Himalaya of Data

Pelle Snickars

The Wayback Machine is truly an incredible piece of 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 newness and 'flatness' of digital culture—even if some would indeed argue that the web means the end of forgetting. All likely, 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. Essentially, the Wayback Machine (run by the Internet Archive) stores screen shots of various GUIs. This means that the web cannot be surfed through its interface, rather specific URLs are always needed. Still, some 150 billion web pages have been crawled since 1996. In fact, archived versions of web pages across time and space appear through the Wayback Machine's digital time capsule almost akin to magic.

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, hence, 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 image of the site. The initial WikiLeaks captures in the beginning of 2007, and there were additional sweeps stored during the following months, vividly illustrates how WikiLeaks gradually developed into a site of almost unprecedented global media attention. The WikiLeaks logo, with it's blue-green hourglass, was, for example, graphically present right from the start, with subsequent headings to the right as '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 is kept (and can still be accessed). Apparently, one of the first online article's to mention what the site was all 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."

Looking and clicking 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 Wikipedia. In short, 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. 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."[1]

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 moved towards a more traditional publication model and no longer accepts either user comments or edits."[2] What did not change, however, is the fact that WikiLeaks was (and is) a distinct archival phenomenon, more or less aptly described as a database of scanned documents, forming 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.[3]

Taking its title from a quote in Geert Lovink's and Oatrice Riemens' influential ten (or twelve) theses on WikiLeaks,[4] this article, then, tries to situate WikiLeaks within a broader archival discourse on data distribution. 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 (more or less) public data can be found online by anyone at all times, what are the implications for, and the contemporary role of archives and libraries (understood in a broad sense)? 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 have entered the cultural circulation of the digital domain in general, as well as more media specific and web 2.0 areas in particular, does hint at various emerging archival models, where free access to hitherto locked material can generate innumerous forms of

new 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 sort of a new archival modality.

This article, then, bears on an ongoing discussion how 'the digital' is changing our understanding what the essential building blocks of archives and libraries are made of today within the so called 'memory sector'. As binary information data can not only be copied back an forth (and tracked *ad nauseum*) 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 dwelve theoretically upon the matter, but rather use respective term in a more general and (perhaps) culturally framed manner. Within the field of Library and Information Science, for example, a lot of research is dealing with similar issues, and 'document theory' has in many ways seen a revival due to 'the digital'.[5] More often than not, 'data mining'—the process of extracting (more or less) hidden patterns from huge amounts of data—is singled out as a computational method with bearing on all these notions. WikiLeaks documents have, for example, been used in various data mining contexts', and has, 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. There are many nuances to this process, but roughly the steps are three: firstly, one has to pre-process raw data, secondly 'mine' the data, and finally interpret the results. Machines can do most of the work, but one (or two) human subjects are often needed.

WikiLeaks, for sure, is a real data mine. The 'organization' has right from the start been all about storage, distributed accessibility and exploration—and it that sense it, actually, does echo Wikipedia. No strings were ever attached between the two 'organizations', however, yet the article 'WikiLeaks' at Wikipedia can, actually, 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 4,000 changes and revisions have up until now been done on this particular piece of text. The article is, in fact, one of the most popular on Wikipedia, regularly ranked as top ten in terms of traffic (on en.wikipedia.org), and 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".[6] These very sentences in the article have been edited, and altered back and forth many, many times—and all changes have, naturally, 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 how users have understood what (*the* database) WikiLeaks was, is, and has been all about. On the one hand, the article 'WikiLeaks' can, thus, 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—or 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 US 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"[7]—hints at an emerging, and occasionally ubiquous 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 UGC on YouTube and Flickr, or for that matter the 20 million digitized heritage objects at the Library of Congress—simply having a look what's 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.[8]

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 catch phrase once went during the 1990s, and we are now increasingly experiencing what such a claim actually implies. The quantitative turn of information overload is becoming

a unavoidable fact of contemporary life, with nothing more important than analyzing 'big social data'—at least for dotcom enterprises—or as the *The New York Times* recently reported: 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 expect growth next year [2012] to match levels last seen in the world economy's boom years: about 19 percent."[9] 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 have, as a consequence, argued that one "can only expect the glut of disclosable information to grow further—and exponentially so."

