

WHERE IS HISTORY BEING WRITTEN?
GEOREFERENCING CONTRIBUTIONS
TO WIKIPEDIA

DAVID KALTSCHMIDT

Diplomarbeit

Dr. Claudia Müller-Birn
Prof. Dr. Robert Tolksdorf
Institut für Informatik
Freie Universität Berlin

David Kaltschmidt: *Where is history being written? Georeferencing
contributions to Wikipedia*
Diplomarbeit, © 2011

SUPERVISORS:
Dr. Claudia Müller-Birn
Prof. Dr. Robert Tolksdorf

LOCATION:
Berlin, Germany

YEAR:
2011

PRODUCED WITH:
L^AT_EX using the ClassicThesis package.

ABSTRACT

Wikipedia is more than an online encyclopedia. It is also a news channel as well as a self-updating history book. A global readership can follow political events as they unfold, written about by local people and later edited by other volunteers. This thesis describes a method to answer the question to what extent local volunteers write about events in their own country. First, the geographic origin of each individual article contribution is determined. In a second step, a given article is annotated with georeferences on a word level. The properties of these annotations then allow for a statistical geographic analysis of a single article or a category of articles.

ZUSAMMENFASSUNG

Als Online-Enzyklopädie ist Wikipedia nicht nur Nachschlagewerk sondern auch ein sich stetig wandelndes Geschichtsbuch. Eine global verteilte Nutzerschaft liest und schreibt über lokale Ereignisse noch während sie passieren. Diese Arbeit beschreibt eine Methode zur Bestimmung des Anteils an Beiträgen, die vom betreffenden Land ausgehen. In einem ersten Schritt werden die geographischen Ursprünge aller Beiträge eines Artikels ermittelt. Mit den daraus erhaltenen Georeferenzen wird der Artikel Wort für Wort annotiert. Basierend auf diesen Annotationen kann dann der lokale Autoren-Anteil bestimmt werden.

CONTENTS

I	THOUGHTS	1
1	INTRODUCTION	2
1.1	Structure	3
2	FOUNDATION	5
2.1	Wikipedia	5
2.1.1	History	6
2.1.2	Wikimedia Foundation	6
2.1.3	Anatomy of an article	6
2.1.4	Wikipedia, wikis and editing	7
2.1.5	User pages	8
2.1.6	Categories	8
2.1.7	Templates	8
2.2	Contributions	8
2.3	Georeferences	9
3	HYPOTHESES	10
3.1	Article creation	10
3.2	Participation	12
3.3	Text survival	13
II	METHODS	14
4	APPARATUS	15
4.1	Data sources	15
4.1.1	Wikipedia website	15
4.1.2	Database dumps	15
4.1.3	MediaWiki API	16
4.1.4	Toolserver	16
4.1.5	Third-party sources/Web services	17
4.2	Available Tools	18
4.2.1	Toolkits	18
4.2.2	Analysis projects	18
4.3	Application design	18
4.3.1	Technologies	19
4.3.2	Models	19
4.3.3	Views	19
4.3.4	Operation	19
4.4	Algorithms	19
4.4.1	Article requirements	19
4.4.2	Date parsing	19
4.4.3	Location parsing	20
4.4.4	Collective authorship	20
4.4.5	Resolving user names to IPs	20
4.4.6	IP Look-up	20

4.4.7	Parsing user pages	21
4.4.8	Geographic resolution	21
4.5	Visualization	21
4.5.1	Maps	22
4.5.2	Maps	22
4.5.3	Line and scatter charts	22
4.5.4	Motion chart	22
4.6	Hypotheses analysis	22
4.7	Possible enhancements	22
4.7.1	Edit relevance	22
4.7.2	Geographic profiling	22
5	EXPERIMENTS	24
5.1	Data sets	24
5.1.1	By category	24
5.1.2	By template	24
5.1.3	Political article vs. place article	24
5.2	Application run	24
5.2.1	Distribution	24
5.2.2	Text survival	24
5.2.3	Localness	24
5.2.4	Motion chart	24
III	RESULTS	25
6	RESULTS	26
7	CONCLUSION	27
7.1	Limitations	27
7.1.1	Political events	27
7.1.2	Article location	27
7.1.3	Cross-language article growth	27
7.2	Further Research	27
IV	APPENDIX	28
	Bibliography	29

LIST OF FIGURES

- | | |
|----------|--|
| Figure 1 | The article <i>2011-2012 Bahraini uprising</i> viewed in a web browser on 00/00/0. 6 |
| Figure 2 | Revision history of the article <i>2011-2012 Bahraini uprising</i> on 00/00/0. 7 |

LIST OF TABLES

LISTINGS

ACRONYMS

Part I

THOUGHTS

INTRODUCTION

If you are open to contributions from others, you generally end up with richer, better, more diverse and expert content than if you try to do it alone.¹

— Alan Rusbridger, editor of THE GUARDIAN

At the end of January 2011, when a wave of public protest spilled from Tunisia into Egypt, a small group of opposition parties and political activists called for a “Day of Rage” via Facebook, a social networking website. By January 25th their Facebook group had more than 80,000 supporters who drew attention to and helped organize the country-wide protests that followed. As people rallied the streets day after day, the Egyptian government first limited access to Twitter, a micro-blogging service, before cutting Egypt off the internet completely on January 28th.[2, 3]

In what came to be known as the Arab Spring, the use of online networks directly influenced the political development. While Facebook played a part in organizing the protests, Twitter acted as an information channel during the demonstrations. As the events unravelled, they were reflected by articles created on Wikipedia, an online encyclopedia. Updated by the minute, the articles covering the protests formed a well of news reports.[4]

Wikipedia’s free access and open editing policy as well as a quality level—putting it “head to head”[5] with Encyclopedia Britannica—turned it into a hugely popular website[6]. The server software used for the website, MediaWiki², ensures that the effort to change an article is minimal. Given an Internet connection and a web browser, anyone can add or edit an account of current events in a related article and publish it in a matter of seconds.

