Information Services & Use 35 (2015) 31–50
DOI 10.3233/ISU-150767
IOS Press

Mining libraries: Lessons learned from 20 years of massive computing on the world's information

Kalev Hannes Leetaru
Senior Fellow, Center for Cyber and Homeland Security, George Washington University, 2000 Pennsylvania Ave, NW, Washington, DC 20052, USA
Tel.: +1 202 994 2437; E-mail: kalev.leetaru5@gmail.com; URL: http://kalevleetaru.com/

Abstract. Over the course of the last two decades I have explored the informational undercurrents of the world's information and the potential of mass data mining of libraries through a myriad of lenses, both technical and methodological. From founding my first Internet startup twenty years ago as an eighth grade middle school student, to running one of the world's largest global monitoring platforms today, my work has debuted a myriad of new datasets, methodologies, and scales to the study of how we understand our global world. A central theme of that work has been around how creative "reimagining" of information through the emerging world of massive computing power can offer powerful and unexpected new lenses onto the world around us, and the incredible future that awaits as libraries transition from being museums of artifacts to becoming conveners of information and innovation that empower a new era of access and understanding of our world. This paper offers a unique detailed view of what it looks like deep in the trenches of the data revolution, surveying a selection of my projects over the last two decades and the lessons they offer libraries and publishers moving forward into our data-driven future.

Keywords: Data mining, libraries, information access

1. Introduction

Imagine a world in which scientists, technologists, scholars, policymakers, journalists and even ordinary citizens work together to leverage massive computing power, digitization, and "big data" to reimagine libraries as centers of information and innovation that help us make sense of the oceans of data confronting society today. In the coming years libraries will be forced to reinvent themselves through a return to their roots, not as museums of physical artifacts for rental, but as conveners of information and invaluable resources who understand and translate that information to the needs of an innovative world. This was the vision I outlined in my opening keynote address at the 2012 General Assembly of the International Internet Preservation Consortium (IIPC)[1] and expanded upon in a guest post for the Knight Foundation last year titled "Reimagining libraries as conveners of information and innovation".[2] Yet, what might this world actually look like? What becomes possible as libraries open their collections to computation, allowing researchers to apply data mining algorithms to catalog, mine, visualize and create new ways of interacting with these vast archives?

[1] http://blogs.loc.gov/digitalpreservation/2012/05/a-vision-of-the-role-and-future-of-web-archives-the-web-archive-in-todays-world/.
[2] http://www.knightfoundation.org/blogs/knightblog/2014/9/30/reimagining-libraries-conveners-information-and-innovation/.

2. Core themes

For nearly two decades my career has closely followed this vision, working with libraries, publishers and vast data archives, coupled with some of the world's most powerful computing platforms, to explore what becomes possible when we render the world's information "computable". The lessons learned from these explorations have a tremendous amount to teach us about the future role of libraries, how we think about information, and how libraries and technologists will interact in the coming years to computationally explore our society's collective knowledge.

To summarize some of the core themes emerging from my personal experiences over the last twenty years:

Collections. Libraries must understand not only their own collection behavior, but also that of the sources they collect in order to advise on the data mining of those materials. *The New York Times* has reduced its total output by nearly half over the last sixty years, meaning that simply plotting the raw number of articles per year mentioning a given keyword will yield an incorrect view of the *Times*' focus – a detailed understanding of the *Times* is required to properly normalize analysis of it.

Computing. Projects of the future will involve massive computing and storage requirements, in many cases spanning multiple platforms and datacenters and drawing together huge technical teams to support them. Many projects will need to blend academic supercomputing facilities with the external commercial cloud, requiring that careful thought be given to data protection and stewardship as material leaves institutional premises. Licensing agreements and data sizes will often necessitate that the elements of a project be distributed across multiple datacenters, requiring careful thought as to how they are recombined for the final analysis.

Entities. In the share-everything social media era, patrons increasingly look upon information as collections of entities and elements rather than as immutable document containers. Users today are rarely looking for a physical book. They are looking for specific pieces of information contained within that book or experiences to be gained from reading it. This has ramifications for how libraries think about, organize and provide access to their collections as entities and experiences rather than as objects.

Funding. Large-scale research programs will increasingly be prototyped through unfunded pilot projects that may require substantial library personnel and material resources without the cost-recovery model of traditional faculty research. Unfunded student projects may rival funded faculty research in terms of scale, while even faculty projects may have such fast turnarounds that they cannot wait for the lengthy gestation period of traditional granting programs. Even non-academic libraries will face increasing demands for data mining or large-scale use of their collections by unfunded "social good" projects that do not have the financial resources to reimburse libraries for their assistance, all the while occurring in a climate of generally decreasing funding for libraries.

Geography. Patrons will increasingly demand the ability to place information spatially, from geographic search to mapping.

Interface. The most valuable information is rendered moot if patrons cannot easily and intuitively access and interact with it. Moving forward libraries must reimagine how they provide access to their collections, especially with respect to modalities like geographic search and advanced visualization, and they should consider partnerships with organizations that specialize in these areas.

Mass archiving. Libraries and institutional repositories will increasingly have to cope with projects that generate outputs of tens- or hundreds of millions of objects totaling many terabytes or even petabytes of data. At the same time, granting agencies and publication venues are increasingly requiring long-term

archiving of data and tools. Numerous questions remain regarding the funding and technological models to support such large long-term archiving.

Mutability. In the printed era, a book on a shelf did not spontaneously change contents – a library had only to acquire an object once and store it for perpetuity. In the web era, material changes constantly, even official U.S. Government publications, meaning libraries must constantly monitor every item for changes on a regular and ongoing basis.

Narrative. As patrons become increasingly accustomed to algorithms filtering and sorting the information they see based on emotion, readability, and other narrative dimensions, they will increasingly demand that libraries provide similar capabilities, but the complex technical landscapes of such analysis require considerable further discussion for libraries.

Networks. The era of social networks has infused the concept of relationships and connectivity among entities into popular culture. Users increasingly wish to understand the network of relationships and connections among information, rather than simply being handed a sorted list of documents.

Open data. Freely available "open data" is the lifeblood of significant innovation, but the vast and chaotic landscape of open data repositories can make it difficult for patrons to locate relevant data. Libraries can help patrons navigate this landscape and provide guidance for unexpected and creative repurposing of such data.

Scale. Libraries have historically provided access to their collections at the level of the individual object, while in the future they will increasingly be called upon to make entire archives available for data mining, and in some cases the entire data holdings of their institution. This has tremendous ramifications for licensing agreements, collection management and technical infrastructure.

Search. It is not enough today for libraries to simply toss material into digital libraries that only support basic search and browse interfaces. As patrons become accustomed to an online world where even images and video are searchable, and where search interfaces support emotional, geographic and even network search, libraries must evolve the interfaces they provide to their patrons. An archive of millions of digitized images is of little use if it must be browsed manually a page at a time.

Speed. As computer and information scientists increasingly provide the driving force for these large-scale projects, the traditional three to five year horizon of past large digital social sciences and humanities projects will give way to projects with timelines measured in months, placing extreme pressures on libraries and reducing their ability to act as collaborative partners for some projects.

