International Journal of Communication 8 (2014), Feature 2666–2678 1932–8036/2014FEA0002

Himalaya of Data

PELLE SNICKARS
Umeå University, Sweden

On January 17, 2007, the Wayback Machine's software crawler captured wikileaks.org for the first time. The crawler's act of harvesting and documenting the Web *meta-stored* a developing site for "untraceable mass document leaking"—all in the form of an "anonymous global avenue for disseminating documents," to quote the archived representational image of the site (Wayback Machine, 2007, para. 6). The initial WikiLeaks captures, and there were additional sweeps stored during the following months, vividly illustrate how WikiLeaks gradually developed into a site of global attention. The WikiLeaks logo, with its blue-green hourglass, was, for example, graphically present from the start, and later headings at the right were "news," "FAQ," "support," "press," and "links"—the latter directing users to various network services for anonymous data publication as i2P.net or Tor. Interestingly, links to the initial press coverage on Wikileaks are kept—which is not always the case at Wayback Machine—and can still be accessed. Apparently, one of the first online articles to mention what the site was about stated: "a new internet initiative called WikiLeaks seeks to promote good government and democratization by enabling anonymous disclosure and publication of confidential government records" (Wayback Machine, 2007, para. 18).

The Wayback Machine is an astonishing crawler software. Through its three-dimensional index, basically anything that has appeared online in the last couple of years can be made visible again. This particular search engine, in fact, serves as a correction to the general flatness of digital memory—even if some would argue that the Web means the end of forgetting. We are only beginning to grasp what it means that so much of what we say, think, and write in print and pixel is, in the end, transformed into permanent (and publicly distributed) digital files—whether leaked or not. Then again, all code is deep, and the Wayback Machine is, arguably, one of the more sophisticated digital methods to extract and visualize the specific historicity of the Web medium, even if it hasn't received the academic attention it deserves. Essentially, the Wayback Machine (run by the Internet Archive) stores superficial screen shots of various graphical user interfaces. This means that the Web cannot be surfed through its interface, because specific URLs are always needed. Still, more than 400 billion Web pages have been crawled since 1996—and wikileaks.org has been captured 1,822 times since 2007 (as of April 2014).

Looking at, reading, and thinking about the first stored captures of wikileaks.org through the Wayback Machine, one cannot help but notice how the site initially wanted to become a new sort of Wikipedia. WikiLeaks strived to "wikify" leaking by way of incorporating advanced cryptographic

technologies for anonymity and untraceability, all in the form of a wiki. Massive amounts of documents were to be combined with "the transparency and simplicity of a wiki interface," at least according to initial FAQs from early 2007. To users, WikiLeaks will "look very much like Wikipedia. Anybody can post to it, anybody can edit it. No technical knowledge is required. Leakers can post documents anonymously and untraceably." Furthermore, it was argued that all users can "publicly discuss documents and analyze their credibility and veracity." As a consequence, users of the site would have the ability to openly "discuss interpretations and context and collaboratively formulate collective publications" (Wayback Machine, 2007, paras. 1, 14).

As is well known, WikiLeaks did not become what it promised back in January 2007. Rather—to quote the site it wanted to resemble—WikiLeaks was originally launched as a user-editable wiki (hence its name), but has "progressively adopted a more traditional publication model and no longer accepts either user comments or edits" (Wikipedia, 2014, para. 5). What did not change, however, is the fact that WikiLeaks was (and is) a distinct archival phenomenon, aptly described as a database of scanned documents, all in the form of a giant information repository. It comes as no surprise that web captures of the site in February 2008—a little more than a year after WikiLeaks was launched—claimed a database of more than 1.2 million documents (Wayback Machine, 2008).

Taking its title from a quote in Geert Lovink's and Patrice Riemens' influential 10 (or 12) theses on WikiLeaks, this article situates WikiLeaks within a broader archival discourse on data distribution (Lovink & Riemens, 2010). What type of archive (or database) is WikiLeaks, and how does the site challenge traditional archives and libraries through new forms of massive information and data retrieval as well as user-oriented exploration? If public data can be found online by anyone at any time, what are the implications for, and the contemporary role of, archives and libraries? Naturally, the controversial nature of the leaked information from WikiLeaks is truly "hot data," which is hardly the case at most heritage institutions. Still, the way the site's massive amounts of freely distributed documents entered the cultural circulation of the digital domain in general, as well as more media-specific areas in particular, hints at new, emerging archival models where free access to hitherto locked material can generate innumerous forms of knowledge (of the past and sometimes even the future), which, after all, is the purpose of most memory institutions.

