**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
Major Activities:

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:**

The Integrated Digital Event Archiving and Library (IDEAL) project, with more than 27 collaborators and 7 collaborating institutions, developed tweet and webpage collections, datasets, services, software, systems, and methods. The related activities include: collecting event-related content, software and system development and refinement, experimentation, evaluation, and working with diverse users (representing key stakeholder groups).

The Internet Archive has expanded its collections and technology support, as well as outreach activities. It hosts, preserves, and provides public access with attribution to web collections created by the project team through its public Wayback Machine interface and Archive-It service. The latter may be browsed by descriptive metadata and searched through Archive-It’s newly deployed full-text Elasticsearch engine. New or updated Internet Archive and Archive-It API documentation and workshops provide project stakeholders with several means to query the data from and about these collections, and to derive datasets for further textual and visual analyses.

Studies have proceeded of important events, including integration of survey and analysis approaches, and publishing findings (e.g., about communications, elections, and political events in Mexico and Tunisia).

**Specific Objectives:**

Doctoral dissertation research led to improved methods for focused crawling and for assigning locations to tweets.

More than 22 computers are connected, mostly in a Hadoop cluster. This network was constructed to support collection, processing, and access already to almost 2 billion tweets across over 1300 collections, along with millions of webpages, covering hundreds of important events.

Regarding collections, prior collections were extended, new ones were launched as events occurred or requests were made by users, the event focused crawler was deployed, and diverse related curation efforts proceeded.

Master’s thesis and class (independent study and graduate courses) research led to improved tweet and webpage techniques for content cleaning and processing, information extraction, classification, clustering, topic analysis, sentiment analysis, indexing, searching, browsing, and visualization.

**Significant Results:**

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

The IDEAL project has developed novel methodology and workflows, tailored to addressing the challenging problem of working with events. At a high level, for collection building, is a workflow to collect tweets about each event or event class, extract URLs, use the URLs present therein as seeds to our event focused crawler, and add resulting webpages to our Web collection. The event focused crawler workflow uses the extracted URLs as seeds to construct an event model that guides the selection and focused crawling for webpages.

Key new methods were developed to analyze and accordingly add value (and metadata) to the collected content. The Xpantrac system finds topics in webpages; it generalizes beyond the webpage content through searching, combining and analyzing results, and summarizing/extracting topics. Regarding our processing of tweets, a new framework was devised to streamline a variety of tweet analysis and transformation workflows. Regarding building tweet classifiers for the hundreds of events studied, a learning optimizer method was devised employing iterative processing with minimal human effort to yield high quality classification of tweets into collections for particular real world events. Regarding the problem that few tweets have associated latitude and longitude values, a methodology was devised for associating locations with tweets based on location indicative words.

Collection building and analysis (of both tweets and webpages) has improved through advances in classification, big data workflows, focused crawling (to identify webpages focused on an event of interest), inferring the location of tweets from their text when GPS data is unavailable, topic analysis, and natural language processing (including Arabic).

Insights gained have been shared regarding juvenile delinquency, school shootings, and the use of information during conflicts, crises, elections, and uprisings. Collections are available to support other research and exploration regarding important events since 2007 such as the above, as well as attacks, bombings, celebrations, climate change, collapses, community activities, crashes, disease outbreaks, earthquakes, eclipses, environmental disruptions, erosion, explosions, fires, floods, hurricanes, innovations, judicial decisions, pollution, power outages, protests, revolutions, shootings, sports, storms, summits, tornadoes, transportation failures, tsunamis, typhoons, and veteran activities.

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

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

Dissemination has been through the reported publications and presentations. Further dissemination was through the project website (http://eventsarchive.org) and the website connected to the tweet collections and descriptions (http://hadoop.dlib.vt.edu/). In addition, we led the 2015, 2016, and 2017 Web Archiving and Digital Libraries (WADL) workshops, with related proceedings.

