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 - Final 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:	09/01/2016 - 08/31/2017
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: 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.

Key outcomes or Other achievements:

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

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

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

In the Fall 2016 class CS5604 (Information Retrieval, IR), the class-wide term project, carried out by students working in six teams (each uploading deliverables into the local institutional repository), was in support of 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

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

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

Edward A. Fox, Martin Klein, and Zhiwu Xie (2017). Guest Editors' Introduction to the Special Issue on Web Archiving. *International Journal on Digital Libraries*. 18 . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = No ; DOI: 10.1007/s00799-016-0203-5

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

Edward A. Fox, Zhiwu Xie, Martin J. Klein (2015). Introduction to the Web Archiving and Digital Libraries 2015 Workshop Issue: Web Archiving and Digital Libraries 2015 (WADL 2015) Overview. *Bulletin of IEEE Technical Committee on Digital Libraries*. 11 (2), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = No ; OTHER: http://www.ieee-tcdl.org/Bulletin/v11n2/papers/intro.pdf

Edward A. Fox, Zhiwu Xie, Martin J. Klein (2017). Web Archiving and Digital Libraries (WADL) 2016: Highlights and Introduction to this Special Issue. *Bulletin of IEEE Technical Committee on Digital Libraries*. 13 (1), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = No ; OTHER: http://www.ieee-tcdl.org /Bulletin/v13n1/papers/intro.pdf

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*. 33 (3), 595-602. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Kavanaugh, A., Sheetz, S., Skandrani, H., and Fox, E. (2017). Media Use by Young Tunisians during the 2011 Revolution vs 2014 Elections. *Information Polity*. 22 ((2017)), 137-158. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.3233/IP-170412

Mohamed Farag and Edward A. Fox (2017). Which webpage should we crawl first? Social media-based webpage source importance guidance. *Bulletin of IEEE Technical Committee on Digital Libraries*. 13 (1), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER: http://www.ieee-tcdl.org /Bulletin//v13n1/papers/farag.pdf

Mohamed M. G. Farag, Edward A. Fox (2015). Building and archiving event web collections: A focused crawler approach. *Bulletin of IEEE Technical Committee on Digital Libraries*. 11 (2), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER: http://www.ieee-tcdl.org/Bulletin/v11n2 /papers/farag.pdf

Mohamed Magdy Gharib Farag, Sunshin Lee, Edward A. Fox (2017). Focused Crawling for Events. *International Journal on Digital Libraries (IJDL)*. 18 1. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1007/s00799-016-0207-1

Saurabh Chakravarty, Eric Williamson and Edward Fox (2017). Classification of Tweets using Augmented Training. *Proc. WADL 2017, a workshop held in conjunction with ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2017, http://2017.jcdl.org/), Toronto, Canada, June 19-23*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Sunshin Lee and Edward A. Fox (2017). Archiving and Analyzing Tweets and Webpages with the DLRL Hadoop Cluster. *Bulletin of IEEE Technical Committee on Digital Libraries*. 13 (1), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER: http://www.ieee-tcdl.org/Bulletin//v13n1/papers/lee.pdf

Licenses

Other Conference Presentations / Papers
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 = OTHER; Acknowledgement of Federal Support = Yes

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
Matthew Bock (2017). *A Framework for Hadoop Based Digital Libraries of Tweets, http://hdl.handle.net/10919/78351*. MS thesis, Virginia Tech. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Saurabh Chakravarty (2017). *A Large Collection Learning Optimizer Framework, http://hdl.handle.net/10919/78302*. MS thesis, Virginia Tech. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

