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

[January 25, 2012 at 20:32]

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:
LaTeX 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

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

Part I

THOUGHTS

1

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, MediaWiki[2], 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 article[3] as well as single articles covering the revolution in each of the affected countries, e.g. Egypt[4] and Libya[5]. 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 extend 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.

2 FOUNDATION

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 research[1] 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 predecessor[2].[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*

Wikipedia is operated by the Wikimedia Foundation, a non-profit organization, founded in Florida on June 20, 2003. It is completely financed by public contributions, such as donations and grants. Individual grants can be quite substantial — among the most generous donors are Google and the Stanton Foundation handing out respectively $2 million and $3.6 million in single donations.[13]

In the Wikimedia Foundation's 2010-11 fiscal year, $8.9 million was spent on website operations, including server hosting and software maintenance. The rest of the $20.0 million of total expenditures went into complementary activities such as fund raising, administration, and the support of local chapters.[14]

The local chapters are self-dependent organizations set up in countries around the globe to locally promote the foundation's cause and collect donations. The first local chapter to be created was Wikimedia Deutschland, founded in Berlin in 2004.[13]

The individual language editions of Wikipedia are not hosted by the local chapters, however. All of Wikipedia's content is centrally stored on servers in Tampa, Florida and in Amsterdam, Netherlands.[13]

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

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. Theses boxes show information in a structured way and can be found on articles 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°01′39″N 50°33′00″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. *Bahrain*, *find article*, instead of providing the geographic coordinates.

2.1.4 *Categories*

At the bottom of each article is an optional list of categories that the article belongs to, e.g. the article *2011-2012 Bahraini uprising* belongs, among others, to "Arab Spring by country"[4] and "2011 protests"[5].

Categories can not only have pages but also sub-categories, e.g. "Arab Spring by country" has the sub-category "2011 Libyan civil war"[6] which in turn has 5 sub-categories and 55 pages. The

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

4 http://en.wikipedia.org/wiki/Category:Arab_Spring_by_country (visited on 01/23/2012)

5 http://en.wikipedia.org/wiki/Category:2011_protests (visited on 01/23/2012)

6 http://en.wikipedia.org/wiki/Category:2011_Libyan_civil_war (visited on 01/23/2012)

categories do not form a tree, however, for there is no restriction on the inclusion of categories — even circular inclusions are possible. As liberal as the topology is inclusion of articles into a category. The category "Arab Spring by country" does not only contain articles covering the Arab Spring by country, but also articles about killed individuals, e.g. "Zakariya Rashid Hassan al-Ashiri"[7].

2.2 MEDIAWIKI AND EDITING

Anyone with a browser and internet access can edit Wikipedia's articles[8]. 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.[9] 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, e.g. adding an article to a category is as easy as putting the category name at the end of the wikitext.

2.2.1 *Templates*

Wikitext has a special syntax for templates[10]. These are reusable containers for text snippets and repetitive material like the info boxes described in Anatomy of an article. When a template is used in a page, the server software replaces the template placeholder — the template name surrounded by curly brackets — with the template's content. The content can be parameterized with key/value pairs so that, for example, an info box about countries can be used by several countries.

By using templates, the information is likely to be more structured than prose. Each invocation of a template also renders in the same way allowing for and encouraging more consistency[11]. More importantly, the usage of a template in an article is lets that article become a member of the group of articles that embed this template. This is an alternative mechanism to group

7 http://en.wikipedia.org/wiki/Zakariya_Rashid_Hassan_al-Ashiri (visited on 01/23/2012)

8 Some articles can be locked because of sustained vandalism or content disputes.[15]

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

10 http://en.wikipedia.org/wiki/Help:Template (visited on 01/23/2012)

11 E.g. when an editor notices that an info box supports the parameter "location" but does not have one yet, the user may feel encouraged to complete it.

articles that is likely to yield more homogenous results than the category system, see Categories.

2.2.2 *Revision history*

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.

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

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 be reverted. Especially in the case of vandalism this mechanism can be used to restore the previous state of the article. Reverts may also appear when different views on the same topic collide. To minimize the potential for *edit wars*[9] Wikipedia urges its users to discuss controversial topics on the article's talk page.

2.2.3 *Authors*

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.[16] The majority of edits comes from registered users, anonymous edits account for a quarter of all edits.[17]

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, either automatically or manually triggered by a real user.[18]

2.2.4 *User pages*

When a Wikipedia user decides to register, a *user page* is created for him on Wikipedia's website. This is a special page that can be edited like any other article. The user can publish personal information either in prose or reuse a template (see Templates).

