TU WIEN Informatics

Art History Data across Languages and Regions

Bias Detection and the Evaluation of Visualization Techniques

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Business Informatics

eingereicht von

Danica Mitrovic, Bsc
Matrikelnummer 11832652

an der Fakultät für Informatik

der Technischen Universität Wien

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

Wien, 25. November 2025

Danica Mitrovic	Wolfdieter Merkl

TU WIEN Informatics

Art History Data across Languages and Regions

Bias Detection and the Evaluation of Visualization Techniques

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Business Informatics

by

Danica Mitrovic, Bsc
Registration Number 11832652

to the Faculty of Informatics

at the TU Wien

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

Vienna, November 25, 2025

Danica Mitrovic	Wolfdieter Merkl

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Danica Mitrovic, Bsc

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.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, habe ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT-Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 25. November 2025

Danica Mitrovic

Danksagung

Ich möchte meinem Betreuer, Professor Dieter Merkl, für seine Anleitung und kontinuierliche Unterstützung während des Schreibens dieser Arbeit meinen Dank aussprechen.

Mein größter Dank gilt meinen Eltern und meiner Schwester, die mich auf jedem Schritt dieses Weges unterstützt und dies möglich gemacht haben.

Abschließend möchte ich all meinen Freunden für ihre unermessliche Unterstützung und Ermutigung danken.

Acknowledgements

I would like to to express my gratitude to my thesis supervisor, professor Dieter Merkl for his guidance and continued support through the process of writing this thesis.

My greatest gratitude goes to my parents and sister who have supported me every step of the way and made this possible.

Finally, I would like to thank all of my friends for their immeasurable support and encouragement.

Kurzfassung

Digitale Tools und Plattformen werden zunehmend in der kunsthistorischen Forschung eingesetzt, wodurch sich die Frage stellt, ob und wie Verzerrungen in den kunsthistorischen Daten selbst die Darstellung der Kunstgeschichte beeinflussen.

Diese Arbeit untersucht das Vorhandensein von sprachlichen und regionalen Verzerrungen in Open-Source-Datenbanken, insbesondere in den mehrsprachigen Ausgaben von Wikipedia, und bewertet Visualisierungstechniken zur Darstellung kunsthistorischer Daten. Durch die Kombination einer quantitativen Voreingenommenheitsanalyse mit einer Visualisierungsbewertung im Rahmen einer strukturierten Nutzerstudie leistet diese Forschung einen Beitrag sowohl zur digitalen Kunstgeschichte als auch zur Methodik der Datenvisualisierung.

Zur Untersuchung der Voreingenommenheit in Open-Source-Datenbanken wurden Daten aus sechs Sprachausgaben von Wikipedia, Wikidata und Getty ULAN anhand einer Reihe von Leistungskennzahlen (KPIs) analysiert: Sprachübereinstimmungsrate, Local Hero Score, Repräsentationsverhältnis und Artikel-Metadaten. Die Arbeit untersuchte, inwieweit jede Wikipedia-Sprachausgabe Künstler bevorzugte, die mit ihrer jeweiligen Sprache und Region in Verbindung stehen. Die Ergebnisse der Analyse bestätigten, dass bestimmte Wikipedia-Ausgaben dazu neigten, Künstler aus ihrer Region zu bevorzugen und Künstler, die nicht aus ihrer jeweiligen Region stammten, unterrepräsentiert waren, und zeigten Unterschiede in der Vollständigkeit und Qualität der Artikel auf. Diese Ergebnisse unterstreichen, wie der Redaktionspool das kunsthistorische Wissen prägt und wie wichtig es ist, Daten gegenseitig zu referenzieren.

Das zweite Forschungsziel konzentrierte sich auf die Bewertung verschiedener Visualisierungstechniken zur Darstellung der Bewegungen und Verteilung von Künstlern. Heatmaps, Flowmaps und Sankey-Diagramme wurden daher im Rahmen einer Nutzerstudie bewertet, an der Teilnehmer mit und ohne kunsthistorische Vorkenntnisse teilnahmen. Die Bewertung erfolgte anhand quantitativer und qualitativer Fragen, mit denen die erzielte Genauigkeit und die subjektive Wahrnehmung analysiert wurden. Die Ergebnisse zeigten zwar, dass keine einzelne Visualisierungstechnik perfekt war, jedoch wurden Heatmaps als bevorzugte Option angesehen, da sie die Dichte und geografische Konzentration klar vermittelten, während Flusskarten bei der Visualisierung von Bewegung und Richtung bevorzugt wurden. Durch das Feedback aus den qualitativen Fragen wurden Verbesserungsmöglichkeiten aufgezeigt und eine Präferenz für interaktive und hybride Visualisierungen deutlich.

Die Ergebnisse dieser Arbeit leisten einen doppelten Beitrag: eine quantitative Analyse der sprachlichen und regionalen Verzerrungen in mehreren Sprachversionen von Wikipedia und eine vergleichende, nutzerzentrierte Bewertung von Visualisierungstechniken zur Vermittlung von Mustern künstlerischer Migration. Die Ergebnisse zeigen, dass ein kontinuierlicher Bedarf an Zusammenarbeit zwischen den Disziplinen Informatik und Kunstgeschichte besteht.

Abstract

Digital tools and platforms have been progressively used in art history research, raising a need to understand if and how biases in the very art historical data shape the representation of art history.

This thesis investigates the presence of language and regional bias in open-source databases, specifically within Wikipedia's multilingual editions, and evaluates visualization techniques for representing art historical data. By combining quantitative bias analysis with visualization evaluation through a structured user study, this research contributes to both digital art history and data visualization methodology.

For the examination of bias in open-source databases, data from six Wikipedia language editions, Wikidata, and the Getty ULAN were analyzed using a set of key performance indicators (KPIs): Language Match Rate, Local Hero Score, representation ratio, and article metadata. The thesis examined the extent to which each Wikipedia language edition favored artists associated with their respective language and region. The results of the analysis confirmed that specific Wikipedia editions tended to favor artists from their region and in turns under-represent artists that were not from their respective region, and found disparities in the article completeness and quality. These findings underscore how the editorial pool shapes art historical knowledge and the high importance of cross-refencing data.

The second research objective focused on the evaluation of different visualization techniques for representing the movements and distribution of artists. Heatmaps, flow maps and Sankey diagrams were thus evaluated through a user study, which included participants with and without prior art historical knowledge. The evaluation was done via quantitative and qualitative questions that analyzed the achieved accuracy and subjective perception. While the results showed that no singular visualization technique was perfect, heatmaps were the preferred option as they communicated clearly the density and geographic concentration, while flow maps were favored in visualizing movement and directionality. Through the feedback from qualitative questions, areas of improvement were highlighted and a preference for interactive and hybrid visualization.

The findings of this thesis offer a dual contribution: a quantitative analysis of language and regional bias across multiple Wikipedia language editions and a comparative user-centered evaluation of visualization techniques for communicating patterns of artistic migration.

The results show that there is an ongoing need for collaboration between the disciplines of computer science and art history.

Contents

CHAPTER 1

Introduction

1.1 Background and Motivation

By digitizing cultural heritage, new opportunities have been created for analyzing art history. Through the usage of digital tools and publicly accessible databases, art historical data can be explored and visualized in new and exciting ways. Platforms such as Wikipedia and Wikidata are used by millions of users as points of reference. By combining them with databases such as the Getty Union List of Artist Names (Getty ULAN)[1] that provide research-curated data, these resources offer previously unprecedented breadth and accessibility. However, the use of these resources raises questions about the presentation of this data. Open source databases reflect the linguistic and regional priorities of their contributors, while research-curated databases favor more well-known figures and traditions. This produces a skewed recording of art history, which consequently increases the risk of creating a biased representation in public understanding and research.

A central issue that this thesis explores is the presence of linguistic and regional bias in open-source databases. Wikipedia has many language editions, and each of them is shaped by a separate editorial community. This leads to differences in aspects such as how artists are recorded, how richly they are cited, and how detailed their biographies are among others. These differences can, in turn, lead to over- and under-representation of certain artists and regions. This hinders the creation of a neutral view of art history. The topic of bias in Wikipedia has been an ongoing issue, and this thesis will work on documenting and measuring it.

The second issue this thesis explores is the visualization of historical data. Different visualization techniques, such as heatmaps and flow maps, have been used in representing

[1] Getty Union List of Artist Names (Research at the Getty), https://www.getty.edu/research/tools/vocabularies/ulan/, Accessed on: 2025-08-27

such data, but little research has been done in evaluating how effective different visualization techniques communicate such data in the field of art historical research. Different visualization techniques emphasize certain aspects of data while obscuring others and thus this thesis aims to evaluate different visualization techniques in order to uncover the most effective one for art historical research.

1.2 Aim of this Thesis

This thesis aims to address both of these challenges. By quantifying language and regional biases in Wikipedia through the application of key performance indicators (KPIs) across six language editions (English, German, French, Spanish, Italian, Serbian), with cross-validation against Getty ULAN. Additionally, this thesis aims to analyze different visualization techniques and their effectiveness for the visualization of the movement of painters. This will be done through the use of a questionnaire. The questionnaire uses task-based accuracy questions, Likert-scale evaluation and qualitative questions to form a structured study.

1.2.1 Research Questions

This thesis aims to answer two main research questions:

- How do linguistic and regional biases in open-source databases in Wikipedia affect the representation of artists and art history?
- Which visualization methods are most effective for representing the movements of artists and the distribution of artistic activity across Europe, as measured by a comparative data analysis?

1.3 Expected results

The expected results of this thesis are measurable insights into the language and regional bias in art historical data across 6 different Wikipedia language editions and insights into the effectiveness of different visualization techniques for communicating patterns of artistic migration across Europe from the 15th to the 20th century.

By using quantitative analysis on data from multiple Wikipedia language versions combined with data from Wikidata and the Getty ULAN, this thesis is expected to confirm the existence of bias based on the language and region of the artist. The hypothesis is that each Wikipedia language edition will show preference for artists associated with its own cultural background. We will quantify this bias through the use of multiple KPIs, such as language match rate, local hero score, representation ratio across languages and article metadata (e.g. number of references and internal links as measures of article quality). Additionally, a comparison with Getty ULAN is expected to highlight the differences in representation between publicly curated databases and research-curated databases.

These results are expected to show how open-source databases mirror the sociolinguistic priorities of their contributor communities, which in turn shape the perceived narrative of art history.

For the research on different visualization techniques, a comparative evaluation is expected to showcase which visualization technique yields the most clarity and usability in showcasing painter migration patterns across Europe. A structured user study with two groups of participants is going to be conducted - we split the participants according to their exposure to art history into those who have studied art history and have prior knowledge, and those without any prior knowledge. The expected results are that heatmaps are the best for visualizing density and offer the most clarity, while flow maps display movement more accurately. Additionally, no single visualization technique is expected to produce optimal results, but an integrated approach combining different techniques will provide the most comprehensive understanding of the data set.

In conclusion, the expected result is a dual contribution. An understanding of how language and regional structure affect the representation of artists in publicly available databases, and an understanding of which visualization technique is best suited for understanding the migration of artists.

1.4 Structure of this Thesis

This thesis is structured to progress from a theoretical framework to an empirical analysis and evaluation, culminating in practical recommendations for digital art historical research.

Chapter 1 – Introduction

This chapter establishes the research problem and motivation. Additionally, the specific research questions are defined, as well as the scope of the work.

Chapter 2 – Theoretical Framework and Related Work

An overview of prior research on the topic of bias in open-source databases and the use of visualization methods in art historical research is presented.

Chapter 3 – Data Sources and Collection

The datasets (Wikipedia, Wikidata, Getty ULAN), extraction methods (e.g., SPARQL), inclusion criteria (time, geography, metadata completeness) and data cleaning and integration methods are documented in detail.

Chapter 4 – Methodology

The CRISP-DM framework and its phases are presented. For each phase, the connection to the thesis is illustrated.

Chapter 5 – Bias Analysis Across Language Edition

Introduces the selected KPIs and the reasoning behind them. Presents results for coverage and metadata differences (references, internal links, categories, sections) across language editions, and interprets these findings in relation to the research questions.

Chapter 6 – Visualization Techniques Assessment

Describes of the different visualizations techniques studied(heatmaps, flow maps, Sankey diagrams), the questions for the user study and the chosen KPIs for qunatifying the effectiveness of the visualization techniques.

Chapter 7 – Results and Analysis of the User Study

Gives us overview of the user study setup as well as documenting the results of the user study, and the connection to art history research with future insight.

Chapter 8 – Discussion

An overview of the findings of the bias analysis and visualization user study, with a special reflection on the limitations and validity.

Chapter 9 – Conclusion and Future Work

A summary of the contribution of this thesis and an outline of subsequent research questions and application.

CHAPTER 2

Theoretical Framework and Related Work

In this chapter, we present the theoretical framework used to explore the topic of this thesis. In particular, we discuss the theoretical foundations used for the exploration of language and regional biases present in open-source data from art history, as well as for the visualization of historical data. Concurrently, we discuss related work that highlights similarities and differences of previous work with respect to this thesis.

2.1 Bias in Open-Source Databases

Open-source databases such as Wikipedia and Wikidata have played an integral role in the distribution and acquisition of knowledge in various fields. Although they are important tools for making knowledge widely accessible to the public and have been used extensively in research, it is known that they contain different types of bias [Hube, 2017]. One of the fundamental principles of Wikipedia is the neutral point of view (NPOV), which aims to present data fairly, proportionately and without editorial prejudice[1]. Although NPOV is one of the leading principles of Wikipedia, the open-source nature of these tools and the fact that they are being curated by a large editorial community can reflect and amplify already existing societal biases or introduce new ones – data does not just portray the world, it also shapes it. In addition, bias in data can distort the perception of reality and perpetuate already existing inequalities. That is one of the reasons why data ethics which govern the usage, collection and management of data are of such importance[2]. This is important because data can further amplify existing inequalities and give greater significance to the perspectives of dominant groups, while marginalizing others [Lopez, 2021] [Chen, 2023].

[1] Wikipedia: Neutral point of view, https://en.wikipedia.org/w/index.php?title=Wikipedia:Neutral_point_of_view&oldid=1279876337, Accessed on: 2025-04-01

[2] 5 Principles of Data Ethics for Business, https://online.hbs.edu/blog/post/data-ethics, Accessed on: 2025-10-22

To understand the biases present in open-source databases, we first need to define the concept of bias. Bias refers to systematic tendencies that lead to uneven or distorted representations of reality in data[Ahmad et al., 2020]. Bias can come in many forms for example, differing levels of editorial attention, the selection of subjects, and uneven representation. In this research, the focus is on data bias in open-source databases such as Wikipedia and Wikidata in the field of art history, more specifically, in the representation of painters in these databases. It is essential to understand the biases present in these databases in order to evaluate the reliability of such databases and their effect on the visibility and portrayal of information as experienced by the general public. This is particularly important in the field of art history as this discipline is already shaped by a more Eurocentric and male-dominated view[3]. As digitization plays an important role in art history (e.g. through platforms like Wikipedia), inherent biases in this type of knowledge can be amplified by the editors in charge of digitizing data, thus creating new layers of distortion.

2.2 Forms of Bias in Open Source Databases

In this section, we discuss some of the most relevant biases present in open-source databases. In particular, we discuss gender, structural and language and regional bias.

2.2.1 Gender Bias

One of the most widely recognized and studied forms of data bias is gender bias. This type of bias refers to the phenomenon in which data collection and analysis are distorted due to systematic favoritism based on gender[4]. It occurs with the over- or under-representation of figures based on gender[5]. Since data is crucial in shaping research, policies, and technologies, this type of bias can have significant real-life implications[Perez, 2019].

