

Analysis of the social network amongst artists on Wikipedia

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Thomas Claus Solich, BSc

Matrikelnummer 1027433

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Mag.rer.soc.oec. Dr.rer.soc.oec. Dieter Merkl

Wien, 8. August 2016

Thomas Claus Solich

Dieter Merkl

Analysis of the social network amongst artists on Wikipedia

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Thomas Claus Solich, BSc

Registration Number 1027433

to the Faculty of Informatics

at the Vienna University of Technology

Advisor: Ao.Univ.Prof. Mag.rer.soc.oec. Dr.rer.soc.oec. Dieter Merkl

Vienna, 8th August, 2016

Thomas Claus Solich

Dieter Merkl

Erklärung zur Verfassung der Arbeit

Thomas Claus Solich, BSc
Senefeldergasse 26/8, 1100 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 8. August 2016

Thomas Claus Solich

Acknowledgements

I would like to show my gratitude to Dieter Merkl for the interesting topic as well as his continuous feedback and valuable support. His suggestions often inspired me for new ideas and to look at things from another perspective.

I would also like to thank my family for their believes in me. I owe my deepest gratitude especially to my parents, Maria and Robert, for their motivation and loving support in every situation. Without them, I would not have come this far. Finally, I want to thank Laura for her love and patience through all the years.

Kurzfassung

Das Internet spielt als Wissensquelle eine immer wichtigere Rolle. Wikipedia, eine der größten frei zugänglichen Enzyklopädien, mit zahlreichen Artikeln in verschiedenen Sprachen, ist für viele Menschen ein wesentlicher Bestandteil bei der Internet-Recherche geworden. Da grundsätzlich jede Person Wikipedia Artikel erstellen und editieren kann, sind diese so verschieden wie die Autoren selbst. Abhängig von den jeweiligen Sprachversionen werden Artikel zu verschiedenen Themen bzw. Artikel zum gleichen Thema mit eventuell anderem Inhalt erstellt. Somit stellt sich die Frage, wie vollständig und verlässlich die auf Wikipedia verfügbaren Artikel und Informationen sind.

Im Rahmen dieser Arbeit wurde analysiert, wie groß die Unterschiede zwischen den Sprachversionen der einzelnen Wikipedia Artikel sind bzw. wie vollständig Artikel im Bezug auf Beziehungsinformationen dargestellt sind. Aufgrund der enormen Anzahl von Artikeln wurde nicht die gesamte Enzyklopädie überprüft, sondern nur Artikel aus dem Bereich Kunst in den drei Sprachversionen, Deutsch, Englisch und Italienisch. Bei den relevanten Artikeln wurden die Unterschiede zwischen den Sprachversionen im Bereich sozialer Netzwerke von Künstlern quantifiziert und verglichen. Es wurde ebenfalls überprüft, ob nationale Künstler in den jeweiligen Wikipedia-Sprachversionen bevorzugt behandelt werden. Um die Vollständigkeit der Artikel zu verifizieren, wurden Beziehungen die in Form von Links in Artikeln enthalten sind, mit Informationen aus zwei anderen Quellen abgeglichen. Diese Quellen sind die Union List of Artist Names (ULAN) sowie Wikidata. ULAN ist ein Verzeichnis welches Einträge über Künstler, kunstverwandte Personen und Objekte enthält. Wikidata ist eine Datenquelle für verschiedene Wikimedia Projekte. Daraus wurden z.B. Beziehungsinformationen und andere, für die Analyse relevante Informationen, erhoben.

In dieser Arbeit wurde eine vergleichende, quantitative, empirische Studie mit bereits bestehenden Materialien und Informationen durchgeführt. Mit einem eigens entwickelten Programm wurden Daten aus Wikipedia, Wikidata und ULAN gesammelt und für weitere Analysen aufbereitet. Zur Auswertung wurden spezielle Key Performance Indikatoren entwickelt, die anschließend zur Beantwortung der Fragestellungen herangezogen wurden. Als weitere Analysemethode wurde die Graphentheorie eingesetzt, mit der soziale Netzwerke noch detaillierter betrachtet werden können. Die Ergebnisse sind mit Praxisbeispielen und Visualisierungen untermauert.

Abstract

The importance of the Internet as a source of information and knowledge is constantly increasing. Wikipedia is one of the largest freely available encyclopaedias. It contains a high number of articles in different languages. Many people start their research for a topic on Wikipedia. There, everybody can create and edit articles. As a consequence, treated topics are as diverse as the user base itself. Depending on the different language versions, articles about different topics are created or articles about the same topic may contain different information. The main question which arises is, how complete and reliable information and articles on Wikipedia are.

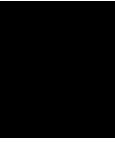
In the course of this thesis differences among the language versions were analysed and quantified with respect to listed relationships in articles. Due to the high number of articles and language versions not the whole encyclopaedia was analysed but only a subset of articles - namely articles about arts in the language versions English, German and Italian. Relevant articles were compared on information about the social network amongst artists. Furthermore, the thesis covered the issue whether articles about artists are more detailed in the "national" Wikipedia version. In the examination only relationships contained as hyperlinks in articles were considered. To verify completeness two other sources of information, the Union List of Artist Names (ULAN) and Wikidata, were adduced. ULAN is a directory which contains (relationship-)information about artists, art-related persons and art-related objects. Wikidata is a source of information for different Wikimedia projects. From there, relationship- but also other information was collected, which was relevant for analysis purposes.

This thesis is a comparative, quantitative, empirical study which is based on existing documents. Data was collected and classified from the three sources with a self-developed program. Apart from the program, key performance indicators were defined to answer the research questions. Besides that, social networks were compared in a more detailed way with graph theoretical metrics. Measures and results were clarified with practical examples and visualisations.

Contents

Kurzfassung	ix
Abstract	xi
Contents	xiii
1 Definition of the content, goals and scope of this master's thesis	1
1.1 Research questions	4
1.2 Scope	4
2 State of the art	7
3 Introduction to Wikipedia, Wikidata, ULAN, Graph Theory and search engines	11
3.1 Wikipedia	11
3.2 Wikidata	20
3.3 ULAN	21
3.4 Graph theory	23
3.5 Information about search engines and basic architecture	27
4 Presentation of the program	31
4.1 Program modules and example	31
4.2 Problems occurred and insights gained	52
5 Definition, calculation and analysis of KPIs	55
5.1 General KPIs	55
5.2 Language specific KPIs	72
6 Comparison of three national artists	99
7 Possible reasons for the differences between language versions	103
8 Conclusion and critical reflection	107
List of Figures	111
	xiii

List of Tables	112
Bibliography	115
Appendix	119



Definition of the content, goals and scope of this master's thesis

Due to the constantly increasing number of Internet-connected devices, many services, the selection and ordering of products or other things have moved to the Internet. Some examples of Internet services which outperformed previously existing products are messenger services for texting (superseded SMS) or video and music streaming services from Amazon, Netflix, Spotify or iTunes which outperformed or revolutionized the sales of previously existing video and music media. Whereas the number of sent WhatsApp messages increases and nearly reached the barrier of 11 trillion messages per year, the number of sent SMS is decreasing in many different countries after years of constant growth ¹.

Similarly, knowledge acquisition has changed. In former times knowledge was transported from one person to another one. People read books to get more information about a topic. Both kinds of knowledge transfer are still valid. However, a new method of gathering information emerged. People use the Internet to get information and acquire knowledge. There are forums where questions can be posted and other users answer them. An example for such a forum is called WetCanvas². There, users discuss art history. Forums are available for nearly every topic. Likewise, there are blog-posts where users describe their experiences in different situations, podcasts and videos where experts or amateurs talk about certain topics or articles, where information is carried together from multiple sources.

A traditional product people tended to use if they had questions, the encyclopaedia, was in fact substituted for the majority of the population. One of the successors of

¹ <http://www.investopedia.com/articles/investing/071515/how-whatsapp-killing-sms-texting.asp>; [accessed 25-February-2016]

² <http://www.wetcanvas.com/forums/forumdisplay.php?f=24>; [accessed 13-July-2016]

the traditional version is a website called Wikipedia³. On Wikipedia users can write articles about different topics. Articles are as diverse as the user base. Diversity, on the user as well as the article side, is one of the key success factors of Wikipedia. Today, there are millions of articles in many different languages. Information and knowledge shall be available to everybody – that is why the language variety is very important. Another important aspect which is also adhered in the slogan of Wikipedia, "the free encyclopaedia", is the free access to it.

Besides many advantages of Wikipedia there is an important fact which has to be considered - not every editor is a professional one or expert in the concerned topic. Consequently, the information contained in articles has to be treated with caution. The articles may contain wrong information or be incomplete. Considering this, Wikipedia is not really accepted in academic papers⁴. Yet, the number of citations of Wikipedia articles is constantly increasing⁴. If a faulty Wikipedia-article is cited in an academic paper the mistake may also be copied by other people who cite it. Once the mistake is copied by someone and the paper is published on the Internet, it is nearly impossible to revert this action and prevent other people from citing it.

Unfortunately, the quality of an extensive article is hard to evaluate – especially automatically. The number of intra-Wikipedia links could be viewed as a quality indicator. A higher number could indicate that the author conducted extensive research work. The picture of the topic the author wanted to provide should be as complete as possible. If the reader had further questions about something in the text, he or she could jump to the corresponding article directly through one of the links.

Art is a very important topic on Wikipedia. There are numerous articles about artists themselves, artistic movements, famous artworks or museums they are exhibited in. In articles about a person often also articles about other persons are linked. Links always imply a relationship between two things. There exists a relationship between e.g. a person and city if the person lived in the city, or two persons if they are in a teacher-student or patron-artist relationship. The relationship between people, especially the teacher-student/patron-artist one, are of main interest in this thesis. Related people contained in Wikipedia will be compared to the Getty Union List of Artist Names (ULAN) and Wikidata. More information about Wikipedia, ULAN and Wikidata will be provided in chapter 3.

The gathering of links and extraction of information from Wikipedia, ULAN and Wikidata was done automatically. For this purpose, the mechanisms of search engines could be adapted. The crawling process started with an article, extracted embedded links, and saved the collected information to a database. After the crawling process for the first article was finished, the process started all over again for one of the articles where a link was contained in the first article. Besides the links also other information was extracted and saved. More details about that will be described in chapter 4. The gathered data was analysed with statistical methods and evaluated with self-defined key performance

³ <https://www.wikipedia.org/>; [accessed 25-February-2016]

⁴ <http://www.zeit.de/digital/internet/2012-04/wikipedia-scholarpedia-verweise>; [accessed 25-February-2016]

indicators.

During the creation of the program and the data collection process some difficulties occurred. One problem was, that information had to be extracted in different formats and normalized to a consistent one. This also led to another difficulty: How can it be determined if data sets from two sources are equal? This question arises soon, as names can be written in different formats (e.g. with or without a second name). Another underestimated factor was the time needed to collect information from Wikipedia, Wikidata and ULAN. It is not allowed or recommended to send requests to servers in short intervals (< 1 second). In the case of Wikipedia where there are millions of articles the built in delay has a major influence on the runtime of the program. Obstacles and insights are described in chapter 4.2 in a more detailed way.

The paper "Art history on Wikipedia, a macroscopic observation" serves as a basis for the thesis. In their paper Goldfarb, Arends, Froschauer and Merkl examined the link structure between different articles/persons in the ULAN, Dbpedia and Wikipedia. They recognized clusters between artists depending on their nationality and the art historical periods. Still, different language versions were not compared extensively and reasons for language differences not discussed. These two points were left for further research. [Goldfarb et al., 2012]

Within this language comparison process, it will be answered, if there are differences in the numbers of linked artists in different languages. It will be clarified if, for example, there are more German artists linked in an article from the German Wikipedia than in the English or Italian one. Is the situation the same for the English and Italian Wikipedia version?

For this thesis, the German, English and Italian language versions were chosen because of different reasons. Firstly, all three of them are amongst the largest language versions in total. The English version is even the largest one. Additionally, there exist similarities between the languages (e.g. not many special characters like À or Ç in the French language which have to be considered). Even though the Italian version has fewer articles than the English or German one, it is interesting to analyse this version with regard to arts, as many famous artists were born in Italy.

The thesis is structured in five main parts. At the beginning, the three data sources are introduced, some theoretical concepts about graph theory are summarized and the architecture of a search engine is presented. In the following the program itself is described together with the tasks and responsibilities of the contained modules. The workflow will be illustrated with a concrete example. The next part contains the specification of the key performance indicators which will be used to measure data quality and answer the research questions. These measures are applied on the collected data to compare the different language versions. In the next section some possible reasons for differences among the language versions are collected. The conclusion and reflection build the end of this thesis.

1.1 Research questions

Within this thesis the following questions shall be answered:

- In which way can differences between Wikipedia language versions be quantified?
- To which extent do the three language versions provide the same information about relationships from artists to other persons? How large are the deviations between the different language versions?
- To which extent are listed relationships from ULAN and Wikidata covered in Wikipedia articles?
- What does the distribution of birth- and death places look like in the different language versions? Do the distributions show content differences in the three language versions?
- Can a preference of national artists be proven in the different language versions of Wikipedia?

Calculations and analyses to answer all questions will be performed on the collected data from Wikipedia, Wikidata and ULAN.

1.2 Scope

Within this thesis, only three language versions of Wikipedia, the German, English and Italian one, will be compared. A comparison process for all Wikipedia language versions would not be possible as there are too many of them. It would take too long to incorporate e.g. special characters or terms which indicate the relevancy of an article of all languages in the program. Additionally, the time needed to gather and analyse data of all relevant articles would take too long. Moreover, as there are many language versions with just a few articles, they probably would just contain articles about local or very famous artists and not many other ones which could be compared. The German, English and Italian language versions are amongst the largest ones on Wikipedia and therefore contain enough articles for a reasonable comparison.

If smaller language versions would be considered too, probably further sources of reference, such as ULAN, would be needed. ULAN does not contain a complete list of artists. This would be a problem for smaller language versions as they probably appertain to a small country where just a few (famous) artists originate. The chances are low, that they are listed in ULAN.

Of course, multiple sources of comparison, besides ULAN, would be of advantage. ULAN is not perfect and probably does not contain (all) information about every artist. Hence, results would be more precise if more sources would be incorporated in the comparison process.

The content for the comparison process is restricted to the area of arts. The main reasons for this are that ULAN is a good and reliable source for comparisons and human relationships (as between artists) are generally very interesting. Anyway, not all kinds of artists will be considered. A list of considered occupations is presented in chapter 4.1. If the target group of Wikipedia articles would be extended, not only additional reliable sources for comparison would be needed but also the data collection and analysis would take much longer.

Moreover, a text analysis mechanism to recognize phrases, names of persons or other structures will not be implemented, as this would exceed the scope of this thesis. Therefore, Wikipedia articles are only analysed on contained links.

Lastly, only results of the analysis process will be published. The database containing the collected information will not be available in public.

State of the art

Due to its popularity, dimensions and content diversity, Wikipedia was already topic of many research projects. In these projects Wikipedia was examined from different perspectives. Some analyses focussed on the content, others on the structure and development, again others on usages of Wikipedia in different areas, for instance the academic one.

Many papers deal with the examination of the overall structure of Wikipedia. One example for such a paper is "Preferential attachment in the growth of social networks: the case of Wikipedia" [Caldarelli et al., 2006]. The authors regarded the network of articles and links as a directed graph. One important property they mentioned is the similarity to the World Wide Web. Still, the growth of Wikipedia is somehow different to the one of the World Wide Web. The growth mechanism occurring on Wikipedia can be explained with the preferential attachment mechanism. This is remarkable as users can add articles anywhere globally. In their paper the authors do not focus on a certain article category but examine Wikipedia in general. In their analysis and graph based perspective, they see articles as nodes and hyperlinks contained in the text as edges connecting the nodes. Another important fact the authors mentioned is, that it is possible to reach nearly every article from any other one. The fact that articles are highly interconnected also has another advantage. It can help to understand the content of an article by "... visiting a connected path along the network". The degree distribution of links between articles exhibits a power law distribution. The theoretical foundation of their analysis is the idea of scale-free networks which first occurred in the Barabási-Albert model. [Caldarelli et al., 2006]

A graph theoretic perspective is very helpful for analysis purposes. Goldfarb, Merkl, Arends and Froschauer incorporate this method in their research process too. In their paper "Art History on Wikipedia, a Macroscopic Observation", they focus on a specific subset of all the article categories contained in Wikipedia, namely articles about art history related persons. The paper addresses the issue that Wikipedia was often analysed in general but not with focus on a certain domain. Within their analysis they also

examined the connection between art history related people regarding their occupation and nationality. They combined information from Wikipedia, ULAN, the Virtual Internet Authority File (VIAF) as well as Dbpedia, a project which contains structured information of Wikipedia. Apart from artists themselves also other groups of people are very important in the context of this analysis. In the field of arts collectors, patrons, politicians or monarchs play important roles too. Information was analysed in form of a graph. With the graph visualization and calculation tool Gephi they visualized the structure and network of Wikipedia articles, did some calculations to analyse and draw conclusions from the graph. In the first visualization the authors found a giant connected component. To increase expressiveness, they took the factor time into consideration. Links covering a difference in time of 0-37.5 years occur much more often than links between larger timespans. This means that persons who lived more years apart are sparsely linked. More often contemporary persons are linked. Additionally, links tend to point into the past and not into the future. The major conclusion is that the in-article links form clusters between people of the same nationality and clusters which focus on art historical periods and schools. Furthermore, the authors found indications that some properties like the node degree distribution of the examined graph are also valid for Wikipedia in general. A language comparison was left for future research projects. [Goldfarb et al., 2012]

A project which has the different language versions of Wikipedia in mind is "Omnipedia". Omnipedia is a project of the Northwestern University. The authors recognized, that language versions sometimes offer content exclusively. Although the language versions convey a feeling for different cultural viewpoints, language barriers might also hinder content to be available in other Wikipedia versions. The program is not yet released for public use (by March 2016). In Omnipedia users can enter queries and the program then graphically illustrates and links all information from an article in different language versions. In some way users can use this tool for comparison purposes. The main purpose however is to show all information from the language versions together to provide an overview about the contents of all articles to the user. Articles are often biased. Based on the language version chosen, they contain information which is more relevant in their own country or surroundings. [Bao et al., 2012]

Some people also compared information contained in Wikipedia to other sources. Clauson, Polen, Boulos and Dzenowagis took the fact that many people lookup health related information on the Internet (and thus also on Wikipedia) as an incentive to compare the scope, completeness and accuracy of drug information on Wikipedia to a traditionally edited online database about drugs. The authors focussed on certain drug categories for comparison. The method of approach was to construct questions which should be answered the best way possible with the two sources. The authors came to the conclusion, that information on Wikipedia is worse than the information contained in the drug database. [Clauson et al., 2008]

Another content comparison project was not drug related but focussed on political coverage. The possibility that everybody can edit articles is surely positive but can have negative consequences too. In 2007, a user even created multiple accounts to confirm

edits he made in articles. Another negative practice already discovered was that in the field of politics, congressional staffers removed negative information in articles about their supervisors. Sometimes they even falsified articles of opponents. Another insight the author got was that articles about politicians are more extensive if they are able to edit their own one. In this article the chosen method of approach was to identify specific facts which articles should contain and then check for a sample of articles if they contain the information or not. The author noticed, that existing articles are quite accurate but sometimes information is incomplete – especially for older or vaguer topics. If Wikipedia contains information it is quite accurate. Nevertheless, the author recommends to rely on other sources and rather use Wikipedia as a tool to get an overview of a topic. [Brown, 2011]

Chelms and Prasanna divided social network analysis in three categories – graph theoretic approaches, data mining as well as semantic approaches. The main idea of the graph theoretic perspective is to unveil certain characteristics of a network like the most important actors. In this context centrality measures are often used to identify such nodes. The second approach in social network analysis, data mining, aims to examine the content published in social networks. In the course of this analysis method sentiment analysis or the use of certain features like hashtags play an important role. Lastly, the method of semantic analysis takes more complex relationship types between actors into account. The main goal of this paper is to summarize different analysis approaches. The authors hereby list pros and cons of the different techniques. Even if the authors' perception of social networks was focused on Twitter, Facebook and others, concepts can also be applied to the analysis of other social networks. [Chelms and Prasanna, 2011]

There exists no paper which really focuses on the comparison of the content with respect to different languages. As mentioned in the paper "Art History on Wikipedia, a Macroscopic Observation" [Goldfarb et al., 2012] – this task was left for further research projects. In contrast to Omnipedia this thesis is not meant to provide users an overview of a certain topic – enriched with information from different language versions of an article. This thesis shall highlight and quantify differences between language versions. Content contained in Wikipedia will be compared to ULAN and Wikidata.

Introduction to Wikipedia, Wikidata, ULAN, Graph Theory and search engines

In the following chapters some basic concepts and information will be presented which are needed in the course of this thesis. The first part deals with the sources of information and used theories, later on technical concepts will be described.

3.1 Wikipedia

Wikipedia is one of the most popular websites worldwide. In December 2015 Wikipedia had around 2.5 billion site visits from desktop computers ¹. In the "Top 500 Global sites on the web" ranking of Alexa Wikipedia is also listed on the seventh place (November 2015) ². There are more than 27 million registered users, around 38.5 million articles in all language versions together and more than 800 million page edits since Wikipedia was set up in 2001 ³.

In January 2016 there were in total 280 different language versions of Wikipedia. The largest among them, based on the number of articles, is the English version. The smallest ones, Afar and Muscogee, had just one article (by February 2016). Based on the number of articles, the five largest language versions are (by February 2016)

¹ <http://www.similarweb.com/website/wikipedia.org>; [accessed 14-January-2016]

² <http://www.alexa.com/topsites>; [accessed 28-December-2015]

³ <https://en.wikipedia.org/wiki/Special:Statistics>; [accessed 28-December-2015]

- the English (5.07 million articles),
- the Swedish (2.82 million articles),
- the Cebuano (1.99 million articles),
- the German (1.91 million articles) and
- the Dutch one (1.86 million articles). ⁴

Based on the number of active users by February 2016

- the English (133.970 users),
- the German (21.514 users),
- the French (17.790 users),
- the Spanish (17.087 users) and
- the Japanese version (11.940 users) are the largest ones. ⁴

On the special Wikipedia page "About" there are also some criteria defined, which declare desired properties of articles. Amongst other criteria, neutrality, verifiability and reliability are the most important criteria. The principle "neutrality" means that authors should not include their own opinion and write articles in an objective manner. Of course, readers should be able to verify the content with the provided sources and references in the article. ⁵

Wikipedia also set up a page to explain how to evaluate the reliability of a source. There it is defined what kind of content is a source or where reliable content may originate (e.g. review articles, reports from news agencies which are responsible for accuracy). Furthermore, also some hints are listed like the fact that the context, the content is derived from, is very important and can completely change the message the content transports. ⁶

Lastly, there is also some information what Wikipedia is not and what authors should keep in mind. It is for example not

- a dictionary
- a place to publish personal thoughts or inventions
- a place to promote something (as for example political programs)

⁴ https://en.wikipedia.org/wiki/List_of_Wikipedias; [accessed 05-January-2016]

⁵ <https://en.wikipedia.org/wiki/Wikipedia:About>; [accessed 05-January-2016]

⁶ https://en.wikipedia.org/wiki/Wikipedia:Identifying_reliable_sources; [accessed 05-January-2016]

- a manual or guide
- a website to publish something about unverifiable events or things which might happen in the future. ⁷

The most important property of Wikipedia however is, that it is not censored. As a consequence, authors are encouraged to write articles in a way which is acceptable for all readers. Content is visible immediately and will be removed if Wikipedia's policies or the law of the United States is violated. The law of the United States is applied as Wikipedia is hosted there. Nevertheless, everyone has to keep in mind, that content is publicly available until it is reported and removed from Wikipedia. ⁷

Wikipedia has many advantages over a conventional encyclopaedia. The most important one is probably the pace of actualizations of articles. They are revised much more often than cased books. Changes are viewable immediately for other users worldwide. The 20th edition of the encyclopaedia Brockhaus was published from 1996 until 1999 (with some expansions in the following years), the 21st edition not until 2005. Another major advantage is the price tag - users do not need to buy a subscription or something else for Wikipedia, it is free for everybody. A new version of the Brockhaus encyclopaedia with more than 20 books costs several hundred Euros. Additionally, it takes much less time to enter the term a user is looking for in a search field and select one of the found articles than to get the right book of the encyclopaedia series and find the right page. Apart from that, Wikipedia can be accessed from anywhere in the world where a user has Internet access. In many cases the structure of Wikipedia articles is also more clear than in conventional encyclopaedias. Besides that, Wikipedia articles sometimes also have a bigger scope than their traditional counterparts. Moreover, different formats of content can be included in Wikipedia articles. In many articles there are images embedded, sometimes even audio files (such as in the article about the Austrian national anthem ⁸) or videos (like in the article about 3D scanners ⁹ where the result of a scan is shown) are incorporated. Even if book versions of encyclopaedias contain a CD or online access to media files the user needs the book as well as a computer or something similar to view or listen to the media files in parallel.

Arends, Froschauer, Goldfarb and Merkl also pointed out in their paper, that a major advantage of web based education systems (in their case for art) is the easy access of websites. Nowadays, people are accustomed to look for information online and find it therefore sometimes easier to search for something online than in a series of encyclopaedias as in this case. [Arends et al., 2011]

Of course, there are also a lot of other encyclopaedias besides Wikipedia and Brockhaus (for instance Baidu Baike or Digital Universe). Some offer cased books and online versions (such as the Encyclopaedia Britannica), others are available either online or as a book

⁷ https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not; [accessed 05-January-2016]

⁸ https://en.wikipedia.org/wiki/Land_der_Berge,_Land_am_Strome; [accessed 06-January-2016]

⁹ https://en.wikipedia.org/wiki/3D_scanner; [accessed 06-January-2016]

version. Depending on the version and condition of a collection, a cased edition of the Encyclopaedia Britannica can cost several hundred euros or online access is available for €69.95 per year (by February 2016) ¹⁰. Another alternative for the categorization of encyclopaedias is their specialization. There are books which are focussed on a certain area of knowledge (including the Oxford Classical Dictionary which is an encyclopaedia for antiquity, the Max Planck Encyclopaedia of Public International Law or the McGraw-Hill Encyclopaedia of Science & Technology), others contain articles about any topic (like the Columbia Encyclopaedia or the Encyclopaedia Britannica).

In 2015, Michael Mandiberg, an American artist, published the English Wikipedia version as an e-book. Whereas a version of the 21st edition of Brockhaus contains around 30 books (depending on the number of extensions), the published e-book of the English Wikipedia version consists of 7.473 volumes of 700 pages each. This illustrated pretty well how much information is really contained in Wikipedia. ¹¹

The ability to store so much data is another advantage of web based solutions, as a lot of information can be made available for the public. [Arends et al., 2011] Not many people have enough space to house so many volumes of a printed Wikipedia encyclopaedia.

In occasion of its 15th birthday Wikipedia published which articles were re-edited the most. In the English Wikipedia version, the article about George W. Bush is leading the list in front of the List of WWE (World Wrestling Entertainment) personnel and the article about the United States. Furthermore, articles about other politicians (e.g. Barack Obama), religion (e.g. Jesus), artists (e.g. The Beatles) or events (e.g. World War II) are contained in the list. [Fitzpatrick, 2016]

3.1.1 Operation

The software Wikipedia is based on is called Mediawiki. It is developed by many users all over the world. A major operator of Mediawiki is the Wikimedia Foundation. Even though Wikipedia is so popular itself, it is just a project of the Wikimedia Foundation. Besides Wikipedia they also take care of other Wikis, for instance Wiktionary, Wikidata or Wikiquote. Due to the free Creative Commons license, Mediawiki is released under, also other companies or private users can use Mediawiki to create their own Wiki. ¹²

Wikipedia runs on less than 300 servers in Tampa (US) and Amsterdam (Netherlands). Even though the number of servers is very low compared to the number of servers of Google or Facebook, some numbers about daily usage point out, how performant Wikipedia's servers are. On a normal day, the servers have to handle more than 50.000 HTTP requests and 80.000 SQL queries per second. According to statistics the servers

¹⁰ https://safel.britannica.com/registrations/signup.do?partnerCode=FAQ_012610; [accessed 06-January-2016]

¹¹ https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia; [accessed 06-January-2016]

¹² https://Wikimediafoundation.org/wiki/Our_projects; [accessed 07-January-2016]

have an uptime of 99%.¹³

The crawling process to collect all the data for this thesis took quite a while – but the reason was the high number of articles and not bad performance of Wikipedia servers. It would have been possible to crawl one or even multiple sites every second. Although, if many people did it that way, the performance would be worse for all users. On these grounds, the program contained a crawl delay of at least three seconds.

When it comes to the creation of content, one of the main advantages, the fact that everybody can contribute to Wikipedia, also turns out into a disadvantage. To edit content on Wikipedia, a user does not even need to be registered. If a user is not logged in, the IP address is saved in the version history. Nevertheless, if the article is faulty it stays this way until another user corrects it. There are some professional editors who proofread articles, but due to the enormous number of articles this may take some time [Ayers et al., 2008].

3.1.2 Funding

Wikipedia does not show ads to earn money. This means that it has to rely on users to donate money to keep the project running. During the fundraising period 2014/2015 users donated more than 75 million US-dollars¹⁴.

In the article “Free but not easy”, published by The Economist, the author pointed out, that the number of servers is much smaller than the number of servers at Facebook or Google. The major part of the funding is used for technology expenses, the major content however comes from editors who are not employed by Wikipedia. [Eco, 2011]

According to Wikipedia Statistics the number of active Wikipedians had its peak in March 2007 with 91.444 users. Since then, the trend was directed slightly downwards. By February 2016, Wikipedia had 69.845 active users.¹⁵

¹³ <http://www.datacenterknowledge.com/archives/2008/06/24/a-look-inside-wikipedias-infrastructure/>; [accessed 07-January-2016]

¹⁴ https://Wikimediafoundation.org/wiki/2014-2015_Fundraising_Report; [accessed 03-January-2016]

¹⁵ <https://stats.Wikimedia.org/EN/TablesWikipediansEditsGt5.htm>; [accessed 09-November-2016]

3.1.3 The major problem

Figure 3.1 illustrates, that the growth of Wikipedia is declining and the number of articles is not growing as fast as in previous years:

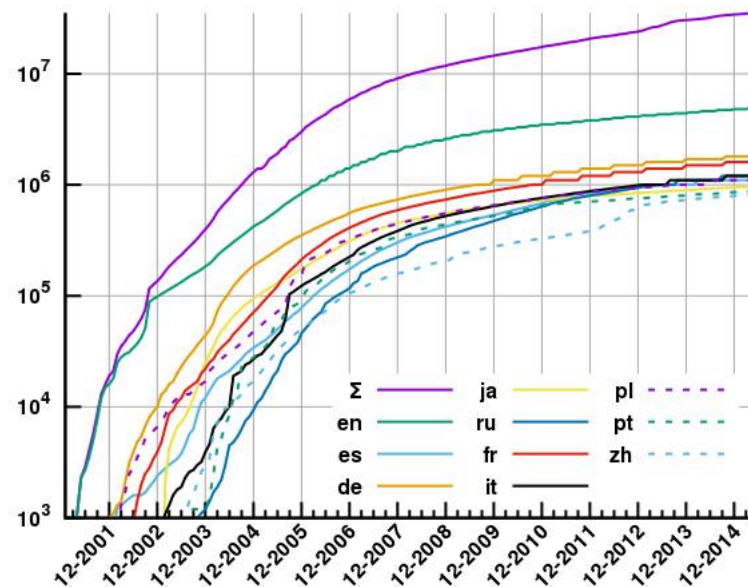


Figure 3.1: Growth of the number of articles on Wikipedia¹⁶

The language abbreviations in Figure 3.1 represent English (en), Spanish (es), German (de), Japanese (ja), Russian (ru), French (fr), Italian (it), Polish (pl), Portuguese (pt) and Chinese (zh).