The templates available to decorate one's user page include the following:

- Spoken languages, e.g. "This user is a native speaker of English."[12]
- Location, e.g. "This user comes from India."[13]
- Expression of personal views, e.g. "This user opposes Imperialism."[14]

Of course, a user can publish just about anything. As a result, the information on a user page has to be taken with a grain of salt.[15]

Like any other article, each user page has a discussion page that can be used to communicate with that user by leaving a message. According to Viégas et al. [19] these pages "hold much of the community interaction".

2.3 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 [20] 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.[16]

12 http://en.wikipedia.org/wiki/Wikipedia:Babel (visited on 01/23/2012)

13 http://en.wikipedia.org/wiki/Template:User_India (visited on 01/23/2012)

14 http://en.wikipedia.org/wiki/User:Serouj/UserBox/Against_Imperialism (visited on 01/23/2012)

15 See user Lihaas, who seems to hail both from India and Pakistan: http://en.wikipedia.org/wiki/User:Lihaas (visited on 01/23/2012)

16 Kramer, Gregorowicz, and Iyer [20] 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.

A similar approach of calculating the longevity of text chunks was followed by Adler and De Alfaro [21]. They adapted standard 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. [22] which offers an API[17] 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.[18]

2.4 GEOREFERENCES

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 [23] 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 [24]. Various visualizations[19] [20] of edit distributions use geolocation databases like MaxMind[21] and Quova[22].

For registered users, the IP address is not stored with the revision. Therefor IP geolocation services cannot be used. Lieberman and Lin [25] 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 [26].

17 http://www.wikitrust.net/vandalism-api (visited on 10/31/2011)
18 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).
19 http://infodisiac.com/blog/2011/05/wikipedia-edits-visualized/ (visited on 10/31/2011)
20 http://sonetlab.fbk.eu/wikitrip/ (visited on 10/31/2011)
21 http://www.maxmind.com (visited on 10/31/2011)
22 http://www.quova.com (visited on 10/31/2011)

3

HYPOTHESES

To gain insight on how Wikipedia is being used during and after political events[1], 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 [27, 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 be 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, therefor 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 [29] 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 [23, 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 sights[5] will dominate this sample. Since my thesis is concerned with political events, I find this point to be worth revisiting[6] 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' proximity[7].

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 [28] found that between two languages the interlinking 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 [30] articles about *geography and places* are third biggest group.

6 In addition, Hardy [23, p. 61] considered only anonymous users. Since creating an article is only allowed for registered users, his method has to be extended.

7 Hardy [23, 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 [31] 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 sample[10] merits further investigation.

In 2007, Kittur et al. [32] 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. [32] 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 [31] 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 [31] 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 [31] 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

4

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. 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*

For a complete article analysis, navigating the website can be tedious as one would have to click through a complete revision history and parse the page's source which is formatted using the HyperText Markup Language (HTML). However, individual pages contain data that is static and can be used throughout the analysis process. This makes it worth writing a specific parser for a technique known as *screen scraping* to extract the information. On a high level, it involves the following steps:

1. Looking at the HTML source of the page and identifying how the HTML tags and attributes that are used to structure the information.
2. Writing a parser that addresses the identifying tags and thereby tokenizes the data.
3. Converting the found tokens into an output format, e.g. JSON.

For a simple HTML table, a parser can be written in a few lines of code. Using this technique, the following static information was gathered:

BOTS A list of bots was built based on the Wikipedia page *List of bots by number of edits*[1]. This list is used to distinguish bots from real authors as contributions done by bots are excluded from the analysis. There are unregistered bots, however, that appear not in the list. For a lack of automated distinction, these are counted as normal authors.[2]

COUNTRIES A list of countries was extracted from the article *ISO_3166-1*[3]. It provides a list of standardized country names that is also respected by Wikipedia's authors when referring to a country by name. In a second pass, the Wikipedia article of each country was parsed for coordinates.[4]

Making both of these sets static was a design decision recognizing the trade-off between having them in memory and querying for each article.

4.1.2 *Database dumps*

Monthly database snapshots of all wikis run by the Wikimedia Foundation, including Wikipedia, are publicly available[5] 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.[6] This huge size makes processing them rather slow.[7] When analyzing only a single article or a category articles,

1 http://en.wikipedia.org/wiki/Wikipedia:List_of_bots_by_number_of_edits (visited on 01/24/2012)

2 A simple heuristic employed by other software to analyze MediaWiki content is treating all contributors whose username contains or whose comments start with "bot" as a bot, e.g. pymwdat, see http://code.google.com/p/pymwdat/source/browse/trunk/toolkit.py?spec=svn13&r=13 (visited on 01/24/2012). This has a potential for false positives and is not used.