Although one of the main principles of Wikipedia is its neutral point of view (NPOV) and it aims to be an inclusive and neutral platform [1], multiple studies have shown that the content of Wikipedia is significantly shaped by its contributors, their background, interests and inherent biases [Wagner et al., 2015]. It is known that approximately only 15% of active editors identified as women (see Figure 2.1 [6]) and thus the editorial process is largely shaped by the predominantly male editorial pool. This directly influences the content of Wikipedia and can reinforce existing gendered patterns of knowledge production. The imbalance in the editorial pool is one of the factors that contributes to the "gender knowledge gap", where topics related to men tend to be overrepresented, while those related to women tend to be underrepresented and marginalized

[3] Transvergence in Art History, https://www.ekac.org/transvergence.html, Accessed on: 2025-11-08

[4] Gender Bias in Data and Tech — COFEM, https://cofemsocialchange.org/gender-bias-in-data-and-tech/, Accessed on: 2025-10-21

[5] Data Bias: understanding the Gender Data Gap, https://hospitalityinsights.ehl.edu/data-bias-gender-data-gap, Accessed on: 2025-10-21

[6] Community Insights/Community Insights 2023 Report - Meta-Wiki, https://meta.wikimedia.org/wiki/Community_Insights/Community_Insights_2023_Report, Accessed on: 2025-10-21

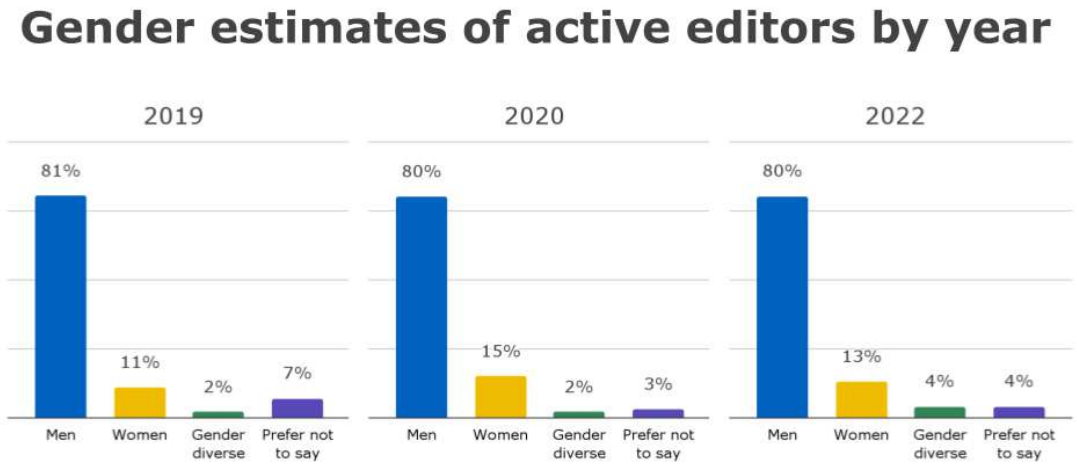


Figure 2.1: Gender estimates of active editors by year[6]

[Thelwall et al., 2019] [Collier and Bear, 2012]. This can also be seen by the fact that 22% of all biographies in Wikipedia are about women [Zhang and Terveen, 2021]. In addition to that, the content of biographies about women also often has a different tone, focusing more on topics such as personal life and relationships in contrast to professional achievements [Graells-Garrido et al., 2015] [Wagner et al., 2016].

2.2.2 Structural Bias

Structural bias refers to the bias in outcomes due to the characteristics of the system itself[7] . Extensive research has been conducted about the structural bias of Wikipedia and it was discovered that this bias stems from a combination of Wikipedia's network architecture and the collective choices of editors. Specifically, in Wikipedia and Wikidata this bias can manifest due to editorial patterns and linkages inside of the Wikipedia network. Research about the Wikipedia hyperlink network has identified bias in the manner articles about men and women were interlinked, showcasing systematic structural bias within [Wagner et al., 2015]. Specifically, it was shown that articles about women were more linked to men than vice versa, which can lead to decreased visibility of women on Wikipedia[Wagner et al., 2015]. The structure of Wikipedia, namely its navigation paths, through filtered knowledge exposure and the creation of polarized views, can lead to bias [Menghini et al., 2019].

[7]Structural Bias — Term, https://fashion.sustainability-directory.com/term/structural-bias/, Accessed on: 2025-10-22

2.2.3 Language and Regional Bias

Understanding language and regional bias is crucial in multilingual databases such as Wikipedia, which has more than 300 language editions that vary in scale and breadth, both in the number of articles and the size of editorial pools in the different language versions[8]. Most articles across different language editions are not just simple translations of one another, but are each curated by a dedicated editor base. As such, they are susceptible to the local editorial cultures and shaped by the historical and political perspectives of its editors. Consequently, this produces different representations of the same artist and cultural topic across language editions.

This fragmentation of the database across languages and regions can lead to the creation of bias, e.g. with some editorial communities having detailed coverage of figures and topics connected to their own cultural background. This occurrence has has been referred to as the "local hero effect" [Callahan and Herring, 2011] [Eom and Shepelyansky, 2013] in recent research. As a prominent part of the editorial base is from the Western world, knowledge will be skewed towards Western topics and figures [Hern, 2015]. In particular, the research found that in 2015, almost 45% of all edits were written by editors from the United Kingdom, the United States, France, Germany, and Italy. Edits from other countries were also often about Europe and North America, further perpetuating the dominance of the content of Western countries.

The background of editors also perpetuates distortion. Participation in projects such as Wikipedia is correlated with education, institutional support, and also with access to the internet. This leads to a lower editorial base in places in the world where there is limited access to these resources, which directly contributes to the lower cultural visibility of these places and cultures.

2.3 Research on Content and Representation Differences across Wikipedia Language Editions

Differences across Wikipedia language editions have been known to the scientific community for a while and have been the topic of many research papers. That body of work provides the theoretical foundation for this thesis. In the paper "Detecting Cross-Lingual Information Gaps in Wikipedia," a computational approach is used to identify and analyze the disparities across Wikipedia language editions [Ashrafimoghari, 2023]. They concluded that while Wikipedia has over 300 language editions, half of all articles on Wikipedia are produced by a small number of languages (ten language editions cover around half of all Wikipedia articles). This highlights the gap of information and knowledge that is present across different languages in Wikipedia [Ashrafimoghari, 2023].

This information gap also stems from the low content overlap across language editions. A study by [Roy et al., 2020] analyzed articles on the same topics across nine Wikipedia

[8] The language geography of Wikipedia, https://internetlanguages.org/en/numbers/wikipedia-language-geography/, Accessed on: 2025-10-22

language editions and quantified differences in coverage and depth. By using different metrics, it was shown that there are substantial gaps and low overlap across languages and that the largest difference could be seen in languages that had a small editorial community compared to the English Wikipedia [Roy et al., 2020].

Articles from 40 different Wikipedia language editions were analyzed in the research done by [Miquel-Ribé and Laniado, 2018]. They used several techniques, such as geolocated article analysis and specific keyword and category matching, to label "cultural context content". Based on their analysis, they concluded that on average, a quarter of all content in a Wikipedia language is based on the respective cultural context of that specific Wikipedia edition, with this number ranging between 7% and 49%. Around half of all articles that were labeled as culturally connected to a specific language edition existed only in that specific language edition. This showcases the cultural content gap present across Wikipedia language editions and explains why content diversity and editorial focus are closely connected to the language and regional community within Wikipedia [Miquel-Ribé and Laniado, 2018].

Apart from differences in the breadth of coverage and depth, there are differences in the structure and quality of articles present across Wikipedia editions. In the work of Lewoniewski, W., Węcel, K., Abramowicz, W. (2017), metrics such as article length, number of references and number of sections among others, was used as an indicator of the quality of articles [Lewoniewski et al., 2017b]. By analyzing these metrics, they revealed significant differences in Wikipedia articles across language editions and domains.

Taken together, these studies show a consistent pattern of Wikipedia not being a neutral source but a database that reflects that the culture, language and origin of its editors in its content. Each language edition functions as its own database, interconnected and influenced by other language editions but operating at the same time as its own repository of knowledge. Thus, we see that measuring the bias present in Wikipedia involves quantifying the degree of overlap, divergences and representation imbalance between different Wikipedia language editions. As each language edition functions as both a reflection and a producer of cultural identity, measuring bias, therefore, involves quantifying the degree of overlap, divergence, and representation imbalance among the respective communities - a task directly addressed in this thesis through quantitative KPIs and visual comparative analysis.

2.4 Digital Tools in Art History

The use of digital tools has reshaped how researchers collect, analyze and present data in the field of art history. Through the use of digital tools and computational analysis, new findings can be revealed. Despite this, the use of such tools needs to be critically examined because of the nature of the investigated data [Surkemper et al., 2019].

Till now, art history has relied on qualitative methods for the exploration and visualization of its findings [Tuscher et al., 2025]. But through the increasing availability of

digitized archives, the field of art history has gradually started to accept visualizations as valuable tools for analyzing, interpreting and communicating complex historical data [Tuscher et al., 2025]. But there is still resistance to new tools in the field of digital humanities such as art history, with many art historians opposing new digital tools as they feel that they are not suitable [Lamqaddam et al., 2018].

To foster a wider acceptance of such tools, a user-centered design must be implemented. A user-centered design process encompasses an iterative design process that focuses on the needs and feedback of the user through the development process [9]. Through the involvement of art historians in the development process, a greater level of trust can be forged and a higher level of usage achieved as these tools would directly support their workflows. This applies to all digital tools, including visualization techniques. Visualizations need be created with the requirements of art historians in mind, by balancing technological innovation with disciplinary practices [Lamqaddam et al., 2018]. Visualizations serve not only as aesthetic tools but are essential to achieve a deep understanding of the data.

It is also of great importance to create a clear distinction between two terms, digitized art history and digital art history [Surkemper et al., 2019]. Digitized art history refers to art historical resources that are made available in a digitized form through the use of tools such as for example an online image collections, while on the other hand digital art history uses computational methods and analytical techniques to answer new questions and gain insights [10].

The theoretical framework of this thesis is based on the intersection of three domains: Digital and Computational Methods in Art History, Visualization Theory and Art Historical Data, and User Studies and Visualization Evaluation. Through combining these domains, a framework is created that supports one of the two main goals of this thesis, which is evaluating different visualization methods on their interpretability and clarity in representing art historical data through a user study. In the following sections, the critiques surrounding the usage of digital tools in this field will be addressed as well as the advantages and potential they have.

2.4.1 Disadvantages and Critiques of Digital Art History

A crucial component in research is the assumption that computational analysis always offers objective results. Researchers often use vast datasets first before asking art historical questions [Surkemper et al., 2019]. This can lead to the creation of topics and clusters of results that are then reinterpreted as historical insight. This approach leads to the avoidance of argumentation as data patterns replace explanation, which in the field of

[9] What is User Centered Design (UCD)? — updated 2025 — IxDF, https://www.interaction-design.org/literature/topics/user-centered-design?srsltid=AfmBOoo4JdiCl_CETVMOqnRG4gkID76lZ4WhvIqdY5JNnd2biaJlCChR, Accessed on: 2025-11-09

[10] Artists Archives and Digital Art History: Imagining the Possibilities, https://artiststudioarchives.org/2016/03/21/artists-archives-and-digital-art-history-imagining-the-possibilities/, Accessed on: 2025-11-26

art history is of utmost importance [Surkemper et al., 2019]. Bishop argues that data is handled as a self-evident, complete solution and not merely a tool for further examination [Bishop, 2018]. This leads to the risk, where art historical research is viewed as a form of data management and where important points, such as ambiguity and uncertainty, that have defined the humanities are erased [Hall, 2013].

This view aligns with the research of Wendy Brown on the neoliberalization of knowledge [Brown, 2015]. She argues that cultural heritage is quantified through metrics, ranking and count, rather than its cultural significance. Knowledge is measured through efficiency and objective measurement instead of reflection and critique, which is central in a field such as art history [Brown, 2015]. This can be seen in research such as "Computerized Face Recognition in Renaissance Portrait Art: A quantitative measure for identifying uncertain subjects in ancient portraits" [Srinivasan et al., 2015] and "History of art paintings through the lens of entropy and complexity"[Sigaki et al., 2018]. The latter uses numerical scales derived from spatial ordering patterns in paintings to reveal trends in visual arts [Sigaki et al., 2018]. These research papers offer until now unseen insights into art historical knowledge through the use of digital tools but cannot be seen as standalone work as they do not encompass the political and cultural background but rather focus on the interaction of metrics as a whole without further reflection and critique.

After outlining the theoretical background for the positioning of this research, it is essential to acknowledge that the integration of digital tools in the field of art history is promising but contested.

2.4.2 Advantages of Digital Art History

The use of digital tools in art historical research can come with its limitations, but it is important to note the positive influence it can have on creating new insights. Digital art history can create new data-informed insights, as seen in cases where machine learning algorithms are used to authenticate artwork (the machine learning model was trained on artwork by Raffaello Sanzio da Urbino)[Ugail et al., 2023]. This type of research can help the more traditional approach of analyzing and authenticating artwork, proving that computation can extend rather than replace art historical insights.

One of the biggest benefits of digital art history and digitized art history is the ability to democratize access to cultural heritage. An example of such initiatives are platforms such as Google Arts & Culture [11] and museum open access initiatives (e.g. The Met's Open Access program[12]) have enabled the digitization of millions of high-resolution images. The outcomes of these initiatives give unprecedented access to artwork to the broader public and researchers. Such initiatives make research using digital tools even possible and support broader public engagement. Digital archives and computer analysis make it possible to study art history at till now unprecedented levels, helping to uncover temporal

[11] Google Arts & Culture, https://artsandculture.google.com/, Accessed on: 2025-10-25

[12] Open Access - The Metropolitan Museum of Art, https://www.metmuseum.org/about-the-met/policies-and-documents/open-access, Accessed on: 2025-10-25

and stylistic patterns that are not accessible through traditional qualitative methods used in art history [Kim et al., 2014].

Additionally, the exploration of patterns through the use of large-scale data analysis complements traditional research and is one of the strong points of digital art history. Research such as "Dissecting landscape art history with information theory" produced new findings into the topic of western landscape paintings by analyzing 14,912 landscape paintings and detecting the evolution of the preferred compositional proportion within landscape paintings [Lee et al., 2020]. Such research enables art historians to ask new research questions that till now would not be impossible through the use traditional methods and local archives. But these analyzes do not replace interpretation; they simply provide new foundations for it.

In this thesis, we will examine one of the tools used in digital art history, which are data visualizations. Visualizations have an important role in the way complex data sets can be analyzed and interpreted. However, it is important to note that visualizations are not neutral representations but are influenced by design choices and assumptions that can skew understanding [Drucker, 2014]. Because of that, it is essential to analyze the potential but also the limitations of using visualizations in the field of art history.

2.4.3 Visualizations in Art Historical Research

Researchers use visualizations to examine complex and often fragmented datasets in order to reveal patterns and relationships that would perhaps remain obscured in traditional text. Visualizations allow for the exploration and interpretation of phenomena such as the migration of artists, the formation of cultural hubs, and the evolution of art styles through regions and centuries. By using visualization techniques such as heatmaps, flow maps, and Sankey diagrams, researchers have the possibily to explore in ways that were not possible through qualitative analysis.

Researchers in the field of humanities such as art history emphasize the need for more user-centered design in the development of visualization tools [Lamqaddam et al., 2018] [Windhager et al., 2019]. A reason for is that art historians approach data with distinct epistemological traditions, which are grounded in contextual interpretation, visual literacy, and narrative reasoning. In order to ensure interpretability and usability iterative evaluations need to be done to ensure that visualizations do not obscure complexity but rather enhance understanding and engagement. Within this framework, visualization becomes a bridge between technological innovation and art historical inquiry, an active site where visual thinking and computational analysis converge.