Barry Newstead, former Chief Global Development Officer of the Wikimedia Foundation, said that being a Wikipedia author is not easy. They have to gain authority amongst other Wikipedia editors, understand policies and write the article in a special format required by Wikipedia. The even greater problem however is that the majority of Wikipedia users does not even know that they can actively edit articles. Another problem, which becomes even more important today is, that many people just use their smartphone to look something up in the Internet or in this case on Wikipedia. The bad thing for Wikipedia is, that smartphones are not well suited for editing an article and doing some research in parallel. To overcome this problem Wikipedia started a cooperation with universities. The first cooperation was established in the town of Pune, India, since the language variety in India is very high but at the time the article in *The Economist* was published, the number of articles in Indian languages was rather low. In the course of lectures students had to write Wikipedia articles to gain points. This was not just an

¹⁶ https://upload.wikimedia.org/wikipedia/commons/2/23/Articlecount_topten_wikipedia.svg; [accessed 10-January-2016]

advantage for Wikipedia but also for the students as they had a larger readership than just their professors. [Eco, 2011]

In their survey “The Rise and Decline of an Open Collaboration System: How Wikipedia’s reaction to popularity is causing its decline”, Halfaker, Geiger, Morgan and Riedl investigated why the number of authors on Wikipedia was declining. The main reasons they identified were the complex editing guidelines, the automatic control algorithms which often just decided to accept or reject an article (if an article is rejected it is reverted instead of giving users the chance to rewrite them) and the harsh exchanges with already established authors from Wikipedia’s inner circle. [Halfaker et al., 2012]

The typical Wikipedia author has certain characteristics. Authors are mostly white males from western industry nations ¹⁷. According to a survey conducted in 2013 by the Public Library of Science (PLOS ONE) only 16% of the authors are women whereas the share of female readers is 50% [Hill and Shaw, 2013]. The gender bias also influences the topics present on Wikipedia. Whereas male topics such as "Hot rods" or "Pinup girls" are contained in the encyclopaedia, female topics like "Pregnancy in art" or "Women in classical music" are missing ¹⁸. The gender composition of the collected data stock reflects this bias too. Whereas the database of collected persons contains 4.818 relevant, female persons with an ULAN-ID, 49.613 male persons with the same criteria were found in the database. The data stock will be described in a more detailed way in the chapter "Crawler".

In addition to that, Graham, Straumann and Hogan conducted a survey and found out, that the availability of broadband Internet is an important factor for the activeness on Wikipedia. The higher the availability, the more users contribute to Wikipedia. Another surprising fact is, that people from poorer regions do not write about their own country or region but rather about richer areas. [Graham et al., 2015]

3.1.4 Trustworthiness

Due to the popularity of Wikipedia many students use it to look up information. Although students might find articles about nearly every topic they need during their education-career, Wikipedia is often not tolerated by professors. In different surveys, which were conducted among different university faculties, the authors concluded that the main reason why Wikipedia is not really accepted in the scientific area is, that everybody can edit articles and write anything he or she wants. This can lead to wrong content in the articles. Even though mistakes might be corrected later on, they are wrong a certain time. As a consequence, only 7% of the respondents of a survey declared, that they frequently use Wikipedia for teaching and research purposes. [Aibar et al., 2013]

Other reasons for a poor acceptance of Wikipedia in academic areas are a lack of knowledge of the editing process on Wikipedia and the way knowledge is created. According to the conducted surveys the collaborative knowledge creation process is not well accepted

¹⁷ <http://science.orf.at/stories/1766259/>; [accessed 15-January-2016]

¹⁸ https://en.wikipedia.org/wiki/Wikipedia:Systemic_bias; [accessed 18-April-2016]

outside academia. This way does not follow the conventional teaching and scholarship methods. [Aibar et al., 2015]

In a student guide provided by the Harvard University it is also pointed out that students can start their learning or research process on Wikipedia. However, as a counterexample for credibility they present the story of a student who wrote in a Wikipedia article that he is the mayor of a Chinese town. Even after some years the entry was still online.¹⁹ Sook Lim of the St. Catherine University in Minnesota (US) conducted a survey about the Wikipedia usage of college students. She noted, that on the one hand many people have concerns about the quality of articles, on the other hand there is evidence that the quality is not so bad at all. The importance and popularity was pointed out as every participant of the survey stated that he or she uses or used Wikipedia (the frequent user group was the largest). The majority of participants however used it for non-academic purposes. The survey also confirmed the problem denoted by the article from *The Economist* – just a minority of the participant group actively edited articles. Even though the majority of respondents had positive experiences with the quality of Wikipedia articles they knew they had to be careful and do not believe everything that was written in the articles. [Lim, 2009]

These findings were also confirmed in another survey which is described in the paper "How today's college students use Wikipedia for course-related research" which was written by Head and Eisenberg. The major findings were that the majority of students uses Wikipedia. The reasons however are more diverse (e.g. an easy to use interface, a good starting point for research work or a source for literature which is linked in the article). One insight the authors gained was, that the usage of Wikipedia also depended on the major the participants chose. Students from engineering, architecture or other sciences tended to use Wikipedia more than students from other majors. [Head and Eisenberg, 2010]

A survey of bitkom (Bundesverband Informationswirtschaft, Telekommunikation und neue Medien) came to the conclusion, that 12% of the participants completely trust articles on Wikipedia, 67% think that articles are mostly trustworthy. Around 18% stated, that they think that articles are rarely trustworthy. The remaining 2% of the participants said, that they never rely on information contained in Wikipedia articles.²⁰ For this thesis Wikipedia will be the main data source. Based on the articles contained in Wikipedia, corresponding ULAN and Wikidata pages will be crawled as well. As already mentioned one of the main questions in this thesis is the "completeness" of Wikipedia compared to ULAN and Wikidata. Of course, this cannot be determined completely as there may be still undiscovered relationships between artists. Besides that, just the links in the articles are considered and the text itself is not indexed or interpreted. Text interpretation would exceed the limits of this master's thesis. The other important question which will be dealt with is, if in the different language versions artists from the

¹⁹ <http://isites.harvard.edu/icb/icb.do?keyword=k70847&pageid=icb.page346376>; [accessed 10-January-2016]

²⁰ <https://www.bitkom.org/Presse/Presseinformation/Vier-von-fuenf-Internetnutzern-recherchieren-bei-Wikipedia.html>; [accessed 11-January-2016]

own country are preferred regarding in-article links.

3.1.5 Other encyclopaedias

If something is successful there are often many attempts to copy it. In 2008 Google started a Wikipedia competitor called “knol”. Knol should also be a platform where users can publish and edit articles. In contrast to Wikipedia a single person was responsible for an article. This person also had a share on earnings through advertisements. The possibility to earn money with interesting and good articles should attract writers. As a consequence, the platform was target for spammers and people who simply copied Wikipedia articles. The project was discontinued in 2012.²¹

Wikipedia’s co-founder, Larry Sangers, also started an own online encyclopaedia called Citizendium. Citizendium should improve the trustworthiness as only registered users can edit articles. Anyhow, the project is not very popular and the article creation nearly stagnated²².

A thing the competitors had in common was that they tried to control article creation and improve the quality. A theory why this did not work out might be, that users did not like the stricter regularities.²¹

Still, smaller and more specialized Wikis are very popular. Besides many other use cases, companies often use Wikis as a knowledgebase, fans of a TV series document everything about characters and the plot or even the German political party, Piratenpartei, publishes information in their own Wiki.²¹

Another Wiki which is based on Mediawiki is Scholarpedia. Scholarpedia focuses on articles from different fields of science for example applied mathematics, experimental high energy physics, neuroscience, physics or some others. It has three main principles:

- articles are peer reviewed,
- articles are written by experts on their fields and
- access to Scholarpedia is free.

In Scholarpedia each article is peer reviewed. In the case of a rejection the reviewer can remain anonymous, on the other hand, if an article is accepted the reviewer is named publicly. Articles are peer reviewed by Scholarpedia editors, invited curators or other invited persons. If an article is rejected it can be improved but in the meantime another article with the same name from different authors can take its place.²³

Curators are responsible for the content of an article. Each article has exactly one curator.

²¹ <http://schmalenstroer.net/blog/2011/11/das-langsame-sterben-der-wikipedia-konkurrenz/>; [accessed 23-January-2016]

²² <http://en.citizendium.org/wiki/CZ:Statistics>; [accessed 23-January-2016]

²³ <http://www.scholarpedia.org/article/Scholarpedia:Peer-review>; [accessed 23-January-2016]

They have to be world-recognized experts on their field. Curatorship can also change in the lifetime of an article (if all contributors of the article agree). With help of this principle the article quality and integrity shall be increased. Curators also have other responsibilities. They can sponsor proposed articles which means that they bail for credibility and the expertise of the author. Furthermore, as mentioned previously, they have to peer review articles and vouch for their correctness.²⁴

Lastly, Scholarpedia promises, that articles will be available for free.²⁵

Besides the free online encyclopaedias there are also paid alternatives to Wikipedia. A very popular example, the Encyclopaedia Britannica was already described earlier. Users have to pay a yearly fee to access content online. Content is created and curated by a group of professional authors.

3.2 Wikidata

Like Wikipedia Wikidata is a project of the Wikimedia Foundation too. Wikidata contains categorized information which can be used in different Wikimedia projects, for instance Wikipedia or Wikivoyage. Wikidata is designed for humans as well as machines. One of the main principles of Wikidata is the free access. Everybody can edit or copy information as it is licensed under the Creative Commons Public Domain Dedication 1.0. Moreover, Wikidata is a collaborative site. Users as well as robots create and edit content together. However, Wikidata is also a little bit different than other Wikimedia projects. Most Wikimedia projects contain entries like an encyclopaedia, but in Wikidata the information is listed in a more structured form. This comes from Wikidata's nature as a support project – information should be reusable easily by other projects. Therefore, programs such as web-crawlers have advantages here as they can parse structured data easier than texts in prose. If information from e.g. a Wikipedia article is linked to Wikidata, it will be updated automatically in the article if it changes in Wikidata. Besides this advantage, Wikidata also maintains the different language links of articles. The language versions are not restricted to Wikipedia articles but also other Wikimedia projects.²⁶

Figure 3.2 shows an extract of the Wikidata page of the Italian artist Titian. Information about him is listed in statements (e.g. the place of birth or the occupation). The statements are not standardized which means that not every person in Wikidata has a statement with the place of birth. The number of references tells readers where the information originates.

²⁴ <http://www.scholarpedia.org/article/Scholarpedia:Curator>; [accessed 23-January-2016]

²⁵ <http://www.scholarpedia.org/article/Scholarpedia:Open-access>; [accessed 23-January-2016]

²⁶ <https://www.Wikidata.org/wiki/Wikidata:Introduction>; [accessed 24-January-2016]



place of birth	Pieve di Cadore	edit
	▼ 2 references	
	imported from	German Wikipedia
	stated in	Integrated Authority File
	GND ID	118622994
	retrieved	12 August 2015
		+ add reference
		+ add

Figure 3.2: Excerpt from the Wikidata page about Titian²⁷

3.3 ULAN

The Getty foundation operates three vocabularies – the ULAN, the Art & Architecture Thesaurus (AAT) and the Getty Thesaurus of Geographic Names (TGN). Besides these three projects also another one, the Cultural Objects Name Authority (CONA), is currently in development. The main goal of the vocabularies is to provide users information about “... objects, artists, concepts and places important to various disciplines that specialize in art, architecture and material culture”.²⁸

ULAN has a pretty long history. Already in 1984 the Getty started to merge information from projects of the Jean Paul Getty Trust.²⁸

The Getty trust is one of the wealthiest non-profit foundations worldwide, with belongings worth more than USD \$8.4 billion. A very important project of the foundation is the operation of the Getty museum in Los Angeles, which has more than one million visitors per year.²⁹

At the beginning, ULAN really was just a list of artists. Later on the structure was adapted to harmonize it with the AAT and TGN and more information about the artists was included. In 1994 the first edition of ULAN was published in a printed and machine-readable version.²⁸

According to ULAN’s definition, artists are “... either individuals or groups of individuals working together. Artists in the ULAN generally represent creators involved in the conception or production of visual arts and architecture; performance artists are included, but typically ULAN does not focus on actors, dancers, or other performing artists.” ULAN is a network of actors. Even though the focus of ULAN is on artists, there are

²⁷ <https://www.wikidata.org/wiki/Q47551>; [accessed 24-January-2016]

²⁸ <http://www.getty.edu/research/tools/vocabularies/ulan/about.html>; [accessed 24-January-2016]

²⁹ <http://www.bloomberg.com/research/stocks/private/snapshot.asp?privcapId=4288215>; [accessed 24-January-2016]

also records for patrons like Pope Paul III ³⁰ was one (for Titian) or other persons, who are e.g. related to an artist as they are members of the same family. Besides records for people there exist also ones for art related things, in particular cities, places and museums such as the Kunsthalle Wien ³¹. ULAN and the other Getty vocabularies are tailored for “professional” users like museums, art libraries or researchers in art. For power-users there are different licensing options to access the data in different formats. Still, the web-access to the vocabularies is free of charge. Therefore, also students or other normal users can access and use it. ²⁸

ULAN has a broad scope of information about artists ranging from the name of an artist in different formats (e.g. different languages) and other general information, for instance, birth and death date, nationality, relationship types, related people and hierarchical positions. For every artist there exists also a section “Sources and Contributors” where it is listed who edited the entry and where the information came from. ²⁸

Figure 3.3 displays an excerpt of the “Sources and Contributors” section of the Austrian artist Friedensreich Hundertwasser. The linked entries refer to the full citation text of the source.

Sources and Contributors:

Friedensreich Hundertwasser [\[AVERY, GRL\]](#)
 [Avery Index to Architectural Periodicals \(1963-\)](#)
 [Library of Congress Authorities database \(n.d.\) n 50036736](#)
 Friedensreich Hundertwassser [\[AVERY\]](#)
 [Avery Authority files \(1963-\)](#)
 Friedrich Stowasser [\[BHA, GRLPSC\]](#)
 [RILA/BHA \(1975-2000\)](#)
 Fritz Hundertwasser [\[VP, WL-Courtauld\]](#)
 [Getty Vocabulary Program rules](#)
 Hundertwasser [\[AVERY Preferred, GRL\]](#)
 [Avery Authority files \(1963-\)](#)
 [Library of Congress Authorities database \(n.d.\) n 50036736](#)
 Hundertwasser, Friedensreich [\[AVERY, BHA Preferred, GRL Preferred, GRLPSC Preferred, Grove Art Preferred\]](#)
 [Avery Authority files \(1963-\)](#)
 [Grove Art artist database \(1989-\)](#)
 [Grove Dictionary of Art online \(1999-2002\)](#) accessed 17 July 2002
 [Library of Congress Authorities database \(n.d.\) n 50036736](#)
 [RILA/BHA \(1975-2000\)](#)

Figure 3.3: Snippet from the sources and contributors of the ULAN page about Friedensreich Hundertwasser³²

³⁰ <http://www.getty.edu/vow/ULANFullDisplay?find=titian&role=&nation=&page=1&subjectid=500114692>; [accessed 01-April-2016]

³¹ http://www.getty.edu/vow/ULANFullDisplay?find=vienna&role=&nation=&prev_page=1&subjectid=500237460; [accessed 01-April-2016]

³² <https://www.getty.edu/vow/ULANFullDisplay?find=&role=&nation=&subjectid=500005179>; [accessed 24-January-2016]

In this master's thesis ULAN will be used as a source for comparison. The relations and related people extracted from ULAN will be compared to the links contained in Wikipedia articles or student relationships from Wikidata. Even though ULAN might also be incomplete or contain wrong information – for the field of arts it is a very good source for comparison purposes. Not everybody can edit articles and information is usually well researched.

3.4 Graph theory

Graph theory is often the basis for social network analysis – no matter if it is a social network amongst artists or a friendship network on Facebook. A graph consists of nodes (which could represent persons) and vertices (relationship between two people) which connect two nodes. With these concepts different scenarios can be modelled. On the observed or created networks different calculations can be performed (e.g. calculate the distance between two individuals or determine the importance of an individual in the network). Some of these calculation methods are also used for analysis purposes in this thesis.

In his report Serrat mentions, that social networks have grown much more than any other form of human organization. Various examples for social networks are financial transactions between people, social contacts, visions or joint memberships in organizations. Social network analysis focuses in particular on the structure of the relationships between different individuals in the network. It is an upcoming area of activity as now, massive amounts of data are available and ready to examine. It was already used a lot in social sciences but reached other fields of study like electronic communication, business organization, health or psychology too. Social network analysis helps to understand the network structure and determine which groups or individuals are important in a network. Once the analysis is done the gained information can be used to improve existing connections between nodes. [Serrat, 2009]

Apart from that, different phenomena can be observed in social networks. The most important ones for this thesis are the small world phenomenon, power laws and shortest paths in networks. In context with the small world phenomenon it will be interesting to see, how long the shortest paths in a network are. Besides these phenomena, the calculation of different centrality measures can provide interesting insights into the network structure.

3.4.1 The small world phenomenon

Already in 1967 Stanley Milgram published findings from his small world experiment. In this experiment some people were selected to send a document to a predefined person they did not know. To accomplish this goal, the participants had to send the document to a person where they thought the probability was higher that the document reaches its destination. The prerequisite was that the sender had to know the receiver by first name. [Travers and Milgram, 1969]

Small in the context of social networks means that two individuals are connected through a short path. Applied on the population of our planet there are four facts which would indicate, that the social network does not exhibit a small world phenomenon:

1. There are billions of people who live on different continents
2. An average person may know many people – however the number is drastically smaller than the population
3. The topology has no central person who all other people know
4. Friendship networks are overlapping (if a person has a friend then they probably have a lot of other friends in common). [Watts, 1999]

These four properties make the existence of a small world phenomenon even more remarkable. Despite that, in sparse, decentralized networks local clustering is high and global separation short which (amongst other factors) makes the appearance of the small world phenomenon possible. The graph theoretical concept "clustering" will be explained later on. [Watts, 1999]

As artists also belong to the normal population the social network amongst artists should also exhibit the small world phenomenon. With the collected data it can be examined whether this assumption is true and how many links there are between two artists. Furthermore, as the number of artists is smaller than the population on the whole planet the distance between random artists should also be smaller.

3.4.2 Power laws

Another phenomenon which is very interesting for the analysis of social networks is the so called power law. In social networks, where the quantity being measured is some kind of popularity, degree distributions mostly follow power laws. This means that there are many people with just a few connections to other people but on the other hand there are some people who are very important and highly linked. [Neidhardt, 2014]

The Italian economist Vilfredo Pareto found out already in 1897 that 20% of the population possessed 80% of the land in Italy. He observed this principle in other scientific and natural areas too. This 80/20 rule is still valid today. For instance, 80% of the revenue is made by 20% of the customers or the best 20% of salesmen generate 80% of the profit of a company. [Ultsch, 2001]

On the basis of the random network model developed by Erdős and Rényi, Barabási and Albert developed a new network model which also takes the link structure of the World Wide Web into account. In their work they stated that in- as well as out-degrees of nodes in the web follow power laws. They observed that if a node already has more links to other nodes than another node, the one with the higher number of links will form new links faster. [Adamic and Huberman, 2000, Broder et al., 2000]

Based on the collected data from Wikipedia it can also be examined whether the number of links follows a power law distribution.

3.4.3 Shortest paths

In graph theory, a sequence of edges which connect two nodes is called a path. Short paths are related to the small world phenomenon, as the distance between two nodes is small. This time however, the idea is to find the shortest and not only short paths. Often, edges in graphs, where shortest path algorithms are applied, are associated with weights which are incorporated in the calculation. These kind of graphs are called weighted graphs. The goal is to find a way from node A to node B on a path with a minimal weight. A very popular algorithm to find the shortest paths in a graph is Dijkstra's algorithm. The algorithm finds the shortest path between a start node A and an end-node Z. In the first step, the distances (weights) from A to all neighbouring nodes are considered. The node with the lowest distance is selected (e.g. B). Then, again all distances to the neighbouring nodes are calculated. This time, the distance of node B to the neighbours is the sum of the distance from A to B and from B to the neighbour. If e.g. A and B are both connected to a third node C, then the distances A to C and A to B to C are compared. Again, the smallest distance to the next node is chosen. This calculation and comparison process continues until the end-node is reached. There exist also other algorithms to find shortest paths in graphs. Nevertheless, the weights can also be ignored and set to 1. Then, just a path with a minimal number of edges has to be found. [Wilson, 2012]

A famous related problem is the so called "Chinese postman problem". The problem was developed by Mei-Ku Kwan. The idea behind it is that a postman should deliver all mailings and visit each road at least once, but keep the traveling distance minimal and again return to the base. The roads can be seen as edges and their lengths are the corresponding weights. [Wilson, 2012]

Before the solution for this problem can be presented, another concept has to be explained. For some graphs there exists a so called Euler-tour. In an Euler-tour all edges are traversed once and the start- and end-node are identical. Per definition a connected graph is Eulerian if the degree of each node is even. Then, an Euler-tour is possible because you "enter" each node as often as you "leave" it. [Bondy and Murty, 2008]

So if the graph from the Chinese postman problem is Eulerian, an Euler-tour is a solution for the stated problem. The Fleury algorithm helps to find Euler-tours in a graph. If, on the other hand, the graph is not Eulerian, the problem is more difficult to solve but even then, an algorithm to find a solution is known. [Wilson, 2012]

The goal of the travelling salesman problem is to find a path with minimal weights and return to the starting point. This time however the requirement is to visit all cities (nodes) instead of streets (edges). The cycle which has to be found is also called Hamiltonian cycle. Even though it does not sound much more difficult than the Chinese postman problem, for many nodes it is nearly impossible to find a solution. The computation would simply take too long. There are, however, some algorithms which work fast but just approximate an optimal solution. [Wilson, 2012]

3.4.4 Clustering

Duncan Watts and Steven Strogatz developed a network model which improved the random network model from Erdős and Rényi to better fit real networks. The Watts Strogatz model finds a middle way because real networks are neither completely random nor completely regular. Regular networks have high average distances and are clustered, random networks have a small average distance but are poorly clustered. In the Watts Strogatz model there is both, small average path lengths and high clustering between nodes. [Sacharidis, 2015b]

High clustering means that if nodes A and B are connected as well as B and C then the probability is high that also A and C are connected.

Short paths, power laws and clustering are the three basic phenomena of social networks.

3.4.5 Centrality measures

Centrality is a very important concept in the area of graph theory and network analysis. The centrality of a node can be regarded as a measure of importance in a graph. In communication networks, centrality can e.g. help to identify important nodes on which the information is routed over. [Borgatti and Everett, 2006]

There are different centrality measures which differ in the way a node is considered as important and consequently also in the way they are calculated.

Some of the well-known centrality measures are

- the Degree centrality: This measure is the most basic one. A node has a high Degree centrality if it has many links to and from neighbours. In the Degree centrality, the centrality is the sum of the in- and out-degree of a node. Nevertheless, it can also be differentiated between the in- and out-degree-centrality, where the centrality of a node is equal to the in- or out-degree. [Opsahl et al., 2010]
- the Eigenvector centrality: The Eigenvector centrality is computed iteratively. The centrality of a node is the sum of centralities of its connected neighbours from the previous iteration-step. This means that if the neighbours of a node have a higher centrality the node itself also gets a higher centrality. [Borgatti et al., 1998]
- the Closeness centrality: For the calculation of the Closeness centrality the distances to every other node are summed up and divided by the number of nodes (average distance). As a consequence, nodes that have a lower average distance are of higher importance. [Faust, 1997]
- the Betweenness centrality: Nodes have a high Betweenness centrality if many shortest paths run through this node. They can control e.g. the communication of many nodes and if they fail the communication paths are much longer. [Newman, 2001]

- the Katz centrality: If a node is connected to important nodes or in general has many links, it also gets a higher centrality with the Katz method. As the Eigenvector centrality, also the Katz centrality is calculated iteratively.³³
- the PageRank centrality: The PageRank centrality measure is one of the most famous ones because it helps Google to determine the rank of a page in the search results. The algorithm was developed by Larry Page and Sergei Brin, the two founders of Google, and is based on the Katz-centrality. Amongst other adaptations, one refinement of the PageRank algorithm compared to the Katz-centrality is, that the centrality of a node is distributed to its neighbours in equal parts based on the number of out-links this node has. If a node has a high out-degree then each neighbouring node gets a smaller centrality value, if it has just a few out-links then each of the neighbours gets a higher centrality-share from this node. [Ding et al., 2009]

The centralities will be computed with the graph visualization and analysis tool Gephi³⁴. For this purpose, the collected information from the database will be converted to a graph. More details about the conversion and calculation of the different measures will be described in chapter 5 - Definition, calculation and analysis of KPIs.

3.5 Information about search engines and basic architecture

Search engines can be used in different ways. End users primarily use them to retrieve information. But besides this purpose, search engines are also used to satisfy other needs. On the one hand they are used to navigate to different websites, on the other hand they help to find websites which offer a certain functionality a user is looking for (to perform a transaction). The queries where users look for information have a share of less than 50% of the total queries. Depending on the use case queries are either called informational, navigational or transactional. Informational queries are ones where users do not have a specific site in mind. Search terms are rather general like "computers". In the case of navigational queries users assume that a certain site exists, they think of a specific website they want to navigate to. Transactional queries shall retrieve and lead users to websites where further actions will take place. Typical examples for actions are online shopping or portals where users can download content. First generation search engines such as AltaVista were only able to satisfy informational queries. The next generation, Google and others, also supported navigational queries. All three types of queries are supported by search engines from the third generation. [Broder, 2002]

Search engine providers display the requested information, but also show advertisements to earn money or analyse user data in other ways like Google Analytics is used to gather

³³ <http://www.sci.unich.it/~francesco/teaching/network/katz.html>; [accessed 12-February-2016]

³⁴ <https://gephi.org/>; [accessed 01-April-2016]

insights about user behaviour.

Today, there exist numerous different search engines. Some of them focus on certain languages, others on certain regions and again others on certain information they retrieve. By January 2016, the most popular multilingual search engines by market share are Google, Bing and Yahoo ³⁵.

An example for geographically limited search engines is search.ch. This search engine focuses on information from Switzerland. A query for “Zürich” returns a map where Zurich is shown, a weather forecast, train departures from Zurich main station, a cinema and TV program as well as search results related to the term “Zürich”. ³⁶

Examples for content focussed search engines are Monster.com (focus on jobs) or Google Scholar (focus on academic and scientific information). But there are also search engines which distribute a user query on several search engines and aggregate the results – these are so called metasearch engines. Some well-known examples for metasearch engines are Dogpile, Clusty or Ixquick.

In their paper, “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, Sergey Brin and Lawrence Page described the basic architecture and principles how Google works. Figure 3.4 depicts an abstract version of this architecture. In their architecture they have an URL server, which takes care of URL handling. It stores URLs and provides them to the crawlers. There are many crawlers which follow the provided URLs and download them. Before the websites are stored they are pre-processed so they e.g. do not take up so much space. In the next step, the indexer parses the stored websites. The words and occurrences (so called hits) are analysed. Besides the word itself, a hit contains a lot of other information such as the position in the document or font properties like capitalization or size. The created hits are then stored in “barrels” and a forward index is created. A basic forward index is a list of documents with each document pointing to a list of hits which are contained in it. As well as the hits creation the indexer also extracts the links in the documents and stores the link text and the endpoints of the link in “anchor” files. The next component, the URL resolver fetches data from the anchor files and processes them. The link text is put in the forward index together with an ID of the document the link points to. In addition, the document IDs are stored in a separate database where the PageRank is computed. Now the sorter creates an inverted index from the forward index. The forward index is sorted by document IDs, the inverted index however is sorted by word IDs. Lastly, the so called “DumpLexicon” combines a list of word IDs with the lexicon which was created by the indexer. The user queries are then answered by the searcher which uses the information from the DumpLexicon and the computed PageRanks. [Brin and Page, 1998]

³⁵ <http://www.statista.com/statistics/216573/worldwide%2Dmarket%2Dshare%2Dof%2Dsearch%2Dengines/>; [accessed 01-April-2016]

³⁶ <http://www.search.ch/?q=z%C3%BCrich>; [accessed 08-February-2016]

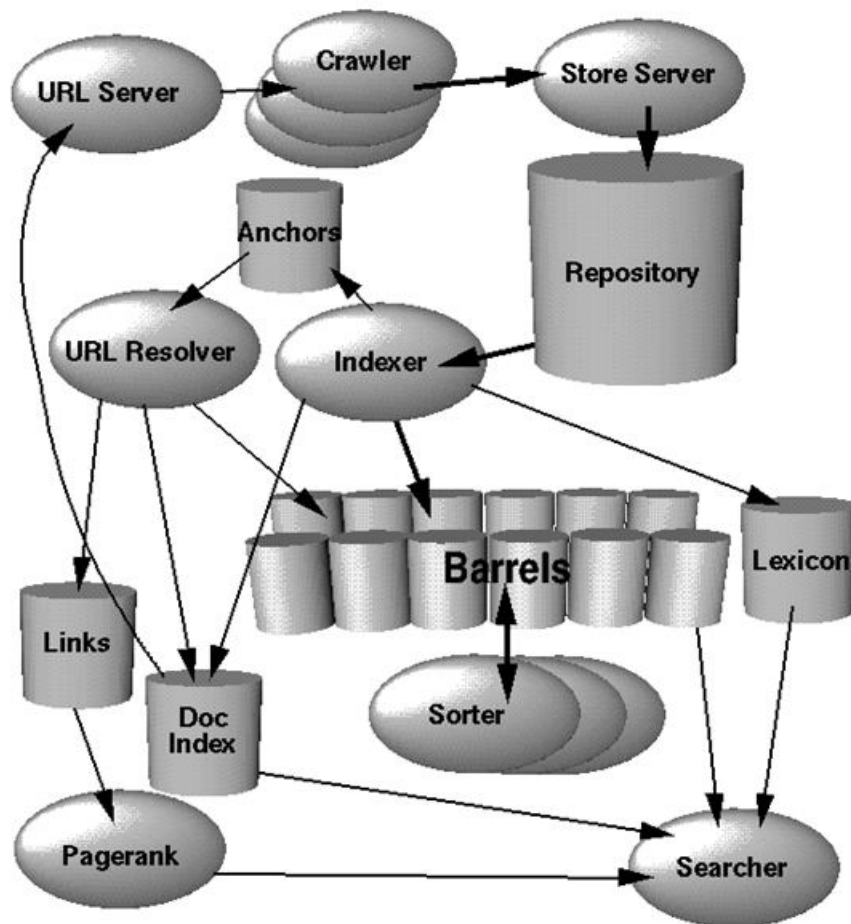


Figure 3.4: High level Google architecture (Source: [Brin and Page, 1998])

The program used in this thesis builds upon the logics and structure of search engines. Some concepts could be adopted, others were not needed or just simplified or slightly changed. As the program should not be able to answer user queries, many processing steps like the computation of the PageRank or the creation of the inverted index could be left out. An example for a step which was simplified was the creation and maintenance of anchor files. In the program there was no separate storage for anchors, URLs were just stored with the other data contained in the articles.

Presentation of the program

As presented in Figure 3.1 the number of articles on Wikipedia is enormous and still growing. Due to the sheer amount it is obvious that a large scale analysis cannot be done manually. Therefore, a program should crawl and index the Wikipedia articles, Wikidata and ULAN pages and save relevant information to a database. All further calculations and analysis tasks are performed with the data from the newly created database. In this chapter, the functionalities of the modules will be described in theory as well as on a practical example.

The program itself is structured in five modules: a crawler, an indexer, a frontier, a database manager and data-transfer objects. Besides these parts, also database queries and the name matching algorithm will be presented, as they are important for the program and further analysis steps. Finally, also problems and insights, which came up during the program run, will be described.

4.1 Program modules and example

Within the process description, also the content and quality of the data processed as well as the tasks of each program module will be described. To clarify some approaches and illustrate taken measures some excerpts of the data sources are presented within the demonstration. Figure 4.1 shows an abstract process chart of the program.

The three swim lanes group the tasks and conditions according to the module(s) they are executed in. The crawler is the controlling module, the frontier and database manager perform the URL handling as well as other database related tasks. The indexer contains the logic to extract information from websites. The frontier and database manager are combined in one swim lane because they work together very closely. Besides that, the frontier can be seen as an intermediary class between the crawler and database manager.