3 http://en.wikipedia.org/wiki/ISO_3166-1 (visited on 01/02/2012)

4 For some countries, coordinates were not present on the page, e.g. http://en.wikipedia.org/wiki/Australia (visited on 01/24/2012). In that case, they were manually added by using that country capital's coordinates. For a discussion on geographic resolution, see Geographic resolution.

5 http://dumps.wikimedia.org (visited on 12/11/2011)

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

7 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).

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[8] which can be used by other programs to remotely use the wiki's features such as changing content and restoring revisions.[9] For analysis of articles, the API offers queries directed at a variety of article properties, e.g. revisions, categories and links. Among the output formats for the responses is the JavaScript Object Notation (JSON). Similar to MediaWiki's Special:Export page[10], the API also offers an article export that includes all revisions.

Listing 1: Example JSON response to a query to list all bots that edited the article *2011-2012 Bahraini uprising*

```
{
    "query": {
        "redirects": [{
                "from": "2011 Bahraini uprising",
                "to": "2011-2012 Bahraini uprising"
        }],
        "pages": {
            "30876395": {
                "pageid": 30876395,
                "ns": 0,
                "title": "2011-2012 Bahraini uprising"
            }
        },
        "allusers": [{
                "userid": "13146235", "name": "28bot"
            }, {
                "userid": "5415725", "name": "718 Bot"
            }, ..., {
                "userid": "13770078", "name": "AWBCPBot"
        }]
    },
    "query-continue": {
        "allusers": {
            "aufrom": "AWeenieBot"
        }
    }
}
```

Some of the queries have a limit on how many results they return on a single request. When there are more results, the response contains a *query-continue* attribute that can be sent with

8 http://www.mediawiki.org/wiki/API:Main_page (visited on 01/24/2012)

9 The full capability of the API can be seen at tried at the *Sandbox* at https://en.wikipedia.org/wiki/Special:ApiSandbox (visited on 01/24/2012), a recent addition to the MediaWiki software.

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

following query so that the next result set can be returned. The following API calls will be important for this thesis:

QUERY INFO This basic query is returns essential information like the article ID, the last revision ID, but also the full wikitext of the last revision.

QUERY REVISIONS Lists all revisions for an article and for each includes a timestamp and the user as well as the comment for the text change.

QUERY CATEGORYMEMBERS For a given category, this query lists the articles and subcategories belong to it. This query can be used to construct groups of articles for analysis in this thesis.

QUERY TEMPLATEEMBEDDERS For a given template name, this request lists all pages that embed it. This query can also be used to build a group of articles.

OPEN SEARCH A method to suggest articles, categories and templates that contain a term. It can be attached to an input field where a user is supposed to enter the name of an article.

PARSE This special query returns the HTML version of the article's wikitext. The content that is returned is exactly the HTML source that is sent to a browser when a visitor looks at this article or user page. This query will be used in cases where it is easier to parse the HTML markup than the wikitext, e.g. the User pages.

4.1.4 *Toolserver*

The Germany based Wikimedia Deutschland e.V. runs Toolserver[11], 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.[12]

Some scripts that are already deployed can be accessed freely, allowing them to be reused. One of these was developed by SoNet[13], a social networking research group based in Italy, for a

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

12 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).

13 http://sonetlab.fbk.eu/ (visited on 12/12/2011)

project called WikiTrip (see Analysis projects). It offers an API[14] 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.[15]

In effect, calling the SoNet API replaces several calls to the original MediaWiki API and therefor speeds up the information retrieval, especially when the number of revisions or authors is high. The returned data object has the following structure:

Listing 2: SoNet API response to a query for the article *2011-2012 Bahraini uprising*

```
{
    "first_edit": {"timestamp":1297734917,"user":"Master&Expert"},
    "count":1778,
    "minor_count":401,
    "count_history":{"today":3,"week":5,"month":90,"year":1778},
    "last_edit":1327370324,
    "totaldays":0,
    "average_days_per_edit":"0.00",
    "edits_per_month":0,
    "edits_per_year":0,
    "edits_per_editor":"4.17",
    "editor_count":426,
    "anon_count":337,
    "editors": {"Bahraini Activist":
        {"all":106,"minor":21,"first":"17 May 2011, 09:45:25",
        "last":"22 January 2012, 10:52:50","atbe" 203811,
        "minorpct":"19.81", "size":"140.54","urlencoded":"
            Bahraini_Activist"},
        ...
    },
    "anons":{"2011-02-15T08:11:52Z":
        ["78.2.29.139","Rovinj Croatia",45.08,13.64],
        ...
    }
}
```

This high density of preprocessed information shows the power of the Toolserver and its direct access to the database. The property *editors* lists all unique authors of an article and their edit count (property *all*). The second exhaustive collection is under the property *anons*. There, all anonymous contributors are listed with their IP address, their geographic region and coordinates. The geographic lookup is uses[16] the GeoCityLite database from Maxmind (see IP Look-up for a discussion).