An example of this would be the research of Tuscher et al. (2025), which reveals how visualizations of nodes, edges, and "artistic wedges" showcase the mutual connections among artists, museums, and cultural centers [Tuscher et al., 2025]. The visualization of networks allows us to see patterns that would be obscured in more traditional narratives. This shows that visualizations are more than just illustrations but rather a mode of

knowledge production through which one can expose patterns of influence and clustering within the art world [Tuscher et al., 2025].

These studies emphasize the usefulness of digital visualizations in art history and allow scholars to gain a deeper understanding of cultural centers, patterns of migration, thus redefining how art history is studied, presented and communicated.

2.5 User Study

As art historical research has adopted more digital tools to interpret and contextualize vast amounts of data, visualizations have been a popular medium for these purposes. As already mentioned, visualizations are not neutral carriers of information, but are susceptible to human perception and interpretation, and as such need to be evaluated [Card et al., 1999]. In this thesis, we evaluate visualization techniques on datasets concerning artistic migration and the formation of cultural hubs using a user study. The user study measures the clarity, comprehension, and overall usability of such visualizations among expert and non-expert audiences. Performance metrics such as computation time and data completeness are quantifiable and important metrics, but they do not reveal how effective these visualizations are in communicating concepts to users. User studies can be used in order to bridge the gap between visualizations and understanding, by observing how users interact with visual representations [Munzner, 2025]. User studies are part of the broader context of human–computer interaction (HCI), the multidisciplinary field that investigates the interaction between humans and computers, and emphasizes the need for clarity, usability and accuracy [Rapp, 2023].

Research emphasizes that viewers of visualizations focus on the mappings of visual variables such as color, shape, and spatial arrangement in connection to the underlying data[Schloss, 2025]. Concepts such as interpretability and memory retention are enhanced in cases where the visual features of such visualization are arranged on principles of cognitive psychology Perceptual and Cognitive Foundations of Information Visualization [Schloss, 2025]. This would indicate that in order to maximize the potential benefits of a visualization they need to be constructed with the user in mind.

For that reason, user studies are instrumental in assessing how visualization design is perceived and provide empirical evidence of what users actually understand and perceive. In the context of this thesis, where visualizations are applied to an art historical dataset, these evaluations are crucial. Overly complex visualizations can decrease comprehension by exceeding users' memory capacity according to the Cognitive load theory [Sweller, 1988]. Therefore, assessing the usability and clarity of visualizations through structured questions using Likert scale evaluations can directly portray which visualizations are the most efficient for the intended purpose.

2.5.1 Empirical Evaluation in Visualization Research

Empirical evaluation in the field of visualization research can be categorized into several different methods, such as analytical, observational and experimental [Lam et al., 2011]. In this thesis, we will rely on the experimental methods, a user study that combines quantitative and qualitative questions, to compare different visualization techniques under controlled conditions.

Experimental approaches often use tasks to measure accuracy, efficiency and workload, which is also done in the user study [Lam et al., 2011]. Through the use of task-based accuracy questions (e.g., identifying regions of artistic migration) and subjective ratings using Likert-scale evaluation of perceived clarity, engagement, and usefulness as well as open-ended questions.

Using both quantitative and qualitative instruments through the use of numerical data such as (Likert scales, accuracy scores, and qualitative feedback (open-ended comments), a 'mixed method' is used in the user study. This method is based on the paper by Onwuegbuzie and Dickinson (2008)[Onwuegbuzie and Dickinson, 2008], where quantitative data collection and analysis is followed by qualitative exploration and interpretation to create a better understanding and interpretation of the results.

Different practices were used in the creation of this user study, such as:

- Controlled experimental design: tasks need to be structured in a way that ensures fair comparison between different visualization types to enable objective measurement[Lam et al., 2011]. This was done by doing the same tasks across each visualization type (heatmap, flow map, Sankey diagram).
- Representative user sampling: Research suggests that the interpretation of visualization tasks varies to a significant degree based on the experience of the viewer and suggests differentiating participants in the sampling and analysis phase [Lee et al., 2017]. This was done by differentiating the participants in the user study by domain expertise (participants who have studied art history and, therefore, are familiar with the topic presented and people who have not studied art history).
- Performance and preference measurement: research suggests that for a more holistic evaluation, both objective and subjective measures must be taken into account[Lezcano Airaldi et al., 2025]. This was done using task-based accuracy questions (objective) and Likert scale evaluation questions concerning themes such as "Ease of Understanding".

2.5.2 Evaluation Criteria

For the quantitative questions, a Likert-scale evaluation was used, leveraging several different criteria. The chosen evaluation criteria are based on several different research papers on the topic of constructing a framework for the evaluation of visualization topics and the specific dataset:

- Ease of understanding: is one of the key usability dimensions as it impacts the adoption by users and task performance [Bresciani and Eppler, 2015].
- Clarity of migration: specific to the chosen dataset, relates to how well spatial patterns are communicated.
- Pattern recognition: measures if the user was able to recognize patterns. The ability to recognize patterns is related to a deeper understanding [Lezcano Airaldi et al., 2025].
- Information load: a metric that measures whether the user perceives the visualization as overwhelming, clustered, or if too little information is represented. If there is too much information, retention of data becomes an issue due to cognitive load[Sweller, 1988].
- Engagement: measures how useful the user thinks the visualizations would be for similar data.
- Further engagement: captures the interest of users for data exploration, because engagement can correlate with many information are retained and willingness to use that particular visualization[Lezcano Airaldi et al., 2025].
- Added value versus text: evaluates whether the user perceived information through visualizations that would not have been noticed in the data presented in textual or table format[Bresciani and Eppler, 2015].

CHAPTER 3

Data Sources, Collection and Processing

3.1 Overview of Data Sources: Wikipedia, Wikidata and Getty ULAN

This thesis relies on three complementary data sources to understand the bias in the representation of painters across different digital platforms and to evaluate visualization methods. In order to construct a comprehensive picture, a combination of publicly curated sources (Wikipedia[1] and Wikidata [2]) and an institutionally curated dataset (GettyULAN [1]) was used.

Each of these data sources has distinct characteristics that, when combined, bring together the breadth of public contributions with the depth of expert curation.

Wikipedia is the central database for the analysis of the language and regional biases present in open-source databases. It is one of the largest publicly available multilingual databases and covers more than 300 languages [Roy et al., 2022]. As an open-source community-curated database, each language edition mirrors the priorities of its editors. However, due to the nature of the unstructured and semi-structured content, large-scale comparison can be challenging and requires additional extraction and cleaning of the data.

Wikidata serves as a secondary knowledge database which stores structured data from Wikipedia, by linking concepts through the use of unique identifiers [2]. While Wikipedia is organized in such a way as to ensure readability for humans, Wikidata is organized around the use of unique identifiers (QID) and the Wikidata Query Service [Odell et al., 2022] [3].

[1] Main Page, https://en.wikipedia.org/w/index.php?title=Main_Page&oldid=1319805659, Accessed on: 2025-11-08
[2] Wikidata, https://www.wikidata.org/wiki/Wikidata:Main_Page, Accessed on: 2025-08-27
[3] Wikidata Query Service, https://query.wikidata.org/, Accessed on: 2025-10-28

By consolidating both data sources, it is possible to align and analyze the content across Wikipedia language editions.

In contrast, Getty ULAN is a structured vocabulary that comprises names, biographies and metadata on historical figures from the world of art and architecture [4]. It is a controlled vocabulary that is peer reviewed and institutionally curated [Harpring, 2010a]. In the context of this thesis, it is used as a reference point in order to highlight not only biases in Wikipedia, but also the presence of systemic gaps.

Combining these sources offers both breadth and depth. As they differ in scope, consistency and reliability, their analysis demands a critical approach to data extraction and cleaning. In the following sections, each source is described in more detail. We also discuss the strategies and tools used for data collection and cleaning.

3.2 Wikipedia

Wikipedia is one of the largest open-source encyclopedias in the world [5]. It was created in 2001 by Jimmy Wales and Larry Sanger and was envisioned as an open-source encyclopedia based on the Wiki software [6]. The Wiki software was introduced in 1995 by Ward Cunningham as a programming language pattern that enables users to access and change the data on a website [6]. In the first year, Wikipedia expanded from English to 18 different languages and over 20,000 articles [7]. Over the past two decades, it has expanded to over 62 million articles and over 300 languages, with the number of articles growing each day. English Wikipedia remains the largest and most edited edition of Wikipedia, with around 7 million articles [8] . In the Figure 3.1 [9], an example of a Wikipedia article can be seen.

The mission of Wikipedia and the Wikimedia Foundation, which hosts Wikipedia, is "to empower and engage people around the world to collect and develop educational content under a free license or in the public domain, and to disseminate it effectively and globally" [10].

[4] Union list of artist names (ULAN) (Getty), https://collectionstrust.org.uk/resource/union-list-of-artist-names-ulan-getty/, Accessed on: 2025-04-01

[5] Wikipedia is 20, and its reputation has never been higher, https://www.economist.com/international/2021/01/09/wikipedia-is-20-and-its-reputation-has-never-been-higher, Accessed on: 2025-08-27

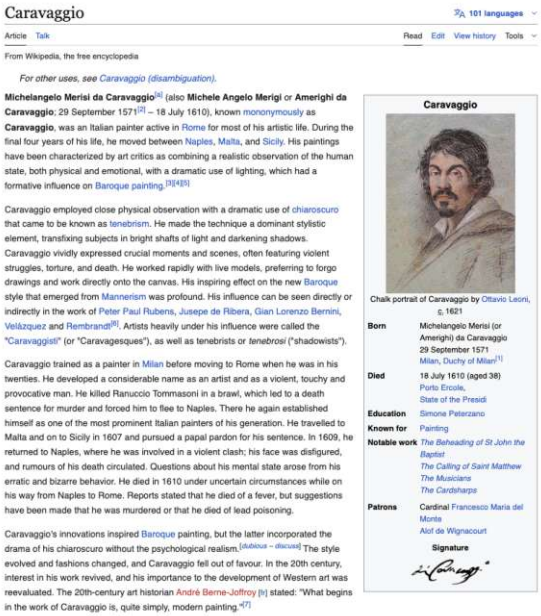
[6] History of Wikipedia, https://en.wikipedia.org/w/index.php?title=History_of_Wikipedia&oldid=1319952874, Accessed on: 2025-11-08

[7] Wikipedia | Definition, Encyclopedia, History, & Facts — Britannica, https://www.britannica.com/topic/Wikipedia, Accessed on: 2025-04-05

[8] Wikipedia: Size of Wikipedia, https://en.wikipedia.org/w/index.php?title=Wikipedia:Size_of_Wikipedia&oldid=1283351169, Accessed on: 2025-04-05

[9] Caravaggio, https://en.wikipedia.org/w/index.php?title=Caravaggio&oldid=1320579564, Accessed on: 2025-11-08

[10] Wikimedia Foundation Mission, https://wikimediafoundation.org/about/mission/, Accessed on: 2025-04-05



Caravaggio

101 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

For other uses, see Caravaggio (disambiguation).

Michelangelo Merisi da Caravaggio[a] (also **Michele Angelo Merigi** or **Amerighi da Caravaggio**; 29 September 1571[2] – 18 July 1610), known mononymously as **Caravaggio**, was an Italian painter active in Rome for most of his artistic life. During the final four years of his life, he moved between Naples, Malta, and Sicily. His paintings have been characterized by art critics as combining a realistic observation of the human state, both physical and emotional, with a dramatic use of lighting, which had a formative influence on Baroque painting.[3][4][5]

Caravaggio employed close physical observation with a dramatic use of chiaroscuro that came to be known as tenebrism. He made the technique a dominant stylistic element, transfixing subjects in bright shafts of light and darkening shadows. Caravaggio vividly expressed crucial moments and scenes, often featuring violent struggles, torture, and death. He worked rapidly with live models, preferring to forgo drawings and work directly onto the canvas. His inspiring effect on the new Baroque style that emerged from Mannerism was profound. His influence can be seen directly or indirectly in the work of Peter Paul Rubens, Jusepe de Ribera, Gian Lorenzo Bernini, Velázquez and Rembrandt[6]. Artists heavily under his influence were called the "Caravaggisti" (or "Caravagesques"), as well as tenebrists or *tenebrosi* ("shadowists").

Caravaggio trained as a painter in Milan before moving to Rome when he was in his twenties. He developed a considerable name as an artist and as a violent, touchy and provocative man. He killed Ranuccio Tommasoni in a brawl, which led to a death sentence for murder and forced him to flee to Naples. There he again established himself as one of the most prominent Italian painters of his generation. He travelled to Malta and on to Sicily in 1607 and pursued a papal pardon for his sentence. In 1609, he returned to Naples, where he was involved in a violent clash; his face was disfigured, and rumours of his death circulated. Questions about his mental state arose from his erratic and bizarre behavior. He died in 1610 under uncertain circumstances while on his way from Naples to Rome. Reports stated that he died of a fever, but suggestions have been made that he was murdered or that he died of lead poisoning.

Caravaggio's innovations inspired Baroque painting, but the latter incorporated the drama of his chiaroscuro without the psychological realism.[dubious – discuss] The style evolved and fashions changed, and Caravaggio fell out of favour. In the 20th century, interest in his work revived, and his importance to the development of Western art was reevaluated. The 20th-century art historian André Berne-Joffroy [fr] stated: "What begins in the work of Caravaggio is, quite simply, modern painting."[7]

Caravaggio

Chalk portrait of Caravaggio by Ottavio Leoni, c. 1621

Born	Michelangelo Merisi (or Amerighi) da Caravaggio 29 September 1571 Milan, Duchy of Milan[1]
Died	18 July 1610 (aged 38) Porto Ercole, State of the Presidi
Education	Simone Peterzano
Known for	Painting
Notable work	*The Beheading of St John the Baptist* *The Calling of Saint Matthew* *The Musicians* *The Cardsharps*
Patrons	Cardinal Francesco Maria del Monte Alof de Wignacourt
Signature	

Figure 3.1: Biographical article on the historical figure Caravaggio[9]

This mission is upheld by the five pillars [11] :

- Wikipedia is an online encyclopedia The first pillar of Wikipedia reflects the encyclopedic character of Wikipedia, which "combines many features of general and specialized encyclopedias, almanacs, and gazetteers" [11]. The goal is not to use Wikipedia for personal gain, such as advertising or vanity press but to create high quality publicly available digital information.

- Wikipedia has a neutral point of view
  The articles in Wikipedia should be written from a neutral point of view. This means that no advocacy or personal view of the editor of the Wikipedia page should influence the article. The goal is to represent the facts and document the majority point of view [11].

- Wikipedia is free content
  The articles on Wikipedia need to respect copyright laws and not plagiarize other sources. They are free to the public and can and should be edited and redistributed [11].

- Wikipedians should interact in a respectful and civil manner
  Wikipedians, the volunteers that contribute to the content of Wikipedia, should not

[11] Wikipedia: Five pillars — Wikipedia, https://en.wikipedia.org/w/index.php?title=Wikipedia:Five_pillars&oldid=1291543181, Accessed on: 2025-07-22

participate in personal attacks and edit wars, overriding the contributions of other editors they are in disagreement with [11].

- Wikipedia does not have firm rules
  Wikipedia's policies and guidelines on content and its interpretation can evolve over time [11].

In the context of this thesis, these pillars are significant, as they underline the strengths and weaknesses of Wikipedia. The multilingual character and public curation of Wikipedia make it an ideal data source for analyzing regional and language biases, while the absence of a centralized editorial body also explains why such biases might emerge in the first place.