4. PRESENTATION OF THE PROGRAM

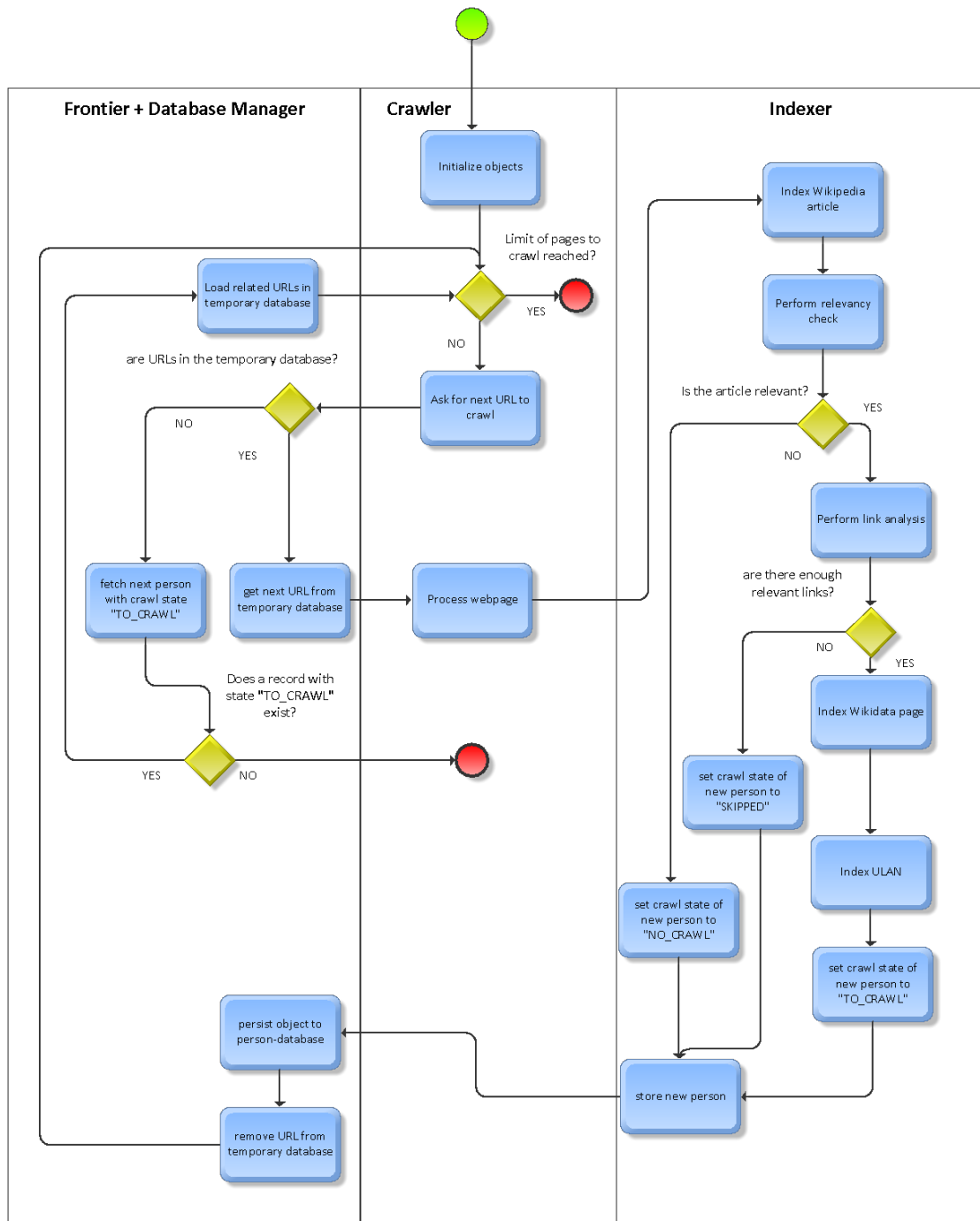


Figure 4.1: Process diagram

4.1.1 Crawler

The crawling module has multiple purposes. Regarding the process operation, it is the main module as it controls the frontier and the indexer and calls certain functions in these modules. In the crawler, the indexing process for a certain URL is initiated. The corresponding URL is supplied by the frontier.

Another purpose of the crawler is to control the runtime of the program: it either stops if the crawler gets the response from the frontier that there are no more pages to crawl or if the runtime is limited and set e.g. to 1000 pages which should be processed.

In contrast to the general architecture of Google just one crawler is on duty. Of course, multiple crawler and indexer threads could be started as well but then too many requests would be sent to the servers. Too many requests would probably lead to a block of the IP address by Wikipedia ¹.

During the program run, the crawler tells the frontier to check whether there are pages to crawl in the temporary URL database. If yes, the frontier returns the next URL to process. This URL is then passed from the crawler to the indexer for further processing.

4.1.2 Frontier and database manager

As URL handling is an important part of web-crawling, an extra class was created to perform these tasks. The main responsibility of the frontier is to maintain URLs to crawl and URLs which were already crawled.

On request of the crawler, the next URL to crawl is provided. Still, the frontier does not maintain just a simple list containing all URLs. As each person has a list of in-article links ("relatedPeople"), the frontier iterates over all persons and loads the in-article links in a temporary database. There, the URLs are then processed. As soon as all URLs from the temporary database were processed and removed, the next person and contained links are processed.

At the initial start of the program, the person and temporary URL database are empty. In this case, the URL for the Wikipedia article "List of painters by name"² is inserted in the temporary URL database, as neither ULAN nor Wikidata provide such an overview. Additionally, the "List of painters by name" is the initial starting URL as Wikipedia is the main data source for this thesis. After the indexing operation of the article "List of painters by name" there exists a record for this article in the person database. This record would need to be processed in the sense of crawling all in-article links. Hence, in the next step, all links contained in the "List of painters by name" are loaded in the temporary URL database and processed consecutively.

Each person has a crawl state (TO_CRAWL, IN_USE, CRAWLED, NO_CRAWL and SKIPPED) to provide some information about the processing status to the frontier. The crawl state "TO_CRAWL" indicates, that all related links of a person shall be

¹ https://en.m.wikipedia.org/wiki/Wikipedia:Database_download; [accessed 02-April-2016]

² https://en.wikipedia.org/wiki/List_of_painters_by_name; [accessed 02-April-2016]

crawled as well. If related links are currently processed, the state of a person is set to "IN_USE". After all related links were crawled, the state of this person is set to "CRAWLED". If an article is irrelevant and not related to art, the assigned crawl state is "NO_CRAWL". Anyway, if just the in-article links indicate, that an article is irrelevant, the status "SKIPPED" is assigned to a person. The classification of articles will be described in detail later on.

Database manipulations are performed by the database manager. The manager reads from the databases, writes to them and passes information to requesting classes like the frontier. In this example, the person database already contains some records, the temporary URL database however is empty.

On request of the crawler the frontier contacts the database manager to return the next person from the person database, where the crawl state is "TO_CRAWL". In this example it is the record of Albrecht Dürer with the URL https://en.wikipedia.org/wiki/Albrecht_D%C3%BCrer. Figure 4.2 exhibits the database entry of Albrecht Dürer.

▼ (1) https://en.wikipedia.org/wiki/Albrecht_D%C3%BCrer	{ 15 fields }
_id	https://en.wikipedia.org/wiki/Albrecht_D%C3%BCrer
className	Person
personName	Albrecht Dürer
> relatedPeople	[423 elements]
> languages	[3 elements]
> languageLinks	[3 elements]
isRelevant	true
sourceArticleUrl	https://en.wikipedia.org/wiki/List_of_painters_by_name_beginning_with_%22D%22
ulanId	500115493
crawledIn	EN
> WData	{ 7 fields }
citizenship	Holy Roman Empire
gender	MALE
> ulan	{ 3 fields }
crawl	TO_CRAWL

Figure 4.2: Database record "Albrecht Dürer"

The crawl state of the record is changed from "TO_CRAWL" to "IN_USE". Now, all 423 URLs, which are stored in the list "relatedPeople", are copied in the temporary URL database. In the temporary URL database each record has three fields:

- the class name of the data transfer object which is "URL" (as "Person" in the case of the record in the person database),
- an identifier (the specific URL) and
- a source URL which is the one of Albrecht Dürer in this example.

Now, as the temporary URL database is filled, the first URL from there is processed – it is [https://en.wikipedia.org/wiki/Self-Portrait_\(D%C3%BCrer\)](https://en.wikipedia.org/wiki/Self-Portrait_(D%C3%BCrer)). Before the URL is returned to the crawler, the person database is scanned if it already contains

a record for the concerned URL. As this is the case, the URL is removed from the temporary database without any further processing. The next URL which was fetched is `https://en.wikipedia.org/wiki/Raphael`. The record from the temporary URL database is depicted in Figure 4.3.



Figure 4.3: Record for the URL `https://en.wikipedia.org/wiki/Raphael` in the temporary URL database

Again, the person database is scanned but this time the query returned nothing which means that this URL had not been parsed yet. The URL referring to the article about Raphael is returned to the crawler for further processing.

Besides the URL handling, the frontier has another responsibility. If a website is not reachable, the indexer tries to fetch the website five times as there might be a temporary malfunction of the Internet connection. However, if the requests are not successful the frontier gets a feedback and writes the unreachable URL together with its source URL to a log file. The program does not terminate after one URL was not reachable. There exists a counter which logs how many URLs were not reachable in a sequence. If the counter reaches a value of five (this value was self-determined), the program stops. If the program would not stop, the whole data stock would be falsified. All relevant people would be processed but no linked URL would be reachable. As then all processed people would have the crawl state "CRAWLED" it would be very hard to determine, where the error occurred and which sites have to be crawled again. URLs might not be reachable if e.g.

- the Internet connection is broken,
- the server is offline,
- the IP address was blocked or
- the requested site was deleted.

On the other hand, the Internet connection might only be temporarily out of order or one specific site not reachable. If the counter for not reachable URLs is still smaller than five and another URL was reachable in the meantime, the counter is set to zero again. After the program run is complete, the crawl state of the source URLs is manually set to `TO_CRAWL` so the unreachable URL is crawled again.

4.1.3 Database manager

As mentioned above, the duties of the database manager are to persist and fetch data transfer objects to and from the database. For processing, the database system MongoDB is used. The document orientation of MongoDB was the main reason why this system was used as the database. The object-document mapper Morphia is used to map database objects to Java classes and vice versa. With the help of Morphia, Java classes can be persisted easily and read operations from the database also return the desired Java classes. The information stored in the class itself can easily be extracted with getter methods.

4.1.4 Data transfer objects (DTO)

To store all extracted information from Wikipedia, Wikidata and ULAN, three classes are used. The main class is called person (even though not all crawled pages dealt with persons) and has one nested Wikidata and ULAN class. The extracted information from the different websites is stored in corresponding fields in the three classes. Besides that, an URL class is used to store the URL together with the source URL in a temporary URL database. The field "source" of an URL is always the URL where the link appeared in the text. The source URL is later stored in the field "sourceArticleUrl" for each person. This allows to reproduce the "path" the crawler took.

4.1.5 Indexer

The indexer class contains the logic to extract the needed information from webpages. Following information is extracted from Wikipedia articles:

- The article name
- Links to other language versions of the article (not all languages were considered, only English, German and Italian articles were analysed in this thesis)
- The link to the Wikidata page
- Links contained in the article text
- The ULAN identifier and link to the ULAN page
- The article categories

Relevant article categories are predefined and were selected according to the focus of this thesis as described in chapter 1. Category terms were translated in the three selected languages so the same ones were used for each of them. For processing and further crawling operations, only articles dealing with arts are considered. Thus, the list of relevant terms, the category section is checked for, contains general keywords, notably, art, paint, graphic, illustrator, movement, style, museum and some famous art movements as Expressionism, Impressionism or Dadaism. The list of terms bases on information

contained in the Getty AAT as well as some other terms which are related to them. They were found by selecting the hierarchy view (e.g. for Baroque ³) of an art movement in the AAT as well as by doing a web search for the original term and browsing through the results. If one of the terms appears in the category section of an article, the article is considered as an interesting one in this context. A restriction of the crawling process is necessary, as otherwise, due to the large number of articles, the process would take many months to finish. Of course, a lot of different terms can be associated with arts. On these grounds it is very hard to create a complete list of terms for the classification list. As this method of classification is just one of three, it is not necessary to include every art related term in the filter list. The complete list of terms for the category classification is presented in the appendix in chapter 8.

Classification of articles with the information given in the category section is not always correct. Therefore, besides the classification with provided article categories also the in-article links are examined. In this analysis-step, the URL texts are analysed whether they contain some predefined keywords. The keywords are composed of, like above, general terms and names of art movements. Besides the predefined terms also the person database is scanned for the links contained in the article. If there already exists a record for a certain URL and this record has the crawl state "CRAWLED" or "TO_CRAWL" this link can also be regarded as relevant. As a consequence, the classification is constantly improving. If the links contain some predefined art related terms or links which were previously classified as relevant, the article will be indexed too.

The Wikidata URL contained in the Wikipedia article is used to locate the corresponding Wikidata page. To see whether there are differences between Wikidata and Wikipedia or ULAN some information is extracted from Wikidata as well. As far as available, student and student of relationships are stored. This information will be compared later on to data collected from ULAN and to links found in the Wikipedia article. Besides contained relationships also the citizenship, gender, birth- and death place, ULAN-ID (to compare it with the one found on Wikipedia) as well as the occupation is fetched. Again, only predefined occupations corresponding to the previously defined categories of artists are considered. These occupations are listed in Table 4.1.

From ULAN just the types of relations and the related people are extracted and saved. A commonality between this program and Google's architecture is the data compression. Google compresses all websites and stores them. In this thesis only the extracted information is saved and not the complete source code of webpages. In contrast to Google's architecture webpages are not downloaded in advance and processed later on by the indexer. The program rather downloads a website and analyses it before the next URL is downloaded.

To better illustrate the information extraction process, some excerpts of the different data sources are presented in the following.

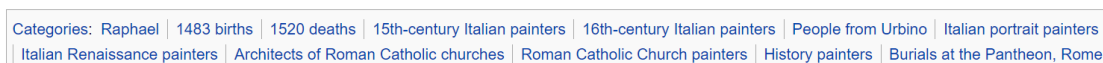
As soon as the indexer gets an URL from the crawler, it starts to fetch the information from the web servers and to parse the fetched data. In this example, the indexer got the

³ <http://www.getty.edu/vow/AATHierarchy?find=baroque&logic=AND¬e=&page=1&subjectid=300021031>; [accessed 03-April-2016]

URL referring to the article about Raphael.

The document behind the URL is fetched in HTML format and stored in a variable for further processing steps. If the website was not reachable, the indexer would try it again five times. If the site was still not reachable, the frontier would get a feedback and increment the counter for unreachable URLs.

As soon as the document is downloaded all needed information can be extracted. At first it is determined whether the article is relevant or not. For this purpose, the category links and heading of the article are scanned for words contained in the previously mentioned list of relevant terms. In the specific case of "Raphael" the categories section is depicted in Figure 4.4.



Categories:	Raphael	1483 births	1520 deaths	15th-century Italian painters	16th-century Italian painters	People from Urbino	Italian portrait painters
	Italian Renaissance painters	Architects of Roman Catholic churches	Roman Catholic Church painters	History painters	Burials at the Pantheon, Rome		

Figure 4.4: Category section from the Wikipedia article about "Raphael" ⁴

In the category section the term "painters" is mentioned several times but also "Renaissance", which is another keyword. The category classification is therefore positive.

Additionally, the in-article links are scanned for predefined terms. Already in the introductory section of the article there are links containing the text "painter" and "Renaissance". Some other links which help to classify the article contain the terms "Mannerism" (artistic style), "Museum" (appearing in "Victoria and Albert Museum") or "Baroque". Besides these terms, also other relevant ones appeared in the links. As a consequence, the link classification is positive too. Nevertheless, relevant terms which could appear in other contexts as well are "art" and "movement". The term "art" could appear in words such as "(political) party" or "article", "movement" also slots in the category politics. Of course, especially the word stem "art" appears in many other words but these were the most obvious ones which had to be filtered out. Another example for a word stem which can appear in other contexts is "graphic" from "graphic designer". It also appears in "geographic", "demographic" or "topographic" – three terms where many articles exist but which are clearly not related to arts. Moreover, some artistic movements like "Renaissance" can appear in the context of e.g. architecture too. The same is true for music, opera or symphony as these terms are art related but not relevant in this context. These articles have to be filtered out with a term-blacklist. If one of these terms appears in the text, the counter for relevant terms appearing in links can simply be reduced.

It was specified, that at least 10% of all links contained in the text must contain a relevant term so the article is considered as relevant. If this limit is not reached, the crawl state is set to "SKIPPED" and contained links in the article are not processed. At the beginning, some concerns of this method were that the predefined terms from the list are rather general. Classification would work much more reliable if the keywords also contained some names of famous artists. Still, it would be hard to draw a line which artists to include in the keyword list and which not. The trade-off made in this case was not to take artist names into the keyword list but to set the limit for required art

⁴ <https://en.wikipedia.org/wiki/Raphael>; [accessed 03-April-2016]

terms to a low level. A limit of 10% seemed to work well, as e.g. an article just has to contain three relevant links if it contains 30 links in total. Database records with the state "SKIPPED" were randomly examined manually on other keywords in the links, which might be relevant but were not considered in the first place. The list of keywords for relevant and also irrelevant topics was iteratively refined during the runtime.

Even if artist names could not be considered for link classification at the beginning – the introduction of the database scan for each link solved this problem. If there already exists a database record for the concerned link and it was already crawled or will be crawled, the link can be regarded as relevant in this case too.

In addition to that, the program checks if an ULAN-ID is available. The authority control section shown in Figure 4.5 is located right above the category section in Wikipedia articles. The link behind the ULAN identifier directly refers to the corresponding ULAN page.



Figure 4.5: Authority control section from the Wikipedia article about "Raphael"⁴

If an ULAN-ID is available, the article is automatically considered as relevant. This special case has to be considered as e.g. patrons of artists do not necessarily have the required keywords in the category section or in the links. Nevertheless, they are relevant for further crawling operations and relationship analysis.

The classification criteria for an article are applied in the following order:

1. Presence of an ULAN-ID
2. Result of the link analysis
3. Terms in the category section

A criterion on a lower level (e.g. 1.) overrules all criteria on higher levels (e.g. 2 and 3). The link analysis is more important for further analyses than the analysis of the category section because in most cases, links provide more information than the assigned categories of an article.

Depending on the relevancy of the article, the crawl state is either set to "TO_CRAWL", "NO_CRAWL" or "SKIPPED". As mentioned previously, the article about Raphael is relevant and therefore got the state "TO_CRAWL". The restriction of the crawling process to art related articles on Wikipedia helped to reduce the overall runtime of the program. The runtime was reduced even more as the Wikidata page of an article was only crawled if the article was classified as relevant. Of course, also a link from an article not dealing with arts might lead to an artist or similar. Anyhow, it is important to draw the line somewhere as it would simply take too long to crawl all articles. With a crawling speed of three to four seconds per article (including the delay), just the processing of

five million English Wikipedia articles would take around 231 days. Even with an offline version of Wikipedia the crawling process would take several months to finish as Wikidata and ULAN should not be crawled at full speed to take care of their servers and the workload. If an article was classified as irrelevant, all links contained in it were not crawled. Even if irrelevant articles might contain links to relevant articles about e.g. an artist, the probability was high that "missed" artists are crawled via a link from another art related article such as an art movement, teacher, student or colleague.

Goldfarb, Arends, Froschauer and Merkl noted, that classification of relevant articles is a very important task. They emphasised in their paper, that there are different kinds of relationships - ranging from student-teacher to artist-patron ones. Art related articles are much more diverse as they can either be about artists themselves or about paintings, museums or artistic movements. [Goldfarb et al., 2012]

In the case of the article about "Raphael", the extracted information from Wikipedia is

- the title of the article (Raphael),
- the links contained in the text,
- the links to the German, English and Italian language versions of the article (as only those are considered in this thesis),
- the link to the Wikidata page, which is located in the "Tools" section on the left side of the webpage and is displayed as "Wikidata item" ⁵, as well as
- the ULAN-identifier (500023578) with the link to the page ⁶.

The record for each article contains an array ("relatedPeople") where found in-article-links are stored. As it could not be determined recursively what the articles behind the found links are about (because the recursion would rather overload the computer before an end could be reached), they were all stored in the database. The assumption was that if an article was classified as relevant, the in-article links are interesting for the further crawling process as well. If one of the in-article links was indexed later on and turned out to be irrelevant, it was not considered in the KPI calculation.

An important advantage of Wikipedia for web crawling and indexing is, that all pages have the same format and structure. Due to the expressiveness of the HTML-tags, their unique identifiers and the common structure, the open source Java library "jsoup" can be used to find information in the HTML documents. With jsoup, HTML tags are selected and properties extracted. The following select statement filters the link to the Wikidata site for an article:

```
Elements wikiDataLink = doc.select("#t-wikibase").select("a[
    href]");
```

⁵ <https://www.wikidata.org/wiki/Q5597>; [accessed 24-February-2016]

⁶ <https://www.getty.edu/vow/ULANFullDisplay?find=&role=&nation=&subjectid=500023578>; [accessed 02-April-2016]

The variable "doc" hereby represents the HTML document which was fetched. The "#" in the first select statement indicates, that all tags with the ID "t-wikibase" are selected from the HTML code. The next select statement "a[href]" indicates, that the "a" tags with the attribute "href" should be filtered. As it could be possible that the select statements return multiple links, the results are stored in a variable of the type "Elements". This variable can then be treated like an array (iterate through it).

As mentioned previously, ULAN and Wikidata pages are just indexed if there exists a link to the corresponding ULAN page in the Wikipedia article or if the article is classified as relevant. If there exists no link to ULAN but the article is still classified as relevant, just Wikipedia and Wikidata are indexed. The other methods of classification apart from the ULAN-ID were important especially for the German Wikipedia version, as just a few of the crawled articles contained links to ULAN pages.

In the case of Raphael following information was extracted from Wikidata:

- Country of citizenship ("Italy")
- Sex or gender ("male")
- Occupation ("painter" and "sculptor" - "architect" is not considered for this thesis)
- Place of birth ("Urbino")
- Place of death ("Rome")
- Student of ("Giovanni Santi", "Timoteo Viti", "Pietro Perugino")
- Student ("Giulio Romano")

To determine the topic of the article, the program checks if the property "gender" from Wikidata is available. This property is a basic indicator if an article deals with a human being or not. Additionally, a few predefined occupations from the identically named property are taken into consideration. In Table 4.1 the selected occupations in all three languages are listed.

English	German	Italian
painter	Maler	pittori
sculptor	Bildhauer	scultore
graphic designer	Grafiker	grafico
illustrator	Illustrator	illustratore

Table 4.1: Relevant occupations in the three selected languages

A consideration of all kind of artists, for example architects or photographers would exceed timely limits as many more articles would need to be processed. Occupations

were selected based on the definition of an artist from ULAN ⁷.

The citizenship is also extracted from Wikidata and not Wikipedia, as Wikidata has an own property for it. To get the citizenship from Wikipedia, the article text had to be parsed and analysed.

On ULAN, the "Related People or Corporate Bodies" section shown in Figure 4.6 was parsed.

Related People or Corporate Bodies:

assisted by [Giovanni da Udine](#)
 (Italian painter, architect, sculptor, and stuccoist, 1487-1564) [500004700]
 assisted by [Penni, Giovanni Francesco](#) ca. 1510-1511
 (Italian painter, ca. 1496-ca. 1528) [500021943]
 assisted by [Sangallo, Antonio da, the younger](#)
 (Italian architect and military engineer, 1484-1546) [500017734]
 child of [Santi, Giovanni](#)
 (Italian painter, 1430 or 1440-1494) [500007294]
 colleague of [Raimondi, Marcantonio](#)
 (Italian printmaker, born ca. 1470/1482, died 1527/1534) [500030773]
 employee was [Perino del Vaga](#) from 1518
 (Italian painter, decorative artist, and draftsman, 1501-1547) [500000030]

Figure 4.6: Snippet of the related people section from ULAN for Raphael⁸

For each entry, the role (e.g. "assisted by"), the name of the person (e.g. "Giovanni da Udine") and the link to the ULAN record of the person are stored.

Parsing ULAN is the hardest part as the website is optimized for viewers and not programs. The HTML tags have no unique identifiers and the relationship entries often have different formats. The program has to understand

- two or one relationship descriptors at a time (e.g. "uncle/aunt of" or just "teacher of")
- different name formats ("surname, firstname" or "firstname surname" or "surname, f.")
- a varying number of "." right behind the name; sometimes there also exists an addition like a timespan of the duration of the relationship right behind the name.

Figure 4.7 shows an extract from the related people section of the Swiss painter and teacher Barthélemy Menn from ULAN. This excerpt visualizes different formats of related

⁷ <http://www.getty.edu/research/tools/vocabularies/ulan/faq.html#artist>; [accessed 02-April-2016]

⁸ <http://www.getty.edu/vow/ULANFullDisplay?find=&role=&nation=&subjectid=500023578>; [accessed 24-February-2016]

people on ULAN. The entry of Eduouard Castres has a format which is easy to parse, for Ferdinand Hodler the situation is more difficult. Right behind Hodler's name there is a short description of his role in Menn's life. The crucial task is to separate where the name ends and where the description starts (if available). In this case, as there is a hyperlink behind Ferdinand Hodler, the text of the link can be used as the name of the related person. This helps to remove the role-description. However, not all related persons have a hyperlink behind their name and therefore some "faulty" names appear in the database as well. These entries have to be re-edited manually if they are apparent.

```
teacher of .... Castres, Edouard
..... (Swiss painter, 1838-1902) [500032021]
teacher of .... Hodler, Ferdinand Hodler became a student of Menn's at the Ecole des Beaux-Arts in Geneva in 1873
..... (Swiss painter, 1853-1918) [500027184]
```

Figure 4.7: Extract from the related people section from ULAN for Barthélemy Menn⁹

Jsoup can still be used to get the whole HTML content or select links or tags for the subheadings which are bold. Then, the text has to be parsed manually.

Another inconsistency in ULAN is that blank spaces are sometimes entered as " " and sometimes encoded with the HTML tag " ".

All gathered information from the three data sources is stored in entities of the different DTO classes.

4.1.6 Data transfer objects

The screenshot from Figure 4.8 below exhibits the populated DTO for Raphael stored in the person database. The primary key of each record is the URL. It would also be possible to use the article name as the identifier but then the differentiation between the language versions of the articles had to be done in another way. Beneath the ID field, the DTO class is listed, the extracted name of the article and the related people (links found in the article). As Raphael is a very famous artist, the article is quite long and contains many links to other articles – exactly 305 (in February 2016). This proportion - one URL crawled and 305 new ones to crawl - illustrates that the number of URLs to crawl grows much faster than the number of pages crawled. Right beneath the related people some language information is saved as well as the flag if the article is relevant and the source article URL. The last information from Wikipedia are the ULAN ID as well as the language the article was crawled in. Apart from the citizenship and gender, all information extracted from Wikidata is saved in the WData object. Together with information about the birth- and death place, also the ULAN identifier, occupation (from the predefined list) and student/student of relationships are stored. The last object is the ULAN class, containing related people, their roles and the links to the corresponding ULAN pages if available.

⁹ <https://www.getty.edu/vow/ULANFullDisplay?find=&role=&nation=&subjectid=500025799>; [accessed 24-February-2016]

▼ (1) https://en.wikipedia.org/wiki/Raphael	{ 15 fields }
_id	https://en.wikipedia.org/wiki/R...
className	Person
personName	Raphael
> relatedPeople	[305 elements]
> languages	[3 elements]
> languageLinks	[3 elements]
isRelevant	true
sourceArticleUrl	https://en.wikipedia.org/wiki/...
ulanId	500023578
crawledIn	EN
▼ WData	{ 9 fields }
_id	https://www.wikidata.org/wiki/...
ulan	500023578
birthplace	Urbino
birthplaceUrl	https://www.wikidata.org/wiki/...
deathplace	Rome
deathplaceUrl	https://www.wikidata.org/wiki/...
> occupation	[2 elements]
[0]	PAINTER
[1]	SCULPTOR
> studentOf	[2 elements]
[0]	Giovanni Santi
[1]	Timoteo Viti
> student	[1 element]
[0]	Giulio Romano
citizenship	Italy
gender	MALE

Figure 4.8: Database entry for the article Raphael from the English Wikipedia version

The continued screenshot from the database entry about Raphael is depicted in Figure 4.9. In this screenshot, an excerpt of the information extracted from ULAN is shown:

▼ ulan	{ 4 fields }
> relatedPeople	[17 elements]
[0]	Giovanni da Udine
[1]	Giovanni Francesco Penni
[2]	Antonio da Sangallo, the younger
[3]	Giovanni Santi
[4]	Marcantonio Raimondi
[5]	Perino del Vaga
[6]	Andrea del Sarto
[7]	Atalanta di Galeotto Baglioni
[8]	Pope Clement VII
[9]	Pope Julius II
[10]	Pope Leo X
[11]	Perugino
[12]	Luca Penni
[13]	Polidoro da Caravaggio
[14]	Giulio Romano
[15]	Tommaso Vincidor
[16]	Pietro di Giacomo Rosselli
> linksToRelatedPeople	[17 elements]
> roles	[17 elements]
[0]	assisted by
[1]	assisted by
[2]	assisted by
[3]	child of
[4]	colleague of
[5]	employee was
[6]	influenced

Figure 4.9: Database entry for the article Raphael showing the ULAN section

The populated DTO is now passed to the database manager to persist it to the person database.

4.1.7 Frontier and database manager

As soon as an URL from the temporary URL database was processed, it is removed from there as it is not needed any more. Thereby, no crawl states are needed in the temporary URL database as only URLs which still need to be processed reside there. In this example, the URL for Raphael was processed and removed from the temporary URL database.

4.1.8 Crawler

The process for the article Raphael is now complete. The crawler checks if the program was started with a runtime limitation of a certain number of pages to crawl. As this is not the case, the crawler starts a new iteration and advises the frontier to get the next URL to crawl.

Some interesting statistics about the program lifespan are listed in Table 4.2.

Runtime of the program	around three months (speed varying between 3-5 seconds per article)
Crawled Wikipedia articles	1.516.254
Crawled Wikidata pages	97.113
Crawled ULAN pages	63.495
Crawled articles with listed gender "male"	77.859
Crawled articles about male persons with an ULAN-ID	49.613
Crawled articles with listed gender "female"	10.593
Crawled articles about female persons with an ULAN-ID	4.818
Persons with related persons listed in ULAN and Wikidata	3.857 (707 in the German, 2.147 in the English and 1.003 in the Italian version)
Persons with related persons listed in ULAN	27.027 (4.690 in the German, 16.077 in the English and 6.260 in the Italian version)
Persons with related persons listed in Wikidata	6.003 (1.401 in the German, 3.208 in the English and 1.394 in the Italian version)
Articles which were classified as irrelevant (category-/heading-/link-analysis; crawl state: "SKIPPED")	298.104
Final size of the database	18.150 MB

Table 4.2: Statistics of the program run

4.1.9 Queries

As part of the data handling, large amounts of data are processed and a lot of information is stored in a database. To write data to the database and fetch it from there, queries are needed. They are important during the crawling and indexing operations but later on also for the calculation of key performance indicators (KPIs).

For the calculation of KPIs, there are three ways of selecting data from the database after the crawling process was finished.

The first one is by accessing the Mongo database directly from the shell. This may be a fast and easy solution but it is not suited for large data sets as the output is hard to read. Figure 4.10 illustrates the output of a console query by ID to find the database entry for the previously crawled artist “Raphael”. The image is cropped as the complete output would be too long to display.

```
> db.getCollection("Person").find(<<_id: "https://en.wikipedia.org/wiki/Raphael">>
{ "_id" : "https://en.wikipedia.org/wiki/Raphael", "className" : "Person", "personName" : "Raphael", "relatedPeople" : [ "https://en.wikipedia.org/wiki/Urbino", "https://en.wikipedia.org/wiki/Marche", "https://en.wikipedia.org/wiki/Rome", "https://en.wikipedia.org/wiki/Italy", "https://en.wikipedia.org/wiki/Painting", "https://en.wikipedia.org/wiki/Architecture", "https://en.wikipedia.org/wiki/High_Renaissance", "https://en.wikipedia.org/wiki/American_English", "https://en.wikipedia.org/wiki/Italians", "https://en.wikipedia.org/wiki/Painter", "https://en.wikipedia.org/wiki/Architect", "https://en.wikipedia.org/wiki/Neoplatonism", "https://en.wikipedia.org/wiki/Michelangelo", "https://en.wikipedia.org/wiki/Leonardo_da_Vinci", "https://en.wikipedia.org/wiki/Vatican_Palace", "https://en.wikipedia.org/wiki/Raphael_Rooms", "https://en.wikipedia.org/wiki/The_School_of_Athens", "https://en.wikipedia.org/wiki/Stanza_della_Segnatura", "https://en.wikipedia.org/wiki/Printmaking", "https://en.wikipedia.org/wiki/Giorgio_Vasari", "https://en.wikipedia.org/wiki/Umbria", "https://en.wikipedia.org/wiki/Florence", "https://en.wikipedia.org/wiki/Giovanni_Santi", "https://en.wikipedia.org/wiki/Federico_III_da_Montefeltro", "https://en.wikipedia.org/wiki/Condottiere", "https://en.wikipedia.org/wiki/Duke_of_Urbino", "https://en.wikipedia.org/wiki/Pope_Sixtus_IV", "https://en.wikipedia.org/wiki/Papal_States", "https://en.wikipedia.org/wiki/Mas
```

Figure 4.10: Extract of a database query via the command line interface

Already when just one result is returned, the readability of the output is rather bad because contained information is listed sequentially. The readability decreases even further with a larger number of results. So it is much more comfortable to use database management tools like "Robomongo" to look something up or edit datasets manually.