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

15 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.

16 https://github.com/volpino/toolserver-scripts/blob/master/php/api.php (visited on 01/24/2012)

4.1.5 *Third-party sources/Web services*

Like the Toolserver scripts in the previous section, other research projects exist that can be reused as data sources. Depending on the project's goal, a variety of preprocessed data is available:

ARTICLE TRAFFIC Wikipedia user Henrik[17] provides a web service that processes Wikipedia's log files[18] to calculate the number page views per article for a given time. These statistics can be viewed through a browser[19] or queried through an API [20].

CATSCAN This web service, offered by Toolserver administrator Duesentrieb[21], finds articles that belong to a given category and its sub-categories (see Categories on why this is non-trivial). It also offers to limit the search to an intersection of categories, e.g. German politicians who are also physicists[22]. The results are presented in the browser or can be downloaded as a file containing comma-separated values (CSV format).

POOR MAN'S CHECKUSER The project *Poor Man's Check User*[23] mapped registered users to IP addresses based on a bug in the session management of the MediaWiki software.[24] For the period the bug has been active, some usernames could be mapped. Naturally, the more edits a user did in this period, the more likely is an appearance in this list. For

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

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

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

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

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

22 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)

23 Project website: http://wikiwatcher.virgil.gr/pmcu (visited on 01/02/2012). The project's name is a reference to the *checkuser* permission that a community-elected group of registered users possesses. It allows a de-masking of the IP addresses for each of a registered user's edit.

24 When a user exceeded a certain time while editing an article without submitting the current changes, the user's session expired on the server. When the edit was submitted after the expiration the user appeared as an anonymous author, being only known by his IP address. When the user then logged in again, the same change was sent again. Scanning all revisions for the same change set therefor allowed for a matching between user name and IP address. This loophole has been closed, however.

the purpose of this thesis, I screen-scraped this table and condensed[25] it to 14,171 unique users.

QUOVA This geo-location web service maps an IP address to a geographic location, see Georeferences.

WIKITRUST Based on Adler et al. [22] an open source online reputation system[26] was set up by the University of California, Santa Cruz, to allow for easy vandalism detection (see Contributions). Given an article ID and a revision ID, the API method *wikimarkup* returns an annotated version of the wikitext of that revision. An annotation consists of a trust value, the revision ID the text got introduced into the article as well as the authors user name or IP address, e.g. revision 473029564 of the article *2011-2012 Bahraini uprising*[27]:

Listing 3: Excerpt of the annotated markup for the revision 473029564 of the article *2011-2012 Bahraini uprising*

```
{{#t:7,468889105,Kudzu1}}The
{{#t:7,470041169,Happysailor}}2011-2012
{{#t:8,413989516,Master&Expert}}Bahraini
{{#t:8,427545590,Kudzu1}}uprising, sometimes called the
{{#t:9,455029613,Sitrawi86}}February 14 Revolution
```

All wikitext following an annotation, up to the next one, was written by that author. The web service provider implemented a custom diff algorithm for the attribution of authorship. This was needed to overcome wiki-specific issues and to maximize tracking, e.g. for text that is removed and re-inserted at a later revision.[28]

4.2 AVAILABLE TOOLS

To process the data from all the data sources, a wide range of software tools are available in the open source community. A simple search for "Wikipedia" on GitHub[29], a source code exchange platform, shows a multitude of small software projects. These come in different programming languages and different

25 Some usernames have multiple entries as each occurrence of the bug created a unique "evidence". Among those, some have been manually verified and ranked. When multiple entries exist, my algorithm picks the top ranked.

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

27 http://en.collaborativetrust.com/WikiTrust/RemoteAPI?method=wikimarkup&pageid=30876395&revid=473029564 (visited on 01/23/2012)

28 http://www.wikitrust.net/frequently-asked-questions-faq#TOC-On-text-author-and-origin (visited on 01/24/2012)

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

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[30] qualify as swiss-army knifes for processing and analyzing Database dumps:

PYWIKIPEDIA As the mother of all Python toolkits, the Python wikipedia robot framework[31] offers an extendable set of classes for all MediaWiki entities like a page, a user, revision, etc, and is typically used to write a bot program for automated editing tasks (see Authors).

PYMWDAT Based on PYWIKIPEDIA, this toolkit offers a convenient downloader for all revisions of an article as well as an extensible dump file analyzer with support for filtering and revert detection.

