COVER SHEET FOR PROPOSAL TO THE NATIONAL SCIENCE FOUNDATION

PROGRAM ANNOUNCEMENT/SOLICITATION NO./CLOSING DATE/if not in response to a program announcement/solicitation enter NSF 11-1

NSF 12-580 **12/17/12**

FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S) (Indicate the most specific unit known, i.e. program, division, etc.)

IIS - INFO INTEGRATION & INFORMATICS

FOR NSF USE ONLY

NSF PROPOSAL NUMBER

1319578

DATE RECEIVED	NUMBER OF COPIES	DIVISION ASSIGNED	FUND CODE	DUNS# (Data Universal Numbering System)	FILE LOCATION
12/17/2012	1	05020000 IIS	7364	003137015	12/17/2012 2:13pm

EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN)

546001805

SHOW PREVIOUS AWARD NO. IF THIS IS
☐ A RENEWAL
☐ AN ACCOMPLISHMENT-BASED RENEWAL

IS THIS PROPOSAL BEING SUBMITTED TO ANOTHER FEDERAL AGENCY? YES ☐ NO ☒ IF YES, LIST ACRONYM(S)

NAME OF ORGANIZATION TO WHICH AWARD SHOULD BE MADE

Virginia Polytechnic Institute and State University

AWARDEE ORGANIZATION CODE (IF KNOWN)

0037549000

ADDRESS OF AWARDEE ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE

Virginia Polytechnic Institute and State University
1880 Pratt Drive
Blacksburg, VA. 240603580

NAME OF PRIMARY PLACE OF PERF

Virginia Polytechnic Institute and State University

ADDRESS OF PRIMARY PLACE OF PERF, INCLUDING 9 DIGIT ZIP CODE

Virginia Polytechnic Institute and State University

Blacksburg ,VA ,240603580 ,US.

IS AWARDEE ORGANIZATION (Check All That Apply) (See GPG II.C For Definitions)
☐ SMALL BUSINESS ☐ MINORITY BUSINESS ☐ IF THIS IS A PRELIMINARY PROPOSAL
☐ FOR-PROFIT ORGANIZATION ☐ WOMAN-OWNED BUSINESS THEN CHECK HERE

TITLE OF PROPOSED PROJECT **III:Small:Integrated Digital Event Archiving and Library (IDEAL)**

REQUESTED AMOUNT	PROPOSED DURATION (1-60 MONTHS)	REQUESTED STARTING DATE	SHOW RELATED PRELIMINARY PROPOSAL NO. IF APPLICABLE
$ **500,000**	**36** months	**08/10/13**	

CHECK APPROPRIATE BOX(ES) IF THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW

☐ BEGINNING INVESTIGATOR (GPG I.G.2)
☐ DISCLOSURE OF LOBBYING ACTIVITIES (GPG II.C.1.e)
☐ PROPRIETARY & PRIVILEGED INFORMATION (GPG I.D, II.C.1.d)
☐ HISTORIC PLACES (GPG II.C.2.j)
☐ EAGER* (GPG II.D.2) ☐ RAPID** (GPG II.D.1)
☐ VERTEBRATE ANIMALS (GPG II.D.6) IACUC App. Date _____
PHS Animal Welfare Assurance Number _____

☒ HUMAN SUBJECTS (GPG II.D.7) Human Subjects Assurance Number **FWA00000572**
Exemption Subsection _____ or IRB App. Date **Pending**
☐ INTERNATIONAL COOPERATIVE ACTIVITIES: COUNTRY/COUNTRIES INVOLVED (GPG II.C.2.j)
_____
☐ HIGH RESOLUTION GRAPHICS/OTHER GRAPHICS WHERE EXACT COLOR REPRESENTATION IS REQUIRED FOR PROPER INTERPRETATION (GPG I.G.1)

PI/PD DEPARTMENT
Computer Science

PI/PD FAX NUMBER
540-231-6075

PI/PD POSTAL ADDRESS
620 Drillfield Drive
2160G Torgersen Hall (0106)
Blacksburg, VA 24061
United States

NAMES (TYPED)	High Degree	Yr of Degree	Telephone Number	Electronic Mail Address
PI/PD NAME **Edward A Fox**	**PhD**	**1983**	**540-231-5113**	**fox@vt.edu**
CO-PI/PD **Kristine Hanna**	**BA**	**1983**	**415-561-6767**	**kristine@archive.org**
CO-PI/PD **Andrea L Kavanaugh**	**PhD**	**1990**	**540-231-1806**	**kavan@vt.edu**
CO-PI/PD **Steven D Sheetz**	**PhD**	**1996**	**540-231-6096**	**sheetz@vt.edu**
CO-PI/PD **Donald J Shoemaker**	**PhD**	**1970**	**540-231-6046**	**shoemake@vt.edu**

Electronic Signature

III:Small:Integrated Digital Event Archive and Library (IDEAL)
Project Summary

We will research the next generation integration of digital libraries and event archiving. We will prove the effectiveness of the 5S (Societies, Scenarios, Spaces, Structures, Streams) approach to intelligent information systems by crawling and archiving events of broad interest, and providing digital library collections and services supporting the diverse interdisciplinary communities of those interested in better understanding of such events. To demonstrate the generality of our methodology and infrastructure we will focus on events falling into two broad categories: 1) related to crises or tragedies as well as recovery (highlighted in the 2012 NSF/CCC report on "Computing for Disasters", and extending our CTRnet project); 2) government/community related events (e.g., elections, demonstrations, planning meetings, local group activities). Thus, in addition to collaborating with the Internet Archive and its partners, we will connect with those interested in emergency preparedness/response, digital government, and the social sciences.

As digital libraries have evolved over the last twenty years, it has become clear that there are many connections with archives, including Web and Internet archives. Yet, there has been limited integration, leading to inefficiencies, limited support for those studying the past (even the fairly recent past), and permanent loss of access to materials needed to understand our culture, heritage, and history. What is archived usually has low recall and precision, high bias, and is hard to analyze or access. We will build a firm foundation for integration and interoperability, construct a system validating our approach, acquire and add value to useful event-focused collections, and show the utility of a broad range of services, targeted to all of the stakeholder communities, including archivists, librarians, researchers, scholars, and the general public.

We will support automatic event detection as well as accept narrow or general requests for event archiving. We will crawl, collect, filter, categorize, preserve, and provide access to Web pages and tweets. Services will include browsing, searching, recommending, notifying, summarizing, identifying topics and themes, analyzing (e.g., sentiments), and visualizing (text and data).

The **intellectual merit** of this work includes providing a foundation for the integration of the library, archive, and information sciences – and their techniques. We will improve methods for detecting, managing, and utilizing information about events, including topic detection/tracking (using news and other websites as well as Twitter and social networks), intelligent focused crawling, and filtering. Partnering with LucidWorks to handle our big data collection, we will apply and extend access and analysis techniques, as well as techniques for visualization/reporting and retrieval/recommendation. We will evaluate both at the micro (e.g., each technique) and macro levels (e.g., end-to-end system operations and support for communities, tasks, and events).

The **broader impacts** of our work include aiding those needing integrated access to information in government or policy making, as well as the general public – as they address emergencies, or analyze and interpret society through events and their effects. The coverage of the 5S framework ensures that the broadest knowledge of events is sought and organized, and sets the tone for related research, regarding techniques, methods, services, and evaluation.

Keywords: computing for disasters; digital library services; community/crisis/government event archiving; focused crawling; text analysis / visualization for big data; twitter/webpage collections

TABLE OF CONTENTS

For font size and page formatting specifications, see GPG section II.B.2.

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

Integrated Digital Event Archive and Library (IDEAL): Project Description

1.0 Introduction

Events lead to our most poignant memories. We remember birthdays, graduations, holidays, weddings, and other events that mark stages of our life, as well as the lives of family and friends. As a society we remember assassinations, natural disasters, political uprisings, terrorist attacks, and wars – as well as elections, heroic acts, sporting events, and other events that shape community, national, and international opinions. Web and Twitter content describes many of these societal events. A side effect of Web 2.0 [1] is that it provides a highly responsive sensor of occurrences in the real world [2], since people from across the globe meet virtually and share related observations and stories online. Extended digital libraries [3-17] can leverage this stream of data, for automatic detection of events, to trigger event archiving, and later to enable event related services that support communities [18-31]. Permanent storage and access to big data collections of event related digital information, including webpages, tweets, images, videos, and sounds, could lead to an important national asset. Regarding that asset, there is need for digital libraries (DLs) – providing immediate and effective access – and archives – with historical collections that aid science and education, as well as studies related to economic, military, or political advantage [32]. So, to address this important issue, we will research an Integrated Digital Event Archive and Library (IDEAL).

1.1 Motivation

When something notable occurs, many users try to locate the most up-to-date information about that event. Later, researchers, scholars, students, and others seek information about similar events, sometimes for cross-event comparisons or trend analysis. Yet, there is little systematic archiving anywhere of information about events, except when national or state events are captured as part of government related Web archives, or when media companies (e.g., CSPAN) build archives of stories they cover. Further, these archives are not integrated. Thus, descriptions of events are fragmented, ad hoc, and incomplete. Though the Internet Archive (see letter) supports some event-oriented archiving [33], coverage is limited. Many important events are ignored, others only captured in part, and often, late onset of archiving, causes crucial early information to be lost. Further, tools for capture are complex, and few archivists master their features, so achieving high recall is expensive. There are few mechanisms to filter out noise in collections. Access to the resulting archives is awkward and inefficient [34]. Thus, suitable technology is lacking.

From the perspective of supporting historical research, or of preserving a record of modern civilization based on what exists in these data streams, this situation is completely inadequate and unacceptable. From the perspective of research on intelligent information systems, there is a broad range of integration and interoperability [14, 35-51] problems that are both intellectually interesting and have broad impact. So, we will research a digital library supporting automatic event detection, tracking, and preservation. By taking input from queries, tweets, news, and blogs, our system will detect events in a user-oriented manner, archive event related digital objects, and provide a broad range of helpful services [52], building upon our partnership with LucidWorks and their Big Data Software [53] (see letter). To the best of our knowledge this is the first attempt to research and evaluate such an integrated digital library and archive.

1.2 Goal and Objectives

We will research an integrated next generation event archiving DL system, compatible with the Open Archival Information System (OAIS) standard (see Fig. 1). The system will monitor web-based and social media activity to automatically detect interesting events, as well as respond to specific and general event archiving requests. When an event is identified, IDEAL will collect, catalog, preserve, and provide access and services to related digital objects, including multimedia, captured from all corners of the Web. Our

system leverages the 5S (Societies, Scenarios, Spaces, Structures, Streams) DL framework [16, 54-59], so it will have a firm theoretical basis, ensuring a comprehensive, efficient, and effective approach to identify and describe relevant content. To make the work feasible but generalizable, we will concentrate on two categories of events: 1) Crises / Tragedies / Recovery (CTR) activities; 2) Government/community events (including about politics and elections, demonstrations, planning meetings, and activities of community groups). These categories of events fit into the **S**ocieties dimension of the 5S framework [59].

Government activities, such as elections and community planning, persist over time, having relevance across many aspects of **S**ociety. They occur in known **S**paces, e.g., polling places or county supervisor meetings, with stakeholders performing well-known **S**cenarios. Most importantly they generate **S**treams of information that capture the nature of life in modern **s**ociety, yet many quickly become unavailable after interest in the event passes, and are lost. During crises, members of **S**ocieties are called upon to perform **S**cenarios that often are uncommon for them. The **S**tream of information related to crises is different from the **S**tream emerging from normal life **S**cenarios, with dramatic bursts of data immediately surrounding the event and rapid drop off in activity after a response has occurred. Thus, multiple parts of **S**ociety engage in (ab)normal **S**cenarios representing shared and isolated **S**paces, generating differing **S**treams of data. Accordingly, by including both of these categories of events, we will demonstrate the generality of our findings. Further, our approach to events will lead to DL representations in **S**tructures, that underlie services fit to stakeholder needs. To ensure the greatest scientific progress, we will evaluate those representations and services at every step of the way, as well as at the sub-system and system levels, and from a user-centric perspective.

The OAIS based architecture for the IDEAL system is presented in Fig. 1. We will ingest tweets and web-based content from social media and the general Web, including news media, government, and other websites. In addition to archiving materials found, we will build a data management system that includes metadata consistent with the 5S framework, along with results from our intelligent crawler, to support comprehensive access to event related content. With the support of two key partners, the IDEAL team will undertake important research investigations, to achieve three complementary objectives:

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 at least 100 new collections, with high precision and recall. Archivists and curators will participate in the process of building the collections, aided by interactive task-oriented tools that will leverage their knowledge, including of sampling practices and Web publishing patterns.

Accessing & Archiving: 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 our partnership with LucidWorks (LW), whose Big Data Software we will collaboratively enhance. Our decades of DL research will be integrated with our partners' reliable and persistent services (see letters).

Analyzing & Visualizing: We will provide a wide range of integrated services beyond the usual (faceted) browsing and searching, including: classification, clustering, recommendation, reporting, sentiment analysis, summarization, text mining, theme and topic identification, and visualization.

1.3 What is an Event?

Events are fundamental to our lives and memories. From the 5S perspective, events involve a spatial (time and location) aspect, with people (set in a social context) carrying out a scenario, and generating a stream of information. We will research additional definitions and characterizations of events. Then we will draw

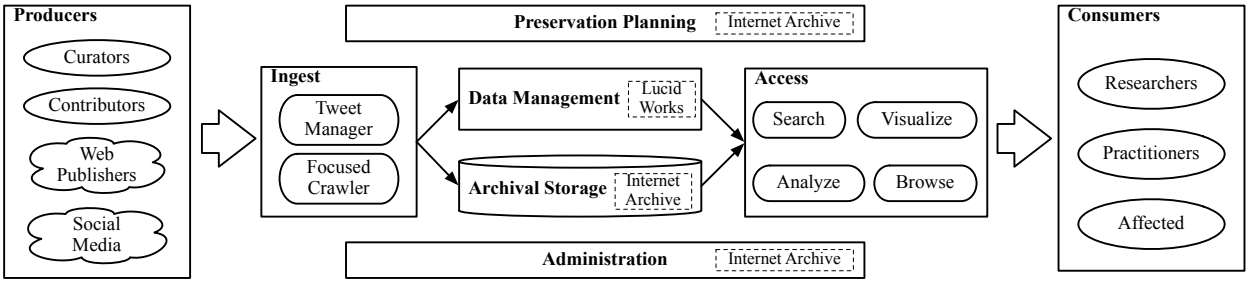

Figure 1: System architecture, as extended OAIS [60] system, integrating DL and archive aspects

on sociology, psychology, political science, etc. to extend current techniques for spotting events, e.g., Topic Detection and Tracking. So far, in our CTRnet project (see Section 2), we have identified and built (or helped with) collections for 42 significant events [33] that our team considered related to crises, tragedies, and community recovery (CTR). While we will continue with the CTR theme as we identify events, we also will broaden into other (slightly overlapping) types, related to government/community/ social situations. This should ensure broad impact. But we will accept requests for archiving an event by individuals or groups, either stated specifically (e.g., the 2011 Japan earthquake and tsunami) or in general (e.g., presidential elections). In the latter case, automatic methods are needed to identify important instances and sub-events. Solving this problem, we will detect significant events in our two areas, monitoring streams of text, pictures, and videos published daily, through formal (news) and informal (social networks) media [61, 62]. We will identify a large collection of both formal and informal sources, characterizing coverage and bias. We will zero in among those sources to build a new collection whenever an event of interest is identified, suitably considering the appropriate societies, spaces, and scenarios to guide our collecting activities.

2.0 Results from Prior NSF Support & Related Work

For the PI and co-PIs, the closest prior project is: NSF IIS-0916733: III:Small:Integrated Digital Library Support for Crisis, Tragedy, and Recovery (CTRnet), E. Fox (PI), N. Ramakrishnan, S. Sheetz, A. Kavanaugh, and D. Shoemaker, $500,000, 8/1/2009 - 7/31/2013. Two Ph.D. students have been funded, but more than 10 others provided volunteer assistance. Many publications and presentations resulted [16, 63-82]. Thus, PI Fox was a panelist in a Nov. webinar on computing for disasters, organized by Lucid-Works [83], and co-PI Sheetz ran a July webinar on Emergency Informatics and Digital Libraries [84].

Building on our digital library work related to the April 16, 2007 mass shooting [85-88], CTRnet has been developing a human and digital network for providing a range of services relating to crises and tragic events. We collected and archived CTR related information, and applied advanced information analysis methods to this domain. Access to key project results is made available through the CTRnet website: http://www.ctrnet.net. We have over 118 twitter collections, each about a different natural or manmade disaster. These and a large number of Web archive collections cover all of the major CTR events in the last three years, as well as older events on special topics, like school shootings. Some 42 collections already are permanently preserved and accessible through our partner, the Internet Archive. Resulting software also is available on request, and will be packaged for those interested, as well as shared as a result of our collaboration with LucidWorks on their Big Data Software. That software runs on clouds as well as a small Virginia Tech system being used for teaching, and on 30 nodes of our System G “green supercomputer”, from which we will provide access to over 10 terabytes of disaster-related data.

2.1 Building Collections for Crisis Events

Web Collections: We developed cyberinfrastructure to collect and archive information about CTR events. This was done in collaboration with our partner, the Internet Archive (IA), a non-profit organization working to archive the Internet. IA provides access to Web crawlers, and hosts collections containing the results of those crawls [33]. At present, there are more than 10 terabytes of data hosted by IA from CTR events [89]. As soon as we identify a CTR event, we list keywords specific to that event, query online news sources, and identify unique URLs found in related tweets identified by querying the Twitter API. Then we use the results as initial seeds for IA crawlers. Later, IA provides the archived data in the form of .warc and .arc files for further processing at Virginia Tech. We save the expanded result as HTML pages, images, and videos.

To improve the precision of the webpage collections, we researched focused crawlers and machine learning techniques. A prototype next-generation focused crawler is being developed as a class team project in a fall 2012 graduate course taught by PI Fox. We also developed a modified version of the LibSVM [90] classifier provided by the data-mining package WEKA [91], and trained a one class classifier [92]. We reduced a noisy large collection of webpages to 3000 documents clearly about school shootings, and extracted appropriate metadata. We continue efforts to build other high precision collections for access through Virginia Tech servers running the LucidWorks [53] Big Data Software. It is an application development platform enabling comprehensive search, discovery, and analysis of content and user interactions, and includes all of the necessary components, pre-integrated and certified. The open source components Solr [93], Lucene [94], Mahout [95], and OpenNLP [96] are adapted for distributed and scalable indexing, searching, browsing, machine learning, and natural language processing.

Tweet Collections: Tweeting has become commonplace during many events. Aware of the rather different work at the Library of Congress, we focused on explored the utility of collecting and analyzing tweets (i.e., posts from Twitter.com) for community support in a pilot study conducted with government officials in Arlington, Virginia [65, 71]. We also studied use of Twitter during an emergency event at UT Austin [66]. More broadly, we investigated the usage of social networking sites after crisis situations [69], and social media use during the mass protests in Iran, Tunisia, and Egypt [68, 72]. As social media play such an important role during emergency events, we have been archiving tweets for both man-made and natural disaster events, and have developed and disseminated our methodology, involving open source tools for collecting, analyzing, and visualizing tweets [67, 70]. See Fig. 2 for an overview of how this connects with our digital library and archiving research. We also integrated tweets of water main breaks with analysis and visualization [97]. The list of webpage and tweet archives we created is accessible through our website [98].

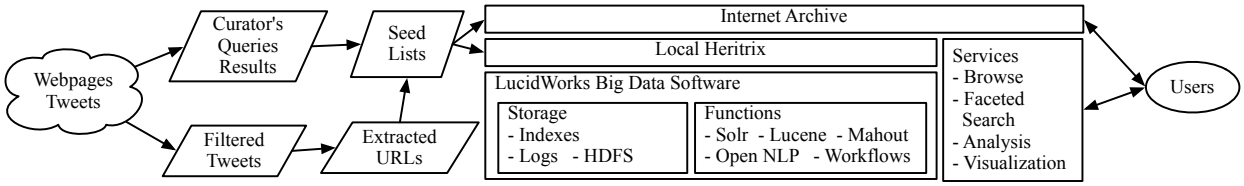

Figure 2: From sources to digital library and archive systems

2.2 Analysis, Access, and Visualization

Providing effective services through CTRnet has required research on a wide range of analyses of the available content. For example, content-based image retrieval (e.g., finding scenes with rubble from earthquakes, or other views of a school where a shooting occurred) requires building representations with

suitable image descriptors [99, 100]. Handling location-related queries and connecting data with locations on maps requires geoparsing and geocoding [101]. Text extraction is needed for name-entity recognition and metadata enhancement [102]. We must model the complex objects we find on the Web (as in news pages with photos and videos) [103] and work with subdocuments [104, 105] to support annotation as well as access to portions of emergency response plans. These efforts, plus significant data cleanup, are required to support hard queries and find chains connecting related documents, both inside and across collections [106-111]. Further, digital library architectures must be enhanced with security features [112-114] so important CTR collections with access restrictions can be included. We will build upon LucidWorks Big Data software, running on System G (an NSF and Virginia Tech project providing a "green" research platform for the development of high-performance software tools and applications with extreme efficiency at scale). Tailored workflows will process warc and tweet based data. For example, workflows will take warc files as input, and will extract, annotate, vectorize, cluster, find similar documents (and statistically interesting phrases), and index the webpages present in the archive.

On our website, using the Google Maps API, we present a CTR event map, organized around locations. To show current key terms from recently posted tweets, we automatically build new word clouds every 10 minutes, for the Japan earthquake disaster, the Libyan Revolution, and Hurricane Sandy [63].

Fig. 3 shows PhaseVis, another visualization service we developed [115]. The Four Phase Model of Emergency Management – mitigation, preparedness, response, and recovery - has been widely used in disaster and emergency management and planning [116]. However, that model has received criticism, contrasting its clear phase distinctions with the complex and overlapping nature of phases indicated by empirical evidence [117]. To investigate how phases actually occur, we designed PhaseVis based on visualization principles, and applied it to Hurricane Isaac tweet data. We first collected tweets about Hurricane Isaac, and then selected a subset of tweets that name major disaster organizations and agencies (i.e., FEMA, Red Cross, and Salvation Army). We trained classifiers using the four phases as categories. Ten-fold cross-validation showed that multi-class SVM performed the best in precision (0.8) while Naïve Bayes Multinomial performed the best in F1 score (0.782). The tweet volume in each category was visualized as a ThemeRiver™ [118], showing the 'What' aspect of a disaster. Other aspects ('When,' 'Where,' and 'Who') also are integrated in the user interface. The classification evaluation and use cases indicate that PhaseVis has potential utility in disasters, aiding those investigating a large tweet dataset.

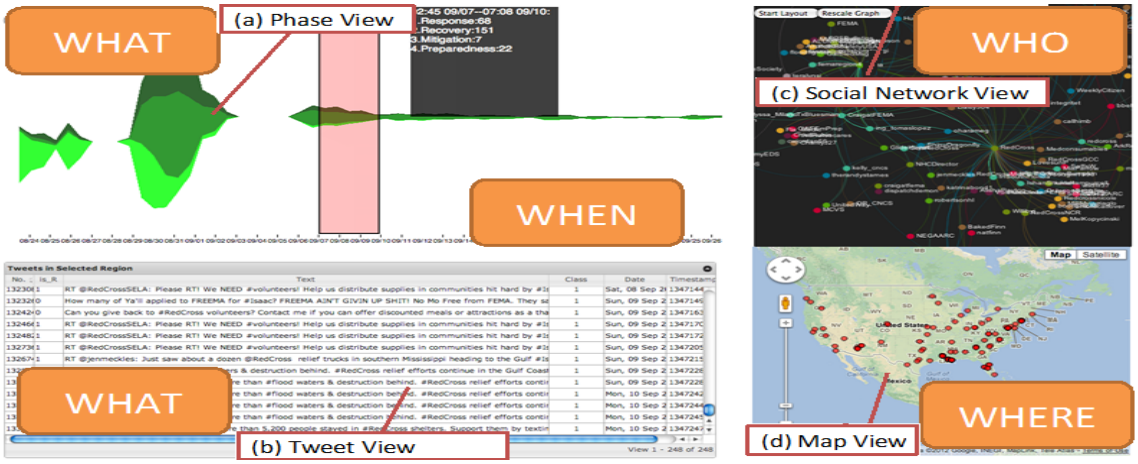

Figure 3: User interface of PhaseVis showing four different views [115]

2.3 Ontology

In next generation digital libraries, knowledge-based approaches to intelligent information systems should be integrated with machine learning; thus we have been developing a comprehensive CTR ontology [73] with in-depth and balanced domain coverage. The focus of one doctoral research study is on high-quality, scalable, semi-automatic methods that involve the least amount of human intervention as well as computational effort. We are both merging ontologies and iteratively enriching the resulting integrated CTR ontology, drawing on 185 concepts from: EM-DAT: The International Disaster Database [119], Disaster Database from University of Richmond [120], Canadian Disaster Database [121], and the DesInventar Disaster Inventory System [122].

2.4 Case Study: Real-time Archiving of Hurricane Sandy

Hurricane Sandy was a devastating superstorm affecting over seven countries between Oct. 22 and Oct. 31, 2012, causing damages exceeding 65 billion dollars. We started collected tweets on Oct. 25. Extracted seeds were added to our Archive-it Hurricane Sandy Collection [123]. Internet Archive helped with Facebook and Google Doc requests for seeds. We also crawled popular news, business, and weather portals twice a day for the first week, and used the Google News real-time coverage. Government and non-profit organization websites, tweets, and Facebook pages were crawled frequently. After the event, in our periodic crawling during recovery, to supplement government sites, we added 83 Facebook pages as seeds to the archive, building on the collection by Steven Clift [124] on local communities working on: shelters, recovery of affected places, and collecting donations. By Nov. 11, we had identified 43,098 unique seeds from Instagram.com, 40,863 from Twitter, and 26,176 from Facebook. Our webpage collection required over 2 terabytes of storage by December. CTRnet continues to expand this and other collections.