Database management tools like Robomongo are the second way to access the database. Robomongo offers a graphical user interface to administer databases (e.g. show information about them, delete them, insert/modify/delete records). Furthermore, records can be shown or selected more specifically. Selects are not limited to basic queries but also more complex scripts can be written and executed via the user interface.

Lastly, data can be accessed in a programmatic approach. Morphia offers some methods how to easily select data from a database. If e.g. the number of linked Italian artists from articles about German artists shall be determined

- firstly, all German artists would have to be selected and
- for each of the "related persons" it had to be selected whether the linked person is an artist and whether he or she is Italian.

The programmatic approach was used to calculate the different KPIs presented in chapter 5.

In contrast to Google's query processor the program does not have to deal with user queries. The queries used in this thesis are all predefined. Of course, queries are also used to retrieve data sets with certain characteristics like all Italian artists but the information is not directly passed to the user. Retrieved information will be analysed and just final, aggregated results are presented.

Queries also play an important role regarding the program performance. The runtime of a query increases together with the number of records in the database. Queries can be limited to return only a certain number of results (to reduce the runtime and conserve resources). The following Java command returns one record from the person-database where the property crawl is equal to "TO_CRAWL".

```
Person p=dsPers.find(Person.class).field("crawl").equal(
    CrawlState.TO_CRAWL).limit(1).get();
```

This command is used to retrieve the next person, whose links are going to be processed. A limit of one result is enough in this case as only one person is processed at a time. On the scale of things that this query was executed many hundred thousand times, a lot of time was saved. The query could stop looking for other results as soon as the next person was found. The analysis of the two queries (with and without limit) also confirmed this. Both queries were executed at a time when the person database contained around 401.000 entries. Figure 4.11 illustrates the query execution statistics for the query without the limit:



Figure 4.11: Analysis for a query without a limit

The field "totalDocsExamined" indicates, that 401.301 documents had to be scanned (the whole database) to find the result for the query. In total, the query returned 107.446 results (field "nReturned"). The time, the query took to finish is displayed in the field "executionTimeMillis" (around 225 seconds). Figure 4.12 shows the same execution statistics but this time for the limited query.

Key	Value
(1)	{ 4 fields }
queryPlanner	{ 6 fields }
executionStats	{ 6 fields }
executionSuccess	true
nReturned	1
executionTimeMillis	143
totalKeysExamined	0
totalDocsExamined	30395
executionStages	{ 13 fields }
serverInfo	{ 4 fields }
ok	1

Figure 4.12: Analysis for a query with a limit

The query only returned one document, had to scan 30.395 documents and therefore only needed 143 milliseconds to execute.

As already mentioned, together with the number of records in the database, also the execution time of certain queries prolonged. A measure to keep the query runtime low, especially for queries like for records with a certain crawl-state which were needed very often, a database index was built on the column "crawl". Figure 4.13 depicts the execution statistics for the same query as above. At the time this query was executed, the Person database contained more than 830.000 records.

Key	Value
(1)	{ 4 fields }
queryPlanner	{ 6 fields }
executionStats	{ 6 fields }
executionSuccess	true
nReturned	124406
executionTimeMillis	88611
totalKeysExamined	124406
totalDocsExamined	124406
executionStages	{ 14 fields }
serverInfo	{ 4 fields }
ok	1

Figure 4.13: Analysis for a query with an index on the field

Even though the database contained more than twice as many records, the execution time was much shorter than for the query without a limit and index. The index can even improve queries with limits. Where a query with limit and without index took 143 milliseconds on a database with 401.000 records, the same query did only take one millisecond on the database with more than 830.000 records but an index on the "crawl"-field.

Queries were also important for the validation of the collected data. Before the data stock was analysed with key performance indicators, queries were used to select "faulty" records. Exemplary anomalies, database records were tested on, are

- names of related people from Wikidata or ULAN which are abnormally long (more than 40 characters),
- articles about people with an ULAN-ID or filled occupation from Wikidata where the in-article links were not crawled or
- different sizes of the lists "relatedPeople", "linksToRelatedPeople" and "roles" of the ULAN-object of a person.

Such anomalies were corrected manually as only few records were affected.

4.1.10 Matching algorithm

After all data was collected, the matching algorithm has to compare related people found in ULAN and Wikidata with the links contained in the Wikipedia article. The basic idea is to iterate over all related people from these data sources and check whether the name of a person linked in the Wikipedia article corresponds to the names found in ULAN and Wikidata.

One of the main questions in this thesis is the comparison of Wikipedia, Wikidata and ULAN. As a consequence, matching of names from the different data sources is a very important part in the analysis and comparison process.

Names could not just be compared as a whole with an equality test, because

- names can have different formats (like a related person from Giovanni Battista Piazzetta: "Tischbein, Johann Heinrich, I" as named in ULAN¹⁰ and "Johann Heinrich Tischbein" as named in Wikipedia¹¹; the detailed description is contained in the section "Indexer") and
- names are not always complete (a person can be listed with a second name in one data source and without it in another data source).

To overcome these problems, names have to be split and the parts compared separately. The matching algorithm will be presented by reference of the database record of Raphael from the English Wikipedia version. First, for each person the names of all related people listed in ULAN are fetched. The ULAN object is shown in Figure 4.14.

¹⁰ <http://www.getty.edu/vow/ULANFullDisplay?find=&role=&nation=&subjectid=500003532>; [accessed 16-June-2016]

¹¹ https://en.wikipedia.org/wiki/Johann_Heinrich_Tischbein; [accessed 16-June-2016]

▼ ulan	{ 4 fields }
▼ relatedPeople	[17 elements]
[0]	Giovanni da Udine
[1]	Giovanni Francesco Penni
[2]	Antonio da Sangallo, the younger
[3]	Giovanni Santi
[4]	Marcantonio Raimondi
[5]	Perino del Vaga
[6]	Andrea del Sarto
[7]	Atalanta di Galeotto Baglioni
[8]	Pope Clement VII
[9]	Pope Julius II
[10]	Pope Leo X
[11]	Perugino
[12]	Luca Penni
[13]	Polidoro da Caravaggio
[14]	Giulio Romano
[15]	Tommaso Vincidor
[16]	Pietro di Giacomo Rosselli

Figure 4.14: Related people from ULAN for Raphael

Each entry is processed separately. The names are split at each blank space. As a consequence, the name of the assistant of Raphael, Giovanni da Udine, is split into three separate parts. Then, each name from all related people linked in the article is fetched and again split after every blank space. In the English Wikipedia article about Raphael, there are 305 in-article links. Figure 4.15 depicts an excerpt of the list "relatedPeople", where also a link referring to the article about "Giovanni da Udine" is contained.

[120]	https://en.wikipedia.org/wiki/Baroque
[121]	https://en.wikipedia.org/wiki/Mannerist
[122]	https://en.wikipedia.org/wiki/Old_master
[123]	https://en.wikipedia.org/wiki/Giulio_Romano
[124]	https://en.wikipedia.org/wiki/Gianfrancesco_Penni
[125]	https://en.wikipedia.org/wiki/Perino_del_Vaga
[126]	https://en.wikipedia.org/wiki/Polidoro_da_Caravaggio
[127]	https://en.wikipedia.org/wiki/Maturino_da_Firenze
[128]	https://en.wikipedia.org/wiki/Giovanni_da_Udine

Figure 4.15: Snippet of the list "relatedPeople" of Raphael

Now, the name-parts from ULAN can be compared to the parts of each Wikipedia-name. For each match a counter is increased. A good method to find the best match is to store the counts for all Wikipedia names (for every name from ULAN and Wikidata). After an ULAN/Wikidata name was compared to all Wikipedia names, the name with the maximal count is selected as the best match. Anyhow, the maximal count should not be the only criterion for a best match selection. The name from Wikipedia has to match the ULAN/Wikidata name with a certain percentage. A percentage of 50% would be too low because if a person had just a first- and surname only one of them would be enough to accept the name as a match. After several experiments an acceptance rate of 80% seemed to work out well.

In the example of Raphael, the separate name parts "Giovanni", "da" and "Udine" are

compared to each article name contained in the list "relatedPeople". Each part of the name is compared separately. So, in the case of an article name which is longer than one word, like "Old master" (from Figure 4.15), "Giovanni" is compared to each part separately as well as "da" and "Udine". For each of the first elements in Figure 4.15, the counter is 0 and the acceptance level is not met:

maximal number of equal parts \geq [number of parts of the name on ULAN] $\times 0.8 \rightarrow$
 $0 \geq 3 \times 0.8 = \text{false}$

For the record of "Polidoro da Caravaggio" the counter is 1 due to the equality of "da". Nevertheless, the inequation is false (1 is not greater than or equal to $3 \times 0.8 = 2.4$). For the related person "Giovanni da Udine" the counter is 3. So, the inequation results to true as $3 \geq 3 \times 0.8$.

If a name consists of e.g. five parts, then at least 4 parts have to match. As a consequence, also names where e.g. a title is missing are recognized as equal. For example, a related person in ULAN has the name "Emperor of Austria Ferdinand I" whereas in the title of the Wikipedia article he is just called "Ferdinand I of Austria". In this example, the counter has a value of 4 whereas the second part of the inequation results in a value of $5 \times 0.8 = 4$. Hence, the inequation is true as $4 \geq 4$.

The separation into parts which are compared individually has the advantage, that the order of the parts of a name does not have to match. Otherwise, also the related person "Marquis of Mantua Ludovico III Gonzaga" from ULAN would not be considered equal to the article name "Ludovico III Gonzaga, Marquis of Mantua" from Wikipedia. The direction of comparison, from ULAN to Wikipedia, and comparison mechanism also allow article names from Wikipedia to be longer than the name of the related person in ULAN. In ULAN an exemplary person is listed under the name "John Erskine, Earl of Mar" whereas the title of the Wikipedia article is longer: "John Erskine, Earl of Mar (1675–1732)". Nevertheless, these two names are matched $5 \geq 5 \times 0.8$. It is important, that this special case also works because article names of persons quite often contain the occupation (like "Charles Robinson (illustrator)") or other information about the person. Another approach of matching in-article links and related people from ULAN is to compare the links behind the related people from ULAN to potentially available links to ULAN contained in records, the in-article links refer to. So if the algorithm tries to match "Giovanni da Udine" from ULAN to one of the in-article links from Wikipedia, the program checks for every person in the list "relatedPeople" of Raphael if it has an own ULAN ID and link to an ULAN page. If the link to the ULAN page of the related person is equal to the link to the ULAN page stored in the list "linksToRelatedPeople" (from the ULAN-object, see Figure 4.9), the related persons can be regarded as equal. As not all relationships from ULAN have an URL behind the name (as some people do not have an own page), this approach is not suitable for every record.

4.2 Problems occurred and insights gained

In this thesis, information was obtained from three data sources. There, information is stored and presented in different formats. Besides the different formats, the amount of data to be processed was enormous. Due to these two circumstances, different problems occurred and insights were gained. Some major ones are presented in this chapter.

In this context, another aspect which has to be considered, is the high number of links in articles. Some articles contain just 30 links or less in the text, others however more than 1000. The blue line in Figure 4.16 visualizes the sum of links gathered from crawled articles from the German Wikipedia, the red line visualizes the linear trend of processed articles. It is clearly visible, that the list of URLs to crawl grew much faster than it could be processed.

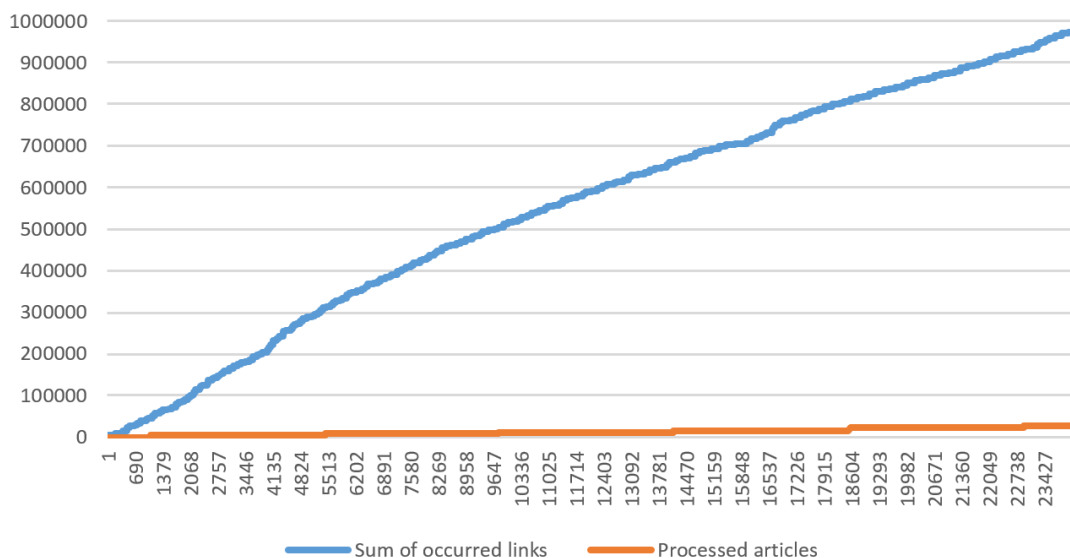


Figure 4.16: Plot of the number of collected links versus the number of processed articles in the German Wikipedia version

If e.g. array-lists would be used to store the URLs, there would soon occur problems with the working memory. Another disadvantage of the in memory list solution is, that stored URLs are lost if the program crashes or exits. Furthermore, it would be also difficult to continue crawling after a program stop.

A database containing crawled URLs and ones that are still to crawl would soon contain many thousand entries. With every insert in the database, the whole database would have had to be scanned to check if there already exists an entry for the URL. Another option would be to store all URLs in the database (even if there are duplicate entries) and just scan the crawled ones before an URL is crawled. This would however increase

the amount of storage needed. Links are saved for each person anyway. On these grounds it is much more efficient to iterate over all related persons and load links of just one person at a time in the temporary URL database. Before an URL from there is crawled and indexed, the program just needs to scan the person database for an occurrence of the URL. The scan process is very fast as the URL is the key of each database record. With this method the effort grows just linearly. The effect of this optimization was very positive.

In the first version of the program, where crawled URLs and URLs to crawl were stored together in a separate database, the URL database contained more than 500.000 entries after processing of just a few thousand articles. An average indexing operation of an article and the subsequent storage of the contained links took around 10 to 15 seconds. The duration could have been reduced with multiple and more powerful computers. Still, with this pace the program would have needed several months to finish. After the adjustment of the program (so that it iterated over all persons and copied the contained links of one person in the temporary database) it took just around one second to index an article and persist all information.

Of course, URLs could also be stored and read from a simple text file. But as the database was already in use it was much more comfortable to write and read a Java-object containing the URL and source (in which article the link was embedded) from the database. Otherwise a file format containing the required fields (URL and source URL) had to be defined and parsed separately during the write/read process. Another advantage of the database solution was, that multiple devices could crawl pages simultaneously without the need to employ a locking strategy.

Time is a crucial factor in the whole process and should not be underestimated. Crawling webpages at full speed is not allowed. The workload of servers should be kept on a low level. This request has to be satisfied as access is free and a single user should not utilize too many resources. If the number of requests exceeds a certain limit the administrators might block the source IP address. Due to the sheer amount of Wikipedia articles it is important to calculate enough time for crawling operations.

In this context, to keep the overhead at a minimal level, classification of content is a very important task which should not be underrated. If the crawling process was not restricted to relevant sites, not only more time would be needed for gathering the data but also for the calculations afterwards. Additionally, if only certain parts and not the whole website is stored, a correct classification already in the data gathering phase is extremely important. Later, if only snippets of the website are available, records are much harder to classify. Therefore, characteristics for classification should already be determined in the development-phase of the program.

Another obstacle which occurred over time was that the database size grew so much that a 32-bit system could not address the database file. Hence, the program and database had to be migrated to another computer and the crawling process had to be continued there.

Two other hurdles were content related. As already mentioned in the last chapter, the first one was the classification whether an article is relevant for this thesis and the second

one is the question whether two names are equal or not. There are no general solutions which can be applied to similar cases. Especially for the name equivalence question an individual formula might do a better job than a standard comparison method where e.g. just the complete name is tested on equivalence.

Definition, calculation and analysis of KPIs

To highlight differences among the three language versions, make results comparable and compare the content among the three sources, the following key performance indicators (KPIs) were developed. They shall answer the main research questions presented in chapter 1.

KPIs are measures which express aspects of the performance of a company, team or similar, which are crucial for the success of this organization. These measures have certain characteristics like their nonfinancial nature, their expressiveness regarding required actions or the significant impact. If measures are from financial nature, they are rather called result indicators. [Parmenter, 2007]

In the context of this thesis a large amount of articles and webpages was parsed. For evaluation purposes not only German but also English and Italian Wikipedia articles were processed. Of course, some KPIs require that an article is at least available in two (or all three) languages. For these KPIs, only articles available in more than one of the three languages are considered.

Calculations are based on the data which was collected from Wikipedia, Wikidata and ULAN. The relevancy of the defined KPIs was evaluated in a discussion with the advisor of this thesis, Professor Dieter Merkl. Professor Merkl can be regarded as domain expert as he has already written several papers and realized different projects about art (history) in context to the Internet as well as Wikipedia.

5.1 General KPIs

As it is important to get a feeling for the overall composition of the data stock, some general KPIs were defined.

5.1.1 Number of artists per language version

The total number of artists per language version is no classical key performance indicator. It is not scaled or normalized. As most of the other KPIS are normalized it is also interesting to know how large absolute numbers are. They are used to convey a feeling on which article base further calculations are performed. Furthermore, it is interesting to see how large deviations between the language versions are.

To get the numbers, all articles from a language version have to be selected (attribute "crawledIn" from the class person). For every element of this result set it has to be evaluated whether the article is about an artist or not. The following indicators were used to determine, whether a database record is about an artist (or closely art-related person) or whether it was just relevant for the crawling process (like articles about museums):

- The crawl state of the record has to be "CRAWLED" (as crawled articles are relevant, see criteria for relevancy listed in chapter "Indexer")
- In the Wikidata element an occupation should be set, a student/student of relationship listed and/or a birth-/death place set
- The property gender should be set
- Again, the presence of an ULAN-ID is a strong indicator for an artist
- Like the availability of the gender and birth-/death place also the attribute citizenship helps to differentiate between persons and things such as museums or galleries.

These criteria were also used in other KPIS for the identification of artists.

The corresponding values for the three language versions are presented in Table 5.1.

English	German	Italian
56.572	24.110	16.431

Table 5.1: Articles about artists per language version

The program ran with the same settings for all three language versions. Nevertheless, the majority of collected artists originates from the English Wikipedia version. The numbers of found and processed artists are from the same size ratio as the number of articles in the language versions. Most artists were found in the English version, which is also the largest one of the three regarding the number of articles. The second most artists were found in the second largest language version, the German one. Lastly, still a large number of artists was found in the Italian version.

10.381 artists are available in all three language versions. In Table 5.2 it is listed how

Languages	Number of artists
English-German	7.076
English-Italian	3.551
German-Italian	440

Table 5.2: Artists available in two language versions

many artists are available only in two language versions. Even though the differences between these numbers are quite large – for KPIs which directly compare two language versions only those artists are considered, who have articles in at least the compared language versions. Otherwise, results could be biased.

Table 5.3 contains the numbers of artists who are available in only one language version.

Language	Number of artists
English	35.564
German	6.213
Italian	2.059

Table 5.3: Artists available in only one language version

5.1.2 Minimum, maximum and average number of links per article

This key performance indicator is meant as an informal one and not really for comparison purposes. It provides some information how interlinked articles are amongst each other. To get the average number of links per article, the number of "relatedPeople" has to be summed up over all crawled articles and then divided by the number of crawled ones.

The average number of in article links can be calculated with finer granularity too. The calculation can be restricted to each of the three language versions to compare the results. A high average number of links can have different reasons. Either there are more articles which can be linked or people from a certain nationality write articles more carefully and enrich it with a lot of information and links. Yet, the hypothesis of more careful writers from certain countries cannot be proven as people from one country can also edit articles in other Wikipedia language versions.

Besides the average number of links per article also the minimal and maximal number of links per language version are emphasized. This shall illustrate the variance among the different articles.

This time, the evaluation is not restricted to artists but to all relevant (crawled) articles. The average number of links per article for all three language versions together is 100.7. In total, the crawled articles contain 13.348.661 links. Table 5.4 shows the results for each language version.

	English	German	Italian
Minimum	0	0	0
Maximum	2.318	1.425	1.004
Average	119.3	47.7	95.5

Table 5.4: Number of links per article

Like in the KPI "Number of artists per language version" the maximum number of links per article relates to the size of the Wikipedia version. Still, even if the Italian version is the smallest of the three, the average number of links per article is higher than in the German version. Figure 5.1, Figure 5.3 and Figure 5.4 show the boxplots of the distribution of links in the crawled articles for the three language versions. All three boxplots are cropped as otherwise the boxes would be very small due to extreme outliers in each language version.

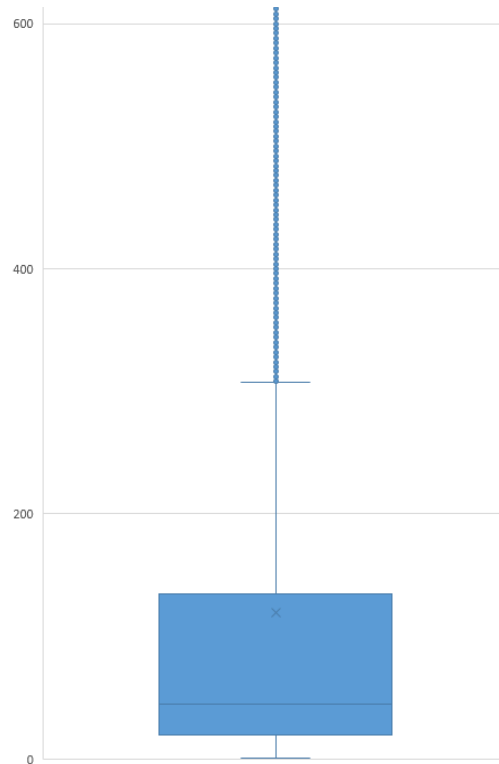


Figure 5.1: Distribution of the number of links per article in the English Wikipedia version

The lower end of the bottom whisker includes all articles with zero links. An example for such an article is the one about Doretta Frenna Smith ¹. Articles with 20 links are already in the box. The median of the box are articles with 45 links. The upper end of the box is formed by articles with 135 links. Articles with more than 308 links are already outliers and not captured by the upper whisker any more. Articles with zero links like the mentioned one above even contain a special note which asks authors to improve the article. Such a note is depicted in Figure 5.2. Of the crawled articles from the English Wikipedia version around 300 ones have more than 1000 links. Due to the fact that for this KPI all crawled articles were considered, the art-related article with the most links is the list of painters in the Web Gallery of Art ². If just relevant persons are considered, the article about Winston Churchill ³ contains the most links (1949 ones). Even though the majority of in-article links refer to political topics he was identified as a relevant person as there exists an ULAN page about him ⁴. There, he is listed as an artist and politician. The Wikipedia article contains a paragraph about his role as an artist too but he is definitely more famous for his political activities.



Figure 5.2: Suggestion to improve an article¹

Also in the German boxplot the bottom whisker starts for articles with zero links. The lower end of the box is formed by articles with 20 links. Articles with 30 links lie at the median of the box. In contrast to the English box the German one is rather short – the upper end already lies at 49 links. The upper whisker reaches to articles with 93 links. Articles with more links are outliers. Only three crawled articles have more than 1000 links. The crawled article with the most links is the "Liste deutscher Museen nach Themen" ⁵. The art-related person with the most links (1054) is Adolf Hitler ⁶. Even though the majority of links is not art-related he also has a record in ULAN ⁷ where he is listed as dictator and painter.

¹ https://en.wikipedia.org/wiki/Doretta_Frenna_Smith; [accessed 12-April-2016]

² https://en.wikipedia.org/wiki/List_of_painters_in_the_Web_Gallery_of_Art; [accessed 12-April-2016]

³ https://en.wikipedia.org/wiki/Winston_Churchill; [accessed 12-April-2016]

⁴ <https://www.getty.edu/vow/ULANFullDisplay?find=&role=&nation=&subjectid=500028788>; [accessed 12-April-2016]

⁵ https://de.wikipedia.org/wiki/Liste_deutscher_Museen_nach_Themen; [accessed 12-April-2016]

⁶ https://de.wikipedia.org/wiki/Adolf_Hitler; [accessed 12-April-2016]

⁷ <https://www.getty.edu/vow/ULANFullDisplay?find=&role=&nation=&subjectid=500119333>; [accessed 12-April-2016]

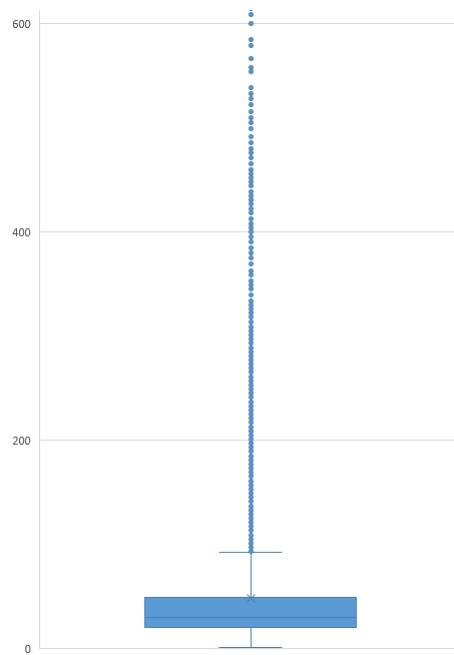


Figure 5.3: Distribution of the number of links per article in the German Wikipedia version

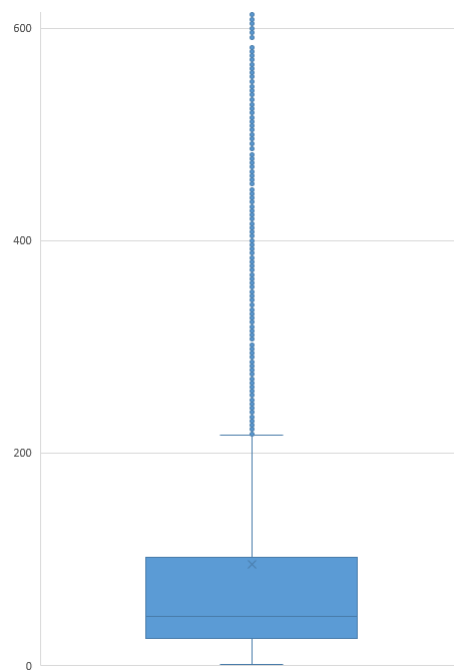


Figure 5.4: Distribution of the number of links per article in the Italian Wikipedia version

Again, the bottom whisker of the Italian boxplot reaches to articles with zero links. The bottom of the box starts with articles with 25 links. The median in the Italian version is a little bit higher than in the English one (46 links). Articles with 102 links build the upper end of the box. The upper whisker ranges to articles with 218 links. The remaining articles with more than 218 links are outliers. From the Italian Wikipedia 12 articles with more than 1000 links were crawled. Regarding art-related persons, the article about Pope John Paul II ⁸ contains the most references to other articles. Like Churchill and Hitler he was classified as relevant because there exists an ULAN record about him ⁹. There, besides the role "pope" also the role "patron" is listed.

One thing all language versions have in common is, that only a small number of articles have a very high number of links, whereas a high number of articles have a small number of links. Apart from that, the boxplots are quite different. The box of the German version is the smallest, the English box the largest. This means, that the article base of the German version is much more homogenous regarding the number of links. Another interesting observation is, that the median of all three language versions is rather low compared to the maximal number of links found.

5.1.3 Average number of links to other artists

Even though articles might have many links to other ones it is also interesting to know, how many of these links lead to articles about artists. Therefore, for this metric, all artists have to be selected from the database (as in the KPI "Number of artists per language version") and in the next step all in-article links have to be scanned for artists. Lastly, the average for all artists is calculated. The KPI conveys information how interlinked the art community is. Examples why two artists are linked could be that they are in a teacher-student relationship or if there are parallels in their work like similar techniques or an engagement in the same artistic movement.

For all three language versions together, the average number of links to other artists is 12. In Table 5.5 the average numbers are presented for the English, German and Italian language version.

English	German	Italian
14	6.44	11.03

Table 5.5: Average number of links to other artists

Despite the fact that more artists from the German version were crawled, the average number of links to other artists is lower than in the Italian version. Another specialty is that the average number of links to other artists in the English version is just slightly

⁸ https://en.wikipedia.org/wiki/Pope_John_Paul_II; [accessed 12-April-2016]

⁹ <https://www.getty.edu/vow/ULANFullDisplay?find=&role=&nation=&subjectid=500278046>; [accessed 12-April-2016]

higher than in the Italian one. One conclusion could be that articles in the English version contain just a few more links to other artists but many ones to other related articles as the average number of links per article is much higher.

The three plots presented in Figure 5.5, Figure 5.6 and Figure 5.7 show the number of links in an article which refer to other artists (x-axis) and the number of occurrences of such articles in the database (y-axis). In all three language versions the majority of articles has just a few links to other artists. The three data sets were tested with R on power law-distributions. Furthermore, the fit was tested on the lognormal-, exponential- and Poisson-distribution. The power law-distribution fitted better for the data set than the exponential- and Poisson-distribution. Nevertheless, the lognormal-distribution fitted best. Nevertheless, the similarity to the power law-distribution is, that the majority of articles only has a few links whereas a few articles have a very high number of links to other artists. Besides the difference in the number of occurrences and the maximal number of links, the plots have similar shapes.