3.2.1 Structure and Data Model

Wikipedia is one of the most extensive and maintained knowledge bases on the internet. Wikipedia's coverage spans a multitude of domains such as technology, geography, art and history and many more. The content is structured in a relational database schema managed by MediaWiki software, which supports the collaborative, multilingual and dynamic nature of Wikipedia [12].

Wikipedia uses a complex relational database structure that stores all components of a Wikipedia article, such as textual content, user metadata, editorial history and categorization logic. In order to do so, it relies on a combination of free-text content, semi-structured elements (such as info-boxes and categories), and metadata [12,13]. At the core of this schema are several tables:

- The page table contains metadata for each Wikipedia page, that contains data such as title, namespace (e.g., article, category or user page) and a unique identifier [14].
- The revision table stores the history of edits associated with each page, such as the time when the edit was made, the ID of the user that made the edit and the links to the edited content [15].
- The text table contains the Wikitext for each revision of the article. These revisions include points such as changes to the main text of the article, embedded templates, info-boxes, categories and references [16].

[12] MediaWiki — Wikipedia, https://en.wikipedia.org/w/index.php?title=MediaWiki&oldid=1301602769, Accessed on: 2025-07-22

[13] The Architecture of Open Source Applications (Volume 2): MediaWiki, https://aosabook.org/en/v2/mediawiki.html, Accessed on: 2025-07-22

[14] Page table — Wikipedia, https://en.wikipedia.org/w/index.php?title=Page_table&oldid=1284638020, Accessed on: 2025-08-27

[15] Manual:Revision table — MediaWiki, https://www.mediawiki.org/wiki/Manual:Revision_table, Accessed on: 2025-08-27

[16] Manual:Text table — MediaWiki, https://www.mediawiki.org/wiki/Manual:Text_table, Accessed on: 2025-08-27

- The user table logs registered users and their associated metadata, an example being the user ID and date of registration[17].
- The category links table connects the pages to the categories under which they are classified, allowing hierarchical and thematic navigation [18].

By using such a schema, Wikipedia is able to maintain full editorial control, track editorial contributions and support dynamic querying of its knowledge base. The content in these tables serves as the foundation for further structured data extraction.

Wikipedia also hosts semi-structured data such as info-boxes, which have structured metadata about entities in a consistent, template-based format. Although stored as raw wikitext, infoboxes follow a predictable key-value structure that can be parsed using template parsers [19].

For this thesis, a central aspect is the multilingual nature of Wikipedia, with each Wikipedia language edition independently curated by volunteer editors. This induces the differences that are present in both size and coverage, but also the emphasis in articles. By comparing the differences across language editions, measuring language and regional bias in art historical coverage is possible. The subsequent sections of this thesis will elaborate on the data retrieval, cleaning and transformation processes that allow the analysis of Wikipedia data using data mining techniques.

3.3 Wikidata

Wikidata is a free, collaborative and multilingual knowledge base that collects structured data to provide support for Wikipedia, Wikimedia Commons, other wikis of the Wikimedia movement and to anyone in the world [20]. It operates as the central structured repository for the Wikimedia ecosystem. Launched by Wikimedia in 2012, it serves as one of the largest public knowledge graphs with data on more than 100 million concepts contributed by over 560,000 editors [Mora-Cantallops et al., 2019]. Through the use of linked open data (LOD), it is supporting researchers, developers and institutions. Linked Open Data is data that is linked through predefined standards such as models and vocabularies and accessible under an open license[21] .

[17] Manual:User table — MediaWiki, https://www.mediawiki.org/wiki/Manual:User_table, Accessed on: 2025-08-27

[18] Manual:Categorylinks table — MediaWiki, https://www.mediawiki.org/wiki/Manual:Categorylinks_table, Accessed on: 2025-08-27

[19] Infobox — MediaWiki, https://www.mediawiki.org/wiki/Infobox, Accessed on: 2025-08-27

[20] Wikidata:Introduction — Wikidata, https://www.wikidata.org/wiki/Wikidata:Introduction, Accessed on: 2025-04-03

[21] Linked Open Data Overview, https://www.canada.ca/en/services/culture/history-heritage/museology-conservation/collections-management/linked-open-data/linked-open-data-overview.html, Accessed on: 2025-10-28

Wikipedia was developed and intended to be readable by humans, while Wikidata is intended to be processed by machines through the use of standardized schema. This difference enables large-scale data analysis and comparison across languages.

In the context of this thesis, Wikidata serves as an important database, providing standardized identifiers (QIDs)[2] for painters, places and events across different language editions of Wikipedia. This makes it possible to do a comparison of language and regional biases by linking equivalent concepts across languages through the use of QIDs. As such, Wikidata serves as a bridge between Wikipedia's narrative, language-specific representations and the neutral, comparable datasets required.

One of the central points of this thesis is the analysis of language and regional biases in Wikipedia's coverage of painters. Through the use of QIDs equivalent entities can be compared across language editions, thus mitigating inconsistencies caused by spelling variations or translation. The verifiability of Wikidata is improved by external identifiers (e.g., ULAN, VIAF, museum catalogues) [Vrandečić et al., 2023].

For the second research objective of evaluating visualization techniques for art historical data, Wikidata provides the standardized and structured data needed. Wikidata has the geographical identifiers (birthplaces, deathplaces) and dates of birth and death that that form the basis of the visualization techniques.

Wikidata is organized around several key principles that make it suitable for this work:

- Community: Wikidata is being edited collaboratively and maintained by a public community of volunteers[22].
- Sharing: The information is freely used and distributed under the Creative Commons CC-0 license.
- Data Model: content is exported in the Resource Description Framework (RDF) standard, formatted using its own data model, and accessible to machines via APIs and the Wikidata Query Service (SPARQL) [Odell et al., 2022].
- Multilinguality: Wikidata items support labels, descriptions and more in multiple languages, using language-independent identifiers (QIDs) that ensure comparability across different language editions of Wikipedia [23] .
- Verifiability: Wikidata uses other sources for verification through the use of references and opposing or disputed opinions may coexist [Vrandečić et al., 2023].
- Data Foundation: Wikidata serves as a data foundation for other Wikimedia projects, including providing picture tags for Wikimedia Commons, linking articles

[22] Wikidata/Notes/Requirements - Meta-Wiki, https://meta.wikimedia.org/wiki/Wikidata/Notes/Requirements, Accessed on: 2025-10-28

[23] Wikidata/Notes/Requirements - Meta-Wiki, https://meta.wikimedia.org/wiki/Wikidata/Notes/Requirements, Accessed on: 2025-10-28

on the same subject in different languages, and contributing data that is presented in Wikipedia articles [Odell et al., 2022].

- Identity provider: Wikidata ideas have consistent, language-neutral identifiers that are connected to other resources (such as social networks, archives and catalogues) by external identifiers[Vrandečić et al., 2023].

3.3.1 Structure and Data Model of Wikidata

Wikidata uses a structured statement-based model that is formed by entities that match RDF standard [Odell et al., 2022]. These RDF triplets match entities, also called items, properties and values [Odell et al., 2022]. They represent concepts from the real world such as people, places, events, or topics. Each of these items has a unique identifier (e.g., Q42 for Douglas Adams) and can include multilingual labels, descriptions and aliases[24] . These statements may also include additional context (qualifiers) and sources supporting the claim (references). These facilitate their use in the semantic web and linked data ecosystems [Mora-Cantallops et al., 2019] [Erxleben et al., 2014]. One of the main features of Wikidata is the option to have cross-lingual querying and support for content alignment across different Wikipedia language editions. This capability directly supports the detection of language and regional biases and enables their visualization in a consistent, structured framework.

3.4 Getty ULAN

The Getty ULAN database also known as the Getty Union List of Artist Names, is a structured database about artists [4]. It is curated by researchers and backed by the Getty Research Institute and includes names, biographies, related people and other metadata about artists, architects, firms, studios, museums, patrons and other people and groups involved in the creation and study of art and architecture [4]. The Getty Vocabularies use Linked Open Data (LOD). With LOD, the metadata about the objects, their creators, patrons, associated places, style, work type, iconography and other terminology concerning their description, history and scholarly research can be extracted and used [25] . The introduction and usage of LOD in the Getty ULAN archives is part of the initiative of the Getty Research Institute called Open Content Program, which has been introduced in 2003 [26] . To promote the study, instruction, and practice of art and art history, the Open Content Program makes high-resolution photos of works of public domain art from the Getty holdings freely available[26]. Additionally, the initiative also seeks to encourage

[24] Library Carpentry Wikidata: Underlying concepts of Wikidata, https://librarycarpentry.github.io/lc-wikidata/02-Wikidata_underlying_concepts.html, Accessed on: 2025-08-27

[25] Getty Vocabularies as LOD (Getty Research Institute), https://www.getty.edu/research/tools/vocabularies/lod/index.html#definition, Accessed on: 2025-04-01

[26] Open Access - The Metropolitan Museum of Art, https://www.metmuseum.org/about-the-met/policies-and-documents/open-access, Accessed on: 2025-10-25

and advance the open-access movement among museums and other cultural heritage institutions.

3.4.1 Structure and Data Model of the Getty Vocabularies

The Getty vocabularies contain terms, names and other information about people, places and concepts that are used to document information related to art and architecture [Harpring, 2010b]. There are multiple vocabularies under the project of Getty Vocabularies[27]:

- Art & Architecture Thesaurus (AAT)®
- Cultural Objects Name Authority (CONA)®
- Getty Iconography Authority (IA)™
- Getty Thesaurus of Geographic Names (TGN)®
- Union List of Artist Names (ULAN)®
- Categories for the Description of Works of Art (CDWA)

The Union List of Artist Names (ULAN)®, Getty Thesaurus of Geographic Names (TGN)® and Art & Architecture Thesaurus (AAT)® are curated by researchers and will be used for the purpose of this thesis. As this thesis will be using over 100 000 records, the Getty ULANs Linked Open Data (LOD) will be used as a point of reference as well. Each of the vocabularies covers different information that is interconnected using unique identifiers. Moreover, the thesauruses used in this thesis are:

- AAT – Art & Architecture Thesaurus - this vocabulary covers generic terms for objects, relationships, materials, styles, roles and techniques in art and architecture[27]
- TGN – Thesaurus of Geographic Names - this vocabulary covers the names of places (both current and historical) that are relevant to art/architecture [27]
- ULAN – Union List of Artist Names - this vocabulary covers the names and authority records for artists, architects, firms, patrons, as well as both individual and corporate bodies [27]

For access to these files, the Getty ULAN SPARQL Query Service can be used [28] or alternatively they can be directly downloaded from the Download Center, which contains the vocabularies in the format of a Relational Table UTF-8 or a XML UTF-8 file [29] .

[27] Getty Vocabularies, https://www.getty.edu/research/tools/vocabularies/index.html, Accessed on: 2025-11-10

[28] Getty Vocabularies LOD: Sample Queries, https://vocab.getty.edu/queries#Finding_Subjects, Accessed on: 2025-11-08

[29] Vocabularies Download Center (Getty Research Institute), https://www.getty.edu/research/tools/vocabularies/obtain/download.html, Accessed on: 2025-11-08

3.5 Data Extraction Techniques

3.5.1 SPARQL Queries

SPARQL (SPARQL Protocol and RDF Query Language) is a semantic query language used to retrieve data stored in the Resource Description Framework (RDF) format or for retrieving Linked Open Data (LOD)[30] [31].

The RDF is a graph-based data model, used for representing interconnected data [32] [33] . It represents data using triplets, each consisting of a subject, predicate and object [33].

- The subject being the resource described by the triple.
- The predicate describes the relationship between the subject and the object.
- The object is a resource that is related to the subject.

In order for data to be linked and visualized across different sources, a labeled and directed graph is formed where each node represents a resource and each edge denotes a relationship between them [34] . Through the use of Uniform Resource Identifiers (URIs) as globally unique identifiers, this allows datasets from different systems, languages or domains to be interlinked in a consistent manner [35] .

This structured format is why linked open-data projects like Wikidata use RDF [36] . Wikidata uses consistent referencing of entities such as historical figures, locations or institutions across multiple languages and domains.

SPARQL is a query language containing triplet patterns which was specifically designed for RDF datasets [Addlesee, 2019]. By specifying patterns to match against the RDF graph, it allows for data extraction. The queries consist of components such as SELECT, WHERE and OPTIONAL, collectively defining the search criteria [31]. This query language was used for the extraction of data from Wikidata, because Wikidata provides a SPARQL Endpoint.

[30] What Is SPARQL?, https://www.ontotext.com/knowledgehub/fundamentals/what-is-sparql/, Accessed on: 2025-10-28

[31] The SPARQL query language — GraphDB 10.8 documentation, https://graphdb.ontotext.com/documentation/10.8/sparql.html, Accessed on: 2025-04-03

[32] RDF - Semantic Web Standards, https://www.w3.org/RDF/, Accessed on: 2025-04-04

[33] What is RDF (Resource Description Framework)? — Search App Architecture, https://www.techtarget.com/searchapparchitecture/definition/Resource-Description-Framework-RDF, Accessed on: 2025-04-04

[34] XML RDF — W3Schools, https://www.w3schools.com/xml/xml_rdf.asp, Accessed on: 2025-04-04

[35] Library Linked Data Incubator Group Final Report, https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/, Accessed on: 2025-10-28

[36] Enhancing Wikidata Performance with RDFox: How to dissect the world's leading RDF database faster | 5 min read | Feb 12, 2024, https://www.oxfordsemantic.tech/blog/enhancing-wikidata-performance-with-rdfox-how-to-dissect-the-worlds-leading-rdf-database-faster, Accessed on: 2025-10-28

3.5.2 Text Mining

Text mining, also known as text data mining, is a process that transforms unstructured text into a structured format to gain new information and extract knowledge[37].

To be able to extract useful data, several key steps need to be performed, such as [Tan et al., 1999]:

- Text document gathering: gathering the data on which the text mining techniques are going to be applied. This is a foundational step and in the context of this thesis, Wikipedia and GettyULAN were used as the source.
- Data Collection and Text Preprocessing: this step includes several tasks such as:
  - Tokenization: Text is split into smaller elements such as words or phrases (also known as tokens)[38].
  - Stop-word removal: Removal of "filler" words that hold no value for the analysis (for example: "the", "and", "is") [39].
  - Stemming and Lemmatization: this are two techniques for preprocessing. They reduce words to their base form in order to find and join all possible variations (for example "studying" becomes "study")[40].
  - Syntactic Parsing: through the analysis of sentences and dividing them into smaller pieces such as nouns and verbs it helps machines understand the sentence structure [41].
- Extracting features: after preprocessing the data, various techniques can be applied in order to extract the data into features. This is done in preparation to serve as the basis for the next stage of analysis. Some examples of techniques that can be used are:
  - Classification and Clustering: Grouping documents or text segments together on the basis of shared characteristics or labeling them based on predefined categories [42].

[37] What Is Text Mining? | IBM, https://www.ibm.com/think/topics/text-mining, Accessed on: 2025-05-27

[38] What is Tokenization? Types, Use Cases, Implementation — DataCamp, https://www.datacamp.com/blog/what-is-tokenization, Accessed on: 2025-10-28

[39] All about Tokenization, Stop words, Stemming and Lemmatization in NLP — Medium, https://medium.com/@abhishekjainindore24/all-about-tokenization-stop-words-stemming-and-lemmatization-in-nlp-1620ffaf0f87, Accessed on: 2025-10-28

[40] Stemming and Lemmatization in Python, https://www.datacamp.com/tutorial/stemming-lemmatization-python, Accessed on: , 2025-10-28

[41] What is Syntactic Parsing? | Activeloop Glossary, https://www.activeloop.ai/resources/glossary/syntactic-parsing/, Accessed on: 2025-10-28

[42] Classification vs Clustering — GeeksforGeeks, https://www.geeksforgeeks.org/machine-learning/ml-classification-vs-clustering/, Accessed on: 2025-11-12

  - Sentiment Analysis: Identifying the emotional tone behind a piece of text through the use of machine learning, text analysis by classifying the tone as positive, negative or neutral [43].
  - Named Entity Recognition (NER): Identifying words and classifying them into predefined categories such as locations, peoples, names etc. [44].
  - Relation Extraction: Identifying the relationships between entities (example, artist and nationality)[45].