3.0 Approach

We will use a modular approach to developing IDEAL system components. This leverages our prior work developing systems like SMART [125, 126], MARIAN [41, 127-131], Envision [132-140], CITIDEL [141-155], ETANA [46, 156-166], and Ensemble [167-176]. Guided by the Internet Archive, we will enhance the support for curators, so higher quality collections can be easily built semi-automatically, and so that there will be a smooth integration of digital library and high-throughput standards based archiving, also allowing historians to work with the Wayback Machine. With LucidWorks, which now has particular interest in topic modeling, clustering, cluster labeling, trend and time series analysis, and visualization, we will extend what is possible with Lucene, Solr, and Mahout, ensuring easy and effective access for general users (both practitioners and those effected) and researchers. Below we give details, citing related studies as appropriate. One aspect of our research involves comparing, extending, enhancing, applying, and integrating advanced methods – with continuous evaluation. Another aspect involves rethinking and extending services, through more detailed modeling of the many sources and publishing practices in Web 2.0 that apply to events of interest.

3.1 Automatic Event Detection

The release of Google Flu Trends demonstrated the effective spotting of flu epidemics based on counts of flu-related queries received from Web users [177]. For such bursty events, the number of event-related queries increases dramatically [2, 178, 179]. “One day after the Assassination of Benazir Bhutto the top 4 most popular queries are all related to that event” [52, 180]. Likewise, changes in Twitter traffic can aid in detection of breaking news [66, 181], such as about an emerging political crisis, or an airplane crash.

Accordingly, IDEAL will use two types of online media for automatic event detection: formal (e.g., popular news websites like CNN and Google news), and informal (e.g., Twitter, Facebook, RSS, blogs, forums, social networks, etc.) [182-184]. Our two-stage approach (see Fig. 4) involves first, formal media for detection, and second, informal media to cross-check and to complement the formal media.

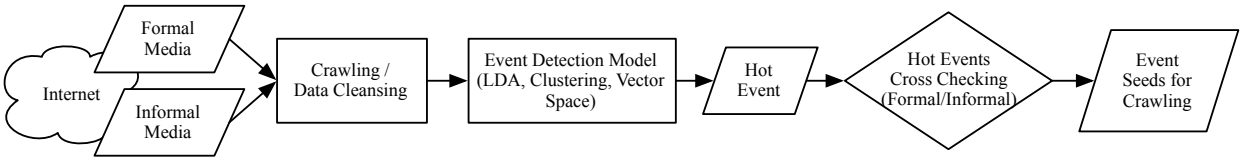

Figure 4: Automatic event detection

Using the techniques of Topic Detection and Tracking (TDT), topics are automatically (e.g., by clustering [185-190] or LDA [191-206]) identified online [178, 207-221], helping users to know "what's new" or "what's going on." However, the vast number of news topics leads to a new problem. In order to find the most important and timely events, we must quickly relate them to our categories of interest, and then rank according to priority. Timeliness can be assessed through analysis of the distribution over time of both content updates and user accesses. Importance ranking is analogous to ranking for Web retrieval, requiring fusing of a variety of signals [222-230], e.g., numbers of news stories and tweets, and numbers of queries and human views [179, 216, 231-239]. Thus, in the context of the vector space model [30, 240-247] we will utilize ranked retrieval [181, 248-252] and Web click-through data [2, 253, 254] (including anonymized user identifiers, queries, query submission time, and URLs of clicked search results [255]). The system will track events as well as themes [61]. Through user studies and experiments with Amazon Mechanical Turk (AMT) [256], we will gather data so we can: compare the effectiveness of approaches and media types, refine techniques, and integrate services for: event detection, event modeling, seed extraction, and seed generation. This will allow user specification of categories as well as of particular events of interest, submission of seeds, and requests for customized crawling.

3.2 Intelligent Focused Crawler

Once an event is identified, the next task is to collect and store related information. For this we collaborate with the Internet Archive (IA) [257]. IA is a part of the IIPC (International Internet Preservation Consortium), which has members from libraries that preserve documents of national interest, or based on current events or websites related to their countries. We aim to research, prototype, and give to IA and IIPC new improved tools to collect and preserve event-related digital content. Collecting just the right information, on a particular topic, from the entire Internet, is a very hard problem, leading to extensive research on intelligent focused crawling [258-262]. We will advance the state-of-the-art and adapt our solution to the needs of IA and IIPC; right now, they use relatively simple crawling methods whose quality depends almost entirely on human guidance through seed selection.

IA Baseline: We will continue to run crawls using IA infrastructure, and their Heritrix crawler. That will ensure reliable capture of information as topics are identified, as well as both immediate and long term access to that information through IA services. The resulting collections also will constitute a baseline for comparison as we build a next generation intelligent focused crawler at Virginia Tech, and deploy it to build collections with ever increasing recall and precision. As our collections improve through further processing and analysis, they also will go to IA for long-term preservation. Ultimately, we expect IA and others in IIPC will adopt our findings and allow suitable technology transfer.

Requirements and Approach: Because of the size diversity and complexity of the content we seek on the WWW, our system must be highly scalable and adaptable to the various types of digital objects [263].

We also must improve methods for topic representation and estimating relevance, leveraging patterns of Web publishing. Regarding representation – supplementing keywords and phrases, URLs, WordNet [264] synsets, metadata records, and ontology nodes or subgraphs – we will develop vectors of descriptors and other features. To accurately estimate the relevance (and priorities) of webpages, we will fuse information from the URLs' text and context [265-272], webpages' text, and ontology concepts [273-281].

Monitoring and tailoring to curators: We will aid collection builders to interact effectively with the system to guide and improve the processing, such as by reducing the search space and resource requirements. They may describe particular interests and provide knowledge about a topic directly to the focused crawler [282-284]. Thus, the user can add a plugin to their browser, so, when they explore, the log of their actions will record their interests regarding genre [19, 285, 286], webpages (i.e., relevant), parts of the webpages (i.e., identify structure), and types of data inside the webpages (i.e., indicate features). Curator profiles also will be enhanced, to streamline future exploration sessions.

Intelligence through modeling: A focused crawler will be more intelligent if it knows information about the event it is crawling (event models), the sources of information about the event (source models), the mechanisms used for disseminating information about the event (publishing venue models), and the entities related to the event (society / organization models). Considering 5S, the organization model will capture societies, the event model will capture the spaces and structures, the source model will capture the streams, and the publishing venue model will capture scenarios. To develop these models, as our focused crawler finds content, it also will build and store a huge graph (including the desired content about related events, connected with the above types of entities, through appropriate relationships) as it works, identifying new elements of the huge graph as it runs into them, and classifying along the way.

Event models: We will develop a model M_E that describes the different aspects of an event. We will capture the phases, geographic issues, and topical coverage of the event, drawing upon the 5S perspective. We will use the SEM (Simple Event Model) [287] ontology to describe all the aspects of an event. Then we will compare different events based on common aspects.

Source models: We will build models of sources (M_S) of information and information dissemination, characterizing bias, scope, coverage, quantity, and quality. We will identify and build a hierarchy of small as well as large commercial news organizations, and government sites, from villages to towns to counties to cities to states, which are covering news of interest. Curators will be able to select suitable sources, depending on the desired sample size as well as other criteria, thus considering language, country, region, and detailed location. Table 1 highlights some of the instances of sources of particular interest.

Table 1: Taxonomy of source models

Informal media	Formal media	
Social Media	News Media	CTR-specific Resources
Micro Blogs • Twitter • Weblogs Other • Facebook • Instagram	National newspaper • New York Times • Washington Post Local newspaper • Roanoke Times • Washington Times	Government-related • FEMA, Ready.gov Volunteer-based • Red Cross/Crescent Other • Google Crisis Response

Publishing venue model: Information dissemination can be understood in terms of sampling and how the various Web 2.0 publishing venues (blogs, Twitter, Facebook, news sites, etc.) discuss events. Each publishing venue model M_V will help us answer questions like: how and why users generate webpages on

the WWW, and how users link to existing content on the WWW. For example, during crisis situations, people may use Twitter to report some information about the crisis and link to a news website for more details. We will build a hierarchy of possible publishing venues used in Web 2.0 and link that to our source model, i.e., for each source, we will identify the possible publishing venues. We will sample appropriately for each chosen publishing venue.

Society / organization models: We will identify individuals, organizations, and a variety of instances of social units, that are involved in an event, including characterizing their respective roles. We will build a graph of these social units and link them to the related source models. For each M_O we can sample to accommodate bias related to political, economic, religious, and social aspects.

Integration: We will integrate all these models to build our intelligent focused crawler. We will make use of these models in two ways: First, we will use these models to direct the focused crawler to new seeds that were not given before and could not be reached with the current set of retrieved webpages. Second, they will direct the focused crawler to webpages that cannot be reached unless through non-relevant webpages (tunneling) [261, 288-290]. We will follow a Bayesian approach [291] to estimate the utility of adding a document (webpage) to our collection. If the webpage is relevant to our event's topic (based on its textual content) then it will be added to our collection. Yet, if the webpage is not topically relevant, still it may point to other relevant webpages through its outgoing URLs. We will consider the webpage publishing venue, source, and organization models to estimate that likelihood. More formally:

$$P(A|D) = c_1 * P(R|D) + c_2 * P(R| M_S, M_O, M_V, M_E)$$

where: A = add to collection, D = webpage, R = relevant to the event, M_S = source model, M_O = organization model, M_V = publishing venue model, M_E = event model. The first part of the equation describes the probability that we should add the webpage based on its topical relevance. The second part describes the suitability of the webpage given the publishing venue, source, organization, and the event models.

Experiments: Each type of model will be individually evaluated. For example, we can compare our models of sources with what results from analyzing some of the large collections others have built and stored at the Internet Archive. Then, when we integrate the set of models into our intelligent focused crawler, we will be able to run experiments to assess the collections resulting from decisions related to each of the models. Thus, we will compare seeds from social media (e.g., tweets) and collections built using human input seeds, to test our hypothesis that seeds collected from social media should lead to a better collection when considering coverage, precision, and recall. Another experiment will use the publishing venue model to classify different parts of a collection and find the distribution of publishing venues used in the collection. This experiment will yield information about the bias found in the collection and therefore the bias found in the seeds used to build the collections.

Evaluation: Guiding our research on focused crawling will be an extensive program of evaluation [292-295]. Controlled studies will allow measuring absolute recall, while live studies will consider relative recall. After merging the collections from multiple seed sets and crawling methods, relevance judgments from Amazon Mechanical Turk (AMT) [256] will facilitate comparisons. Measures considered will include scalability and use of resources like time and space, as well as of effectiveness, through precision, recall, and F1 (a combination measure) [296].

3.3 Text and Data Analysis

As we build more and better event archives, they should generate increased interest from a variety of stakeholder groups, who will expect easy-to-use and effective services. Those will require [161] extensive offline automatic analysis and processing of the archived information, so online interactive access can be

fast and effective, and relevant and meaningful information can be presented in an easy-to-understand fashion [297]. As was explained in Section 2.3, we have explored a variety of types of analysis. For example, given the low precision we observe in most existing archive collections, we can train classifiers to build smaller collections with higher precision. Intelligent focused crawling (see Section 3.2) also will lead to improved collections. Other types of analysis are explained below, for many of the planned types of services [91, 96, 298-308]. We will use and collaborate with LucidWorks software in improving their software to extend and support more of the analysis functionalities.

3.3.1 Theme and Topic Identification

Another important line of research is the identification of themes [309, 310] within events. This affords a more refined view of events, as well as recognition of thematic connections among events. We will use natural language processing as well as clustering and classification techniques to identify important themes within events [91, 96, 298-301].

In the same line of study, we are working on an automatic topic identification approach based on sampling of relevant information using search engine APIs. Our software lists topics independently of a prepared large document corpus, given an electronic document (e.g., webpage). We are experimenting with our prototype to compare its performance with that of four human topic indexers. We use a slightly modified version of Wolfram and Olson's metric (inter-indexer consistency density [311]). Each piece of a segmented electronic document is sent to a search API sequentially to retrieve the relevant descriptions of matching webpages from the Web. Returned descriptions are viewed as an expanded micro-corpus of a single document that is transformed into a term-document matrix as in the Vector Space Model [312, 313]. As a baseline, we apply the tf*idf term weighting scheme [314] to identify significant words. We plan to extend this approach to find RDF (Resource Description Framework) triples [315], and to connect with our focused crawler for improved archive development.

3.3.2 Categorization

We will extend the event ontology [316, 317] developed in support of sentiment analysis (see below) to help users explore (e.g., browse and search) data based on event types [318]. We also will apply taxonomy building [319, 320] and classification techniques [195, 321-323] for this step. We will evaluate software like ORANGE [299], RAPIDminer [298], and WEKA [91] to aid in these processes.

3.3.3 Sentiment Analysis

Formal and informal media [324, 325] often reflect bias and include opinion data. Sentiment analysis applies natural language processing [326, 327], e.g., to determine the attitude of a speaker or a writer with respect to some topic. The attitude may be an author's judgment or evaluation, affective state (i.e., emotional state when writing), or the intended emotional communication (i.e., desired emotional effect on the reader). Sentiment analysis can identify the most influential opinion holders, help monitor how trends of opinions change over time in social media, and aid in understanding the story behind an incident. Popular events are related to increases in sentiment strength. For example, if an event is related to protests then word usage in tweets or blogs increases in negative sentiment strength. So, sentiment analysis can play a role in automatic event detection. Also sentiment analysis can help with early forecasting of an event. Possibly vulnerable persons might be identified, to save anxiety or even lives.

We will research how to support archive studies based on sentiment, e.g., identifying different perspectives, including polarity (positive vs. negative), intensity (degree of emotion), and subjectivity (impartiality). We will extend our prior work on ontologies (Section 2.3), expanding to cover other areas beyond CTR, providing a supporting lexicon [328]. We will find relationships and patterns in textual data, to be used in an automated decision support systems or assessed by a human analyst. We will evaluate

and compare approaches to both supervised and unsupervised sentiment classification [326-334], and will leverage systems and tools like WEKA [91], OpinionFinder [305, 335], and SentiStrength [304, 336].

It is really difficult for computers to characterize the tone and meaning of a document. To address this problem, big data is required. Leveraging the 5S approach, including the spatial aspects of events, we will collect large streams of data, e.g. tweet feeds, news contents, reviews, blogs, and Web forums. We will structure documents based upon the modeling approaches proposed earlier. Sentiment analysis will be performed on each different type of event scenario, and for each type of society (see Table 2).

3.4 Data, Information, and Context Visualization

Users of IDEAL will benefit greatly from rich support for archive visualization. Visualizations will help with management of query results, analysis results, individual collections, and the complete IDEAL archive. Supported view types will include: map, time-line, geographical, and hierarchical. The best visualizations turn data into engaging stories that facilitate personal connections, as well as explanations.

Web 2.0 refers to collaborative and interactive value-added services, extending the Web's content hosting services [1]. It is an expression of the voice of the people that can be recorded and preserved as never before possible. For example, covering the events "Egypt and Tunisia revolutions", there are about 3 million entries in Web 2.0: 55% blogs, 32% news, and 13% social media [337-339]. In addition, we have collected over 11 million tweets on *Egypt Revolution* and 5 million tweets on *Libya Revolution*. To provide additional interactive services in CTRnet, we have prototyped a variety of visualizations [66, 67, 72] as well as a tool, PhaseVis, which is shown in Fig. 3. These have led us to plan a much more extensive set of visualization services for IDEAL.

As visualizations related to the live Web increase in number and quality, it is becoming clear that visualization services are expected for archives as well [340]. Our research will focus on innovative and creative approaches to archive visualization, supplementing existing techniques [341-346]. Since event information is heterogeneous and since different stakeholders have varied needs, we will research two modes of visualization for events, discussed in the next subsections [347].

3.4.1 Data Visualization

Building upon our prior research with Web and digital library [348] visualization [48, 85, 110, 138, 153, 349-353], we will devise next generation data visualization services that work directly on archives (e.g., webpages, images, and videos), supplementing flexible browsing and searching services. These will be available almost as soon as an event occurs. For example, Twitter data collected, by using both the Twitter Search API [354] and the Streaming API [355], will be processed through a pipeline of tools: the command-line data processing tool Gawk [356]; statistics packages such as R, Excel, and SPSS for data analysis [357] or integrations of Excel with visualization like NodeXL [358, 359]; text processing with Leximancer [360] or WordStat [361] for keyword analysis; and Gephi for network analysis and visualization [362, 363]. The plots by Snub of tweets related to Revolution Egypt demonstrate the value of such visualization tools in understanding data [363]. Through user and AMT studies we will evaluate a variety of services, finding which ones work best for each stakeholder group and type of task or activity.

3.4.2 Information and Context Visualization

Building on the variety of analyses described in Section 3.3, we will extend from data to information visualization [342, 364, 365]. We will research the use of interactive visual representations that amplify cognition [366], ultimately mapping to conventional 2D computer screen spaces [347] to support large communities of interactive users, though class projects also may lead to special studies with our CAVE [367, 368] or Gigapixel display [369-373].

Semantic approaches [374] will lead to triples, and visualizations using word clouds, semantic graphs, and diagrams of relationships among entities. Geospatial information will fit with a variety of mashups and hyperlinked maps [375, 376]. Timeline-based interactions [118, 377, 378] will support a suite of perspectives, e.g., chronology of an event or series of events, changes in interest among themes, shifts in polarity or intensity of various sentiments, or flows of causality. Other approaches combine a set of visualizations to aid conceptualization, e.g., word clouds, co-occurrences, year histograms, and item listings [379]. To show how different events are linked, along with related documents, we will extend our Stepping Stones and Pathways (SSP) approach [106-110]. Other visualization methods will build upon analyses related to structuring or browsing by events. Our evaluations will be of specific tasks, as well as of the higher goals of exploration, discovery, and understanding.

3.5 User Support

IDEAL will support a variety of users, who generally fit into one of the stakeholder groups described in Table 2. It will help them with the challenges they face regarding events, affording a rich set of benefits. This perspective will guide additional user-centered evaluation and subsequent research.

Table 2: Perspectives on consumers/stakeholders

	Stakeholder Types	Stakeholder Examples	Examples of Challenges	Examples of Benefits
RESEARCHERS	**Researcher**	University/institute faculty, researchers	Finding event data, Analyzing & visualizing knowledge	Access to event DL and its content & services
	Student	Graduate and undergraduate	Tailored summaries, Knowledge management	Access to a rich set of support services
	Librarian / Archivist	University, federal, state &city libraries	Preservation, Scalability, Interoperability	Automatic event detection, Archiving
PRACTI-TIONERS	**Public Sector**	Government agencies (all levels)	Intellectual property rights, Multi-nationals	Accountability, Strategic planning, Analyses
	Social Service Provider	Non-profit organizations	Early response to an event as well as help with recovery	Fast event notification
AFFEC-TED	**Directly / Indirectly Affected People**	Victims, families, and friends	Event tracking, Situation awareness, Communications, and Access	Notifications, Summaries, and Visualizations

For example, regarding librarians and archivists responsible for curating collections about events of interest, we will aid them to speed up and broaden their accomplishments, through event detection (see Section 3.1). Further, as requested, we will extend the work discussed in Section 2.1, as explained in Section 3.2, to build high quality collections. Similarly, for other stakeholder groups, as is summarized in Table 2, we will help them face difficult challenges, and reap desired benefits (illustrated in the right two columns of that table, respectively). Thus, we will extend the research discussed in Sections 2.2 - 2.4, leading to a rich set of services for analysis (Section 3.3) and visualization (Section 3.4).

3.6 Dissemination and Validation

Boards: We will disseminate our results through twice-yearly meetings with our local advisory and external advisory boards (see Table 3), staggered so there is an in-person or webinar discussion every three months. At these meetings we will discuss the results of our ongoing evaluations of the elements of the architecture, as they are developed and deployed, and as progress is made on project plans.

Based on theory: We will continuously evaluate our products and processes at two levels. We will base our evaluation on five constructs of Diffusion of Innovation Theory [380]: 1) Relative advantage captures the extent that a new technology improves on existing techniques; 2) Complexity is the perception that a new technology is easy-to-use; 3) Compatibility is the extent that a new technology fits with the task goals of the people using the technology; 4) Trialability is the idea that the technology is testable by its users; 5) Observability is the extent that using the technology can be observed by others.

First we will explicitly evaluate each technology we develop, e.g., event identification, in the context provided by the 5S framework [59], for each of our stakeholder groups (societies). Thus, political scientists interested in an election will focus on scenarios and streams of information different from government officials monitoring that election. Gathering and analyzing rich qualitative data from users (representatives of a community of stakeholders) about their role in events and their interaction with event information will help us support situated use of IDEAL [381, 382]. We will employ a variety of techniques including: one-on-one and focus-group interviews [383, 384]; surveys and questionnaires; case studies of common scenarios and particular use cases for IDEAL; analysis of logs resulting from our instrumenting each service related to user actions; and usability studies of each subsystem and service. These will help in designing tasks for evaluating the relative advantage, complexity, and compatibility of IDEAL components. We will extend the techniques used in prior user studies [385-387].

Second, we will continuously evaluate the overall project by examining the integration among elements of the architecture, e.g., ensuring seamless transition from event identification to seed generation to intelligent crawling. We also will solicit and implement recommendations from our internal and external advisory boards for additional evaluation activities and methods. Thus, Chris Barrett will be able to leverage our data and services to feed into models and simulations of spread related to demonstrations and revolutions (see letter). Likewise, our Library will connect us with those interested in archiving and analyzing events, through their connections with social scientists, including in data management planning (see letter). Patrick Meier and Carlos Castillo (see letter) will connect us with humanitarian organizations, through their center in Qatar, which PI Fox will visit yearly, due to a related Web archiving project [388].

Our prior experience with qualitative and innovation diffusion research will help in designing and conducting these evaluations. [65, 385, 386]. The participants in the evaluations will be a representative sample of our stakeholder groups and will include members of the advisory boards, partners of Internet Archive (including Board members from Alberta, ODU, Stanford, and USC), customers of LucidWorks, students at Virginia Tech, and Amazon Mechanical Turkers, selected as is appropriate.

Ongoing evaluative case study: One of the goals of this project is to connect with the needs of stakeholders. Starting in the first year of the project, we will validate and iteratively refine our research by working closely with colleagues and students at Virginia Tech. We will provide technical assistance to colleagues affiliated with the Center for Peace Studies and Violence Prevention (Peace Center – see letter by Director Hawdon), which is located within the Department of Sociology at Virginia Tech. That is, we will evaluate our tools along the way, using real tasks required by members of the Peace Center. One project proposed by the Peace Center, in collaboration with faculty members at the University of Turku, Finland, is to investigate the presence and influence of on-line hate groups, and related events. Several

perpetrators of recent mass shootings, including the Norwegian terrorist Anders Breivik, Russian mass murderer Dmitry Vinogradov, and Finnish school shooter Pekka-Eric Auvinen, have been actively involved in online hate groups. The project, funded by the Kone Foundation for one year, beginning 1/1/2013 (with possible extension to three years), aims to investigate the social networks of hate group members, their rhetoric, the techniques they use to disseminate their messages, and the influence these groups have on young people who are exposed to them. Using tweet analysis, Web crawling, and the sampling of media articles related to mass shooting events, our proposed project will help with the collection, analysis, and visualization of such data. Elements of the IDEAL architecture are consistent with the goals of this project, e.g., finding and crawling websites of hate groups; thus we can test our techniques by collaborating with the students and faculty in the Peace Center, as we support their research.

After we focus on users among the many interested groups at Virginia Tech in year one, we will gradually broaden as additional partners emerge as a result of our tutorials, workshops, papers, and presentations. We also will provide direct team support for focused collection, analysis, and visualization studies. In year 1 we expect to carry out three such studies, followed by 6 in year 2, and 9 in year 3. As our tools and services improve, we will shift from internal comparative evaluations to live user tests, and then to log and retrospective analysis of those working with our expanding collections. These will be coupled with educational activities: PI Fox teaches yearly a capstone course for CS seniors, CS4624 (Multimedia, Hypertext and Information Access), as well as a yearly graduate course, usually CS5604, Information Retrieval, or else CS6604, Digital Libraries; in all of these there are term projects wherein teams of students will connect with the proposed research, through system building and experiments.

Broader dissemination: Starting in year two, we will broaden our dissemination efforts and recruit from stakeholder groups interested in specific events or series of events, to collaborate in evaluating our tools and techniques. In support of this goal we will propose tutorials and workshops, such as for the American Sociological Association Annual Meeting and the Information Systems for Crisis Response and Management (ISCRAM) conference. We will make similar proposals for conferences on digital government, digital libraries, Web archiving, curation, and preservation. The result will be a growing base of researchers as well as general users of IDEAL. Thus our dissemination efforts will facilitate the validation of our tools, as we refine their usefulness, ease-of-use, and consistency with users' scenarios.

4.0 Management Plan

We plan for IDEAL to start soon after the July 2013 termination of CTRnet, allowing unbroken archiving of CTR events, as well as expansion of that archiving to cover key government / community events. We also will dramatically expand the scope of our research, providing broader user support (see Table 2) and an extensive integrated program of studies related to event DLs and archives (recall Section 3).

We will retain 4 of the 5 co-PIs involved in CTRnet (see Budget Justification for overview of our areas of focus in this project), continue our partnership with the Internet Archive, and expand to connect with the broader International Internet Preservation Consortium. We will continue with local and external advisory committees/boards, but will expand them to gain advice from others interested in government and community related Web archiving. Table 3 shows the people who have agreed to serve.

Regarding research on particular methods and parts of our infrastructure, we expect to leverage our CTRnet project to have a partially operational IDEAL by the end of year 1, and an improved version by the end of each subsequent year – as we dramatically improve its services, efficiency, and effectiveness. Guided by co-PI Sheetz, we will apply a modified Agile methodology that includes: continual collaboration with stakeholders, accommodating changing requirements, test/scenario driven development,

frequent delivery of software, and empowering teams and individuals [389]. The methodology will be modified due to the nature of the development team, with continuing students building and maintaining software for key scenarios, and supplemental elements and enhancements added through independent studies and graduate theses. Project teams from courses will evaluate system components (often by comparing data streams generated by the system to those generated by stakeholders), and benefit from awareness of Diffusion of Innovation Theory constructs.