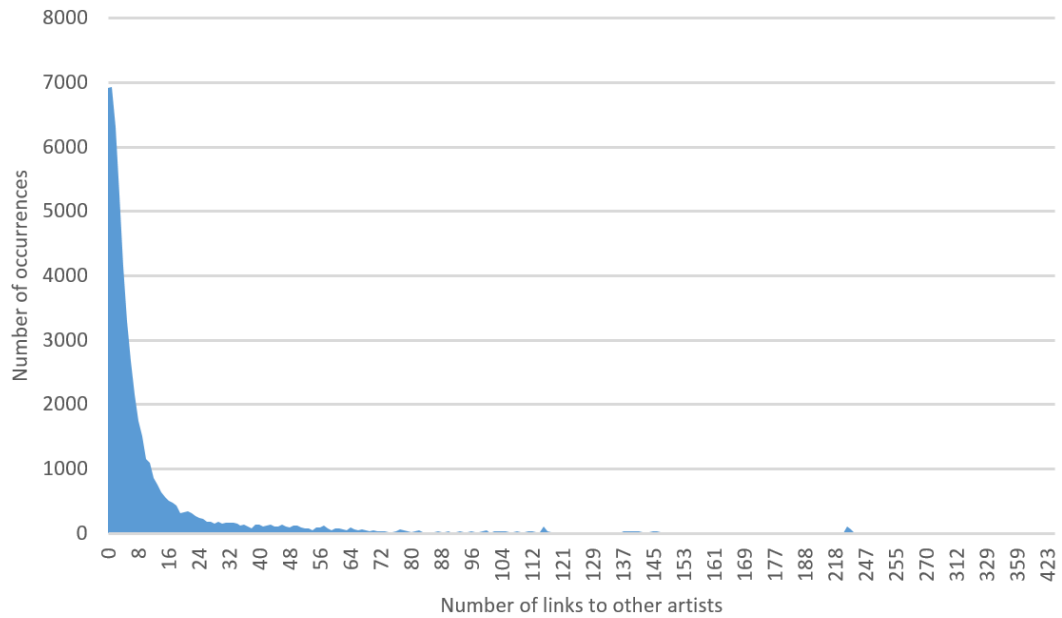


Figure 5.5: Plot for the average number of links to other artists for the English Wikipedia

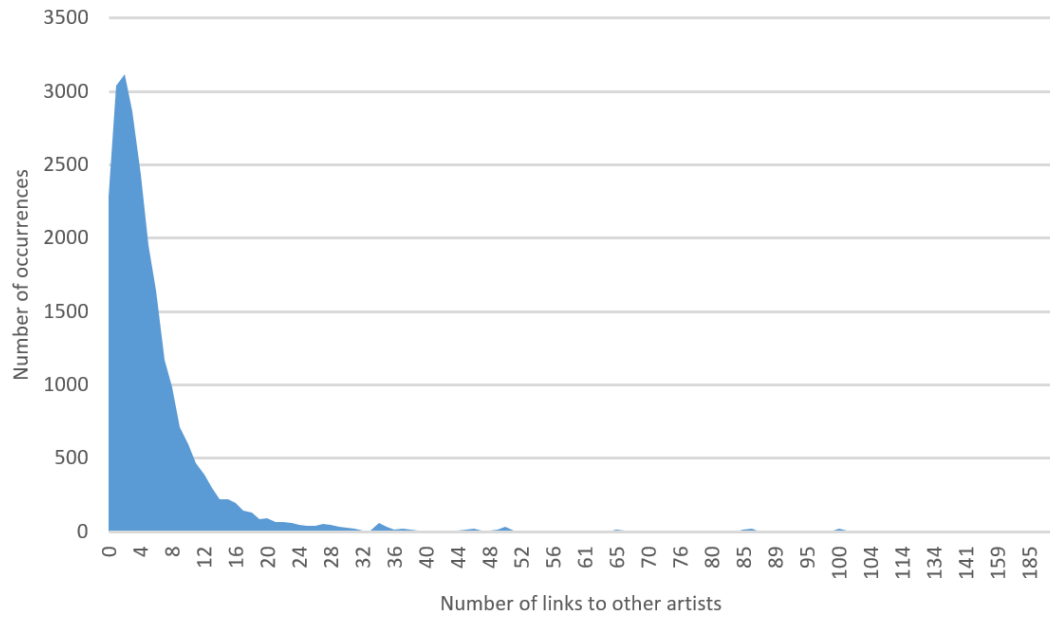


Figure 5.6: Plot for the average number of links to other artists for the German Wikipedia

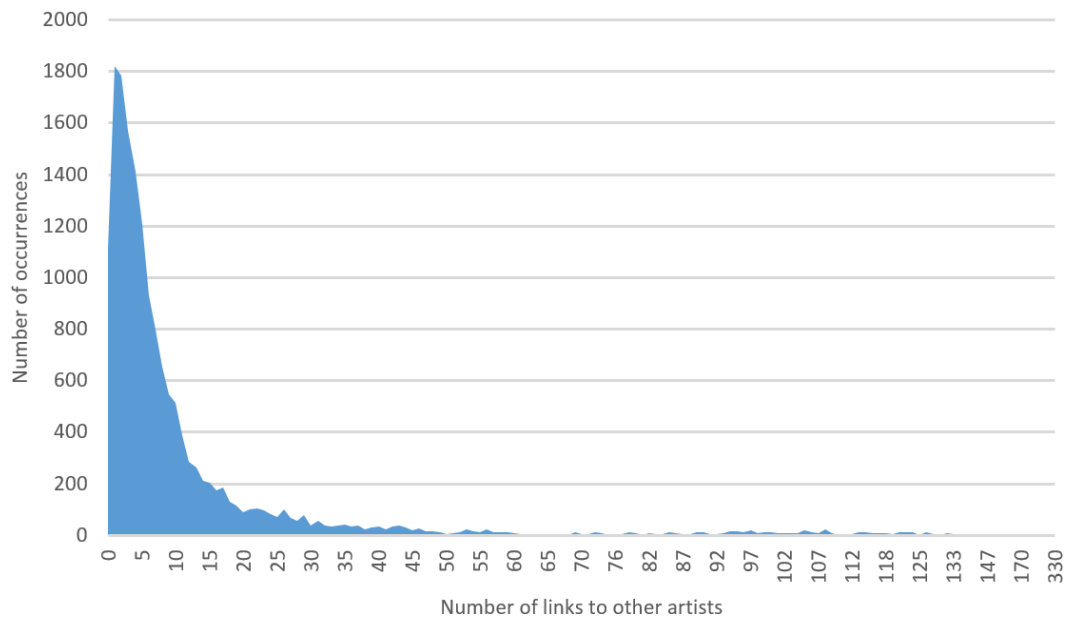


Figure 5.7: Plot for the average number of links to other artists for the Italian Wikipedia

5.1.4 Distribution of birth- / death places

As far as available, also the birth- and death place of a person was extracted from Wikidata. The birthplace of a person usually has no influence on the creativity. Hence, besides the fact that more people live and are born in cities, the probability that an artist is born in a city or a small village on the countryside should be equal. One hypothesis however is, that cities attract more people so the majority of artists died in cities and not the countryside. A reason for the attraction might be that wealth was usually concentrated in cities rather than on the countryside. Wealthy people could buy artworks and sponsor artists. As a consequence, artists were attracted to cities as the chances to realize their vision were higher there.

The distribution of birth- and death places will be analysed on

- the aggregate-level: all artists from the three language versions will be analysed together.
- the language specific level: it will be examined whether accumulations can be observed on the language level.

Popular death places will probably be Paris or Rome. In Paris, arts and culture always played an important role. New artistic styles were developed and current ones perfected. Therefore, it was always a focal point for artists. An explanation for Rome as an important death place could be, that the centre of the Catholic Church is located there. The Catholic Church sponsored many artists. One question is, if there exist famous birth- and death places in e.g. Austria or Germany too. It will be interesting to see how they compare to cities like Paris or Rome in the popularity rankings.

In Table 5.6 the most popular (based on the number of occurrences) birth- and death places gathered are listed. In this evaluation all three language versions were considered as one. Absolute numbers are listed next to the name of the place.

Rank	Birthplace	Death place
1	Paris (1.903)	Paris (3.364)
2	London (951)	Rome (1.896)
3	Florence (885)	London (1.377)
4	Vienna (753)	New York City (880)
5	Antwerp (684)	Vienna (853)

Table 5.6: Most popular birth- / death places in total

Altogether more than 9100 distinct birth- and 5000 death places were collected. The gap between these numbers may arise from different reasons. Firstly, many artists could still be alive and therefore only a birthplace is stored in Wikidata. The other reason however might be, that according to the hypothesis, there exists a concentration of death places. Table 5.7 shows the most popular birth- and death places in the English language version, Table 5.8 the ones for the German and Table 5.9 for the Italian version.

Rank	Birthplace	Death place
1	Paris (1.629)	Paris (2.857)
2	London (874)	London (1.258)
3	Florence (624)	Rome (1.258)
4	Antwerp (617)	New York City (803)
5	New York City (593)	Amsterdam (603)

Table 5.7: Most popular birth- / death places in the English language version

The heatmap shown in Figure 5.8 illustrates the distribution of birthplaces of artists across Europe. In this map, only birthplaces are marked which were found for English articles. Of course, all three language versions contain artists with birth- and death places on different continents. However, to illustrate and highlight differences, only Europe is shown in the following figures. Birth- and death places are not displayed as single dots as otherwise the illustration would be rather unclear as there are simply too many of them. Just areas where many artists were born or died are emphasized in form of “heated” areas. In the English Wikipedia there are numerous marked areas where artists were born. There exists a certain concentration on cities. Anyhow, there are many heated areas spread over different countries, not necessarily in cities.

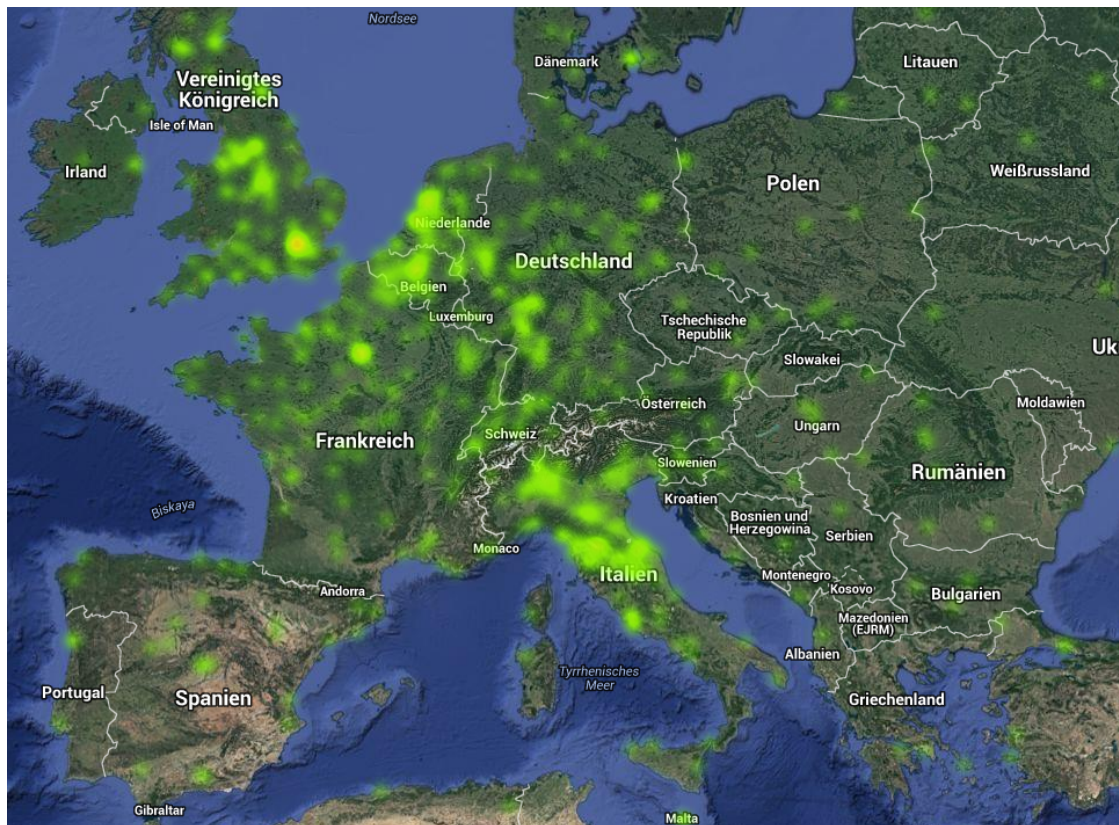


Figure 5.8: Distribution of birthplaces of artists from the English Wikipedia version

The distribution of death places exhibits a completely different picture. As illustrated in Figure 5.9 just a few places are highlighted in the map, mostly large cities. The redder an area is marked on the map, the higher the concentration of occurrences. The heatmap supports the hypothesis that there exists a concentration of death places to cities whereas birthplaces are distributed quite evenly. Especially Paris, London and Rome are coded lightly yellow to red.

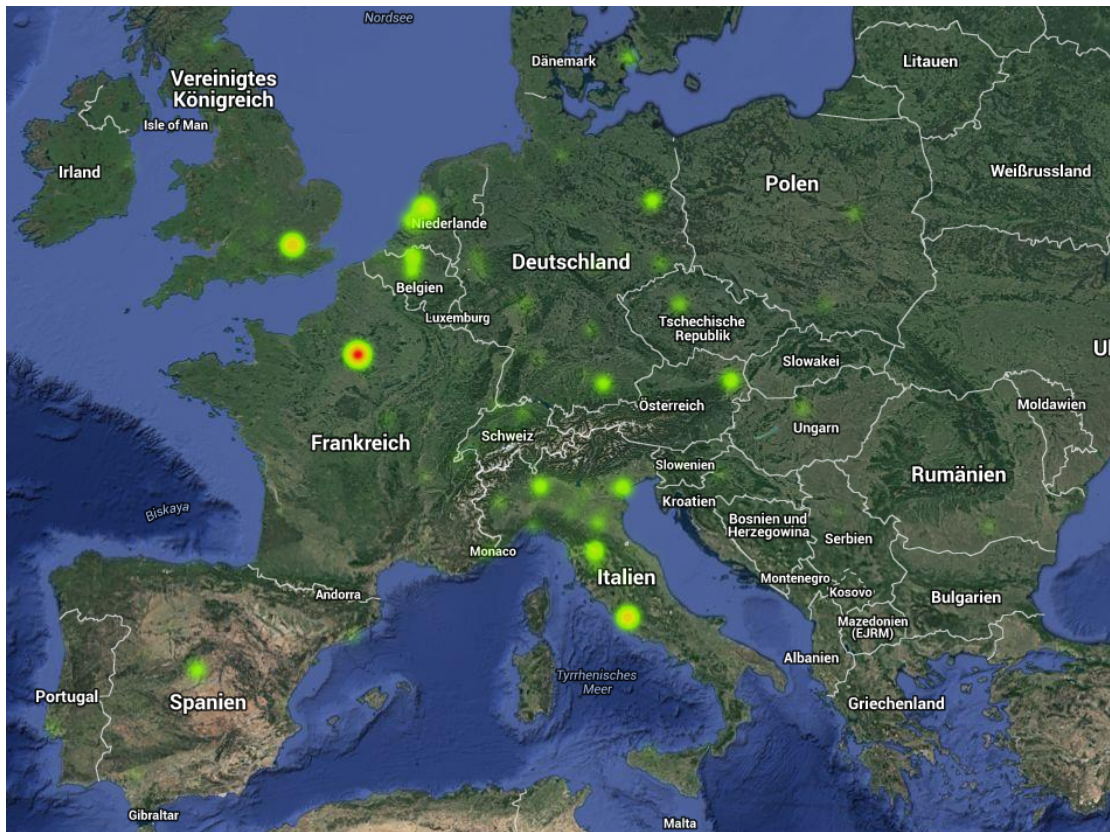


Figure 5.9: Distribution of death places of artists from the English Wikipedia version

To make it comparable, only European birth- and death places from the German and Italian version are displayed in the following figures.

The most popular birth- and death places for art-related persons found in the German Wikipedia are listed in Table 5.8.

Rank	Birthplace	Death place
1	Paris (554)	Paris (1.073)
2	Vienna (550)	Munich (695)
3	Berlin (494)	Berlin (669)
4	Munich (283)	Vienna (624)
5	London (250)	Rome (426)

Table 5.8: Most popular birth- / death places in the German language version

Like in the English version, the birthplaces shown in Figure 5.10 are scattered across Europe and especially Germany. In contrast to the English birthplaces shown in Figure 5.8 there are fewer and smaller heated spots in e.g. the United Kingdom. In return, there are more and larger heated areas in Germany.

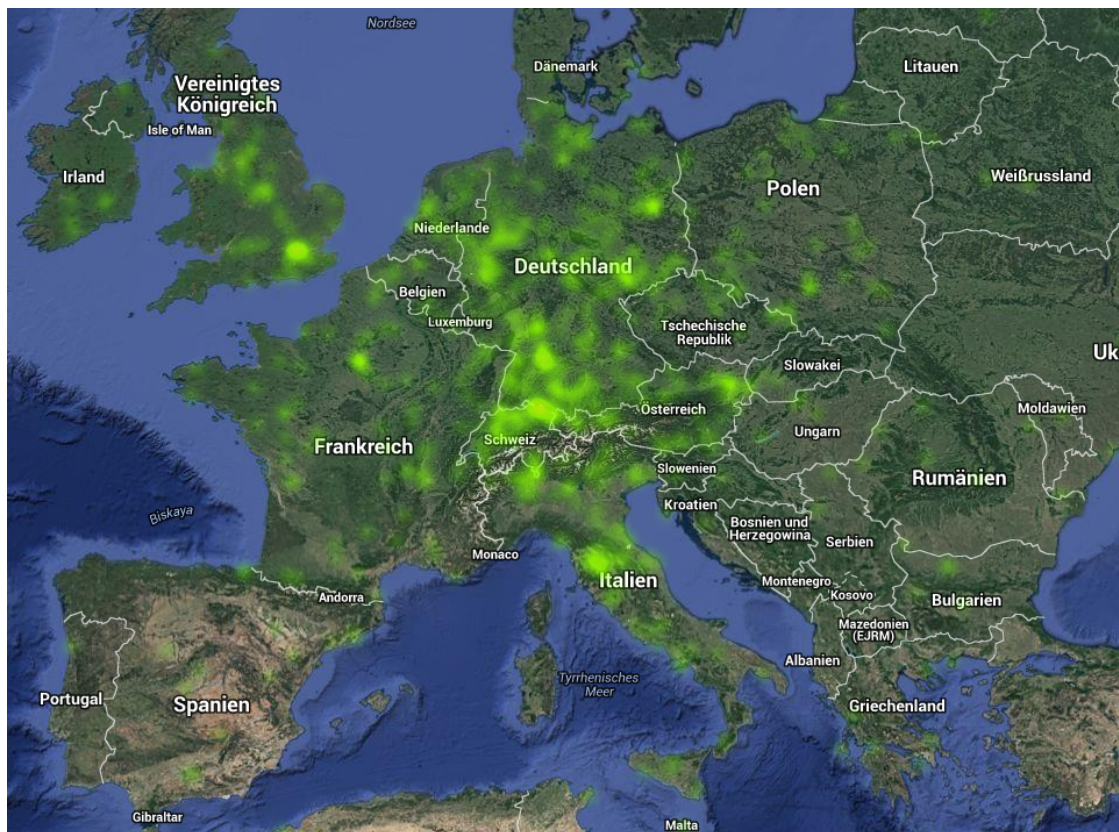


Figure 5.10: Distribution of birthplaces of artists from the German Wikipedia version

Whereas birthplaces were scattered across Europe, there is again a concentration of death places to cities such as Paris, Vienna, Munich or Berlin. The distribution of death places is presented in Figure 5.11. Besides the mentioned large cities there are still many heated spots across Germany, even in smaller cities like Stuttgart or Dresden.

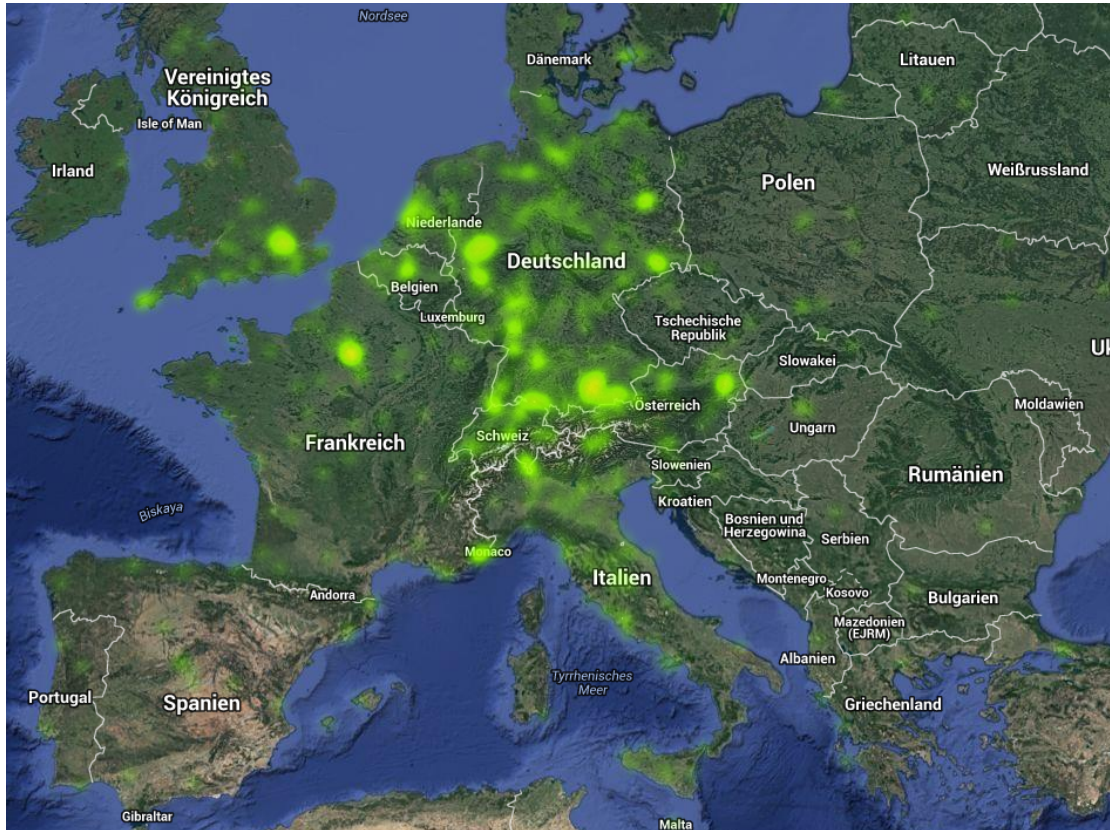


Figure 5.11: Distribution of death places of artists from the German Wikipedia version

In Table 5.9 the most popular birth- and death places for people from the Italian Wikipedia are listed.

Rank	Birthplace	Death place
1	Paris (608)	Rome (1.194)
2	Florence (578)	Paris (1.131)
3	Rome (436)	Florence (551)
4	Milan (345)	Milan (522)
5	Venice (302)	London (420)

Table 5.9: Most popular birth- / death places in the Italian language version

The heatmap of birthplaces from art-related persons contained in the Italian Wikipedia version, as shown in Figure 5.12, has an even higher concentration across Italy. Apart from that, the heated Italian regions are much larger than the German or English one in their corresponding figures. The colour is already lightly yellow where it was still mostly green in the German and English version.

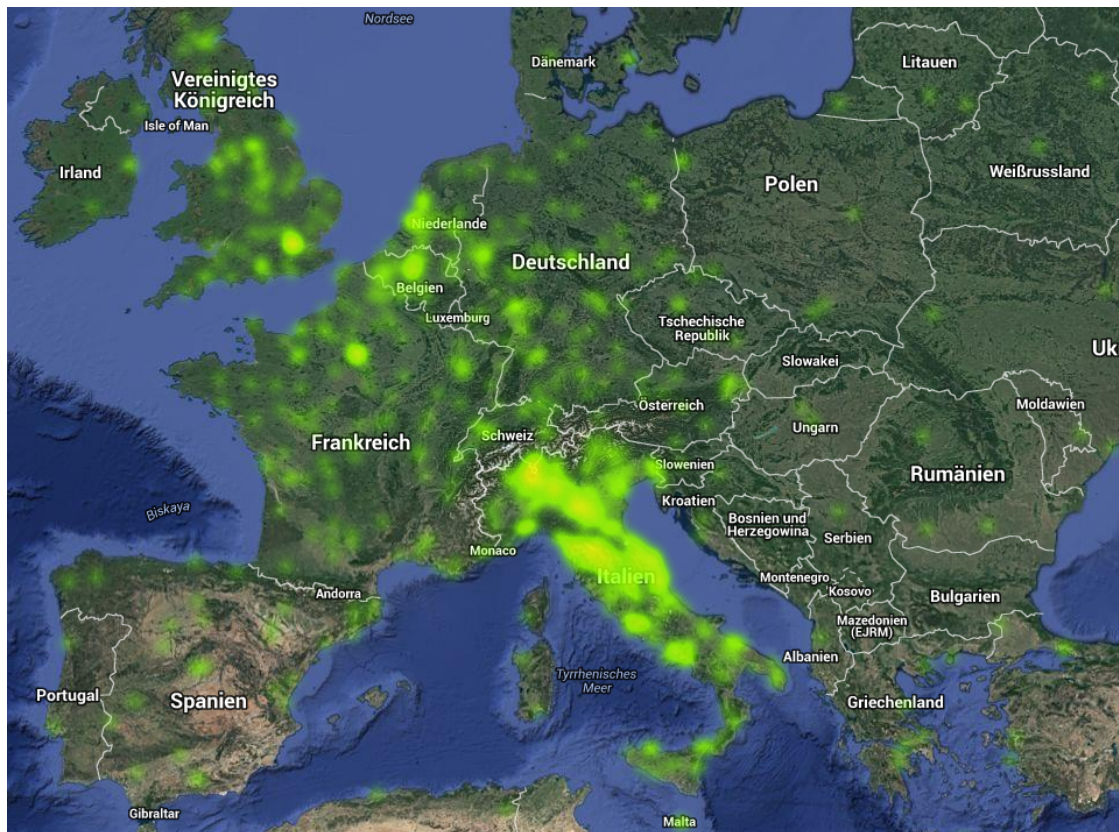


Figure 5.12: Distribution of birthplaces of artists from the Italian Wikipedia version

Figure 5.13 shows the heatmap of death places of art-related people from the Italian Wikipedia version. As in the English and German version there exists a concentration of death places to cities too. Mostly cities in Italy are highlighted. Besides them, also cities like Paris or London are very popular. They are among the top five places in the other language versions too.

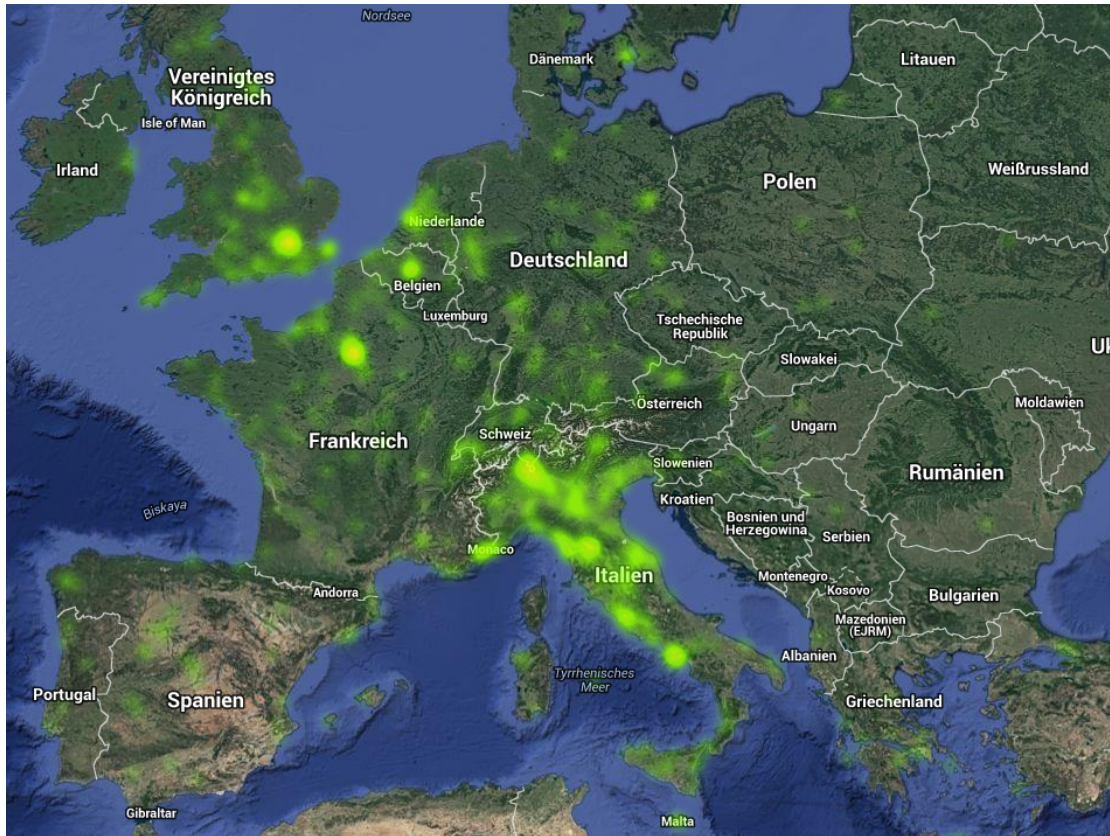


Figure 5.13: Distribution of death places of artists from the Italian Wikipedia version

The four tables and corresponding figures exhibit clearly, that the content of the Wikipedia language versions is different. All three versions have some places in common. Still, amongst the five most important ones there are always at least some differences. In e.g. the German version, the German cities and the Austrian city Vienna are ranked higher than in the other language versions. For the Italian version the situation is even more extreme as, besides Paris, all other birthplaces amongst the top five are Italian cities. Whereas New York is among the top five birth- and death places in the English version, the city is not as popular in the other two language versions. Apart from the detected differences another insight is, that the hypothesis of the concentration of death places is confirmed.

5.2 Language specific KPIS

The following KPIS focus on the comparison of language and nationality differences. In total, the German, English and Italian Wikipedia version have more than 1.9, 5.1 and 1.2 million articles (by April 2016) ¹⁰. Due to the restriction of the crawling process to art related articles, the person database contains 233.960 German, 1.072.016 English and 210.278 Italian records. The English version of Wikipedia has much more articles in total than any of the other versions. Therefore, the number of English records in the person database is also much higher than for any of the other versions. Yet, not all database records are about artists. As mentioned in the KPI "Number of artists per language version" the database contains 24.110 artists from the German, 56.572 from the English and 16.431 from the Italian Wikipedia version. Table 5.2 illustrates, that not all of them are available in two or even all three language versions. As the KPIS "Overlap" and "Size comparison" require, that an article is available in both (or all three) language versions, only those articles were considered in the calculations, which are at least available in the compared versions.

Nearly no German article contained a link to the corresponding ULAN page. That is why the ULAN objects were copied from the English and Italian records to the corresponding German ones. Otherwise, the matching rate could not be computed.

One hypothesis is, that the German and Italian language version have many commonalities (more than e.g. the German and the English version) due to the geographical closeness of the German and Italian speaking regions. The frequencies of mentioned artistic styles from all three Wikipedia versions are compared in Figure 5.14. The numbers of occurrences are normalized by the total sum of occurrences of all styles in the concerned language version. Therefore, the y-axis indicates the share of occurrences of a style in percent. Occurrences of URLs with slightly changed titles but the same artistic movement, for example, "Dada" and "Dadaism", were grouped together manually.

There are some similarities between the bars of all language versions. Famous artistic movements like Baroque, Bauhaus or Renaissance are popular in all three languages. Besides that, there are also similarities between two language versions. For instance, Impressionism has a high relative importance in the German and Italian version but is less popular in the English Wikipedia. Likewise, the situation is the same for Surrealism. Cubism or Rococo are approximately on the same popularity level in the German and English version. Whereas Art Nouveau is rather unimportant in the German Wikipedia, it is almost equally popular in the English and Italian version. Another exception is, that the German and English version have a few styles which are much more popular than the other ones. In the English version it is Renaissance whereas Expressionism, Impressionism and Renaissance are exceptionally popular in the German Wikipedia. In the Italian version the popularity is a little bit more balanced. There exist a few styles which are more popular than the others but not a single one which is clearly the most popular one. Some styles like the Danube School or Hyperrealism occur very rarely in all three versions.