LEVITATION A creative project to turn Database dumps into a Git[32] repository. As a source code management system, Git offers a more space efficient way to store the sequence of revisions of the articles, since it only stores the difference between revisions. Once converted to a repository, moving from revision to revision is much faster than processing the large dump files. Git's diff mechanism together with its *blame* command can be used as an alternative way to attribute authorship to passages of article content.[33]

4.2.2 *Analysis projects*

In addition to the toolkits, a handful of research projects exist that process Wikipedia's content. These are purpose built applications that have a much narrower focus but are very skillful in combining and using different data sources such as MediaWiki API, the Toolserver or other Third-party sources/Web services:

30 Although none of these were used for the content analysis of this thesis, their study proved very insightful on how to process Wikipedia's content.

31 http://pywikipediabot.sourceforge.net/ (visited on 01/24/2012)

32 http://git-scm.com/ (visited on 01/24/2012)

33 In fact, git operates on a line level, making the attribution rather coarse. To get blaming functionality on a word level, I patched the source, see my fork at https://github.com/davkal/levitation/commit/5fca0001d26cb67fde6ff9d8a5f2b1414cf7681e (visited on 01/24/2012).

WIKIPRIDE Python web-application[34] to visualize contributions of groups of editors that registered in the same month.[35]

WIKI TRIP JavaScript application[36], written at SoNet, that uses the MediaWiki API as well as its own Toolserver scripts, to visualize the evolution of a single article over time including: anonymous vs. registered contributors, male vs. female registered users, anonymous edits by country.[37]

4.3 APPLICATION DESIGN

For the analysis of articles I developed an application that could draw data from the different sources and process them in a timely fashion. Following the impressive WikiTrip application design (see Analysis projects) I decided to build a web application that runs entirely in a web browser.[38]

4.3.1 *Technologies*

As a web application the software heavily relies on HTML version 5, JavaScript (JS) and Cascading Stylesheets (CSS). It uses a range of open-source toolkits and libraries for a variety purposes:

BOOTSTRAP Twitter's web application toolkit[39] controls the basic layout, the styling of sections and form fields, and the navigation bar at the top.

JQUERY The JS library jQuery[40] is used to asynchronously retrieve data from the various data sources, dynamically insert elements into the layout as well as parsing screen-scraped webpages by making use of its selectors for addressing elements. The AUTOCOMPLETE[41] widget of jQuery's UI library is used to display article suggestions based on what has been entered into the article search field.

UNDERSCORE Underscore[42] is a "utility-belt library" for functional programming in JS. Its *map* and *groupBy* methods are being heavily used in the analysis of content.

34 https://github.com/declerambaul/WikiPride (visited on 12/11/2011)
35 Project website: http://meta.wikimedia.org/wiki/Research:WikiPride (visited on 12/11/2011)
36 https://github.com/volpino/wikipedia-timeline (visited on 12/11/2011)
37 Live demo: http://sonetlab.fbk.eu/wikitrip/ (visited on 12/11/2011)
38 Application development was done mainly using the Google Chrome browser. The application should run in any HTML5-capable browser.
39 http://twitter.github.com/bootstrap/ (visited on 01/25/2012)
40 http://jquery.com/ (visited on 01/25/2012)
41 http://jqueryui.com/demos/autocomplete/ (visited on 01/25/2012)
42 http://documentcloud.github.com/underscore/ (visited on 01/25/2012)

BACKBONE Built upon Underscore, Backbone[43] provides a way to structure a JS application. All models, collections and views in the application are encapsulated by Backbone objects that can communicate with each other via events.

DATEJS Date.js[44] is a library that was being used to parse dates in articles about events.

LZ77 A JS implementation[45] of the LZ77 text compression algorithm[46]. It is used to shrink the size of results when stored in the browser's limited data store[47].

D3.JS Data-driven documents (d3)[48] is a library to visualize big data sets. This application uses d3's box plots[49] to summarize the quantitive distribution of analysis results.

GOOGLE CHART TOOLS Google's[50] chart tools[51] provide a wide range of chart types that I used in the application, including line chart, scatter chart and motion chart.

4.3.2 *Models*

Looking at Wikipedia's website as a system, visible items like pages, authors and revisions can be abstracted into classes of objects that can be modeled using BACKBONE (for a complete overview of all models, see *model chart in appendix*).

When a model is instantiated, it knows where to retrieve the data that will populate its attributes, e.g. a revision collection knows that it can get all revisions of an article's history using the query *revisions* from the MediaWiki API. Once the data is retrieved, the attributes of each individual revision object are set which in turn triggers an event, telling a single revision model fetch the annotated wikitext from the WikiTrust API (see Third-party sources/Web services). While some events trigger the retrieval of more detailed information, others indicate that a model's data is fully populated ready for display in the application.

