Abstract

*Wikipedia: The Free Encyclopedia* was launched in January 2001, and its articles now represent a major resource for understanding the world. Many of these articles have been negotiated and edited for a decade or more, and the history of that editing can provide insight into the recent history of ideas. This paper describes the development of a tool called WIScker that works with the Wikipedia Application Programming Interface (API) to scrape, or “wisck,” the revision history of any *Wikipedia* article, in order to build a corpus for subsequent text analysis and visualization. As an example, we examine a fourteen-year revision history of the article “Terrorism,” first introduced into *Wikipedia* in October 2001, the month after 9/11, and subsequently expanded to provide a more historically informed, though still politically motivated, entry.

Background: The Origin of WIScker

In the first week of each fall semester, the Humanities Computing (HUCO) Masters program at the University of Alberta runs the “Intensity Challenge” for newly-admitted graduate students. Every morning one of the HUCO professors gives the students a new challenge, one with no obvious solution. Then, working in groups, the students have until the end of the day to come up with a solution and present it. Some of the reasons for the Challenge are to introduce students to each other and the professors, introduce the

---

* University of Alberta
** Bucknell University
*** McGill University
**** The University of Illinois at Urbana-Champaign

students to the work the professors do, get the students to start doing research together, and encourage students to think on their feet, without the pressure of grades hanging over them.

Professor Geoffrey Rockwell’s challenge, in September of 2014, was to explain the difference between Humanities Computing and the Digital Humanities using text analysis. It was here that the idea behind WIScker was formed. During the challenge, Zachary Palmer and Robert Budac worked as a team and presented on the difference between Humanities Computing and Digital Humanities by using text analysis to compare several past versions of the two corresponding Wikipedia articles. From here, the idea of using past versions of Wikipedia articles as a way of studying the evolution of ideas on a topic blossomed. With financial support from the INKE team\textsuperscript{2}, Andrea Budac developed the Wikipedia Idea Scraper (WIScker) and presented it with Palmer at the Canadian Society for Digital Humanities (CSDH/SCHN) in 2015. Since then the WIScker has been enhanced, tested, and linked to Voyant, a web-based text analysis environment,\textsuperscript{3} so that users can scrape versions of a Wikipedia article over time and pass the collection of versions to tools like Voyant in order to study the evolution of the article.

**Wikipedia and Its Study**

*Wikipedia*, self-described as “the free encyclopedia that anyone can edit,”\textsuperscript{4} has become a significant authority for the collective understanding of ideas. With its accessibility and popularity, *Wikipedia* is often the first and last source of information for the plugged-in public. Even if the widely repeated—though, according to Viégas et al. (2007), “naive”—understanding of *Wikipedia*’s unreliability were true, its utility, ubiquity, and openness make it a uniquely influential source for the transmission, reproduction, and formation of ideas.

*Wikipedia* provides a wealth of data for scientists to mine and visualize in a variety of ways. In 2004 the History Flow project created visualizations showing the

---

\textsuperscript{2} The Imagining New Knowledge Environments (INKE) project was led by Dr. Ray Siemens (University of Victoria) and funded by the Social Science and Humanities Research Council of Canada. See the website at http://inke.ca accessed July 12, 2017.


contributions of *Wikipedia* authors over time. From *Wikipedia data* the History Flow researchers showed various responses to different types of vandalism, and evolving negotiations over appropriate content (Viégas, Wattenberg, and Dave 2004). In addition to looking at temporal activity, researchers have looked at network connections between various data points. An early study generated a category map by defining the similarity of categories based on their co-occurrence in articles. The map revealed major category clusters, such as television, and significant top-level categories, such as culture, geography, history, science, and people (Holloway, Bozicevic, and Börner 2007). Even relatively naive methods for visualizing data from *Wikipedia* have produced interesting results. The Chromograph project coded edit comments into colors based on the first three letters of the comment, showing evidence of temporal bursts in activity around particular pages, systemic activity around particular topics, and reactions to vandalism (Wattenberg, Viégas, and Hollenbach 2007). Additional work to improve the visualization of various data points gathered from *Wikipedia* has continued (Biuk-Aghai and Cheang 2011; Kimmerle et al. 2010; Pang and Biuk-Aghai 2011; Yousaf, Li, Zhang, and Hou 2012).