¹⁰ https://en.wikipedia.org/wiki/List_of_Wikipedias; [accessed 17-April-2016]

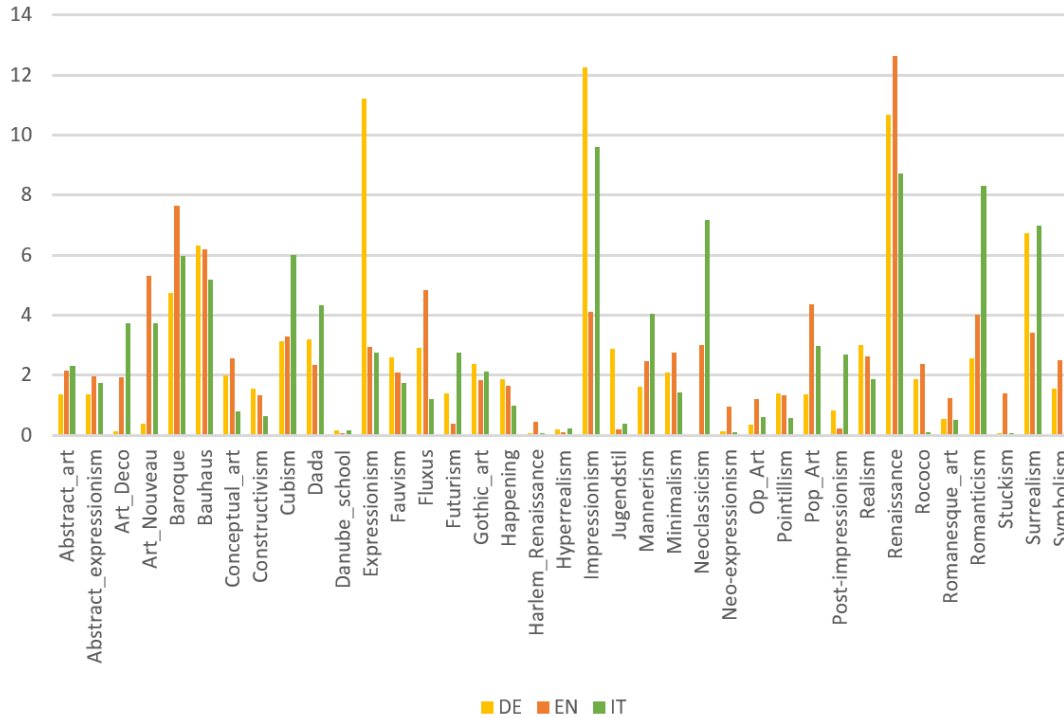


Figure 5.14: Artistic movements occurring in the three language versions

5.2.1 Average number of links to artists with the same citizenship as the current artist

This metric bases on the KPI "Average number of links to other artists". This time however, the citizenship is considered in the evaluation. Only those artists are counted, who have the same citizenship as the one where the link appeared in the text.

Nevertheless, nationalities may change over time. Some artists had e.g. the citizenship of the Holy Roman Empire which does not exist anymore. In especially this case an automatic separation is very difficult as the Holy Roman Empire was very large and covered areas of many of today's countries. Artists who have an empire as citizenship were assigned manually to one of today's countries, depending on the location of the birthplace on a current map.

The values of this KPI for the three language versions are shown in Table 5.10.

English	German	Italian
6.4	2.7	5.2

Table 5.10: Average number of links to artists with the same citizenship as the current one

Surprisingly, the average number of links to artists with the same citizenship is much lower in the German than in the Italian version. The average number in the Italian Wikipedia is nearly twice as large as the German one. Nevertheless, the ratio of the two results indeed relates to the KPI "Average number of links to other artists". As expected, the average artist in the English version has much more links to other artists with the same citizenship than the average artist in the German one. If there exist more articles about artists in a Wikipedia language version, the authors consequently have more possibilities to create links to them. Even though fewer artists were found in the Italian version, the average artist contains nearly as many links to other ones with the same citizenship as the average artist in the English version.

5.2.2 Average number of linked artists from the same nationality as the articles are crawled in

This KPI shall illustrate the assumption, that more artists from the same nationality as the Wikipedia language version are linked in articles than artists with other citizenships. The evaluation process is relatively straight-forward: First, all artists are selected from the database. In the next step their in-article links are examined:

- Does the link lead to another artist and
- does he or she has the same citizenship as the Wikipedia language version?

One aspect which has to be considered is, that a Wikipedia version has no real citizenship. So, different citizenships were grouped and assigned to a Wikipedia language version. If a language is official or co-official in a certain country, this citizenship was considered in the KPI calculation. One restriction was, that the language must not be co-official only in a certain region of that country (like German in South Tyrol). The following countries of citizenship were assigned to the German, English and Italian language version of Wikipedia:

- German version: Germany, Austria, Liechtenstein, Switzerland, Belgium, Luxembourg ¹¹
- English version: Anguilla, Antigua, Barbuda, Australia, Bahamas, Barbados, Belize, Bermuda, Botswana, British Virgin Islands, Cameroon, Canada, Cayman Islands, Dominica, England, Fiji, Gambia, Ghana, Gibraltar, Grenada, Guyana, Ireland, Jamaica, Kenya, Lesotho, Liberia, Malawi, Malta, Mauritius, Montserrat, Namibia, New Zealand, Nigeria, Papua New Guinea, St. Kitts and Nevis, St. Lucia, St. Vincent and the Grenadines, Scotland, Seychelles, Sierra Leone, Singapore, Solomon Islands, South Africa, Swaziland, Tanzania, Tonga, Trinidad and Tobago, Turks

¹¹ http://www.bbc.co.uk/languages/european_languages/languages/german.shtml; [accessed 02-May-2016]

and Caicos Islands, Uganda, United Kingdom, United States of America, Vanuatu, Wales, Zambia, Zimbabwe ¹²

- Italian version: Italy, San Marino, Vatican City, Switzerland ¹³

To make language versions comparable it is important to calculate the average for each artist. Nonetheless, it is hard to compare the English version to the other ones as English is an official or co-official language in much more countries than German or Italian. That is why the probability that artists with the same “citizenship” are linked is much higher. The results of this evaluation are shown in Table 5.11.

English	German	Italian
3.4	2.6	3.0

Table 5.11: Average number of links to artists with the same citizenship as the Wikipedia version

Whereas the difference in the previous KPI is rather large, the German and Italian language versions are nearly on par in this KPI. Still, even if the English Wikipedia version has more "citizenships" than the German one, the average number of linked artists with these citizenships is not much higher than in the German version. One explanation might be that much more artists were found in the English version. Among them, also (foreign) ones are contained who do not have a single reference to an artist with the same citizenship as the English Wikipedia version. Concretely, there are 31.604 such persons without a reference to an "English" artist. Again, the average artist from the Italian Wikipedia version has just marginally fewer links to artists with the same citizenship as the Wikipedia version than the average artist from the English version.

5.2.3 Completeness regarding ULAN and Wikidata

For the calculation of the completeness rates, related people found in ULAN and Wikidata are compared to relationships contained in Wikipedia articles. The difficulty within this comparison process is, that names are not always consistent. As mentioned previously, names can have different formats. Therefore, the names gathered from ULAN were reformatted already during the crawling process. If names did not contain any ",", they were not reformatted. If they contained a comma, the part after the comma was prepended. This is exemplified by the patron of the Italian painter and sculptor Andrea del Verrocchio ¹⁴. His patron is listed in ULAN as "Medici, Lorenzo de' ". For further

¹² https://www.ncsu.edu/grad/handbook/official_language_english.htm; [accessed 02-May-2016]

¹³ http://www.bbc.co.uk/languages/european_languages/languages/italian.shtml; [accessed 02-May-2016]

¹⁴ http://www.getty.edu/vow/ULANFullDisplay?find=Andrea+del+Verrocchio&role=&nation=&prev_page=1&subjectid=500003951; [accessed 03-May-2016]

comparisons the name was persisted in the database as "Lorenzo de' Medici". Names had to be reformatted this way because Wikipedia articles are mostly named in the format "Firstname Lastname". For the comparison of names from ULAN, Wikipedia and Wikidata the matching algorithm described in chapter 4.1.10 was used. Again, each language version was compared separately. For each artist in each language version the matches are counted and summed up. Lastly, the sum is divided by the total number of related people listed in ULAN/Wikidata. These shares are then averaged for all artists from one language version. Table 5.12 shows, how many relationships from the considered source were found in the corresponding Wikipedia articles.

	English	German	Italian
Completeness regarding ULAN	30%	20.8%	17.7%
Completeness regarding Wikidata	56%	39%	35.6%
Total completeness regarding ULAN	33% for articles which are available in only one language version, 39% for articles which are available in multiple language versions		

Table 5.12: Completeness rates to other sources of information

Even though the Italian version had advantages in the first KPIs, the completeness rates regarding the relationships listed in ULAN and Wikidata are higher in the German version. Of the listed relationships in ULAN on average around 20.8% are also present as links in the German articles. In the Italian version only around 17.7% are covered as links in the articles. Naturally, as the Getty Research Institute has its office in the USA, ULAN is probably more popular in the English speaking regions. Therefore, and due to the highest number of artists and the highest number of links to artists in the English Wikipedia, the completeness rate is the highest in this version.

The completeness rates regarding relationships from Wikidata are higher than the ones of ULAN – around 39% of the collected relationships are contained as links in German articles. In the Italian version around 35.6% of the relationships are incorporated as links in the articles. The English Wikipedia outperformed both other versions as more than half of the listed relationships in Wikidata are also included as links in Wikipedia articles.

For the total completeness regarding ULAN, related people from all available Wikipedia language versions of an article were combined and matched to related people listed in ULAN. Maybe completeness has to be seen in a broader context. Maybe the different language versions should not be seen as separate Wikis but rather as one big encyclopaedia. Probably, the goal would then be to be complete in total and not in one single language version. The KPI "Total completeness regarding ULAN" shows, that if the links of multiple language versions are combined, more matches to the relationships from ULAN are achieved. If only one language version (regardless which one) of an article was

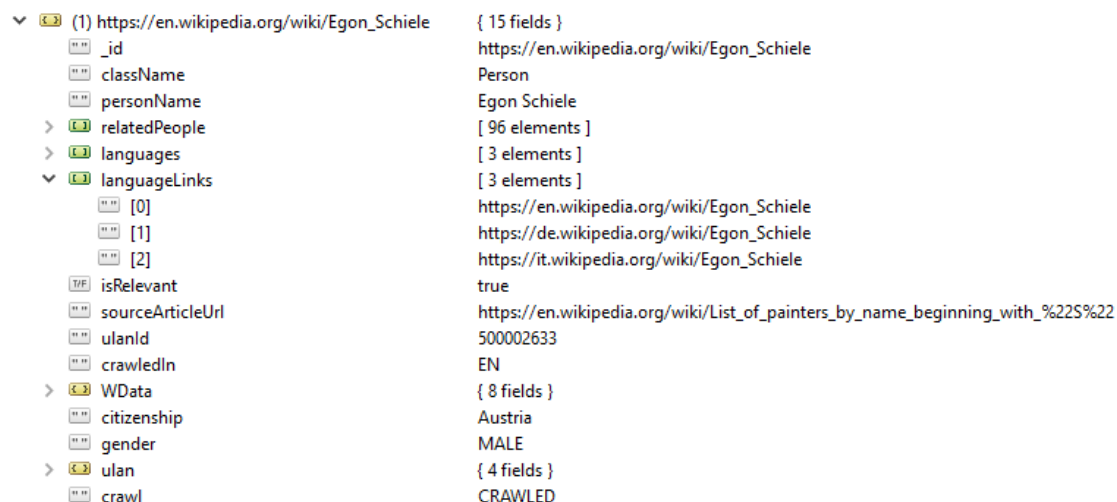
available, usually 33% of the listed relationships from ULAN were contained in the Wikipedia article. If multiple language versions were compared to the relationships from ULAN, 39% of the them were found in the articles. This rate is much higher than the completeness rate of any of the three language versions.

5.2.4 Overlap of links between language versions

Another question which can be answered with the gathered data is, how many links Wikipedia articles from different language versions have in common. The calculation is done for one article at a time and at the end the results are averaged. Only crawled articles will be included in the calculation of this KPI. In contrast to the previous language specific KPIs which only focus on one language version at a time, this KPI compares articles in two language versions the same time. On these grounds only articles which are available in the two compared languages will be considered in the calculation. The calculation is explained by the example of the English article about Egon Schiele.

The IDs (URLs) of the other language versions are stored in the list "languageLinks" as depicted in Figure 5.15.

As only crawled articles are considered for this KPI, all links contained in the list



(1) https://en.wikipedia.org/wiki/Egon_Schiele	{ 15 fields }
_id	https://en.wikipedia.org/wiki/Egon_Schiele
className	Person
personName	Egon Schiele
relatedPeople	[96 elements]
languages	[3 elements]
languageLinks	[3 elements]
[0]	https://en.wikipedia.org/wiki/Egon_Schiele
[1]	https://de.wikipedia.org/wiki/Egon_Schiele
[2]	https://it.wikipedia.org/wiki/Egon_Schiele
isRelevant	true
sourceArticleUrl	https://en.wikipedia.org/wiki/List_of_painters_by_name_beginning_with_%22S%22
ulanId	500002633
crawledIn	EN
WData	{ 8 fields }
citizenship	Austria
gender	MALE
ulan	{ 4 fields }
crawl	CRAWLED

Figure 5.15: Database record for the English article about Egon Schiele

"relatedPeople" were crawled as well. A snippet of Egon Schiele's list "relatedPeople" is shown in Figure 5.16. For each of these in-article links the language versions are fetched. If, for example, the English article about Egon Schiele is compared to the German version, the program iterates over all records in the list "relatedPeople" from the English article. Each of them is compared to all links from the German article. If e.g. the link to the English article about Gustav Klimt is processed, the program iterates over all links from the German article about Egon Schiele and fetches the database record for each of those links. Each record also contains a list with URLs ("languageLinks") pointing to the other

▼ (1) https://en.wikipedia.org/wiki/Egon_Schiele	{ 15 fields }
_id	https://en.wikipedia.org/wiki/Egon_Schiele
className	Person
personName	Egon Schiele
▼ relatedPeople	[96 elements]
[0]	https://en.wikipedia.org/wiki/Tulln_an_der_Donau
[1]	https://en.wikipedia.org/wiki/Austro-Hungarian_Empire
[2]	https://en.wikipedia.org/wiki/Vienna
[3]	https://en.wikipedia.org/wiki/Austria
[4]	https://en.wikipedia.org/wiki/Academy_of_Fine_Arts_Vienna
[5]	https://en.wikipedia.org/wiki/Painting
[6]	https://en.wikipedia.org/wiki/Expressionism
[7]	https://en.wikipedia.org/wiki/Painters
[8]	https://en.wikipedia.org/wiki/Gustav_Klimt
[9]	https://en.wikipedia.org/wiki/Self_portrait

Figure 5.16: Excerpt of the in-article links of Egon Schiele

language versions of an article. Therefore, the algorithm can look up whether the English URL of Gustav Klimt is contained in the list of language versions of any of the links from the German Egon Schiele-article. Lastly, the occurrences are summed up and divided through the total number of in-article links of the article in the English version. After all articles are processed, the percentages are averaged.

The direction of comparison reveals an interesting difference. Sometimes, a Wikipedia article can be reached via two different URLs. An example for such an article is the one about the "Gemäldegalerie" in Berlin. In the English Wikipedia version, the article can be reached via

- <https://en.wikipedia.org/wiki/Gem%C3%A4ldegalerie> or
- https://en.wikipedia.org/wiki/Gem%C3%A4ldegalerie,_Berlin

Even though both URLs reference to the same article, in some articles the first and in some the second URL can be embedded in the text. In the English article about Albrecht Dürer¹⁵ even both URLs for the Gemäldegalerie are used. Another example for an article which is reachable via two URLs is the one about Rembrandt. In e.g. the German language version the article can be reached via

- https://de.wikipedia.org/wiki/Rembrandt_van_Rijn or
- <https://de.wikipedia.org/wiki/Rembrandt>.

An article is reachable via two URLs if a redirection is set up. Redirections are set up to forward a user (if he/she enters the URL containing just Rembrandt) to the webpage

¹⁵ https://en.wikipedia.org/wiki/Albrecht_D%C3%BCrer; [accessed 04-April-2016]

containing `Rembrandt_van_Rijn` in the URL. The article containing just `Rembrandt` in the URL does not have any content besides the redirection statement. Of course, there might also be duplicate articles with slightly changed titles on Wikipedia. Yet, the administrators and authors of Wikipedia constantly try to reduce the number of duplicates.

Set up redirections or duplicate articles therefore influence the overlap. Under the assumption that duplicates are negligible, differences in the overlap result from set up redirections. As for each article the share of equal links is finally divided by the number of links in the considered article, percentages vary depending on the perspective of the comparison. The percentages listed in Table 5.13 show how large the overlap between two language versions is. For example, the overlap of 27.4% listed in row two and column three means, that 27.4% of the English links also appear in the German articles.

	English	German	Italian
English	-	27.4%	35%
German	29.3%	-	36%
Italian	29%	27%	-

Table 5.13: Results of the overlap calculations

As mentioned above, for each language pairing two values are listed due to the direction of comparison. The overlap from the German to the Italian version is the highest one of all but also the difference between the two directions is the largest. In contrast to the German-Italian overlap, the German-English overlap is a little bit lower. Interestingly, the English-Italian overlap is higher than the other way round. This is remarkable because on average, English articles contain more links. As the share of common links is divided by the total number of links in the article one would expect to see a higher number on the side with shorter articles.

5.2.5 Size comparison of the different language version articles

One might expect, that, if an artist is e.g. from Austria, the article in the German language version is more detailed than in e.g. the Italian version. This KPI, the size comparison regarding in-article links, shall answer this question in some way. Like for the KPI "Overlap of links between language versions" articles are only considered for the calculation, if they are available in both (or all three) language versions which are compared.

Each language pairing (the German with the Italian, the German with the English, the Italian with the English and all three language versions together) will be analysed. The calculations are done separately for each pairing. In each of these four calculation runs, there exists a variable for each language version. If an article from the German and Italian version is compared and the German version has more in-article links, the corresponding variable is incremented. After all artists, who occur in the two/three language versions,

were analysed, the variable with a higher value marks the language version where more articles have a higher number of in-article links. It is important to take the citizenship of artists into account in the calculation. The hypothesis is, that articles are longer in the national version. So if e.g. the German and Italian version are compared but more Italian artists were crawled, the Italian version would dominate the comparison. The values of the comparison are listed in Table 5.14.

	English	German	Italian
English	-	66.4%	58%
German	33.6%	-	35.5%
Italian	42%	64.5%	-

Table 5.14: Size comparison between the different language versions

Interestingly, 64.5% of the articles about identified art-related persons, who have articles in the German and Italian Wikipedia, contain more links in the Italian version. Of those persons 24% had a "German" citizenship and 42% an "Italian" one. The citizenships of the Wikipedia versions were previously defined in the KPI "Average number of linked artists from the same nationality as the articles are crawled in". 76% of the articles about German artists were longer in the German Wikipedia version whereas 89% of the articles about Italian artists were longer in the Italian version. So, even if the results of this KPI at first sight convey the feeling that Italian articles are longer than the German ones, one has to consider, that more articles about Italian artists were analysed. Articles about national artists were mostly longer in the corresponding Wikipedia version.

The comparison of the German and English version did not return surprising results. Of the found articles which deal with art-related persons and are available in German and English, two thirds are longer in the English version and only one third is longer in the German Wikipedia. 19% of the considered artists had a citizenship like the German, and 28% one like the English Wikipedia. If the artist's citizenship is taken into account, the majority of articles contains more links in the national version. Only 10% of the articles about an English artist are longer in the German Wikipedia version. On the contrary, 21% of the articles about German artists are longer in the English Wikipedia version.

Unexpectedly, the difference between the English and Italian version is rather small. Whereas the difference in size between e.g. the German and English version is greater than 30% the difference between the English and Italian version is only 16%. Those artists who have a citizenship like the English Wikipedia usually have longer articles in the English version and vice versa. 10% of the articles about English artists were longer in the Italian version whereas 18% of the articles about Italian artists were longer in the English version.

Table 5.15 shows the results of a size comparison of all three language versions together. Of the articles crawled in all three language versions the English one contained the most links in 57.5% of the time. The Italian article contained most links in 28.5% of the time whereas the German one was the longest in only 14% of the time. 1.865 of the processed

	English	German	Italian
#1	57.5%	14.0%	28.5%
#2	28.5%	27.9%	43.6%
#3	14.0%	58.1%	27.8%

Table 5.15: Size comparison of the three language versions together

persons had a citizenship like the German Wikipedia, 3.460 one like the English and 3.443 one like the Italian Wikipedia. The rows with the headings #2 and #3 show, in how many cases the concerned language version had the second or third most links.

5.2.6 Graph metrics

It is also interesting to consider Wikipedia as a network and calculate different graph metrics like the average link distance between various artists. A short distance could indicate the existence of the small world phenomenon in the field of painters, sculptors, graphic designers and illustrators. Of course, again only articles about artists and not about e.g. artistic style should be considered. This is necessary as artists who worked on the same art movement are probably directly linked through the article of the movement itself. This would falsify the results.

The network of articles can be seen as a graph where the articles about persons are nodes and the links to other articles are edges. Edges do not have different weights. In the case of a search for the shortest path the goal is to find a path with a minimal length. As it is unknown whether the graph is an Eulerian one, the Fleury algorithm, cannot be used. Nevertheless, it would have been possible to use Dijkstra's algorithm.

There exist programs which are specialized on the processing of graphs. Therefore, the algorithms to calculate the shortest paths and other metrics are not self-developed. Concretely, Gephi will be used for the computations.

Data pre-processing and conversion to a graph

The open source software Gephi does not only create visualizations of graphs but can also compute different metrics for them. Prior the calculations the person database entries have to be converted to nodes and the links contained in the articles to edges connecting the nodes. Gephi accepts files in the graph modelling language (GML) format. Other graph editors like yEd or NetworkX work with GML files too. GML files contain a textual specification of graphs. An example for a simple GML file is shown in Figure 5.17.

```
graph [
  comment "This is a sample graph"
  directed 1
  IsPlanar 1
  node [
    id 1
    label
    "Node 1"
  ]
  node [
    id 2
    label
    "Node 2" ]
  node [
    id 3
    label
    "Node 3"
  ]
  edge [
    source 1
    target 2
    label "Edge from node 1 to node 2"
  ]
]
```

Figure 5.17: Excerpt from a sample GML file¹⁶

This GML file represents a planar, directed graph with three nodes and one edge. For each edge, one node is specified as the source and a second node as a target node. Nodes are addressed with their IDs. It is important to consider the direction of the links. A link in an article is not bi-directional but points only from one person to another one. An algorithm was developed to create the GML file from the data stock. All crawled persons (from the different language versions) are fetched from the database. In the next step, for each person a node section is written in the GML file and for each related person an edge is added. Figure 5.18 depicts an excerpt of the GML file for the German Wikipedia version.

¹⁶ <https://www.fim.uni-passau.de/fileadmin/files/lehrstuhl/brandenburg/projekte/gml/gml-technical-report.pdf>; [accessed 07-February-2016]

```

graph [
  comment "DE graph"
  directed 1
  node [
    id https://de.wikipedia.org/wiki/Colijn_de_Coter
    label "Colijn de Coter"
  ]
  edge [
    source https://de.wikipedia.org/wiki/Colijn_de_Coter
    target https://de.wikipedia.org/wiki/Ulrich_Thieme
  ]
  edge [
    source https://de.wikipedia.org/wiki/Colijn_de_Coter
    target https://de.wikipedia.org/wiki/Felix_Becker_(Kunsthistoriker)
  ]
  edge [
    source https://de.wikipedia.org/wiki/Colijn_de_Coter
    target https://de.wikipedia.org/wiki/Max_J._Friedl%C3%A4nder
  ]
]

```

Figure 5.18: Excerpt of the GML file for the German Wikipedia version

The created GML file can then be imported into Gephi. As mentioned above, it is important to specify the graph as a directed one during the import. After the GML file was imported, the graph is shown as a black square. Initially, there are too many nodes which are displayed too close together to distinguish them. This is exemplified by Figure 5.19 where the graph of the German Wikipedia version is shown right after the import in Gephi.



Figure 5.19: Graph for the German Wikipedia version, right after the import in Gephi

There are different layout mechanisms, for instance, "Force Atlas" or "Fruchterman Reingold" which can be applied to reposition the nodes and edges in the graph. Yet, even

if the nodes and edges are not rearranged, different metrics can be calculated with Gephi. The results of these calculations can be applied as properties or filter criteria to the nodes and edges contained in the graph (e.g. the node size or colour can vary depending on the value of the metric, nodes can be filtered depending on the value of a property).

Figure 5.20 shows the same graph as above, this time however with a degree filter and different layout mechanisms applied. Exemplary, the node size corresponds to the Eigenvector centrality of that node. The larger the size the higher the centrality. The colour of a node represents the degree – the lighter it is the lower the degree.

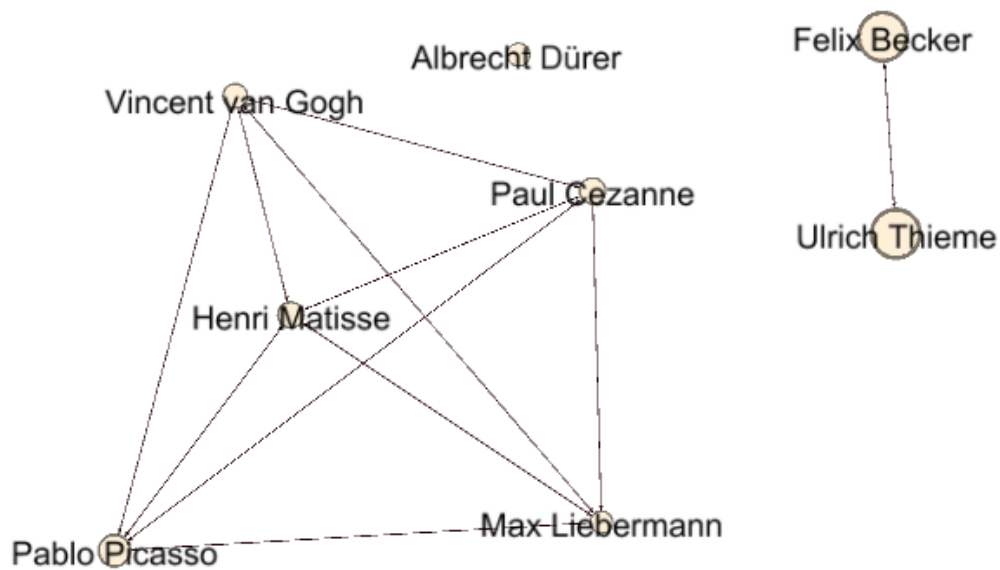


Figure 5.20: Graph for the German Wikipedia version after applying layout and filtering mechanisms

Even though not all articles were considered in the crawling process and for each language version a separate GML file was produced, the graphs still contain a lot of nodes and edges. To get a clear image of the graph, different parts can be filtered out so not all nodes and edges are shown. Different parameters like in the example above, the node degree, the shown k-cores or ego networks can be used as filtering criteria. Nevertheless, it is important to perform metric calculations before applying the filters because as the calculations only base on the filtered graph.

Gephi will not only be used to filter and rearrange the graph but also to calculate some metrics for the graphs of the different language versions. In this context the calculation of different centrality measures or degree distributions will provide interesting information for the comparison process.

The creation of the German and Italian GML files took nearly the same amount of time. Yet, due to the larger amount of articles and links it took much more time to create the

English GML file. As 56.572 art-related persons were found in the English Wikipedia version and each article on average contains around 120 links, nearly 6.8 million links had to be analysed.

In the Italian Wikipedia version fewer artists were found than in the German one. The Italian articles however contain more links on average. The graph of the German Wikipedia version consists of 24.110 nodes and 155.173 edges between them. The Italian graph contains 16.431 nodes which are linked by 181.314 edges. The English one comprises the most nodes and edges, in particular 56.572 nodes and 791.955 edges.

The calculated metrics of the different graphs are listed in Table 5.16.

	German		Italian		English	
Average degree	12.9		18.2		22.8	
Network diameter	20		16		19	
Average path length	5.7		4.8		5.4	
Number of shortest paths	338.067.645		173.814.822		1.790.204.573	
Connected components	Weakly	CC: 1.363;	Weakly	CC: 744;	Weakly	CC: 5.009;
	Strongly	CC: 8.217	Strongly	CC: 5.026	Strongly	CC: 20.613
Clustering coefficient	12.7%		17.6%		15%	

Table 5.16: Metrics of the three language graphs

The calculated average degree of the graphs supports the values of the KPI "Average number of links to other artists". Nodes from the English Wikipedia have, on average, the highest degree, followed by nodes from the Italian and German Wikipedia. The reason why the average degree is higher than the KPI is, that not only outgoing but also incoming links are considered in the calculation of this metric. The lower or higher linking can also be seen in the ratio between nodes and edges. For instance, in the German graph the ratio is around 1:6 whereas it is around 1:14 in the English one. The difference between the Italian and English graph is even smaller than between e.g. the German and the Italian one. The small difference coincides with the calculated average number of links to other artists, where the average Italian article has just a few links less than an English one.

The diameter is the length of the longest shortest path in the network [Sacharidis, 2015a]. A denser link structure positively influences the network diameter. Due to the fact that the average article in the German version has a lower degree than a node in the other graphs, the average shortest paths are also longer in this version. This can actually be seen in the average path length, which is approximately one hop longer than in the Italian version. Despite the fact that the English network is much larger regarding the number of nodes and edges, paths are shorter than in the German graph. It is remarkable that just around five hops are needed in the English graph to get from a random node to another one. Surprisingly, the average path length is even lower in the Italian version. The same is true for the network diameter, which is the shortest in the Italian graph. It

seems that in the Italian graph the network of art-related people is linked more densely than in the other versions.

The number of edges (links) influences, how many shortest paths there are in a graph. In total more than 1.7 billion shortest paths have been found by Gephi in the English graph. In the German graph, fewer shortest paths have been found. However, if the number of shortest paths of the German graph was multiplied by five (because the English graph contains five times as many edges) it would be nearly equal to the number of shortest paths found in the English graph. Even though the Italian graph has more edges than the German one, Gephi found fewer shortest paths than in the German and of course the English graph.

There are two types of connected components – strongly and weakly connected ones. A component is strongly connected if there exists a path from every node to every other one, taking the edge direction into account. If the direction of the edges is ignored and there exists a connection from every node to every other one, the component is weakly connected. [Sacharidis, 2015a] Despite the fact that Italian articles are higher linked on average, the number of weakly connected components is higher in the German version. This issues from the total higher number of nodes in the German version. The proportion between the German and English graph turns out as expected. Due to the fact that the English graph is much larger than the German one, the numbers of connected components are higher in the English graph. The difference between the Italian and English graph is quite large. Whereas the English one contains more than three times as many nodes and four times as many edges as the Italian one, the number of connected components is more than six (weakly connected components) and more than four (strongly connected components) times larger.

The clustering coefficient is an indicator of how many neighbours of a node are connected. The fact that the clustering coefficient for the Italian version is higher than for the German graph supports the previous insight that the network of art-related people in the Italian version is linked more densely. Still, the clustering coefficients for both language versions are rather low. This means that generally not many neighbours are connected to each other. Like the average shortest path length, the clustering coefficient can indicate the existence of a small world phenomenon. Still, as the average path lengths are low but the clustering coefficients too it is not quite clear, whether the graphs exhibit a small world phenomenon. The same is true for the English graph. Whilst the clustering coefficient is higher than in the German graph it is, surprisingly, lower than in the Italian one.

The following graphs are all filtered so only those nodes are shown which have the highest metric-values. The number of nodes displayed varies to improve readability.

In Figure 5.21 the nodes with the highest Betweenness centralities from the German Wikipedia version are shown. Only the 16 most important nodes are shown. A node has a high Betweenness centrality if many shortest paths lead over it. The node size indicates the value of the centrality – the larger the node the higher the centrality. Some of the nodes with high Betweenness centralities are famous international artists. Yet, also German artists like Max Liebermann or Albrecht Dürer have a high centrality. The graph contains many German persons. As these nodes are on many shortest paths it means that they are often linked in other articles. This proves the theory that authors of e.g. the German Wikipedia link more German persons in articles.