Table 3: External and internal advisory boards

External	Paul Doscher, LucidWorks (see letter)	Patrick Meier, iRevolution (see letter from Carlos Castillo)	Eric Van de Velde, EVdV Consulting
	Kristine Hanna, IA (see letter)	Michael Nelson, ODU	Kris Kasianovitz, Stanford
	Susan Metros, USC	Geoff Harder, U. Alberta	
Internal	Tyler Walters, Library Dean (see letter)	James Hawdon, Director CPSVP (see letter)	Gardner Campbell, Learning Technology
	Purdom Lindblad, Library	John Ryan, Head, Sociology	Sanmay Das, Comp. Sci.
	Gail McMillan, Library	Russell T. Jones, Psychology	Chris North, Comp. Sci.
	Chris Barrett, NDSSL (see letter)	Timothy Luke, Political Sci.	Scott Midkiff, CIO

5.0 Summary of Broader Impact and Intellectual Merit

The Integrated Digital Event Archive and Library (IDEAL) will support our planned intelligent information system research, which is ambitious but feasible, since we build on almost 30 years of related work, including that on our CTRnet project, which already has had significant impact (see Section 2).

The **broader impact** of our work (recall Table 2) is on institutions providing library and archiving services, as well as their patrons, and on the communities working on information retrieval, library & information science, or Web science. The techniques and methods we develop, findings from our evaluations, software we build, and services we provide, will all be open and shared. The archive we develop, supported by the continuously improving IDEAL system, should serve a wide range of stakeholder communities, including sociologists, psychologists, political scientists, government workers, policy makers, those concerned with crises or tragedies, historians, and the general public. More broadly, we expect to usher in a new era of permanent archiving of interesting events that will promote deeper understanding and future study based on what now is highly transient digital information.

The **intellectual merit** of our project includes theory, algorithms, techniques, methods, software, systems, and evaluation results ranging across many areas of information retrieval, information visualization, library and information science, archiving, and the Web. Now that the 5S framework [59] has been shown to provide a solid theoretical foundation for digital libraries [16], we will extend it to support Web 2.0 archiving and services. We will solve the challenging problems associated with detecting interesting events, in our two chosen areas, or in user-specified sub-areas. We will find effective solutions for intelligent focused crawling about events, and classification techniques to ensure that event archives have high recall and precision. We will test state-of-the-art approaches to archive analysis and visualization, and discover better methods with regard to event archives for supporting identification of themes, understanding sentiments, and categorization. Our contributions also will include integration and evaluation at the micro and macro levels, so our theoretical unification of the area will lead to improvements at the system and services levels.

References

[1] T. Oreilly, "What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software," *Communications & Strategies,* vol. No. 65, 1st Quarter, 2007. http://ssrn.com/paper=1008839

[2] Q. Zhao, T.-Y. Liu, S. S. Bhowmick, and W.-Y. Ma, "Event detection from evolution of click-through data," in Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, Philadelphia, PA, USA, 2006, pp. 484-493.

[3] W. Y. Arms, *Digital Libraries*. Cambridge, MA: MIT Press, 2000.

[4] E. Fox, H. Suleman, D. Madalli, and L. Cassel, "Digital Libraries," in *Practical Handbook of Internet Computing*, M. Singh, Ed., 2004.

[5] E. A. Fox and O. Sornil, "Digital Libraries," in *Modern Information Retrieval*, R. Baeza-Yates and B. Ribeiro-Neto, Eds. Harlow, England: ACM Press / Addison-Wesley-Longman, 1999, pp. 415-432.

[6] E. A. Fox and O. Sornil, "Digital Libraries," in *Encyclopedia of Computer Science, 4th edition*, A. Ralston, E. D. Reilly, and D. Hemmendinger, Eds. London: Nature Publishing Group, 2000, pp. 576-581.

[7] E. A. Fox and S. Urs, "Digital Libraries," in *Annual Review of Information Science and Technology (ARIST), Ch. 12*. vol. 36, B. Cronin, Ed. Amsterdam: American Society for Information Science, 2002, pp. 503-589.

[8] V. Srinivasan, S. Yang, and E. A. Fox, "Digital Libraries," in *Encyclopedia of Database Systems* Berlin: Springer, 2008.

[9] E. A. Fox, "Digital Libraries ("hot topics")," *IEEE Computer,* vol. 26, no. 11, pp. 79-81, 1993.

[10] E. A. Fox, R. M. Akscyn, R. K. Furuta, and J. J. Leggett, "Digital Libraries (Introduction to Special Issue)," *Communications of the ACM,* vol. 38, no. 4, pp. 22-28, April 1995. http://doi.acm.org/10.1145/205323.205325

[11] S. Griffin. (1999). *Digital Libraries Initiative* [home page for Phase I]. Available: http://www.dli2.nsf.gov/dlione/

[12] E. A. Fox, "The Digital Libraries Initiative: Update and Discussion: Guest editor's introduction to Special Section," *Bulletin of the American Society of Information Science,* vol. 26, no. 1, pp. 7-11, 1999.

[13] M. Lesk, *Understanding Digital Libraries, 2nd Edition*. San Francisco: Morgan Kaufmann, Elsevier, 2005.

[14] G. Marchionini and E. A. Fox, "Progress toward digital libraries: Augmentation through integration; Guest Editor's Introduction to Special Issue on Digital Libraries," *Information Processing and Management,* vol. 35, no. 3, pp. 219-225, 1999.

[15] C. L. Borgman, "What are digital libraries? Competing visions," *Information Processing and Management,* vol. 35, no. 3, pp. 227-243, January 1999.

[16] E. A. Fox, Ed., *Digital Libraries: A 5S Approach*. Blacksburg, VA: 502 page Virginia Tech textbook used in CS6604, in process to appear first as 4 smaller books and then as a single large book, published by Morgan-Claypool, 2011.

[17] E. A. Fox, "Sourcebook on Digital Libraries: Report for the National Science Foundation," Dept. of Computer Science, Virginia Tech, Blacksburg, VA, Technical Report TR-93-35, December 1993. http://fox.cs.vt.edu/pub/DigitalLibrary/

[18] J. Jeon and Y. Liu, "Semi-supervised learning for automatic prosodic event detection using co-training algorithm," in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing, Stroudsburg, PA, USA, 2009, pp. 540-548.

[19] N. Zhang, L.-Y. Duan, Q. Huang, L. Li, W. Gao, and L. Guan, "Automatic video genre categorization and event detection techniques on large-scale sports data," in Proceedings of the 2010 Conference of the Center for Advanced Studies on Collaborative Research, Toronto, Ontario, Canada, 2010, pp. 283-297.
[20] A. Agarwal and O. Rambow, "Automatic detection and classification of social events," in Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, Massachusetts, 2010, pp. 1024-1034.
[21] D. Jurgens and K. Stevens, "Event detection in blogs using temporal random indexing," in Proceedings of the Workshop on Events in Emerging Text Types, Borovets, Bulgaria, 2009, pp. 9-16.
[22] K. Ellingsen, "Salient event-detection in video surveillance scenarios," in Proceedings of the 1st ACM workshop on analysis and retrieval of events/actions and workflows in video streams, Vancouver, British Columbia, Canada, 2008, pp. 57-64.
[23] A. Rosenberg and J. Hirschberg, "Detecting pitch accents at the word, syllable and vowel level," in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, Boulder, Colorado, 2009, pp. 81-84.
[24] P. R. Pietzuch, B. Shand, and J. Bacon, "A framework for event composition in distributed systems," in Proceedings of the ACM/IFIP/USENIX 2003 International Conference on Middleware, Rio de Janeiro, Brazil, 2003, pp. 62-82.
[25] R. Castellanos, H. Kalva, O. Marques, and B. Furht, "Event detection in video using motion analysis," in Proceedings of the 1st ACM international workshop on analysis and retrieval of tracked events and motion in imagery streams, Firenze, Italy, 2010, pp. 57-62.
[26] K. Watanabe, M. Ochi, M. Okabe, and R. Onai, "Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs," in Proceedings of the 20th ACM international conference on information and knowledge management, Glasgow, Scotland, UK, 2011, pp. 2541-2544.
[27] A. Stewart, M. Smith, and W. Nejdl, "A transfer approach to detecting disease reporting events in blog social media," in Proceedings of the 22nd ACM conference on hypertext and hypermedia, Eindhoven, The Netherlands, 2011, pp. 271-280.
[28] C. Poppe, S. D. Bruyne, and R. V. d. Walle, "Generic architecture for event detection in broadcast sports video," in Proceedings of the 3rd international workshop on automated information extraction in media production, Firenze, Italy, 2010, pp. 51-56.
[29] D. Tjondronegoro, Y.-P. P. Chen, and B. Pham, "A statistical-driven approach for automatic classification of events in AFL video highlights," in Proceedings of the 28th Australasian conference on Computer Science - Volume 38, Newcastle, Australia, 2005, pp. 209-218.
[30] G. Kumaran and J. Allan, "Using names and topics for new event detection," in Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada, 2005, pp. 121-128.
[31] G. Kumaran and J. Allan, "Text classification and named entities for new event detection," in Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, Sheffield, United Kingdom, 2004, pp. 297-304.
[32] P. Dwivedi, "Archive - Where it started and the Problems of Perpetuity," in Proceedings of the 18th IEEE Symposium on Mass Storage Systems and Technologies (MSS), 2001, pp. 353-353.
[33] InternetArchive. (2011). *Spontaneous Events (part of Archive-It's Collections by Topic)*. Available: http://www.archive-it.org/public/topic.html?topic=spontaneousEvents

[34] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. Balakireva, S. Ainsworth, and H. Shankar, "Memento: Time Travel for the Web," LANL, Los Alamos, arXiv preprint arxiv:0911.1112, 2009. http://arxiv.org/abs/0911.1112

[35] A. Paepcke, C.-C. K. Chang, H. Garcia-Molina, and T. Winograd, "Interoperability for Digital Libraries Worldwide," *Communications of the ACM,* vol. 41, no. 4, pp. 33-43, 1998.

[36] A. Paepcke, S. B. Cousins, H. G. Molina, S. W. Hassan, S. K. Ketchpel, M. Roscheisen, and T. Winograd, "Using Distributed Objects for Digital Library Interoperability," *IEEE Computer Magazine,* vol. 29, no. 5, pp. 61-68, 1996.

[37] C. Lynch and H. Garcia-Molina, "Interoperability, Scaling, and the Digital Libraries Research Agenda: A Report on the May 18-19, 1995 IITA Digital Libraries Workshop," IITA, Reston, VA, 1995. http://www-diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html

[38] P. Miller, "Interoperability. What is it and Why should I want it?," *Ariadne,* no. 24, 2000. http://www.ariadne.ac.uk/issue24/interoperability/intro.html

[39] A. M. Ouksel and A. Sheth, "Semantic Interoperability in Global Information Systems," *SIGMOD Record,* vol. 28, no. 1, pp. 5-12, 1999.

[40] C. Lagoze and H. V. d. Sompel, "The Open Archives Initiative: building a low-barrier interoperability framework," in Proc. of the 1st ACM/IEEE-CS Joint Conf. on Digital Libraries (JCDL'2001), June 24-28, Roanoke, Virginia, 2001, pp. 54-62. http://doi.acm.org/10.1145/379437.379449

[41] M. A. Goncalves, R. K. France, and E. A. Fox, "MARIAN: Flexible Interoperability for Federated Digital Libraries," in *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2001, September 4-9, Darmstadt, Germany, Springer Lecture Notes in Computer Science 2163*, 2001, pp. 173-186. http://www.springerlink.com/index/83V86UNDXFDHP5AV

[42] Y. Petinot, C. L. Giles, V. Bhatnagar, P. B. Teregowda, and H. Han, "Enabling Interoperability For Autonomous Digital Libraries : An API To CiteSeer Services," in ACM/IEEE Joint Conference on Digital Libraries (JCDL 2004), Tucson, AZ, June 7-11, 2004, pp. 372-373.

[43] S. Payette, C. Blanchi, C. Lagoze, and E. A. Overly, "Interoperability for Digital Objects and Repositories: The Cornell/CNRI Experiments," *D-LIb Magazine,* vol. 5, no. 5, 1999. http://www.dlib.org/dlib/may99/payette/05payette.html

[44] E. Fox, Q.-F. Chen, and R. K. France, "Integrating Search and Retrieval with Hypertext," in *Hypertext/ Hypermedia Handbook*, E. Berk and J. Devlin, Eds. New York: McGraw-Hill, 1991, pp. 329-355.

[45] A. Meissner, T. Luckenbach, T. Risse, T. Kirste, and H. Kirchner, "Design Challenges for an Integrated Disaster Management Communication and Information System," in The First IEEE Workshop on Disaster Recovery Networks (DIREN 2002), June 24, New York City, 2002.

[46] U. Ravindranathan, R. Shen, M. Goncalves, W. Fan, E. A. Fox, and J. W. Flanagan, "ETANA-DL: A Digital Library For Integrated Handling Of Heterogeneous Archaeological Data," in *Proceedings of the Fourth ACM/IEEE Joint Conference on Digital Libraries: Global Reach and Diverse Impact, JCDL2004, Tucson, AZ, June 7-11*, 2004, pp. 76-77.

[47] A. Raghavan, D. Rangarajan, R. Shen, M. A. Goncalves, N. S. Vemuri, W. Fan, and E. A. Fox, "Schema mapper: A visualization tool for DL integration," in *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, June 7-11, 2005, Denver, CO*, 2005, p. 414.

[48] R. Shen, N. S. Vemuri, W. Fan, R. d. S. Torres, and E. A. Fox, "Exploring digital libraries: Integrating browsing, searching, and visualization," in *6th ACM/IEEE Joint*

Conference on Digital Libraries (JCDL), June 11-15, 2006, Chapel Hill, NC, 2006, pp. 1-10.

[49] R. Shen, "Integration," in *Digital Libraries: A 5S Approach*, E. A. Fox, Ed. Blacksburg, VA: Virginia Tech (in 502 page textbook used in CS6604, in process to appear first as 4 smaller books and then as a single large book, published by Morgan-Claypool, 2012-2013), 2011, p. Ch. 5.

[50] D. Wang, S. Zhu, T. Li, Y. Chi, and Y. Gon, "Integrating Document Clustering and Multidocument Summarization," *ACM Trans. Knowl. Discov. Data,* vol. 5, no. 3, pp. 1-26, 2011.

[51] D. Fahland, T. M. Gläßer, B. Quilitz, S. Weißleder, and U. Leser, " HUODINI - Flexible Information Integration for Disaster Management," in 4th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Delft, NL, 2007.

[52] S. Aixin and H. Meishan, "Query-Guided Event Detection From News and Blog Streams," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on,* vol. 41, no. 5, pp. 834-839, 2011.

[53] LucidWorks. (2012). *LucidWorks Big Data*. Available: http://www.lucidworks.com/products/lucidworks-big-data

[54] E. Fox and M. A. Gonçalves. (2011, December 1). *5S Framework for Digital Libraries*. Available: http://www.dlib.vt.edu/projects/5S-Model/

[55] R. Shen, *Applying the 5S Framework To Integrating Digital Libraries (Doctoral Dissertation)*. Blacksburg, VA, USA: Virginia Tech, 2006. http://scholar.lib.vt.edu/theses/available/etd-04212006-135018/

[56] M. A. Goncalves, E. A. Fox, L. T. Watson, and N. A. Kipp, "Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries," *ACM Transactions on Information Systems,* vol. 22, no. 2, pp. 270-312, April 2004. http://doi.acm.org/10.1145/984321.984325

[57] Q. Zhu, "5SGraph: A Modeling Tool for Digital Libraries," Master's Thesis, Virginia Tech – Department of Computer Science, Blacksburg, 2002. http://scholar.lib.vt.edu/theses/available/etd-11272002-210531/

[58] M. A. Goncalves and E. A. Fox, "5SL - A Language for Declarative Specification and Generation of Digital Libraries," in *Proc. JCDL'2002, Second ACM / IEEE-CS Joint Conference on Digital Libraries, July 14-18*, G. Marchionini, Ed. Portland, Oregon: ACM, 2002, pp. 263-272.

[59] E. A. Fox, M. A. Goncalves, and R. Shen, *Theoretical Foundations for Digital Libraries: The 5S (Societies, Scenarios, Spaces, Structures, Streams) Approach*. San Francisco: Morgan & Claypool, 2012.

[60] CCSDS. (2012). *Reference Model for an Open Archival Information System (OAIS)* Available: http://public.ccsds.org/publications/archive/650x0m2.pdf

[61] R. Papka, "On-Line New Event Detection, Clustering, and Tracking," PhD Doctoral Dissertation, University of Massachusetts, Amherst, MA, USA, 1999.

[62] Y. Sen, C. Xueqi, C. You, Z. Jin, X. Hongbo, and F. Gaolin, "Detect Events on Noisy Textual Datasets," in Proceedings of the 12th International Asia-Pacific Web Conference (APWEB), Busan, Korea, 2010, pp. 372-374.

[63] CTRnet. (2011). *Crisis, Tragedy, and Recovery Network website*. Available: http://www.ctrnet.net/

[64] S. Yang, A. Kavanaugh, N. P. Kozievitch, L. T. Li, V. Srinivasan, S. D. Sheetz, . . . E. A. Fox, "CTRnet DL for Disaster Information Services," in Proceedings of the 11th annual international ACM/IEEE joint conference on digital libraries (JCDL 2011), June 13-17 Ottawa, Ontario, Canada, 2011, pp. 437-438. http://www.ctrnet.net/sites/default/files/jcdl212p-yang-submitted.pdf

[65] A. Kavanaugh, E. A. Fox, S. Sheetz, S. Yang, L. T. Li, T. Whalen, . . . L. Xie, "Social Media Use by Government: From the Routine to the Critical," in Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times (dg.o 2011), June 12-15 College Park, Maryland, 2011, pp. 121-130. http://www.ctrnet.net/sites/default/files/dgo.2011.Paper Final.pdf

[66] L. T. Li, S. Yang, A. Kavanaugh, E. A. Fox, S. D. Sheetz, D. Shoemaker, . . . V. Srinivasan, "Twitter Use During an Emergency Event: the Case of the UT Austin Shooting," in Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times (dg.o 2011), June 12-15, College Park, Maryland, 2011, pp. 335-336. http://www.ctrnet.net/sites/default/files/dgo2011-camerarady2.pdf

[67] S. Yang and A. L. Kavanaugh, "Half-Day Tutorial: Collecting, Analyzing and Visualizing Tweets using Open Source Tools," in Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times (dg.o 2011), June 12-15, College Park, Maryland, 2011, pp. 374-375. http://www.ctrnet.net/sites/default/files/Twitter tutorial submitted.pdf

[68] A. Kavanaugh, S. Yang, S. Sheetz, L. T. Li, and E. Fox, "Between a Rock and a Cell Phone: Social Media Use during Mass Protests in Iran, Tunisia and Egypt," Virginia Tech, Department of Computer Science, Blacksburg, VA, Technical Report TR-11-10, 2011. http://eprints.cs.vt.edu/archive/00001149/

[69] S. D. Sheetz, E. A. Fox, A. L. Kavanaugh, D. J. Shoemaker, A. Fitzgerald, and S. Palmer, "Why Students Use Social Networking Sites After Crisis Situations," in 8th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2011, Lisbon, Portugal, 8-11 May, poster 11, 2011. http://www.ctrnet.net/sites/default/files/ISCRAM2011WhyFacebookPoster.pdf

[70] S. Yang and A. Kavanaugh, "Collecting, Analyzing and Visualizing Tweets using Open Source Tools," Virginia Tech, Department of Computer Science, Blacksburg, VA, Technical Report TR-11-14, 2011. http://eprints.cs.vt.edu/archive/00001160/01/Dg.o.2011TutorialHandout-techreport.pdf

[71] A. Kavanaugh, A. Nastev, E. A. Fox, S. Sheetz, D. Shoemaker, L. Xie, . . . V. Srinivasan, "Social Media for Cities, Counties and Communities. Final Grant Report to VT CCSR," 2011. http://www.ctrnet.net/sites/default/files/CCSR White Paper Report VT IBM Kavanaugh Natsev.pdf

[72] A. Kavanaugh, S. Yang, S. D. Sheetz, L. T. Li, and E. A. Fox, "Microblogging in Crisis Situations: Mass Protests in Iran, Tunisia, and Egypt," in Workshop in conjunction with the ACM Conference on Human Factors in Computing Systems (CHI'11), Vancouver, Canada, 2011, p. 6 pages. http://www.ctrnet.net/sites/default/files/Twitter Use in Iran Tunisia Egypt.Kavanaugh.Final_0.pdf

[73] S. Sheetz. (2011). *A Database Driven Initial Ontology for Crisis, Tragedy, and Recovery (working paper)*. Available: http://www.ctrnet.net/sites/default/files/CTR_Ontology_ResearchInProgressV2.pdf

[74] V. Srinivasan, B. Dewanjee, E. A. Fox, D. J. Shoemaker, S. D. Sheetz, A. Kavanaugh, and N. Ramakrishnan, "CTRnet: A Distributed Digital Library for Rescue and Recovery (winner of best poster in category award)," in International Outreach NOW Conference, Virginia Tech, Blacksburg, Virginia, 2009.

[75] E. Fox, "Digital Libraries, Scholarship, and the Humanities (keynote presentation as Guest of Honor for the conference)," in *Libraries in the Digital Age (LIDA) 2010, 24-28 May*, University of Zadar, Zadar, Croatia, Year. http://www.ffos.hr/lida/lida2010/

[76] E. Fox, "Introduction to Digital Libraries (refereed 1/2 day tutorial)," in *ACM/IEEE Joint Conf. on Digital Libraries, JCDL 2011, June 13-17*, Ottawa, Year.

[77] E. Fox, "Introduction to (Teaching / Learning about) Digital Libraries (refereed 1/2 day tutorial)," in Research and Advanced Technology for Digital Libraries, Proc. 14th European Conference, ECDL2010, Sept. 6-10, Glasgow, 2010.

[78] E. Fox, "Introduction to Digital Libraries (refereed full-day tutorial)," in JCDL 2010, June 21-25, Gold Coast, Australia, 2010.

[79] A. Kavanaugh, E. Fox, S. Sheetz, S. Yang, L. T. Li, T. Whalen, . . . L. Xie, "Social Media for Cities, Counties and Communities," Department of Computer Science, Virginia Tech TR-11-09, July 2011. http://eprints.cs.vt.edu/archive/00001148/

[80] E. A. Fox. (2009). *CTRnet (Crisis, Tragedy, & Recovery Network) (*http://www.ctrnet.net*): A global human network and distributed digital library. Invited presentation at University of Iowa, Oct. 29*. Available: http://fox.cs.vt.edu/talks/2009/20091029IowaCTRnet.pptx

[81] E. A. Fox. (2009). *CTRnet: A Crisis, Tragedy, & Recovery Network) (*http://www.ctrnet.net*). Virginia College of Osteopathic Research Day, Oct. 16*. Available: http://fox.cs.vt.edu/talks/2009/20091016CTR-VCOM.ppt

[82] K. Hanna, E. A. Fox, J. Jones, and P. Srinivasan. (2008). *Capturing Crisis: A Digital Library to Study Tragedy and Recovery from Around the World. Presentations on panel at CNI Fall Meeting, Washington, DC, Dec. 9*. Available: http://www.cni.org/tfms/2008b.fall/Abstracts/PB-digital-hanna.html, http://www.cni.org/tfms/2008b.fall/Abstracts/handouts/cni_digital_hanna.pdf

[83] LucidWorks. (2012). *Webinar on Computing for disasters: saving lives with big data*. Available: http://pro.gigaom.com/webinars/lucid-imagination-computing-for-disasters-saving-lives-with-big-data/

[84] CTRnet. (2012). *Webinar on Emergency Informatics and Digital Libraries*. Available: http://www.ctrnet.net/webinars

[85] E. A. Fox, C. Andrews, W. Fan, J. Jiao, A. Kassahun, S.-C. Lu, . . . L. Boutwell, "A Digital Library for Recovery, Research, and Learning from April 16, 2007 at Virginia Tech," *Traumatology,* vol. 14, no. 1, pp. 64-84, 2008.