Another topic of interest among *Wikipedia* researchers has been conflict and cooperation. Yasseri et al. (2014) have identified and analyzed the most controversial *Wikipedia* articles across several languages, based in part on the frequency of these edits. Kittur, Suh, and Chi (2008) and Kittur, Chi, and Suh (2009) used Amazon’s Mechanical Turk service to recruit volunteers to classify the trustworthiness and level of conflict for a set of *Wikipedia* pages. The research showed that some subjects, such as philosophy and religion, were more prone to editorial conflicts despite comprising only a small percentage of the overall content in *Wikipedia*. Some of this work was used to develop visual interfaces which could display the edit history of a *Wikipedia* page, making it easier for readers to assess the quality of the content (Suh, Chi, Kittur, and Pendleton 2008). Ferron and Massa (2011) examine *Wikipedia*’s articles and talk pages of traumatic incidents as spaces of collective cultural commemoration.

Each of these projects relies on the reading of many—if not all—versions of a *Wikipedia* article and its attendant discussion pages to understand the history of the page. The WIScker project makes this process simple, intuitive, and accessible.

That said, ours is not the first publicly available web tool that allow people to investigate *Wikipedia*’s editing complexities. The Open Wikipedia Ranking from the Università degli Studi di Milano graphs *Wikipedia* pages and sorts them by centrality to
identify the most “important” pages in various fields. In 2007, Virgil Griffith created the WikiScanner, a searchable database linking anonymous *Wikipedia* edits to the organizations that may have been responsible. It was built with a tool that reads the IP addresses of otherwise anonymous *Wikipedia* editors so you can try to match the IP address. In doing so, the WikiScanner can detect when anonymous edits originate from an IP address belonging to an official institution. Often, as independent use of the tool uncovered, the editing institutions had an obvious interest in managing the public perception of the edited pages—usually pages for the institution itself, a competitor, or a political issue on which it had an ideological stance. Griffith eventually removed the WikiScanner from his website in 2013, citing hosting costs, but Federico Scrinzi and Paolo Massa, researchers at the Fondazione Bruno Kessler, have since released an equivalent tool called the WikiWatchdog that searches for anonymous *Wikipedia* edits by originating domain name or IP address.

The WikiScanner and WikiWatchdog differ from our WIScker because they focus on who does the edits and thus track only the versions of articles before and after the edits. They are designed to track who is editing. The WIScker, by contrast, collects the long-term evolution of a particular page to study the content that is edited. We believe using the two types of tools together can help study how changes made by particular editors can change an article over time.

**WIScker and its Abilities**

WIScker is written in JavaScript, making it easy to run; all that one needs is a browser that supports JavaScript. It is also licensed as open-source software, so it is easily downloadable and editable for those who wish to do so. All one has to do is to download and run the tool locally is to save the web page (http://cloud.tapor.ca/wiscker/) to one’s local drive.

---


Using WIScker is a simple process; to start, one only needs to copy the URL of the Wikipedia article one wants to scrape into the top left field. WIScker allows for scraping of the article by time or by frequency. So, for example, one can grab the first revision every month for 5 years, or every 25th revision of the article.

As well, one can choose how the text should be displayed. First, there are two formatting options for the metadata that is provided by the Wikipedia API, such as author and revision date. To facilitate text analysis, XML is the default. The XML format is structured so that external programs like Voyant, which are XML-aware, can use the metadata. The alternative to XML is a “human readable” format which is meant to be browsed. Next, there are two formatting options for the text of the Wikipedia article. The default, “plain text,” is cleaner and easier to read but has extra data such as descriptions of links and references removed, whereas those are kept with “Wikipedia formatting.”

To run the tool one presses the “scrape” button. Assuming there are no errors in the form, the tool should build a corpus and output it underneath the article on the right. In order to facilitate copying this text, a “select text” button is included. We have also added a “launch in Voyant” button that currently opens the corpus in Voyant Tools 2.2. As long as
as the corpus is in XML format, Voyant Tools should be able to split it into multiple documents, based on revisions.