Hence, the importance of WikiLeaks as a sort of new archival modality. This article, then, bears on an ongoing discussion of how the digital is changing our understanding of what the essential building blocks of archives and libraries are made oftoday within the so-called memory sector. As binary information, data can not only be copied back and forth (and tracked ad nauseam), the very relation between notions as documents, archival records, data, and information—not to mention knowledge—has become fluid and extremely complicated to pin down. This article will, however, not dwell theoretically upon the matter, but rather use the terms in a more general and culturally framed manner. Within the field of library and information science, for example, a lot of research deals with similar issues, and document theory has in many ways seen a revival due to the digital (Lund, 2010).

Data mining—the process of extracting hidden patterns from huge amounts of data—and topic modeling—algorithms that uncover hidden thematic structure in document collections and thus develop new ways to search, browse, and summarize large archives of texts—are singled out as new computational methods to move between these notions. WikiLeaks documents have, for example, been used in various data mining contexts and have, in general, become an increasingly important tool to transform data into information. Data mining is, in short, the process of using computation power to retrieve new techniques for knowledge discovery, and the same goes for topic modeling. There are many nuances to these processes, but roughly three steps are involved in data mining: One has to first preprocess raw data; second, mine the data; and, third, interpret the results. Machines can do most of the work, but one (or two) human subjects are always needed.

WikiLeaks is a real data mine. The organization has, from the start, been about storage and distributed accessibility—and, in that sense, it does echo Wikipedia. No strings were ever attached between the two organizations, however. Still, the article on WikiLeaks at Wikipedia can be seen as an illustrative case in point of linked relations. Begun approximately at the same time as the Wayback Machine captured wikileaks.org for the first time, some 5,100 changes and revisions have been made on this particular piece of text. The article is, in fact, a popular one on Wikipedia and was (at least for a while) regularly ranked among the top 10 in terms of traffic (on en.wikipedia.org), and it is visited each month on average by a quarter of a million users. The article, initially, states that the "wikileaks.org domain name was registered on 4 October 2006" and that it published its first "document in December 2006. . . . The creators of WikiLeaks have not been formally identified . . . [but] it has been represented in public since January 2007 by Julian Assange" (Wikipedia, 2014, para. 4). These very sentences in the article have been edited and altered many times—and all changes have been kept by the Wikipedia version tracker (with quite a few edits done by nonhuman, automatic bots). The article "WikiLeaks" on Wikipedia might, hence, be understood and perceived as an archive of an ongoing conversation about how users have understood what (*the* database) WikiLeaks was, is, and has been all about. On the one hand, the article "WikiLeaks" can be seen as a framework for understanding how knowledge came to be—and (often) be (mis)understood at Wikipedia—and, on the other hand, the article (or site) also functions as an archive preserving data and information on the very same discourse.

Documents as Data

More data is better data—so they say. WikiLeaks is not Google, but they both operate within the same digital domain and according to a similar computational and numeric logic of data distribution. The so-called Cablegate, with 250,000 leaked U.S. embassy cables during late autumn 2010, for example—described by WikiLeaks as "the largest set of confidential documents ever to be released into the public domain" (Wikileaks, 2011, para. 1)—hints at an emerging and occasionally ubiquitous computational culture, spearheaded by Google's vision of distributing data, but which WikiLeaks (as well as other major digital archives) also currently form an important part of. Given the sheer size of contemporary online database collections—from the vast information repositories of data at WikiLeaks and shared files (or,

actually, torrents) at The Pirate Bay, to billions of pieces of user-generated content on YouTube and Flickr, or, for that matter, the 20 million digitized heritage objects at the Library of Congress—not to mention the data harvesting operations orchestrated by the National Security Agency (NSA) (and revealed in 2013 by Edward Snowden)—simply having a look at what is inside such vast databases or digital archives is no longer possible. "Digital archives can house so much data that it becomes impossible for scholars to evaluate the archive manually, and organizing such data becomes a paramount challenge," as some humanities–computer science researchers have stated (Simeone, Guiliano, Kooper, & Bajcsy, 2011, para. 1).