WikiLeaks is, indeed, a novel digital phenomenon, unthinkable without the complex media ecosystem created by advanced computer networks and related 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 over way 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), occasionally using the 'wisdom of the crowd' online. The notion of a particular 'search', then, is arguably not the answer to the more or less *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 never be used by anyone. Estimations done in the 1990s by the film archive organization FIAF, for example, then stated that approximately 95 percent of film reels kept in various international film archives will never be looked at. Preservation and access (in that order) are, in short, seen as the basic principles behind the nature of archives. The library sector works according to similar principles, even if a difference exist between public libraries and, say, national libraries, were the latter need to follow the law and (often according to a legal deposit) collect and keep everything that has been 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 altering logics; an archive, for example, regularly sorts out documents and records in an organized manner. A complete set of material is never kept, a principle very 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 etcetera). Archives, on the contrary, often organize material on a more collective basis in wider 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 15 years has, however, naturally challenged such traditional conceptions. The memory sector is currently involved in dramatic changes, fundamentally altering the basis and conceptions of what heritage institutions should devote themselves to. Challenges regularly comes from the digital domain, constantly fueled by everything from new UIs, apps, and APIs, to radical accessibility on various *web n.0* platforms (to use the notion of Peter Lunenfeld) as YouTube, Flickr and Wikipedia Commons—with the subsequent implosion of the legal deposit law (since there are no gatekeepers)—not to mention the usage of P2P file sharing technology for long term digital preservation. Hence, if cloud based storage solutions are the latest tech fad altering the publishing and media industries, WikiLeaks massive data distribution, can be regarded as yet another digital challenge affecting the ways archives (and libraries) conceive of themselves.

As a user generated archive—it is various individuals who in the first place have scanned and uploaded all secret documents—WikiLeaks have on the one hand distributed these unsorted chunks of documents as major data sets, but on the other, also been organized according to a contrary logic than traditional archives, with access, distribution and (semi-)openness as guiding principles. Broadly speaking, WikiLeaks has accentuated a contemporary trend of not only unlocking various archival holdings, making them widely accessible in digital form, but at the same time also detaching 'the archival record' (or document) from its traditional place and location. If archives and libraries for centuries where 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 towards 'the archive' as a distributed data stream.

In short, new digital archives are always dynamic by nature; essentially they are made of copies of copies that need to migrate from one format to the other in order to be preserved. Yet, it goes without saying that 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 US

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). Naturally, WikiLeaks has over the years used many hosting services (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."[10] Hence, like with (illegal) file sharing, once information (or content) has been uploaded and distributed online, there is literally no way trying to manage and master 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 self description). Sharing data through P2P protocols, in essence, means data is dispersed in such a way that its coded nature makes it (more or less) impossible to control.

One needs to remember, however, that almost all 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'—i. e. 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. Based on code, the former material can be analyzed in myriad of ways, whereas the latter often takes the form of a representational image file (as 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. Suffice to say, 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, hence, be understood and addressed as the flip side of 'digital' openness and transparency—indeed, dark for some, especially the US State Department—accentuating how computers have become crucial for coding (and decoding) contemporary information culture. As binary code, data can easily be shared and effortlessly multiplied, still computers and their programs also often needs to be used for decoding the exponentially increased information; i. e. no one human can actually read, say, the more than 90,000 leaked documents related to the war in Afghanistan, but they can be searched by a computer (or a network of them).

As a dispersed computational archive, distributed documents through WikiLeaks are in many ways uploaded into an information circuit, where the context of data content quickly becomes fleeting and arbitrary, and material more or less detached from its place of origin. The 'embassy cables', for instance, which date from 1966 and contain confidential communications between 274 embassies and the US State Department, are so many and heterogeneous (apparently comprising 261,276,536 words) that WikiLeaks has made a graphic of the 'Cablegate dataset', as well as giving tips on how to explore the data.

One of the lessons to be learned from WikiLeaks with regards 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 at an exponential rate. Thus, some resources will require different modes and more archival stabilization than others. Still, 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 also a smart way to preserve material, documents and records *through* digital technology rather than being hindered, put back and interfered by new IT. In fact, the LOCKSS technology is an open source, peer-to-peer, decentralized digital preservation infrastructure with a lot of 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 numerous information, but 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, still absolutely mesmerizing video—described by one if its producers, Mike Dewar, as follows:

> 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.[11]

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 wikileaked documents—as the *New Tork 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"[12]—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.