Students. In the past large data-driven projects were exclusively the realm of large funded faculty projects with well-established protocols and communication channels. But moving forward students will play an increasing role in the data mining of library collections. In an era when a single undergraduate student can digitize tens of thousands of pages, take a quarter-million photographs, write three hundred histories, and blend all of this into an interactive website in the scope of a senior thesis or build the largest political event dataset in the world as a graduate dissertation, libraries must reexamine their outreach and interaction with students to better facilitate such research. In particular, academic libraries have sometimes had an adversarial relationship with student data miners in the past, refusing to support projects that were not part of established funded faculty research.

Technologists. Libraries will increasingly find themselves working closely with computer and information scientists and other technologists, requiring them to become familiar with the terminology, methodology, workflow and literatures of those fields.

Translation. As the digital era brings the world ever more closely together, the academic community is increasingly realizing that it must move beyond purely English and Western information sources to understand the rest of the world. At the same time, public libraries are increasingly finding a greater

diversity in the cultural backgrounds and languages spoken by their patrons. These trends place renewed emphasis on the collection practices and interfaces of libraries. Technologies such as machine translation offer powerful opportunities to help globalize library collections.

Such bullet points, however, provide little perspective on what it is like to actually conduct such research. What has it looked like to data mine the world's information over the past two decades, how have things changed, what trends have remained constant, and what can libraries learn from this? The rest of this article focuses on a selection of highlights tracing my own experiences with large-scale "big data" analysis from founding my first Internet startup in eighth grade twenty years ago through my work today, charting a course through the ever-changing world of the intersection of information and computing.

3. The beginning

My entrance into the world of "big data" and applying massive computing power to explore information began in 1995 when I founded my first Internet startup while still in eighth grade, just a year after the Mosaic web browser was unveiled. Though my startup primarily focused on building software for creating websites, I also worked extensively on web mining, applying algorithms to construct networks of the organizations, topics and emotions surrounding the growing web. In my senior year of high school I joined the National Center for Supercomputing Applications (NCSA), the birthplace of Mosaic, as one of the few high school interns in its history. One of my first projects there was the VIAS Project, a "domain-specific information retrieval, archival, and processing system".[3] VIAS was designed to scour the open web and monitor all relevant mailing lists and USENET groups about a topic of interest, extract an array of metadata from the monitored coverage, and make all of this available through a realtime interactive search interface. It applied a vast library of metadata algorithms to extract everything from person and organization names to locations and topics to bibliographic citations. The latter was of especial interest – a common application of VIAS was to compile realtime topically-focused bibliographies of fields of research, how those works were being discussed, and the topics and people most closely associated with them. Even all the way back in 1999 web users wanted to go beyond searching web pages towards macro-level trends and to connect the electronic world of the web back to the print world of journals and books using online bibliographic citations.

In 2001 I became heavily involved with NCSA's virtual reality facilities, creating a platform called ShadowLight[4] that was used for nearly half a decade by senior and graduate-level university architectural design courses, local middle school outreach programs, and even the United States Army. The goal of the ShadowLight program was to extend the presentation of and interaction with information beyond the mouse and keyboard and flat computer screen into fully-immersive environments where users could literally walk around, look under, and even physically touch and move objects in space around them. While ShadowLight's primary focus was to enable the creation of worlds inside virtual reality,[5] blending organic and inorganic design principles, it was also used to import external data, such as simulations and external data feeds, seamlessly integrating them into a single fused-information environment. Multiple users could import a model of a battlefield and have it appear to float in space in front of them, watch as realtime information flowed in from a myriad of realtime sensor streams, and actually physically walk

[3] http://www.kalevleetaru.com/Publish/NCSA_VIAS.pdf.
[4] http://kalevleetaru.com/Publish/SIGGRAPH_2004_Poster.pdf.
[5] http://kalevleetaru.com/Publish/SPIE_2005_Paper.pdf.

around the model, drawing shared annotations on it, and reaching out to grab objects and physically move them to new places in the display. What was most remarkable about the system was that by replacing the mouse and keyboard with immersive virtual reality, a single user interface was found to be accessible to a broad range of highly-disparate users, from middle school students to professional and student architects to military engineers and strategists, even when presenting large volumes of highly-complex realtime indicators. Fast forward more than a decade to today's virtual reality renaissance, where the multi-million dollar facilities[6] that ran ShadowLight have been replaced with affordable consumer devices, and there is tremendous potential for how libraries might leverage virtual reality to present their enormous archives of material to patrons of the future.

My undergraduate senior thesis in 2004 focused on prototyping the future of mass-scale digital history projects: digitizing, photographing and documenting the history of the University of Illinois' physical buildings and spaces.[7] I personally digitized more than fifteen thousand pages of material and integrated more than seventy thousand pages of historical documents,[8] took more than a quarter of a million photographs capturing every corner of the campus,[9] wrote detailed histories of more than three hundred buildings and spaces,[10] created a virtual tour of campus,[11] and even built a special virtual museum of the university's math models collection.[12] One of the most interesting lessons from the project was that technology had advanced to the point where a single undergraduate could, within the scope of a senior thesis, create one of the larger digital history projects in existence at the time, involving what was in 2004 an enormous amount of digital information. It also required rethinking how to blend textual narrative and digital objects into digital storytelling user interfaces.

Yet, perhaps of greatest interest were the new opportunities all of this data offered for mining the university's history. As one prototype, I wrote a tool to extract every date reference from all seventy thousand pages of digitized material dating back over one hundred and fifty years, resulting in an archive of more than half a million dates. This was used to create a "Today in the History of the University" feature that displays every mention of the present day/month across the University's history.[13] The same year I also developed a suite of tools at NCSA, among them a system called the "Editable Web Browser", designed to allow non-technical users to more easily contribute to and manage vast web-accessible collections of content. My resulting undergraduate research yielded three issued U.S. patents that have been collectively cited by over fifty other patents. Today with the rise of social media there are a myriad opportunities for libraries to engage with their patrons both in helping to digitize historical material and in tools and platforms to present, organize and utilize that data.

4. Building on the foundation

2005: I was the technology lead for one of the NCSA teams building analytic, archival and visualization prototypes for the NARA Electronic Records Archives[14] program. One set of visualizations

[6] http://www.kalevleetaru.com/Publish/NCSA_Cave.pdf.
[7] http://uihistories.library.illinois.edu/.
[8] http://uihistories.library.illinois.edu/cgi-bin/rview?REPOSID=8.
[9] http://uihistories.library.illinois.edu/photoarchive/.
[10] http://uihistories.library.illinois.edu/cgi-bin/cview?SITEID=1&ID=1.
[11] http://uihistories.library.illinois.edu/virtualtour/.
[12] http://www.mathmodels.illinois.edu/cgi-bin/cview?SITEID=4&ID=342.
[13] http://uihistories.library.illinois.edu/cgi-bin/uihist_todayinhistory.
[14] http://www.archives.gov/era/.

explored creating large-scale network visuals[15] on a room-sized computer display[16] to show the complex web of interconnections among the people and topics of large email collections. One of the focal areas of my team's portion of the project was to demonstrate how the digital archives of the future would move beyond simple online file repositories towards offering advanced cloud-based analytic and visualization tools to allow patrons to look holistically across all of that digital material. One of the visions we ultimately presented to NARA was the notion of a data.gov-like system that would make available a range of advanced analytics tools, contributed by volunteer open source developers, allowing data mining across the entire U.S. Government by ordinary citizens, scholars, journalists and even entrepreneurs. While the eventual data.gov platform did not ultimately adopt this model of cloud-based computing for government data, I have worked extensively over the last two years with Roger Macdonald and the Internet Archive to prototype such a system with the Archive's holdings to great success, called the Virtual Reading Room.[17]