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

Kohler, Rachel; Tasooji, Reza; Sullivan, Patrick (2016). *CS 5604 INFORMATION STORAGE AND RETRIEVAL Front-End Team Fall 2016 Final Report*. Team term project in CS5604, Information Retrieval, Virginia Tech, Department of Computer Science, Blacksburg, VA, http://hdl.handle.net/10919/73711. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Williamson, Eric R.; Chakravarty, Saurabh (2016). *CS5604 Fall 2016 Classification Team Final Report*. Team term project in CS5604, Information Retrieval, Virginia Tech, Department of Computer Science, Blacksburg, VA, http://hdl.handle.net/10919/73713. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Li, Liuqing; Pillai, Anusha; Wang, Ye; Tian, Ke (2016). *CS5604 Fall 2016 Solr Team Project Report*. Team term project in CS5604, Information Retrieval, Virginia Tech, Department of Computer Science, Blacksburg, VA, http://hdl.handle.net/10919/73710. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Wagner, Mitchell J.; Abidi, Faiz; Fan, Shuangfei (2016). *CS5604: Information and Storage Retrieval Fall 2016 - CMT (Collection Management Tweets)*. Team term project in CS5604, Information Retrieval, Virginia Tech, Department of Computer Science, Blacksburg, VA, http://hdl.handle.net/10919/73739. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Li, Liuqing; Harb, Islam; Galad, Andrej (2017). *CS6604 Spring 2017 Global Events Team Project*. Team term project in CS6604, Digital Libraries, Virginia Tech, Department of Computer Science, Blacksburg, VA, http://hdl.handle.net/10919/77867. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Bartolome, Abigail; Islam, MD; Vundekode, Soumya (2016). *Clustering and Topic Analysis in CS 5604 Information Retrieval Fall 2016*. Team term project in CS5604, Information Retrieval, Virginia Tech, Department of Computer Science, Blacksburg, VA, http://hdl.handle.net/10919/73712. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Dao, Tung; Wakeley, Christopher; Weigang, Liu (2017). *Collection Management Webpages - Fall 2016 CS5604*. Team term project in CS5604, Information Retrieval, Virginia Tech, Department of Computer Science, Blacksburg, VA, http://hdl.handle.net/10919/76675. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Sunshin Lee (2017). *Geo-Locating Tweets with Latent Location Information, http://hdl.handle.net/10919/75022*. Ph.D. dissertation, Virginia Tech. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Bartolome, Abigail; Bock, Matthew; Vinayagam, Radha Krishnan; Krishnamurthy, Rahul (2017). *Sentiment and Topic Analysis*. Team term project in CS6604, Digital Libraries, Virginia Tech, Department of Computer Science, Blacksburg, VA, http://hdl.handle.net/10919/77883. 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), http://hdl.handle.net/10919/74272*. (2015). Virginia Polytechnic Institute and State University. Acknowledgement of Federal Support = Yes

Mohamed Magdy Gharib Farag. *Intelligent Event Focused Crawler, http://hdl.handle.net/10919/73035*. (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	1
Kavanaugh, Andrea	Co PD/PI	1
Sheetz, Steven	Co PD/PI	1
Shoemaker, Donald	Co PD/PI	1
Bohland, James	Faculty	0
Farag, Mohamed	Faculty	0
Kanan, Tarek	Faculty	0
Lee, Sunshin	Faculty	1
Logan, Nneka	Faculty	0
Murray-Tuite, Pamela	Faculty	0
Nicholls, Natsuko	Faculty	0
Pereira, Denilson	Faculty	1
Salehi-Isfahani, Djavad	Faculty	0
Sandoval-Almazan, Rodrigo	Faculty	0
Sforza, Peter	Faculty	0
Skandrani, Hamida	Faculty	0
Tedesco, John	Faculty	0

Name	Most Senior Project Role	Nearest Person Month Worked
Xie, Zhiwu	Faculty	0
Yang, Senugwon	Faculty	0
Chakravarty, Saurabh	Other Professional	1
Mansour, Riham	Other Professional	0
Mather, Paul	Other Professional	0
Bartolome, Abigail	Graduate Student (research assistant)	1
Bock, Matthew	Graduate Student (research assistant)	1
Li, Liuqing	Graduate Student (research assistant)	1
Niu, Shuo	Graduate Student (research assistant)	0
Song, Ziqian	Graduate Student (research assistant)	0
Sun, Yue	Graduate Student (research assistant)	0
Alayadi, Abdulaziz	Undergraduate Student	0
Chon, Jieun	Undergraduate Student	0
Eyler, Megan	Undergraduate Student	0
Won, Stephen	Undergraduate Student	0

Full details of individuals who have worked on the project:

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

Contribution to the Project: PI, 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?

Name	Type of Partner Organization	Location
Alexandria University	Academic Institution	Egypt
Autonomous University of the State of Mexico, Toluca	Academic Institution	Toluca, Mexico
George Washington University	Academic Institution	Washington, D.C.
Internet Archive	Other Nonprofits	San Francisco, CA, USA
Louisiana State University	Academic Institution	Baton Rouge, LA
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

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

Impacts

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

The success of repeated offerings of CS5604, Information Retrieval (IR), being run using the pedagogial method of problem/project based learning, with the problem of how to develop an advanced information retrieval system in support of 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 Deparment of Computer Science for our infrastructure, University Libraries has built a very similar infrastructure, and the campus IT groups have launched several clusters to support other similar types of investigations.

What is the impact on information resources that form infrastructure?

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

What is the impact on technology transfer?

The Internet Archive is a partner, working with the 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.