43 http://documentcloud.github.com/backbone/ (visited on 01/25/2012)
44 http://www.datejs.com/ (visited on 01/25/2012)
45 https://github.com/olle/lz77-kit (visited on 01/25/2012)
46 http://en.wikipedia.org/wiki/LZ77_and_LZ78 (visited on 01/25/2012)
47 See http://dev.w3.org/html5/webstorage/ (visited on 01/25/2012) for the limitations.
48 http://mbostock.github.com/d3/ (visited on 01/25/2012)
49 http://mbostock.github.com/d3/ex/box.html (visited on 01/25/2012)
50 Although not open source, some of the charts are developed by the open-source community.
51 http://code.google.com/apis/chart/interactive/docs/index.html (visited on 01/25/2012)

4.3.3 *Views*

The rendering of models is encapsulated in views. Following the publish-subscribe pattern, a view listens to changes in a certain model. When a change event is observed, the view renders itself. For example, the *article view* is rendered multiple times because it draws data from different sources (the page ID is available sooner than, say, the first sentence of the article). Most views, however, rely on several models as they analyze various aspects of an article and then display the results in a chart (see Visualization).

4.3.4 *Main routine*

This section describes an application run on a high level, for noteworthy algorithms for the sub-routines, see Algorithms.

When loaded in a browser, the application expects an article title, a category or a template name as input. In the case of an article being entered, a click on "Analyze!" starts following routine, querying the various Data sources:

1. QUERY INFO is being called to check if the article title is a valid title. A successful query also returns the wikitext of the article's latest revision.
2. A call to PARSE retrieves the HTML version of the latest revision as well as the article's links to other language editions. Both the wikitext and the HTML are then being parsed for location and dates.
3. The SoNet Toolserver script is called to retrieve all authors of the article.
4. For all registered usernames in the author collection, a subroutine is started to locate the users.
5. QUERY REVISIONS is being called to retrieve all revisions (excluding wikitext).
6. The a subset of the revisions, the annotated wikitext is loaded using WikiTrust API.
7. For each day of the article's existence, the page view statistics are loaded.
8. For each language present in the language link collection, the first revision is loaded.
9. All loaded data is analyzed and the results are stored and rendered in the browser.

Depending on the article's number of revisions, this process takes around one minute and involves 100–500 API calls, most of which are done in parallel. During this process, data is rendered whenever it has been arrived and processed. However, when a category or template name has been entered, only the article overview will be rendered while still all of the above steps are being taken. On a high level, this *group mode* works as follows:

1. Fetch a list of articles by calling QUERY CATEGORYMEMBERS or QUERY TEMPLATEEMBEDDERS.
2. Run the analysis routing on all articles.
3. Compute group results and show them in the browser.

4.4 ALGORITHMS

The individual algorithms described mostly deal with extracting data from the various Data sources. To classify an article as treating an event, the article had to be parsed for a location and a date or date interval. All algorithms will be written in a Python-esque pseudo-code.

4.4.1 *Article requirements*

The presence of a date and a location is a direct requirement for an article to qualify for further analysis when the application is in group mode. All of the following criteria have to be satisfied:

- The article has a location in the form of geographic coordinates or by name, e.g. Cairo.
- The article has a date, e.g. 9 November 1989, or a date interval, e.g. May 2007 - August 2007.
- The article was created after the event started (with a 3 day tolerance) to sort out events that have been scheduled.
- The event (start) date is not before 2002 to make sure that Wikipedia was available as a medium.
- The article does not use certain templates or is part of certain categories, e.g. *Category:Living People*. This has been included for the purpose of filtering out articles that passed the previous tests but a clearly not an event article.

Even when all of the requirements above are met, some articles may not have a full set of results after their analysis. They will however be included in the computations for which results are present. *link to group analysis*

4.4.2 *Date parsing*

The way dates are mentioned in articles are as diverse as the people that write them. And when it comes to date intervals, e.g. May 8–12 2007, even a specialized date parsing library like DATE.JS can only be of limited use. I wrote a custom parser that for a given text returns first occurrence of an interval or a single date, i.e. if no interval could be found, the parsing is repeated for a single date.