[86] E. A. Fox, W. Fan, C. North, N. Ramakrishnan, and D. Shoemaker. (2007). *The 4/16 Research Library*. Available: http://www.dl-vt-416.org

[87] CDDC. (2007). *The April 16 Archive*. Available: http://www.april16archive.org/, http://www.cddc.vt.edu/

[88] M. Hughes, M. Brymer, W. T. Chiu, J. A. Fairbank, R. T. Jones, R. S. Pynoos, . . . R. C. Kessler, "Post-traumatic Stress among Students after the Shootings at Virginia Tech.Preprint: Psychological Trauma: Theory, Research, Practice, and Policy. Advance online publication. doi: 10.1037/a0024565," Virginia Tech, Blacksburg, VA, 2011. http://hdl.handle.net/10919/11318

[89] CTRnet. (2012). *CTRnet Web Collections*. Available: http://archive-it.org/organizations/156

[90] C. C. Chang and C. J. Lin. (2001). *LIBSVM: a library for support vector machines*. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[91] WEKA. (2011, December 1). *WEKA, Machine Learning Group at University of Waikato*. Available: http://www.cs.waikato.ac.nz/ml/weka/index.html

[92] V. Srinivasan and P. Angara, "Classification," in *Digital Libraries: A 5S Approach*, E. A. Fox, Ed. Blacksburg, VA: Virginia Tech (in 502 page textbook used in CS6604, in process to appear first as 4 smaller books and then as a single large book, published by Morgan-Claypool, 2012-2013), 2011, p. Ch. 8.
[93] Solr. (2012). *Apache Solr*. Available: http://lucene.apache.org/solr/
[94] Lucene. (2012). *Apache Lucene*. Available: http://lucene.apache.org/
[95] Mahout. (2012). *Apache Mahout*. Available: http://mahout.apache.org/
[96] OpenNLP. (2011, December 1). *Toolkit for Processing of Natural Language Text*. Available: http://opennlp.apache.org/
[97] S. Lee, N. Elsherbiny, and E. A. Fox, "A digital library for water main break identification and visualization," in Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, Washington, DC, USA, 2012, pp. 335-336.
[98] CTRnet. (2012). *CTRnet Web and Tweet Archives*. Available: http://www.ctrnet.net/node/2937
[99] R. d. S. Torres, N. a. P. Kozievitch, and U. Murthy, "Content-based Image Retrieval," in *Digital Libraries: A 5S Approach*, E. A. Fox, Ed. Blacksburg, VA: Virginia Tech (in 502 page textbook used in CS6604, in process to appear first as 4 smaller books and then as a single large book, published by Morgan-Claypool, 2012-2013), 2011, p. Ch. 9.
[100] N. Kozievitch, S. Codio, J. Francois, E. Fox, and R. Torres, "Exploring CBIR concepts in the CTRnet Project," Technical Report IC-10-32, Institute of Computing, University of Campinas, 2010.
[101] L. T. Li and R. d. S. Torres, "Geospatial Information," in *Digital Libraries: A 5S Approach*, E. A. Fox, Ed. Blacksburg, VA: Virginia Tech (in 502 page textbook used in CS6604, in process to appear first as 4 smaller books and then as a single large book, published by Morgan-Claypool, 2012-2013), 2011.
[102] S. H. Park, V. Srinivasan, and P. Angara, "Text Extraction," in *Digital Libraries: A 5S Approach*, E. A. Fox, Ed. Blacksburg, VA: Virginia Tech (in 502 page textbook used in CS6604, in process to appear first as 4 smaller books and then as a single large book, published by Morgan-Claypool, 2012-2013), 2011, p. Ch. 15.
[103] N. a. P. Kozievitch and R. d. S. Torres, "Complex Objects," in *Digital Libraries: A 5S Approach*, E. A. Fox, Ed. Blacksburg, VA: Virginia Tech (in 502 page textbook used in CS6604, in process to appear first as 4 smaller books and then as a single large book, published by Morgan-Claypool, 2012-2013), 2011, p. Ch. 4.
[104] U. Murthy, L. M. Delcambre, R. d. S. Torres, and N. a. P. Kozievitch, "Subdocuments," in *Digital Libraries: A 5S Approach*, E. A. Fox, Ed. Blacksburg, VA: Virginia Tech (in 502 page textbook used in CS6604, in process to appear first as 4 smaller books and then as a single large book, published by Morgan-Claypool, 2012-2013), 2011, p. Ch. 6.
[105] U. Murthy and E. Fox. (2006). *The Superimposed Information Project at Virginia Tech (homepage)*. Available: http://si.dlib.vt.edu/
[106] F. A. Das Neves and E. A. Fox. (2003). *Extending retrieval with stepping stones and pathways (web site for funded NSF proposal; see also first year report of Aug. 18)* [NSF Proposal]. Available: http://fox.cs.vt.edu/SSP/
[107] F. A. Das Neves, "Stepping Stones and Pathways:Improving Retrieval by Chains of Relationships between Documents," Doctoral dissertation, Computer Science, Virginia Tech, Blacksburg, VA, 2004. http://scholar.lib.vt.edu/theses/available/etd-11012004-003013/restricted/dissertation.PDF
[108] F. Das Neves, E. Fox, and X. Yu, "Connecting topics in document collections with stepping stones and pathways," in Proceedings of the 14th ACM international conference on information and knowledge management, 2005, pp. 91-98.

[109] X. Yu, F. Das-Neves, and E. A. Fox, "Hard Queries can be Addressed with Query Splitting Plus Stepping Stones and Pathways," *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering,* vol. 28, no. 4, pp. 29-38, 2005. http://sites.computer.org/debull/A05dec/yu.pdf

[110] E. A. Fox, F. Das Neves, X. Y. Yu, R. Shen, S. Kim, and W. Fan, "Exploring the computing literature with visualization and stepping stones & pathways," *Communications of the ACM,* vol. 49, no. 4, pp. 52-58, Apr 2006. http://doi.acm.org/10.1145/1121949.1121982

[111] L. Venkatachalam, "Scalability of Stepping Stones and Pathways," MS Thesis, Dept. of Computer Science, Virginia Tech, Blacksburg, VA, 2008. http://scholar.lib.vt.edu/theses/available/etd-05072008-225923/

[112] N. I. ElSherbiny, "Security," in *Digital Libraries: A 5S Approach*, E. A. Fox, Ed. Blacksburg, VA: Virginia Tech (in 502 page textbook used in CS6604, in process to appear first as 4 smaller books and then as a single large book, published by Morgan-Claypool, 2012-2013), 2011, p. Ch. 14.

[113] N. I. ElSherbiny, "Secure Digital Libraries," MS thesis, Dept. of Computer Science, Virginia Tech, Blacksburg, VA, 2011. http://scholar.lib.vt.edu/theses/available/etd-06302011-161547/

[114] E. A. Fox and N. I. ElSherbiny, "Security and Digital Libraries," in *Digital Libraries: Methods and Applications*, K. H. Huang, Ed. Rijeka, Croatia: InTech, 2011, pp. 151-160. http://www.intechopen.com

[115] PhaseVis. (2012). *Four Phases Visualization Prototype*. Available: http://spare05.dlib.vt.edu/~ctrvis/phasevis

[116] M. E. Baird, "The "Phases" of Emergency Management," ed. Background Paper Prepared for the Intermodal Freight Transportation Institute (IFTI) University of Memphis, 2010. http://www.memphis.edu/cait/pdfs/Phases_of_Emergency_Mngt_FINAL.pdf

[117] D. M. Neal, "Reconsidering the phases of disasters," *International Journal of Mass Emergencies and Disasters,* vol. 15, no. 2, pp. 239-264, 1997.

[118] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "ThemeRiver: Visualizing Thematic Changes in Large Document Collections," *IEEE Transactions on Visualization and Computer Graphics,* vol. 8, no. 1, pp. 9-20, Jan. 2002 2002.

[119] CRED. (2011, September 26). *EM-DAT the International Disaster Database*. Available: http://www.emdat.be/

[120] DisasterDatabaseProject. (2011, December 1). *University of Richmond, The Disaster Database Project*. Available: http://learning.richmond.edu/disaster/index.cfm

[121] Canada. (2011, September 26). *Canadian Disaster Database*. Available: http://www.publicsafety.gc.ca/prg/em/cdd/srch-eng.aspx

[122] DesInventar. (2011, September 26). *Inventory system of the effects of disasters*. Available: http://www.desinventar.org/

[123] CTRnet. (2012). *Hurricane Sandy Web Collection*. Available: http://archive-it.org/collections/3358

[124] Hurricane_Sandy. (2012). *Local Online Groups for Community Recovery from Hurricane Sandy*. Available: https://docs.google.com/document/d/1OIHiEz1jtZnGJqdBgXS9Y273-DRunmpCa-TNrkC7oU4/edit

[125] E. Fox, "Implementing SMART for Minicomputers Via Relational Processing with Abstract Data Types," *ACM SIGSMALL Newsletter (Jt. Proc. SIGSMALL Symp. on Small Systems and SIGMOD Workshop on Small Data Base Systems, Orlando, FL, Oct. 1981),* vol. 7, no. 2, pp. 119-129, 1981.

[126] E. A. Fox, "Some Considerations for Implementing the SMART Information Retrieval System under UNIX," Cornell Univ. Dept. of Comp. Science, Ithaca, NY, Technical report TR 83-560, Sept. 1983.
[127] E. A. Fox, R. K. France, E. Sahle, A. Daoud, and B. E. Cline, "Development of a Modern OPAC: From REVTOLC to MARIAN," in *Proc. 16th Annual Int'l ACM SIGIR Conf. on R&D in Information Retrieval, SIGIR '93* Pittsburgh: ACM Press, 1993, pp. 248-259. http://www.acm.org/pubs/articles/proceedings/ir/160688/p248-fox/p248-fox.pdf
[128] R. K. France, M. A. Goncalves, and E. A. Fox. (2002). *MARIAN Digital Library Information System (home page)*. Available: http://www.dlib.vt.edu/products/marian.html
[129] M. A. Goncalves, R. K. France, E. A. Fox, and T. E. Doszkocs, "MARIAN: Searching and Querying Across Heterogeneous Federated Digital Libraries," in *Proceedings of the First DELOS Network of Excellence Workshop on Information Seeking, Searching and Querying in Digital Libraries, Dec. 11-12, Zurich, Switzerland*: DELOS, 2000. http://www.ndltd.org/talks/delosfox.ppt
[130] M. A. Goncalves, P. Mather, J. Wang, Y. Zhou, M. Luo, R. Richardson, . . . E. A. Fox, "Java MARIAN: From an OPAC to a Modern Digital Library System," in *Proceedings of 9th String Processing and Information Retrieval Symposium (SPIRE 2002), September 11-13, Lisbon, Portugal*, 2002, pp. 194-209.
[131] J. Zhao, "Making Digital Libraries Flexible, Scalable, and Reliable: Reengineering the MARIAN System in JAVA," Master of Science Thesis, Department of Computer Science, Virginia Tech, Blacksburg, VA, 1999. http://scholar.lib.vt.edu/theses/available/etd-070499-204531/unrestricted/SGML-etd/
[132] E. A. Fox, "Envision-ing a Computer Science Digital Library," in Digital Libraries of the Future Panel, ACM Multimedia 93, Anaheim, CA, 1993.
[133] E. A. Fox, "A Digital Library Connecting Envision, KMS, and Mosaic with Interfaces, Communications, and Data Interchange," *SIGOIS Bulletin,* vol. 15, no. 1, p. 6, 1994.
[134] E. A. Fox, D. Hix, L. Nowell, D. Brueni, W. Wake, L. Heath, and D. Rao, "Users, User Interfaces, and Objects: Envision, a Digital Library," *J. American Society Information Science,* vol. 44, no. 8, pp. 480-491, 1993.
[135] L. S. Heath, D. Hix, L. T. Nowell, W. C. Wake, G. A. Averboch, E. Labow, . . . E. A. Fox, "Envision: A User-Centered Database of Computer Science Literature," *Communications of the ACM,* vol. 38, no. 4, pp. 52-53, April 1995. http://doi.acm.org/10.1145/205323.376383
[136] L. Nowell and D. Hix, "User interface design for the project Envision database of computer science literature," in *Proc. Twenty-second Annual Virginia Computer Users Conference (NHVCUC 92), Blacksburg, VA*, 1992, pp. 29-33.
[137] L. Nowell and D. Hix, "Visualizing search results: User interface development for the project Envision database of computer science literature," in *Proceedings of HCI International '93, 5th International Conference on Human Computer Interaction, Advances in Human Factors/Ergonomics*. vol. 19B: Elsevier, 1993, pp. 56-61.
[138] L. T. Nowell and E. A. Fox, "Envision: Information Visualization in a Digital Library (demonstration)," in *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'95, Seattle, WA, July 10*: ACM, 1995, p. 365.
[139] L. T. Nowell, R. K. France, and E. A. Fox, "Visualizing search results with Envision (demonstration)," in *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval, ACM SIGIR'96, Zurich, Switzerland*: ACM, 1996, pp. 338-339.

[140] J. Wang, A. Agrawal, A. Bazaz, S. Angle, E. A. Fox, and C. North, "Enhancing the ENVISION Interface for Digital Libraries," in *Proc. JCDL'2002, Second ACM / IEEE-CS Joint Conference on Digital Libraries, July 14-18, Portland, Oregon*: ACM Press, 2002, pp. 275-276. http://doi.acm.org/10.1145/544220.544279

[141] CITIDEL. (2001). *Computing and Information Technology Interactive Digital Educational Library, CITIDEL Homepage* [Web site]. Available: http://www.citidel.org

[142] CITIDEL. (2004). *Virginia Instructional Architect for Digital Undergraduate Computing Teaching (VIADUCT)*. Available: http://www.citidel.org/?op=viaduct_front

[143] E. Fox, "Advancing Education through Digital Libraries: NSDL, CITIDEL, and NDLTD," in Digital Library: IT Opportunities and Challenges in the New Millennium, Beijing, China, 2002, pp. 107-117.

[144] E. Fox and K. Garach, "CITIDEL Collection Building," Virginia Tech, Blacksburg, VA, Technical Report TR-03-14, 2003. http://eprints.cs.vt.edu/archive/00000660/

[145] E. A. Fox, "Case Studies in the US National Science Digital Library: DL-in-a-Box, CITIDEL, and OCKHAM," in 6th International Conference on Asian Digital Libraries (ICADL 2003), Kuala Lumpur, Malaysia, 2003, pp. 17-25.

[146] J. Impagliazzo, "Using CITIDEL as a Portal for IT Education," in Informing Sciences Conference, Cork, Ireland, 2002.

[147] J. Impagliazzo, "Using CITIDEL as a Portal for CS Education," in CCSCNE Conference, 2002.

[148] J. Impagliazzo, L. Cassel, and D. Knox, "Using CITIDEL as a Portal for CS Education," *Journal of Computing in Small Colleges,* vol. 17, no. 6, pp. 161-163, 2002.

[149] J. Impagliazzo, D. Knox, and L. Cassel, "Using the NSDL and CITIDEL to Enhance Your Teaching," in Innovation and Technology in Computer Science Education (ITiCSE), University of Macedonia, Thessaloniki, Greece, 2003.

[150] J. Impagliazzo, J. Lee, and L. Cassel, "Enhancing Distance Learning Using Quality Digital Libraries and CITIDEL," in Quality Education at a Distance - IFIP Distance Learning Conference, Deakin University, Geelong, Australia, 2003.

[151] N. Kampanya, R. Shen, S. Kim, C. North, and E. A. Fox, "Citiviz: A Visual User Interface to the CITIDEL System," in *Research and Advanced Technology for Digital Libraries: Proc. 8th European Conference on Digital Libraries, ECDL 2004, September 12-17, University of Bath, UK, Lecture Notes in Computer Science, vol. 3232*, R. Heery and L. Lyon, Eds.: Springer-Verlag GmbH, Berlin, 2004, pp. 122-133. http://www.springerlink.com/openurl.asp?genre=article&issn=0302-9743&volume=3232&spage=122

[152] J. A. N. Lee, J. Impagliazzo, L. N. Cassel, E. A. Fox, C. L. Giles, D. Knox, and M. A. Pérez-Quiñones, "Enhancing distance learning using quality digital libraries and CITIDEL," in *Quality Education @ a Distance*, 2003, pp. 61-71.

[153] S. Perugini, K. McDevitt, R. Richardson, M. Perez-Quinones, R. Shen, N. Ramakrishnan, . . . E. A. Fox, "Enhancing Usability in CITIDEL: Multimodal, Multilingual, and Interactive Visualization Interfaces," in *Proceedings Fourth ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL2004), Tucson, AZ, June 7-11*: ACM, 2004, pp. 315-324.

[154] E. Tressler, "Crawlifying and Resource Discovery for CITIDEL," Virginia Tech Department of Computer Science, Blacksburg, VA, Undergraduate Research Report, May 2002.

[155] B. Zhang, M. A. Goncalves, and E. A. Fox, "An OAI-Based Filtering Service for CITIDEL from NDLTD," in *Proceedings 6th International Conference on Asian Digital Libraries, ICADL 2003, Digital Libraries: Technology and Management of Indigenous Knowledge for Global Access; Kuala Lumpur, Malaysia, Dec. 8-12; Springer, Lecture Notes in*

Computer Science 2911, T. Mohd, T. Sembok, H. B. Zaman, H. Chen, S. R. Urs, and S. H. Myaeng, Eds., 2003, pp. 590-601.

[156] J. W. Flanagan, E. A. Fox, and W. Fan. (2003). *Managing complex information applications: An archaeology digital library, ETANA-DL homepage*. Available: http://digbase.etana.org

[157] U. Ravindranathan, "Prototyping Digital Libraries Handling Heterogeneous Data Sources - An ETANA-DL Case Study," Master's Thesis, Computer Science, Virginia Tech, Blacksburg, VA, 2004. http://scholar.lib.vt.edu/theses/available/etd-04262004-153555/

[158] U. Ravindranathan, R. Shen, M. Goncalves, W. Fan, E. A. Fox, and J. W. Flanagan, "ETANA-DL: managing complex information applications – an archaeology digital library (Demo)," in *Proceedings of the Fourth ACM-IEEE Joint Conference on Digital Libraries - Global Reach and Diverse Impact, JCDL 2004, June 7-11, Tucson, AZ*, H. Chen, M. Christel, and E. P. Lim, Eds. Tucson, AZ, 2004, p. 414. http://feathers.dlib.vt.edu/

[159] U. Ravindranathan, R. Shen, M. A. Goncalves, W. Fan, E. A. Fox, and J. W. Flanagan, "Prototyping Digital Libraries Handling Heterogeneous Data Sources – The ETANA-DL Case Study," in *Research and Advanced Technology for Digital Libraries: Proc. 8th European Conference on Digital Libraries, ECDL2004, September 12-17, U. Bath, UK, Lecture Notes in Computer Science, vol. 3232*, R. Heery and L. Lyon, Eds.: Springer-Verlag GmbH, Berlin, 2004, pp. 186-197. http://springerlink.metapress.com/openurl.asp?genre=article&issn=0302-9743&volume=3232&spage=186

[160] A. Raghavan, "Schema Mapper: A Visualization Tool for Incremental Semi-automatic Mapping-based Integration of Heterogeneous Collections into Archaeological Digital Libraries: The ETANA-DL Case Study," MS thesis, Computer Science, Virginia Tech, Blacksburg, VA 24061, 2005. http://scholar.lib.vt.edu/theses/available/etd-05182005-114155/

[161] R. Shen, M. A. Goncalves, W. Fan, and E. A. Fox, "Requirements Gathering and Modeling of Domain-Specific Digital Libraries with the 5S Framework: An Archaeological Case Study with ETANA," in *Proc. European Conference on Digital Libraries, ECDL 2005, Vienna, Sept. 18-23*: Springer, 2005, pp. 1-12. http://dx.doi.org/10.1007/11551362_1, http://fox.cs.vt.edu/talks/2005/20050919ECDLmodeling.ppt

[162] E. A. Fox and N. S. Vemuri, "ETANA-DL: Leveraging DL Technologies to Support Archaeology," in Presentation in ETANA (Electronic Tools and Ancient Near Eastern Archives) Workshop I, Nov. 17; Edward A. Fox and Linda Cantara, co-chairs, ETANA Workshop II, Nov. 18, in ASOR 2006, Washington, DC, 2006.

[163] D. Gorton, R. Shen, N. S. Vemuri, W. Fan, and E. A. Fox, "ETANA-GIS: GIS for archaeological digital libraries," in Proc. 6th ACM/IEEE Joint Conference on Digital Libraries (JCDL), June 11-15, 2006, Chapel Hill, NC, 2006, pp. 379-379.

[164] J. L. S. Rajkumar, "ETANA-CMV: A coordinated multiple view visual browsing interface for ETANA-DL," MS thesis, Computer Science, Virginia Tech, Blacksburg, VA, 2006. http://scholar.lib.vt.edu/theses/available/etd-01052007-154321/

[165] N. S. Vemuri, R. Shen, S. Tupe, W. Fan, and E. A. Fox, "ETANA-ADD: An Interactive Tool for Integrating Archaeological DL Collections," in Proc. 6th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2006), June 11-15, Chapel Hill, NC, 2006, pp. 161-162.

[166] R. Shen, N. S. Vemuri, W. Fan, and E. A. Fox, "Integration of Complex Archaeology Digital Libraries: An ETANA-DL Experience," *Information Systems,* vol. 33, no. 7-8, pp. 699-723, Nov.-Dec. 2008 2008.
[167] Ensemble. (2009). *Ensemble Distributed Digital Library for Computing Education homepage*. Available: http://www.computingportal.org
[168] P. Brusilovsky, L. Cassel, L. Delcambre, E. Fox, R. Furuta, D. D. Garcia, . . . M. Yudelson, "Enhancing Digital Libraries with Social Navigation: The Case of Ensemble," in Research and Advanced Technology for Digital Libraries, Proc. 14th European Conference, ECDL2010, Glasgow, Sept. 6-10, 2010, pp. 116-123.
[169] E. A. Fox, Y. Chen, M. Akbar, C. A. Shaffer, S. H. Edwards, P. Brusilovsky, . . . L. Cassel, "Ensemble PDP-8: Eight Principles for Distributed Portals," in Proc. JCDL/ICADL 2010, June 21-25, Gold Coast, Australia, 2010, pp. 341-344.
[170] P. Tanapaisankit, M. Song, and E. A. Fox, "Developing a Concept Extraction Technique with Ensemble Pathway," in Proceedings of JCDL 2011, Ottawa, June 13-17, 2011, pp. 405-406.
[171] F. M. Shipman, L. Cassel, E. Fox, R. Furuta, L. Delcambre, P. Brusilovsky, . . . D. D. Garcia, "Ensemble: A Distributed Portal for the Distributed Community of Computing Education (refereed poster)," in Research and Advanced Technology for Digital Libraries, Proc. 14th European Conference, ECDL2010, Glasgow, Sept. 6-10, 2010, pp. 506-509.
[172] L. B. Cassel, E. Fox, F. Shipman, P. Brusilovsky, W. Fan, D. Garcia, . . . S. Potluri, "Ensemble: enriching communities and collections to support education in computing: poster session, Consortium for Computing Sciences in Colleges," *Journal of Computing Sciences in Colleges,* vol. 25, no. 6, pp. 224-226, June 2010.
[173] E. A. Fox, M. Akbar, Y. Chen, M. Stewart, C. A. Shaffer, S. H. Edwards, and P. Fan, "Ensemble: Enriching Communities and Collections to Support Education in Computing (reviewed poster)," in 2010 Conference on Higher Education Pedagogy, Inn at Virginia Tech, Feb. 18-19, 2010, p. 102. http://www.cider.vt.edu/conference/2010/2010proceedings.pdf
[174] Y. Chen, E. Fox, S. Edwards, and C. Shaffer, "The National Science Digital Library resources to enhance learning: lessons from CITIDEL and the Ensemble pathways project (refereed poster 66)," in Conference on Higher Education Pedagogy @ VT, February 18, Virginia Tech, 2009.
[175] B. S. CarpenterII, R. Furuta, F. Shipman, A. Huie, D. Pogue, E. A. Fox, . . . L. M. L. Delcambre, "Multiple Sources with Multiple Portals: A Demonstration of The Ensemble Computing Portal in Second Life," in JCDL 2010, June 21-25, Gold Coast, Australia, 2010, pp. 397-398.
[176] G. W. Hislop, L. N. Cassel, R. Furuta, L. M. L. Delcambre, P. Brusilovsky, E. A. Fox, and D. D. Garcia, "Ensemble - the online community center for computing educators: demonstration. June 2010. Consortium for Computing Sciences in Colleges," *Journal of Computing Sciences in Colleges,* vol. 25, no. 6, pp. 74-75, June 2010.
[177] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature,* vol. 457, no. 7232, pp. 1012-1014, 2009. http://dx.doi.org/10.1038/nature07634
[178] M. Hu, A. Sun, and E.-P. Lim, "Event detection with common user interests," in Proceedings of the 10th ACM workshop on Web information and data management, Napa Valley, California, USA, 2008, pp. 1-8.
[179] C. Wang, M. Zhang, L. Ru, and S. Ma, "Automatic online news topic ranking using media focus and user attention based on aging theory," in Proceedings of the 17th ACM conference on information and knowledge management, Napa Valley, California, USA, 2008, pp. 1033-1042.

[180] Y. Chen, S. Yang, and X. Cheng, "Bursty topics extraction for web forums," in Proceedings of the 11th international workshop on Web information and data management, Hong Kong, China, 2009, pp. 55-58.
[181] S. Phuvipadawat and T. Murata, "Breaking News Detection and Tracking in Twitter," in Proceedings of the IEEE/WIC/ACM Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010, pp. 120-123.
[182] W. J. Corvey, S. Vieweg, T. Rood, and M. Palmer, "Twitter in mass emergency: what NLP techniques can contribute," in Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, Los Angeles, California, 2010, pp. 23-24.
[183] K. Starbird and L. Palen, ""Voluntweeters": self-organizing by digital volunteers in times of crisis," in Proceedings of the 2011 annual conference on human factors in computing systems Vancouver, BC, Canada, 2011, pp. 1071-1080.
[184] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: what twitter may contribute to situational awareness," in Proceedings of the 28th international conference on human factors in computing systems, Atlanta, Georgia, USA, 2010, pp. 1079-1088.
[185] J. Allan, A. Feng, and A. Bolivar, "Flexible intrinsic evaluation of hierarchical clustering for TDT," in Proceedings of the 12th international conference on Information and knowledge management, New Orleans, LA, USA, 2003, pp. 263-270.
[186] B. G. Ahn, B. V. Durme, and C. Callison-Burch, "WikiTopics: what is popular on Wikipedia and why," in Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, Portland, Oregon, 2011, pp. 33-40.
[187] Z.-L. Wu and C.-h. Li, "Topic Detection in Online Discussion Using Non-negative Matrix Factorization," in Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops, 2007, pp. 272-275.
[188] P. Fung, G. Ngai, and C.-S. Cheung, "Combining optimal clustering and Hidden Markov models for extractive summarization," in Proceedings of the ACL 2003 workshop on multilingual summarization and question answering - Volume 12, Sapporo, Japan, 2003, pp. 21-28.
[189] V. Hatzivassiloglou, L. Gravano, and A. Maganti, "An investigation of linguistic features and clustering algorithms for topical document clustering," in Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval Athens, Greece, 2000, pp. 224-231.
[190] K. Y. Kamath and J. Caverlee, "Transient crowd discovery on the real-time social web," in Proceedings of the 4th ACM international conference on Web search and data mining, Hong Kong, China, 2011, pp. 585-594.
[191] C.-C. Pan and P. Mitra, "Event detection with spatial latent Dirichlet allocation," in Proceedings of the 11th annual international ACM/IEEE joint conference on digital libraries, Ottawa, Ontario, Canada, 2011, pp. 349-358.
[192] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, Seattle, Washington, USA, 2006, pp. 178-185.
[193] D. Li, B. He, Y. Ding, J. Tang, C. Sugimoto, Z. Qin, . . . T. Dong, "Community-based topic modeling for social tagging," in Proceedings of the 19th ACM international conference on information and knowledge management, Toronto, ON, Canada, 2010, pp. 1565-1568.
[194] T. Masada, D. Fukagawa, A. Takasu, T. Hamada, Y. Shibata, and K. Oguri, "Dynamic hyperparameter optimization for Bayesian topical trend analysis," in Proceedings of the 18th ACM conference on information and knowledge management, Hong Kong, China, 2009, pp. 1831-1834.