5. Information in context

Also in 2005 I began collaborating with my father Hannes Leetaru on a series of explorations of the emotional response to climate change and the energy sector from tracking changing global sentiment[18] to informing community engagement guidelines.[19] Very early on, it became clear that users wanted to go beyond a simple daily compilation of coverage of climate change and energy. Instead, they wanted all of that coverage sorted in order of emotional response, from the most positive (the latest innovations, success stories and applications) to the most negative (new research on limitations or issues and failed projects). In short, users wanted the ability to sort information by its view on each energy sector, rapidly triaging for material on either end of the emotional spectrum. They also wanted the ability to aggregate across news coverage to identify news outlets that tend to emphasize innovations and success stories versus those that focus on the limitations of a given technology, in order to know which outlets to follow more closely. Similarly, users wanted to create maps showing which parts of the world were most closely associated with positive versus negative news about each energy sector and the specific subtopics of particular focus in each area. Legislators and scholars were interested in examining how sentiment changed over time, especially around major events, legislation and marketing campaigns, to examine their impact on popular perception. All of these are queries that libraries are likely to receive in greater volume as information access becomes increasingly algorithmically mediated.

When we think of "information" we tend think of words or numbers on a page, devoid of emotional or narrative context. Yet, information does not exist in a vacuum: the context within which a statement or action appears and the interface through which we access it can entirely change how we understand it. Even the act of sharing information with others does not occur in a contextual vacuum and instead draws deeply upon shared context – though the increasing brevity of the mobile era requires ever-more information to understand that context. In the past one might recommend or disprove of a book by writing a lengthy review describing in detail one's thoughts and reactions to each section of the book.

[15] http://kalevleetaru.com/Publish/NARA_Email_Network_Visualization_Analysis.pdf#page=10.
[16] http://www.kalevleetaru.com/Publish/NCSA_Tiled_Display_Wall.pdf.
[17] http://www.knightfoundation.org/blogs/knightblog/2014/1/7/internet-archives-virtual-reading-room-empowers-data-mining-societal-scale/.
[18] http://carboncapturereport.org/.
[19] http://www.wri.org/publication/guidelines-community-engagement-carbon-dioxide-capture-transport-and-storage-projects.

Today that recommendation or condemnation might instead appear as a simple web hyperlink, social media share or retweet that could equally imply endorsement of the content being shared or derision of it. Instead of the sharing of information being a self-contained act, understanding sharing behavior now requires extensive understanding of the surrounding commentary and shared background knowledge of the persons sharing and receiving the information.

This has especial concern to libraries as the algorithms and user interfaces that drive social media platforms play an increasingly outsized role in the information to which users are exposed. The lack of a "dislike" button on Facebook[20] has arguably helped contribute to a dichotomy where Facebook "tend[s] to focus on positive social interactions"[21] while Twitter emphasizes exposure to a raw unfiltered view of events.[22] As algorithms increasingly guide the information environments to which patrons are exposed online, narrowing information-seeking behavior along economically-incentivized pathways, libraries must reexamine their role in the information ecosystem. Users accustomed to environments that transparently customize the information they receive based on their views and beliefs will likely demand greater ability to access information based not just on topic, but on the perspective and narrative surrounding that topic. Underlying this emerging world of information access is the core concept of accessing information not merely by topical focus, but through the emotional, narrative, and network lens through which it presents that information and the interfaces used to access it. As information increasingly takes on specific alignment with the topics it covers, patrons will demand access mediated through these lenses.

2005 to 2006: During those years I took over the technical reins of another major digital history project called RiverWeb, which focused on the history and evolution of the Mississippi River in North American life. One of the initiatives I spearheaded was the digitization and integration of a vast archive of digitized newspapers, book chapters, autobiographical letters, city directories and Geographic Information System (GIS) datasets documenting the history of East St. Louis and its relationship to the Mississippi River. A key focus of this initiative was on the integration of advanced data mining tools and visualizations to allow historians to trace the networks of people, locations, dates and events that define a place over time. In one experiment I extracted all of the people, locations and dates from a collection of digitized newspaper articles and created a massive interactive network diagram showing their interconnections. As with my UIHistories project mentioned earlier, a key lesson learned from this project was that as one digitizes large volumes of content and makes it available in digital libraries, patrons increasingly want to move beyond simple keyword search towards visualizing the entities and networks contained within. In short, patrons want to move beyond text towards understanding what that text is saying about their topic of interest.

In 2006 as part of my graduate research, I created the Profile of a Campus project,[23] which collected all of the University of Illinois' available open datasets on the functioning and research of the institution, along with a snapshot of the institution's web space. This ranged from financial and operations statistics to human resources and phone directories to graduation rates to a list of all publications by the university's engineering faculty. It even included a massive web crawl of the top one hundred pages from each of the university's three hundred departments and centers. This collection of pages was processed to identify the core topics and themes of each page and its links to all other university pages and

[20] http://www.slate.com/articles/technology/future_tense/2014/12/facebook_dislike_button_why_mark_zuckerberg_won_t_allow_it.html.

[21] http://www.huffingtonpost.com/2013/04/03/facebook-dislike-button-bob-baldwin_n_3006997.html.

[22] http://www.usatoday.com/story/tech/2014/09/02/facebook-twitter-ferguson-icebucketchallenge/14818505/.

[23] http://www.kalevleetaru.com/profileofacampus/.

brought together into a single massive network diagram showing how the institution's web presence was connected and the picture it presented of the institution as a whole. The datasets used in this project came from across the university and were used in very different contexts from that for which they were originally created, but by virtue of the university making them available as open data, a graduate student was able to mash them together to create a powerful new way of looking across the functioning of the entire institution. This is one of the greatest potentials for the open data world, of the ability for even students to create powerful new views onto society by remixing data in unexpected ways, and is an area in which libraries can play a key role in by making open data easily accessible and helping patrons to navigate the landscape of available repositories.

2007: This year I began to build the digitization, communications, coding, data management and technical infrastructures for the Comparative Constitutions Project (CCP)[24] that provides the source texts for Google Constitute.[25] The CCP project collects and digitizes constitutions from throughout the world and uses a large globally-distributed team of human analysts to read each constitution and fill out a massive survey instrument recording the constitutional legal structure of the nation. One element of the project involved creating ways of achieving better visibility into the functioning of such a large geographically-distributed team of analysts who needed to collaborate closely and maintain in constant communication. This resulted in a series of visualizations that automatically constructed biographical profiles of each analyst, identifying with whom they most frequently communicated throughout different times of day, the topics they struggled with and those that they were deemed as experts on, and their informational role within the network. A key finding of this system was that the vast commentary and dialog around a document can yield critical insight into how it is understood by a community and just how different this view can be compared with self-identified or administrator-assigned metadata. For libraries, this suggests they should find ways of moving beyond the static expert-assigned metadata of their OPAC catalogs towards systems that are able to better capture the changing cultural impact of a work over time.