### Products

#### Books

#### Book Chapters

#### Inventions

#### Journals or Juried Conference Papers


**Licenses**

**Other Conference Presentations / Papers**


**Other Products**

*Software or Netware.*

http://hadoop.dlib.vt.edu/ is the website connecting to project tweet collections and related descriptions

*Software or Netware.*

http://www.eventsarchive.org/ is the project website, publicly available over WWW, with title "Events Archiving"

**Other Publications**


Patents

Technologies or Techniques

Thesis/Dissertations


## Participants/Organizations

### What individuals have worked on the project?

<table>
<thead>
<tr>
<th>Name</th>
<th>Most Senior Project Role</th>
<th>Nearest Person Month Worked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fox, Edward</td>
<td>PD/PI</td>
<td>1</td>
</tr>
<tr>
<td>Bailey, Jefferson</td>
<td>Co PD/PI</td>
<td>1</td>
</tr>
<tr>
<td>Kavanaugh, Andrea</td>
<td>Co PD/PI</td>
<td>1</td>
</tr>
<tr>
<td>Sheetz, Steven</td>
<td>Co PD/PI</td>
<td>1</td>
</tr>
<tr>
<td>Shoemaker, Donald</td>
<td>Co PD/PI</td>
<td>1</td>
</tr>
<tr>
<td>Bohland, James</td>
<td>Faculty</td>
<td>0</td>
</tr>
<tr>
<td>Farag, Mohamed</td>
<td>Faculty</td>
<td>0</td>
</tr>
<tr>
<td>Kanan, Tarek</td>
<td>Faculty</td>
<td>0</td>
</tr>
<tr>
<td>Lee, Sunshin</td>
<td>Faculty</td>
<td>1</td>
</tr>
<tr>
<td>Logan, Nneka</td>
<td>Faculty</td>
<td>0</td>
</tr>
<tr>
<td>Murray-Tuite, Pamela</td>
<td>Faculty</td>
<td>0</td>
</tr>
<tr>
<td>Nicholls, Natsuko</td>
<td>Faculty</td>
<td>0</td>
</tr>
<tr>
<td>Pereira, Denilson</td>
<td>Faculty</td>
<td>1</td>
</tr>
<tr>
<td>Salehi-Isfahani, Djavad</td>
<td>Faculty</td>
<td>0</td>
</tr>
<tr>
<td>Sandoval-Almazan, Rodrigo</td>
<td>Faculty</td>
<td>0</td>
</tr>
<tr>
<td>Sforza, Peter</td>
<td>Faculty</td>
<td>0</td>
</tr>
<tr>
<td>Skandrani, Hamida</td>
<td>Faculty</td>
<td>0</td>
</tr>
<tr>
<td>Tedesco, John</td>
<td>Faculty</td>
<td>0</td>
</tr>
<tr>
<td>Name</td>
<td>Most Senior Project Role</td>
<td>Nearest Person Month Worked</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>--------------------------------------------------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>Xie, Zhiwu</td>
<td>Faculty</td>
<td>0</td>
</tr>
<tr>
<td>Yang, Senugwon</td>
<td>Faculty</td>
<td>0</td>
</tr>
<tr>
<td>Chakravarty, Saurabh</td>
<td>Other Professional</td>
<td>1</td>
</tr>
<tr>
<td>Mansour, Riham</td>
<td>Other Professional</td>
<td>0</td>
</tr>
<tr>
<td>Mather, Paul</td>
<td>Other Professional</td>
<td>0</td>
</tr>
<tr>
<td>Bartolome, Abigail</td>
<td>Graduate Student (research assistant)</td>
<td>1</td>
</tr>
<tr>
<td>Bock, Matthew</td>
<td>Graduate Student (research assistant)</td>
<td>1</td>
</tr>
<tr>
<td>Li, Liuqing</td>
<td>Graduate Student (research assistant)</td>
<td>1</td>
</tr>
<tr>
<td>Niu, Shuo</td>
<td>Graduate Student (research assistant)</td>
<td>0</td>
</tr>
<tr>
<td>Song, Ziqian</td>
<td>Graduate Student (research assistant)</td>
<td>0</td>
</tr>
<tr>
<td>Sun, Yue</td>
<td>Graduate Student (research assistant)</td>
<td>0</td>
</tr>
<tr>
<td>Alayadi, Abdulaziz</td>
<td>Undergraduate Student</td>
<td>0</td>
</tr>
<tr>
<td>Chon, Jieun</td>
<td>Undergraduate Student</td>
<td>0</td>
</tr>
<tr>
<td>Eyler, Megan</td>
<td>Undergraduate Student</td>
<td>0</td>
</tr>
<tr>
<td>Won, Stephen</td>
<td>Undergraduate Student</td>
<td>0</td>
</tr>
</tbody>
</table>

