle.net/10919/70947. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Bock, Matthew; Cantrell, Michael; Shahin, Hossameldin (2016). *Classification Project in CS5604, Spring 2016*. Virginia Tech, CS5604 Team Term Project Report website, Blacksburg, VA, http://hdl.handle.net/10919/70929. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Li, Tianyi; Nakate, Pranav; Song, Ziqian (2016). *Collaborative Filtering for IDEAL*. Virginia Tech, CS5604 Team Term Project Report website, Blacksburg, VA, http://hdl.handle.net/10919/70948. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Ma, Yufeng; Nan, Dong (2016). *Collection Management for IDEAL*. Virginia Tech, CS5604 Team Term Project Report website, Blacksburg, VA, http://hdl.handle.net/10919/70930. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Won, Stephen (2016). *IDEAL Tweet Collection Categorization*. Virginia Tech, CS4624 Team Term Project Report website, Blacksburg, VA, http://hdl.handle.net/10919/70964. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Biondi, Luciano; Walker, Omavi; Yeshiwas, Dagmawi (2016). *IDEALvr Word Cloud: IDEAL Data Visualization using Virtual Reality*. Virginia Tech, CS4624 Team Term Project Report website, Blacksburg, VA, http://hdl.handle.net/10919/70931. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Kafley, Somn; Steele, Derek; Singh, Samyak (2016). *Searchable IDEAL Climate Change Collections*. Virginia Tech, CS4624 Team Term Project Report website, Blacksburg, VA, http://hdl.handle.net/10919/70936. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Xia, Long; Jiang, Tingting; Galad, Andrej; Maharshi, Shivam (2016). *Solr Project with IDEAL, in CS5604 (Information Storage and Retrieval)*. Virginia Tech, CS5604 Team Term Project Report website, Blacksburg, VA, http://hdl.handle.net/10919/70928. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Mehta, Sneha; Vinayagam, Radha Krishnan (2016). *Topic Analysis project in CS5604, Spring 2016: Extracting Topics from Tweets and Webpages for IDEAL*. Virginia Tech, CS5604 Team Term Project Report website, Blacksburg, VA, http://hdl.handle.net/10919/70933. 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

Mohamed Magdy Gharib Farag. *Intelligent Event Focused Crawler*. (2016). Virginia Tech. Acknowledgement of Federal Support = Yes

Websites

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Fox, Edward	PD/PI	1
Bailey, Jefferson	Co PD/PI	0
Kavanaugh, Andrea	Co PD/PI	1
Sheetz, Steven	Co PD/PI	1
Shoemaker, Donald	Co PD/PI	1
Bohland, James	Faculty	0
Logan, Nneka	Faculty	0
Murray-Tuite, Pamela	Faculty	0

Nicholls, Natsuko	Faculty	0
Pereira, Denilson	Faculty	0
Salehi-Isfahani, Djavad	Faculty	0
Sandoval-Almazan, Rodrigo	Faculty	0
Sforza, Peter	Faculty	0
Skandrani, Hamida	Faculty	0
Tedesco, John	Faculty	0
Xie, Zhiwu	Faculty	0
Yang, Senugwon	Faculty	0
Mansour, Riham	Other Professional	0
Mather, Paul	Other Professional	1
Newman, Joseph	Other Professional	0
Bock, Matthew	Graduate Student (research assistant)	0
Farag, Mohamed	Graduate Student (research assistant)	7
Kanan, Tarek	Graduate Student (research assistant)	0
Lee, Sunshin	Graduate Student (research assistant)	7
Li, Liuqing	Graduate Student (research assistant)	0
Nakate, Pranav	Graduate Student (research assistant)	1
Song, Ziqian	Graduate Student (research assistant)	0
Sun, Yue	Graduate Student (research assistant)	0
Alayadi, Abdulaziz	Undergraduate Student	1
Almzayyen, Abdalsalam	Undergraduate Student	0
Ayoub, Souleiman	Undergraduate Student	0
Chon, Jieun	Undergraduate Student	0
Eyler, Megan	Undergraduate Student	0
Won, Stephen	Undergraduate Student	1

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

Jefferson J Bailey
Email: jefferson@archive.org
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 0

Contribution to the Project: Served as our contact at Internet Archive, providing and receiving data from VT, and facilitated VT use of Archive-It services and systems. Researched crawling and focused crawling, as well as better support for researchers using Internet Archive data.