2008: During this year I collaborated with another researcher to examine the mutability of U.S. Government records in the digital era. The White House was found to have engaged in a long-standing and constant process of retroactive editing of its official press releases, altering critical details and changing their publication dates to rewrite key aspects of the U.S. involvement in Iraq. The study was published in *The New York Times*,[26] whose Editorial Board built upon it the following day in an article titled "Orwell Comes to Iraq".[27] The project shone a light on the transformation of government records in the digital era from permanent records with legal status to ethereal collections of digital bits that can be constantly rewritten over time. It also demonstrated the enormous value of web archives, making extensive use of the Internet Archive to compare versions of each press release over time in identifying when each was changed.

Later that year I compared the operation of the Google Books and Open Content Alliance (OCA) mass digitization programs.[28] A key finding was that there was considerably more technical information available on the functioning of Google Books than OCA, but that the majority of it was being published in the technical computer science literature, rather than in library or information science venues. As

[24] http://comparativeconstitutionsproject.org/.
[25] https://www.constituteproject.org/.
[26] http://www.nytimes.com/2008/11/25/washington/25documents.html.
[27] http://theboard.blogs.nytimes.com/2008/11/25/president-bushs-coalition-of-the-willing-or-orwell-comes-to-iraq/.
[28] http://journals.uic.edu/ojs/index.php/fm/article/view/2101/2037.

computer scientists and Internet companies play increasingly central roles in the digitization, management, analysis and interaction with the world's information, this trend will only intensify, meaning that the library community must engage ever more closely with the technical community on its own turf of technical conferences and journals. I also used a vast archive of global news material to examine how the topical focus of news coverage of the collapse of WorldCom[29] exhibited strong geographic stratification, demonstrating the ability of large-scale thematic data mining to uncover macro-level trends.

6. Information bias

2009: In the following year I utilized another large historical web archive to explore the sourcing behavior of the *Drudge Report*,[30] one of the web's earliest "new media" news outlets. Despite *Time Magazine* naming its founder as one of the world's most influential people, *The Washington Post* identifying the site as its single largest traffic driver, and both ABC and CBS News claiming that the report single-handedly drove the political news cycle, there had been little formal study of the site in its role as a media driver and the way it selected stories to feature. A key finding of the study was that as information consumers we rarely bother to try to understand the selection process through which we are provided with information and the biases that process may have on how we understand a given topic – an issue of especial importance in the era of increasing algorithmic selection of information relevance. This also holds true for understanding how web archives select the material they archive and the frequency with which they archive a site, an issue I expanded upon in my 2012 IIPC keynote address.[31] The Internet Archive's sporadic snapshots of the *Drudge Report* were found to be too few to be usable for this study, which ultimately relied on a specialized web archive focusing exclusively on the *Drudge Report* which offered a fixed snapshot interval of every two minutes over more than six years.

2010: This year I returned to the notion of understanding where our information comes from with the first unclassified quantitative examination of how Western Open Source Intelligence agencies gather news material from throughout the world.[32] The collection of foreign news content by the U.S. Government was found to be highly biased, with significantly greater collection of material from Russia than from the entire Latin American region, Spain and Portugal combined, while European news agencies were found to be the primary source of its coverage of Africa. Mirroring the findings of my *Drudge Report* study, U.S. Government collection of news was found to center on collecting what was most easily accessible without ever stopping to quantitatively assess the resulting collection and sourcing behavior to understand the resulting biases[33] and coverage holes.[34]

Later that year I explored the changing face of news coverage of higher education by examining *The New York Times* coverage of research universities over the last half century.[35] A key finding was how much change even a major news source like *The New York Times* can undergo over time: the total number of articles published in the paper per year has shrunk linearly over the past sixty years, decreasing

[29] http://www.kalevleetaru.com/Publish/Open_Source_Intelligence_FBIS_WorldCom.pdf.

[30] http://journals.uic.edu/ojs/index.php/fm/article/view/2500/2235.

[31] http://blogs.loc.gov/digitalpreservation/2012/05/a-vision-of-the-role-and-future-of-web-archives-conclusions-and-the-role-of-archives/.

[32] https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/volume-54-number-1/the-scope-of-fbis-and-bbc-open-source-media.html.

[33] http://foreignpolicy.com/2015/04/15/why-we-cant-just-read-english-newspapers-to-understand-terrorism-big-data/.

[34] http://www.foreignpolicy.com/articles/2014/09/26/why_big_data_missed_the_early_warning_signs_of_ebola.

[35] http://kalevleetaru.com/SoundbiteUniversity/.

by more than half. In turn, the number of articles mentioning research universities has remained nearly constant over those sixty years, suggesting at first that coverage of higher education has remained unchanged. However, because the total size of the paper has shrunk by half, the overall density of higher education references in *The New York Times* has actually doubled, presenting a very different picture. Much of the social sciences and humanities literature that explores news trends uses absolute counts of the number of articles mentioning a given topic over time, which fails to account for the dramatic change in total volume of many sources. For libraries, this means they must understand not only their own collection behaviors, but also the behaviors of the sources they collect in order to best advise their patrons.

2011: This year my Culturomics 2.0[36] study used a massive large-memory supercomputer to analyze a collection of more than one hundred million global news articles, extracting over ten billion people, places and things, and more than one hundred trillion relationships[37] to create an enormous network diagram over 2.4 petabytes in size spanning half a century of news. In what was at the time the largest deployment of sentiment analysis, more than fifteen hundred emotions and themes were assessed from all one hundred million articles, yielding over one hundred and fifty billion emotional-thematic measurements. A key discovery of the research was that the average positive/negative "tone" of global news coverage about a country offers a short-horizon, high-accuracy forecast of its political stability over subsequent weeks: the first study to quantitatively demonstrate at scale the link between latent news emotion and future political stability for risk assessment. It was also the first to combine the geography[38] and emotion[39] of news media at global scale to produce a series of animations visualizing worldwide news emotion over a quarter century and demonstrating that the geographic affinity of public figures in news coverage offers high-resolution estimates of their travels. Yet perhaps most relevant to the library community, the study became a poster child of the "post theory" era of "big data" in which massive collections of data, enormous computing power, and innovative algorithms are used to find the patterns in data that lead to discoveries.[40] In contrast to the theory-driven research that has historically driven academia, where data is used merely to test human-derived hypotheses, Culturomics 2.0 "helped usher in the era of 'petascale humanities"' where computers themselves "identify useful or interesting patterns if provided with sufficiently large data repositories".[41] As academic research increasingly turns to large data archives, libraries will play an ever-more central role in facilitating and helping to manage these vast archives.