[195] D. Li, S. Somasundaran, and A. Chakraborty, "A combination of topic models with max-margin learning for relation detection," in Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing, Portland, Oregon, 2011, pp. 1-9.
[196] K. Farrahi and D. Gatica-Perez, "Discovering routines from large-scale human locations using probabilistic topic models," *ACM Trans. Intell. Syst. Technol.,* vol. 2, no. 1, pp. 1-27, 2011.
[197] A. Pozdnoukhov and C. Kaiser, "Space-time dynamics of topics in streaming text," in Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, Chicago, Illinois, 2011, pp. 1-8.
[198] B. B. Klebanov, E. Beigman, and D. Diermeier, "Discourse topics and metaphors," in Proceedings of the Workshop on Computational Approaches to Linguistic Creativity, Boulder, Colorado, 2009, pp. 1-8.
[199] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.,* vol. 3, pp. 993-1022, 2003.
[200] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in Proceedings of the 1st Workshop on Social Media Analytics, Washington D.C., 2010, pp. 80-88.
[201] Y. Liu, A. Niculescu-Mizil, and W. Gryc, "Topic-link LDA: joint models of topic and author community," in Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, Quebec, Canada, 2009, pp. 665-672.
[202] D. Andrzejewski and X. Zhu, "Latent Dirichlet Allocation with topic-in-set knowledge," in Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, Boulder, Colorado, 2009, pp. 43-48.
[203] J. T. Chien and C. H. Chueh, "Topic-Based Hierarchical Segmentation," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 20, no. 1, pp. 55-66, 2012.
[204] Z. Liu, M. Li, Y. Liu, and M. Ponraj, "Performance evaluation of Latent Dirichlet Allocation in text mining," in Proceedings of the 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2011), Shanghai, China, 2011, pp. 2695 -2698.
[205] D. Lan, W. L. Buntine, and J. Huidong, "Sequential Latent Dirichlet Allocation: Discover Underlying Topic Structures within a Document," in Proceedings of the IEEE 10th International Conference on Data Mining (ICDM 2010), Sydney, Australia, 2010, pp. 148-157.
[206] W. Hongjun, L. Zhishu, and C. Yang, "Weighted Latent Dirichlet Allocation for Cluster Ensemble," in Proceedings of the 2nd International Conference on Genetic and Evolutionary Computing (WGEC '08), 2008, pp. 437-441.
[207] NIST. (2011, December 1). *Topic Detection and Tracking Evaluation*. Available: http://www.nist.gov/speech/tests/tdt/
[208] G. Cselle, K. Albrecht, and R. Wattenhofer, "BuzzTrack: topic detection and tracking in email," in Proceedings of the 12th international conference on intelligent user interfaces, Honolulu, Hawaii, USA, 2007, pp. 190-197.
[209] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi, "Topic detection and tracking with spatio-temporal evidence," in Proceedings of the 25th European conference on IR research, Pisa, Italy, 2003, pp. 251-265.
[210] R. Nallapati, "Semantic language models for topic detection and tracking," in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Proceedings of the HLT-NAACL 2003 student research workshop - Volume 3, Edmonton, Canada, 2003, pp. 1-6.
[211] C. L. Wayne, "Topic detection and tracking in English and Chinese," in Proceedings of the 5th international workshop on on information retrieval with Asian languages, Hong Kong, China, 2000, pp. 165-172.

[212] V. Lavrenko, J. Allan, E. DeGuzman, D. LaFlamme, V. Pollard, and S. Thomas, "Relevance models for topic detection and tracking," in Proceedings of the 2nd international conference on Human Language Technology Research, San Diego, California, 2002, pp. 115-121.
[213] M. Mori, T. Miura, and I. Shioya, "Topic Detection and Tracking for News Web Pages," in Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, 2006, pp. 338-342.
[214] A. J. Kurtz and J. Mostafa, "Topic detection and interest tracking in a dynamic online news source," in Proceedings of the 3rd ACM/IEEE-CS joint conference on digital libraries, Houston, Texas, 2003, pp. 122-124.
[215] B. Li, W. Li, and Q. Lu, "Topic tracking with time granularity reasoning," vol. 5, no. 4, pp. 388-412, 2006.
[216] C. Wang, M. Zhang, S. Ma, and L. Ru, "Automatic online news issue construction in web environment," in Proceedings of the 17th international conference on World Wide Web, Beijing, China, 2008, pp. 457-466.
[217] R. Schirru, "Topic-based recommendations in enterprise social media sharing platforms," in Proceedings of the 4th ACM conference on recommender systems, Barcelona, Spain, 2010, pp. 369-372.
[218] C. Shah, W. B. Croft, and D. Jensen, "Representing documents with named entities for story link detection (SLD)," in Proceedings of the 15th ACM international conference on information and knowledge management, Arlington, Virginia, USA, 2006, pp. 868-869.
[219] F. Fukumoto and Y. Suzuki, "Topic tracking based on bilingual comparable corpora and semisupervised clustering," vol. 6, no. 3, p. 11, 2007.
[220] A. Leuski and J. Allan, "Improving realism of topic tracking evaluation," in Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, Tampere, Finland, 2002, pp. 89-96.
[221] M. Franz, T. Ward, J. S. McCarley, and W.-J. Zhu, "Unsupervised and supervised clustering for topic tracking," in Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, New Orleans, Louisiana, United States, 2001, pp. 310-317.
[222] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw, "Combining the Evidence of Multiple Query Representations for Information Retrieval," *Information Processing and Management,* vol. 31, no. 3, pp. 431-448, May-June 1995.
[223] E. A. Fox and J. A. Shaw, "Combination of multiple searches," in *The Proceedings of the Second Text REtrieval Conference (TREC-2)*, D. K. Harman, Ed.: National Institute for Standards and Technology, NIST Special Publication 500215, 1994, pp. 243-252.
[224] C. C. Vogt and G. W. Cottrell, "Fusion Via a Linear Combination of Scores," *Information Retrieval,* vol. 1, pp. 151-173, 1999.
[225] E. M. Voorhees, N. K. Gupta, and B. Johnson-Laird, "Learning collection fusion strategies," in Proceedings of the 18th Annual International ACM SIGIR Conference on research and development in Information Retrieval, July 9 - 13, 1995, Seattle, WA, 1995, pp. 172-179. http://www.acm.org/pubs/citations/proceedings/ir/215206/p172-voorhees/
[226] W. Xi and E. Fox, "SimFusion: Measuring Similarity using Unified Relationship Matrix," in Proceedings of the 28th annual intl ACM SIGIR conf on R&D in information retrieval, 2005.
[227] W. Xi, B. Zhang, Z. Chen, Y. Lu, S. Yan, W.-Y. Ma, and E. A. Fox, "Link Fusion: A Unified Link Analysis Framework for Multi-type Inter-related Data Objects," in *Proc Thirteenth Intl World Wide Web Conf, WWW2004, New York, U.S.A. 19-22 May*, 2004, pp. 319-327.

[228] B. Zhang, Y. Chen, W. Fan, E. A. Fox, M. A. Goncalves, M. Cristo, and P. Calado, "Intelligent Fusion of Structural and Citation-Based Evidence for Text Classification (poster)," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2005, Salvador, Brazil*, 2005, pp. 667-668. http://doi.acm.org/10.1145/1076034.1076181

[229] B. Zhang, Y. Chen, W. Fan, E. A. Fox, M. A. Gonçalves, M. Cristo, and P. Calado, "Intelligent GP fusion from multiple sources for text classification," in 2005 ACM International Conference on Information and Knowlege Management, 2005, pp. 477-484.

[230] B. Zhang, M. A. Goncalves, W. Fan, Y. Chen, E. A. Fox, P. Calado, and M. Cristo, "Intelligent Fusion of Structural and Citation-Based Evidence for Text Classification," Virginia Tech, Blacksburg, VA, Technical Report TR-04-16, 2004. http://eprints.cs.vt.edu/archive/00000693/

[231] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, . . . S. Sigelman, "Tracking and summarizing news on a daily basis with Columbia's Newsblaster," in Proceedings of the 2nd international conference on Human Language Technology Research, San Diego, California, 2002, pp. 280-285.

[232] R. K. Pon, A. F. Cardenas, D. Buttler, and T. Critchlow, "Tracking multiple topics for finding interesting articles," in Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining, San Jose, California, USA, 2007, pp. 560-569.

[233] P. Duygulu, J.-Y. Pan, and D. A. Forsyth, "Towards auto-documentary: tracking the evolution of news stories," in Proceedings of the 12th annual ACM international conference on multimedia New York, NY, USA, 2004, pp. 820-827.

[234] D. Frey, R. Gupta, V. Khandelwal, V. Lavrenko, A. Leuski, and J. Allan, "Monitoring the news: a TDT demonstration system," in Proceedings of the 1st international conference on human language technology research, San Diego, 2001, pp. 1-5.

[235] R. D. Brown, "A server for real-time event tracking in news," in Proceedings of the 1st international conference on human language technology research, San Diego, 2001, pp. 1-3.

[236] C. Flynn and J. Dunnion, "Topic Detection in the news domain," in Proceedings of the 2004 international symposium on Information and communication technologies, Las Vegas, Nevada, 2004, pp. 103-108.

[237] A. Pons-Porrata, R. Berlanga-Llavori, and J. Ruiz-Shulcloper, "Building a hierarchy of events and topics for newspaper digital libraries," in Proceedings of the 25th European conference on IR research, Pisa, Italy, 2003, pp. 588-596.

[238] B. Li, W. Li, Q. Lu, and M. Wu, "Profile-based event tracking," in Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, Salvador, Brazil, 2005, pp. 631-632.

[239] P. Kim and S. H. Myaeng, "Usefulness of temporal information automatically extracted from news articles for topic tracking," vol. 3, no. 4, pp. 227-242, 2004.

[240] B. Chen, H.-M. Wang, and L.-S. Lee, "A discriminative HMM/N-gram-based retrieval approach for Mandarin spoken documents," vol. 3, no. 2, pp. 128-145, 2004.

[241] Q. Ma and K. Tanaka, "Topic-structure-based complementary information retrieval and its application," vol. 4, no. 4, pp. 475-503, 2005.

[242] D. Mimno and A. McCallum, "Expertise modeling for matching papers with reviewers," in Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining, San Jose, California, USA, 2007, pp. 500-509.

[243] G. S. Mann, D. Mimno, and A. McCallum, "Bibliometric impact measures leveraging topic analysis," in Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries, Chapel Hill, NC, USA, 2006, pp. 65-74.

[244] S.-H. Lin and B. Chen, "Topic modeling for spoken document retrieval using word- and syllable-level information," in Proceedings of the 3rd workshop on searching spontaneous conversational speech, Beijing, China, 2009, pp. 3-10.
[245] L. Li, B. Roth, and C. Sporleder, "Topic models for word sense disambiguation and token-based idiom detection," in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 1138-1147.
[246] X. Jin, S. Spangler, R. Ma, and J. Han, "Topic initiator detection on the World Wide Web," in Proceedings of the 19th international conference on World Wide Web, Raleigh, North Carolina, USA, 2010, pp. 481-490.
[247] X. Li, J. Yan, W. Fan, N. Liu, S. Yan, and Z. Chen, "An online blog reading system by topic clustering and personalized ranking," *ACM Trans. Internet Technol.,* vol. 9, no. 3, pp. 1-26, 2009.
[248] L. Wang, J. Lin, and D. Metzler, "A cascade ranking model for efficient ranked retrieval," in Proceedings of the 34th international ACM SIGIR conference on research and development in information, Beijing, China, 2011, pp. 105-114.
[249] C. Brandt, T. Joachims, Y. Yue, and J. Bank, "Dynamic ranked retrieval," in Proceedings of the 4th ACM international conference on Web search and data mining, Hong Kong, China, 2011, pp. 247-256.
[250] T. Elsayed, J. Lin, and D. Metzler, "When close enough is good enough: approximate positional indexes for efficient ranked retrieval," in Proceedings of the 20th ACM international conference on information and knowledge management, Glasgow, Scotland, UK, 2011, pp. 1993-1996.
[251] B. T. Bartell, G. W. Cottrell, and R. K. Belew, "Automatic combination of multiple ranked retrieval systems," in Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval, Dublin, Ireland, 1994, pp. 173-181.
[252] A. Lad and Y. Yang, "Learning to rank relevant and novel documents through user feedback," in Proceedings of the 19th ACM international conference on information and knowledge management, Toronto, ON, Canada, 2010, pp. 469-478.
[253] G.-R. Xue, H.-J. Zeng, Z. Chen, Y. Yu, W.-Y. Ma, W. Xi, and W. Fan, "Optimizing web search using web click-through data," in Proceedings of the 13th ACM international conference on information and knowledge management, Washington, D.C., USA, 2004, pp. 118-126.
[254] L. Chen, Y. Hu, and W. Nejdl, "Using subspace analysis for event detection from web click-through data," in Proceedings of the 17th international conference on World Wide Web, Beijing, China, 2008, pp. 1067-1068.
[255] G. Pass, A. Chowdhury, and C. Torgeson, "A picture of search," in Proceedings of the 1st international conference on scalable information systems, Hong Kong, 2006, p. 1.
[256] Amazon. (2011). *Amazon Mechanical Turk*. Available: http://www.mturk.com
[257] Internet_Archive. (2000). *Internet Archive* [Web page]. Available: http://www.archive.org
[258] S. Chakrabarti, M. Van den Berg, and B. Dom, "Focused crawling: a new approach to topic-specific Web resource discovery," *Computer Networks,* vol. 31, no. 11-16, pp. 1623-1640, 1999.
[259] S. Chakrabarti, K. Punera, and M. Subramanyam, "Accelerated focused crawling through online relevance feedback," in Proceedings of the 11th International World Wide Web Conference, Honolulu, Hawaii, USA, 2002, pp. 148-159. http://doi.acm.org/10.1145/511446.511466

[260] Z. Zhuang, R. Wagle, and C. L. Giles, "What's there and what's not?: Focused crawling for missing documents in digital libraries," in Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries (JCDL 2005), Denver, CO, USA, 2005, pp. 301-310.
[261] L. Dongfei and L. Jia, "PPSpider: Towards an Efficient and Robust Topic-Specific Crawler Based on Peer-to-Peer Network," in Proceedings of the 2nd International Workshop on Computer Science and Engineering (WCSE), 2009, pp. 101-105.
[262] Y. Guojun, X. Xiaoyao, and L. Zhijie, "The design and realization of open-source search engine based on Nutch," in Proceedings of the International Conference on Anti-Counterfeiting Security and Identification in Communication (ASID), 2010, pp. 176-180.
[263] J. Xiaolong, J. Jianmin, and M. Geyong, "Automatic digital preservation solutions enabled by Web services and intelligent agents," in Proceedings of eChallenges, Warsaw, Poland, 2010, pp. 1-8.
[264] C. Fellbaum, *WordNet: An electronic lexical database*. Cambridge, Massachusetts: The MIT Press, 1998.
[265] D. Hati, B. Sahoo, and A. Kumar, "Adaptive focused crawling based on link analysis," in Proceedings of the 2nd International Conference on Education Technology and Computer (ICETC), 2010, pp. V4-455-V4-460.
[266] G. Almpanidis and C. Kotropoulos, "Combining text and link analysis for focused crawling," in Proc. (Part 1) Pattern Recognition and Data Mining, Third International Conference on Advances in Pattern Recognition, ICAPR 2005, Bath, UK, August 22-25, Springer LNCS 3686, 2005, pp. 278-287.
[267] G. Pant, "Deriving link-context from HTML tag tree," in Proceedings of the 8th ACM SIGMOD workshop on research issues in data mining and knowledge discovery (DMKD), San Diego, California, USA, 2003, pp. 49-55.
[268] A. Pal, D. S. Tomar, and S. Shrivastava, "Effective Focused Crawling Based on Content and Link Structure Analysis," *arXiv preprint arXiv:0906.5034,* 2009.
[269] C. Xiaoyun and Z. Xin, "HAWK: A Focused Crawler with Content and Link Analysis," in Proceedings of the IEEE International Conference on e-Business Engineering (ICEBE), Xi'an, China, 2008, pp. 677-680.
[270] G. Pant and P. Srinivasan, "Link contexts in classifier-guided topical crawlers," *IEEE Transactions on Knowledge and Data Engineering,* vol. 18, no. 1, pp. 107-122, 2006.
[271] M. Jamali, H. Sayyadi, B. B. Hariri, and H. Abolhassani, "A method for focused crawling using combination of link structure and content similarity," in Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI), Hong Kong, China, 2006, pp. 753-756.
[272] C. Wang, Z. Guan, C. Chen, J. Bu, J. Wang, and H. Lin, "On-line topical importance estimation: an effective focused crawling algorithm combining link and content analysis," *Journal of Zhejiang University-Science A,* vol. 10, no. 8, pp. 1114-1124, 2009.
[273] A. Batzios, C. Dimou, A. L. Symeonidis, and P. A. Mitkas, "BioCrawler: An intelligent crawler for the semantic web," *Expert Systems with Applications,* vol. 35, no. 1-2, pp. 524-530, 2008.
[274] A. B. Can and N. Baykal, "MedicoPort: A medical search engine for all," *Computer methods and programs in biomedicine,* vol. 86, no. 1, pp. 73-86, 2007.
[275] C. Cesarano, A. d'Acierno, and A. Picariello, "An intelligent search agent system for semantic information retrieval on the internet," in Proceedings of the 5th ACM international workshop on Web information and data management (WIDM), New Orleans, Lousiana, USA, 2003.
[276] S. Chang, G. Yang, Y. Jianmei, and L. Bin, "An efficient adaptive focused crawler based on ontology learning," in Proceedings of the 5th International Conference on Hybrid Intelligent Systems (HIS), 2005, pp. 73-78.

[277] Y.-J. Chen and V.-W. Soo, "Ontology-Based Information Gathering Agents," in Proceedings of the 1st Asia-Pacific Conference on Web Intelligence: Research and Development 2001, pp. 423-427.
[278] M. Ehrig and A. Maedche, "Ontology-focused crawling of Web documents," in Proceedings of the 2003 ACM symposium on applied computing, Melbourne, Florida, 2003, pp. 1174-1178.
[279] S. Ganesh, M. Jayaraj, V. Kalyan, S. Murthy, and G. Aghila, "Ontology-based Web Crawler," in Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC), Las Vegas, Nevada, USA, 2004, p. 337.
[280] A. Maedche, M. Ehrig, S. Handschuh, L. Stojanovic, and R. Volz. (2002). *Ontology-Focused Crawling of Web Documents and RDF-based Metadata*. Available: http://projekte.l3s.uni-hannover.de/pub/bscw.cgi/S4893f6f4/d5269/Maedche_Ehrig-Focused_Crawler-ISWC2002sub.pdf
[281] J. Tane, C. Schmitz, and G. Stumme, "Semantic resource management for the web: an e-learning application," in Proceedings of the 13th international World Wide Web conference, alternate track papers & posters, New York, NY, USA, 2004, pp. 1-10.
[282] H. Liu, E. Milios, and J. Janssen, "Focused Crawling by Learning HMM from User's Topic-specific Browsing," in Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI), Beijing, China, 2004, pp. 732-732.
[283] H. Liu, E. Milios, and J. Janssen, "Probabilistic models for focused web crawling," in Proceedings of the 6th annual ACM international workshop on Web information and data management (WIDM) Washington D.C., U.S.A, 2004, pp. 16-22. http://doi.acm.org/10.1145/1031453.1031458
[284] M. Yuvarani, N. C. S. N. Iyengar, and A. Kannan, "LSCrawler: A Framework for an Enhanced Focused Web Crawler Based on Link Semantics," in Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, 2006, pp. 794-800.
[285] G. de Assis, A. Laender, M. Gonçalves, and A. da Silva, "Exploiting Genre in Focused Crawling," in *String Processing and Information Retrieval*. vol. 4726, N. Ziviani and R. Baeza-Yates, Eds.: Springer Berlin / Heidelberg, 2007, pp. 62-73. http://dx.doi.org/10.1007/978-3-540-75530-2_6
[286] P. Braslavski, "Marrying Relevance and Genre Rankings: An Exploratory Study," in *Genres on the Web*. vol. 42, A. Mehler, S. Sharoff, and M. Santini, Eds.: Springer Netherlands, 2011, pp. 191-208. http://dx.doi.org/10.1007/978-90-481-9178-9_9
[287] W. R. Van Hage, V. Malaisé, R. Segers, L. Hollink, and G. Schreiber, "Design and use of the Simple Event Model (SEM)," *Web Semantics: Science, Services and Agents on the World Wide Web,* vol. 9, no. 2, pp. 128-136, 2011.
[288] T. Peng, C. Zhang, and W. Zuo, "Tunneling enhanced by web page content block partition for focused crawling," *Concurrency and Computation: Practice and Experience,* vol. 20, no. 1, pp. 61-74, 2008.
[289] N. Luo, W. L. Zuo, and F. Y. Yuan, "Gray Tunneling Based on Block Relevance for Focused Crawling," in Proceedings on Intelligent Systems and Knowledge Engineering, ISKE 2007.
[290] N. Luo, W. Zuo, F. Yuan, and C. Zhang, "A New Method for Focused Crawler Cross Tunnel," in *Rough Sets and Knowledge Technology*. vol. 4062, G.-Y. Wang, J. Peters, A. Skowron, and Y. Yao, Eds.: Springer Berlin / Heidelberg, 2006, pp. 632-637. http://dx.doi.org/10.1007/11795131_92
[291] R. L. Winkler, *An introduction to Bayesian inference and decision*. New York: Holt, Rinehart and Winston, 1972.

[292] P. Srinivasan, F. Menczer, and G. Pant, "Defining evaluation methodologies for topical crawlers," in Proceedings of the SIGIR Workshop on Defining Evaluation Methodologies for Terabyte-Scale Collections, Toronto, Canada, 2003.
[293] A. Cami and N. Deo, "Evaluation of a Graph-based Topical Crawler," in International Conference on Internet Computing, 2006, pp. 393-399.
[294] P. Srinivasan, F. Menczer, and G. Pant, "A general evaluation framework for topical crawlers," *Information Retrieval,* vol. 8, no. 3, pp. 417-447, 2005.
[295] F. Menczer, G. Pant, and P. Srinivasan, "Topical web crawlers: Evaluating adaptive algorithms," *ACM Transactions on Internet Technology (TOIT),* vol. 4, no. 4, pp. 378-419, 2004.
[296] C. J. Van Rijsbergen, *Information Retrieval*. London: Butterworth-Heinemann, 1979. http://www.dcs.gla.ac.uk/Keith/Preface.html
[297] F. Tak-chung, D. C. M. Sze, P. K. C. Leung, H. Kei-yuen, and C. Fu-lai, "Analysis and Visualization of Time Series Data from Consumer-Generated Media and News Archives," in Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Silicon Valley, USA, 2007, pp. 259-262.
[298] RAPIDMiner. (2011, December 1). *Open Source System for Data Mining*. Available: http://rapid-i.com/content/view/181/190/
[299] ORANGE. (2011, December 1). *Open Source Visualization and Analysis Tool*. Available: http://orange.biolab.si/
[300] NLTK. (2011, December 1). *Natural Language Toolkit*. Available: http://www.nltk.org/
[301] GATE. (2011, December 1). *The University of Sheffield Open, GATE, Source Solution for Text Processing*. Available: http://gate.ac.uk/
[302] OpenEphyra. (2011, December 1). *The Ephyra Question Answering System*. Available: http://www.ephyra.info/
[303] OSQA. (2011, December 1). *The Open Source Q&A System*. Available: http://www.osqa.net/
[304] SentiStrength. (2011, December 1). *SentiStrength is a sentiment analysis (opinion mining) program.* Available: http://sentistrength.wlv.ac.uk/
[305] OpinionFinder. (2011, December 1). *OpinionFinder System*. Available: http://www.cs.pitt.edu/mpqa/opinionfinder.html
[306] MEAD. (2011, December 1). *MEAD is a public domain portable multi-document summarization system*. Available: http://www.summarization.com/mead/
[307] N. Rotem. (2011, December 1). *Open Text Summarizer*. Available: http://libots.sourceforge.net/
[308] Extractor. (2011, December 1). *The World of Relevant Information in the Palm of Your Hand*. Available: http://www.extractor.com/
[309] L. K. Kit, C. K. Man, and E. Baniassad, "Isolating and relating concerns in requirements using latent semantic analysis," *SIGPLAN Not.,* vol. 41, no. 10, pp. 383-396, 2006.
[310] C.-L. Yeh and Y.-C. Chen, "Creation of topic map by identifying topic chain in Chinese," in Proceedings of the 2004 ACM symposium on document engineering, Milwaukee, Wisconsin, USA, 2004, pp. 112-114.
[311] D. Wolfram and H. A. Olson, *A method for comparing large scale inter-indexer consistency using IR modeling*. New York: ACM Press, 2007.
[312] S. K. M. Wong, Z. Wojciech, and P. Wong, "Generalized vector space model in information retrieval," in *the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Montreal, Canada, Year, pp. 18-25.