Funding Support: Collaborative funding from NSF to Internet Archive

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 research related to community and digital government aspects of the project, as well as political crises, such as conflict around elections in Mexico and the revolutions in Tunisia and Egypt. Data include Twitter collections, webpages of news and information, and citizen surveys about use and impact of information sources and technologies during and after political events or crises.

Funding Support: This project

International Collaboration: Yes, Egypt, Iran (Islamic Republic Of), Mexico, Tunisia
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

James Bohland
Email: jayjon@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Serving as liaison between IDEAL and the Global Forum for Urban and Regional Resilience, hosting our presentation, exploring mutual collaboration with GFURR personnel

Funding Support: N/A

International Collaboration: No
International Travel: No

Nneka Logan
Email: nlogan@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Helped with curating the Starbucks's "#RaceTogether" tweet collection

Funding Support: Virginia Tech

International Collaboration: No
International Travel: No

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

Contribution to the Project: Helped curate a collection related to transportation problems and concerns

Funding Support: N/A

International Collaboration: No
International Travel: No

Natsuko Nicholls
Email: nnatsuko@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaborating regarding Library curation of collections and liaising with faculty around campus interested in IDEAL

Funding Support: VT University Libraries

International Collaboration: No
International Travel: No

Denilson Alves Pereira
Email: denilson@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Dr. Pereira, visiting for a year from University of Lavras in Brazil, has helped with CS5604 (projects connected with IDEAL) and advised on research connected with entity resolution and disambiguation.

Funding Support: N/A

International Collaboration: No
International Travel: No

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

Contribution to the Project: Exploring how our project could help with understanding the job search behavior of youth in Saudi Arabia, guiding our collection and analysis of tweets posted by the followers and friends of relevant accounts

Funding Support: Support through Harvard University, indirectly from Saudi Arabia

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

Peter M. Sforza
Email: sforza@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaboration with regard to Virginia's DMV car crash data set and our related data. Advised regarding data analysis and publications.

Funding Support: N/A

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: 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

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

Contribution to the Project: Collaborating on surveys, analysis, and publication

Funding Support: N/A

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: Ph.D. member of Library faculty, collaborating on related Web archiving projects, and serving as co-chair of WADL 2016 as well as co-editor of a related IJDL special issue. Implemented an MOU between VT University Libraries to help with cluster expansion. PI on a new IMLS project that will leverage the IDEAL data to guide library infrastructure decision making.

Funding Support: Virginia Tech University Libraries

International Collaboration: No
International Travel: No

Senugwon Yang
Email: seungwon@vt.edu
Most Senior Project Role: Faculty
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, and then starting on the faculty at LSU. Still continues to provide guidance related to his prior and other related work.

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

Paul Mather
Email: pmather@vt.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 1

Contribution to the Project: Helped us with upgrading DLRL cluster including adding 30 4TB HDDs and building Hadoop backup system, as a system engineer employed by Virginia Tech University Libraries

Funding Support: VT University Libraries

International Collaboration: No
International Travel: No

Joseph Newman
Email: jmano7@vt.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 0

Contribution to the Project: Collaboration with regard to Virginia's DMV car crash data set and our related data. Advised regarding data analysis and publications.

Funding Support: N/A

International Collaboration: No
International Travel: No

Matthew Bock
Email: mattb93@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Worked in the classification team in CS5604, continuing that work, especially with regard to improving results using Spark

Funding Support: N/A

International Collaboration: No
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. Helped with webpage collections and their processing. Made significant progress especially with regard to event modeling and focused crawling. Prepared publications and completed dissertation about event focused crawling.