7. Sentiment analysis

2012: The following year I conducted three massive studies exploring the intersection of geography and emotion beyond news media to encyclopedias, books and social media. In the spring of 2012 I collaborated with supercomputer vendor Silicon Graphics International (SGI) to map the emotional undercurrents of Wikipedia's four million English language articles.[42] The resulting two hundred year

[36] http://firstmonday.org/ojs/index.php/fm/article/view/3663/3040.
[37] http://www.hpcwire.com/2011/09/09/nautilus_harnessed_for_humanities_research_future_prediction/.
[38] http://blogs.loc.gov/digitalpreservation/2015/04/mapping-words-lessons-learned-from-a-decade-of-exploring-the-geography-of-text/.
[39] http://www.knightfoundation.org/blogs/knightblog/2014/10/22/our-global-dreams-and-fears-news-emotion/.
[40] http://www.wired.co.uk/news/archive/2013-01/25/big-data-end-of-theory.
[41] http://www.hpcwire.com/2013/09/12/can_supercomputers_predict_the_future_/.
[42] http://www.sgi.com/go/wikipedia/.

animation of Wikipedia's view of world history[43] was the first large-scale exploration of the emotion of encyclopedias, finding that despite an emphasis on factually-oriented content, encyclopedias still contain considerable measurable emotional language. Yet, perhaps of greatest importance to libraries, the study found that Wikipedia's user-contributed geographic metadata was heavily biased and did not reflect the actual geography of its articles.[44]

A few months later in the summer of 2012, in collaboration with noted American South scholar Vernon Burton, I applied sentiment and thematic analysis to more than one billion pages of digitized books from the HathiTrust and the Internet Archive to explore the spread of ideas and emotions over space during the American Civil War, producing the first at-scale micro-level view of emotion and narrative across millions of books. This project emphasized the critical disconnect between the current state of sentiment analysis tools and their application to historical material. The majority of current sentiment tools have their origins in the last three decades, with most built in just the last fifteen years, meaning they record the emotional connotations of words as they stand in the early 21st century. Many words, however, have undergone substantial change in connotation over time, especially when looking across the one hundred and fifty years since the American Civil War. Even when adjusting language use to the Civil War era, words can take on very different meanings depending on the speaker, having a positive connotation when spoken by a Northerner and a negative connotation when spoken by a Southerner, for example. However, assessing sufficient information about the speaker to determine such use can be difficult with historical material. Libraries, which often house large historical collections from a diversity of perspectives and authors, should take such issues into consideration when applying sentiment tools to their holdings, as discussed further in a moment.

Finally, in the fall of 2012 I worked with SGI once again, this time to create the first realtime geographic maps of the emotion of social media,[45] which included the first live maps of the national emotional response to a natural disaster[46] and a presidential election.[47] This required considerable adaptation of sentiment algorithms to the high level of typographical error, shorthand expression, and assumed shared background context of social media. In particular, social media exhibits many of the same issues as historical archives: many words have different connotations and meanings depending on their speaker and it can be difficult to assess the background and intent of speakers on Twitter given the paucity of available biographical information about a given user.

Two other co-authored studies of mine in 2012 explored the concepts of "readability" scores and "group-based discourse" in the understanding of large collections. Following in the footsteps of my 2008 thematic analysis of global news coverage of the WorldCom collapse,[48] I helped examine the posturing of global news coverage of the 2003 U.S. invasion of Iraq,[49] assessing the degree to which the presses of the world portrayed the invasion in purely clinical terms or whether it was positioned in terms

[43] https://www.youtube.com/watch?v=KmCQVIVpzWg.

[44] http://blogs.loc.gov/digitalpreservation/2015/04/mapping-words-lessons-learned-from-a-decade-of-exploring-the-geography-of-text/.

[45] http://www.sgi.com/go/twitter/.

[46] https://www.youtube.com/watch?v=g3AqdIDYG0c.

[47] https://www.youtube.com/watch?v=oVaBws-3BVs.

[48] http://www.kalevleetaru.com/Publish/Open_Source_Intelligence_FBIS_WorldCom.pdf.

[49] http://www.kalevleetaru.com/Publish/ISA_2011_Stake-In-This-War-Worldwide-Test-In-Group-Out-Group-Open-Source-Intelligence.pdf.

of "us versus them" in which the U.S. was seen as a shared enemy invading not only Iraq, but all other countries that shared its cultural background. A second study examined the linguistic complexity of two hundred years of the world's constitutions,[50] finding that those with less repetition and greater use of singular words significantly increased the comprehensibility of the constitution in terms of being able to successfully codify its legal structure. Both studies demonstrated the importance of inventorying the posturing and "readability" of materials in order for libraries to better sort material not simply by topic, but by how that topic is presented.

2013: In March I successfully defended my doctoral dissertation,[51] which expanded upon my Culturomics 2.0 work to explore the ability of the narrative and emotional undercurrents of news coverage to successfully forecast future small-bore unrest. The dissertation applied an array of data mining algorithms to a collection of more than 4.8 million news articles totaling 1.35 billion words comprising all international coverage from Agence France-Presse, Associated Press and Xinhua dating back more than thirty years. It also resulted in the creation of a massive database of "events" recording the factual activities described in those 1.35 billion words of news content, from riots and protests to peace appeals and aid promises, along with the precise geographic coordinates and over sixty details about each event. At the time of its release it was the largest political event dataset in the world, dwarfing even a $125 million dollar U.S. Government program.[52]

8. New era of data mining

Perhaps most significant from the perspective of libraries is that this dissertation represented a new era of "big data" research in which massive-scale data analytic projects were no longer solely the realm of large multi-million-dollar campus research labs led by faculty or large commercial entities. Such projects could now just as easily be a single student's personal dissertation research unaffiliated with a campus lab, unfunded, and not part of traditional library–faculty engagement. This presents uncharted territory for academic libraries in which they will be increasingly called upon to support student research outside the bounds of traditional grant-funded faculty research and lacking the equivalent single points of contact and well-established cost-recovery mechanisms to reimburse the necessary library personnel and resource time. Whether it is a student digitizing and publishing tens of thousands of pages of historical material and capturing hundreds of thousands of photographs like my 2004 undergraduate thesis, or data mining billions of words of news content and producing the world's largest political event dataset like my 2013 graduate dissertation, students will play an ever-increasing role in the data ecosystem of the future. Libraries must consider how they reach out to students, how they help facilitate the copyright clearances and licensing agreements necessary to work with large content collections, and how to help with the long-term archival and preservation of those projects, all without the traditional cost-recovery streams historically provided through large faculty research projects.

Just a few weeks after my dissertation defense, I unveiled the GDELT Project,[53] which is a realtime index over the global news media, monitoring local news outlets in every corner of the world in more

[50] http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2191145.
[51] http://www.kalevleetaru.com/Publish/Leetaru_Dissertation_Can_We_Forecast_Conflict-Dissertation.pdf.
[52] http://www.foreignpolicy.com/articles/2014/01/03/half_a_billion_clicks_can_t_be_wrong_mapping_crisis_zones.
[53] http://gdeltproject.org/.

than one hundred languages to identify the people,[54] locations,[55] counts,[56] themes,[57] emotions,[58] narratives,[59] events[60] and patterns undergirding global society.[61] Today it is the largest event dataset ever created, totaling more than 300 million georeferenced event records stretching back thirty years, the largest deployment of streaming machine translation,[62] live translating all monitored news in sixty-five languages, the largest deployment of sentiment analysis,[63] assessing more than forty-five hundred emotions and themes from every article, the largest deployment of multilingual geocoding,[64] identifying and disambiguating every mention of location across each article, the largest real-time, multilingual news-based data mining program,[65] and the largest single program to preserve the online journalism of the non-Western world.[66] Among its special collections is the largest socio-cultural graph over academic literature and the first large-scale content analysis of a web archive,[67] the largest socio-cultural graph over human rights documentation,[68] and the first large-scale emotional analysis of television news.[69] The entire output of the project is released as fully-open data, with every piece of computed metadata available for download and use, feeding projects from the United Nations to the U.S. Institute of Peace. While GDELT occupies a unique role as the flag bearer of the emerging era of massive data-driven social sciences and humanities with its pioneering debut of new datasets, methodologies and scales to the field, the intense interest it has driven over the past two years means that libraries will find themselves fielding an increasing number of requests for help in accessing large digital datasets as non-traditional disciplines like the social sciences and humanities show greater interest in data-driven scholarship. As with student research, academic libraries in particular will play a crucial role in helping to connect scholars with data vendors, assist in licensing and other arrangements, offer guidance on data management and preservation, and help orchestrate the data ecosystem of their campuses.