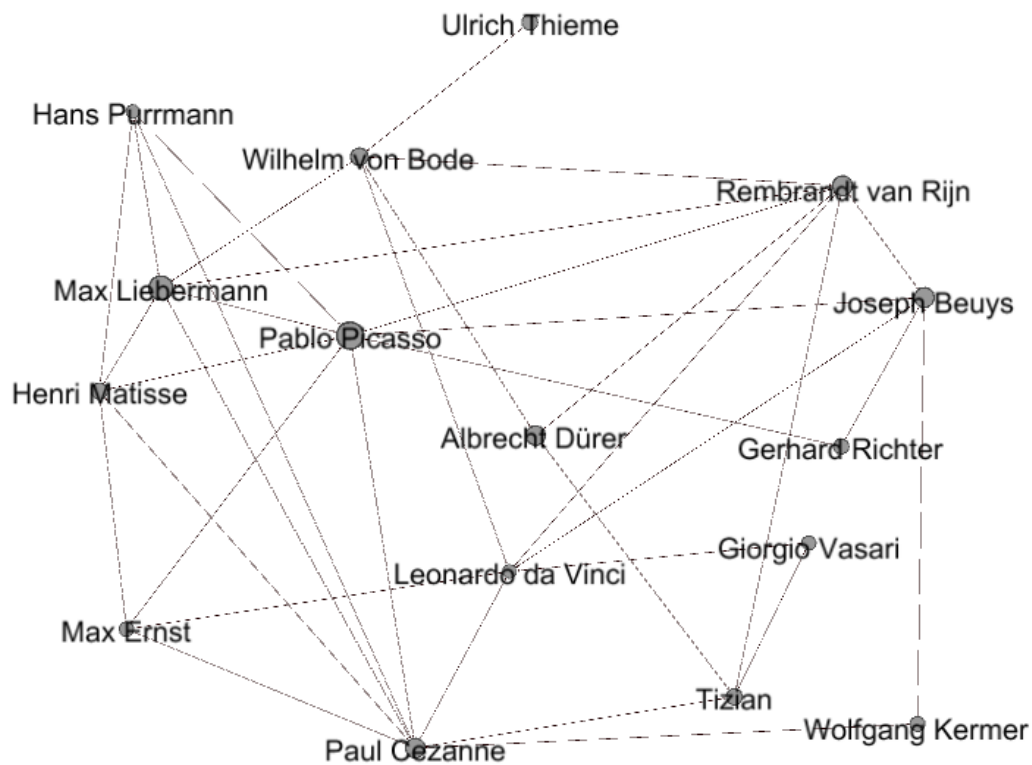


Figure 5.21: Articles with highest Betweenness centrality from the German Wikipedia version

Figure 5.22 exhibits the network of articles from the Italian Wikipedia with the highest Betweenness centralities. The size of the nodes points out, that the article about e.g. Leonardo da Vinci has a higher centrality value than the one about Andy Warhol. Other important articles are ones about famous artists like Renato Guttuso, Titian or Caravaggio. The majority of nodes with high Betweenness centralities are from articles about Italian persons. Interestingly, Canaletto and Volpedo are not directly connected to the other nodes shown in this graph.

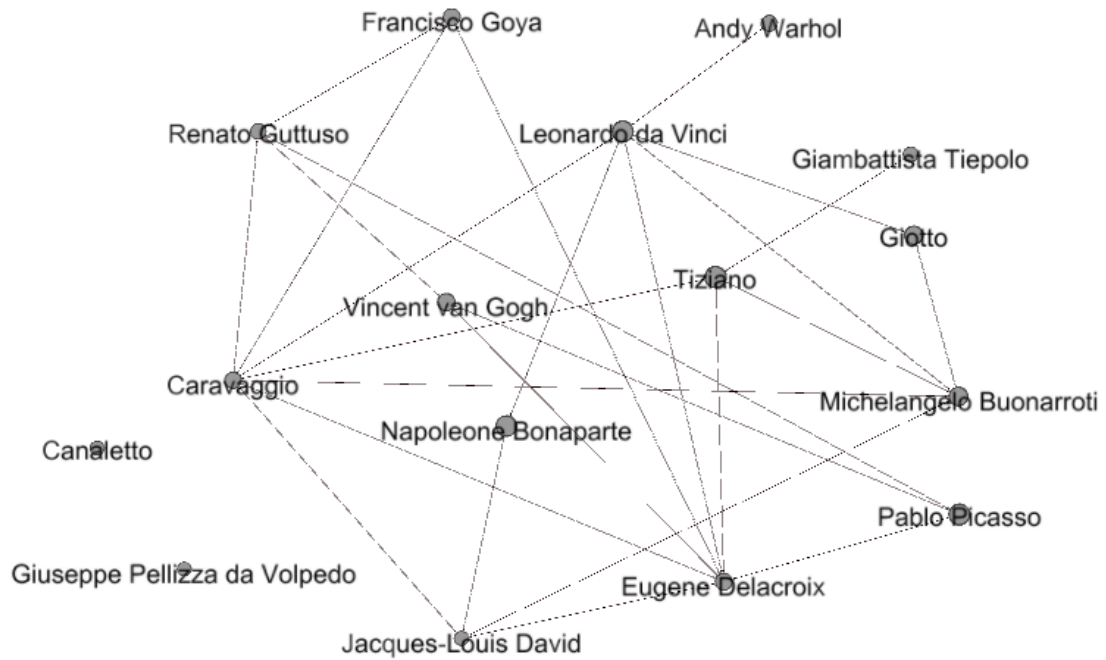


Figure 5.22: Articles with highest Betweenness centrality from the Italian Wikipedia version

The Betweenness centrality graph of the English Wikipedia shown in Figure 5.23 contains many famous international artists like Leonardo da Vinci or Vincent van Gogh. English artists contained in the graph are the British artists William Morris, William Blake, John Ruskin and the American ones Andy Warhol and Mary Cassatt. The node size indicates that Pablo Picasso has the highest Betweenness centrality. Of the three filtered Betweenness centrality graphs the English one is the one with the most foreign persons.

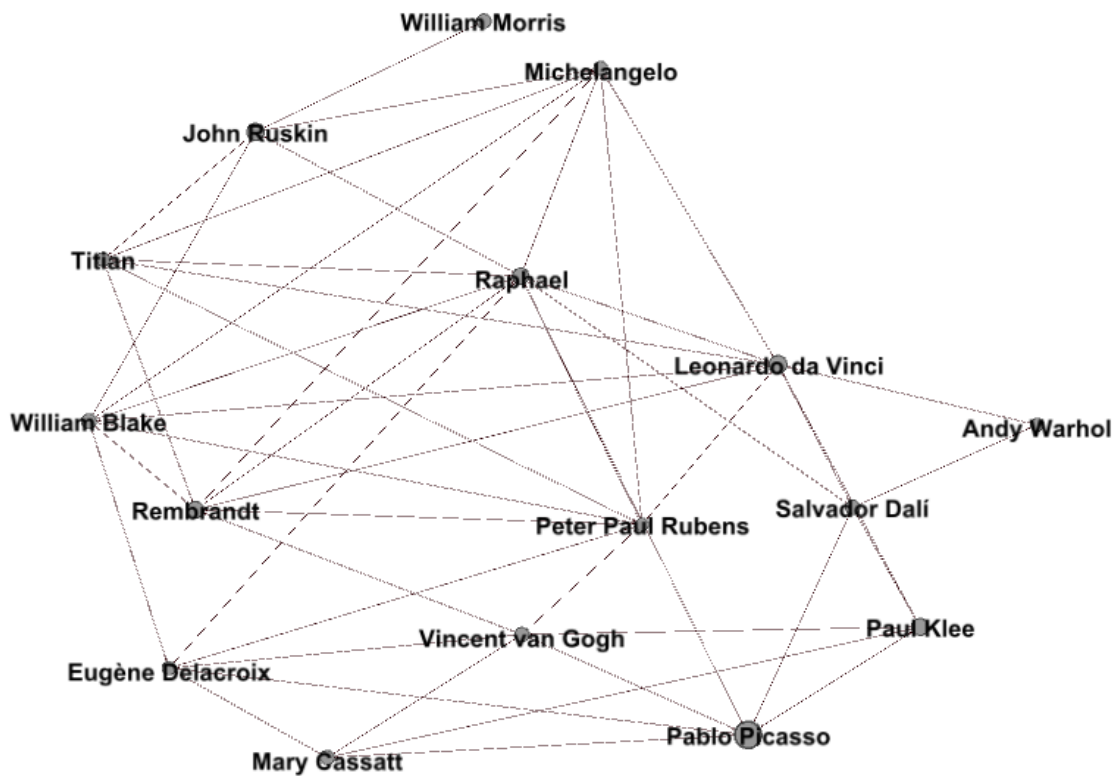


Figure 5.23: Articles with highest Betweenness centrality from the English Wikipedia version

In the following three figures the nodes with the highest PageRank centralities from the German, Italian and English Wikipedia version are shown. The node size again represents the importance of the node based on the calculated PageRank value. The PageRank and Betweenness centrality graphs have nodes in common. Still, many contained nodes are different. Whereas the Betweenness centrality bases on shortest paths, the PageRank focuses on the link structure. Max Liebermann who has a high Betweenness centrality in the German graph is not amongst the nodes with the highest PageRank centralities. In return, new artists like Raphael or the German art historian August Schmarsow show up in the graph. Another speciality is, that the nodes of Pablo Picasso, Theodore Blake Wirgman and Adolf Hitler are isolated from the other, most popular PageRank nodes. Such an isolation can occur if e.g. Picasso is linked in many other articles and therefore got a high PageRank but is not directly connected to any of the other important nodes shown in this graph.

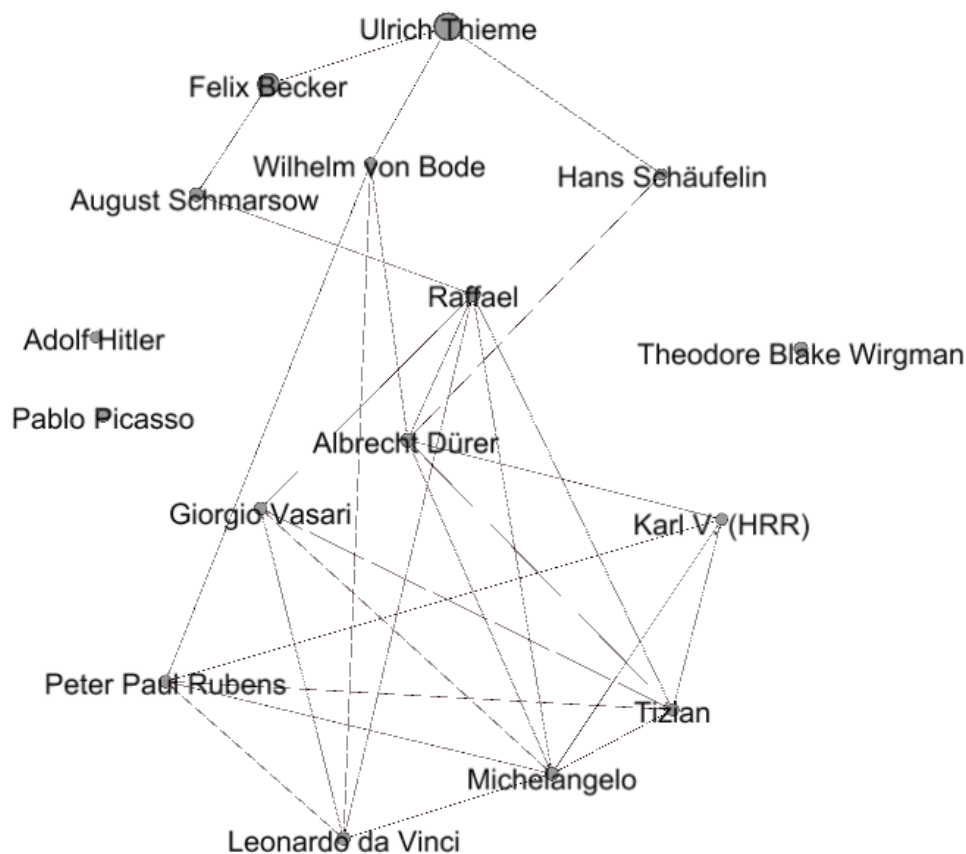


Figure 5.24: Articles with highest PageRank centralities from the German Wikipedia version

The situation is the same for the Italian PageRank graph. Titian or the article about Giotto show up in both graphs but e.g. Pierro della Francesca, Pope Sisto IV or Pope Sisto V are new in the PageRank graph. A peculiarity is, that the Italian PageRank graph contains more edges than the Betweenness centrality graph (57 versus 26 edges). Apart from that, it also contains more edges than the German one (57 versus 39 edges). Whereas in the German PageRank graph some nodes like the one of Ulrich Thieme are from larger size (due to a higher centrality), the picture in the Italian graph looks more homogenous. In general, there are more nodes from larger size in the Italian graph than in the German one. Both PageRank graphs have some nodes in common. However, it is obvious that in each language version national people are more important.

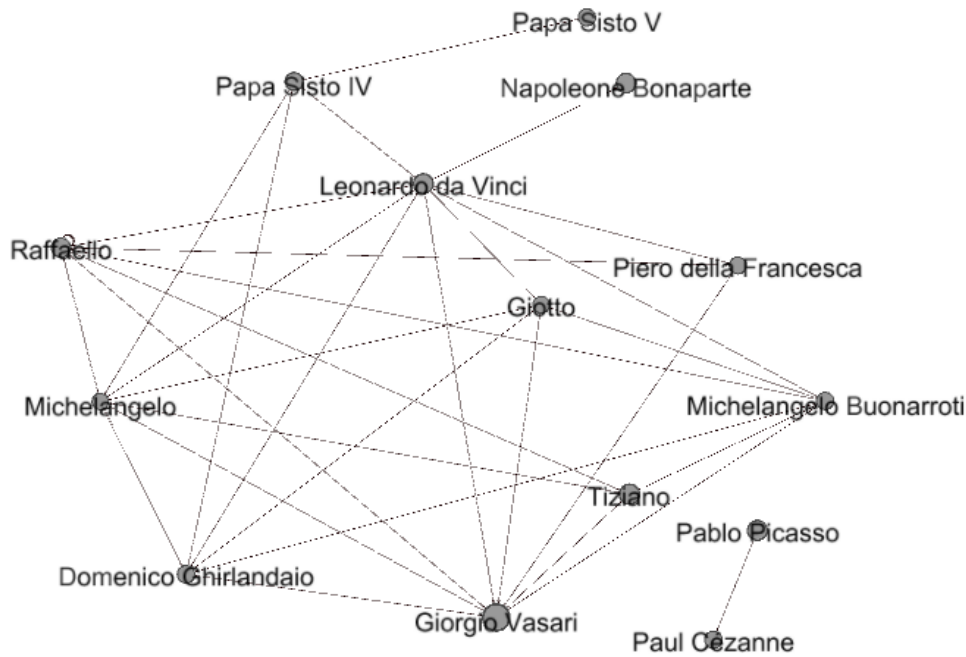


Figure 5.25: Articles with highest PageRank centralities from the Italian Wikipedia version

Figure 5.26 contains the nodes of the English graph with the highest PageRank centralities. A difference to the other two graphs is, that the graph again contains a lot of foreign persons. Whereas the German graph contains many German personalities the English one contains e.g. some Italian and Dutch persons. This anomaly could already be observed in the previous English graph. Apart from the nodes of the Dutch artist Arnold Houbraken and art historian Michael Bryan the node sizes are similar to each other.

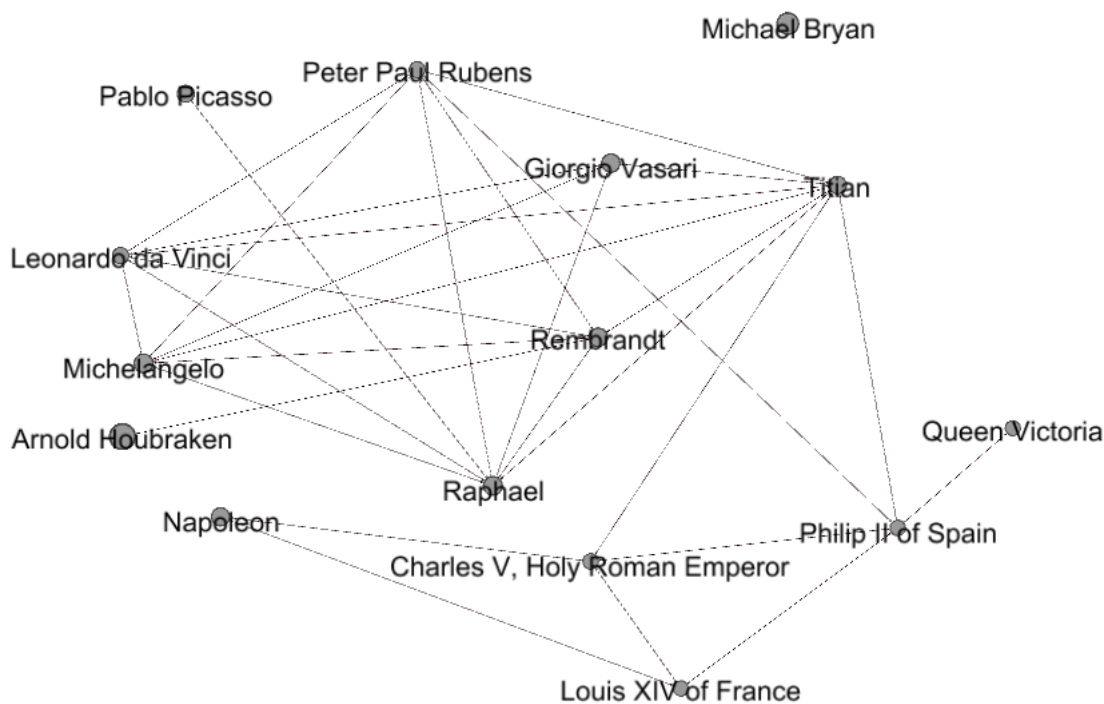


Figure 5.26: Articles with highest PageRank centralities from the English Wikipedia version

Another centrality measure is the Eigenvector Centrality. The Eigenvector centrality of a node is computed by taking the centralities of its neighbours into account.

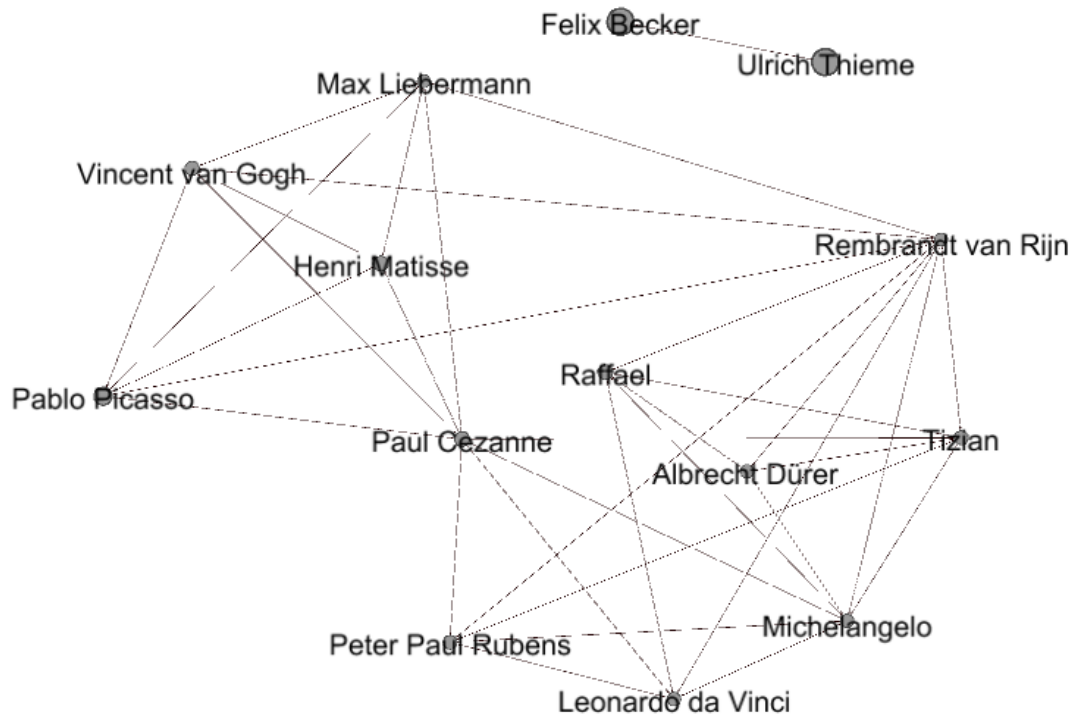


Figure 5.27: Articles with highest Eigenvector centralities from the German Wikipedia version

Figure 5.27 presents the 14 nodes from the German graph which have the highest Eigenvector centralities. There are again some nodes like the one of Paul Cézanne which do not show up in the Betweenness or PageRank graphs but are contained in the Eigenvector graph. The distribution of the Eigenvector centralities is somehow similar to the other two graphs. The picture is quite homogenous – the node size shows that many nodes have an equal centrality value. There are two nodes, the ones of Ulrich Thieme and Felix Becker, which have higher centralities. In the PageRank graph Ulrich Thieme and Felix Becker have a high centrality too, in the Betweenness graph Picasso and Max Liebermann are the ones with the highest centralities. The three graphs illustrate very well that the methods of calculation for centrality values return quite different results.

A part of the Italian Eigenvector-graph is shown in Figure 5.28. Only those nodes, which have the highest Eigenvector centralities, are shown. Interestingly, all nodes are from articles about popes. Another remarkable property of the graph is the very high number of links among the nodes. This graph contains only 17 nodes but 272 edges between them. Every node is directly linked to all other ones. That is why they all have a high Eigenvector centrality. This circumstance is also visible in the size of the nodes as they are equal.

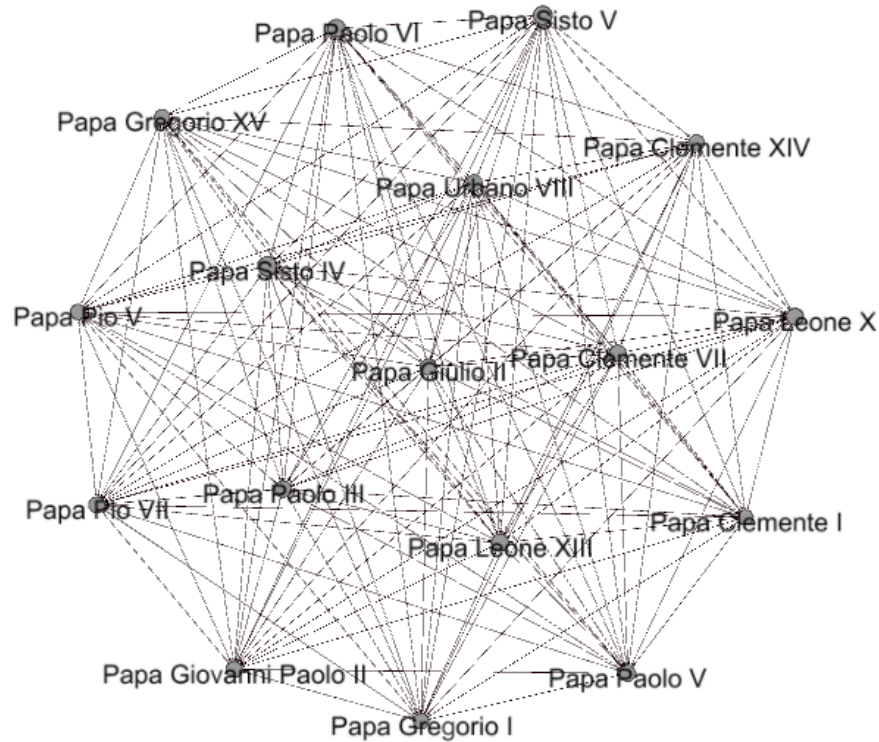


Figure 5.28: Articles with highest Eigenvector centralities from the Italian Wikipedia version

The Eigenvector centrality graph of the English Wikipedia version, as depicted in Figure 5.29, looks a little bit like the Italian one. First of all, the graph contains a lot of nodes. Even if just the upper 0.02% of nodes with the highest centralities is shown, the graph still contains 20 nodes and 380 edges. They all have a centrality value of more than 0.99 where 1 is the maximum. Secondly, the graph exhibits a similar phenomenon like the Italian one. It exclusively contains Indian people whereas the Italian graph only contains popes. Some examples of contained people are Yusuf Arakkal, Jitish Kallat or Bose Krishnamachari.

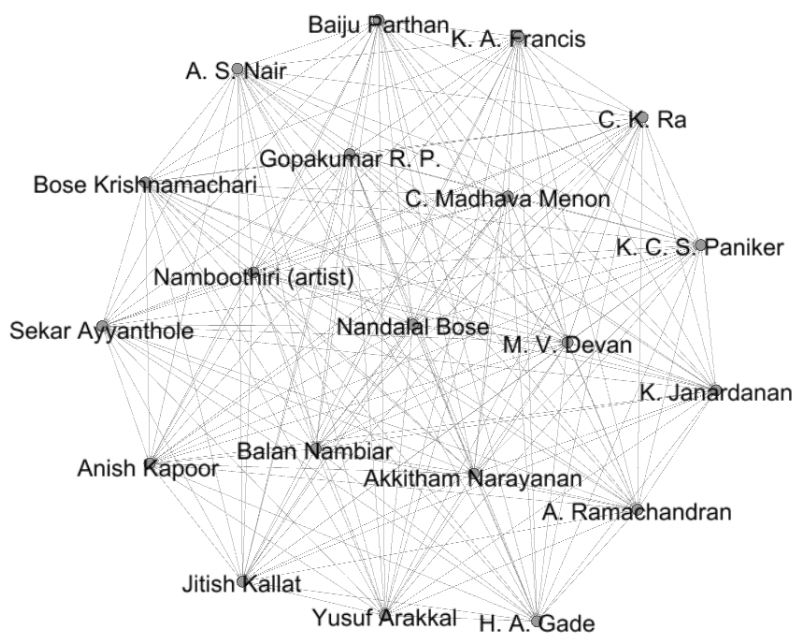


Figure 5.29: Articles with highest Eigenvector centralities from the English Wikipedia version

As shown in Figure 5.30 many partners are linked in the article about Bose Krishnamachari. In the article of each of these artists there exists a partner-box too. Because of the Eigenvector centrality calculation algorithm the growth of centralities of those nodes is reinforced as every node gets the sum of centrality values of all neighbours.

V · T · E	Painters from Kerala	[hide]
Akkitham Narayanan · A. Ramachandran · A. S. Nair · Baiju Parthan · Bose Krishnamachari · C. K. Ra · C. Madhava Menon · C. N. Karunakaran · Gopakumar R. P. · Mini Sivakumar · Jitish Kallat · K. A. Francis · Kavitha Balakrishnan · K. C. S. Paniker · K. G. Subramanyan · Artist Janardanan · M. V. Devan · Artist Nambuthiri · N. N. Rimzon · Paris Viswanathan · Raja Ravi Varma · Riyaz Komu · Sekar Ayyanthole · S. Jithesh · Sosa Joseph · T. K. Padmini · T. V. Santhosh · V. S. Valiathan · Yusuf Arakkal · Balan Nambiar		

Figure 5.30: Partners of Bose Krishnamachari¹⁷

¹⁷ https://en.wikipedia.org/wiki/Bose_Krishnamachari; [accessed 13-May-2016]

The reduction of the graph to nodes with the highest in-degree provides some interesting information too. This perspective shows to which nodes most links point to. The filtered German graph is presented in Figure 5.31. Rembrandt van Rijn has the least in-degree with a value of 238. Some nodes like the ones of Ulrich Thieme, Titian, or Albrecht Dürer already appeared in previous graphs. The fact that some nodes appeared in nearly all graphs show their high importance in the German Wikipedia in the field of painters, sculptors, illustrators and graphic designers. Yet, there is a new node in this graph too. Wassily Kandinsky did not appear previously.

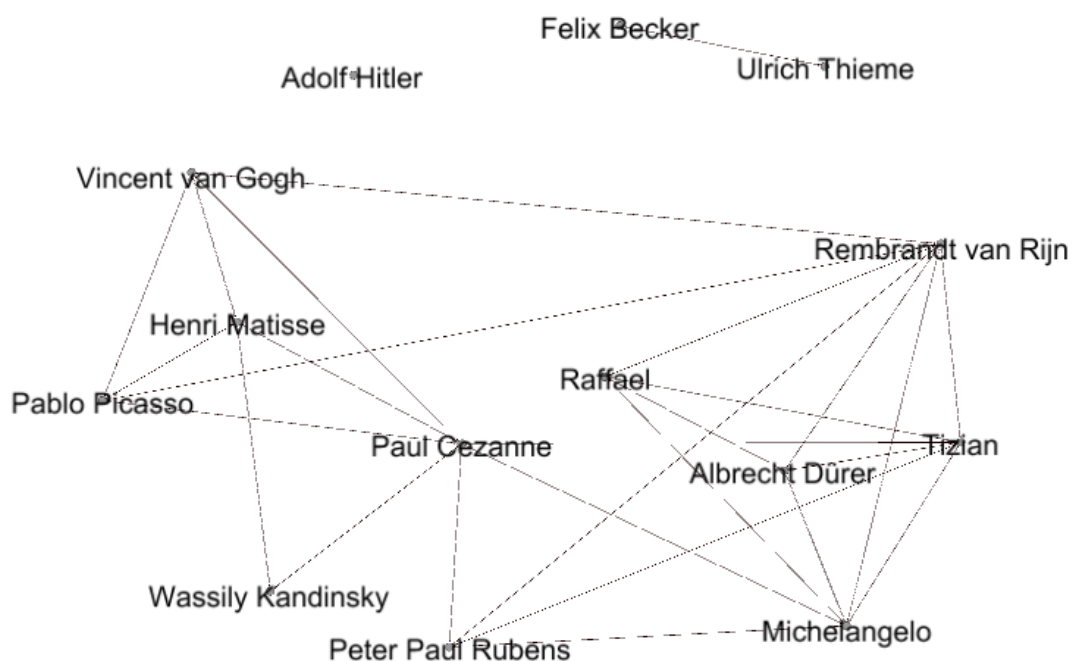


Figure 5.31: Articles with highest in-degree from the German Wikipedia version

The nodes with the highest in-degree in the Italian graph are shown in Figure 5.32. To be among the top nodes it requires at least 247 in-links like for the article about Papa Urbano VIII. This is more than the minimal value in the German graph. The high importance of the church in Italy can be seen through the appearance of many popes in the Italian graphs. Two obvious reasons why popes are so important in the Italian Wikipedia are, that the church (and therefore also popes) were important sponsors for artists. Apart from that, the church played and still plays an important role in the life of many Italians. Seemingly this influences authors, so many links to popes are included in articles.

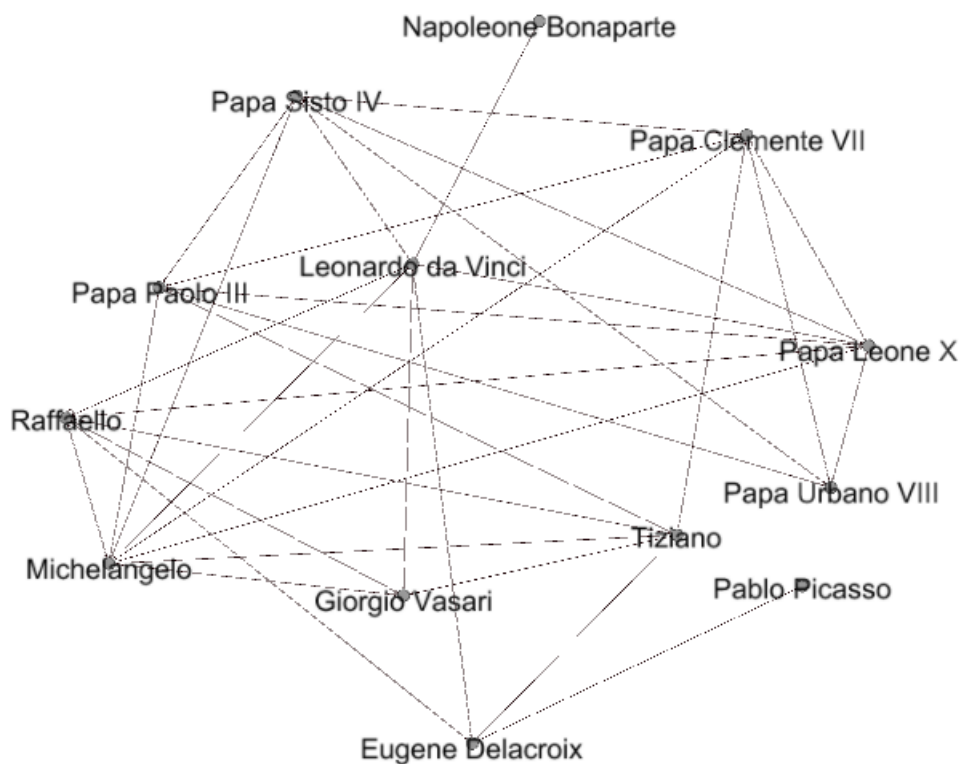


Figure 5.32: Articles with highest in-degree from the Italian Wikipedia version

In Figure 5.33 the nodes with the highest in-degree from the English graph are presented. The graph does not only contain artists but also important art-related people like Michael Bryan or Louis XIV of France. The minimal in-degree of these nodes is 588 which means that a lot of other articles contain links which point to the article about e.g. Jacques-Louis David in this case. This is far more than the minimum number of in-links in the German and Italian graph.