Some articles embed info box templates (see Templates) that produce annotated markup using the HCARD[52] microformat. The annotations can be easily addressed with CSS selectors[53] and the values conform to the standard for the representation of date and time, ISO 8601[54]. In listing 4 the first block tries to parse the microformat annotations. When they are not present, the first info box is checked for a date field. As a last resort, the first sentence and paragraph are scanned for dates to not filter out eligible articles like *2007 Georgian demonstrations*[55]

Listing 4: Date candidates algorithm

```
if 'dtstart' in HTML:   # checking for hcard
    start = datejs.parse(dtstart.text)
    if 'dtend' in HTML:
        end = datejs.parse(dtend.text) # proper interval
    elif 'ongoing' in dtstart.next.text:
        end = today() # ongoing event
    else:
        end = start + 1 # single day event
elif 'date' in templates.infoboxes[0]:
    start, end = custom.parse(templates.infoboxes[0].date) # wikitext
if not start:
    start, end = custom.parse(article.first_sentence) # HTML
if not start:
    start, end = custom.parse(article.first_paragraph) #HTML
```

The custom parser then checks for a range of formats including lazy ones like December 14–19, 2008. Over a long iterative

52 http://en.wikipedia.org/wiki/HCard (visited on 01/25/2012)

53 This requires the template to be properly programmed. Some templates mistakenly mark start and end of an interval with the same annotation: `dtstart`. On other pages the error lies with the authors who misused the date template, e.g. "Municipal Library Elevator Coup" happened on 28 January 1908 which was added as `Start date|1908|28|01` to the info box but unpredictably rendered as April 1, 1908 possibly due to a mixup in the order, see http://en.wikipedia.org/wiki/Municipal_Library_Elevator_Coup (visited on 01/25/2012).

54 http://en.wikipedia.org/wiki/ISO_8601 (visited on 01/25/2012)

55 The article does not contain an info box and a date is only mentioned in the second sentence: "The demonstrations peaked on November 2, 2007,..." with more dates to follow in the same paragraph. This particular example already shows how idiosyncratic dates can be codified. See http://en.wikipedia.org/wiki/2007_Georgian_demonstrations (visited on 01/25/2012).

process, I identified the following tokens from which to construct the date patterns as regular expressions:

Listing 5: Date tokens

```
ords = ['th', 'st', 'nd', 'rd'];
tokens = {
    M: "({0})".format(MonthNames.join('|')), # months
    D: "(\d{1,2})" + "({0})?".format(ords.join('|')), # day
    Y: "(\d{4})", #year
    T: "(/|-|--|\sto\s|\sand\s)", # interval delimiter
    O: "('*ongoing'*?|'*present'*)", # ongoing event
    F: "From", # ongoing event
    S: "[,\s]*", # whitespace
    P: "\|", # pipe
    A: "([^- --]*)" # other text
};
```

Using these tokens, I produced patterns to match all encountered date formats, e.g. the pattern for December 14–19, 2008 is `MDTDY`, or to capture the even less conform "From 15 October 2011" (meaning the event is ongoing) the pattern is `FDMY`. The patterns are ranked by accuracy so that "12 May 2001 - present" is matched before "May 2001 - present". The coarsest pattern to match is `Y`, a single year (4 digits).

4.4.3 *Location parsing*

Like dates, locations and even coordinates can be codified in numerous ways. Most of the coordinate templates used in the info boxes produce annotated markup, thereby making machine-readable. I still wrote a custom parser for all the cases where the marker (`geo`) is not produced.

Some event articles however, do not have coordinates are clearly describing actions at a location, e.g. *Maspero demonstrations*[56]. From articles like these, location candidates are scraped and then resolved:

Listing 6: Article's locate algorithm

```
function locate(article):
    if 'geo' in article.HTML: # checking for machine-readable coords
        location = custom.parse(article.geo.text)
        return
    else: # look for location candidates
        candidates = []
        # all links from the first info box's location field (wikitext)
        candidates.extend(templates.infoboxes[0].location.links)
        # all flags in the first info box
        candidates.extend(templates.infoboxes[0].flags) #HTML
        # all links from the first paragraph
        candidates.extend(article.first_paragraph.links) # HTML
    if len(candidates) and article.isMainArticle:
```

56 http://en.wikipedia.org/wiki/Maspero_demonstrations (visited on 01/25/2012)

```
14          candidates = candidates[:10]
15          until location:
16              location = locate(retrieve(candidates.pop()))
17          return location
```

The above mentioned *Maspero demonstrations* article is exemplary for the candidate list mechanism. Its info box's location field offers three links to articles of a place: Maspiro, Cairo, Egypt. They are checked until an article with coordinates is found, in this case after the second try, Cairo[57].

4.4.4 *Collective authorship*

Most of the authorship processing is being done by SoNet's Toolserver script[58], see Toolserver. The PHP-script directly queries a live copy of the Wikipedia database for all revisions of the requested article. For each revision's author, an edit counter is incremented, and if the author was anonymous, the IP is being resolved to a geographic location. A collection of authors and all resolved locations are then returned in a JSON object.