Indeed, as a massive provider of data, WikiLeaks has acted as a transformative symbol of the digital information society at large, hinting at the *data avalanche* currently overwhelming us all. Everything that can be digitized will be digitized, the catchphrase went during the 1990s, and we are now increasingly experiencing what such a claim implies. The quantitative turn of information overload is becoming an unavoidable fact of contemporary life, with nothing hotter than analyzing big social data—at least for dot-com enterprises—or, as *The New York Times* and other major newspapers have repeatedly reported in numerous articles: Despite concerns of a global economic slowdown, companies that construct and operate data centers that run the Internet and store vast amounts of corporate and government data are expecting to grow even more. In short, data is nothing less than the new raw material of the information economy, even if online players are just beginning to learn how to use and process it. In relation to WikiLeaks, Lovink and Riemens (2010) have, as a consequence, argued, "The glut of disclosable information can only be expected to continue grow—and exponentially so." (para. 17)

At the same time, the NSA surveillance has also revealed an urgent need to think twice around data production and consumption, not the least on a personal level. In an article in the German *Frankfurter Allgemeine Zeitung* during the summer of 2013, Evgeny Morozov, for example warned against an increased "information consumerism," that even laws and regulations won't stop:

> European politicians can try imposing whatever laws they want but as long as the consumerist spirit runs supreme and people have no clear ethical explanation as to why they shouldn't benefit from trading off their data, the problem would persist. NSA surveillance, Big Brother, Prism: all of this is important stuff. But it's as important to focus on the bigger picture—and in that bigger picture, what must be subjected to scrutiny is information consumerism itself—and not just the parts of the military-industrial complex responsible for surveillance. As long as we have no good explanation as to why a piece of data shouldn't be on the market, we should forget about protecting it from the NSA, for, even with tighter regulation, intelligence agencies would simply buy—on the open market—what today they secretly get from programs like Prism. (para. 30)

The NSA Prism surveillance program as well as WikiLeaks are, of course, digital phenomena unthinkable without the complex media ecosystem created by advanced computer networks and technologies. Still, in terms of increased data, the contemporary flood of information is, by no means, new. On the contrary, libraries and archives, for example, have during the last century repeatedly complained of far too many books and even more documents and records. The major difference today is that in digitized form such material can be analyzed by powerful computers, and even scrutinized collectively as major cultural data sets (rather than on a singular basis only), and occasionally using the "wisdom of the crowd." The notion of a particular search, then, is arguably not the answer to the *infinite* digital archive.

Then again, traditional archives are often thought of as collections of historical records that have accumulated over the years. Archives store information—on shelves or in stacks—and most material will, in fact, never be used by anyone. General estimations done in the 1990s by the International Federation of Film Archives, for example, stated that nine out of 10 film reels kept in various international film archives will never be viewed. Preservation and access (in that order) are seen as the basic principles behind the nature of archives. The library sector works according to similar principles, even if a major difference exists between public libraries and, say, national libraries, where the latter need to follow the law and (often according to a legal deposit) collect and keep everything that is published.

Within the digital domain, archives and libraries (as well as museums) are often seen as belonging to a heritage sector—with more similarities than differences. Still, these institutions do follow different logics. An archive, for example, regularly sorts documents and records in an organized manner. A complete set of material is never kept—a principle different from a legal deposit, where all material should be preserved according to law. Since national libraries (at least) keep everything, library catalogues tend to be detailed and based on singular items (a book, a film, a photograph, etc.). Archives, on the contrary, often organize material on a more collective basis in broader categories. As a consequence, archival collections are generally broader and more unsorted than library ones.

The general transition from analogue to digital production and distribution of books, media material, documents, and various archival records during the last 20 years has, however, challenged such traditional conceptions. The memory sector is currently involved in dramatic changes, fundamentally altering the basis and conceptions of what heritage institutions in essence should devote themselves to. Challenges regularly come from the digital outside, constantly fueled by everything from new user interfaces, apps, and application programming interfaces to radical accessibility on various Web n.0 platforms such as YouTube, Flickr, Twitter, Snapchat, Pinterest—or Wikipedia Commons—with the subsequent implosion of the legal deposit law (since there are no gatekeepers within the digital domain)—not to mention the use of peer-to-peer (P2P) file-sharing technology for long-term digital preservation. Hence, if cloud-based storage solutions and streaming alternatives have during the last years altered media and tech industries, WikiLeaks' massive data distribution can be regarded as yet another digital resource (or platform) affecting the ways archives (and libraries) conceive of themselves.