9. Visualization and mapping

In May 2013 I returned to my previous digital history focus in a collaboration with Columbia University to mine and visualize the vast NARA archive of declassified State Department cables[70] covering the

[54] http://www.foreignpolicy.com/articles/2014/06/09/everybody_loves_bashar_gdelt_putin_assad_media.
[55] http://www.dlib.org/dlib/september12/leetaru/09leetaru.html.
[56] http://blog.gdeltproject.org/typhoon-ruby-realtime-aid-tracking-pilot-with-ocha/.
[57] http://dlib.org/dlib/september14/leetaru/09leetaru.html.
[58] http://www.knightfoundation.org/blogs/knightblog/2014/10/22/our-global-dreams-and-fears-news-emotion/.
[59] http://www.foreignpolicy.com/articles/2014/10/24/don_t_blame_cnn_for_the_ebola_panic_media_coverage.
[60] http://www.foreignpolicy.com/articles/2014/05/29/did_the_arab_spring_really_spark_a_wave_of_global_protests_GDELT.
[61] http://www.bbc.com/news/technology-28895098.
[62] http://blogs.loc.gov/digitalpreservation/2015/04/libraries-looking-across-languages-seeing-the-world-through-mass-translation/.
[63] http://www.knightfoundation.org/blogs/knightblog/2014/10/22/our-global-dreams-and-fears-news-emotion/.
[64] http://blogs.loc.gov/digitalpreservation/2015/04/mapping-words-lessons-learned-from-a-decade-of-exploring-the-geography-of-text/.
[65] http://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/.
[66] http://www.knightfoundation.org/blogs/knightblog/2015/3/27/looking-across-languages-seeing-world-through-mass-translation-local-news/.
[67] http://dlib.org/dlib/september14/leetaru/09leetaru.html.
[68] http://blog.gdeltproject.org/announcing-the-new-human-rights-global-knowledge-graph-hr-gkg/.
[69] http://blog.gdeltproject.org/visualizing-the-emotions-of-american-television-news/.
[70] http://diploglobe.declassification-engine.org/.

period 1973–1976. While the collection was already available in searchable digital format for historians, the lack of a visualization interface allowing spatial and temporal visualization of the cables made it difficult for scholars to make sense of the collection. A unique geographic network visualization tool from Google Ideas was adapted to allow the complete set of cables to be displayed in a country-centric view highly aligned with how historians understand the flow of diplomacy. A key lesson from the project was that simply releasing large datasets for research will often yield little impact until the necessarily visualization and interface tools are available to make it possible for users to intuitively interact with them.

In the fall I collaborated with Roger Macdonald and the Internet Archive's Television News Archive to create the first large-scale map of the geography of television news.[71] More than four hundred thousand hours of closed captioning of American television news totaling over 2.7 billion words was geocoded to produce an animated daily map of the geographic focus of television news from 2009–2013. Once again, this involved taking information created for one reason (closed-captioning streams intended for the hard of hearing) and repurposing it for another (as a textual proxy for the contents of television news programming). As libraries explore making their digital collections available for research, they should take into consideration these kinds of unexpected applications, ensuring that their technical architectures allow data to be made available in flexible ways. In this case, the Internet Archive had had the foresight to enable closed-captioning streams to be accessed as plaintext ASCII files, one per broadcast, making it trivial to apply a wide range of data mining tools to them.

That same fall I was also approached by NBCUniversal's Syfy channel to create the first live emotion-controlled television show[72] using the discussion on Twitter[73] about Syfy's new Opposite Worlds show to generate a live "leaderboard" ranking all of the show's contestants in order from most to least "popular" that was used to drive key elements of the show.[74] This presented enormous challenges, both technical, in adapting sentiment technology to television use,[75] and methodological, around how to present a topic as complex as sentiment analysis to a non-technical television audience and how to drive a television show based on live audience feedback. The resulting system, which debuted in the spring of 2014, set numerous records for social engagement with a television show and was used again in the fall for Syfy's flagship Face Off show. Creating a visual interface that masked the enormous complexity of sentiment analysis through an intuitive display that users could easily understand, even if they were just tuning into the show for the first time, posed considerable challenges. It required reaching deeply into visual metaphors popular with Syfy's demographics, including sporting and gaming "leaderboards", as well as a color scheme that was not overly-rooted in its own emotional connotations. This is a crucial concept for libraries to consider as an ever-greater portion of the world's information is accessed through electronic systems: interfaces must adapt to users, rather than users being forced to learn new interfaces.

From a technology standpoint, it turns out that many sentiment analysis tools today still use simple hand-categorized lists of "positive" and "negative" words and count up how many times words from each list appear in a given tweet,[76] changing little from the first computerized sentiment analysis system,

[71] http://blog.archive.org/2013/12/13/mapping-400000-hours-of-u-s-tv-news/.

[72] https://web.archive.org/web/20140802030028/http:/technorati.com/how-viewer-social-media-control-is-the-future-of-television/.

[73] https://blog.twitter.com/2014/opposite-worlds-collide-on-twitter-with-syfytv.

[74] http://www.latimes.com/entertainment/tv/showtracker/la-et-st-social-media-tv-20140309-story.html.

[75] http://www.wired.com/2014/06/how-to-teach-heartless-computers-to-really-get-what-were-feeling/.

[76] http://www.wired.com/2014/06/how-to-teach-heartless-computers-to-really-get-what-were-feeling/.

General Inquirer,[77] which debuted in the era of punch card computers more than half a century ago. Other systems use machine learning algorithms to autonomously "learn" the emotional connotations of words, but these approaches often "learn" that "Monday", "treadmills", "airports", "economists", "hospitals", "orthodontists", "doctors" and "dentists" are all highly-negative concepts that provoke dread, which becomes problematic when such systems code "Monday night football" or "I'm getting married on Monday" as extremely negative concepts. Language changes over time,[78] with the emotional connotations of words evolving even over the course of a few years. Many systems widely used in the academic literature today are based on training data that predates the social media era and thus do not include emoticons or "lol"'s, while "cool" means "cold". Nearly all current systems are based on modern language use, causing them to yield highly skewed results when applied to documents from the 1800–1923 period that constitutes the majority of today's historical digital libraries.