- Applying analytical techniques: after the structured features are extracted, analytical models can be used. In the case of this thesis, a dictionary-based rule-matching method was used to detect nationalities mentioned in Wikipedia articles.
- Interpreting the results: this step consists of the evaluation of the results in the context of the research objectives of this thesis. In this thesis, it involved the validation of the accuracy of nationalities that have been extracted and the integration of the results into the already existing dataset.

3.6 Criteria for Artist Inclusion (Time Periods, Geography, Metadata Completeness)

The data sources used for this thesis have large amounts of data, so in order to ensure that the scope of this research is analytically meaningful, a subset of that data was used. The scope of the chosen data is that the data focus on painters who were born between the 15th and 20th centuries. With respect to geography, all regions of the world were considered, without limitation to national or cultural origin. This time period was chosen to ensure consistency in metadata quality and completeness, as earlier records (e.g., prior to the 14th century) can be sparse or inconsistently annotated. Additionally, that time period captures key historical and stylistic periods in European and global art history, including the Renaissance, Baroque, Neoclassicism and Romanticism.

Metadata completeness was considered when selecting painters for the final analysis. Only entries with at least basic biographical information (consisting of name, date of birth, and a valid Wikidata QID) were considered. Since different databases were used, it was necessary to link and enable further extraction tasks from Wikipedia and Getty ULAN.

This subset of data provides a balance between data availability, practical feasibility for computational processing and historical breadth. It will be used as a base for analyzing how painters are represented across language editions of Wikipedia and for visualizing their cultural and geographical context through time.

[43] What Is Sentiment Analysis? | IBM, https://www.ibm.com/think/topics/sentiment-analysis, Accessed on: 2025-10-28

[44] What Is Named Entity Recognition? | IBM, https://www.ibm.com/think/topics/named-entity-recognition, Accessed on: 2025-10-28

[45] Identifying entities and their relations in text — Medium, https://nikhilsrihari-nik.medium.com/identifying-entities-and-their-relations-in-text-76efa8c18194, Accessed on: 2025-11-10

3.7 Data Cleaning and Integration

For a statistical and visual analysis of data, the reliability, completeness and accuracy of data are of crucial importance. In this thesis, the collection and integration of data involved structured and semi-structured data from three sources: Wikipedia (in six language editions: English, French, German, Italian, Spanish and Serbian), Wikidata and Getty ULAN. The process followed the CRISP-DM framework and involved multiple iterations of data cleaning and validation.

3.7.1 Data Collection

The data set was queried through Wikidata SPARQL, targeting all individuals categorized as:

- Instance of: human(wdt:P31 wd:Q5)
- Occupation: painter(wdt:P106 wd:Q1028181)
- Birth date: between 1400 and 1999

This query ensures that only individuals explicitly identified as humans and painters in Wikidata are included in the data collection.

In order to enable the language and regional bias analysis, the same SPARQL Query Structure was reused across six different Wikipedia language editions: English, German, French, Spanish, Italian and Serbian. The endpoint was only modified in the SPARQL query to point to the corresponding Wikipedia language edition. This enabled the collection of language-specific metadata and the comparison of how artists are represented across different language editions.

3.7.2 Data Cleaning

After obtaining the data from Wikipedia and Wikidata, there were missing and inconsistent data points including conflicting and missing nationality, duplicates due to alternate spelling and missing birth and death dates. In order to address these challenges, several different steps were performed:

- Deduplication using the Wikidata QIDs: duplicate entities were consolidated with merging.
- Normalization: in the Serbian Wikipedia the text is in Cyrillic, which needed normalization.
- Error Correction and detection of inconsistencies in the data was flagged and corrected, e.g. incorrect labeling.

3.7.3 Data Enrichment using Getty ULAN

One of the biggest issues with the dataset concerning this thesis was the absence of a structured nationality list for many historical figures. In order to fill in these gaps, painters were cross-referenced with Getty ULAN.

Through the use of the Getty ULAN LOD and the Wikidata QID, entries were matched and files were enriched with nationality data. As Getty ULAN has preferred and non-preferred nationality columns, all were taken into consideration for the data enrichment.

Scraping Wikipedia for Missing Nationality

In the cases where the nationality was missing in both Wikidata and Getty ULAN, the nationality of entities was scraped from the unstructured Wikipedia article corresponding to the entity taking into consideration the Wikipedia language edition.

The scraping process involved parsing the lead paragraph and info-box of each article and applying pattern matching while accounting for grammatical structures specific to each language edition.

The scraped nationalities were stored in a separate field and by sampling of the data manually verified in order to minimize the occurrence of false data due to errors in parsing.

3.7.4 Dataset Schema

After the cleaning and enrichment of the data for each Wikipedia language edition, a dataset was created with a unified schema containing the following columns:

- qid: Unique identifier from Wikidata for each artist.
- name: Full name of the artist as recorded in Wikidata.
- birth_date: Date of birth
- death_date: Date of death
- birth_place: Birth location as extracted from Wikidata.
- birth_coords: Geographical coordinates (latitude and longitude) of the birthplace.
- death_place: Death location
- death_coords: Geographical coordinates (latitude and longitude) of the death place.
- nationality: Nationality as listed in Wikidata (may be missing for some records).

- in_getty_ulan: Boolean flag which indicates whether an artist is listed in the Getty ULAN or not.
- getty_nationalities: List of nationalities in the Getty ULAN for the specific entity.
- yes_no_scrape: Boolean flag which shows whether the artist's nationality was extracted by scraping.
- scraped_nationality: Nationality obtained from unstructured Wikipedia text using scraping and NLP tools.

3.7.5 Metadata for Bias Analysis

For bias analysis, different Wikipedia language editions were used and Wikipedia articles were scraped to extract structural and bibliographic metadata such as: number of references, number of categories, number of sections and number of internal links. These attributes were chosen because they were used in prior studies as proxies for article completeness and quality [Lewoniewski et al., 2017b].

The metadata was retrieved by scraping the HTML structure of the corresponding Wikipedia articles across language editions. Because of the inconsistent formatting of Wikipedia articles, elements such as reference tags (<ref>), section headers (<h2>, <h3>) and internal link elements (<a href>) needed to be identified. In order to address this issue, a custom scraping logic had to be implemented.

Through the use of conditional extraction rules language-specific patterns were handled in order to uncover language-specific patterns (for example: localized reference section titles such as "Quellen" in German or "Références" in French).

The scraping process had to be iteratively refined to ensure consistency across language editions.

3.7.6 Tools and Libraries Used

The scraping and natural language processing tasks were implemented using a combination of libraries and tools:

- SPARQLWrapper: used for executing SPARQL queries remotely, in the case of this thesis using the through the Wikidata endpoint.
- BeautifulSoup: used for the parsing of HTML content from Wikipedia pages[46] .

[46] beautifulsoup4: Screen-scraping library, https://www.crummy.com/software/BeautifulSoup/bs4/, Accessed on: 2025-07-20

- re (Regular Expressions): used in order to extract words (e.g., nationalities) from free-text fields [47] .
- langdetect: used to verify and handle multilingual inputs during scraping[48] .
- pandas and NumPy: used for transforming and joining data from multiple sources[49].

3.8 Limitations

Several limitations that can affect the validity and scope of the results must be considered, despite best efforts employed for gathering and cleaning the data set.

- Limited ULAN-Wikidata Integration: There is limited integration between the Getty ULAN libraries and Wikidata. Many painters that are listed in Wikidata and Wikipedia lack a corresponding entry in GettyULAN. This made cross-referencing all data on validity with a high-quality data set such as the Getty ULAN difficult. Alack of entries in the Getty ULAN was most noticeable for lesser known painters.
- Inconsistent Wikipedia Article Structure: Wikipedia articles vary in structure across language editions but also inside of its own language edition. This made automated scraping difficult and prone to inconsistent and false results. To mitigate that, an iterative correction was made on the scraping logic with special consideration to the variations in structure.
- Multiple and Historical Nationalities: Because of the nature of the dataset, many inconsistencies in the nationalities of painters were found. As painters lived during the time of many political changes the nationalities changed through history making the mapping of historical nationalities to present-day nationalities prone to potential misrepresentations.
- Bias in the datasets: Since Wikidata and Wikipedia are open-source public datasets, their collaborative editorial nature can reflect systematic biases. Over-representation of national figures in local language versions, under-representation of women or uneven emphasis on certain periods are documented biases in such datasets. These biases are the object of analysis but also hinder the possibility to treat the datasets as neutral.

Despite the limitations present a robust dataset was created, through iterative validation and manual oversight.

[47] Regular expressions — JavaScript | MDN, https://developer.mozilla.org/en-US/docs/Web/JavaScript/Guide/Regular_expressions, Accessed on: 2025-11-12

[48] langdetect: Language detection library ported from Google's language-detection, https://github.com/Mimino666/langdetect, Accessed on: 2025-07-20

[49] NumPy, https://numpy.org/, Accessed on: 2025-07-20

CHAPTER 4

Methodology

In this thesis, the data analytics model CRISP-DM (Cross Industry Standard Process for Data Mining) [Schröer et al., 2021] was used to categorize and visually represent data from various sources. This methodology was chosen because it provides an overarching analytical structure and reduces the chances of failure due to poor planning [Wirth and Hipp, 2000]. Moreover, the independence of this model from industries and technology is one of its advantages [Schröer et al., 2021]. The independence comes from the fact that it is a high-level model that is not optimized for a specific task but provides structure in data mining research. Through the use of multiple overarching phases with non-industry-specific tasks, it is able to provide reproducibility, one of the fundamental principles of data science [Schwalbe, 2016]. In the Figure 4.1 and Figure 4.2 an overview of the various phases and their outputs is given.

CRISP-DM is particularly suited for projects involving heterogeneous data sources and iterative exploration, as is the case here with the integration of Wikipedia, Wikidata and GettyULAN data. The methodology will be applied across two major aspects of the analysis:

- The analysis of language and regional bias, where the editorial and structural disparities in open-source databases such as Wikipedia are analyzed.

- Visualization evaluation, in which the interpretability, clarity and effectiveness of three different visualization methods are analyzed through a user study. The visualization methods in question are heatmaps[1] , flow maps[Phan et al., 2005] and Sankey diagrams[6]. The dataset used is historical data that was extracted from Wikipedia.

[1] A Complete Guide to Heatmaps, https://www.atlassian.com/data/charts/heatmap-complete-guide, Accessed on: 2025-07-17

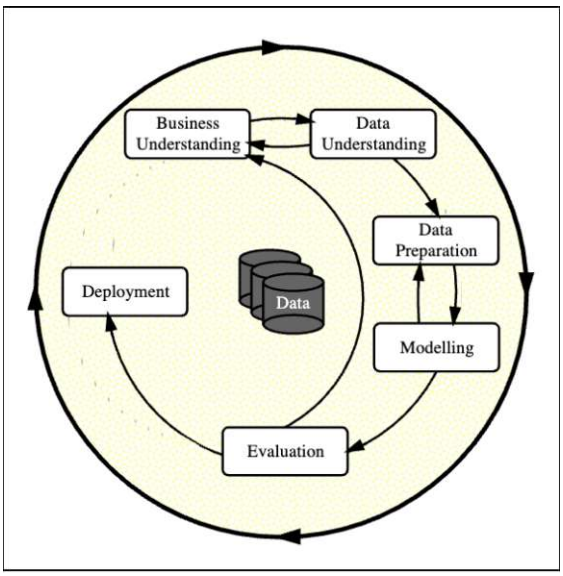


Figure 4.1: Phases of the CRISP-DM Model [Wirth and Hipp, 2000]

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	*Data Set* *Data Set Description* **Select Data** *Rationale for Inclusion / Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data*	**Select Modeling Technique** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Description* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

Figure 4.2: Overview of the CRISP-DM phases and their outputs [Wirth and Hipp, 2000]

This chapter presents the CRISP-DM phases and describes how each is applied in this research.

- Business Understanding - the first phase of the CRISP-DM process serves to understand the project, its overarching goal, and scope [Wirth and Hipp, 2000]. Concerning this thesis, there are two major aspects.

  The first core aspect is the analysis of regional and language bias that is present in the data of painters across six Wikipedia language editions: English, German, French, Spanish, Italian and Serbian. This entails an investigation of how painters are represented quantitatively and qualitatively. Different Key Performance Indicators (KPIs) are used to assess the disparities present in the editorial practices and data. Through the use of article metadata, such as the number of references, sections, internal links, categories, and the number of painters present in different Wikipedia language editions, the thesis aims to reveal patterns of over- and under-representation of historical figures that reflect editorial bias and limitations in publicly curated databases such as Wikipedia.

  The second core aspect is the evaluation of visualization techniques as tools for explaining historical data. For this, three different types of visualizations were used: heatmaps, flow maps, and Sankey diagrams. They were developed to visualize the migrations of painters between their places of birth and death. A user study is then done in which participants evaluate the visualizations based on interpretability, clarity and usability. The goal is to determine which visualizations are most effective for communicating historical data to both expert and non-expert audiences.

- Data Understanding - in this phase, the collection of data and understanding of the scope and nature is done [Wirth and Hipp, 2000]. This includes data collection from Wikipedia (in six language editions: English, German, French, Spanish, Italian, and Serbian), Wikidata and the Getty Union List of Artist Names (GettyULAN) through the use of text mining and the Wikidata Query Service. These sources provide both structured and semi-structured metadata on painters.

  In order to ensure analytical depth and the feasibility of the data, the scope of the collected data is painters born between the 15th and 20th centuries. This time period was selected to balance metadata completeness with historical importance, as earlier records often lack consistency.

  The main attributes that are used are: the name of the artist, date of birth, date of death, place of birth and place of death. Additional metadata, such as the number of references, sections, categories and internal links, was retrieved to support the bias analysis. These elements were collected using a combination of Wikidata SPARQL queries, text mining and HTML scraping techniques. In the Chapter 3, "Data Sources and Collection," a more detailed overview of this phase is provided.

- Data Preparation - this is a crucial step in the data analysis process. This phase covers multiple iterations of data cleaning, selection of attributes, construction of new attributes, as well as formatting of the dataset[Wirth and Hipp, 2000].

  In this thesis, special consideration was taken to avoid discrepancies and overlapping data, as the data is being taken from different sources. Handling of the data, including handling missing or conflicting values, normalizing place names and nationalities across languages, and deduplicating entries using Wikidata QIDs, were all crucial steps in this phase. One of the challenges was the multilingual nature of the dataset, an example being Serbian Wikipedia's use of Cyrillic text.

  As one of the key attributes relevant for this thesis is the nationality of painters, extra steps were taken to ensure consistency. One of the key components of this step was the enrichment of nationality data from GettyULAN, a researcher-curated database. For entities where the nationality data was missing or was ambiguous, additional scraping techniques were applied to extract it from Wikipedia's lead paragraphs, language-specific patterns.

  Even though multiple steps have been taken to ensure the quality of the data, there are limitations that remain: the variable quality of the articles present in Wikipedia, limited GettyULAN coverage, as well as inherent systematic biases present in Wikipedia and Wikidata just to name a few.

- Modeling - in this phase, different modeling techniques and algorithms are applied in an iterative manner [Wirth and Hipp, 2000]. Between the iterations, parameters are modified in order to get appropriate values. In this thesis, the modeling was done according to two different strands of the research: for the exploration of and measurement of language and regional bias, and for the assessment of different visualization techniques.