[313] G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," *Communications of the ACM,* vol. 18, no. 11, pp. 613--620, 1975-11,November 1975 1975. http://doi.acm.org/10.1145/361219.361220
[314] TF-IDF. (2012). *TF-IDF weighting scheme*. Available: http://en.wikipedia.org/wiki/Tf%E2%80%93idf
[315] W3C. (2012, December 12,). *Resource Description Framework (RDF): Concepts and Abstract Syntax*. Available: http://www.w3.org/TR/rdf-concepts/
[316] I. Niles and A. Pease, "Towards a standard upper ontology," in Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001, Ogunquit, Maine, USA, 2001, pp. 2-9.
[317] P. Velardi, P. Fabriani, and M. Missikoff, "Using text processing techniques to automatically enrich a domain ontology," in Proceedings of the international conference on Formal Ontology in Information Systems, Ogunquit, Maine, USA, 2001, pp. 270-284.
[318] H.-T. Zheng, C. Borchert, and H.-G. Kim, "Exploiting corpus-related ontologies for conceptualizing document corpora," *Journal of the American Society for Information Science and Technology,* vol. 60, no. 11, pp. 2287-2299, 2009. http://dx.doi.org/10.1002/asi.21145
[319] S. C. Gates, W. Teiken, and K.-S. F. Cheng, "Taxonomies by the numbers: building high-performance taxonomies," in Proceedings of the 14th ACM international conference on information and knowledge management, Bremen, Germany, 2005, pp. 568-577.
[320] X. Jin, Y. Li, T. Mah, and J. Tong, "Sensitive webpage classification for content advertising," in Proceedings of the 1st international workshop on data mining and audience intelligence for advertising, San Jose, California, 2007, pp. 28-33.
[321] A. Platt, S. S. R. Mengle, and N. Goharian, "Improving classification based off-topic search detection via category relationships," in Proceedings of the 2009 ACM symposium on Applied Computing, Honolulu, Hawaii, 2009, pp. 869-874.
[322] F. Peng, D. Schuurmans, and S. Wang, "Language and task independent text categorization with simple language models," in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, Edmonton, Canada, 2003, pp. 110-117.
[323] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu, "Parameter free bursty events detection in text streams," in Proceedings of the 31st international conference on very large data bases, Trondheim, Norway, 2005, pp. 181-192.
[324] B. Al-Ani, G. Mark, and B. Semaan, "Blogging in a region of conflict: supporting transition to recovery," in Proceedings of the 28th international conference on Human factors in computing systems, Atlanta, Georgia, USA, 2010, pp. 1069-1078.
[325] B. Al-Ani, G. Mark, and B. Semaan, "Blogging through conflict: sojourners in the age of social media," in Proceedings of the 3rd international conference on Intercultural collaboration, Copenhagen, Denmark, 2010, pp. 29-38.
[326] T. Nasukawa and J. Yi, "Sentiment analysis: capturing favorability using natural language processing," in Proceedings of the 2nd international conference on knowledge capture, Sanibel Island, FL, USA, 2003, pp. 70-77.
[327] K. Eguchi and V. Lavrenko, "Sentiment retrieval using generative models," in Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, 2006, pp. 345-354.
[328] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comput. Linguist.,* vol. 37, no. 2, pp. 267-307, 2011.
[329] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach," in Proceedings of the 20th

ACM international conference on information and knowledge management, Glasgow, Scotland, UK, 2011, pp. 1031-1040.
[330] N. O'Hare, M. Davy, A. Bermingham, P. Ferguson, P. Sheridan, C. Gurrin, and A. F. Smeaton, "Topic-dependent sentiment analysis of financial blogs," in Proceedings of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion, Hong Kong, China, 2009, pp. 9-16.
[331] T. T. Thet, J.-C. Na, C. S. G. Khoo, and S. Shakthikumar, "Sentiment analysis of movie reviews on discussion boards using a linguistic approach," in Proceedings of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion, Hong Kong, China, 2009, pp. 81-84.
[332] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002, pp. 417-424.
[333] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in Proceedings of the ACL-02 conference on empirical methods in natural language processing - Volume 10, 2002, pp. 79-86.
[334] C. Lin, Y. He, and R. Everson, "A comparative study of Bayesian models for unsupervised sentiment detection," in Proceedings of the 14th Conference on Computational Natural Language Learning, Uppsala, Sweden, 2010, pp. 144-152.
[335] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, . . . S. Patwardhan, "OpinionFinder: a system for subjectivity analysis," in Proceedings of HLT/EMNLP on Interactive Demonstrations, Vancouver, British Columbia, Canada, 2005, pp. 34-35.
[336] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment in short strength detection informal text," *J. Am. Soc. Inf. Sci. Technol.,* vol. 61, no. 12, pp. 2544-2558, 2010.
[337] J. Park, B. Ahn, R. Myung, K. Lim, W. Lee, and M. Cha, "Revolution 2.0 in Tunisia and Egypt: Reactions and sentiments in the online world," in Proc. ICWSM-11, Barcelona, 17-21 July, 2011. http://www.icwsm.org/2011/documents/IDC2011.pdf
[338] M. Khalifa. (2011). *The role of Information Technology in defeating the Arab regimes: Facebook 2-0 Arab Presidents*. Available: http://www.ifla.org/files/faife/Spotlight_1.pdf
[339] J. Ghannam. (2011). *Social Media in the Arab World: Leading up to the Uprisings of 2011: A Report to the Center for International Media Assistance*. Available: http://cima.ned.org/publications/social-media-arab-world-leading-uprisings-2011-0
[340] E. T. Meyer, A. Thomas, and R. Schroeder. (2011). *Web Archives: The Future(s)*. Available: http://ssrn.com/abstract=1830025
[341] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in Proceedings of the IEEE Symposium on Visual Languages, Boulder, CO, USA, 1996, pp. 336-343.
[342] R. Spence, *Information Visualization: Design for Interaction (2nd Edition)*: Prentice-Hall, Inc., 2007.
[343] E. R. Tufte, *Beautiful evidence*: Graphics Press, Cheshire, CT, 2006.
[344] E. R. Tufte, *Envisioning information* vol. 21: Graphics Press, Cheshire, CT, 1990.
[345] E. R. Tufte, *The visual display of quantitative information (2nd ed.)* vol. 7: Graphics Press, Cheshire, CT, 2001.
[346] E. R. Tufte, *Visual explanations: images and quantities, evidence and narrative*. Cheshire, CT: Graphics Press, 1997.

[347] R. Lengler and M. J. Eppler, "Towards a periodic table of visualization methods for management," *Proceedings of Graphics and Visualization in Engineering (GVE 2007),* 2007.
[348] K. Börner and C. Chen, Eds., *Visual Interfaces to Digital Libraries*. Springer Verlag LNCS 2539, 2002.
[349] L. Nowell, "Graphical Encoding for Information Visualization: Using Icon Color, Shape and Size to Convey Nominal and Quantitative Data," Ph.D. Dissertation, Dept. of Computer Science, Virginia Tech, Blacksburg, VA, 1997.
[350] L. Nowell, E. A. Fox, L. Heath, D. Hix, W. Wake, and E. Labow, "Seeing Things Your Way: Information Visualization for a User-Centered Database of Computer Science Literature," Virginia Tech Dept. of Computer Science, Blacksburg, VA, Technical Report TR-94-06, January 1994.
[351] R. Shen, J. Wang, and E. A. Fox, "A Lightweight Protocol between Digital Libraries and Visualization Systems," in *JCDL Workshop on Visual Interfaces to Digital Libraries, Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries* Portland, USA: ACM Press, 2002, p. 425.
[352] J. Wang, *VIDI: A lightweight protocol between visualization systems and digital libraries*. Blacksburg, VA: Virginia Tech, Department of Computer Science Masters thesis, 2002. http://scholar.lib.vt.edu/theses/available/etd-07012002-145841/
[353] E. A. Fox, G. Abdulla, and W. Heagy, Eds., *Quantitative Analysis and Visualization Regarding Interactive Learning with a Digital Library in Computer Science, ACM Digital Libraries '97 (256 pages)*. Philadelphia, PA: ACM, 1997.
[354] Twitter. (2012). *Using the Twitter Search API*. Available: https://dev.twitter.com/docs/using-search
[355] Twitter. (2012). *The Streaming APIs*. Available: https://dev.twitter.com/docs/streaming-apis
[356] A. D. Robbins, "GAWK: Effective AWK Programming," *Boston: Free Software Foundation,* vol. 3, 2004.
[357] M. J. Norusis, *SPSS advanced statistics 6.1*: SPSS, 1994.
[358] SRF. (2011). *NodeXL: Network Overview, Discovery and Exploration for Excel*. Available: http://nodexl.codeplex.com/
[359] D. Hansen, B. Shneiderman, and M. A. Smith, *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*. Burlington, MA: Morgan Kaufmann, 2010.
[360] A. Smith, "Leximancer (Vsn. 2.21)," *Brisbane, Australia: University of Queensland, Centre for Human Factors and Applied Cognitive Psychology,* 2005.
[361] N. Peladeau, "WordStat 5.0 [computer software]," *Montreal: Provalis Research,* 2005.
[362] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," in Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media California, USA, 2009, pp. 361-362.
[363] A. Bruns. (2011, December 1). *Visualising Twitter Dynamics in Gephi, Part 1*. Available: http://www.mappingonlinepublics.net/2010/12/30/visualising-twitter-dynamics-in-gephi-part-1/
[364] C. North, "Information Visualization," in *Handbook of Human Factors and Ergonomics, 3rd Edition*, G. Salvendy, Ed. New York: John Wiley & Sons, 2005, pp. 1222-1246.
[365] J. S. Risch, D. B. Rex, S. T. Dowson, T. B. Walters, R. A. May, and B. D. Moon, "The STARLIGHT Information Visualization System," in IEEE International Conference on Information Visualization, London, 1997, pp. 42-49.
[366] S. K. Card, J. D. Mackinlay, and B. Shneiderman, *Readings in information visualization: using vision to think*. San Francisco: Morgan Kaufmann, 1999.

[367] F. A. Das Neves and E. A. Fox, "A study of user behavior in an immersive virtual environment for digital libraries," in *Proceedings of the Fifth ACM Conference on Digital Libraries: DL '00, June 2-7, 2000, San Antonio, TX* New York: ACM Press, 2000, pp. 103-111.
[368] VTCAVE. (2011). *VT CAVE Information Center*. Available: http://www.cave.vt.edu/
[369] A. Sabri, R. Ball, S. Bhatia, A. Fabian, and C. North, "High-Resolution Gaming: Interfaces, Notifications and the User Experience," *Computer Games (Special Issue on HCI Issues),* vol. 19, no. 2, pp. 151-166, March 2007.
[370] R. Ball and C. North, "Realizing Embodied Interaction for Visual Analytics through Large Displays," *Computers & Graphics (C&G),* vol. 31, no. 3, pp. 380-400, 2007.
[371] R. Ball and C. North, "An Analysis of User Behavior on High-Resolution Tiled Displays," in Tenth IFIP International Conference on Human-Computer Interaction (INTERACT 2005), 2005, pp. 350-364.
[372] R. Ball, M. Varghese, B. Carstensen, E. D. Cox, C. Fierer, M. Peterson, and C. North, "Evaluating the Benefits of Tiled Displays for Navigating Maps," in IASTED International Conference on Human-Computer Interaction, 2005.
[373] C. North. (2011). *Gigapixel Display Laboratory*. Available: http://www.cs.vt.edu/labs/gigapixel
[374] L. Dali and D. Mladenic, "Visualization of Web Page Content Using Semantic Technologies," in Proceedings of the 14th International Conference on Information Visualisation (IV), London, UK, 2010, pp. 280-284.
[375] K. H. Huang, C. Tsai, and C. C. Cheng, "Information visualization of the digital archives: implementing an exploratory learning environment with GIS interface," in Proceedings of the International Conference on Active Media Technology (AMT), Takamatsu, Kagawa, Japan, 2005, pp. 497-498.
[376] S. B. Liu and L. Palen, "The new cartographers: Crisis map mashups and the emergence of neogeographic practice," *Cartography and Geographic Information Science,* vol. 37, no. 1, pp. 69-90, 2010.
[377] V. Kumar, R. Furuta, and R. B. Allen, "Metadata visualization for digital libraries: interactive timeline editing and review," in *Proceedings of the Third ACM Conference on Digital Libraries*, 1998, pp. 126-133.
[378] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. Keim, "EventRiver: Visually Exploring Text Collections With Temporal References," *IEEE Transactions on Visualization and Computer Graphics,* vol. 18, no. 1, pp. 93-105, 2012.
[379] S. Hinton and M. Whitelaw, "Exploring the digital commons: an approach to the visualisation of large heritage datasets," in Electronic Visualization and the Arts (EVA), London, 5-7 July 2010. http://ewic.bcs.org/content/ConWebDoc/36049
[380] E. Rogers and D. Kincaid, *Communication Networks: Toward a New Pradigm for Research*. New York: The Free Press, 1981.
[381] B. Nardi, *Context and Consciousness: Activity Theory and Human-computer Interaction*. Cambridge, MA: MIT Press, 1995.
[382] L. Suchman, *Plans and Situated Actions: the problem of human-machine communication*. Cambridge, England: Cambridge Press, 1987.
[383] J. W. Cresswell, *Qualitative Inquiry and Research Design: Choosing among Five Approaches*. Thousand Oaks, California: Sage Publications, Inc., 2007.
[384] C. Glesne, *Becoming qualitative researchers: An Introduction*. White Plains, New York: Longman, 1992.
[385] U. Murthy, "Digital Libraries with Superimposed Information: Supporting Scholarly Tasks that Involve Fine Grain Information," PhD Dissertation, Dept. of Computer Science,

Virginia Tech, Blacksburg, VA, 2011. http://scholar.lib.vt.edu/theses/available/etd-04142011-175752/

[386] U. Murthy, L. T. Li, E. Hallerman, E. A. Fox, M. A. Perez-Quinones, L. M. Delcambre, and R. d. S. Torres, "Use of subimages in fish species identification: a qualitative study," in Proceedings of the 11th annual international ACM/IEEE joint conference on digital libraries, Ottawa, Ontario, Canada, 2011, pp. 185-194. http://doi.acm.org/10.1145/1998076.1998112

[387] U. Murthy, L. T. Li, and A. Kavanaugh, "What if anyone in the world can be a participant in your user study?," in Proceedings of Grace Hopper Celebration of Women In Computing (GHC 2011), 9-12 Nov., Portland, OR, 2011.

[388] E. A. Fox, M. Samaka, C. L. Giles, and J. Impagliazzo, "Funded proposal to Qatar National Research Fund, Project No. NPRP 4-029-1-007: Establishing a Qatari Arabic-English Library Institute," Virginia Tech, Blacksburg, VA, 2010.

[389] K. Beck, M. Beedle, A. v. Bennekum, A. Cockburn, W. Cunningham, M. Fowler, . . . D. Thomas, "Manifesto for Agile Software Development," 2001. http://www.agilemanifesto.org

EDWARD A. FOX

(fox@vt.edu, http://fox.cs.vt.edu/)

A. Professional Preparation

MIT	Electrical Engineering (Computer Science Option), B.S., 1972
Cornell	Computer Science, M.S., 1981; Ph.D., 1983

B. Appointments

1/98-	Director, Digital Library Research Laboratory, VPI&SU (Virginia Tech)
4/95-	Professor, Dept. of Computer Science, VPI&SU (Virginia Tech), 24061 USA
6/90-12/02	Associate Director for Research, VPI&SU (Virginia Tech) Computing Center
5/88-4/95	Associate Professor, Dept. of Computer Science, VPI&SU
9/83-5/88	Assistant Professor, Dept. of Computer Science, VPI&SU
8/82-4/83	Manager of Information Systems, Intl Inst. Tropical Agriculture, Ibadan, Nigeria
8/78-8/82	Instructor, Research Assistant, Teaching Assistant, Dept. of CS, Cornell
6/72-7/78	Data Processing Manager, Vulcraft, Div. of NUCOR Corp., Florence, SC
7/71-6/72	Data Processing Instructor, Florence Darlington Technical College

C. Publications (over 440, with h-index 49 according to Google Scholar)

Publications (Selected Related):

1. Kavanaugh, A.L., Fox, E.A., Sheetz, S.D., Yang, S., Li, L.T., Whalen, T., Shoemaker, D. J., Natsev, P., Xie L. Social Media Use by Government: From the Routine to the Critical. Government Information Quarterly (GIQ) 29(4): 480-491, 2012
2. Kavanaugh, A.L., Sheetz, S.D., Hassan, R., Yang, S., Elmongui, H.G., Fox, E.A., Magdy, M., Shoemaker, D. Between a Rock and a Cell Phone: Communication and Information Use During the Egyptian Uprising. Proceedings of the 9th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2012). Apr. 22-25, 2012. Vancouver
3. Seungwon Yang, Kiran Chitturi, Gregory Wilson, Mohamed Magdy, and Edward A. Fox. A Study of Automation from Seed URL Generation to Focused Web Archive Development: The CTRnet Context. Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2012), Washington D.C., June 10-14, 2012, 341-342
4. Seungwon Yang, Andrea Kavanaugh, Nadia P. Kozievitch, Lin Tzy Li, Venkat Srinivasan, Steven D. Sheetz, Travis Whalen, Donald Shoemaker, Ricardo da A. Torres, and Edward A. Fox. CTRnet DL for Disaster Information Services. Proceedings of JCDL 2011, Ottawa, June 13-17, 2011, 437-438
5. Sheetz, Steve, Fox, Edward A., Fitzgerald, A., Palmer, S., Shoemaker, D., Kavanaugh, Andrea. Why Students Use Social Networking Sites After Crisis Situations. Proceedings of the 8th International ISCRAM Conference, Lisbon, Portugal, May 8-11, 2011

Publications (Selected Other):

1. Rao Shen, Marcos Andre Goncalves, and Edward A. Fox. Key Issues Regarding Digital Libraries: Evaluation and Integration. Morgan & Claypool Publishers, San Francisco, 2013 (in press)
2. Edward A. Fox, Marcos Andre Goncalves, and Rao Shen. Theoretical Foundations for Digital Libraries: The 5S (Societies, Scenarios, Spaces, Structures, Streams) Approach. Morgan & Claypool Publishers, San Francisco, 2012
3. Uma Murthy, Edward Fox, Naren Ramakrishnan, Andrea Kavanaugh, et al. Building an Ontology for Crisis, Tragedy, and Recovery. NKOS Workshop, ECDL 2009, 1 Oct. 2009, Corfu, Greece
4. Edward A. Fox et al., A Digital Library for Recovery, Research, and Learning from April 16, 2007 at Virginia Tech. Traumatology, 14(1): 64-84, 2008
5. Hui Han, C. Lee Giles, Eren Manavoglu, Hongyuan Zha, Zhenyue Zhang, and Edward A. Fox. Automatic Document Metadata Extraction using Support Vector Machines. Proc. JCDL'2003, Third Joint ACM / IEEE-CS Joint Conference on Digital Libraries, May 27-31, 2003, Houston, 37-48

D. Synergistic Activities

1. Work on disaster computing: Member, Steering Committee, NSF-funded CCC Vision Workshop on Computing for Disaster Management, Washington, D.C., April 24-25, 2012; speaker for webinars on 7/24/2012 and 11/7/2012; PI on two NSF funded projects and co-PI on another project
2. Service to computing research: Member, Board of Directors, Computing Research Association
3. Service to digital libraries since 1991, including General Chair, JCDL2001; Program Chair ACM DL 1996 and 1999; Chair, Steering Committee, Joint Conf on Digital Libraries (JCDL); Chairman, IEEE Tech. Comm. Digital Libraries; Executive Director, Networked Digital Library of Theses and Dissertations; Member, Steering Committee, Intl Conf of Asian Digital Libraries (ICADL); Member, JISC Advisory Group, Programmes on Digital Repositories, Digital Asset Management & Preservation; Member, Advisory Board, EU's DELOS Network of Excellence on Digital Libraries
4. Service to information retrieval since 1985, including Vice-chair and Chair, 1987-1995, ACM Special Interest Group on Info. Retrieval (SIGIR); program chair SIGIR'95; many program committees
5. Editorial Board service (current): ACM Trans. Info. Systems, ACM Computers in Entertainment, Inf. Proc. Mgmt., Int. J. on Digital Libraries, JEMH, JIIS, JUCS, MTAP, TOISJ

E. Collaborators & Other Affiliations

Collaborators and Co-Editors (selected recent, other than listed below; see also DBLP list)

A. Abbott (VT), J. Almeida (UNICAMP), C. Andrews (VT), D. Archer (Portland St.), G. Athanasopoulos (U. Athens), R. Beck (Villanova), P. Bogen (ORNL), L. Boutwell (VT), S. Britell (Portland St.), P. Brusilovski (U. Pitt.), W. Cameron (Villanova), S. Carpenter (Penn. St.), L. Cassel (Villanova), H. Chen (U. Arizona), H. Chung (VT), W. Chung (UNC Fayetteville), F. Corubolo (U. Trento), F. Decker (Portland St.), L. Delcambre (Portland St.), S. Edwards (VT), A. Falcão (UNICAMP), T. Falcao (UNICAMP), W. Fan (VT), B. Friedman (VT), R. Furuta (TAMU), D. Garcia (Berkeley), C.L. Giles (Penn St.), K. Goldman (Google), K. Hall (VT), E. Hallerman (VT), K. Hanna (Internet Archive), J. Heines (U. Mass.), G. Hislop (Drexel), K. Hoyle (VT), M. Hsiao (VT), H. Hsieh (Iowa), J. Impagliazzo (Hofstra), Y. Ioannidis (U. Athens), J. Jeffers (U. Mass.), J. Jiao (VT), G. Kakaletris (U. Athens), A. Kassahun (VT), A. Kavanaugh (VT), K. Kuchibhotla (Villanova), A. Laender (UFMG, Brazil), N. Leite (UNICAMP), J. Levy (VT), S.C. Lu (VT), N. Lynberg (VT), N. Ma (VT), D. Maier (Portland St.), N. Manola (U. Athens), M. Marathe (VT), G. Marchionini (UNC-CH), J. McCall (Robert Gordon U.), P. McElmurray (VT), G. McMillan (VT), C. Meghini (ISTI CNR), K. Miller (VT), S. Misra (VT), B. Moreira (UFMG), H. Mortveit (VT), S. Murthy (Portland St.), A. Natsev (IBM), C. North (VT), S. Oh (UNC-CH), J.P. Papa (UNICAMP), M. Perez-Quinones (VT), J. Pomerantz (UNC-CH), S. Price (Portland St.), J. Racer (VT), K. Raheb (U. Athens), N. Ramakrishnan (VT), E. Ramos (UNICAMP), A. Rauber (TU Vienna), A. Scarpa (VT), C. Shaffer (VT), S. Sheetz (VT), F. Shipman (TAMU), D. Shoemaker (VT), N. Short (VT), B. Siegfried (Villanova), D. Soergel (U. Buffalo), R. Suryavanshi (Villanova), M. Tungare (VT), J. Velasco-Martin (UNC-CH), G. Wang (VT), L. Watson (VT), T. Whalen (VT), B. Wildemuth (UNC-CH), K. Williams (Cape Town), W. Xi (Google), L. Xie (Australian National U.), Y. Yuan (Villanova), M. Yudelson (CMU), B. Zhang (Microsoft)

Graduate Advisor: G. Salton (deceased 1995)

Thesis Advisor (>40): G. Abdulla, M. Akbar, S. Betrabet, J. Bourne, Q. Chen, Y. Chen, A. Daoud, F. Das Neves, N. ElSherbiny, S. Feizbadi, R. France, M. Goncalves, D. Gorton , R. Kelapure, T. Kanan, S. Kim, N. Kipp, A. Krowne, S. Lee, W. Lee, J. Leidig, B. Liu, M. Luo, Y. Ma, U. Murthy, S. Park, A. Raghavan, D. Rangarajan, U. Ravindranathan, R. Richardson, J. Shaw, R. Shen, O. Sornil, V. Srinivasan, H. Suleman, L. Tinoco, N. Vemuri, L. Venkatachalam, J. Wang, L. Wang, M. Weaver, S. Winett, S. Yang, X. Yu, B. Zhang, J. Zhao, Y. Zhou, Q. Zhu

Postgraduate-Scholar Sponsor (4): R. Gaur (MDI, Gurgaon), Nádia P. Kozievitch (UNICAMP), Lin Tzy Li (UNICAMP), D. Madali (Indian Stat. Inst.), R. da S. Torres (UNICAMP), S. Urs (U. Mysore)

Kristine Hanna, Director, Archiving Services

SUMMARY OF QUALIFICATIONS

- 20+ years project management experience on large scale and high profile programs and projects
- 13+ years of experience managing and developing client and partner relationships
- 13+ years of experience managing the development of Web products and services
- Nearly 6 years of experience managing Internet Archive web archiving projects
- Co-founded GirlGeeks.com, a new media company, and built it into a global Web brand with Fortune 1000 clients and a thriving online community.
- Received "Women Who Make a Difference" Award by the California State Senate.
- Nominated twice for an Emmy Award from the Academy of Television, Arts, and Sciences for "Outstanding Visual Effects" for The Young Indiana Jones Chronicles.
- Attended USC School of Cinema and Television

2006 – Present **Internet Archive; San Francisco, CA**

Director, Archiving Services

Works with memory institutions to develop Web archiving services and solutions that will help preserve the worldwide Web and digital content. Collaborates with partners and leads an internal Web development team to build and manage programs and projects. Directs the Archive-It subscription program used by over 180 memory institutions around the world. Leads a team that has grown the self-sustaining service from 0 to 180 partners with recurring revenue of $1,000,000 and growing each month. Acted as Product Manager for 10 Web application releases of the Archive-It Web application. Works with the partners and the internal engineering team to determine features, functionality, schedule, and timing. Acts as a digital steward, interacting on NDIIPP/NSDA programs to increase awareness of Web archiving and digital preservation. Works with these same partners on the sustainability for these programs as well as identifying "at risk content categories". Acts as an evangelist for digital archiving, speaking at conferences in the US and around the world and sitting on steering committees to encourage other institutions to harvest and preserve their digital collections.

2004 – 2005 **Digital Think; San Francisco, CA**

Senior Manager, Creative Services

Successfully managed projects from concept to delivery, including personnel, budget, scope, schedule, technical specifications, escalations and client satisfaction. Drove decision-making and manage risks for successful execution. Liaisoned with clients to determine business, learning, and design requirements. Clients included Microsoft and Intel. 3D Simulation received 2005 Brandon Hall Excellence in Learning Award for Custom Content. Defined and spearheaded strategy/business objectives for new video/flash and 3D simulations product offerings. Developed and implemented a plan for the organization, with successful results. Hired, trained, managed, and motivated a highly skilled team of Creative Directors and Designers, developed their skill sets and created goals that paralleled their career paths. Managed cross functionally in a matrix environment with internal departments, remote employees, and a global workforce.

2002 – 2004 **Austin, TX**

Senior Director, Content and Talent Development

Directed the Marketing department, conceiving and producing online media, including: Webcasts, Webinars, streaming videos and other content to successfully execute client's recruiting and staffing business objectives. Clients included Dell, Boeing, EA, Adobe, and Roche Diagnostics. Conceptualized and defined a strategy for diversity recruiting Web-based products and services for clients. Collaborated with Product Development, Marketing, and existing client base.