It remains important to stress that WikiLeaks has not only been about providing a platform for whistleblowers and disseminating secret documents, its distributed data has also been packaged (and framed) towards 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, and at wikileaks.org tips are frequent on 'how to explore the data'. In addition, the 'Collateral Murder' video from April 2010 suggest that edited usage (and re-usage) form important parts of the WikiLeaks concept.

Perceived as sort of an 'archive' WikiLeaks, hence, by and large likens an archive that potentially, through crowd sourcing—and especially its professional press partners—can be dissected and analyzed, filtered and sorted into something more akin to a usable 'document'. What WikiLeaks partnering media organizations have essentially done to the leaked information is breaking the data down in smaller pieces. Regarding, Cablegate, for instance, each cable "is essentially very structured data", as *The Guardian* aptly puts on their data blog. The leaked information features distinct categories as, for example, 'source', 'recipients', 'subject field' and 'tags'. "Each cable

was tagged with a number of keyword abbreviations", the paper informs, and *The Guardian* has put together a downloadable Google glossary spreadsheet of the most important keywords.[13] As a provider of massive amount of data, WikiLeaks and its media partner acting as kind of 'archival organizations', have as a consequence repeatedly invited users to work with and explore the distributed data—and in that sense the 'wiki' culture once evoked by WikiLeaks is still prevalent. "Take our data, mash it up and create great visualizations", the Guardian Datastore on Flickr declares. At present it includes hundreds of visualizations, a vivid illustration of the cultural circulation and re-use of the leaked information—not to mention the more important observation of WikiLeaks close relation to a wider discourse on user-generated content and the *web n.0* phenomenon. On the Guardian Datastore at Flickr, a certain David Placr has, for example, produced a number of haunting visualizations on deaths in Bagdad, based on the distributed data from WikiLeaks. City maps range from "total deaths as a result of the War in Iraq", with major red circles spread all over this trouble 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)."[14]

There are, in fact, innumerous examples which stress that WikiLeaks was never strictly devoted to distributing raw data only, even if it has preferred to perceive itself as an 'organization' that acts 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)", as Lovink and Riemens have poignantly remarked.[15] Packages of selected content do remain an integral and important part of WikiLeaks, which of course, is most obvious in regards to the media organizations (*Le Monde, El Pais, The Guardian*, *The New York Times* and *Der Spiegel*) that WikiLeaks have co-operated with. Whether those newspapers (and adjacent media outlets) can be regarded as WikiLekas 'media partners' remains an open questions, however. Some, as *The New York Times* have rejected this being the case (and only published a few hundred documents), while *The Guardian*, for example, has promoted its eloquent data blog more or less as a direct consequence of the partnership.

Naturally, WikiLeaks have also gotten their fair amount of criticism regarding these media partnerships, not the least from activist circles. As a consequence, an FAQ online features the rhetorical question why WikiLeaks has chosen "established 'old media' as

your initial media partners for the release? 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 scrutinizing the material." According to WikiLeaks, the established media partners simply have the resources "necessary to spend many weeks ahead of publication making a start on their analysis."[16] An illustrative case is *The Guardian's* data blog (related to its Flickr initiative), basically an interactive guide to the WikiLeaks 'embassy cables', with the exhortation to users to "download the key data and see how it breaks down." According to the newspaper, the information released has "produced a lot of stories but does it produce any useful data? We explain what it includes." Plenty of infographics are, hence, present—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, 2001. In addition, a number of data packages can be downloaded and presented directly using various Google services (as docs and fusion tables).

Most interestingly, however, is that *The Guardian* in an informative passage actually does explain what the leaked data includes, with a full description of the various "layers of data." The cables themselves come "via the huge Secret Internet Protocol Router Network, or SIPRNet", a worldwide US 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 will no longer get trapped. Over the past decade, an increasing number of US embassies have been 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* states. And from there, it can 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 covers more than three million people. In other words, (too) many people had access, and that someone would act as potential 'leaks' was more or less bound to happen in an age of digitally instant reproducibility.[17]

Conclusion

The Guardian is 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 is, however, quite different. Few memory institutions today invite users to download, visualize and work with digitized data the way WikiLeaks and its media 'partners' have, even if research initiatives like Cultural Analytics or the grant request Digginig into Data are steps taken in this direction. 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 is devoting a lot of energy into digitizing their collections, and given the increasing role that computerized technology plays for (humanistic) research in general (whether it wants it or not), the issue does remains 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, for example, 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 all likely (at least in due time) fade away. Massive amounts of leaked data simply suggests other archival methods and practices than traditional extraction of miniscule data from archives, gleaned bit by bit. As the report, *Our Cultural Commonwealth* stated already 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*."[18]