  For the bias analysis, the modeling process consisted of the implementation of different Key Performance Indicators (KPIs) that captured the bias present across different Wikipedia language editions. The KPIs that were chosen were: the Local Hero Score (LHS), Language Match Rate(LMR) and the Getty ULAN coverage ratio. KPIs that were indicators of article quality were the number of references, internal links, categories, and sections. Each of these KPIs was chosen based on the relevance for the thesis and applied across six language editions of Wikipedia: English, German, French, Spanish, Italian and Serbian. In order to show the disparities in the data these KPIs were aggregated across nationality groups and compared.

  For the visualization assessment part, three different visualization techniques were developed and iteratively refined in order to model the data. For this thesis, heatmaps, flow maps and Sankey diagrams were chosen based on their ability to show complex data. Each visualization was developed and iteratively refined through different design variations, such as map backgrounds, marker sizes, flow sizes and color variations. These variants were compared internally through early

user feedback and further refined on the aspect of readability before being finalized for the user study.

The visualizations are the foundation of the user study, in which participants evaluate each visualization on clarity, usability and effectiveness in communicating the data.

- Evaluation - After several models that produce reliable results are made, this phase evaluates how well the outputs align with the goals that were set in the business understanding phase [Wirth and Hipp, 2000]. In the context of this thesis, the evaluation is based on the detection and analysis of regional and language bias and the assessment of different visualization techniques through a user study.

  For the analysis of bias, the evaluation consists of interpreting the values computed from the KPIs and determining if they reveal disparities across the six Wikipedia language editions. For a satisfactory result, a specific threshold is not defined, using instead a comparative insight, such as language edition having lower reference counts or disproportionately featuring painters from their nationality group. These findings are expected to validate the hypothesis that bias is present and manifests across Wikipedia different language editions.

  For the assessment of different visualization methods, the evaluation is based on user feedback from a user study. The user study was conducted, participants interacted with three types of visualizations: heatmaps, flow maps and Sankey diagrams. The participants were asked to rate them on clarity, readability and interpretive usefulness. The study includes both Likert-scale ratings, task-based accuracy questions, and open-ended questions. The responses from the user study were used in order to evaluate the effectiveness of each visualization in communicating historical data. A successful result would be identifying the visualization technique that shows the highest user satisfaction and the lowest mistake rate.

  The evaluation phase additionally serves as a reflection. In the case where the KPIs or user study produce an inconsistent result or are not aligned with the goals decided in the business understanding phase, the process needs to be repeated and a revisit of the business understanding and modeling phase needs to be done. This feedback loop ensures that the research remains grounded, iterative, and adaptive.

- Deployment - this is the final phase of the CRISP-DM model [Wirth and Hipp, 2000]. In the context of this thesis, a well-documented report in form of a master thesis is created.

  The thesis is a structured output of all the phases and presents the computed bias indicators, the user study and the evaluated visualization methods.

  The deployment phase includes the reflection on limitations, documentation of methodological choices, and the formulation of potential future research directions. This involves the proposition on how the bias and visualization evaluation framework can be adapted and potentially scaled. Additionally, it raises potential questions for future research.

CHAPTER 5

Data Analysis and Visualization

5.1 Language and Regional Bias Across Wikipedia Language Editions

In this section, several key performance indicators (KPIs) are going to be discussed in order to highlight the potential language and regional bias that is present across different Wikipedia language editions.

Wikipedia is a free online encyclopedia built on the principle of open-source knowledge generation[1]. Contributors from all over the world with diverse linguistic and cultural backgrounds who work voluntarily can write and edit Wikipedia articles. The open concept of Wikipedia has managed to generate a vast amount of articles and to solidify Wikipedia as the largest online encyclopedia, but this has also introduced inconsistencies in the data, as well as potential bias.

To investigate the potential bias, in this thesis we will focus on painters recorded in Wikidata who were born between the 14th and 20th centuries. We will then examine how these painters are represented across six Wikipedia language editions. These languages were chosen because they are some of the biggest language editions of Wikipedia, giving a wide range of data. The Serbian language edition was chosen because it is a relatively small language edition to showcase the disparity between the data.

One of the crucial parameters in this study is nationality, as it has a strong impact on the linguistic and regional aspects of this study. To account for this, we categorize artists into six nationality groups: English, German, French, Italian, Spanish, and Serbian. This grouping enables us to explore how artists are represented not only in their "native" language editions but also across foreign ones.

To measure potential bias, we use a range of different KPIs that represent the quality of the article, as well as the quantitative representation of the artists.

[1] Wikipedia | Definition, Encyclopedia, History, & Facts — Britannica, https://www.britannica.com/topic/Wikipedia, Accessed on: 2025-04-05

Through the analysis of these KPIs across Wikipedia language editions and nationality groups, the aim is to assess the visibility and structural consistency of articles about painter biographies. By using such a comparative approach, we aim to expose patterns of bias. An example would be whether some language editions systematically provide less detail for foreign artists, or whether certain nationality groups receive disproportionately sparse or rich coverage regardless of language. These findings contribute to the broader conversation on cultural heritage and representation and the completeness of multilingual datasets such as Wikipedia.

5.1.1 Nationality Groups

The data set we are examining is of an art historical nature. As one of the more important points for consideration in the search for language and regional bias is the cultural heritage, special consideration was given to the nationality of the painters. As we are examining painters that were born between the 14th and 20th century, it is crucial to also consider the geopolitical history.

In the dataset that we are analyzing, the nationality is not translated and mapped to the current political landscape and as such, can not be overtaken as-is. Special consideration was also given to include older geopolitical entities. An example being the inclusion of entities such as "Heiliges Römisches Reich Deutscher Nation", which is considered in the scope of this thesis as part of the "German nationality group". This is one of the limitations of this research, as the nationality of the painters could be contested; nevertheless, special consideration was given to capture as complete an image as possible.

By capturing both contemporary and historical language-nation associations, the nationality groups become a robust indicator of how language editions of Wikipedia reflect or neglect their associated cultural heritage.

The nationality groups are as follows:

- English: United Kingdom, England, Scotland, Wales, Northern Ireland, Kingdom of Great Britain, British, English, United States, USA, American, Canada, Canadian, Australia, Australian, New Zealand, New Zealander, Ireland, Irish, South Africa, South African, Jamaica, Jamaican, Trinidad and Tobago, Trinidadian, Tobagonian, Barbados, Barbadian, The Bahamas, Bahamian, Belize, Belizean, Saint Kitts and Nevis, Kittitian, Nevisian, Antigua and Barbuda, Antiguan, Barbudan, Saint Lucia, Saint Lucian, Saint Vincent and the Grenadines, Vincentian, Grenada, Grenadian, Dominica, Dominican (Dominica), British Empire, British Raj, British Ceylon, British North America, British rule in Myanmar, British Hong Kong, British Cyprus, Colony of New South Wales, Colony of New Zealand, Irish Free State, Irish Republic, Scottish, Welsh, Guyana, Guyanese, Montserratian, Fiji, Sierra Leone, Nigeria, Kenya, Zimbabwe, Zambia, Botswana, Guernsey, Jersey, Isle of Man, Northern Irish, Ontarian, Puerto Rican, Hawaiian

- German: Deutschland, Deutsches Kaiserreich, Deutsches Reich, Deutsche Demokratische Republik, Deutscher Bund, Heiliges Römisches Reich Deutscher Nation, Königreich Preußen, Königreich Sachsen, Königreich Bayern, Bayern, Freistaat Preußen,Bremen, Weimarer Republik, Königreich Württemberg, Bundesrepublik Deutschland bis 1990, Großherzogtum Baden, Böhmen, Germany, German, Deutsch, Austrian, Österreich, Austria, Switzerland, Swiss, Liechtenstein, Liechtensteiner, Luxembourg, Luxembourgish, Luxembourger, South Tyrol, South Tyrolean, Bohemia, Sudeten German, East Prussia, West Prussia, Anhalt, Austro-Hungarian, Bavarian, Brandenburg-Preußen, Deutschordensstaat, Lübeck, Habsburgermonarchie, Hamburg, Preußen, Reichsstadt Nürnberg, Reichsstadt Strassburg, Republik Baden, Sachsen, Schleswig-Holstein, Württemberg, West German

- French: royaume de France, France, Troisième République française, duché de Bourgogne, monarchie constitutionnelle française, Première République française, Premier Empire français, duché de Savoie, Français, Belgium, Belgian, Switzerland, Swiss, Canada, Canadian, Quebec, Québécois, Monaco, Monegasque, Luxembourg, Luxembourgish, Luxembourg, Deuxième République, Monégasque, Belgique, Belge, Suisse, Luxembourgeois, Québec, Canadien, Parisian, Normandie, Alsatian, Languedocian, Lorraine, Provençal, Royaume-Uni, République française, République florentine, Protectorat français au Maroc, Protectorat français d'Annam, Protectorat français de Tunisie, Établissements français de l'Inde, République libanaise sous mandat français, État du Grand Liban

- Italian: Abruzzese, Bergamo, Bologna, Bolognese, Calabrian, Cremona, Ducato di Ferrara, Ducato di Firenze, Ducato di Mantova, Ducato di Milano, Ducato di Modena e Reggio, Ducato di Parma e Piacenza, Ducato di Urbino, Ferrarese, Firenze, Florentine, Friulian, Genoese, Granducato di Toscana, Italia, Italian, Italiano, Ligurian, Lucchese, Mantuan, Marchesato di Mantova, Milanese, Modenese, Paduan, Piedmontese, Piemonte, Principato vescovile di Trento, Regno Lombardo-Veneto, Regno d'Italia, Regno delle Due Sicilie, Regno di Napoli, Regno di Sardegna, Regno di Sicilia, Repubblica Cisalpina, Repubblica Italiana, Repubblica Ligure, Repubblica Transpadana, Repubblica di Firenze, Repubblica di Genova, Repubblica di Lucca, Repubblica di Siena, Repubblica di Venezia, Repubblica genovese, San Marino, Sicilian, Sienese, Signoria di Bologna, Signoria di Correggio, Signoria di Perugia, Stato Pontificio, Territorio Libero di Trieste, Torinese, Tuscan, Umbrian, Venetan, Venetian, Venezia, provincia di Bologna, storia di Genova, Italy, Italian

- Spanish: Spain, Spanish, Mexico, Mexican, Argentina, Argentine, Argentinian, Colombia, Colombian, Chile, Chilean, Peru, Peruvian, Venezuela, Venezuelan, Ecuador, Ecuadorian, Guatemala, Guatemalan, Cuba, Cuban, Bolivia, Bolivian, Dominican Republic, Dominican, Honduras, Honduran, Paraguay, Paraguayan, El Salvador, Salvadoran, Nicaragua, Nicaraguan, Costa Rica, Costa Rican, Uruguay,

Uruguayan, Panama, Panamanian, Equatorial Guinea, Equatoguinean, Español, España, Corona de Aragón, Corona de Castilla, Países Bajos Españoles, Virreinato de Nueva España, Reino de España (1700-1873), Imperio español, Monarquía Hispánica, Castilian, Castilla la Nueva, Catalan, Galician, Gallego, Andalusian, Reino de Granada, Reino de Castilla, Reino de Valencia, Reino de Aragón, Reino de Mallorca, Virreinato del Perú, Argentino, Chileno, Cubano, República Dominicana, México, Mexicano, Panamá, Perú

- Serbian: Srbija, Serbian-Montenegrin, Serbia, Србија, Југославија, Краљевина Југославија, Краљевина Срба, Хрвата и Словенаца, Социјалистичка Федеративна Република Југославија, Држава Словенаца, Хрвата и Срба, Краљевина Србија, Кнежевина Србија, Српски, Serbian, Yugoslav, Савезна Република Југославија, Србија и Црна Гора, Демократска Федеративна Југославија

5.2 Representation

The following KPIs aim to showcase the imbalance in the representation of painters based on their nationality and the specific Wikipedia language edition.

5.2.1 Local Hero Score (LHS)

$$\text{LHS} = \frac{\text{Number of artists present in their native language edition}}{\text{Total number of artists from that nationality in the dataset}}$$

The Local Hero Score (LHS) is a metric that illustrates the coverage of native artists in their language compared to other language editions of Wikipedia [Eom and Shepelyansky, 2013]. Native artists are, for the purpose of this thesis, defined as artists who come from a country that speaks the language of the Wikipedia edition as a first language, or that has been considered through history as part of the native speaking country. The LHS is indicative of how well the Wikipedia editors from that language base documented their own cultural figures in the field of visual arts. A high score is indicative of editorial prioritization of local content, while low scores could reveal prioritization of more internationally known figures [Eom and Shepelyansky, 2013]. These scores reveal patterns of bias and editorial capacity. A low Local Hero Score could indicate underrepresentation, due to possibly fewer contributors and less digitized national heritage. On the other side, high scores can be indicative of a greater number of contributors and a more digitized national heritage.

In the context of this thesis, six different Wikipedia language editions were used: English, Spanish, French, German, Italian, and Serbian. In the Table 5.1. we can see an overall high score for the languages, the lowest being 84.12% for the Italian edition and the highest 96.81% for the English edition.

The English Wikipedia has the largest editorial base and the greatest number of articles compared to any other Wikipedia language, and that is reflected in the highest score with 96.81. This highlights the advantages of larger editorial communities for documenting

Language	In Native Wikipedia	Total Artists	Local Hero Score (%)
English	17,817	18,404	96.81
Spanish	6,146	6,813	90.21
German	19,152	20,312	94.29
French	15,004	17,076	87.87
Serbian	707	790	89.49
Italian	6,793	8,075	84.12

Table 5.1: Local Hero Score across Wikipedia Language Editions

knowledge. In contrast, the Italian Wikipedia, which has a rich history in the field of art, has the lowest score with just 84.12%. This imbalance could point to the fact that Italian painters are underrepresented in their own language edition, but have a high level of representation in other Wikipedia language editions. Similarly, the French Wikipedia has a score of 87.87% , underperforming in contrast to the German, Spanish and Serbian Wikipedias. This could also indicate that, while French painters are covered, a significant amount of coverage is spread across different Wikipedia language editions due to their international influence.

From the perspective of bias, these results show two contrasting tendencies. A high LHS value indicates strong representation in their community, while a low LHS value indicates reliance on foreign editorial communities. This shows that the ability to narrate a country's artistic heritage on open platforms like Wikipedia can shift away from its own language community.

By comparison, across language editions, these tendencies become more pronounced. German (94.29%) and Spanish (90.21%) Wikipedia's manage strong local representation despite smaller global editorial bases, reflecting the presence of active and committed editorial communities. This is particularly seen in the case of the Serbian Wikipedia, where, despite its small size, it achieves a high score of 89.49%. This suggests that even small editorial communities can prioritize the documentation of local historical figures.

The Local Hero Score (LHS) thus highlights the uneven editorial commitment across language editions for how well communities represent their own artists, while the Language Match Rate (LMR) evaluates how well the painters are represented across different language editions. This can reveal further layers of regional and language bias.

5.2.2 Language Match Rate (LMR)

$$\text{NMR} = \frac{\text{Number of articles where artist nationality matches the language edition}}{\text{Total number of artist articles in that language edition}}$$

The Language Match Rate (LMR) is a metric that illustrates what percentage of painters in a given language edition are from the country associated with that language. It indicates the degree to which national self-representation is present within Wikipedia. While the

Language	No of Painters	No of artists speaking the language	LMR (%)
English	46455	18011	38.77
German	35252	19165	54.37
French	30724	15590	50.74
Spanish	15678	6148	39.21
Italian	14413	6793	47.13
Serbian	1863	707	37.95

Table 5.2: Language Match Rate

Local Hero Score (LHS) examined how thoroughly each language documents its own artists, the LMR is asking a different question:

What share of a language edition's coverage is dedicated to its own nationals?