**Tracing the Public Negotiation of an Idea**

What is the point of WIScker, and how can one use it? Tracing the changes in an idea through its negotiated public representation in the *Wikipedia* ties into INKE’s ongoing experimentation in the “epidemiology of ideas.” The INKE project has been experimenting with various text analysis tools and techniques on large corpora to trace the history and development of concepts (ideas). We, the authors, call this work “epidemiology” as it is analogous to the work of examining patterns in the spread and correlation of diseases in order to model or predict disease proliferation.

Using epidemiological techniques to study the spread of non-contagious phenomena has precedent both within and outside medical epidemiological scholarship. The medical researchers Christakis and Fowler have extended epidemiological methods to noncommunicable diseases like obesity and habitual behaviors like smoking (Christakis and Fowler 2009). Citation analysis has long been used to trace influences within academia. Bollen et al. (2009) attempted to create a more true-to-life model of academic influence by tracking not the citations in the final work but the individual clicks around a journal repository website during preliminary research.

Other projects that have adapted similar techniques to cultural and intellectual contexts include Underwood and Goldstone’s (2012) use of topic modeling to connect concepts in literary history. So and Long’s (2013) network analysis of modernist poetry is a more focused attempt to structurally analyze a particular school of thought through the social interactions of its participants. These examples demonstrate the utility of the concept as well as the necessity of adapting the methodologies of epidemiology, rather than mapping them across disciplines. For example, Shalizi and Thomas (2011) arguably mistake homophily for contagion—claiming that a social network caused an idea, when it is likely that the social ties that form the network were instead caused by the shared idea. Understanding this fallacy is even more important in tracing an epidemiology of ideas than it is for examining non-contagious disease and behavior.

**“Terrorism” Example**
To give a concrete example of how WIScker can be used, we looked at the *Wikipedia* article “Terrorism” and how it changed between 2001 and 2014. We chose “terrorism” as it is an idea that has gotten a lot of attention since September 11, 2001. We hypothesized that the idea of terrorism has been negotiated publicly on sites like *Wikipedia* since 9/11 (whereas before the *Wikipedia* it would have been negotiated more in specialized literature). This is partly confirmed by the fact that the *Wikipedia* article has edits locked down and subject to review under the “pending changes” setting due to “problematic” edits—a sign of contestation.

As a first pass we set WIScker to capture one version a year, beginning with the first entry on October 2, 2001. Figure 2 shows what the opening XML of the resulting corpus looks like. Note that the date of this first entry is after 9/11; prior to October 2001, there is no record of a “Terrorism” entry in *Wikipedia*.

```
<?xml version='1.0'?>
<collection>
<header>
<wikipedia_article_title>Terrorism</wikipedia_article_title>
<scrape_time>Tue Apr 28 2015 13:15:17 GMT-0600 (MDT)</scrape_time>
<start_scrape_date>2001/10/02</start_scrape_date>
<end_scrape_date>2015/03/02</end_scrape_date>
<time_interval>every year (365 days)</time_interval>
</header>
<version>
<revision_date>2001-10-01</revision_date>
<revision_time>T01:49:01Z</revision_time>
<author>The Cunctator</author>
<body>Terrorism refers to the calculated use of violence or the threat of violence by individuals or non-governmental groups (what about the same act if committed by and agent of the state?), often against the civilian population, to instill fear in an audience for purposes of obtaining political goals. One who carries out such acts is a terrorist,
```

Figure 2. XML from WIScker

The next step was to load the corpus into Voyant (http://voyant-tools.org) using the XML ingestion options. We set it up so that Voyant would treat each <version> element as a document and use the content of the <revision_date> as the title for the document. This gave us a corpus of 14 documents, one for each year from 2001 to 2014.\(^8\)

\(^8\) This corpus can be accessed and tried by anyone at http://bit.ly/wiscker.
Figure 3. XML Options for Voyant 2.2 that WIScker automatically fills in

Figure 4. Cirrus word cloud of terrorism corpus
Figure 4 shows a word cloud, with common stop words removed. Figure 5 shows another word cloud, this time with the word “terror” and its derivatives also removed, since we already know the topic of the entry.