This form of news production turns the encyclopedia into a news channel that is constantly updated and corrected by an army of volunteers. The result is a self-governed news source that lends itself the aura of authority and credibility of a knowledge reference. At the same time a technophile public that uses the Internet as an efficient means of news acquisition, can check facts on Wikipedia and act upon the consumed information.[7, p. 424–427] Therefore the collective authorship of such a news

¹ The Economist. *The people formerly known as the audience*. URL: <http://www.economist.com/node/18904124> (visited on 08/10/2011)

² <http://www.mediawiki.org> (visited on 10/31/2011)

medium could have a direct influence on the political decision-making process. As ordinary people become producers of journalism the need arises to analyze these contributions. To investigate, this thesis focuses on the geographic origins of contributions to Wikipedia articles.

Political events are often limited to a country or region. This is reflected by the Wikipedia articles covering the Arab Spring: there is an overarching parent article³ as well as single articles covering the revolution in each of the affected countries, e.g. Egypt⁴ and Libya⁵. The events in the latter country also exemplify how divided the political actors can be. While nearly all revolutionaries welcomed the airstrikes, one faction was concerned about foreign meddling and another one just opposed the deployment of ground troops.[8]

The collective authorship could be equally divided. Despite Wikipedia's core policy to oblige everyone to write from a *neutral point of view*⁶ (NPOV), people regularly express opinions. The collision of opinions in a collectively written article can result in a prolonged series of an edit and its subsequent reversal by another person. The resulting edit pattern is known as an *edit war*. [9] These clashes of opinion create a potential for further investigation into the geopolitics of article contribution. Where do the first reports of an event originate? As later iterations of revisions turn these reports into historical accounts, are these editors from the same country? And more generally, to what extent is a collection of these articles written by volunteers located at the respective location of the event.

In this thesis I will propose a method to help answer these questions. By trying to determine the geographic origin of each edit to an article I will be able to calculate the geographic distribution of contributors. This distribution will then be used to answer the questions above for either a single article or a collection. *Include complete summary and key findings?*

1.1 STRUCTURE

complete over time, name the basic chapters and their function, one part = one paragraph

The chapter **FOUNDATION** provides background information about **Wikipedia**, article editing (**Contributions**) and the applica-

³ http://en.wikipedia.org/wiki/2010-2011_Middle_East_and_North_Africa_protests (visited on 10/31/2011)

⁴ http://en.wikipedia.org/wiki/Egyptian_Revolution_of_2011 (visited on 10/31/2011)

⁵ http://en.wikipedia.org/wiki/2011_Libyan_uprising (visited on 10/31/2011)

⁶ http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view (visited on 12/08/2011)

tion of geographic data ([Georeferences](#)). The first part ends with [HYPOTHESES](#) where I propose the research questions that this thesis hopes to answer.

In [APPARATUS](#) I will describe the tools available to be used in the method that will be applied to a host of articles in [EXPERIMENTS](#). The findings will be presented in [RESULTS](#). Followed by a discussion of their feasibility in [CONCLUSION](#).

Wikipedia is a phenomenon that has attracted researchers across all fields, notably computer science and sociology, who have written over 1,000 reports on the subject to date. Nielsen [10] compiled an overview of Wikipedia research¹ and divides these publications into four categories:

CONTENT PRODUCTION Covering all aspects of voluntary production such as motivation, collaboration, coverage and bias, quality and vandalism, actuality and geography.

INFORMATION USE Treating how the resulting corpus is being used, e.g. Wikipedia citation in research, use in court, trend-spotting, natural language processing and automatic translation tools, thesaurus construction or categorization.

IMPROVEMENT These are studies concerned with the improvement of both the software used by Wikipedia and the content, e.g. automatic linking, improved editors as well as quality and trust indicators.

COMMUNICATION Studies in this category look at Wikipedia as an online collaboration tool for education and research.

This thesis falls into the first category, content production, as it examines the geography of article contributions that will become part of the Wikipedia's corpus. After a short overview of Wikipedia from a user's perspective, I will introduce its model of collective authorship and present prior research of concerning location and geography.

2.1 WIKIPEDIA

Wikipedia is an online encyclopedia with editions in over 260 languages. Counting 3.6 million articles, the English version is by far the biggest. However, other language editions differ sharply in size and usage.[11] If articles covering the same topic exist in other language editions, these are connected by interwiki links.

¹ Another resource is the Wikimedia Foundation's own directory of Wikipedia research projects at <http://meta.wikimedia.org/wiki/Research:Projects> (visited on 10/12/2011).

2.1.1 History

Wikipedia was officially started on 15 January 2001 by Jimmy Wales and Larry Sanger. Wales previously founded Nupedia, a free and peer-reviewed online encyclopedia written only by experts. However, the speed of content production was extremely low. Wikipedia was founded as a feeder project to collectively write on articles before these entered Nupedia's review process. Wikipedia then quickly created other language editions and dwarfed its predecessor².^[12]

After being mentioned on Slashdot, a technology news website, in March 2001 Wikipedia quickly attracted new users. This tech-savvy folk created new articles at a staggering rate of 1,500 articles per month in the first year. These articles then quickly started showing up in Google's search results, attracting even more new users. The non-English editions grew slower but as a group accounted for 75% of all articles in 2007. By 2011 the combined article count passed 20 million.^[12]

2.1.2 Wikimedia Foundation

Foundation structure, independence of language wikis, funding

2.1.3 Anatomy of an article

All Wikipedia articles share a similar layout: a large content area topped by the article title. Article titles can change over time, e.g. *2011 Bahraini uprising* was renamed to *2011-2012 Bahraini uprising*³. For these cases, Wikipedia has a redirecting mechanism that forwards the visitor to the final article and displays a small note below the title (see figure 1).