As a user-generated archive—and it is worth recalling that it was actual people who first scanned and uploaded all secret documents—WikiLeaks has distributed these unsorted chunks of documents as major data sets. However, Wikileaks has also been organized according to a logic contrary to traditional archives, with access, distribution and (semi-)openness acting as its guiding principles. Broadly speaking, WikiLeaks has accentuated a contemporary trend of not only unlocking various archival holdings—not unlike initiatives associated with *open data*, making them widely accessible in digital form—but also detaching the archival record (or document) from its traditional place and location.

If archives and libraries for centuries were physical spaces where static documents and records were kept—sometimes with the task of being stored for eternity—WikiLeaks seems to suggest a rapid transition toward the archive as a distributed data stream. New digital archives are always dynamic by nature; essentially, they are made of copies of copies that need to migrate from one format to another in order to be preserved. Yet storing and safekeeping such digital documents and records become difficult in an age of instant reproducibility and dissemination. During Cablegate for example, WikiLeaks lost its support of many U.S. partners (including host Amazon.com), but all confidential information remained available online through mirror sites and torrent peer-to-peer sharing programs (as a so-called insurance.aes256 file).

WikiLeaks has, over the years, used many hosting services, with quite a few of them being Swedish, and during Cablegate the organization posted a message online that stated:

> We've decided to make sure everyone can reach our content. As part of this process we're releasing archived copy of all files we ever released—that's almost 20,000 files. The archive linked . . . contains a torrent generated for each file and each directory. (Wikileaks, 2010, para. 21)

Hence, as with (illegal) file sharing, once information (or content) has been uploaded and distributed online, there is literally no way to manage or master the data. The reason is as simple as it is technologically complex, and WikiLeaks servers are also constantly "migrating throughout the globe" (if one is to believe their own description). Sharing data through P2P protocols, in essence, means data is dispersed in such a way that its coded nature renders it impossible to control.

One needs to remember, however, that most documents released by WikiLeaks have been scanned Xerox copies of printed material—that is, digitized content. Leaving aside the prevalent discussion on the importance of materiality in relation to the digital—that is, leaked physical documents gone virtual—there remains an important distinction between the natively digital and the digitized; between digital objects and content born in new media online, as opposed to, for example, scanned documents that have migrated into the digital domain (as the leaked documents). Based on code, the former material can be analyzed in myriad ways, whereas the latter often takes the form of a representational image file, as is

the case with most WikiLeaks documents. In short, *digitized* material is not digital. Still, it goes without saying that politics are involved in any representation of data. There are also inherent and implicit structures in digital data, especially when detached from the realm of computer code.

As a widely disseminated archive, WikiLeaks can be understood and addressed as the flip side of digital openness and transparency—indeed, dark for some, especially the U.S. State Department—accentuating how computers have become crucial for coding (and decoding) contemporary information culture. Data can easily be shared and effortlessly multiplied, but computers and their programs need to be used for decoding the exponentially increased information. No one human can actually read the more than 90,000 leaked documents related to the war in Afghanistan; they can only be searched by a computer (or a network of them). As a dispersed computational archive, distributed documents through WikiLeaks are in many ways launched and uploaded into an information circuit, where the context of data content quickly becomes fleeting and arbitrary, and material is detached from its place of origin. The embassy cables, for instance, which date from 1966 and contain confidential communications between 274 embassies and the U.S. State Department, were so many and heterogeneous—apparently comprising 261,276,536 words—that WikiLeaks made a graphic of the Cablegate data set as well as gave tips on how to explore the data.

One of the lessons to be learned from WikiLeaks with regard to the heritage sector is, therefore, to place a structure of stability over the archival document or record, in what seems to be an endless flow of infinite possibilities within the digital domain. The digital (and sometimes the digitized) object can always be enhanced with new layers of protocol or code, and potential meanings and context can easily increase exponentially. Thus, some resources will require different modes and more archival stabilization than others. New archival strategies can, of course, also use the technology at hand—and, in a sense, "follow the medium." As the project LOCKSS (Lots of Copies Keep Stuff Safe) indicates, for example, distributing a set of documents over a P2P file-sharing network is a smart way to preserve material, documents, and records *through* digital technology rather than being hindered, put back, and interfered with by new IT. In fact, the LOCKSS technology is an open-source, P2P, decentralized digital preservation infrastructure with some resemblances to WikiLeaks.