From an archival perspective WikiLeaks can, thus, be regarded as a prototype for this kind of development. New productive ways to explore data is 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 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 also taps in, and relates to an emerging scholarly trend. *The New York Times* has, for example, during the last year run a series of articles on how

technology is changing the humanistic 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'"—rather they look towards 'data', all in an effort to explore how digital technology as an accelerating force is changing the overall understanding of the liberal arts. New methodologies, powerful technologies, vast amount of data and stored digitized materials "that previous humanities scholars did not have", act as a revisionist call of what humanities research is all about.[19] The article did not mention WikiLeaks as forerunner and predecessor to the current transformation, but, for sure, it could have. 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 have stated: "to organize and interpret this Himalaya of data is a collective challenge."[20]

[1] See the various captures of wikileaks.org on 17 January 2007 through the Wayback Machine – http://web.archive.org/web/20070114162346/http://www.wikileaks.org/index.html (30 September 2011).

[2] See the article 'Wikileaks' on Wikipedia – http://en.wikipedia.org/wiki/Wikileaks (30 September 2011).

[3] See, "Wikileaks: About" 16 Februari 2008 – http://web.archive.org/web/20080216000537/http://www.wikileaks.org/wiki/Wikileaks:About (30 September 2011).

[4] Geert Lovink & Patrice Riemens, "Twelve Theses on Wikileaks" 7 December 2010 –http://www.eurozine.com/articles/2010-12-07-lovinkriemens-en.html/ (30 September 2011).

[5] For a general discussion, see for example, Nils Windfeld Lund "Document, text and medium: concepts, theories and disciplines" *JDoc* no. 5, 2010 – www.emeraldinsight.com/0022-0418.htm (30 September 2011).

[6] 'Wikileaks' on Wikipedia – http://en.wikipedia.org/wiki/Wikileaks (30 September 2011).

[7] "Secret US Embassy Cables" Wikileaks 10 February 2011 – http://wikileaks.org/cablegate.html (30 September 2011).

[8] Michael Simeone *et al*., "Digging into data using new collaborative infrastructures supporting humanities-based computer science research" *First Monday* no. 5, 2011 – http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3372/2950 (15 September 2011).

[9] James Glanz, "Survey Sees Major Expansion of World's Data Centers" *New York Times* 27 September 2011.

[10] See, "All released leaks archived" Wikileaks 28 November 2010 – http://www.wikileaks.org/file/wikileaks_archive.7z (30 September 2011).

[11] Mike Dewar, "Visualisation of Activity in Afghanistan using the Wikileaks data" Vimeo 16 August 2010 – http://vimeo.com/14200191 (30 September 2011).

[12] C. J. Shivers *et al*., "View Is Bleaker Than Official Portrayal of War in Afghanistan Activity in Afghanistan" *New York Times* 25 July 2010.

[13] For a discussion, see Simon Rogers, "WikiLeaks embassy cables: download the key data and see how it breaks down" *The Guardian* 3 December 2010 – http://www.guardian.co.uk/news/datablog/2010/nov/29/wikileaks-cables-data# (30 *September* 2011).

[14] For an illustration, see David Placr's photostream on Flickr – http://www.flickr.com/photos/55213715@N04/5122416361/in/photostream/ (30 September 2011).

[15] Lovink & Riemens 2010.

[16] FAQ Wikileaks – http://www.wikileaks.org/static/html/faq.html (30 September 2011).

[17] For a discussion, see Simon Rogers, "WikiLeaks embassy cables: download the key data and see how it breaks down" *The Guardian* 3 December 2010 – http://www.guardian.co.uk/news/datablog/2010/nov/29/wikileaks-cables-data# (30 September 2011).

[18] *Our Cultural Commonwhealth* ed. Marlo Welshons, American Council of Learned Societies, 2006, 16 – http://www.acls.org/cyberinfrastructure/ourculturalcommonwealth.pdf (30 September 2011).

[19] Patricia Cohen, "Digital Keys for Unlocking the Humanities' Riches" *New York Times* 16 November 2010.

[20] Lovink & Riemens 2010.