1997 – 2002 **GirlGeeks; San Francisco, CA**

CEO and Founder

Founded the company and established the vision of GirlGeeks, working with a team to develop the look and feel of the brand, messaging, and content for Website, Web community, and Web services and products. Collaborated with the management team of six to conceptualize, produce and analyze revenue-producing training and career products and services for a technically-savvy consumer base and Fortune 1000 clients. Produced online marketing campaigns and offline events for Fortune 1000 companies—both online and traditional. Clients included Microsoft, Intel, IBM, Dell, Cisco, Motorola, and Hewlett-Packard. Hired and trained managers, technical staff, and creative teams. Coached and developed them to optimal performance as individuals and as a team. Hired, fired and conducted performance reviews. Project Manager for three large scale Websites, including design, architecture, features, functionality, content, community, user experience, and site optimization. Received numerous awards, including "Best Bay Area Website" by American Women in Radio and Television. Executive Producer of online animation, video and audio segments, 3D simulations and live interactive Webcasts, seminars; as well as traditional offline seminars at conferences and tradeshows.

1992 – 1997 **Lucasfilm, Inc.; San Rafael, CA**

Visual Effects Producer

Budgeted $1M projects and supervised post production, successfully delivering 40 one-hour television episodes and 5 two-hour TV specials. Produced visual effects, collaborating with Executive Producer and Designers to bring the overall storyline to life while leveraging budget and personnel. Received two Emmy nominations as 'Visual Effects Producer'. Supervised post-production processes and personnel, including telecine, digital imagery, stock footage, online assemblies, and audio laybacks. Designated and deployed graphic applications and platforms to align with specific shots, sequences, and personnel. Received an honorary Emmy team award for Sound Composition. Promoted from, and simultaneously worked in, various positions throughout employment period.

Andrea L. Kavanaugh

Education

Bates College, B.A. English (French minor), 1973
University of Pennsylvania, M.A. Annenberg School for Communication, 1985
Virginia Tech (VT), Ph.D. Environmental Design and Planning, 1990

Appointments

Associate Director and Senior Research Scientist (2002-present), Center for Human-Computer Interaction (CHCI), Computer Science Department, VT
Director of Research, Blacksburg Electronic Village, Information Systems VT, 1993-01
Adjunct Professor, Communication Studies Department, VT 1997-98
Instructor, Center for Interdisciplinary Studies, VT 1992
Instructor, Communication Studies Program, Hollins University 1989-90

Awards and Fellowships

Fulbright Research Fellowship 1991-92 (North Africa)
Cunningham Dissertation Fellowship 1987-88 (Virginia Tech)
US Department of Education, Graduate Studies Fellowship, National Defense Foreign Language NDFL Title VI (full tuition and stipend), University of Pennsylvania, 1983-85) Language and Area Studies (Middle East Studies/Persian)

Recent professional service

Board of Directors, Treasurer, Secretary, Digital Government Society, 2008-2013
Board of Directors, International Telecommunications Society, 2002-2008
Chair, Committee for Web Communication, Digital Government Society, 2006-07
Conference Program Committee: Digital Government, HyperText, ASE/IEEE Social Informatics, ASE/IEEE Social Computing, Communities & Technologies, Online Deliberation, Information Systems for Crisis Response and Management (ISCRAM)
Professional Societies: ACM, Digital Government Society

Cumulative research productivity

Three books, 23 refereed journal articles, 11 refereed conference proceedings, 21 refereed or invited book chapters; 25 refereed presentations at professional meetings; over 60 refereed or invited professional presentations, demonstrations and talks.

Recent Related Publications

Kavanaugh, A., Sheetz, S., Hassan, R. Yang, S., Elmongui, H., Fox, E., Magdy, M. and Shoemaker, D. (Forthcoming). Between a Rock and a Cell Phone: Communication and Information Technology Use during the 2011 Egyptian Uprising. *International Journal of Information Systems for Crisis Response and Management,* Special Issue on Social Media Use in Crises.

Kavanaugh, A., Neidig, S., Ahuja, A., Gad, S., Perez-Quinonez, M., Ramakrishnan, N. and Tedesco, J. (Invited for special issue, 2013). (Hyper)Local News Aggregation: Designing for Social Affordances. *Government Information Quarterly.*

Kavanaugh, A. (Forthcoming) Physical versus Web Communities: The Arc of Social Computing. In Reda Alhajj and Jon Rokne (Eds.), *Encyclopedia of Social Network Analysis and Mining,* Surrey, UK: Springer.

Kavanaugh, A. (2012) The Blacksburg Electronic Village, pp. 593-601. In W. Bainbridge (Ed.) *Leadership in Science and Technology: A Reference Handbook,* Thousand Oaks, CA: Sage.

Kavanaugh, A., Fox, E., Sheetz, S., Yang, S., Li, L.T., Shoemaker, D., Natsev, A. and Xie, L. (2012). Social Media Use by Government: From the routine to the critical. *Government Information Quarterly, 29*(4): 480-491.

Other Relevant Publications

Kavanaugh, A., Pérez-Quiñones, M., Tedesco, J. and Sanders, W. (2010) Toward a Virtual Town Square in the Era of Web 2.0, pp. 279-294. In J. Hunsinger, L. Klastrup and M. Allen (Eds.) *Handbook of Internet Research*. Surrey, UK: Springer.

Kavanaugh, A., Kim, B.J., Schmitz, J. and Pérez-Quiñones, M. 2008. Net Gains in Political Participation: Secondary effects of the Internet on community. *Information, Communication and Society, 11* (7).

Kavanaugh, A., Zin, T.T., Rosson, M.B., Carroll, J.M. and Schmitz, J. 2006. Local groups online: Political learning and participation. *Computer Supported Cooperative Work, 16* (September): 375-395.

Kavanaugh, A., Carroll, J.M., Rosson, M.B., Reese, D.D. & Zin, T.T. 2005. Participating in Civil Society: The case of networked communities. *Interacting with Computers* 17, 9-33.

Kavanaugh, A. Reese, D.D., Carroll, J.M., & Rosson, M.B. 2005. Weak Ties in Networked Communities. *The Information Society 21* (2), 119-131.

Synergistic Activities

I have been leading sponsored research for three decades employing quantitative and qualitative research methods to evaluate the diffusion, adoption, use and impact of information and communication technology. I have been investigating most recently the use of information technology for citizen-to-citizen deliberation, and the development of tools to support social interaction among diverse users and groups. I have led funded collaborations with local government and voluntary associations, including those that serve the needs of socio-economically disadvantaged citizens in the Appalachian region.

Recent Doctoral and Masters students (chair or co-chair)

Szu-Chia Lu, Computer Science, Virginia Tech (VT), MS, 2010
Vineeta Chaube, Computer Science, VT, MS, 2010
B. Joon Kim, Public Administration and Policy, VT, PhD, 2009
Candida Tauro, Computer Science, VT, MS, 2008
Jaideep Godara, Industrial & Systems Engineering, VT, MS, 2006
Jason Snook, Computer Science, VT, PhD, 2005

Collaborators in past 48 months

E. Fox, F. Quek, S. Sheetz, M. Perez-Quinones, P. Isenhour, D. Tatar, A. Puckett, N. Ramakrishnan, D. Shoemaker, D. Kafura, D. Gracanin, D. Dunlap, W. Sanders, J. Tedesco, V. Chaube, S. Ahuja, B. Hanrahan, I. Bukovic, J. Godara, A. Fabian, H.N. Kim, B.J. Kim, J. Gabbard, S. McCrickard, M. Sampat, M.A. Evans, B. Jones, C. Evia (Virginia Tech); J. Carroll, M.B. Rosson (Penn State); F. Casalegno (MIT), K. Hampton (Rutgers), Y. Arens, E. Hovey (U of Southern California), J. Fountain (U Mass, Amherst), K. Hanna (Internet Archives). My PhD advisors: F. Ventre, C. Goodsell, C. Bostian, P. Knox, T. Luke (Virginia Tech), and B. Wellenius (The World Bank).

Steven D. Sheetz

Education

Ph.D. in Business Administration, Major in Information Systems, Minor in Linguistics, University of Colorado at Boulder, 1996.
Masters of Business Administration, Major in General Business, University of Northern Colorado, 1987.
Bachelors of Science, Major in Computer Science, Minor in Economics, Texas Tech University, 1984.

Academic and Professional Appointments

2005 – Present, Director, Center for Global e-Commerce, Pamplin College of Business, Virginia Polytechnic Institute and State University.
2002- Present, Associate Professor, Department of Accounting and Information Systems, Virginia Polytechnic Institute and State University.
1997 – Present, Information Systems Consultant, Blacksburg, Virginia.
1996 – 2002, Assistant Professor, Department of Accounting and Information Systems, Virginia Polytechnic Institute and State University.
1994 – 1995, Visiting Instructor of Information Systems, at the University of Colorado at Colorado Springs.
1990 – 1994, Graduate Research Assistant, University of Colorado at Boulder.
1992 – 1993, Part Time Instructor, at the University of Northern Colorado, Greeley, CO.
1991 – 1992, Graduate Part Time Instructor of Information Systems at the University of Colorado, Boulder, CO.
1986 – 1990, Systems Analyst, National Systems and Research Co., Loveland, CO.
1985 – 1986, Programmer/Analyst, National Systems and Research Co., Loveland, CO.

Five Publications Relevant to the Proposal (i)

S.D. Sheetz, A. Kavanaugh, F. Quek, B.J. Kim, and S.C. Lu, "Expectation of connectedness and cell phone use in crisis", Int. J. Emergency Management, Vol. 7, No. 2, 2010, Pages 124-136.

S.D. Sheetz, D. Henderson, and L. Wallace, "Understanding Developer and Manager Perceptions of Function Points and Source Lines of Code," Journal of Systems and Software, Vol. 82, 2009, Pages 1540-1549.

E. A. Fox, C. Andrews, W. Fan, J. Jiao, A. Kassahun, S. Lu, Y. Ma, C. North, N. Ramakrishnan, A. Scarpa, B. H. Friedman, S. D. Sheetz, D. Shoemaker, V. Srinivasan, S. Yang, and L. Boutwell. "A Digital Library for Recovery, Research, and Learning From April 16, 2007, at Virginia Tech." Traumatology, (2008), vol. 14: pp. 64 - 84.

D.P. Tegarden, L.F. Tegarden, S.D. Sheetz, "Cognitive Factions in a Top Management Team: Surfacing and Analyzing Cognitive Diversity using Causal Maps," DOI 10.1007/s10726-007-9099-1, November 2007, Group Decision and Negotiation.

D.P. Tegarden and S.D. Sheetz. “Group Cognitive Mapping: A Methodology and System for Capturing and Evaluating Managerial and Organizational Cognition. Omega, Vol. 31, 2003, Pages 113-125.

Five Other Significant Publications (ii)

D. Henderson, S.D. Sheetz, F. Belanger, “Explaining Developer Attitude Toward Using Formalized Commercial Methodologies: Decomposing Perceived Usefulness”, Information Resources Management Journal, Vol. 25, 2012, pages 1-20.

E. V. Wilson and S. D. Sheetz, “A demands-resources model of work pressure in IT student task groups,” Computers and Education: An International Journal, 55, 2010, Pages 415-426.

E.V. Wilson and S.D. Sheetz. “Context Counts: Effects of Work vs. Non-Work Context on Participants’ Perceptions of Fit in Email vs. Face-to-Face Communication.” Communications of the Association of Information Systems (CAIS), Volume 22, Article 17. (2008). Online at: http://aisel.aisnet.org/cais/vol22/iss1/17/

D. P. Tegarden, and S. D. Sheetz, “Cognitive Activities in OO Development," International Journal of Human-Computer Studies, Volume 54, Number 6, Spring 2001, pp. 779-798.

D. P. Tegarden, S. D. Sheetz, and D. E. Monarchi, “A Software Complexity Model of Object-Oriented Systems,” Decision Support Systems, The International Journal, Volume 13, Spring 1995, pp. 241-262.

Synergistic Activities

- Member of Information Systems for Crises, Response, and Management Association. 2010-Present.
- Reviewer for Journal of Management Information Systems, IEEE Transactions on Professional Communication, Omega, Decision Support Systems

Collaborators and Other Affiliations

(i) Collaborators

R. Barki (Virginia Tech), D. Tegarden (Virginia Tech), V. Wilson (Arizona State), G. Irwin (Colorado State University), L. Wallace (Virginia Tech), R. Beck (Villanova)

(ii) Graduate Advisor

Ph.D. Advisor: Kenneth A. Kozar, University of Colorado.

(iii) Graduate Students (Ph.D)

D. Henderson (Chair), Lemuria Carter, Youngwha Lee, John Briggs (Chair), Susan Kruck, Joeseph Ferki, Freda McBride

DONALD J. SHOEMAKER

A. Professional Preparation

University of Georgia: Athens, Georgia; M.A. Sociology, 1968; Ph.D. Sociology, 1970; NDEA Title IV Fellow, 1966-1969; Millsaps College: Jackson, Mississippi; B.A. Sociology, 1966; University of Mississippi: Oxford, Mississippi; Undergraduate Study, 1962-63

B. Appointments

Professor, Virginia Polytechnic Institute and State University, Department of Sociology, 1997
Director, Center for the Study of Violence in Society, 2003-2008
Visiting Exchange Professor, Department of Sociology, University of the Philippines, Diliman, Spring Semester, 1997; June-August, 1998; Spring Semester, 2001
Associate Professor, Virginia Polytechnic Institute and State University, Department of Sociology, 1977-1997. Assistant Professor, 1974-1976. One-third time with the Community Resource Development Office at Virginia Tech, Winter, 1983
Visiting Exchange Professor, Department of Sociology, Xavier University in Cagayan de Oro City, Philippines, January-March, 1987
Adjunct Professor, Roanoke College, Fall, 1986 (taught a course on juvenile delinquency)
Assistant Professor, University of Southern Mississippi, 1970-1974

C. Publications (Selected Related)

1. Kavanaugh, A.L., Fox, E.A., Sheetz, S.D.,Yang, S., Li, L.T., Whalen, T. Shoemaker, D.J., Natsev, P., Xie, L., Social Media Use by Government: From the Routine to the Critical. Government Information Quarterly (GIQ) 29(4): 480-491, 2012
2. Kavanaugh, A.L., Sheetz, S.D., Hassan, R., Yamg, S., Elmongui, H.G., Fox, E.A., Magdy, M., Shoemaker, D., Between a Rock and a Cell Phone: Communication and Information Use During the Egyptian Uprising. Proceedings of the 9th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2012), Vancouver, April 22-25, 2012
3. Seunwang, Yang, Andrea Kavanaugh, Nadia P. Kozievitch, Lin Tzy Li, Venkat Srinivasan, Steven D. Sheetz, Travis Whalen, Donald Shoemaker, Ricardo da A. Torres, and Edward A. Fox, CTRnet DL for Disaster Information Services. Proceedings of JCDL 2011, Ottawa, June 13-17, 2011, 437-438
4. Sheetz, Steve, Fox, Edward A., Fitzgerald, A., Palmer, S., Shoemaker, D., and Kavanaugh, Andrea. Why Students Use Social Networking Sites After Crisis Situations. Proceedings of the 8th International ISCRAM Conference, Lisbon, Portugal, May 8-11, 2011
5. Kavanaugh, A.L., A. Nastev, E. Fox, S. Sheetz, D. Shoemaker, L. Yie, S. Yang, V. Srinivasan, L.T. Li, and T. Whalen, Social Media for Cities, Counties, and Communities (CCSR Planning Grant, July 2010-December, 2010), Final Report

Publications (Selected Other)

1. Shoemaker, D.J., Theories of Delinquency: An Examination of Explanations of Delinquent Behavior. New York: Oxford University Press, 1984), pp. 281. Second edition, 1990, pp. 329. Third edition, 1996, pp. 284. Fourth edition, 2000, pp.294. Fifth edition, 2005. Sixth edition, 2010, pp. 398.
2. Shoemaker, D.J., Juvenile Delinquency, Lanham, MD: Rowman & Littlefield, 2009, pp. 447 (second edition in preparation).
3. Fox, E.A., C. Andrews, W. Fan, J. Jiao, A. Kassahun, S. Lu, Y. Ma, C. North, N. Ramakrishnan, A. Scarpa, B.H. Friedman, S.D. Sheetz, D. Shoemaker, V. Srinivasan, S. Yang, and L. Boutwell, "A Digital Library for Recovery, Research, and Learning from April 16, 2007 at Virginia Tech." Traumatology, Vol. 14 (1), 2008:64-84.
4. Gutierrez, F.C. and D.J. Shoemaker, "Self-Reported Delinquency of High School Students in Metro Manila: Gender and Social Class." Youth & Society, Vol. 40, 2008:55-85.
5. Shoemaker, D.J. and D. McDonald, "An Evaluation of the Drug Court of the Twenty-third Judicial Circuit Court of Virginia: A Response to the War on Drugs." Criminal Law Bulletin, Vol. 39, Number 5, 2003: 569-583.

D. Synergistic Activities

1. Proposal reviewer for NSF, March, 2007
2. Member, Editorial Board, Journal of Research in Crime and Delinquency, 2002-2011
3. Member, Editorial Board, Philippine Journal of Law and Justice, 2000-2005
4. Member, Editorial Advisory Board, Youth and Society, 1985-1995
5. Proposal reviewer for SEA Grant Program, 1990

E. Collaborators & Other Affiliations (recent and partially including those mentioned above)

W. Timothy Austin, Indiana University of Pennsylvania; Paul Friday, University of North Carolina-Charlotte; Filomin Gutierrez, University of the Philippines, Diliman; Danielle McDonald, Northern Kentucky University; Zin Ren, California, State University, Sacramento; Timothy W. Wolfe, Mount Saint Mary College

Advisor

Raymond Payne (deceased)

Thesis and Dissertation Advisees (recent)

Sinan Demarik, Danielle McDonald, John McMullen, Andrea Nash, Roderick Neal, Virginia Rothwell, Jamie Spradlin, Elizabeth Ward

Total number of graduate advisees, 21

Budget Justification Page

VT Budget Justification:

Faculty:
PI Fox will work 0.45 months each year, on average.
Co-PIs Kavanaugh, Sheetz, and Shoemaker will about 0.43 months each year.
Drs. Fox and Kavanaugh have CY appointments.
Drs. Sheetz and Shoemaker have AY appointments.
Dr. Fox will direct the project and supervise technical efforts, including digital libraries, information retrieval, machine learning, analysis services, information visualization, logging and log analysis, and focused crawling. He will attend conferences like JCDL, TPDL, and SIGIR.
Dr. Sheetz will focus on ontologies, databases, focus groups, and agile software development activities. Regarding collections, he will focus on
CTR (Crises, Tragedies, and community Recovery), and will continue to represent us at ISCRAM conferences. He will guide our validation in accordance with Diffusion of Innovation Theory.
Dr. Kavanaugh will focus on digital government, community support (building on her work with the Blacksburg Electronic Village), HCI, social sciences, surveys, questionnaires, etc. She will attend digital government meetings.
Dr. Shoemaker will focus on sociology, violence, demonstrations, communities, etc. He will liaise with the Center for Peace Studies and Violence Prevention, and with local advisory board members from sociology, psychology, and political science.

Graduate Research Assistants:
Two students will be funded for three years.
One student interested and suitable is working on a Ph.D. with support for tuition through the VT-MENA program, in collaboration with Egypt.

Fringe Rates are:
29.25% Regular Faculty
31.25% Special Research Faculty
17% Part Time Faculty
7.75% Summer Faculty / Wage Employee
8.5% GRA

Equipment:
Two computers will be purchased in year 1 for the GRAs.
No overhead is charged on this.

Travel:
Year 1 has $800 for domestic, $800 for foreign travel expenses.
Years 2 and 3 have escalations due to inflation, more activities.
Travel will include for working with Internet Archive, and attending conferences (papers, presentations, workshops, tutorials, etc.) such as about: digital libraries, CTR, digital government, sociology.
Examples include JCDL, TPDL, and ISCRAM.

Subcontract:
Please see the letter from the Internet Archive.
They will partner with us, and provide permanent archiving of content.
Their usual rates are much higher; we are getting a large discount.
Overhead is only charged on the first $25K (out of the 3*16K = $48K).

Budget Justification Page

Other Direct Costs - Other:
This covers tuition for one GRA, since the VT-MENA program covers the other.
No overhead is charged on this.

Indirect: 61% is charged on all but tuition and equipment.

- -

IA Budget Justification:

Please see the letter from the Internet Archive.
They will partner with us, and provide permanent archiving of content.
Their usual rates are much higher; we are getting a large discount.
Overhead is only charged on the first $25K (out of the 3*16K = $48K).

Budget Justification Page

Please see the letter from the Internet Archive. They will partner with us, and provide permanent archiving of content. Their usual rates are much higher; we are getting a large discount. Overhead is only charged on the first $25K (out of the 3*16K = $48K).

Current and Pending Support
See GPG Section II.D.8 for guidance on information to include on this form.

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

Investigator: Edward A. Fox	Other agencies (including NSF) to which this proposal has been/will be submitted:

Support: ☒ Current ☐ Pending ☐ Submission planned in near future ☐ Transfer of support
Project/Proposal Title: **Collaborative Project: Ensemble: Enriching Communities and Collections to Support Education in Computing**
Source of Support: **NSF**
Total Award Amount: **$509,897 (VT)** Total Award Period Covered: **09/15/08-8/31/13**
Location of Project: **Virginia Tech**
Person-months committed to project: Cal: **0.00** Acad: **0.00** Sumr: **0.00**

Support: ☒ Current ☐ Pending ☐ Submission planned in near future ☐ Transfer of support
Project/Proposal Title: **SDCI NMI New: From Desktops to Clouds - A Middleware for Next Generation Network Science**
Source of Support: **NSF**
Total Award Amount: **$1,350,000** Total Award Period Covered: **8/1/2010-7/31/13**
Location of Project: **Virginia Tech**
Person-months committed to project: Cal: **0.5** Acad: **0.00** Sumr: **0.00**

Support: ☒ Current ☐ Pending ☐ Submission planned in near future ☐ Transfer of support
Project/Proposal Title: **Computing in Context**
Source of Support: **Villanova University (pass through from NSF TUES)**
Total Award Amount: **$22,500** Total Award Period Covered: **8/15/2012 – 7/31/2014**
Location of Project: **Virginia Tech**
Person-months committed to project: Cal: **0.07** Acad: **0.00** Sumr: **0.00**

Support: ☒ Current ☐ Pending ☐ Submission planned in near future ☐ Transfer of support
Project/Proposal Title: **Establishing a Qatari Arabic-English Library Institute**
Source of Support: Qatar National Research Fund Project No. NPRP 4-029-1-007
Total Award Amount: **$1,010,485** Total Award Period Covered: **4/1/2012-3/31/2014**
Location of Project: **Virginia Tech**
Person-months committed to project: Cal: **0.80** Acad: **0.00** Sumr: **0.00**

Support: ☒ Current ☐ Pending ☐ Submission planned in near future ☐ Transfer of support
Project/Proposal Title: **III:Small:Integrated Digital Library Support for Crisis, Tragedy, and Recovery**
Source of Support: **NSF**
Total Award Amount: **$500,000** Total Award Period Covered: **8/1/2009-7/31/2013**
Location of Project: **Virginia Tech**
Person-months committed to project: Cal: **1.0** Acad: **0.00** Sumr: **0.00**

Support: ☐ Current ☒ Pending ☐ Submission planned in near future ☐ Transfer of support
Project/Proposal Title: **CSR:Large:Big Data for Predictive Disaster Management**
Source of Support: **NSF**
Total Award Amount: **$500,002** Total Award Period Covered: **8/1/2013 – 7/31/2018**
Location of Project: **Virginia Tech**
Person-months committed to project: Cal: **1.2** Acad: **0.00** Sumr: **0.00**

Support: ☐ Current ☒ Pending ☐ Submission planned in near future ☐ Transfer of support
Project/Proposal Title: **III:Small:Integrated Digital Event Archiving and Library (IDEAL)**
Source of Support: **NSF**
Total Award Amount: **$500,000** Total Award Period Covered: **8/10/2013-8/9/2016**
Location of Project: **Virginia Tech**
Person-months committed to project: Cal: **0.45** Acad: **0.00** Sumr: **0.00**

***If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.**

Current and Pending Support

(See GPG Section II.D.8 for guidance on information to include on this form.)

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

Investigator: Kristine Hanna

Other agencies (including NSF) to which this proposal has been/will be submitted.

Support: ☐ Current ☒ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title:
III:Small:Integrated Digital Event Archiving and Library (IDEAL)

Source of Support: NSF
Total Award Amount: $500,000 Total Award Period Covered: 8/10/2013 - 8/9/2016
Location of Project: Blacksburg, VA
Person-Months Per Year Committed to the Project. Cal: 0.0 Acad: 0.0 Sumr: 0.0

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title:

Source of Support:
Total Award Amount: $ Total Award Period Covered:
Location of Project:
Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title:

Source of Support:
Total Award Amount: $ Total Award Period Covered:
Location of Project:
Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title:

Source of Support:
Total Award Amount: $ Total Award Period Covered:
Location of Project:
Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title:

Source of Support:
Total Award Amount: $ Total Award Period Covered:
Location of Project:
Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

*If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

Investigator: **Andrea L. Kavanaugh**	Other agencies (including NSF) to which this proposal has been/will be submitted:

Support: ☒ Current ☐ Pending ☐ Submission planned in near future ☐ Transfer of support
Project/Proposal Title: **CTRNet: Integrated Digital Library Support for Crisis, Tragedy and Recovery**
Source of Support: **NSF IIS-III** (IIS-0916733)
Total Award Amount: **$499,999** Total Award Period Covered: **8/16/09-8/15/12**
Location of Project: **Blacksburg, Virginia**
Person-months committed to project: Cal: **1.8** Acad: **0.00** Sumr: **0.00**

Support: ☒Current ☐Pending ☐ Submission planned in near future ☐ Transfer of support
Project/Proposal Title: **Participation on the Town Square in the Era of Web 2.0**
Source of Support: **NSF Social-Computational Systems (SES-1111239)**
Total Award Amount: **749,999** **8/16/11-8/15/14**
Location of Project: **Blacksburg, Virginia**
Person-months committed to project: Cal: **1.8** Acad: **0.00** Sumr: **0.00**

Support: ☐ Current ☒Pending ☐ Submission planned in near future ☐ Transfer of support
Project/Proposal Title: **III (Small) IDEAL: Integrated Digital Event Archive and Library**
Source of Support: **NSF III: Small (this proposal)**
Total Award Amount: **$500,000** Total Award Period Covered: **8/16/13-8/15/16**
Location of Project: **Blacksburg, Virginia**
Person-months committed to project: Cal: **0.42** Acad: **0.00** Sumr: **0.00**

Support: ☐ Current ☒Pending ☐ Submission planned in near future ☐ Transfer of support
Project/Proposal Title: **Improved Confidentiality through Community-enabled Applications**
Source of Support: **NSF Secure and Trustworthy Cyberspace (SaTC): Small**
Total Award Amount: **$500,000** Total Award Period Covered: **8/1/13-7/31/16**
Location of Project: **Blacksburg, Virginia**
Person-months committed to project: Cal: **0.5** Acad: **0.00** Sumr: **0.00**

Support: ☐ Current ☐ Pending ☐ Submission planned in near future ☐ Transfer of support
Project/Proposal Title:
Source of Support:
Total Award Amount: Total Award Period Covered:
Location of Project:
Person-months committed to project: Cal: **0.00** Acad: **0.00** Sumr: **0.00**

Support: ☐ Current ☐Pending ☐ Submission planned in near future ☐ Transfer of support
Project/Proposal Title:
Source of Support:
Total Award Amount: Total Award Period Covered:
Location of Project:
Person-months committed to project: Cal: **0.0** Acad: **0.00** Sumr: **0.00**

***If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.**

Current and Pending Support
See GPG Section II.D.8 for guidance on information to include on this form.