As noted earlier, words often have very different connotations to both their authors and speakers. A word might be highly derogatory when spoken by one demographic, but used to indicate affection by another. To someone who almost drowned as a child, the word "ocean" might prompt traumatic memories, while to someone who was on summer swim team and boated as a child, it might evoke utopian images of summer childhood. Emotion around conflict is especially complex. The classic example is a sporting match: the following day the hometown newspapers of both teams will report the same factual details of the game, but the winning team's paper will likely portray the game as a tremendous success, while the losing team's paper will portray it as an utter failure. In war, a military attack that leads to substantial loss of life on the receiving side, but not the other will often be portrayed in glowingly positive terms by the media of the side conducting the attack as indicative that it is winning the war. In short, there is no universal emotional undercurrent to any text and any attempt to assess the emotional dimensions of text must take all of these issues into account.

For libraries, this means being wary of the myriad of tools today claiming to provide high-accuracy "sentiment mining" of their collections or offering "emotional search" tools. The current state of sentiment mining technology is not yet sufficient to accurately assess a broad range of emotions across the vast thematic, temporal and spatial range of most library collections. At the same time, such tools can yield highly usable results on selected genres of material and libraries should be open to experimenting with such technologies on more narrow collections and listen closely to patron feedback to explore whether particular materials are more amenable to such analysis. As users become increasingly used to having their information transparently filtered to emphasize certain emotions[79] libraries will need to explore the ramifications and abilities of such filtering within their own collections and patron communities.

2014: In the fall of 2014 I wanted to further explore the potential of large-scale emotional-thematic assessment of library collections. I launched two of the largest deployments of sentiment analysis technology, one applied to American television news and the second applied to academic literature. Following in the footsteps of my fall 2013 collaboration with the Internet Archive's Television News Archive to explore the geography of television news, I once again worked with the Archive to assess more than twenty-three hundred emotions and themes from five years of American television news totaling more than five hundred and forty thousand hours. This included both an interactive timeline visualization

[77] http://www.computer.org/csdl/proceedings/afips/1963/5062/00/50620241.pdf.
[78] http://ideas.ted.com/20-words-that-once-meant-something-very-different/.
[79] http://www.huffingtonpost.com/2013/04/03/facebook-dislike-button-bob-baldwin_n_3006997.html.

tool[80] and downloadable datasets of all of the assessed emotions[81] Instead of building yet another sentiment dictionary, this project brought together eighteen of the major sentiment packages used in the academic literature to capture a vast array of concepts from "abashment" to "wrath".[82] By bringing together existing sentiment packages, users are able to tap into the emotions, tools and literatures of their respective disciplines or which are of greatest interest to their given query.

10. Socio-cultural content mining

In the second of the two collaborations, I worked with researchers from the U.S. Army to conduct the first large-scale socio-cultural content analysis of academic literature and the open web.[83] More than twenty one billion words of academic literature were processed, including the entire contents of JSTOR, DTIC, CORE, CiteSeerX and the Internet Archive's 1.6 billion PDFs relating to Africa and the Middle East, along with a parallel project conducting the first at-scale content analysis and mapping of human rights reports.[84] Given the magnitude of data being analyzed and the multiple source platforms being used, the project necessitated a blended data management model, where computing was split, with the Internet Archive's 1.6 billion PDFs being analyzed using the Virtual Reading Room,[85] while the remaining content was processed using Google Cloud.

No project had attempted socio-cultural content mining of full-text social sciences and humanities journal articles or open web content at this scale before, so a substantial portion of the project was devoted to prototyping and pioneering the technical workflows and methodologies for mining content at this scale. Unlike science and engineering publications, which feature comprehensive summaries and make use of concise detail-rich language that is traditionally standardized across a field, the literature of the humanities and social sciences tends to emphasize flowery language, is rarely standardized and lacks concise summaries. Many of the tools and techniques developed for fields like medical literature mining therefore struggle with its diffuse nature. Even citation extraction becomes far more complex, with an incredible diversity of citation styles and relatively loose editorial enforcement of those styles: papers in the same issue of a journal can sometimes use very different citation formats. Similarly, producing the first large-scale socio-cultural analysis of open web content required creating new workflows for interacting with the Internet Archive's web archive and dealing with the enormous diversity of publication styles found in nearly twenty years of the web's PDFs.

For libraries, the project offers a glimpse at the future of large-scale data-driven social sciences and humanities research, involving multiple data platforms, massive numbers of publishers and computation distributed across multiple computing platforms with highly complex workflows, large numbers of cross-disciplinary algorithms and analytic tools, and advanced visualization requirements. While in this case I oversaw all data and technical aspects of the project by myself, in the general case libraries should be prepared to work with large institutionally-distributed teams spanning many different disciplines when engaging with these kinds of projects. In some cases libraries may be asked only for an introduction to connect scholars with data vendors and publishers, while in other cases libraries may be called upon to

[80] http://analysis.gdeltproject.org/cgi-bin/iatvemotions/iatvemotions.
[81] http://blog.gdeltproject.org/visualizing-the-emotions-of-american-television-news/.
[82] http://www.knightfoundation.org/blogs/knightblog/2014/10/22/our-global-dreams-and-fears-news-emotion/.
[83] http://dlib.org/dlib/september14/leetaru/09leetaru.html.
[84] http://blog.gdeltproject.org/announcing-the-new-human-rights-global-knowledge-graph-hr-gkg/.
[85] http://www.knightfoundation.org/blogs/knightblog/2014/1/7/internet-archives-virtual-reading-room-empowers-data-mining-societal-scale/.

provide far more extensive support and data management resources, especially for projects that do not have substantive technical collaborators to assist in those areas.

11. Projects of the future: Rapid turn-around

Yet, perhaps of greatest impact to libraries is that in the past such endeavors have largely involved multi-year timelines with significant planning periods and evenly spaced milestones that afforded plenty of time for discussion and collaborative steering. As the computer and information science disciplines play greater roles in such research, with their emphasis on short-turnaround, rapid-fire publication schedules, the projects of the future will increasingly occur in time horizons of less than a year, with very short planning periods and rapid milestones. This will place greater pressure on libraries to act as data clearinghouses, rapidly connecting scholars with data collections and quickly stepping back, and thus may limit the staff resources they are able to contribute to fully-collaborative endeavors. It also means libraries must contemplate a future that revolves around massive data- and computation-driven projects orders of magnitude larger than they have ever seen and where the entire project launches and completes in a timespan shorter than the preplanning phases of past projects.

12. Re-imagining the book

In my final collaboration of 2014 I expanded upon this concept of rapid-turnaround "whole of archive" research in a massive exploration with the Internet Archive to reimagine the concept of the book. Instead of thinking of books as textual documents, what if we thought of them as containers of imagery? In doing so, what would it look like to extract every illustration, drawing, chart, map or photograph from the world's digitized books, along with their surrounding text,[86] and make it possible to navigate the world's books not as paragraphs of text, but as a visual tapestry of history?[87] In the end, the images of more than six hundred million digitized book pages from over one thousand libraries spanning five hundred years of history were extracted and uploaded to Flickr to create a fully-searchable visual archive of the imagery of half-a-millennium of books.[88]

While the final results of the project ultimately debuted to the public in August 2014, I had first approached the Internet Archive with the idea in fall of 2013 to begin a dialog around the technical feasibility and logistics of how to actually process the Archive's entire public domain book catalog given its breathtaking size (the raw page scan imagery can total up to one gigabyte per book). While numerous previous projects had extracted imagery from digitized books, none had ever attempted an endeavor of this sheer scale, both in total volume of material being processed and the range of time periods, genres and collections being processed. In addition, previous projects had extracted only large imagery and had not attempted to contextualize that imagery by connecting it to the text immediately surrounding it on the page. A key lesson learned from those efforts was that users do not want to manually scroll through millions of images a page at a time – the web has accustomed users to interactive keyword search that returns relevant results instantly. It was also unclear how useful the resulting images would be when

[86] http://blog.flickr.net/en/2014/08/29/welcome-the-internet-archive-to-the-commons/.
[87] http://blog.archive.org/2014/08/29/millions-of-historic-images-posted-to-flickr/.
[88] https://www.flickr.com/photos/internetarchivebookimages/.

processing the entirety of a book collection as varied as the Archive's – previous initiatives had largely used hand-selected subsets of works known to have particularly high-quality illustrations.