Figure 1: The article *2011-2012 Bahraini uprising* viewed in a web browser on 01/23/2012.

Occasionally the content section can be topped by one or more warning boxes to inform the visitor that the article is violating an editing policy, e.g. the information of the article may be outdated because it is subject to current events. When an article spans several sections a table of contents is added below the first introductory paragraphs. In addition to prose, some articles feature info boxes on the right hand side. These boxes show information in a structured way and can be found on ar-

² Only 24 articles were completed in Nupedia's review process. The project was officially ended in 2003.

³ http://en.wikipedia.org/wiki/2011-2012_Bahraini_uprising (visited on 01/23/2012)

ticles of similar topics giving the visitor a quick glance on key information without having to read the text.

These information may include dates and geographic coordinates. E.g. the article *2011-2012 Bahraini uprising* has the time interval “14 February 2011 – ongoing” and is tagged with the coordinates $26^{\circ}01'39''\text{N}$ $50^{\circ}33'00''\text{E}$, pointing to the centre of Bahrain. Even when no coordinates are present in the article, it still may be associated to a location. In that case the info box just presents the place, e.g. *Bahrainfind article*, instead of providing the geographic coordinates.

2.1.4 Wikipedia, wikis and editing

Anyone with a browser and internet access can edit Wikipedia’s articles⁴. In collaboration, people all over the world contribute and improve the content. This is made possible by MediaWiki, the the server software that makes Wikipedia a wiki. The software allows website visitors to add and modify the page content in the browser using *wikitext*, simplified markup language.⁵ Its syntax can be used to structure a text into sections, embed images and links to other pages, much like HTML. The syntax is explicitly kept simple to keep the entry barrier to editing low.

Each submission of an edit in the browser creates a new revision of the article and is stored in the revision history, see figure 2. Naturally, each article available today started from an empty page and is the result of a succession of edits. Each entry in the revision history consists of the new wikitext, the date of submission, the user and an optional comment explaining the change. Each revision can not only be examined by other users and but also reverted. To minimize the potential for *edit wars*[9] Wikipedia urges its users to discuss controversial topics on the article’s talk page.

Figure 2: Revision history of the article *2011-2012 Bahraini uprising* on 01/23/2012.

Contributions to an article can be done anonymously or as a registered user. A registered user gains privileges like the ability to create articles or the use of the social network features in Wikipedia. With the initial registration a *user page* is created where the user is allowed to publish a profile and interact with other registered users.[14] The majority of edits comes from reg-

⁴ Some articles can be locked because of sustained vandalism or content disputes.[13]

⁵ For the syntax see http://en.wikipedia.org/wiki/Help:Wiki_markup (visited on 12/12/2011).

istered users, anonymous edits account for a quarter of all edits.[15]

A third group of editors are automatic programs known as *bots*. They perform routine tasks ranging from spell-checking over curse word detection to automatic reverts on vandalism. Currently the English Wikipedia alone has nearly approved 1,500 bot tasks running.[16] Some editors may be registered as a bot

2.1.5 *User pages*

write about user pages, prose, boxes

2.1.6 *Categories*

Wikipedia's category system, not a tree but rather a graph

2.1.7 *Templates*

Wikipedia's template system, can be used to get group of similar articles

2.2 CONTRIBUTIONS

Wikipedia's articles are continuously edited by its users. The nature of an edit can range from simple spelling or grammar correction, over improving the content of a sentence to writing or removing of paragraphs or even whole articles. This collective authorship makes it difficult to determine an individual author's contributions, in other words, it is not easy to tell who wrote what.

Research in this area tends to be motivated by the desire to identify individual authors with a good reputation in order to assign a trust score to them. This is based on the assumption that trusted authors consistently produce high quality contributions that outlive contributions of lower quality. Kramer, Gregorowicz, and Iyer [17] devised a method to assign trust scores to the authors of an article by examining the wealth of information contained in the article's revision history. They looked at an article as being a set of phrases. The author who first wrote a sentence gets the credit for that phrase and will gain trust if it survives future edits.⁶

A similar approach of calculating the longevity of text chunks was followed by Adler and De Alfaro [18]. They adapted stan-

⁶ Kramer, Gregorowicz, and Iyer [17] define a sentence as an n-gram—a sequence of n words—and use a sliding window model to follow it across revisions to prevent simple rearrangements of text from counting as a new sentence.

dard text-diff algorithms to the peculiarities of the wiki revision system, e.g. keeping track of text chunks that were removed at one point and then reinserted in later revisions. Based on these algorithms a reputation system was implemented by Adler et al. [19] which offers an API⁷ that can annotate a Wikipedia article. The annotated text is the result of splitting the original text into chunks and attributing them with their respective authors, the number of the revision where the chunk was added and a trust value for the author.⁸

2.3 GEOREFERENCES

"Combined approaches (i.e., where quantitative spatial analysis models are calibrated with surveyed locations) may prove useful." [20, p. 85]

In order to analyze the localness of contributions, it is necessary to geotag them, i.e. applying geospatial metadata like coordinates to each contribution, derived from the author's location. In his doctoral thesis Hardy [20] used the Wikipedia corpora to study the spatial behavior of article production. The dataset was limited to anonymous users and articles that were geotagged.

For each anonymous contribution an IP address, belonging to the point of Internet access, is stored in the revision that is created. Various methods to determine the geographic location from a given IP address have been studied by Muir and Oorschot [21]. Various visualizations^{9 10} of edit distributions use geolocation databases like MaxMind¹¹ and Quova¹².