Funding Support: NSF

International Collaboration: No
International Travel: No

Tarek G. Kanan
Email: tarekk@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Doctoral student, assisting with search and Arabic research, using SOLR and other tools. Completed his dissertation and then began teaching in the Software Engineering Department of Al Zaytonah University of Jordan. Continued helping with related publications.

Funding Support: Qatar

International Collaboration: Yes, Qatar
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. Assisted with publications and presentations.

Funding Support: This project

International Collaboration: No
International Travel: No

Liuqing Li
Email: liuqing@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Volunteer to help with project, especially with supporting the upgrading and use of our Hadoop cluster

Funding Support: N/A

International Collaboration: No
International Travel: No

Pranav Nakate
Email: npranav@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 1

Contribution to the Project: Researched about big data processing of School Shooting Archives (Tweets, Webpages) as a independent study

Funding Support: Own funding

International Collaboration: No
International Travel: No

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

Contribution to the Project: Working with co-PI Kavanaugh, analyzed the Virtual Town Square data, helping integrate community content with IDEAL. Participated in Collaborative Filtering team in CS5604, helping with IDEAL and with publishing the related team report.

Funding Support: N/A

International Collaboration: No
International Travel: No

Yue Sun
Email: syue88@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

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

Funding Support: N/A

International Collaboration: No
International Travel: No

Abdulaziz Saleh Alayadi
Email: abdul94@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Worked on the job seeking collection from Saudi Arabia. Identified relevant accounts and relevant tweets. Categorized the tweets for further analysis. Evaluated the Twitter accounts and did a gender analysis using Arabic language texts.

Funding Support: Support through Harvard, flow through originally from Saudi Arabia.

International Collaboration: No
International Travel: No

Abdalsalam Almzayyen
Email: almzayyen@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 0

Contribution to the Project: Volunteer helping with project activities

Funding Support: N/A

International Collaboration: No
International Travel: No

Souleiman Ayoub
Email: siayoub@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 0

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

Jieun Chon
Email: g471000@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 0

Contribution to the Project: Resumed earlier volunteer help with IDEAL, e.g., helping at the Virginia Science Festival

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, and providing guidance afterward too.

Funding Support: N/A

International Collaboration: No
International Travel: No

Stephen Won
Email: step63n1@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Helped with big tweet data collection management and classification. Also assisted with tagging and annotating at the collection level

Funding Support: Own funds

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
Internet Archive	Other Nonprofits	San Francisco, CA, USA
Universite Laval	Academic Institution	Quebec, Canada
University of Tunis - Manouba Campus	Academic Institution	Tunis, Tunisia
University of the Philippines, Diliman	Academic Institution	Philippines

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

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.

Universite Laval

Organization Type: Academic Institution
Organization Location: Quebec, Canada

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Collaboration is led by Sehl Mellouli, full professor, Head of Department of Organizational Information Systems, Faculty of Business Administration, Universite Laval, G1V 0A6, Quebec, Quebec, Canada, Tel: 418-656-2131 (ext. 11449), http://www.fsa.ulaval.ca/personnel/mellouls/

University of Tunis - Manouba Campus

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 the use of social media and other information and communication technology in Tunisia during and since the political crisis of the 2011 revolution.

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 with co-PI Shoemaker in development.

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. PI Fox earned one of VT's two 2016 XCaliber Awards, "for making extraordinary contributions to technology enriched active learning", based on his connecting this project with multiple courses over the 2014-2016 time period, see https://vtnews.vt.edu/articles/2016/04/fs-edwardfox.html. Using problem based learning, with the IDEAL project as the base, leveraging its data collections, and aiming to improve the services being developed for IDEAL, students have learned computational linguistics (NLP) and information retrieval in a much more effective manner. Connecting with the IDEAL project has engaged students in each of the courses in research as well as in learning and applying state-of-the-art techniques and technologies in the big data area.