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

Investigator: Steven D. Sheetz	Other agencies (including NSF) to which this proposal has been/will be submitted:

Support: ☐ Current ☒ Pending ☐ Submission planned in near future ☐ Transfer of support
Project/Proposal Title: **III:Small:Integrated Digital Event Archiving and Library (IDEAL)**
Source of Support: **NSF (this proposal)**
Total Award Amount: **$500,000** Total Award Period Covered: **08/10/2013-08/09/2016**
Location of Project: **Virginia Tech**
Person-months committed to project: Cal: **0.00** Acad: **0.00** Sumr: **0.43**

Support: ☒ Current ☐ Pending ☐ Submission planned in near future ☐ Transfer of support
Project/Proposal Title: **III:Small: CTRnet: Integrated Digital Library Support for Crisis, Tragedy, and Recovery**
Source of Support: **NSF**
Total Award Amount: **$500,000** Total Award Period Covered: **8/16/2009 – 8/15/2013**
Location of Project: **Virginia Tech**
Person-months committed to project: Cal: **0.00** Acad: **0.00** Sumr: **0.40**

Support: ☐ Current ☐ Pending ☐ Submission planned in near future ☐ Transfer of support
Project/Proposal Title:
Source of Support:
Total Award Amount: Total Award Period Covered:
Location of Project:
Person-months committed to project: Cal: Acad: Sumr:

***If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.**

Current and Pending Support

(See GPG Section II.C.2.h for guidance on information to include on this form.)

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

Investigator: Donald Shoemaker	Other agencies (including NSF) to which this proposal has been/will be submitted.

Support: ☒ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title: CTRNet: Integrated Digital Library Support for Crisis, Tragedy, and Recovery

Source of Support: NSF

Total Award Amount: $ 500,000 Total Award Period Covered: 08/01/09 - 07/31/13

Location of Project: Blacksburg, virginia

Person-Months Per Year Committed to the Project. Cal: 0.00 Acad: 0.00 Sumr: 0.50

Support: ☐ Current ☒ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title: III: Small-Integrated Digital Event Archiving Library (IDEAL)

Source of Support: NSF

Total Award Amount: $ 500,000 Total Award Period Covered: 08/10/13 - 08/29/16

Location of Project: Virginia Tech

Person-Months Per Year Committed to the Project. Cal: 0.00 Acad: 0.00 Sumr: 0.43

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Summ:

*If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.

Facilities, Equipment and Other Resources (Virginia Tech)

The work for this project will be carried out using the facilities available to the Digital Library Research Laboratory (DLRL), Department of Computer Science, Center for Human-Computer Interaction, and Virginia Tech's computing infrastructure.

In May 2006, the Department of Computer Science gained almost 40,000 square foot of new space with a new building (Knowledge Works II) housed in the Virginia Tech Corporate Research Center (CRC). The department's undergraduate program is housed in McBryde Hall, together with a newly renovated Undergraduate Learning Center. In addition, the department also maintains numerous research labs in Torgersen Hall, like DLRL, directed by PI Fox.

Digital Library Research Laboratory (DLRL)

The Virginia Tech Digital Library Research Laboratory (DLRL) is intended to support digital library research activities on campus. It has a server room, a conference room, two offices, an area for printers and bookshelves, and about a dozen carrels. There also is a DLRL area down the hall, close to the office of Dr. Fox, with 4 more carrels. Its neighbors in the same building, dedicated late in 2000 as Torgersen Hall after the most recent prior university president, include roughly 10 other research groups, providing a fertile ground for collaboration. For example, Virginia Tech Digital Library and Archives, part of VT University Libraries, is also in the $30M building, as is the Discovery Analytics Center. As VT's Advanced Communications and Information Technology Center, the building hosts high speed wireless and Ethernet connections, as well as very fast connection to the campus backbone, and from there to the fastest national networks. Other building labs include those for software engineering, visualization and virtual environments (including a CAVE), digital discourse, etc. The neighboring Laboratory for Advanced Scientific Computing and Applications houses a Sun Fire X4600 system with 8 AMD quad core 8356 2.3 GHz 64-bit CPUs, for a total of 32 cores. The Sun Fire system has 64 GB of shared memory and 500 GB storage.

The DLRL is occupied mainly by graduate students engaged in digital library research, though there frequently have been visitors for long periods, e.g., from South Korea, India, Japan, and Brazil. It houses primarily a variety of PCs, some running Linux and others running Windows, plus a number of MacBook Pro and other Macintosh systems. Six new systems were purchased in 2012. There are two high-powered Alien systems for graphics related work. In addition, the DLRL has systems in a Computing Center machine room. One, for research and public access to information on crisis, tragedy, and recovery, is a load-balanced server with two dual-core processors for the front-end, and two quad-core processors for the back-end. Storage includes 16G RAM and 8 terabytes of disk. A newer server has 48 cores and 256G of RAM.

Department of Computer Science

Researchers in the Department of Computer Science at Virginia Tech have access to several high-end computing platforms. Various facilities across several laboratories are available for use

by students in this project. The following systems can be leveraged for use on the project as needed:

- A hydra 9-node Intel Dual-Core 2-processor XEON 3 GHz, 8 GB main memory, 1 Gbps interconnect, Linux cluster
- Two high-performance server-class heterogeneous machines, each with 2x Intel Xeon E5620 2.40 GHz quad-cores, 48~GB of RAM, and nVidia Fermi C2070 (6~GB GDDR5 Memory, 448 cuda cores)
- Four Dell Precision 750, dual 2.5 GHz Intel Core 2 servers with 4 GB main memory, NVIDIA GPU cards
- Two Dell Optiplex 620, 3.4 GHz servers with 3 GB main memory
- 4-node/64-core cluster, with quad-quad-core processor node (AMD Barcelona 2.0 GHz)
- 25-node PlayStation3 Cluster (Cell/BE processor at 3.2 GHz), with quad-core and oct-core Intel head nodes
- Multiple Dell oct-core, quad-core and dual-core systems (Precision, Optiplex) used for code development, documentation, data archiving and maintenance.

Center for Human-Computer Interaction

HCI Labs: An NSF Research Infrastructure Grant (CDA9303152) plus local support, totaling $2M, led to the development of laboratory facilities for conducting networked multi-user interactive experiments. The laboratory includes an electronic conference room with networked computers and telecommunication capabilities linked with high-speed connections to several individual experimental rooms. All of the rooms are situated around an instrumented control room containing computing and video control and capture capabilities. The facility offers unique multiuser evaluation with single-point two-way glass observation. A usability methods research control center allows real-time capture and integration of behavioral data from computer interaction, video observation, and experimenter comments. Other facilities include equipment for analyzing video data (including process control tools for controlling and synchronizing video recording and observer-initiated critical incident reports, and digital editing facilities for analyzing and editing video and audio records) and for developing multimedia information content. A new addition to the facilities is a four-port phone-server for VoiceXML client access to computer information.

Computer and networking support: Project staff will have access to networked workstations and servers running Windows, Mac OSX, Linux, and Solaris. Both wired and wireless connectivity is available in the Center's lab facilities. Infrastructure software includes database, file, proxy, and web server systems, as well as server software for custom collaboration tools developed by the Center. Custom and off-the-shelf tools for audio and video capture, processing and transcription support data collection activities. Analysis tools include custom session log processing software, as well as qualitative and quantitative data analysis packages. Development tools are also available for a variety of platforms and languages.

Office space: The Center for Human-Computer Interaction manages 2,395 square feet of well equipped laboratory (Aware lab, Gigapixel lab, 3DInteraction/CAVE, HCI lab).

College and University Facilities

The Virginia Tech College of Engineering (VT COE) provides some of the most advanced technology available, including wireless internet and hundreds of windows-based PCs in various laboratories, several other laboratories incorporating equipment from Sun Microsystems, Apple, SGI and other Unix-based machines. These resources and the resources of the Center for High-End Computing (CHECS) will be available for use by students participating in research and related classes. These machines include:

System G is a state-of-the-art energy-aware supercomputer with 325 nodes (2600 cores) currently being deployed by the Department of Computer Science. Each node is a Mac Pro computer, with two 4-core 2.8 GHz Intel Xeon processors and 8 GB of memory. System G employs QDR Infiniband interconnect that achieves 40 Gbps data transfer speeds. System G currently has a peak performance of 22.8 TeraFlops. The specialty of this system is that it employs over 11,000 power and thermal sensors, which makes it an ideal test bed for techniques that enable emerging green-computing. Thirty nodes of this system are being used in DLRL's CTRnet project, running Big Data Software made available by LucidWorks.

HokieSpeed is a new heterogeneous supercomputing instrument based on a combination of central processing units (CPUs) and graphical processing units (GPUs). In terms of raw performance, HokieSpeed is expected to deliver 35 times better peak performance, 70 times better peak power efficiency, and 14 times better peak space efficiency than our recently decommissioned supercomputer, System X.

Rlogin is a 20 node compute cluster where each node has two Xeon X5647 processors, 2.93GHz, with 4 cores per CPU (i.e., 8 cores per node) and 12GB of RAM, along with a 40 GB hard drive. Rlogin is mainly used for student computing and experimentation.

Data Management Plan

Data management and sharing for research results created by this project will conform to NSF policy on the dissemination and sharing of research results as defined in NSF's Award and Administration Guide Chapter VI (D4).

1. Types of Data
We will have terabytes of data from collections of web content on events (e.g., crises, government processes, and community activities). These data include text, images, and video derived from formal online media (popular news websites like CNN, Google news), and informal media (Twitter, Facebook, RSS, blogs, forums, social networks, YouTube).

Other data are related to analyses, reports, and visualizations produced by the interactive tools and services we provide.

2. Data and Metadata Standards
For our collections of data and metadata (e.g., multi-document summaries, aggregated information extracted from documents obtained from multiple sources), we will use open source, standards-based data sharing systems and applications.

3. Policies for Access and Sharing and Provisions for Protection of Privacy/Security
Virginia Tech operates its networks and servers under secure measures and ensures privacy of users through standard password procedures and updates.

Any human subjects data we collect related to usability evaluation will be reviewed and approved by the Institutional Review Board (IRB) for research on human subjects at Virginia Tech. All human subjects data from any evaluations will be stored in the locked office of the PI or Co-PIs on the project. The data of participating respondents will be anonymized and remain confidential.

The data collections on events and related tools and methods will be shared via our website (http://www.ctrnet.net) and our partner, the Internet Archive. Our website will leverage LucidWorks Big Data Software in conjunction with our software to support access and analysis.

4. Policies for Re-use, Redistribution
The techniques and methods we develop, findings from our evaluations, software we build, and services we provide, will all be open and shared.

5. Plans for Archiving and Preservation of Access
We will archive and preserve access to our data through our website (http://www.ctrnet.net) and our partner, the Internet Archive (IA). IA will make data from us available permanently. IA is a part of the International Internet Preservation Consortium (IIPC) that has members from libraries that preserve documents of national interest, or based on current events or web sites related to their countries. We will give to IA and IIPC new tools to collect and preserve event related digital content. Any data that pertains to human subjects (e.g., raw data from results of usability evaluation) will be archived with ICPSR at the University of Michigan (http://www.icpsr.umich.edu/icpsrweb.ICPSR/index.jsp) by the end of the project, or when all related publications have been finalized.

List of Project Personnel and Partner Institutions

1. Edward Fox; Virginia Tech (Computer Science, Information Technology, Center for Human Computer Interaction, Discovery Analytics Center); PI
2. Andrea Kavanaugh; Virginia Tech (Computer Science, Center for Human Computer Interaction); Co-PI
3. Steven Sheetz; Virginia Tech (Accounting and Information Systems, Center for Human Computer Interaction); Co-PI
4. Donald Shoemaker; Virginia Tech (Sociology); Co-PI
5. Kristine Hanna; Internet Archive (Director, Archiving Services); Co-PI, Subawardee, External Advisory Committee Member
6. Paul Doscher; LucidWorks (CEO); Unpaid Collaborator (providing software and support – see letter), External Advisory Committee Member
7. Carlos Alberto Alejandro Castillo Ocaranza; Qatar Computing Research Institute (Senior Scientist - Social Computing); Unpaid Collaborator (offering to host an intern – see letter)
8. Patrick Meier; Qatar Computing Research Institute (Director, Social Innovation) and iRevolution author; External Advisory Committee Member
9. Geoff Harder; University of Alberta (Digital Initiatives Coordinator); External Advisory Committee Member
10. Susan Metros; University of Southern California (Associate Vice Provost, Deputy CIO); External Advisory Committee Member
11. Eric Van de Velde; EVdV Consulting; External Advisory Committee Member
12. Chris Barrett; Virginia Tech (Director, Network Dynamics and Simulation Science Laboratory – see letter); Internal Advisory Committee Member
13. Tyler Walters; Virginia Tech (Dean, University Libraries – see letter); Internal Advisory Committee Member
14. Gail McMillan; Virginia Tech (Director of Digital Library and Archives); Internal Advisory Committee Member
15. Purdom Lindblad; Virginia Tech (Digital Humanities Librarian); Internal Advisory Committee Member
16. James Hawdon; Virginia Tech (Sociology and Director, Center for Peace Studies and Violence Prevention – see letter); Internal Advisory Committee Member
17. John Ryan; Virginia Tech (Dept. Head, Sociology); Internal Advisory Committee Member
18. Timothy Luke; Virginia Tech (Political Science, University Distinguished Professor); Internal Advisory Committee Member
19. Sanmay Das; Virginia Tech (Computer Science, Discovery Analytics Center); Internal Advisory Committee Member
20. Russell Jones; Virginia Tech (Psychology); Internal Advisory Committee Member
21. Chris North; Virginia Tech (Computer Science, Center for Human Computer Interaction); Internal Advisory Committee Member
22. Scott Midkiff; Virginia Tech (Information Technology, CIO); Internal Advisory Committee Member

Other Supplementary Documents

1. Letters from External Partners and Supporters:

Internet Archive – Kristine Hanna, Director, Archiving Services
LucidWorks – Paul Doscher, CEO
Qatar Computing Research Institute – Carlos Castillo, Senior Scientist

2. Letters from Virginia Tech Supporting Parties:

Tyler Walters, Dean, VT University Libraries
Chris Barrett, Director, Network Dynamics & Simulation Science Lab (NDSSL)
James Hawdon, Director, Center for Peace Studies and Violence Prevention

December 11 2012

Professor Ed Fox
Department of Computer Science, Virginia Tech
114 McBryde Hall, M/C 0106
Blacksburg VA 24061

Dear Dr. Fox:

The Internet Archive is pleased to support and participate in the Integrated Digital Event Archive and Library (IDEAL). If NSF funding is provided we would be happy to have Virginia Tech as an Archive-It partner, and to work together to develop IDEAL. Also, I would be delighted to serve on the Advisory Committee and act as the Subaward/ Subrecipient Principal Investigator on the project. We look forward to collaboration on technology transfer so that the digital library research and development at Virginia Tech can be integrated into improved and extended archiving practices.

Archive-It is a web archiving service first deployed at the Internet Archive in early 2006. Curation of partner collections is managed directly by each partner. Archive-It is used by over 230 organizations in 46 U.S. states including university libraries, state archives, national organizations and local city governments.

The Archive-It subscription period of performance for this account is three years, starting when your project begins. And crawling is available throughout this time frame using 10 available capture frequencies, including a "manual on demand" option; as well as the features and functionality inside the web application.

The details of this $48K quote, beyond our research collaboration, are as follows:

- The account level is quoted at $16,000 per year (a discounted rate)
- This account level can archive up to 20 million documents and 2 terabytes of data per year
- The data is indexed for full text search and is browse-able by URL
- The data is hosted and stored at the Internet Archive data centers in perpetuity.
- The data is accessible and online 24/7
- The account comes with basic and advanced training, help documentation, an online user manual as well as Partner Specialist support via email
- Virginia Tech can download the archived data through our online access page.

Sincerely,

Kristine Hanna
Director, Archiving Services
Internet Archive
415 561 6799 x 5
Kristine@archive.org (Project PI and Administrative Contact)

LucidWorks™

11 December 2012

Professor Edward A. Fox
Department of Computer Science
114 McBryde Hall, M/C 0106
Virginia Tech
Blacksburg, VA 24061

Dear Professor Fox:

I am writing in support of your proposal to the National Science Foundation entitled "Integrated Digital Event Archive and Library (IDEAL)". This will continue and extend the support we have provided since August for your CTRnet project.

First, I agree to serve on the IDEAL Advisory Board, and thus to help guide project research and outreach. Second, we agree to provide, free of charge, software and technical support for LucidWorks software related to "big data". Third, we will work with your project team to disseminate your research software to our staff and others using our software, to ensure broad technology transfer, thus leading to an enhanced software suite for working with big data (including collecting, crawling, archiving, analysis, and visualization). Fourth, to deepen understanding and promote the wider use of tools for working with big data, we will make available the related educational materials you develop, both to our staff and to our clients / customers, and provide feedback to help enhance the coverage and utility of those materials.

I thank you for participating in our November Webinar on Computing for Disasters, and look forward to your continuing contribution to applying advanced information systems research to help those interested in crises, tragedies, and community recovery. We hope that your broadening of that work to deal with the integration of digital libraries and archives, and to handling other types of events, will lead to improved technology and services for those developing collections focused on events, in addition to those building domain or topic-oriented archives.

As you know, LucidWorks' mission is to transform the way people access data to enable information-driven decisions. The original goal with our LucidWorks Big Data product was to work with a small number of noteworthy organizations to solve their complex application challenges. Specifically, we were looking for applications that, before LucidWorks Big Data, were too difficult to attempt. Computing for Disasters fit perfectly. Topping this, the project appealed to us because of its humanitarian focus. We were pleased to partner with your organization on this project. Your IDEAL project is an extension of the work we are already doing with you, making it a natural fit for moving our relationship further to benefit society as a whole.

Sincerely,

Paul Doscher
Chief Executive Officer

3800 Bridge Parkway, Suite 101, Redwood City, CA 94065 – 650.353.4057

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation
P.O Box 5825
Tornado Tower Floor 10
West Bay, Doha-Qatar
Tel +974 4454 0629
Fax +974 4454 0630
www.qcri.qa

Prof. Ed Fox
Department of Computer Science
Virginia Tech
Blacksburg, VA 24061

December 12, 2012

Dear Ed,

I am writing to express my enthusiastic support for the NSF proposal entitled "Integrated Digital Event Archive and Library (IDEAL)" as it corresponds in important ways with our work in crisis informatics and humanitarian relief.

As you know, at Qatar Computing Research Institute in Doha, our work focuses on improving coordination on the type of responses that humanitarian organizations like UN OCHA engage in. This includes making sense of big data in crisis situations to provide additional situational awareness. A collaboration with your project would be especially helpful in the areas of ontologies, comparing and making sense of large social media data collections, developing methods and tools for analyses of big data, and working with data in multiple formats and languages (e.g., Arabic, English, French, etc.).

As you know, we host student interns at the Institute and would be willing to have a student from the IDEAL project spend a semester or summer working with us to advance our collaboration. Additionally, my colleague Patrick Meier has agreed to serve on the advisory board of the project.

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation

P.O Box 5825
Tornado Tower Floor 10
West Bay, Doha-Qatar
Tel +974 4454 0629
Fax +974 4454 0630
www.qcri.qa

These are very interesting research problems and important humanitarian issues. We look forward to working with you.

Sincerely,

Carlos Alberto Alejandro Castillo Ocaranza
Senior Scientist – Social Computing
Qatar Computing Research Institute

Tyler Walters, Dean
University Libraries
Virginia Tech
P.O. Box 90001
Blacksburg, Virginia 24062-9001
540.231.5595

December 14, 2012

To Whom It May Concern,

I am writing to voice my enthusiastic support for the grant proposal entitled "**IDEAL: Integrated Digital Event Archiving and Library**." This is an important project that moves Web archiving technologies and paradigms forward specifically and advances the field's general practices more broadly. Several characteristics make the IDEAL project significant in its intellectual merit and potential to make a broad impact. The 5S approach to digital library design is applied in the project, building upon earlier leading edge work by Prof. Fox and his research team. It also builds upon the earlier NSF-funded CTR project, which advanced digital library work related to crisis/tragedy/recovery. Building on these precursors positions the IDEAL project well to succeed. As Dean of the University Libraries of Virginia Tech, I will be serving on the IDEAL Local Advisory Board and am looking forward to guiding this initiative. Other important project components are addressed below.

The automated detection of events, in particular, and the subsequent crawling of web sites and archiving of web-based content is a vital technological approach to web archiving. With large-scale and targeted (i.e. event-related) web archiving on the rise, improved automated methods are sorely needed and we look forward to the project realizing its deliverables. The project goals not only include developing these automatic detection methods, but also provide a service that supports third- party requests to crawl and archive event-related Web materials. Moreover, the project also will collect, filter, catalog, preserve, and provide access to information found in the open Web and via Twitter. Together with the additional services of browsing, searching, recommending, classifying, clustering, linking/associating, analyzing, and visualizing make the IDEAL project truly a unique resource. In addition, the library professionals involved in the project will be suggesting events that we might archive, whenever they know of particular events that will directly support research groups at Virginia Tech.

Virginia Tech has deep experience with living through and recovering from tragic events. The 4/16/07 event in which 33 people met with a fatal end on the VT campus has made an indelible mark in the community's social fabric. Recovering from tragic events such as this has been a very important process for the advancement of the VT community. Both Prof. Fox's Digital Library Research Laboratory and the University Libraries have invested in collecting and managing digital documentation related to this and other tragedies. Data produced during the project will be managed and made available according to the data management plan jointly developed by the Laboratory and the Libraries. For two years the Libraries have provided

Virginia Tech with data management planning services and has been collaborating with Dr. Fox and his project personnel to devise the plan.

If I may amplify any of these comments, please contact me. We look forward to the IDEAL project's initiation.

Sincerely,

Tyler Walters
Dean, University Libraries
Virginia Tech

VirginiaTech
Virginia Bioinformatics Institute

Virginia Bioinformatics Institute
Network Dynamics & Simulation Science Laboratory
1880 Pratt Drive, RB XV (0477)
Blacksburg, Virginia 24061
540/231-8252 Fax: 540/231-2891
ndssl.vbi.vt.edu

December 13, 2012

Professor Edward A. Fox

Department of Computer Science
Virginia Tech
114 McBryde Hall, M/C 0106
Blacksburg, VA 24061

Dear Dr. Fox,

The Network Dynamics and Simulation Science Laboratory (NDSSL) is pleased to support the proposed Integrated Digital Event Archive and Library (IDEAL). If NSF funding is provided, I will be pleased to serve on your Advisory Committee, and to work to connect NDSSL activities with IDEAL.

The NDSSL is pursuing an advanced research and development program for interaction-based modeling, simulation, and associated analysis, experimental design, and decision support tools for understanding large biological, information, social, and technological systems.

As you explain in the proposal section on Dissemination and Validation, data collected and analyzed by IDEAL could feed into models and simulations relatedto the spread of crime, demonstrations, and revolutions. Thus, we could help validate your work, and you could help provide information to feed into simulations we would run of various social processes. That would complement our other collaborations with you, such as in developing an NSF funded cyberinfrastructure project (http://ndssl.vbi.vt.edu/cinet/).

Sincerely,

Chris Barrett
Director, Network Dynamics and Simulation Science Laboratory
Professor, Virginia Bioinformatics Institute
Professor, Department of Computer Science

VirginiaTech
College of Liberal Arts
and Human Sciences

James Hawdon,
Professor, Department of Sociology
Director Center for Peace Studies and Violence Prevention

205a Norris Hall
Blacksburg, Virginia 24061
540/231-7476 hawdonj@vt.edu

December 7, 2012
National Science Foundation
Review Panel.

I am writing in support of the proposal, "Integrated Digital Event Archive and Library (IDEAL)." The project will use new web-crawling techniques and develop a web-based program to automatically detect interesting crisis, government, and community events from the internet and then collect, catalog, and preserve the digital records of the event. This project will significantly enhance our ability to study critical incidents by providing novel, automated methods for collecting data from the internet. While the internet is an excellent source of event-related information, the sheer volume of information available on the web requires a systematic and automated means of collecting it. In addition, the program will provide a means of preserving the information by archiving all forms of content and media related to the event. Finally, the research team will provide a user-friendly means of accessing and studying the event-related digital objects collected on an event.

This research and the program resulting from it will prove extremely valuable for researchers. The amount, quality, and accessibility of the data that can be gathered by the proposed program are impressive. These data will be valuable to researchers who are studying the causes of crises, the management of crises, and the response to crises. I am confident that the techniques developed in this project will transform the nature of web-based event research. It is highly likely this project will benefit researchers studying mass tragedies, critical incidents, disasters, and other stakeholders conducting event analyses.

I am excited that the IDEAL project will be able to assist the efforts of the Center for Peace Studies and Violence Prevention at Virginia Tech. The IDEAL project will undoubtedly be useful for research projects on which I am working. In addition, I will be able to use information gathered through the IDEAL program in my course on Peace and Violence as Critical Incidents. Given my enthusiastic support for this research, I would be happy to serve on the project's Advisory Committee.

Sincerely,

James Hawdon
Professor
Director Center for Peace Studies and Violence Prevention