![Figure 4](image4.png)
![Figure 5](image5.png)

Figure 5. Cirrus word cloud of terrorism corpus minus the word “terror” and its derivatives

What we understand from looking at the high-frequency words in this case is that terrorism has something to do with violent acts involving politics and governments. The version from 2001 emphasizes this fact by stating explicitly that terrorism is a political act rather than a military one. The 2014 entry, on the other hand, mentions in the first paragraph that terrorism is considered a war crime, which implies that it is a military act. More interestingly, from the point of view of studying the negotiated change in an idea, is that loading it into Voyant allows us to find words that have a spike or skew over time. Figure 6 shows a screenshot of Voyant where we have added columns for “peakedness” and “skew” in the “Words in the Entire Corpus” panel. We have then sorted on peakedness to show words that “peak”—that is, words that show up in one version but not others, as shown in figure 7.
Take the word “independent,” which shows up in the 2005 version in paragraphs like this one:

The Fenians/IRA and the IMRO may be considered the prototype of all 'nationalist terrorism', and equally illustrate the expression that "one man's terrorist is another man’s freedom fighter". Both groups achieved their goal, an independent Ireland and an independent Macedonia. (“Terrorism” 2015)

This appeared in the history of terrorism section in 2005 that was then moved to a longer article, meaning the brief history in the main article was shortened. The part about achieving goals through terrorism was also edited out; in the 2014 “History of Terrorism” article in Wikipedia (https://en.wikipedia.org/wiki/History_of_terrorism), assertions about the success of terrorist groups like the IRA in establishing “independent” states have been removed.
The word “freedom” also has an interesting trajectory, which can be seen in figure 8. It is not used for the first two years, then it briefly appeared in the phrase “freedom fighter,” and increased in frequency in 2006 through further discussion of freedom fighters.
We can also use negative peakedness and skew to see words that indicate sections added to the article at different points. Using these tools, we find that there is a section on perpetrators, added in 2005, with data coming from studies about the education of perpetrators. Sorting by skew lets us see more complex patterns of introduced terms, which can be seen in figure 9. For example, we can see that the term “tactics” has an increase in relative frequency over time.
Conclusions

In sum, WIScker provides a tool for examining the negotiation of ideas in and through Wikipedia. It does not provide a tool for big data gathering and analysis. Instead it provides a small utility suitable for gathering a corpus around an idea in one context: Wikipedia. It also does not do any analysis—it just gathers a corpus for use in analytical tools like Voyant. Further, it illustrates how small JavaScript tools made available through web pages that can be saved locally (and edited as needed) are a useful lightweight way of sharing tools among digital humanists.9

---

9 Another example would be TiddlyWiki, http://tiddlywiki.com/, which provides a simple database tool.
We believe WIScker has promise as a tool that students can use to learn about corpus building and analysis through examining *Wikipedia*. We hope that studying such diachronic corpora will discourage the blind trust in *Wikipedia* that galls so many experts. Seeing the way the texts of *Wikipedia* articles shift can encourage critical thinking about it as a source of information.

A disadvantage of the way WIScker was implemented is that it cannot easily be scaled up. You can manually combine corpora to look at small sets of articles, but the WIScker does not do this automatically. One can imagine a larger-scale project that would scrape a large collection of the variants of an entire class of *Wikipedia* articles, if not all articles. Such a “big data” collection could then be mined for large-scale changes in style and/or content across many article chains. One could also track passages that are moved from one article to another or editors working across different articles. To create such a collection one would need a different type of scraper that can work in the background, or one could work directly with a downloaded version of the full *Wikipedia* database (see https://en.wikipedia.org/wiki/Wikipedia:Database_download). Connecting to Voyant would also involve a different architecture to support a much larger collection of texts in the interface of Voyant.

Perhaps WIScker signals another change coming to the field we now call the digital humanities. WIScker is very much a tool for making small corpora off the web dynamically. WIScker lets us reflect critically on one important web service, *Wikipedia*. It helps us read changes in articles closely. But now big data is becoming the next frontier for digital humanists and the web is ceasing to be so central to our work. As big data collections like Twitter, the Internet Archive, and HathiTrust become accessible to research, they allow us to ask different types of questions of different types of evidence, namely big data. They allow us to explore the epidemiology of ideas on a different scale, leaving the reading of the web behind for the automated reading of millions of books (Crane 2006). WIScker can be seen as one of many different types of scrapers that build corpora on demand for text mining.

**Links**

WIScker: http://cloud.tapor.ca/wiscker/
Code: https://github.com/abudac/LNKE_WIScker
References