From the response, my application then creates an author collection. If an author is a registered bot, the author is excluded. From the location list, a second collection is created to manage all author locations for the current article.

The attribution of text passages to authors is done by the web service WikiTrust, see Third-party sources/Web services. It returns an annotated markup that can easily be parsed with the following regular expression (JS):

```
/{{#t:\d+,\d+,[^}]*}}/g;
```

Barring bots, all edits are considered relevant, i.e. reverts or blanking[59] are not treated in a special way.

4.4.5 *Resolving user names to IPs*

- *registered vs. unregistered vs. bots vs. admins*
- *incorporate key findings of [23] as laid out in chapter 2.4*
- *IPs of unregistered users: Geo lookup*
- *Autoren-Profile: Information Extraction*
- *Geographische Zuordnung vom user profile*

57 http://en.wikipedia.org/wiki/Cairo (visited on 01/25/2012)

58 https://github.com/volpino/toolserver-scripts/blob/master/php/api.php (visited on 01/25/2012)

59 The illegitimate removal of all content of an article, see http://en.wikipedia.org/wiki/Wikipedia:VANDTYPES#Types_of_vandalism (visited on 01/25/2012)

4.4.6 *IP Look-up*

- *Services*
- *Accuracy*
- *Active prevention by proxies and anonymizers:*
  J.A. Muir and PC 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 gathers, 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 country[1]
- *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*

7

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 disaster[1].

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 [28].

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] *Wikimedia Foundation*. URL: http://en.wikipedia.org/wiki/Wikimedia_Foundation (visited on 01/24/2012).

[14] *Wikimedia Foundation Financial Report 2010-11*. URL: https://upload.wikimedia.org/wikipedia/commons/4/48/WMF_AR11_SHIP_spreads_15dec11_72dpi.pdf (visited on 01/24/2012).

[15] Wikipedia. *Protection policy*. URL: http://en.wikipedia.org/wiki/Wikipedia:Protection_policy (visited on 11/16/2011).

[16] Wikipedia. *Why create an account?* URL: http://en.wikipedia.org/wiki/Wikipedia:Why_create_an_account%3F (visited on 08/10/2011).

[17] Fabian Kaelin. *Research:Anonymous edits*. URL: http://meta.wikimedia.org/wiki/Research:Anonymous_edits (visited on 12/10/2011).

[18] *Wikipedia Bots*. URL: http://en.wikipedia.org/wiki/Wikipedia:Bots (visited on 12/10/2011).

[19] F.B. Viégas et al. „Talk Before You Type: Coordination in Wikipedia." In: *Proceedings of HICSS*. Vol. 40. 2007.

[20] 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.

[21] 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.

[22] B.T. Adler et al. „Assigning trust to wikipedia content." In: *Proceedings of the 4th International Symposium on Wikis*. ACM. 2008, pp. 1–12.

[23] D. Hardy. „Volunteered geographic information in Wikipedia." PhD thesis. UNIVERSITY OF CALIFORNIA, SANTA BARBARA, 2011.

[24] J.A. Muir and P.C.V. Oorschot. „Internet geolocation: Evasion and counterevasion." In: *ACM Computing Surveys (CSUR)* 42.1 (2009), p. 4.

[25] M.D. Lieberman and J. Lin. „You are where you edit: Locating Wikipedia users through edit histories." In: *ICWSM'09* (2009), pp. 106–113.

[26] 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.

[27] D.K. Lewis. *Philosophical papers*. Vol. 2. Oxford University Press, USA, 1987.

[28] 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.

[29] B. Hecht and D. Gergle. „The Tower of Babel meets Web 2.0: User-generated content and its applications in a multilingual context.“ In: *Proceedings of the 28th international conference on Human factors in computing systems*. ACM. 2010, pp. 291–300.

[30] 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.

[31] 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.

[32] 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.

[33] J.A. Muir and PC van Oorschot. *Internet geolocation and evasion*. Tech. rep. Citeseer, 2006.

[34] M. Duckham and L. Kulik. „A formal model of obfuscation and negotiation for location privacy.“ In: *Pervasive Computing* (2005), pp. 152–170.

[35] L. Xiao et al. „Information extraction from the web: System and techniques.“ In: *Applied Intelligence* 21.2 (2004), pp. 195–224.

[36] O. Etzioni et al. „Unsupervised named-entity extraction from the web: An experimental study.“ In: *Artificial Intelligence* 165.1 (2005), pp. 91–134.

[37] 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.

[38] B. Snook et al. „On the complexity and accuracy of geographic profiling strategies.“ In: *Journal of Quantitative Criminology* 21.1 (2005), pp. 1–26.