For registered users, the IP address is not stored with the revision. Therefor IP geolocation services cannot be used. Lieberman and Lin [22] found an interesting approach by assuming users prefer to edit geographic articles in their proximity. The approximated user location was derived from the center of the convex hull around those articles.

Registered users are also given the opportunity to create a personal profile in their *user page*. The user can choose prose or structured boxes to reveal information like his general interests, spoken languages, but also his location. When entity names can be extracted from location information they can lead to coordinates as shown by Hecht and Gergle [23].

⁷ <http://www.wikitrust.net/vandalism-api> (visited on 10/31/2011)

⁸ Adler et al. also released a Firefox add-on that highlights untrustworthy passages when viewing Wikipedia articles: <https://addons.mozilla.org/en-US/firefox/addon/wikitrust/> (visited on 11/15/2011).

⁹ <http://infodisiac.com/blog/2011/05/wikipedia-edits-visualized/> (visited on 10/31/2011)

¹⁰ <http://sonetlab.fbk.eu/wikitrip/> (visited on 10/31/2011)

¹¹ <http://www.maxmind.com> (visited on 10/31/2011)

¹² <http://www.quova.com> (visited on 10/31/2011)

3

HYPOTHESES

To gain insight on how Wikipedia is being used during and after political events¹, and ultimately whether articles covering the events are written by people that are most affected, I will propose a set of hypotheses aimed at different aspects of article production.

For this thesis I will use the event definition proposed by Lewis [24, p. 243]:

“An event is a localized matter of contingent fact.
[...] An event occurs in a particular spatiotemporal region.”

It follows that events must have clear spatial and temporal boundaries. The spatial boundaries give it a location, distinguishing the where. The temporal boundaries, namely the start and the end date, divide the event into three intervals: *before*, *during* and *after*.² Let further me further denote the *beginning* of an event as the first three days of an event.³ In conjunction with location this division into time intervals allows for a detailed look into [Article creation](#), the level of [Participation](#) in following intervals as well as [Text survival](#).

3.1 ARTICLE CREATION

The use of Wikipedia concerning a political event starts with the creation of the article describing the event. Due to the website’s popularity I would expect the delay between the start date and the date of article creation to be rather short. Moreover, considering the rising year-on-year Wikipedia usage in numbers of pages viewed[6], this delay should become shorter and shorter suggesting an increased use of Wikipedia as a news channel. This leads to the following hypotheses:

HYPOTHESIS 1. Articles are created with only a short delay after the start date of the event.

¹ I will not try to argue what makes an event political but rather identify a set of events by picking a suitable category of articles.

² For ongoing events the end date will be the date of the analysis.

³ The interval was picked arbitrarily but acknowledges the fact that event dates in Wikipedia articles rarely carry a time attribute, therefore a shorter interval, say 24 hours, is less feasible.

HYPOTHESIS 2. The more recent an article, the shorter is the delay between the event start and article creation.

A user has the chance to create a new article in any of the 260-odd language editions of Wikipedia. Although the English version is by far the biggest and most used, it would be interesting to see whether it is the prime choice to create the first article for a new event. Shortly after the first article has been created, articles covering the same topic will be created across various editions of Wikipedia. These articles are then being linked, mostly manually, via inter-wiki links.⁴ When studying knowledge diversity across language editions, Hecht and Gergle [26] found that the English edition is not the superset of concepts of all editions as was previously believed. This means authors retain knowledge they consider important only to their compatriots. For a citizen of a country where English is not the first language creating an article becomes a political decision: should the author make the information available to fellow citizens or to a world-wide readership. However, picking English, even though it is not an official language of the country where the event is happening, would further support Wikipedia's role as a news channel, thus:

HYPOTHESIS 3. Articles are being created first in the English Wikipedia.

Regarding the localness of contributions, Hardy [20, p. 57] has established that Wikipedians write about places in their proximity more often than distant ones. His sample included only articles that have a geotag. Naturally, articles about geographic places like towns and sights⁵ will dominate this sample. Since my thesis is concerned with political events, I find this point to be worth revisiting⁶ for the people most affected should be the ones creating the article, thus:

HYPOTHESIS 4. Articles about political events are created by people in the events' proximity⁷.

Hypotheses 1–4 will only be tested against articles that were created as a reaction to an event that has already started. This

⁴ Adar, Skinner, and Weld [25] found that between two languages the inter-linking is not symmetrical, i.e. the number of out-links does not match the number of in-links. Links are either missing on one side or the respective topics are not congruous and the user intentionally left out one direction.

⁵ According to Kittur, Chi, and Suh [27] articles about *geography and places* are third biggest group.

⁶ In addition, Hardy [20, p. 61] considered only anonymous users. Since creating an article is only allowed for registered users, his method has to be extended.

⁷ Hardy [20, p. 57] defined proximity not in absolute terms, rather he considered the likeliness of authors being located less far than the average distance between an article and all its contributors.

excludes scheduled events like elections, e.g. *Russian legislative election, 2011*⁸ which was created 335 days before the election date, almost a year in advance.

3.2 PARTICIPATION

A Wikipedia article usually has more than one author. Once it has been created, users from around the globe can edit an article collectively. Viégas, Wattenberg, and Dave [28] tried to find patterns in the revision history that would reveal certain aspects of collaboration or the lack thereof, e.g. discussions and vandalism.⁹ In respect to authorship, the researchers found the proportions of anonymous contributions differed strongly from page to page while showing no preference to any topic. This inconclusive result and the age of the sample¹⁰ merits further investigation.

In 2007, Kittur et al. [29] found that a core of registered users is still doing the bulk of all edits. However, anonymous users contribute considerable amounts of text. For accounts of political events, due to their dynamic nature, I expect a strong participation by unregistered users while the events are still unfolding:

HYPOTHESIS 5. In the beginning of the event anonymous users contribute more than registered users.