Exploring Data

In late August 2010, a group of *hackademics* started working on the more than 90,000 WikiLeaks documents known as the Afghanistan War Logs, trying to produce a video visualization of the leaked data. These documents naturally contain a vast amount of information, but they were basically used to track events in Afghanistan, including deaths, civilian injuries, and friendly fire over the course of six years. The result was a graphically simple but mesmerizing video—described as follows by one if its producers, Mike Dewar:

> This is a visualization of activity in Afghanistan from 2004 to 2009 based on the WikiLeaks data set. Here we're thinking of activity as the number of events logged in a small region of the map over a one month window. These events consist of all the different types of activity going on in Afghanistan.
>
> The intensity of the heatmap represents the number of events logged. The color range is from 0 to 60+ events over a one month window. We cap the color range at 60 events so that low intensity activity involving just a handful of events can be seen—in lots of cases there are many more than 60 events in one particular region. The heat map is constructed for every day in the period from 2004-2009, and the movie runs at 10 days per second (Dewar, 2010, paras. 1–2).

Even if the general media debate around these war logs during the summer of 2010 was centered on missing aspects of the Afghanistan war in the leaked documents—as *The New York Times* firmly stated: "The archive is clearly an incomplete record of the war. It is missing many references to seminal events and does not include more highly classified information" (Chivers et al., 2010, para. 19)—what the video visualization vividly illustrated is how surges of activity grew drastically as the war progressed. The Afghanistan map literally becomes increasingly (blood)red.

WikiLeaks has, thus, not only been about providing a platform for whistleblowers and disseminating secret documents. Its distributed data has also been packaged (and framed) toward maximum usage and media attention. Naturally, Dewar and his programmer colleagues released the code to generate the Afghanistan video as *open source*, inviting others to continue working on it. At wikileaks.org, tips are frequent on how to explore the data. In addition, the *Collateral Murder* video from April 2010 suggests that edited usage (and reusage) form important parts of the WikiLeaks concept. Perceived as a potential archive, WikiLeaks is likened to a real archive that, through crowd sourcing—and especially its professional press partners—can be dissected and analyzed, filtered and sorted into something more akin to a more usable document. What WikiLeaks' partnering media organizations essentially did at the peak of media attention around Wikileaks a few years ago was to break down the leaked information into smaller pieces. Regarding, Cablegate, for instance, each cable "is essentially very structured data," as *The Guardian* aptly put it on its data blog in late 2010. The leaked information featured distinct categories as, for example, source, recipients, subject field, and tags. Each cable was also tagged with a number of keyword abbreviations, and *The Guardian* even put together a downloadable Google glossary spreadsheet of the most important keywords (Rogers, 2010).

During 2010 and 2011, WikiLeaks and its media partners in many ways acted as archival organizations providing massive amounts of data to be used in various ways. It repeatedly invited users to work with and explore the distributed data—and, in that sense, the wiki culture once evoked by WikiLeaks became prevalent. Take our data, mash it up, and create great visualizations, the Guardian Datastore on Flickr, for example, declared at the time. Hundreds of visualizations were produced—a vivid illustration of

the cultural circulation and reuse of the leaked information—in addition to the often neglected WikiLeaks relation to a wider discourse on user-generated content and the Web n.0 phenomenon. On the Guardian Datastore at Flickr, David Placr, for example, produced a number of haunting visualizations on deaths in Bagdad, based on the distributed data from WikiLeaks. City maps ranged from "total deaths as a result of the War in Iraq," with red circles spread all over this troubled city, to "'enemy' deaths as a result of the War in Iraq shown as a red circle, compared to total deaths (the larger clear circle)" (Placr, 2011, para. 1).

In fact, WikiLeaks was never strictly devoted to distribute raw data only, even if it preferred to perceive itself (in the media) as an organization that acted as a neutral provider of classified information.

> One of the main difficulties with explaining WikiLeaks arises from the fact that it is unclear (also to the WikiLeaks people themselves) whether it sees itself and operates as a content provider or as a simple conduit for leaked data (the impression is that it sees itself as either/or, depending on context and circumstances) (Lovink & Riemens, 2010, para. 5).