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, directing project, teaching students involved, researching, publishing, presenting

Funding Support: this project

International Collaboration: No
International Travel: Yes, Canada - 0 years, 0 months, 6 days
What other organizations have been involved as partners?

<table>
<thead>
<tr>
<th>Name</th>
<th>Type of Partner Organization</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexandria University</td>
<td>Academic Institution</td>
<td>Egypt</td>
</tr>
<tr>
<td>Autonomous University of the State of Mexico, Toluca</td>
<td>Academic Institution</td>
<td>Toluca, Mexico</td>
</tr>
<tr>
<td>George Washington University</td>
<td>Academic Institution</td>
<td>Washington, D.C.</td>
</tr>
<tr>
<td>Internet Archive</td>
<td>Other Nonprofits</td>
<td>San Francisco, CA, USA</td>
</tr>
<tr>
<td>Louisiana State University</td>
<td>Academic Institution</td>
<td>Baton Rouge, LA</td>
</tr>
<tr>
<td>Universite Laval</td>
<td>Academic Institution</td>
<td>Quebec, Canada</td>
</tr>
<tr>
<td>University of Tunis - Manouba Campus</td>
<td>Academic Institution</td>
<td>Tunis, Tunisia</td>
</tr>
<tr>
<td>University of the Philippines, Diliman</td>
<td>Academic Institution</td>
<td>Philippines</td>
</tr>
</tbody>
</table>

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

Impacts

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

The success of repeated offerings of CS5604, Information Retrieval (IR), being run using the pedagogical method of problem/project based learning, with the problem of how to develop an advanced information retrieval system in support of IDEAL, should help others teaching IR to improve the learning in their classes by similar connection to research.

The Xpantrac system demonstrated a promising approach to associating topics with webpages, as an alternative to techniques like LDA. Topic analysis is a key application in IR, and is usually carried out through machine learning methods.

The event focused crawler extends the scope of Web crawling to situations where webpages are sought about an event, rather than about a topic or organization or website.

The methodology of using location indicative words to infer location for tweets that lack latitude and longitude values should expand the utility of tweet collections and related social network studies, by enabling analyses and visualizations that involve locations or geospatial reasoning.

The integration of processing of tweets and webpages, all related to important events, in one system with linked workflows, should broaden the scope of studies that largely just use only one of these two sources for digital library and Web archiving research.

What is the impact on other disciplines?

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

What is the impact on the development of human resources?

IDEAL has led to 4 dissertations, 3 theses, and 49 student reports across 12 offerings in 5 different courses. The application of problem/project based learning has been very popular with students, who are highly motivated, and apply their skills in other courses as well as internships and work after graduation. Students working on the project are now in faculty positions at Louisiana State University and Radford University as well as universities in Egypt and Jordan. A number of those involved are woman, and a number come from underrepresented groups.

The project has involved more than 27 collaborators and 7 collaborating institutions. People in diverse fields have been exposed to advanced data analytics and visualization, enhancing their appreciation of science and understanding about working with data.

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

What is the impact on institutional resources that form infrastructure?

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

What is the impact on information resources that form infrastructure?

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

What is the impact on technology transfer?

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

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

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

What is the impact on society beyond science and technology?

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

Changes/Problems
Changes in approach and reason for change
Nothing to report.

Actual or Anticipated problems or delays and actions or plans to resolve them
Nothing to report.

Changes that have a significant impact on expenditures
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.