HYPOTHESIS 6. For the duration of the event there are more local contributions than distant ones.

When the political event is considered “over” its end date in the article changes from “present” to a calendar date. In this unbounded and final phase I would expect the flood of contributions to subside and the content to consolidate when editors tighten prose or remove text they believe to be irrelevant. Looking at the whole lifespan of an article I would also expect registered users to outnumber anonymous ones as suggested by Kittur et al. [29] and the spatial distribution of contributions to become less local. Thus the final hypotheses:

HYPOTHESIS 7. Articles of a political event that has ended will continuously shrink in size.

⁸ http://en.wikipedia.org/wiki/Russian_legislative_election,_2011 (visited on 01/07/2012)

⁹ Using their history flow visualization Viégas, Wattenberg, and Dave [28] first identified patterns in single articles and later tried to statistically confirm their prevalence by analyzing the complete English corpus.

¹⁰ Viégas, Wattenberg, and Dave [28] used a dataset from May 2003.

HYPOTHESIS 8. After an event has ended, there will be more contributions from registered users than from anonymous ones.

HYPOTHESIS 9. After an event has ended, the spatial distribution of the contributors will become less local.

3.3 TEXT SURVIVAL

In 3.2 the contributions are only treated in volume giving credit to each contributor. However, when multiple authors write the same article, they do not only add text but also modify or even delete parts. A user who reads an article will only see the text that has survived all edits after it was added. Viégas, Wattenberg, and Dave [28] found that early contributions have a high survival rate. Recognizing this *first-mover advantage*, I suspect that accounts of political events show a strong localness in the beginning. Thus the key hypotheses from 3.2 have to be extended to reflect the spatial distributions of the contributions that make up the article:

HYPOTHESIS 10. For the duration of the event the article text contains more local contributions than distant ones.

HYPOTHESIS 11. After an event has ended, the spatial distribution of the surviving contributions will become less local.

This concludes the statement of the hypotheses. To test them, I will develop an APPARATUS and the EXPERIMENTS in the next part.

Part II

METHODS

APPARATUS

This chapter describes data sources to get Wikipedia content like articles and revision history as well as tools to retrieve and analyze those.

Intro aspects of apparatus

4.1 DATA SOURCES

For an automated analysis, simply browsing Wikipedia's website is not really feasible. *why?* The bulk of Wikipedia's content like articles, revisions, discussions is stored on its database servers. Unfortunately, these databases are not directly accessible over the Internet. The Wikimedia Foundation, however, makes a lot of the data available in the form of database dumps or through an application programming interface (API).

4.1.1 Wikipedia website

used to get static data useful throughout the analysis process, screen scraping, static data

BOTS *list of bots*

COUNTRIES *country names*

4.1.2 Database dumps

Monthly database snapshots of all wikis run by the Wikimedia Foundation, including Wikipedia, are publicly available¹ as database dump files in the XML file format. For each of the wikis a variety of dumps is available that include all articles and, optionally, their revision history, all categories, interwiki-links, etc. Despite this openness, some database tables are not publicly available. The dump files of the database tables *users* and the *watchlist* are kept private.

The dump files can be quite large, e.g. a compressed dump of all articles of the English Wikipedia in their current revision has a size 7.3 GB.² This huge size makes processing them rather

¹ <http://dumps.wikimedia.org> (visited on 12/11/2011)

² The uncompressed size is 31.0 GB, see http://en.wikipedia.org/wiki/Wikipedia:Database_download (visited on 12/11/2011)

slow.³ When analyzing only a single article or a category articles, the MediaWiki API can deliver the same information contained in the dumps in a much more targeted manner.

4.1.3 *MediaWiki API*

Wikipedia runs on the open source software MediaWiki. This PHP-based wiki package offers a well documented API which can be used by other programs to remotely use the wiki's features such as changing content and restoring revisions. For analysis of articles, the API offers queries directed at a variety of article properties, e.g. revisions, categories and links. Similar to MediaWiki's Special:Export page⁴, the API also offers an article export that includes all revisions.

Sandbox, Example call

The following queries will be important for this thesis:

- *query info (pageid info, wikitext, langlinks, ...)*
- *parse (HTML version of wikitext)*
- *parse (HTML user pages)*
- *open search (autocomplete search)*
- *revisions (all revisions, first revisions of all languages)*
- *category members*
- *template embedders*

Category and template embedders will be used to get groups of articles

4.1.4 *Toolserver*

The Germany based Wikimedia Deutschland e.V. runs Toolserver⁵, a platform for software tools that can access a continuously updated copy of Wikipedia's databases. Among these replicated databases is the English Wikipedia and other major language editions. However, the deployment of self-made software scripts is restricted and requires an account on Wikimedia's Toolserver.⁶

³ The project WikiHadoop addresses this problem by offering a stream task format to be used in Hadoop (MapReduce) infrastructure, see <https://github.com/whym/wikihadoop> (visited on 12/11/2011).

⁴ The page <https://en.wikipedia.org/wiki/Special:Export> (visited on 12/11/2011) allows for exporting of articles from the English Wikipedia.

⁵ <http://toolserver.org> (visited on 12/11/2011)

⁶ I applied for a Toolserver account outlining my necessary database queries, usage profile as well as my affiliation with the Freie Universität Berlin. The application was submitted on 2011-12-21 and has not been processed yet (2012-01-23).