A high LMR suggests that the Wikipedia language edition tends to focus more on painters associated with the language, while a low LMR suggests that the Wikipedia language edition has a broader coverage of international figures.

This metric can help show patterns of editorial bias based on the background of the historical figures. A Wikipedia language edition that has a high LMR may suggest editorial bias towards its own cultural heritage. This can come at the cost of a more balanced representation, favoring local figures over global representation.

From Table 5.3, we can observe that German Wikipedia has the highest LMR at 54.37 %, meaning that more than half of the artist biographies featured are of German-speaking origin. The French and Italian Wikipedia editions follow closely with LMRs of 50.74 % and 47.13 %, this implies a strong editorial focus on historical figures that are aligned with the language of the Wikipedia edition. This suggests active and engaged editorial communities that view cultural preservation as one of its central tasks.

In contrast to that, the English Wikipedia shows a lower LMR of 38.77 %. A possible reason for that can be the widespread use of English as a second language. This makes the English Wikipedia edition the default entry point for editors and readers interested in non-local subjects, which in turn can dilute national focus.

The Spanish and Serbian language editions have a lower LMR, with 39.21% and 37.95%, respectively. The Serbian language edition results suggest that even smaller editorial communities can heavily focus on covering their own historical figures, but may lack the editorial capacity to raise the coverage across a certain threshold. The Spanish edition has a relatively low score, taking into consideration the large editorial community. Taking into account the widespread use of the Spanish language and dispersed editorial efforts across Latin America and Spain, it is fair to assume that the Spanish Wikipedia community might be sharing similar experiences as the English Wikipedia users.

To conclude, the LMR helps highlight gaps in coverage and content and editorial biases within Wikipedia's decentralized structure.

5.2.3 Artist Coverage Across Languages in Wikipedia

The goal of this metric is to show the number of artists in their respective language editions and to show if they are underrepresented or overrepresented in different Wikipedia language editions. While the Local Hero Score (LHS) captures the completeness of national coverage and the Language Match Rate (LMR) reflects the share of local versus foreign painters, this metric focuses instead on how widely painters appear across different Wikipedia language editions.

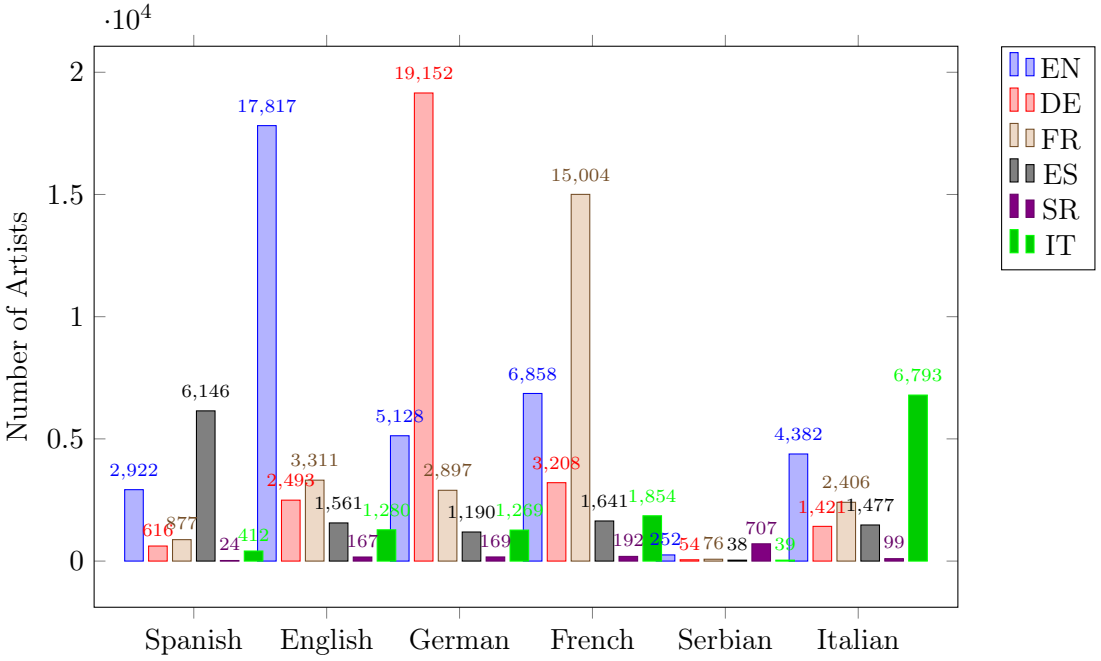


Figure 5.1: Distribution of painters across six Wikipedia language editions by nationality group.

The Figure 5.1 illustrates the presence of painters based on their nationality across six Wikipedia language editions (English, German, French, Spanish, Italian and Serbian) and reveals notable differences. Each nationality is most prominently featured in their respective native language Wikipedia, which reflects the phenomenon of the local hero effect [Eom and Shepelyansky, 2013] [Callahan and Herring, 2011]. Most pronounced is the disparity in the Serbian Wikipedia edition, where in the Serbian language edition 707 artists are listed, while in others, only a fraction: 24, 167, 169, 192 and 99 painters respectively. This illustrates how smaller language editions can have a high level of self-representation, but are underrepresented in other language editions.

The English Wikipedia has, in contrast, the broadest representation of painters that do not originate from a primarily English-speaking country. In the English Wikipedia version, French and German painters appear respectively 3311 times and 2493 times. This indicates the dominance of the English Wikipedia version as a global knowledge platform.

The German and French Wikipedia versions show a similar pattern. While in each of their respective language editions historical figures are well represented, there is still a high representation of other groups, although much smaller. The German Wikipedia version is leading with 19152 figures from their language group, with the next biggest group being the figures from the English nationality group with 5128. This suggests a strong editorial focus on the representation of local historical figures.

The Spanish and Italian Wikipedia editions have a similar count for their local figures, with respectively 6146 and 6793 entries. While they have a strong representation of local figures, international figures are much less represented. This is more apparent in the Spanish version, where German and French painters, for example, have only 616 and 877 entries. These patterns indicate that these editorial communities, while active in the representation of their own historical figures, often lack the international scope.

The data reveals a clear pattern of regional representation bias across Wikipedia language editions. Each language edition heavily features painters whose nationalities correspond with the linguistic sphere of that edition. As observed in the figure 5.1, the German Wikipedia has a dominant count of German painters with 19152, while the French Wikipedia leads in coverage of French-speaking painters with 17817.

The strong representation of painters in their own language edition is directly in contrast to the representation of painters in other language editions where their numbers are significantly lower. An example of that being the German Wikipedia features 19152 German-speaking painters, while the presence of German painters in other Wikipedia language editions such as the English and Spanish Wikipedia is much lower with 2493 and 616, respectively. This can also be seen for Serbian artists who are prominently represented in the Serbian Wikipedia with 707, but barely appear in others. This shows how smaller or less globally dominant language communities may experience lower visibility.

Additionally, the data implies that visibility is highly skewed in favor of dominant global languages. The English-speaking painters are present in large numbers across all language editions, for example, German with 5128, French with 6858, and Italian with 4382. In contrast, French and Italian artists receive relatively limited attention in the English Wikipedia, despite the historical and artistic significance of painters from that background.

To conclude, these results reveal patterns of bias across different language editions. Wikipedia editions with larger editorial communities often succeed in documenting their historical figures while still managing to represent foreign painters, although in much smaller numbers. On the other hand, smaller editorial communities such as Serbian, while actively documenting their own historical figures, are relatively underrepresented in other Wikipedia language editions.

5.2.4 Representation of painters in Wikidata and GettyUlan

To understand how comprehensively painters are represented across both crowd-sourced, Wikipedia and Wikidata, and institutional data infrastructures such as GettyULAN, we measure the overlap between these systems.

Language	Number of painters in the dataset	Number in GettyULAN	Coverage Ratio
English	46289	20876	45
German	35022	16189	46.2
French	30575	16276	53.2
Spanish	15589	7451	47.8
Italian	14413	8264	57.34
Serbian	1820	809	44.4

Table 5.3: Connection of GettyULAN and Wikidata

The Table 5.3 illustrates the degree to which painters listed in various Wikipedia language editions are also represented in the Getty Union List of Artist Names (ULAN). The Coverage Ratio highlights the percentage of Wikipedia painters matched to GettyULAN entries. This was possible by using the QID, the Wikidata unique item ID to match to the corresponding ULAN IDs, which are unique identifiers in the Getty Union List of Artist Names (ULAN).

The highest alignment can be seen in Italian and French editions (57.34 % and 53.2%, respectively), which suggests a relatively strong alignment between the structured cultural heritage maintained in GettyULAN and the editorial focus of these Wikipedia editions. Significantly lower coverage can be seen in the Serbian (44.4%) and English (45%). For the Serbian Wikipedia a reason could be that a vast majority of figures covered are from the Serbian nationality group which is not internationally recognized. For the English Wikipedia a possible reason could be that as it has a vast coverage of artists again not all of them are internationally recognized and as such not available in the Getty ULAN.

Such an uneven and generaly low coverage with none exceeding 58% can indicate how artists are connected and documented in different databases. The highest coverage ratio for Italian and French editions could come from the fact that Italian and French artists are in the archival systems. In comparison, the lower coverage ratio for Spanish and Serbian artists can reflect the underrepresentation in ULAN. While the Getty ULAN and Wikipedia and Wikidata are connected through clear identifiers, there is room for inconsistencies in linking practices. For example, many local or regional painters might be prominent in a national Wikipedia edition but not yet recognized in global authority files like GettyULAN.

There are two different types of bias present here: first, a bias where certain cultures and nationalities are better represented in historical documentation, and second, an infrastructural bias, where language editions may vary in their integration with external authority data like GettyULAN. These biases limit the discoverability and interoperability of knowledge across platforms, reinforcing systemic imbalances in the cultural memory embedded in linked data ecosystems.

5.3 Evaluating Representation through Wikipedia Quality Indicators

In order to assess the potential bias present across different Wikipedia language editions, a look into various quantifiable metrics is done in this thesis, such as: number of references, number of internal links, number of sections and the number of categories. These metrics serve as quantifiable indicators of the editorial attention given to articles. While the quality of articles in this thesis is not going to be assessed through qualitative measures, there is research that points to the fact that metrics such as reference counts, number of links, number of categories and more can be connected to the quality of an article. [Lewoniewski et al., 2017b]

By grouping painter biographies according to nationality and comparing them across Wikipedia language editions, this thesis aims to identify disparities that reflect bias. These indicators alone do not define the quality of an article, but combined, they give an overview of structural inequality in Wikipedia. The following sections delve into each indicator in more detail, as supported by figures.

The results are presented in a table format and each indicator is summarized according to nationality, count, mean, median, standard deviation, minimum, and maximum. The nationality represents the nationality group that the painter belongs to, and the count is the number of painters of that nationality that appear in the specific Wikipedia language edition. The min and max represent the minimum and maximum count per article. The mean, median, and standard deviation are important indicators as they together create a clear picture of the distribution of the specific indicator.

The mean serves to show the overall average; for example, outliers with a very high or extremely low reference count can skew the value[2] . The mean is sensitive to extreme values, while the median is less sensitive and represents the middle value in a dataset. In order to indicate the spread of variability in the data, the standard deviation is used. A high value suggests strong inconsistencies between articles within a group, while a low value indicates a more uniform editorial effort[2]. By combining these metrics, the general tendencies and disparities in article quality between groups and language editions are shown.

5.3.1 References across language editions

One of the main metrics that show the reliability and verifiability of Wikipedia articles is the number and quality of references they contain[Lewoniewski et al., 2017a]. Previous research on the evaluation of Wikipedia in multiple languages has shown that metrics, such as the number of references, play an important role in determining the quality of articles. [Lewoniewski et al., 2017b] In order to verify information and ensure that the

[2]13.1: Basic Statistics – Mean, Median, Average, Standard Deviation, Z-Scores, and P-Value, https://eng.libretexts.org/Bookshelves/Industrial_and_Systems_Engineering/Chemical_Process_Dynamics_and_Controls_(Woolf)/13%3A_Statistics_and_Probability_Background/13.01%3A_Basic_statistics-_mean_median_average_standard_deviation_z-scores_and_p-value, Accessed on: 2025-10-26

content presented is based on facts and not personal opinions, references are used. This allows readers to trace the claims in articles to verified sources. Thus, the number of references in Wikipedia articles can serve as a potential indicator of the quality and correctness of an article in Wikipedia. [Lewoniewski, 2023]

In the context of this thesis, we delve into the representation of artists in multilingual data sources and the potential bias linked to the background of painters and the source in which it is presented. A disparity in references across languages can lead to different levels of accuracy of information and amounts of data presented across multiple Wikipedia language editions.

A disparity in the number of references also affects the way articles are featured and ranked in search engine results [Vincent and Hecht, 2021] [Nicholson et al., 2021]. A lack of references can thus further lead to the reinforcement of bias in data in less significantly represented painters and language editions.

This thesis takes the distribution of references as a potential marker of the quality and representation of the article. The number of references across Wikipedia language editions is compared with attention given to how artists from that nationality group are represented in the corresponding Wikipedia language edition. In the following tables, the distribution of references is shown across Wikipedia language editions and analyzed based on the nationality groups of the painters represented.

References in the English Wikipedia edition

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
English	17816	11.30	6.0	18.30	0	461
French	6861	7.45	4.0	12.81	0	206
German	5132	6.18	3.0	11.99	0	334
Italian	4385	4.42	2.0	8.74	0	230
Spanish	2922	6.82	3.0	12.91	0	231
Serbian	252	6.42	4.0	6.98	0	43

Table 5.4: Reference counts for painter biographies across nationality groups in the English Wikipedia edition.

The results in (Table 5.4) indicate that the English Wikipedia edition places the strongest emphasis on referencing compared to other language editions. Painters in the English nationality group average 11.3 references with a median of 6. This shows that while many articles have a moderate reference count, there is a number of articles that contain a high number of references, some reaching into the hundreds. By contrast, biographies of painters from other national groups, such as French (mean = 7.45, median = 4.0) or German (mean = 6.18, median = 3.0), receive fewer references on average. Italian painters are especially underrepresented, with a low mean reference count (mean = 4.42

and median = 2.0). This indicates more editorial attention toward painters from the English nationality group.

The median values indicate bias, while the English nationality group has the highest average, most articles remain sparsely referenced with 6 references. This shows that a subset of articles on famous painters receives editorial attention, skewing the representation. The high standard deviation reinforces this with a value of 18.30. Essentially, while famous painters may be extensively referenced, the majority of entries remain less referenced. This imbalance reflects both the visibility bias and the systemic bias in volunteer contributions.

References in the French Wikipedia edition

Nationality	Mean	Median	Std. Dev.	Max
French	8.54	5.0	14.36	455
English	6.75	3.0	11.41	236
German	5.37	2.0	14.64	455
Italian	4.53	2.0	10.42	285
Spanish	6.27	2.0	13.72	221
Serbian	3.82	2.0	6.28	33

Table 5.5: Reference counts for painter biographies across nationality groups in the French Wikipedia edition.

Table 5.5 presents the distribution of reference counts in painter biographies across six nationality groups in the French Wikipedia. French painters, who make up by far the largest group in the dataset (n = 15,034), receive on average the highest number of references (mean = 8.54, median = 5.0). English (n = 3,311, mean = 6.75) and Spanish painters (n = 877, mean = 6.27) also show comparatively higher referencing levels. In contrast, Italian (n = 2406, mean = 4.53) and German painters (n = 2897, mean = 5.37) are less well referenced on average, while Serbian painters, the smallest group (n = 76), receive the fewest references (mean = 3.82).

The high standard deviations for the French, German and Italian group, respectively 14.36, 14.64 and 13.72, indicates that while some articles have a large number of references, up to 455, the vast majority remain sparsely referenced.