The first phase of the project was therefore a system prototype that involved processing the entirety of the Archive's collections using low-resolution page imagery designed for e-reader devices and focusing on fine tuning the algorithms to identify imagery and extract usable text around them.[89]

The results of the pilot were so successful that the process was immediately repeated, this time with the full resolution page scan imagery. Instead of attempting to build algorithms to recognize images from page scans, the project made creative reuse of the OCR XML files produced as a byproduct of making the books full-text searchable. One of the greatest challenges of the project was the technical demands of processing such an enormous volume of scanned page imagery, extracting the full-resolution images from each of those pages, and then saving those back to disk. The enormous disk IO ultimately became the primary bottleneck of the project and the overwhelming majority of technical development was spent on absolute minimization of data movement.[90]

Finally, once all of the images were extracted, they had to be saved somewhere and made available for patron access. In most projects, large volumes of data may be processed, but the final output that must be saved and made available is relatively small in size. In this project, the total volume of output material (14.7 million high resolution images) was many terabytes in size and required an entirely separate workflow to be able to permanently store and make available all of that material. This is an area that libraries have not for the most part had to cope with yet. At academic institutions, most institutional repositories are designed to archive relatively small PDF or ZIP files, not single projects comprising tens of millions of objects totaling terabytes and with an array of associated index and metadata database files. All of the extracted images and metadata were made available for download on the Internet Archive's website, but to make them truly accessible to the general public required an interface and technical infrastructure very different from that used by the rest of the Internet Archive. Thus, I reached out to Flickr to host and make the collection publicly accessible. Hosting within Flickr also allowed the collection to leverage Flickr's enormous investment in user interface, mobile access, social sharing and community engagement.

The project ultimately debuted to the public in August 2014, but the entire project from start to finish took less than a month and a half of my time – the rest of the time was spent in discussions with users, staging of the images for public download, and uploading and customizing the interface to the images in Flickr. The project presented a host of technical challenges of particular relevance to libraries contemplating large-scale analysis of their collections, from the limiting nature of IO at these scales, to the fact that some projects will require libraries to not just provide access to their collections, but also offer long-term hosting for the output of projects using those collections. Yet, perhaps the greatest lesson for libraries is the critical importance of interface. Even the greatest collections will sit largely unused if patrons cannot easily and intuitively interact with them and if they do not have sufficient contextualizing metadata to make it possible to rapidly search the collection at a high degree of resolution and detail. Libraries should explore partnerships with organizations that offer specialized interfaces and tools that can help them make their content available in specialized formats without libraries themselves having to build every single tool.

Finally, most recently, in the spring of 2015, GDELT 2.0 launched.[91] This now brings together twenty-four emotional and thematic assessment packages that collectively assess more than forty five hundred

[89] http://blogs.loc.gov/digitalpreservation/2014/12/unlocking-the-imagery-of-500-years-of-books/.
[90] http://blogs.loc.gov/digitalpreservation/2014/12/unlocking-the-imagery-of-500-years-of-books/.
[91] http://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/.

emotions and themes[92] and live translates sixty five languages in real-time,[93] along with new prototype visualizations to the world's news.[94] In pushing beyond the limitations of the English- and Western-centric view of the world[95] the project is exploring how "big data" can reshape public diplomacy,[96] assist in disaster response,[97] and offer a new lens onto global conflict.[98] As an open data initiative, analysis and visualization platforms from Google BigQuery to Tableau to CartoDB to Palantir to Lux have all made GDELT accessible within their systems, offering an incredible diversity of interfaces and access modalities. Libraries simply do not have the resources to try to build every possible tool for interacting with their collections – partnering with other organizations can allow them to offer highly-customized experiences to their collections leveraging the latest technologies.

13. Conclusion

Over the course of the last two decades I have explored the informational undercurrents of the world's information and the potential of mass data mining of libraries through a myriad of lenses both technical and methodological. From founding my first Internet startup twenty years ago as an eighth grade middle school student, to running one of the world's largest global monitoring platforms today, my work has debuted a myriad of new datasets, methodologies and scales to the study of how we understand our global world. A central theme of that work has been around how creative "reimagining" of information through the emerging world of massive computing power can offer powerful and unexpected new lenses onto the world around us and the incredible future that awaits as libraries transition from being museums of artifacts to becoming conveners of information and innovation[99] that empower a new era of access and understanding of our world.

About the author

One of Foreign Policy Magazine's Top 100 Global Thinkers of 2013, Kalev H. Leetaru is a Senior Fellow at the George Washington University Center for Cyber & Homeland Security and a member of its Counterterrorism and Intelligence Task Force, as well as a Council Member of the World Economic Forum's Global Agenda Council on the Future of Government. From 2013–2014 he was the Yahoo! Fellow in Residence of International Values, Communications Technology & the Global Internet at Georgetown University's Edmund A. Walsh School of Foreign Service, where he was also adjunct faculty. Kalev has been an invited speaker throughout the world, from the United Nations to the Library of Congress, Harvard to Stanford, Sydney to Singapore, while his work has appeared in the presses of more than 100 nations and from *Nature* to *The New York Times*. In 2011 *The Economist* selected his Culturomics 2.0

[92] http://blog.gdeltproject.org/introducing-the-global-content-analysis-measures-gcam/.
[93] http://blog.gdeltproject.org/gdelt-translingual-translating-the-planet/.
[94] http://blog.gdeltproject.org/announcing-the-gdelt-live-trends-dashboard/.
[95] http://foreignpolicy.com/2015/04/15/why-we-cant-just-read-english-newspapers-to-understand-terrorism-big-data/.
[96] https://isd.georgetown.edu/sites/isd/files/From%20Big%20Data%20to%20Global%20Diplomacy.pdf.
[97] http://www.internationalpeaceandconflict.org/profiles/blogs/a-force-for-good-how-digital-jedis-are-responding-to-the-nepal.
[98] http://www.chathamhouse.org/event/using-big-data-understand-global-conflict.
[99] http://www.knightfoundation.org/blogs/knightblog/2014/9/30/reimagining-libraries-conveners-information-and-innovation/.

study as one of just five science discoveries deemed the most significant developments of 2011, while the following year HPCWire awarded him the Editor's Choice Award for Edge HPC (High Performance Computing) "representing the highest level of honor and recognition given to the thought leaders in the HPC community" and in 2013 noted "his research helped usher in the era of petascale humanities", while later that year his work on Twitter was recognized by Harvard's Neiman Lab as the top social media study of 2013.