Some scripts that are already deployed can be accessed freely, allowing them to be reused. One of these was developed by SoNet⁷, a social networking research group. *link to wikitrif* It offers an API⁸ to get simple article statistics like article ID, text length as well as complex data structures like a list of unique editors including their gender if they are registered users and chose to reveal their gender in their Wikipedia account.⁹

4.1.5 Third-party sources/Web services

Some research projects can be reused as data sources. Depending on the project's goal, preprocessed statistics can become available:

ARTICLE TRAFFIC Wikipedia user Henrik¹⁰ provides a web service that processes Wikipedia's log files¹¹ to calculate the number page views per article for a given time. These statistics can be viewed through a browser¹² or via an API¹³.

CATSCAN This web service offered by Toolserver administrator Duesentrieb¹⁴ finds articles that belong to a given category and its sub-categories¹⁵. It also offers to limit the search to an intersection of categories, e.g. German politicians who are also physicists¹⁶. The results are presented in the browser or can be downloaded as a file containing comma-separated values (CSV format).

POOR MAN'S CHECKUSER *Username to IP resolution, link to georeferencing*

QUOVA *IP to location resolution, link to georeferencing*

WIKITRUST *annotations*

⁷ <http://sonetlab.fbk.eu/> (visited on 12/12/2011)

⁸ The API is documented here: <https://github.com/volpino/toolserver-scripts/tree/master/php> (visited on 12/12/2011)

⁹ Try http://toolserver.org/~sonet/api_gender.php?article=Egypt&lang=en (visited on 12/11/2011) to get a list of all registered users who edited the article *Egypt* of the English Wikipedia.

¹⁰ <http://en.wikipedia.org/wiki/User:Henrik> (visited on 12/12/2011)

¹¹ These are available at <http://dumps.wikimedia.org/other/pagecounts-raw/> (visited on 12/12/2011)

¹² E.g. http://stats.grok.se/en/201105/2011_Egyptian_Revolution (visited on 12/12/2011)

¹³ E.g. http://stats.grok.se/json/en/201105/2011_Egyptian_Revolution (visited on 12/12/2011)

¹⁴ <http://meta.wikimedia.org/wiki/User:Duesentrieb> (visited on 12/13/2011)

¹⁵ See [Categories](#) on why this is non-trivial.

¹⁶ [https://toolserver.org/~daniel/WikiSense/CategoryIntersect.php?wikilang=de&wikifam=.wikipedia.org&basecat=Politiker+\(Deutschland\)&basedeep=3&mode=cs&tagcat=Physiker&tagdeep=2](https://toolserver.org/~daniel/WikiSense/CategoryIntersect.php?wikilang=de&wikifam=.wikipedia.org&basecat=Politiker+(Deutschland)&basedeep=3&mode=cs&tagcat=Physiker&tagdeep=2) (visited on 12/13/2011)

4.2 AVAILABLE TOOLS

In the open source community a wide range of software tools are available. A simple search for “Wikipedia” on GitHub¹⁷, a source code exchange platform, shows a host of small software projects. These come in different programming languages and different feature sets and usually help in downloading articles in batches and extract data from big dump files. Developed by vigilantes and researchers alike, these programs facilitate both data retrieval and processing.

4.2.1 Toolkits

A group of openly available software packages qualify as swiss-army knives for processing dumps and sending requests to the API:

PYWIKIPEDIA *the mother of all python wp toolkits, classes for all mediawiki entities like page, user, revision*

PYMWDAT *relies on pywikipedia, dump file analysis, revisions, revert detection, filtering, helps in understanding*

LEVITATION *original project using git scm for blaming, my fork word level*

4.2.2 Analysis projects

This section describes useful projects, exemplary on how to combine different data sources...

WIKIPRIDE Python web-application¹⁸ to visualize contributions of groups of editors that registered in the same month.¹⁹

WIKI TRIP JavaScript application²⁰ that uses the Wikipedia API as well as a Toolserver backend, see [Toolserver](#), to visualize the evolution of a single article over time including: anonymous vs. registered contributors, male vs. female registered users, anonymous edits by country.²¹

4.3 APPLICATION DESIGN

¹⁷ <https://github.com/search?q=wikipedia&type=Repositories> (visited on 12/12/2011)

¹⁸ <https://github.com/declerambaul/WikiPride> (visited on 12/11/2011)

¹⁹ Project website: <http://meta.wikimedia.org/wiki/Research:WikiPride> (visited on 12/11/2011)

²⁰ <https://github.com/volpino/wikipedia-timeline> (visited on 12/11/2011)

²¹ Live demo: <http://sonetlab.fbk.eu/wikitrip/> (visited on 12/11/2011)

following wikitrip design of a browser application, based on the following technologies

4.3.1 Technologies

4.3.2 Models

4.3.3 Views

4.3.4 Operation

- *fetch article, wikitext and HTML*
- *parse for dates and location (delegate to subpage) and evaluate article relevancy*
- *fetch authors (toolserver)*
- *fetch revisions*
- *fetch first revision of each language*
- *fetch page views*
- *locate users*
- *fetch annotated source from WikiTrust for selected revisions*
- *render/store results*

group mode

- *fetch article list*
- *analyze all articles*
- *render group results*

4.4 ALGORITHMS

intro, complementary HTML and wikitext, HTML was easier offering annotations like dtstart

4.4.1 Article requirements

list requirements

4.4.2 Date parsing

link to Template

4.4.3 *Location parsing*

link to Template

4.4.4 *Collective authorship*

Introduce types of authors (roles) as well as methods to determine contribution/attribution

- *Autoren*
- *Bots*
- *text attribution, using wikitrust annotations, not for all revisions*

Are all edits relevant? Edit wars? Bots?