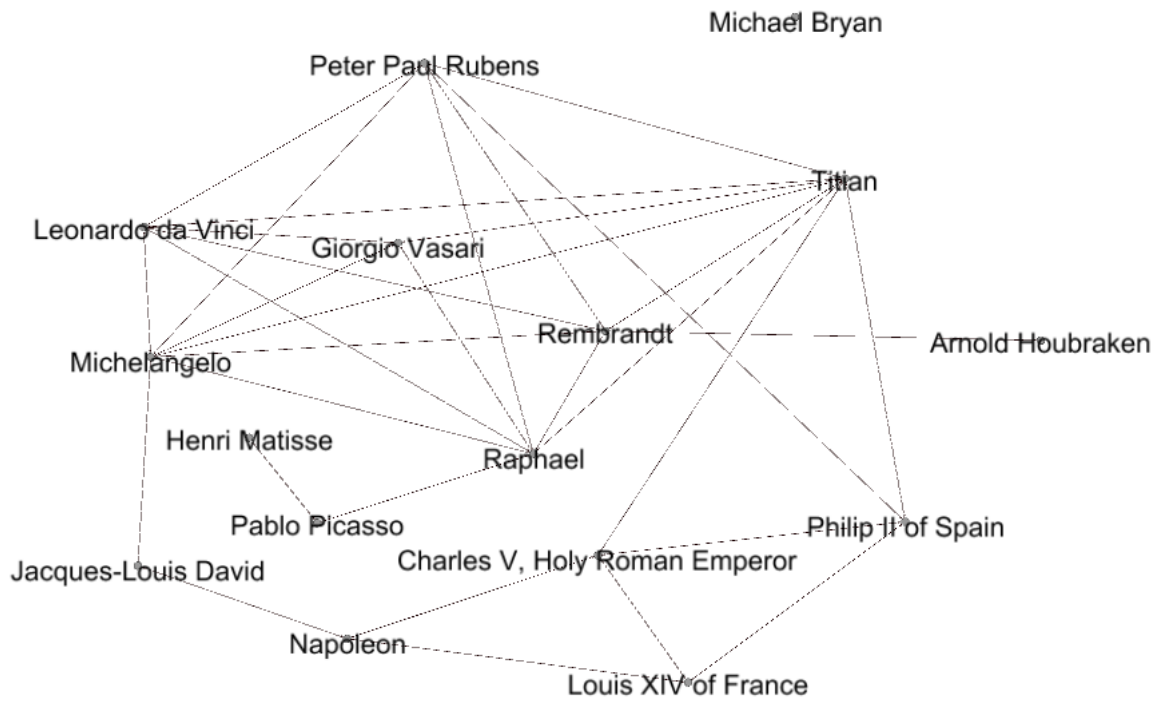


Figure 5.33: Articles with highest in-degree from the English Wikipedia version

Comparison of three national artists

To illustrate the differences between the language versions even better the KPIs are calculated for a German, Italian and English artist who have articles in all three language versions. The selection criteria for artists were, that they have related people listed in ULAN and Wikidata. Apart from that, the selection was performed randomly. The German painter Otto Modersohn is famous for his landscape paintings ¹. In ULAN respectively Wikidata two related persons are listed. Caravaggio, who was born and lived in Italy, focused on large-scale, realistic religious paintings ². In ULAN, six relationships were listed whereas one teacher was found in Wikidata. For the English Wikipedia version Thomas Eakins, an American painter, was chosen. He was one of the most important American Realism painters ³. In Wikidata for Thomas Eakins only one student and teacher are listed but in ULAN 23 relationships are contained. The overlap is calculated from the language of origin of the artist to the other versions. The size comparison is not listed separately in the tables as it directly corresponds to the number of links per article.

The evaluation of the KPIs for Otto Modersohn, which are listed in Table 6.1, once again emphasizes, that local artists have longer and more detailed articles in the corresponding Wikipedia version.

Especially the number and citizenship of linked artists strongly depends on the language version. In the German article nearly 29% of the links point to other artists whereas the shares for the other two languages are 23%. Furthermore, a lot of artists with the same citizenship as the Wikipedia version are linked in the German article. The level of detail is reflected in the completeness regarding ULAN and Wikidata too.

¹ <http://www.modersohn-museum.de/zum%20werk.html>; [accessed 10-May-2016]

² <http://www.britannica.com/biography/Caravaggio>; [accessed 10-May-2016]

³ <http://www.britannica.com/biography/Thomas-Eakins>; [accessed 20-May-2016]

6. COMPARISON OF THREE NATIONAL ARTISTS

Otto Modersohn	English	German	Italian
Number of links in article	35	87	17
Number of links to other artists	8	25	4
Number of links to artists with the same citizenship as the current artist	5	16	1
Number of linked artists from the same nationality as the articles are crawled in	0	16	0
Completeness regarding ULAN	50%	50%	50%
Completeness regarding Wikidata	0%	100%	0%
Overlap	24%	-	9%

Table 6.1: KPIs for the articles about Otto Modersohn

The KPI "Overlap" indicates, that 9% of the German links were found in the Italian article and 24% in the English one. At first, this might sound very low. Due to the fact that the number of equal links is divided by the total number of contained links in the focus-language (in this case German), the shares are rather low. Nevertheless, if the overlap is calculated from Italian to German and English to German, the shares are 47% and 51%.

Caravaggio	English	German	Italian
Number of links in article	274	286	390
Number of links to other artists	112	54	96
Number of links to artists with the same citizenship as the current artist	37	24	47
Number of linked artists from the same nationality as the articles are crawled in	1	1	48
Completeness regarding ULAN	50%	33.3%	50%
Completeness regarding Wikidata	100%	100%	100%
Overlap	41%	23%	-

Table 6.2: KPIs for the articles about Caravaggio

The calculated KPIs for the articles about the Italian artist Caravaggio are listed in Table 6.2. The high number of in-article links in all language versions shows, that Caravaggio is very famous in different countries. For all three language versions the number of links to other artists is very high. Not only with respect to the number of in-article links but also in total the English article contains most relationships to other

artists. Even though this disagrees with the set up hypothesis this special situation might issue from Caravaggio's extraordinary popularity. Still, the Italian article contains more links to artists with the same citizenship as Caravaggio and as the Italian Wikipedia version. Apart from the German article the other two contain half of all listed relationships from ULAN. Regarding the completeness of Wikidata relationships all language versions are on par. Compared to the articles about Otto Modersohn the overlap-shares are very high, especially the one to the English article.

Thomas Eakins	English	German	Italian
Number of links in article	146	79	42
Number of links to other artists	32	20	2
Number of links to artists with the same citizenship as the current artist	17	5	0
Number of linked artists from the same nationality as the articles are crawled in	6	0	0
Completeness regarding ULAN	30%	13%	4%
Completeness regarding Wikidata	100%	50%	50%
Overlap	-	15%	9%

Table 6.3: KPIs for the articles about Thomas Eakins

Lastly, the articles about Thomas Eakins were analysed with the help of the developed KPIs. The results are listed side by side in Table 6.3. The results fully approve the set up hypotheses. The English article outperforms the other two ones in every KPI. Not only does the English article contain more links but also more links to related artists. The English article contains far more relationships listed in ULAN and Wikidata than the German and Italian article. The overlap is smaller (or equal to the German-Italian one for Otto Modersohn) than the ones for the other artists.

Possible reasons for the differences between language versions

There may be several reasons for differences or similarities between the different Wikipedia language versions. Mostly they are hard to prove and on these grounds only assumptions. The first possible explanation for similarities between the German and Italian Wikipedia version but differences to the English version is the geographical closeness between Italy, Austria, Switzerland and Germany but the greater distance to Great Britain. Besides the distance, Great Britain is an island and separated from the European mainland through the Channel.

Especially in former times travelling was not always as affordable, comfortable and secure as it is today. Therefore, journeys were often restricted to the mainland. Besides that, the distance people could travel in former times was also restricted through the availability of the means of transportation. Distance-restrictions would be indicative, that certain art movements rather spread over the mainland and closer surroundings like neighbouring countries than over countries which are far away from each other or separated through the sea. Nevertheless, most artistic styles somehow spread over nearly every country over time. The similarities and differences were pointed out in the bar plot in chapter 5.2.

Another reason for similarities between the German and Italian Wikipedia version could be, that in former times empires covered large parts of Europe.

The Carolingian empire (800-924) for example reached over several countries. It covered large parts of France, Germany, Austria and Italy. The empire which then emerged, the Holy Roman Empire (962-1806), covered at around 1181 a large part of Italy, the majority of Austria, Corsica, Germany, Switzerland, parts of France, nearly the whole Czech Republic and parts of the Netherlands, Belgium, Luxemburg and Slovenia. [Albig et al., 2014] Yet, as many pieces of art from that time were destroyed or lost, it is

nearly impossible to prove this hypothesis.

Another explanation why some artists from foreign countries are higher interlinked and more detailed than others are exhibitions in that country. Unfortunately, like the hypothesis above, also this one is nearly impossible to observe. To show the influence of an exhibition, a city had to be selected where a certain exhibition will take place. In the next step, it had to be evaluated, if occurring artists already have Wikipedia articles and how extensive these articles are. If not, it had to be observed if a Wikipedia article is written about an artist from the exhibition. This could also be done through backwards comparison as every Wikipedia article has a version list. The major problem however is to determine, where the authors are from and if they really visited the exhibition. This may be done by performing interviews with Wikipedia authors.

Differences between the German, English and Italian Wikipedia version might also originate from education. In every country there are probably more exhibitions about national artists. So it is easier for e.g. art teachers to show students artworks of national artists rather than international ones. Of course, there are always exhibitions about famous international artists. Yet, there are just a few exhibitions about less famous artists from other countries. In the following, two well-known museums for every language version are selected. For each of them, the exhibitions of 2015 which focus on paintings are analysed. Other exhibitions which focus on e.g. archaeology or where the citizenship of the artists is not determinable are not mentioned in this list. The results are shown in Table 7.1.

Museum	Location of the museum	Name of the exhibition	Country of citizenship of artist(-s)
Albertina ¹	Austria	Arnulf Rainer	Austria
		Miro	Spain
		Karl Prantl	Austria
		Birgit Graschopf	Austria
		Degas, Cézanne, Seurat	France
		Warhol bis Richter	Multiple
		Sturtevant	USA
		Von der Schönheit der Natur	Austria
		Lee Miller	USA
		Bacon, Warhol, Richter	Multiple
		Drawing Now: 2015	Multiple
		Abstraktion in Österreich	Austria
		Spurensuche	Multiple
Hamburger Kunsthalle ²	Germany	Nolde in Hamburg	Germany
		Franz Ludwig Catel	Germany
		When there is hope	Multiple
		Verzauberte Zeit	France, Netherlands

			Feministische Avant- Multiple garde der 1970er Jahre
Uffizi Gallery ³	Italy		Gherardo delle Notti: Netherlands (the artist Bizzarre Paintings and was heavily influenced Merry Suppers by Michelangelo during his visit in Rome ³)
			Piero di Cosimo 1462- Italy 1522 – Eccentric Painter between the Renais- sance and Mannerism
Musei Capitolini ⁴	Italy		Leonardo da Vinci Italy
			Raffaello, Parmiginino, Italy and Barocci
Victoria and Albert Museum ⁵	England		Horst: Photographer of USA Style
			Constable: The Making England of a Master
			Russian Avant-garde Russia Theater: War Rev- olution and Design
			1913-1933
			Captain Linnaeus Tripe: England Photographer of India and Burma, 1852-1860
			Julia Margaret Cameron England
J. Paul Getty Museum ⁶	USA		Spectacular Rubens: Belgium The Triumph of the Eucharist
			Drawing in the Age of Germany / Netherlands Rubens / Beglium
			Josef Koudelka: Nation- Czech Republic ality Doubtful
			In Focus: Play England and USA
			Zeitgeist: Art in the Germany / Austria Germanic World, 1800- 1900
			J. M. W. Turner: Paint- England ing set free
			Renaissance Splendors Italy of the Northern Italian Courts

Light, Paper, Process:	Germany, mostly USA
Reinventing Photography	
A Kingdom of Images:	Mostly France
French Prints in the Age of Louis XIV, 1660–1715	
Andrea del Sarto: The Renaissance Workshop in Action	Italy
Degas: "Russian Dancers" and the Art of Pastel	France
In Focus: Animalia	Mostly USA, England, South Africa and others

Table 7.1: List of museums, their exhibitions and citizenship of artists

Table 7.1 exhibits, that museums mostly show national artists or ones from countries close or related (through e.g. the language like Great Britain and America) to the exhibiting one. Of course, sometimes also international artists are exhibited. The results however sustain the theory that showing national artists to students is easier than international ones. As a consequence, this can also be a reason why articles about national artists are more detailed in the corresponding language version.

¹ <http://www.albertina.at/jart/prj3/albertina/main.jart?rel=de&reserve%2Dmode=active&content%2Did=1202307119337&av=2015%2D01%2D01&ab=2015%2D12%2D31>; [accessed 17-April-2016]

² <http://www.hamburger-kunsthalle.de/index.php/archiv.html>; [accessed 17-April-2016]

³ <http://www.uffizi.org/museum/exhibitions/past-exhibitions-at-the-uffizi/>; [accessed 17-April-2016]

⁴ [http://en.museicapitolini.org/mostre_ed_eventi/mostre/\(p\)/in_archivio/\(y\)/2015](http://en.museicapitolini.org/mostre_ed_eventi/mostre/(p)/in_archivio/(y)/2015); [accessed 17-April-2016]

⁵ <http://www.vam.ac.uk/page/p/past-exhibitions-and-displays/>; [accessed 17-April-2016]

⁶ <http://www.getty.edu/visit/exhibitions/past.html>; [accessed 17-April-2016]

Conclusion and critical reflection

The Internet is an important knowledge base. There exists an answer for nearly every question or problem. However, sometimes it is difficult to evaluate if the answers are correct. This question was addressed in the context of this thesis. Of all available websites on the Internet, Wikipedia is a very famous, important and freely accessible source of information. Therefore, Wikipedia was selected as subject for this analysis. The goal of this thesis was to examine articles contained on Wikipedia on their extent and content. Due to the sheer amount of articles not each of them could be analysed, thus the main focus was on articles about selected areas of arts. In concrete terms relationships between art-related people were studied. Within the analysis process, three language versions, German, English and Italian, were compared. Articles about art-related persons from these language versions were fetched from Wikipedia. Additional sources of information and sources of reference were the Wikimedia-project "Wikidata" and the database "ULAN". ULAN is a source tailored for professional users but it is also available for the general public. For comparison of relationship information ULAN was the main source of reference.

Based on the results of the analysis, the English Wikipedia version contained most articles about relevant persons. Furthermore, articles from this version were the most accurate ones regarding contained relationships. Nevertheless it has to be considered, that typically articles about artists are more detailed in the "national" Wikipedia version. Even though ULAN pages were rarely linked in German Wikipedia articles more relationships coincided with the Wikipedia article than in the Italian version. Apart from ULAN also more relationships listed in Wikidata were contained in German articles than in their Italian counterparts.

The developed KPIs and visualisations helped to clarify differences between the three language versions. The KPI "Overlap" illustrated very well, that on average two language versions have maximal 36% of the links in common. As listed in Table 5.15 English articles are the longest in nearly 60% of the time. If smaller language versions were compared to

e.g. the English one, the result could be even more extreme. There are also differences in the completeness regarding ULAN and Wikidata. In general, the English version is the most accurate one. Another insight was, that more relationships listed in Wikidata are covered in Wikipedia articles than relationships listed in ULAN. Regarding the question whether national artists are preferred in the corresponding Wikipedia versions, the heatmaps illustrate very well, that this is true. This implies that the coverage of foreign artists is not as high as the coverage of national ones. Especially the heat maps depicted very well that content differs, which is also shown in other comparisons in chapter 5.

Another fact which is highlighted in this thesis is, that no source of information is perfect. Every source should be compared with another one. Even though articles on e.g. Wikipedia often contain a lot of information – there is still some room for improvement. Nevertheless, it is amazing how much content is already available on Wikipedia. Still, an issue which has to be considered as a reader of Wikipedia is the gender bias amongst authors. The mostly male authors may have a different view towards certain topics or might not even write articles about something. A possible cause might be that male authors simply have a different perception than female ones.

When starting a research process on Wikipedia, users should, if possible, read articles in different language versions. This might convey a more detailed picture of the topic already at the beginning of the research process. Definitely, the articles are a good starting point for further research. Contained information might not be complete in a single language version but rather in all language versions together. If the topic can be assigned to a certain nationality (like an invention which took place in that country or if a person originates from that country) the article probably contains most information in the national Wikipedia version. Even if Wikipedia articles are not always correct or complete, they are a very important source of information.

There are some points which have to be considered after reading this thesis. First of all, the situation might look completely different for other artists like musicians, actors, architects or in completely other areas. Articles might contain more relationships or language differences might be smaller.

Another issue is the classification algorithm. Some artists might have been missed or persons might have been wrongly classified as artists or art relevant. Surely, there are ways to improve the classification algorithm. For example, the complete article text itself and not only contained links could be analysed to identify relationships. Definitely, this would have an impact on the results of the KPIs. Yet, text analysis mechanisms would have exceeded the scope of this thesis.

The most important point however is, that the goal of this thesis is not to diminish the value of Wikipedia. Wikipedia is a very important and helpful website. Knowledge is available for everybody. The thesis shall only highlight differences and draw reader's attention to compare different sources of information during research processes. As mentioned above, another idea might be that the goal is not to have a single, complete language version but to be complete in total.

The insights of this thesis can be incorporated in future works too. Wikipedia articles

could e.g. be improved by automatically adding missing relationships. For missing relationships standard sentences containing the role and name of a person could be inserted in articles. As soon as authors see links which were created automatically (and which might not reference to an existing article) they could write an article about the referenced person or correct the link so it references to the correct name of the article. Besides Wikipedia also Wikidata could be improved. There, information could be added even easier as it is listed in a property / value manner and not in prose text. Another idea would be to automatically add notes in less complete articles which reference to a language version with more information. Users may then be attentive that an article does not convey a complete picture about a topic. These notes could also be created not for readers but for authors. They could get remarks in an article in their national language version which points them to a more complete article in another language version. The goal would be to encourage authors to extend the shorter article. Based on Sir Francis Bacon's famous quote "Knowledge is power" all people should have easy access to information. Still, information should be complete and correct. Therefore, people should always consult multiple sources to find correct answers before making important decisions.

List of Figures

3.1	Growth of the number of articles on Wikipedia	16
3.2	Excerpt from the Wikidata page about Titian	21
3.3	Snippet from the sources and contributors of the ULAN page about Frieden- reich Hundertwasser	22
3.4	High level Google architecture (Source: [Brin and Page, 1998])	29
4.1	Process diagram	32
4.2	Database record "Albrecht Dürer"	34
4.3	Record for the URL https://en.wikipedia.org/wiki/Raphael in the temporary URL database	35
4.4	Category section from the Wikipedia article about "Raphael"	38
4.5	Authority control section from the Wikipedia article about "Raphael"	39
4.6	Snippet of the related people section from ULAN for Raphael	42
4.7	Extract from the related people section from ULAN for Barthélemy Menn . .	43
4.8	Database entry for the article Raphael from the English Wikipedia version .	44
4.9	Database entry for the article Raphael showing the ULAN section	44
4.10	Extract of a database query via the command line interface	46
4.11	Analysis for a query without a limit	47
4.12	Analysis for a query with a limit	48
4.13	Analysis for a query with an index on the field	48
4.14	Related people from ULAN for Raphael	50
4.15	Snippet of the list "relatedPeople" of Raphael	50
4.16	Plot of the number of collected links versus the number of processed articles in the German Wikipedia version	52
5.1	Distribution of the number of links per article in the English Wikipedia version	58
5.2	Suggestion to improve an article	59
5.3	Distribution of the number of links per article in the German Wikipedia version	60
5.4	Distribution of the number of links per article in the Italian Wikipedia version	60
5.5	Plot for the average number of links to other artists for the English Wikipedia	62
5.6	Plot for the average number of links to other artists for the German Wikipedia	63
5.7	Plot for the average number of links to other artists for the Italian Wikipedia	63
5.8	Distribution of birthplaces of artists from the English Wikipedia version . . .	66

5.9	Distribution of death places of artists from the English Wikipedia version . .	67
5.10	Distribution of birthplaces of artists from the German Wikipedia version . .	68
5.11	Distribution of death places of artists from the German Wikipedia version . .	69
5.12	Distribution of birthplaces of artists from the Italian Wikipedia version . . .	70
5.13	Distribution of death places of artists from the Italian Wikipedia version . . .	71
5.14	Artistic movements occurring in the three language versions	73
5.15	Database record for the English article about Egon Schiele	77
5.16	Excerpt of the in-article links of Egon Schiele	78
5.17	Excerpt from a sample GML file	82
5.18	Excerpt of the GML file for the German Wikipedia version	83
5.19	Graph for the German Wikipedia version, right after the import in Gephi . .	83
5.20	Graph for the German Wikipedia version after applying layout and filtering mechanisms	84
5.21	Articles with highest Betweenness centrality from the German Wikipedia version	87
5.22	Articles with highest Betweenness centrality from the Italian Wikipedia version	88
5.23	Articles with highest Betweenness centrality from the English Wikipedia version	89
5.24	Articles with highest PageRank centralities from the German Wikipedia version	90
5.25	Articles with highest PageRank centralities from the Italian Wikipedia version	91
5.26	Articles with highest PageRank centralities from the English Wikipedia version	92
5.27	Articles with highest Eigenvector centralities from the German Wikipedia version	93
5.28	Articles with highest Eigenvector centralities from the Italian Wikipedia version	94
5.29	Articles with highest Eigenvector centralities from the English Wikipedia version	95
5.30	Partners of Bose Krishnamachari	95
5.31	Articles with highest in-degree from the German Wikipedia version	96
5.32	Articles with highest in-degree from the Italian Wikipedia version	97
5.33	Articles with highest in-degree from the English Wikipedia version	98

List of Tables

4.1	Relevant occupations in the three selected languages	41
4.2	Statistics of the program run	45
5.1	Articles about artists per language version	56
5.2	Artists available in two language versions	57
5.3	Artists available in only one language version	57

5.4	Number of links per article	58
5.5	Average number of links to other artists	61
5.6	Most popular birth- / death places in total	64
5.7	Most popular birth- / death places in the English language version	65
5.8	Most popular birth- / death places in the German language version	68
5.9	Most popular birth- / death places in the Italian language version	70
5.10	Average number of links to artists with the same citizenship as the current one	73
5.11	Average number of links to artists with the same citizenship as the Wikipedia version	75
5.12	Completeness rates to other sources of information	76
5.13	Results of the overlap calculations	79
5.14	Size comparison between the different language versions	80
5.15	Size comparison of the three language versions together	81
5.16	Metrics of the three language graphs	85
6.1	KPIs for the articles about Otto Modersohn	100
6.2	KPIs for the articles about Caravaggio	100
6.3	KPIs for the articles about Thomas Eakins	101
7.1	List of museums, their exhibitions and citizenship of artists	106
A1	Terms for a positive classification	119
A2	Terms for a negative classification	120

Bibliography

- [Eco, 2011] (2011). Free but not easy. *The Economist*. The Economist Newspaper Limited in London. published November 5, 2011.
- [Adamic and Huberman, 2000] Adamic, A. L. and Huberman, A. B. (2000). Power-law distribution of the world wide web. volume 287 of *Science*, pages 2115–2115. American Association for the Advancement of Science.
- [Aibar et al., 2013] Aibar, E., Lerga, M., Lladós, J., Meseguer, A., and Minguillón, J. (2013). Wikipedia in Higher Education: an Empirical Study on Faculty Perceptions and Practices. 5th International Conference on Education and New Learning Technologies (EDULEARN13), pages 4269–4275. IATED.
- [Aibar et al., 2015] Aibar, E., Lladós-Masllorens, J., Meseguer-Artola, A., Minguillón, J., and Lerga, M. (2015). Wikipedia at university: what faculty think and do about it. volume 33 of *The Electronic Library*, pages 668–683. Emerald Group Publishing Limited.
- [Albig et al., 2014] Albig, J.-U., Rademacher, C., Berhorst, R., Friederichs, H., Hombach, M., Kindel, C., Klüver, R., Mesenhöller, M., Rietz, C., Stratenwert, I., and Strempel, J. (2014). Karl der Große und das Reich der Deutschen. volume 70 of *Geo Epoche*. Gruner + Jahr AG & Co KG.
- [Arends et al., 2011] Arends, M., Froschauer, J., Goldfarb, D., and Merkl, D. (2011). Analysing user generated content related to art history. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, i-KNOW '11*, pages 12:1–12:8, New York, NY, USA. ACM.
- [Ayers et al., 2008] Ayers, P., Matthews, C., and Yates, B. (2008). *How Wikipedia works: And how you can be a part of it*. No Starch Press.
- [Bao et al., 2012] Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., and Gergle, D. (2012). Omnipedia: bridging the Wikipedia language gap. In *Proceedings of the SIGCHI (Special Interest Group on Computer-Human Interaction) Conference on Human Factors in Computing Systems, CHI '12*, pages 1075–1084, New York, NY, USA. ACM.

- [Bondy and Murty, 2008] Bondy, A. J. and Murty, R. U. S. (2008). *Graph theory with applications*, volume 240 of *Graduate Texts in Mathematics*. Springer Verlag London.
- [Borgatti et al., 1998] Borgatti, P. S., Jones, C., and Everett, G. M. (1998). Network measures of social capital. volume 21 of *Connections*, pages 27–36.
- [Borgatti and Everett, 2006] Borgatti, S. P. and Everett, M. G. (2006). A graph-theoretic perspective on centrality. volume 28 of *Social networks*, pages 466–484. Elsevier.
- [Brin and Page, 1998] Brin, S. and Page, L. (1998). The Anatomy of a Large-scale Hypertextual Web Search Engine. In *Proceedings of the Seventh International Conference on World Wide Web 7, WWW7*, pages 107–117, Amsterdam, The Netherlands. Elsevier Science Publishers B. V.
- [Broder, 2002] Broder, A. (2002). A taxonomy of web search. volume 36 of *SIGIR (Special Interest Group on Information Retrieval) Forum*, pages 3–10, New York, NY, USA. ACM.
- [Broder et al., 2000] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web. volume 33 of *Computer networks*, pages 309–320, New York, NY, USA. Elsevier North-Holland, Inc.
- [Brown, 2011] Brown, R. A. (2011). Wikipedia as a data source for political scientists: Accuracy and completeness of coverage. volume 44 of *PS: Political Science & Politics*, pages 339–343. Cambridge University Press.
- [Caldarelli et al., 2006] Caldarelli, G., Capocci, A., Colaioni, F., Servedio, V., Buriol, L., Donato, D., and Leonardi, S. (2006). Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. volume 74 of *Physical Review E*. American Physical Society.
- [Chelmiss and Prasanna, 2011] Chelmiss, C. and Prasanna, K. V. (2011). Social networking analysis: A state of the art and the effect of semantics. 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), pages 531–536. IEEE.
- [Clauson et al., 2008] Clauson, A. K., Polen, H. H., Boulos, K. M. N., and Dzenowagis, H. J. (2008). Scope, completeness, and accuracy of drug information in Wikipedia. volume 42 of *Annals of Pharmacotherapy*, pages 1814–1821. SAGE Publications.
- [Ding et al., 2009] Ding, Y., Yan, E., Frazho, A., and Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. volume 60 of *Journal of the American Society for Information Science and Technology*, pages 2229–2243. Wiley Online Library.
- [Faust, 1997] Faust, K. (1997). Centrality in affiliation networks. volume 19 of *Social networks*, pages 157–191. Elsevier.

- [Fitzpatrick, 2016] Fitzpatrick, A. (2016). The 15 Most Edited Wikipedia Articles of All Time. *Time Magazine*. published January 14, 2016.
- [Goldfarb et al., 2012] Goldfarb, D., Arends, M., Froschauer, J., and Merkl, D. (2012). Art History on Wikipedia, a Macroscopic Observation. *ACM Web Science 2012 (WebSci12)*, pages 163–168. ACM.
- [Graham et al., 2015] Graham, M., Straumann, K. R., and Hogan, B. (2015). Digital Divisions of Labor and Informational Magnetism: Mapping Participation in Wikipedia. volume 105 of *Annals of the Association of American Geographers*, pages 1158–1178. Taylor & Francis.
- [Halfaker et al., 2012] Halfaker, A., Geiger, S. R., Morgan, J. T., and Riedl, J. (2012). The rise and decline of an open collaboration system: How Wikipedia’s reaction to popularity is causing its decline. volume 57 of *American Behavioral Scientist*, pages 664–688. SAGE Publications.
- [Head and Eisenberg, 2010] Head, J. A. and Eisenberg, B. M. (2010). How today’s college students use Wikipedia for course-related research. volume 15 of *First Monday*.
- [Hill and Shaw, 2013] Hill, B. M. and Shaw, A. (2013). The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation. volume 8 of *PLoS ONE*, pages 1–5. Public Library of Science.
- [Lim, 2009] Lim, S. (2009). How and why do college students use Wikipedia? volume 60 of *Journal of the American Society for Information Science and Technology*, pages 2189–2202. Wiley Subscription Services, Inc., A Wiley Company.
- [Neidhardt, 2014] Neidhardt, J. (2014). Web science. lecture notes in E-Commerce. Vienna University of Technology.
- [Newman, 2001] Newman, E. M. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. volume 64 of *Physical review E*. American Physical Society.
- [Opsahl et al., 2010] Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. volume 32 of *Social Networks*, pages 245–251.
- [Parmenter, 2007] Parmenter, D. (2007). *Key performance indicators: Developing, Implementing, and Using Winning KPIs*. John Wiley & Sons.
- [Sacharidis, 2015a] Sacharidis, D. (2015a). Basic concepts. lecture notes in Web Science. Vienna University of Technology.
- [Sacharidis, 2015b] Sacharidis, D. (2015b). Network models. lecture notes in Web Science. Vienna University of Technology.

- [Serrat, 2009] Serrat, O. (2009). *Social network analysis*. Asian Development Bank.
- [Travers and Milgram, 1969] Travers, J. and Milgram, S. (1969). An Experimental Study of the Small World Problem. volume 32 of *Sociometry*, pages 425–443. American Sociological Association, Sage Publications, Inc.
- [Ultsch, 2001] Ultsch, A. (2001). Eine Begründung der Pareto-80/20 Regel und Grenzwerte für die ABC-Analyse. accessible via <http://www.mathematik.uni-marburg.de/%7Edatabionics/papers/ultsch01begrueundung.pdf> [accessed 30-January-2016].
- [Watts, 1999] Watts, J. D. (1999). Networks, Dynamics, and the Small-World Phenomenon 1. volume 105 of *American Journal of sociology*, pages 493–527. JSTOR.
- [Wilson, 2012] Wilson, J. R. (2012). *An introduction to graph theory*. Pearson Education India.

Appendix

The words and word-stems in the following table helped to positively classify articles as relevant.

abstract	abstrakt	art (without part, dart, mart)
arte_romanica	artist	astratt
barocco	barock	baroque
bauhaus	bild (without bildung, bildgebend)	bildhauer
conceptual	concettual	constructivis
contemporary	costrutt	cubis
dada	danub	espressioni
expressioni	fauvis	fluxus
galler	gothic	gotic
gotik	graph (without demo, topo, geo)	heidelberg
illustrat	image (without imaging)	impression
impressioni	konstrukt	konzeptkunst
kubis	kunst (without eis and flug)	künstl
maler	manieris	manneris
minimalis	modern	modernis
museo	museum	neoclassic
neoklass	painter	pittor
pointil	pop_	portr
portrait	puris	realis
renaissance	rinascata	ritratto
rococo	rokoko	romanti
romanticis	sculpt	scult
simboli	stucki	stuckis
surrealis	symbolis	zeich
zeitgen		

Table A1: Terms for a positive classification

The following table contains words which should not occur in an article (e.g. due to scope restrictions or due to similarities from word-stems). These terms negatively influenced

the classification of an article.

archit	album	canzone
chart	comic	figure_skating
film	filoso	foto
game	hit	lied
musi	oper	philosoph
photo	polit	railway
simphon	song	symphon

Table A2: Terms for a negative classification