2. Our research has laid the foundation for further work on event archiving, as well as supporting research on global challenges, using Internet Archive data going back to 1997 as well as data collected at VT since 2007, in a new project: NSF IIS-1619028, III: Small: Collaborative Research: Global Event and Trend Archive Research (GETAR).

3. Through a tutorial at JCDL 2016, and our diverse set of publications and presentations, we have advanced the digital library and Web archiving fields, aiding education as well as research and practice.

4. 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.

5. 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.

6. 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. This should help others connect social media data with locations.

What is the impact on other disciplines?

1. We have supported related work in library and information science, digital humanities, social sciences, economics, global change, and additional fields, 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 by co-PI Donald J. Shoemaker and Timothy W. Wolfe on Juvenile Justice, as well as a related course: Sociology 4424 – Juvenile Delinquency.

3. Our project will help serve as foundation for research related to the power grid, transportation systems, and human adaptation in the new project: NSF CMMI-1638207, CRISP Type 2/Collaborative Research: Coordinated, Behaviorally-Aware Recovery for Transportation and Power Disruptions (CBAR-tpd).

4.Our project will help serve as part of the foundation to guide the development of infrastructure in university libraries in the 2016 project: IMLS LG-71-16-0037-16: Developing Library Cyberinfrastructure Strategy for Big Data Sharing and Reuse

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 2016 offerings of CS4624 (Multimedia, Hypertext, and Information Access) and CS5604 (Information Storage and Retrieval) carried out term projects in groups related to the front-end, clustering, classification, collaborative filtering, collection management, tweet collection categorization and tagging, a VR interface, collections about climate change, Solr, and topic analysis. Ten 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 information retrieval, HCI, 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. In the Fall 2016 offering of CS5604, 6 term project teams are learning by working with IDEAL data and software.

2. Also at Virginia Tech, two other students learned by carrying out independent studies. Pranav Nakate, as part of his MS work, worked with our data related to school shootings, preparing a final report, final presentation, and a conference poster. Stephen Won, for undergraduate research, improved the infrastructure for tweet management, and helped with categorization and tagging of our tweet collections.

3. Mohamed Magdy G. Farag, project GRA, completed his doctoral studies in August 2016, learning and contributing to the understanding of event focused crawling, and has since embarked upon a career in academia.

4. Sunshin Lee, project GRA, passed his research defense in July 2016, advancing his academic work, moving him closer to expected completion of his dissertation later in 2016. He has contributed to knowledge related to improved analysis (geo-location) of the tweets we have collected, as well as tweets in other collections where location can be inferred.

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 both 2015 and 2016 in the Information Storage and Retrieval course as well as the Multimedia, Hypertext, and Information Access course. University Libraries contributed funding so the storage is over 150 terabytes.

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 Violence Prevention. In addition, there has been a very broad collaboration with VT's University Libraries, including an MOU for ongoing collaboration, with the library investing funds and personnel to help expand our Hadoop cluster so our work can be integrated with library services to the VT community and beyond.

What is the impact on information resources that form infrastructure?

1. 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 1.3 billion tweets, spread across more than 1000 tweet collections related to hundreds of events and broader categories, over 65 webpage collections at the Internet Archive, and over 11 TB of webpages being categorized according to event type and event instance.

2. We have advanced research and interest in Web archiving, an important part of global information infrastructure. In particular, we led work on publishing from the 2015 workshop on Web archiving (WADL 2015), and on planning and running WADL 2016. Related, we have led work on a special issue (in process) of the International Journal on Digital Libraries on this same topic,

What is the impact on technology transfer?

1. Educational modules and publications were developed that facilitate learning about subjects related to the IDEAL project, regarding big data, computational linguistics, information retrieval, and machine learinng.

2. In addition to technology transfer between VT and the Internet Archive (and thence to the broad archiving community), 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 Internet Archive 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. More broadly, this effort is preserving tweets and webpages about important events that otherwise might disappear from the historical record.

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.