4.4.5 *Resolving user names to IPs*

- *registered vs. unregistered vs. bots vs. admins*
- *incorporate key findings of [20] as laid out in chapter 2.3*
- *IPs of unregistered users: Geo lookup*
- *Autoren-Profile: Information Extraction*
- *Geographische Zuordnung vom user profile*

4.4.6 *IP Look-up*

- *Services*
- *Accuracy*
- *Active prevention by proxies and anonymizers:*
J.A. Muir and P.C. van Oorschot. Internet geolocation and evasion. Tech. rep. Citeseer, 2006
J.A. Muir and P.C.V. Oorschot. „Internet geolocation: Evasion and counterevasion.“ In: ACM Computing Surveys (CSUR) 42.1 (2009), p. 4
M. Duckham and L. Kulik. „A formal model of obfuscation and negotiation for location privacy.“ In: Pervasive Computing (2005), pp. 152–170

4.4.7 Parsing user pages

- Automatic annotation of entities: <http://wikipedia-miner.cms.waikato.ac.nz/demos/annotate/>, also has services for categories. Alternative: <http://tagme.di.unipi.it/>
- IE approach with Machine Learning L. Xiao et al. „Information extraction from the web: System and techniques.“ In: Applied Intelligence 21.2 (2004), pp. 195–224
- unsupervised IE: O. Etzioni et al. „Unsupervised named-entity extraction from the web: An experimental study.“ In: Artificial Intelligence 165.1 (2005), pp. 91–134
- if city is mentioned, determine country (needs disambiguation, e.g. Berlin)
- coordinates are optional?

4.4.8 Geographic resolution

- settle for a country
- some examples on accuracy for different countries
- clustering of origins: areas of influence

4.5 VISUALIZATION

- A. Kjellin et al. „Evaluating 2D and 3D visualizations of spatiotemporal information.“ In: ACM Transactions on Applied Perception (TAP) 7.3 (2010), pp. 1–23
- Identify. Characteristics of an object.
- Locate. Absolute or relative position.
- Distinguish. Recognize as the same or different.
- Categorize. Classify according to some property (e.g., color, position, or shape).
- Cluster. Group same or related objects together.
- Distribution. Describe the overall pattern.
- Rank. Order objects of like types.
- Compare. Evaluate different objects with each other.
- Associate. Join in a relationship.
- Correlate. A direct connection.

4.5.1 *Maps*

- *Darstellung der geographischen Analyse*
- *per Wort, Satz, Artikel, Wort*

4.5.2 *Maps*

4.5.3 *Line and scatter charts*

4.5.4 *Motion chart*

4.6 HYPOTHESES ANALYSIS

for each hypothesis, what data is gathered, how is it crunched

H1

H2

H3

H4

H5

H6

H7

H8

H9

H10

H11

4.7 POSSIBLE ENHANCEMENTS

4.7.1 *Edit relevance*

detect reverts, vandalism

4.7.2 *Geographic profiling*

- M.D. Lieberman and J. Lin. „You are where you edit: Locating Wikipedia users through edit histories.“ In: ICWSM'09 (2009), pp. 106–113
- B.J. Hecht and D. Gergle. „On the localness of user-generated content.“ In: Proceedings of the 2010 ACM conference on Computer supported cooperative work. ACM. 2010, pp. 229–232
- from other fields such as criminal research:
B. Snook et al. „On the complexity and accuracy of geographic profiling strategies.“ In: Journal of Quantitative Criminology 21.1 (2005), pp. 1–26
- feasibility, maybe just as enhancer

5

EXPERIMENTS

5.1 DATA SETS

- *Using Wikipedia's category system and template embedders, link to foundation sections*
- *Choosing the "right" category*
- *Revolutions by country¹*
- *Can it be representative?*
- *Articles are categorized by people.*

5.1.1 *By category*

5.1.2 *By template*

5.1.3 *Political article vs. place article*

link to appendix with complete list

5.2 APPLICATION RUN

describe one run, time it takes, requests made, small article, big article, images

5.2.1 *Distribution*

5.2.2 *Text survival*

5.2.3 *Localness*

5.2.4 *Motion chart*

¹ http://en.wikipedia.org/wiki/Category:Revolutions_by_country

Part III

RESULTS

6

RESULTS

- *for each set, show hypothesis results, values*
- *box charts*

CONCLUSION

- *summarize method*
- *summarize results*
- *which hypotheses got confirmed?*
- *wikipedia as news medium vs history book*

7.1 LIMITATIONS

- *Mobile contributions, smartphones*
- *Privacy*

7.1.1 Political events

The same hypotheses may be applicable to other types of articles than political ones. The key requirements are that they have a location attribute and a time interval. This is easily fulfilled by disaster articles, e.g. Fukushima Daiichi nuclear disaster¹.

7.1.2 Article location

Although *location* is central to more abstract concepts like *Culture*² these subjects clearly defy being attributed with *a* location. Nevertheless, an analysis of the spatial distribution of contributors could be interesting.

7.1.3 Cross-language article growth

The growth rates of articles covering the same topic across various language editions could be analyzed to further investigate issues like language barrier—locals contributing only in their language—and information arbitrage as suggested by Adar, Skinner, and Weld [25].