Packages of selected content remain an integral and important part of WikiLeaks, most obviously in regard to the media organizations (*Le Monde, El Pais, The Guardian*, *The New York Times,* and *Der Spiegel*) that WikiLeaks cooperated with. Whether those newspapers (and adjacent media outlets) can be regarded as WikiLeaks' media partners remains an open question, however. Some, such as *The New York Times,* have rejected this being the case—and consequently only published a few hundred documents. But *The Guardian*, for example, promoted its data blog as a direct consequence of the partnership.

WikiLeaks received a fair amount of criticism regarding these media partnerships—especially from activist circles. As a consequence, an FAQ online featured the rhetorical question about why WikiLeaks chose "established 'old media'" as media partners. The FAQ states the following (with a rather strange phraseology):

> Wikileaks makes to a promise to its sources: that will obtain the maximum possible impact for their release. Doing this requires journalists and researchers to spend extensive periods of time scrutinising the material. The established partners chosen were among the few with the resources necessary to spend many weeks ahead of publication making a start on their analysis (Wikileaks, 2014, para. 12).

An illustrative case at the time was *The Guardian*'s data blog—related to its Flickr initiative—which basically was an interactive guide to the WikiLeaks embassy cables. The exhortation to users was to "download the key data and see how it breaks down." According to the newspaper, the information released had "produced a lot of stories but does it produce any useful data? We explain what it includes" (Rogers, 2010, para. 1). Plenty of infographics were presented—ranging from a world map with top locations where the cables were uploaded to a storyline of cables sent in the weeks around 9/11. In

addition, several data packages could be downloaded and presented directly using various Google services (as Docs and Fusion Tables).

Most interestingly, however, was that *The Guardian* explained in detail what the leaked data included, and what technology was used with a full description of the various "layers of data." The cables themselves came "via the huge Secret Internet Protocol Router Network, or SIPRNet," a worldwide U.S. military Internet system, apparently "kept separate from the ordinary civilian internet and run by the Department of Defense in Washington. Since the attacks of September 2001, there has been a move in the US to link up archives of government information," in the hope, according to *The Guardian*, that key intelligence would no longer get trapped. Over the past decade, an increasing number of U.S. embassies were linked to the SIPRNet sharing military and diplomatic information. "An embassy dispatch marked SIPDIS is automatically downloaded on to its embassy classified website," *The Guardian* stated. And from there, they could be "accessed not only by anyone in the state department, but also by anyone in the US military who has a security clearance up to the 'Secret' level, a password, and a computer connected to SIPRNet," which covered more than 3 million people (Rogers, 2010, para. 4). In other words, (too) many people had access. That someone would act as a potential leak was, thus, bound to happen in an age of digitally instant reproducibility.

Conclusion

The Guardian is still a major newspaper with a huge staff—not the least in relation to an anemic heritage sector. Compared to the traditional archival sector, which also increasingly deals with large cultural data sets, WikiLeaks' insistence on exploring the leaked data was—and still is—different. Few memory institutions today invite users to download, visualize, and work with digitized data in the way WikiLeaks and its media partners have, even if research initiatives like Cultural Analytics or the grant request Digging into Data points in this direction, as does the field of digital humanities in a broader sense.

Heritage users are often scholars, and, given the conservative culture of scholarship in general, and humanistic research in particular, this is not surprising. Still, since heritage institutions devote a lot of energy into digitizing their collections, and given the increasing role of computerized technology in humanistic research in general (whether it wants it or not), the issue does remain somewhat puzzling. If the computer is the cultural machine of our age, then the same goes for archives, libraries, and their potential users. Exceptions can, of course, be found. The field of digital humanities is rapidly picking up speed—often closely linked to the cultural heritage sector—and the discursive idea of the lone scholar working in isolation with his or her own archiving solutions will likely (at least in due time) fade away.

Massive amounts of leaked data suggest other archival methods and practices than traditional extraction of minuscule data from archives, gleaned bit by bit. As the report *Our Cultural Commonwealth* stated in 2006, humanistic researchers and users of

> massive aggregations of text, image, video, sound, and metadata will want tools that support and enable discovery, visualization, and analysis of patterns; tools that facilitate collaboration; an infrastructure for authorship that supports remixing, recontextualization, and commentary—in sum, *tools that turn access into insight and interpretation* (Welshons, 2006, p. 16, emphasis in original).