7.2 FURTHER RESEARCH

¹ http://en.wikipedia.org/wiki/Fukushima_Daiichi_nuclear_disaster (visited on 10/31/2011)

² <http://en.wikipedia.org/wiki/Culture> (visited on 10/31/2011)

Part IV

APPENDIX

BIBLIOGRAPHY

- [1] The Economist. *The people formerly known as the audience*. URL: <http://www.economist.com/node/18904124> (visited on 08/10/2011).
- [2] The Economist. *Protest in Egypt: Another Arab regime under threat*. 2011. URL: <http://www.economist.com/node/18013760>.
- [3] sueddeutsche.de. *Krise in Ägypten - Die Kinder des 6. April und der Tag der Entscheidung - Politik*. 2011. URL: <http://www.sueddeutsche.de/politik/krise-in-aegypten-die-kinder-des-april-rufen-zum-protest-1.1053426>.
- [4] Wikipedia. *Revision history of 2011 Egyptian revolution*. URL: http://en.wikipedia.org/w/index.php?title=2011_Egyptian_revolution&dir=prev&action=history.
- [5] J. Giles. „Internet encyclopaedias go head to head.“ In: *Nature* 438.7070 (2005), pp. 900–901. ISSN: 0028-0836.
- [6] Wikipedia. *Wikipedia Page Views*. URL: <http://stats.wikimedia.org/EN/TablesPageViewsMonthly.htm> (visited on 08/10/2011).
- [7] A. Chadwick. *Routledge handbook of Internet politics*. Taylor & Francis, 2009. ISBN: 0203962540.
- [8] The Economist. *Libya: A civil war beckons*. 2011. URL: <http://www.economist.com/node/18290470>.
- [9] B. Suh et al. „Us vs. them: Understanding social dynamics in Wikipedia with revert graph visualizations.“ In: *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*. IEEE. 2007, pp. 163–170.
- [10] F.Å. Nielsen. „Wikipedia research and tools: Review and comments.“ In: (2011).
- [11] Wikipedia. *Statistics*. URL: <http://stats.wikimedia.org/EN/Sitemap.htm> (visited on 08/10/2011).
- [12] Wikipedia. *History of Wikipedia*. URL: http://en.wikipedia.org/wiki/History_of_Wikipedia (visited on 12/10/2011).
- [13] Wikipedia. *Protection policy*. URL: http://en.wikipedia.org/wiki/Wikipedia:Protection_policy (visited on 11/16/2011).
- [14] Wikipedia. *Why create an account?* URL: http://en.wikipedia.org/wiki/Wikipedia:Why_create_an_account%3F (visited on 08/10/2011).

- [15] Fabian Kaelin. *Research:Anonymous edits*. URL: http://meta.wikimedia.org/wiki/Research:Anonymous_edits (visited on 12/10/2011).
- [16] *Wikipedia Bots*. URL: <http://en.wikipedia.org/wiki/Wikipedia:Bots> (visited on 12/10/2011).
- [17] M. Kramer, A. Gregorowicz, and B. Iyer. „Wiki trust metrics based on phrasal analysis.“ In: *Proceedings of the 4th International Symposium on Wikis*. ACM. 2008, pp. 1–10.
- [18] B.T. Adler and L. De Alfaro. „A content-driven reputation system for the Wikipedia.“ In: *Proceedings of the 16th international conference on World Wide Web*. ACM. 2007, pp. 261–270.
- [19] B.T. Adler et al. „Assigning trust to wikipedia content.“ In: *Proceedings of the 4th International Symposium on Wikis*. ACM. 2008, pp. 1–12.
- [20] D. Hardy. „Volunteered geographic information in Wikipedia.“ PhD thesis. UNIVERSITY OF CALIFORNIA, SANTA BARBARA, 2011.
- [21] J.A. Muir and P.C.V. Oorschot. „Internet geolocation: Evasion and counterevasion.“ In: *ACM Computing Surveys (CSUR)* 42.1 (2009), p. 4.
- [22] M.D. Lieberman and J. Lin. „You are where you edit: Locating Wikipedia users through edit histories.“ In: *ICWSM’09* (2009), pp. 106–113.
- [23] B.J. Hecht and D. Gergle. „On the localness of user-generated content.“ In: *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM. 2010, pp. 229–232.
- [24] D.K. Lewis. *Philosophical papers*. Vol. 2. Oxford University Press, USA, 1987.
- [25] E. Adar, M. Skinner, and D.S. Weld. „Information arbitrage across multi-lingual Wikipedia.“ In: *Proceedings of the second ACM international conference on Web search and data mining*. ACM. 2009, pp. 94–103.
- [26] B. Hecht and D. Gergle. „The Tower of Babel meets Web 2.0: User-generated content and its applications in a multi-lingual context.“ In: *Proceedings of the 28th international conference on Human factors in computing systems*. ACM. 2010, pp. 291–300.
- [27] A. Kittur, E.H. Chi, and B. Suh. „What’s in Wikipedia?: mapping topics and conflict using socially annotated category structure.“ In: *Proceedings of the 27th international conference on Human factors in computing systems*. ACM. 2009, pp. 1509–1512.

- [28] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. „Studying cooperation and conflict between authors with history flow visualizations.“ In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. CHI '04. Vienna, Austria: ACM, 2004, pp. 575–582. ISBN: 1-58113-702-8. DOI: <http://doi.acm.org/10.1145/985692.985765>.
- [29] A. Kittur et al. „Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie.“ In: *World Wide Web* 1.2 (2007), p. 19.
- [30] J.A. Muir and PC van Oorschot. *Internet geolocation and evasion*. Tech. rep. Citeseer, 2006.
- [31] M. Duckham and L. Kulik. „A formal model of obfuscation and negotiation for location privacy.“ In: *Pervasive Computing* (2005), pp. 152–170.
- [32] L. Xiao et al. „Information extraction from the web: System and techniques.“ In: *Applied Intelligence* 21.2 (2004), pp. 195–224.
- [33] O. Etzioni et al. „Unsupervised named-entity extraction from the web: An experimental study.“ In: *Artificial Intelligence* 165.1 (2005), pp. 91–134.
- [34] B. Snook et al. „On the complexity and accuracy of geographic profiling strategies.“ In: *Journal of Quantitative Criminology* 21.1 (2005), pp. 1–26.
- [35] A. Kjellin et al. „Evaluating 2D and 3D visualizations of spatiotemporal information.“ In: *ACM Transactions on Applied Perception (TAP)* 7.3 (2010), pp. 1–23.