From an archival perspective, WikiLeaks can be regarded as a prototype for this kind of development. New productive ways to explore data are, thus, one experience that can be drawn from the site. Data-literate scholars and experts in statistical methods and data analysis technologies are still hard to find within the heritage sector. But sites such as WikiLeaks, and the way data is being handled and transformed, explored and analyzed as a consequence of distributive strategies online, seem to suggest an increased need for such personnel.

The issue thus taps in and relates to an emerging scholarly trend. *The New York Times* has, for example, run a series of articles on how technology is changing the humanistic and academic landscape. According to one of the texts, members of new generation of "digitally savvy humanists" do not look for inspiration anymore in the next "political or philosophical 'ism'"—but rather look toward data to explore how digital technology as an accelerating force is changing the overall understanding of the liberal arts. New methodologies, powerful technologies, vast amounts of data, and stored digitized materials "that previous humanities scholars did not have" act as a revisionist call of what human research is all about (Cohen, 2010, para. 3). WikiLeaks can be seen as forerunner of and predecessor to the current transformation. Coming to terms with WikiLeaks is, in fact, a task as demanding as it is provocative (at least for some)—or as Lovink and Riemens (2010) have stated: "to organize and interpret this Himalaya of data is a collective challenge" (para. 17).

References

Chivers, C. J., Gall, C., Lehren, A. W., Mazzetti, M., Perlez, J., & Schmitt, E. (2010, July 25). View is bleaker than official portrayal of war in Afghanistan. *The New York Times*. Retrieved from http://www.nytimes.com/2010/07/26/world/asia/26warlogs.html?pagewanted=all

Cohen, P. (2010, November 16). Digital keys for unlocking the humanities' riches. *The New York Times*. Retrieved from http://www.nytimes.com/2010/11/17/arts/17digital.html?pagewanted=all&_r=0

Dewar, M. (2010, August 16). Visualisation of activity in Afghanistan using the WikiLeaks data. Vimeo clip at http://vimeo.com/14200191

Lovink, G., & Riemens, P. (2010). Twelve theses on Wikileaks. Retrieved from http://www.eurozine.com/articles/2010-12-07-lovinkriemens-en.html

Lund, W. N. (2010). Document, text and medium: Concepts, theories and disciplines. *Journal of Documentation,* 5. doi:10.1108/00220411011066817

Morozov, E. (2013, July 27). Information consumerism. The price of hypocrisy. *Frankfurter Allgemeine Zeitung*. Retrieved from http://www.faz.net/aktuell/feuilleton/debatten/ueberwachung/information-consumerism-the-price-of-hypocrisy-12292374.html

Placr, D. (2011). Photostream on Flickr. Retrieved from http://www.flickr.com/photos/55213715@N04/5122416361/in/photostream

Rogers, S. (2010, December 3). WikiLeaks embassy cables: Download the key data and see how it breaks down. *The Guardian*. Retrieved from http://www.guardian.co.uk/news/datablog/2010/nov/29/wikileaks-cables-data#

Simeone, M., Guiliano, J., Kooper, R., & Bajcsy, P. (2011). Digging into data using new collaborative infrastructures supporting humanities-based computer science research. *First Monday,* 5. Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3372/2950

Wayback Machine. (2007). Wikileaks. Retrieved from http://web.archive.org/web/20070501000000*/http://wikileaks.org

Wayback Machine. (2008). Wikileaks. Retrieved from http://web.archive.org/web/20080216000537/http://www.wikileaks.org/wiki/Wikileaks:About

Welshons, M., (2006). *Our cultural commonwealth.* American Council of Learned Societies. Retrieved from http://www.acls.org/cyberinfrastructure/ourculturalcommonwealth.pdf

WikiLeaks. (2010, November 28). All released leaks archived. Retrieved from http://www.wikileaks.org/file/wikileaks_archive.7z

WikiLeaks. (2011, February, 10). Secret U.S. embassy cables. Retrieved from http://wikileaks.org/cablegate.html

WikiLeaks. (2014). FAQ. Retrieved from http://www.wikileaks.org/static/html/faq.html

Wikipedia. (2014). Wikileaks. Retrieved from http://en.wikipedia.org/wiki/Wikileaks