The most editorial attention received painters from the French nationality group (mean = 8.54, median = 5.0), indicating editorial favoritism based on the background of the painters.

References in the German Wikipedia edition

The Table 5.6 presents the distribution of reference counts in painter biographies across six nationality groups in the German Wikipedia. Painter biographies typically contain

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
German	19152	5.37	3.0	9.55	0	409
French	3208	4.37	2.0	9.94	0	159
English	2493	5.54	3.0	11.63	0	174
Italian	1421	4.29	3.0	6.80	0	127
Spanish	643	5.02	2.0	12.67	0	126
Serbian	54	3.80	2.0	4.14	0	18

Table 5.6: Reference counts for painter biographies across nationality groups in the German Wikipedia edition.

between 3.8 and 5.54 references. Articles on painters from the German and English nationality groups have the highest mean value, respectively 5.37 and 5.54.

The low median values (between 2–3 references) show that many biographies remain low in reference counts, despite some extreme outliers that have over 400 references. Figures from the French, Italian, and Serbian groups receive slightly fewer references on average with a mean value between 3.8 and 4.37, indicating an uneven distribution of editorial effort between different nationalities. While the English nationality group has the highest mean value, it is important to remark that in the German Wikipedia, the German painters are the most numerous and better referenced than non-German groups, implying bias. The German edition has a long-tail distribution concerning referencing pattern, seen in how a small number of very detailed pages are there with a broad base of minimally referenced articles.

The standard deviation for the Italian and Serbian nationality groups is lower on average than the others with a value of 6.8 and 4.14, respectively. This indicates that, while there is a longtail distribution, it is much lower; meaning that on average the articles have a comparably more similar reference count.

This structural difference illustrates how Wikipedia editions diverge not only in content but also in editorial practices, reinforcing systemic bias.

References in the Spanish Wikipedia edition

Table 5.7 represents the distribution of reference counts in painter biographies across six nationality groups in the Spanish Wikipedia. Painters in the Spanish nationality group, which is the largest group ($n = 6146$), have a mean value of 6.14 while the median value is 3.0. This shows that while there are outliers with the number of references indicated also by a high standard deviation of 10.54, most of the articles remain sparsely referenced.

Among the other painters, the most well-referenced are painters from the English nationality group, with a mean value of 8.4 and the highest median value of 4. Here is also the highest standard deviation with 16.73 and a maximum of 361, confirming a long-tailed

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
Spanish	6,146	6.14	3.0	10.54	0	194
French	1,641	4.92	2.0	11.31	0	237
English	1,560	8.40	4.0	16.73	0	361
Italian	1,477	3.81	2.0	7.05	0	118
German	1,190	4.61	2.0	9.47	0	147
Serbian	38	2.66	2.0	3.81	0	19

Table 5.7: Reference counts for painter biographies across nationality groups in the Spanish Wikipedia edition.

distribution where a subset of painters is well-referenced, while most have a small number of references.

French and German painters have a mean value of 4.92 and 4.61 and a median value of 2. This shows that these painters are less well-referenced on average, despite their centrality to European art history.

Serbian painters are the smallest group ($n = 38$) and have the lowest values, with a mean value of 2.66 and a median value of 2. The small maximum value of 19 indicates marginal coverage and limited editorial investment.

The overall high standard deviations present across all groups indicate a skewed distribution where a minority of articles are very well sourced, while many remain lightly referenced, which serves as evidence of uneven editorial attention across national groups.

These results confirm the presence of bias. Although painters from the Spanish nationality group have the highest mean value in the Spanish Wikipedia editions across other language editions, they are still less referenced on average than painters from the English nationality group. Painters from the English nationality group receive comparatively richer referencing, reflecting also the wider availability of English-language sources.

References in the Italian Wikipedia edition

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
Italian	6793	5.46	3.0	9.83	0	236
French	1855	4.95	2.0	9.62	0	180
English	1280	7.28	3.0	16.35	0	204
German	1269	4.52	2.0	10.40	0	180
Spanish	412	5.31	2.0	12.41	0	140
Serbian	39	4.05	2.0	5.30	0	24

Table 5.8: Reference counts for painter biographies across nationality groups in the Italian Wikipedia edition.

Table 5.8 represents the distribution of reference counts in painter biographies across six nationality groups in the Italian Wikipedia. The highest number of entries are the painters from the Italian nationality group ($n = 6793$). Although they have the highest count, they have a mean value of 5.46, which is substantially lower than the mean value of painters of the English nationality group with 7.28. This higher mean number is driven by a group of highly referenced individuals, as for both the Italian and English painters the median and maximum values are comparable.

The French, German and Spanish articles show lower reference levels (means between 4.52 and 5.31), while the biographies of Serbian painters are the least referenced overall. The distribution is skewed in all groups, with median values clustered at only 2–3 references, despite some very well-referenced outliers reaching more than 200 citations. This points to a long-tail pattern: a few highly sourced articles contrast with a large base of minimally referenced entries.

These results indicate that, while Italian painters are well represented in quantity, they lack referencing depth. Painters from the English nationality group achieve a higher mean value. This pattern can also be observed in the Spanish and German Wikipedia versions, while the local painters are comparatively well-referenced to other language versions, the painters from the English language group lead. This is potentially due to the vast availability of sources in the English language.

References in the Serbian Wikipedia edition

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
Serbian	715	4.71	3.0	5.72	0	48
English	264	9.00	4.0	14.43	0	114
German	250	5.46	3.0	10.60	0	118
French	237	6.03	3.0	10.62	0	118
Italian	115	6.30	3.0	12.57	0	114
Spanish	41	9.22	7.0	11.43	0	60

Table 5.9: Reference counts for painter biographies across nationality groups in the Serbian Wikipedia edition.

The data in Table 5.9 show that English and Spanish artists receive the most references on average (means around 9), while Serbian painters, with 715 entries, are referenced less frequently with a mean value of 4.71. While this is a relatively low reference count, it is higher than in other Wikipedia language editions, except for the English Wikipedia language edition, where the mean value is higher with 6.42.

The imbalance between English and Spanish painters toward Serbian painters also points to editorial bias, where more global figures receive more attention compared to local painters. Although a pattern of high mean values can be observed for painters of the English nationality groups across all language editions, the high mean values of painters

from the Spanish nationality group are an anomaly. As the count of Spanish painters is quite low, with 41 a possible reason could be that some editors who have a special interest in painters from that group have given those figures additional editorial attention.

Painters from the German, French and Italian nationality groups have a comparable mean value ranging from 5.46 to 6.30, indicating that a uniform level of editorial attention was given.

Although local painters are present the most, with a count of 715, they are under-referenced, possibly also a result of generally less available digital knowledge on less globally prominent historical figures. This can indicate how smaller Wikipedia editions also reproduce broader cultural biases and focus on international figures rather than counterbalancing them.

5.3.2 Results and Conclusion on References

After examining the reference counts across different Wikipedia language editions, differences can be observed. The English Wikipedia has the highest number of references in both mean values and maximum numbers of references. This indicates a strong editorial base.

Local painters have a higher reference count in their "local" Wikipedia language edition, which can be observed in the French, German, Spanish, Italian, and Serbian language editions. An example of this is that painters from the French nationality groups have a much more extensive referencing in the French Wikipedia language edition than other nationality groups. Nevertheless, the high standard deviations show that a large number of biographies are sparsely referenced, with some outliers achieving high reference counts.

While the painters from the English nationality group have the highest mean value in their Wikipedia language editions, across other language editions, they still have the second-highest, if not the highest, mean reference count. A possible reason for that can be that they are more internationally recognized and as such have a high number of works that can be referenced.

Painters from the Italian nationality group remain in all Wikipedia language editions as one of the lesser referenced groups, except in their own language edition.

This reveals two important findings. First, there are structural inequalities between language editions. The English Wikipedia emphasizes referencing much more than other language editions, such as Serbian or Italian. Secondly, there are inequalities within the language editions based on the linguistic and regional background of the painters. More dominant nationalities, which have more native speakers, receive a higher reference count on average, which can reinforce visibility.

The uneven distribution of references across languages and nationalities indicates broader editorial biases. As references serve as indicators for article quality and representation, these biases can influence how knowledge is preserved and accessed in Wikipedia.

5.3.3 Internal links across language editions

Internal links in Wikipedia connect pages to other Wikipedia entries. They connect Wikipedia articles with other articles concerning topics such as people, places and events within Wikipedia [3]. They serve as a navigable knowledge graph and reflect how well connected the article is to the broader domain within Wikipedia.

In the context of this thesis, the number of internal links is used as a metric to measure bias and editorial attention given in the form of how well the article is integrated inside of Wikipedia. Although it is not a definitive metric, it can indicate that more editorial effort was given in constructing the article. By measuring the number of internal links of articles concerning painters across multiple Wikipedia language editions and grouping them by nationality group, we can use it as an indicator for the potential bias in the dataset. The number of internal links also has a role in the construction of knowledge graphs and algorithmic ranking [Sun et al., 2021]. Lack of links can lead to less visibility in search engines and less visibility inside Wikipedia.

Internal links in the English Wikipedia edition

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
English	17816	99.80	66.0	143.64	17	3718
French	6861	86.49	61.0	107.63	16	2701
German	5132	84.05	61.0	119.54	17	3165
Italian	4385	71.03	51.0	86.50	18	1994
Spanish	2922	87.86	59.0	103.82	18	1869
Serbian	252	91.48	71.5	71.81	18	493

Table 5.10: Internal link counts for painter biographies across nationality groups in the English Wikipedia edition.

In Table 5.10 we can see that the biographies of the English Wikipedia edition are highly interconnected, with the mean averaging between 71.5 and 99.8. This illustrates both the size of the English edition and the editorial culture, which emphasizes hyperlinking to related people, movements, and historical contexts. In addition, this emphasizes the concept of Wikipedia as a knowledge graph and increases the navigability of painter biographies and their interconnectedness in a network of cultural and historical references.

Painters in the English nationality group outperform the other groups with a mean of 99.8. This suggests a bias in the local editorial traditions. Serbian painters, while having the lowest count with 252 painters, outperform other nationality groups with a mean of 91.48. Additionally, they have the highest median value, implying that overall Serbian painters' biographies have a high number of internal links. This could indicate that,

[3] Help:Link — Wikipedia, https://en.wikipedia.org/w/index.php?title=Help:Link&oldid=1301529598, Accessed on: 2025-07-23

although Serbian articles are relatively short and less referenced, editors compensate by embedding them more deeply in the knowledge graph. Other nationality groups have a relatively similar mean value, ranging from 84.05 to 87.86, implying a generally even distribution of editorial efforts. However, Italian painters are much less densely linked with a mean value of 71.03. This suggests that while Italian painters are very important in the domain of art history, they are far less interlinked in the English edition's article network. This can also be connected to how well Italian painters are referenced.

Painters from the German, French and Spanish nationality groups have similar mean results ranging from 84.05 to 87.86, indicating that they are to a similar degree embedded into the knowledge graph of Wikipedia.

From a bias perspective, these results indicate that linguistic dominance translates not just into higher reference density but also into greater network centrality, amplifying systemic visibility biases across nationalities.

Internal links in the French Wikipedia edition

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
French	15,034	153.63	124.0	136.27	25	3,592
English	3,311	150.99	129.0	122.36	25	1,960
German	2,897	148.11	124.0	122.51	25	1,481
Italian	2,406	145.90	125.0	95.84	25	1,491
Spanish	877	149.33	123.0	133.02	25	1,460
Serbian	76	148.58	133.0	83.89	25	547

Table 5.11: Internal link counts for painter biographies across nationality groups in the French Wikipedia edition.

The Table 5.11 shows that painters in the French nationality group are most densely interlinked, resulting from a mean value of 153.63. Other groups also have a high mean value, ranging from 145.9 to 150.99. This shows that in the French Wikipedia, internal links are seen as an important feature of article quality, which can be seen in how consistent the practice is across all groups.

Even though the mean value is consistently high across all nationality groups, the high standard deviations and extreme maximum values show that there is a highly uneven distribution within those groups. With the mean values ranging between 145.9 to 150.99, with outliers having values of 1400 and more. These articles reinforce the disproportionate visibility and navigability of prominent painters toward lesser-known historical figures.

From the perspective of bias, the bias based on nationality is evident in the fact that, on average, painters from the French nationality group have the highest mean value. A more apparent bias is noticeable in the prominence of selected well-known painters and how thoroughly they are interlinked, while lesser-known figures are minimally connected, further limiting their discoverability within the knowledge graph of Wikipedia.

Internal links in the German Wikipedia edition

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
German	19152	65.56	58.0	48.94	0	1542
French	3208	69.25	59.0	58.99	15	941
English	2493	72.96	62.0	66.01	15	1612
Italian	1421	69.87	61.0	46.81	15	605
Spanish	643	63.89	52.0	66.37	15	941
Serbian	54	61.57	59.0	22.03	15	131

Table 5.12: Internal link counts for painter biographies across nationality groups in the German Wikipedia edition.

In the Table 5.12 it is seen that in the German Wikipedia, biographies are moderately interlinked with a mean value ranging between 61.57 and 72.96. This is noticeably lower than in the French Wikipedia, indicating less editorial attention to interlinking. Internal linking across nationalities is evenly distributed, with English (mean = 73.0), French (mean = 69.3), and Italian (mean = 69.9) painters only slightly ahead of German painters. Spanish (mean = 63.9) and Serbian (mean = 61.6) painters are the least interlinked. Since the values are relatively close, this suggests a baseline of editorial consistency.

Similar to the French Wikipedia edition, high standard deviations and high maxima indicate internal disparities within nationalities.

From the perspective of bias, there is less nationality-based bias, as the German nationality group is not the most prominently interlinked in this Wikipedia. But the overall lower values compared to the French and English Wikipedia suggest that the German Wikipedia is less integrated into the broader network of the encyclopedia. While there is editorial uniformity across nationality, articles are less visible and interconnected, making them less discoverable.

Internal links in the Spanish Wikipedia edition

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
Spanish	6,146	74.10	62.0	60.25	9	1,561
French	1,641	97.17	86.0	61.32	16	962
English	1,560	101.18	88.0	69.33	16	1,038
Italian	1,477	85.74	75.0	50.06	9	569
German	1,190	95.07	84.0	59.98	16	865
Serbian	38	74.50	72.5	36.06	16	149

Table 5.13: Internal link counts for painter biographies across nationality groups in the Spanish Wikipedia edition.

In Table 5.13, it can be seen that between nationalities based on the mean value, there are significant variations in the number of internal links. Painters in the English nationality group are most densely interlinked, with a mean value of 101.18, followed closely by French and German painters with mean values of 97.17 and 95.07. The painters from the Spanish nationality group, while being the largest group in the dataset, have the lowest mean value of 74.1. This suggests that even in their own language edition, Spanish painters receive relatively lower attention compared to foreign painters. The painters from the Serbian and Italian nationality groups are also on the lower end of mean values, with 74.5 and 85.74, respectively. This corresponds to how overall these groups perform across Wikipedia language editions.

The high standard deviation across all nationality groups suggests that, as in the French and English Wikipedia, the dataset is dominated by a relatively small number of highly interlinking figures. This reflects how editorial efforts are selective towards prominent figures.

Internal links in the Italian Wikipedia edition

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
Italian	6793	102.59	80.0	92.66	15	1672
French	1855	106.07	91.0	78.62	15	989
English	1280	120.40	94.0	125.49	15	1822
German	1269	108.26	84.0	105.68	15	1404
Spanish	412	112.40	87.5	115.86	15	1617
Serbian	39	91.10	87.0	30.44	15	162

Table 5.14: Internal link counts for painter biographies across nationality groups in the Italian Wikipedia edition.

The Table 5.14 suggests that across all nationality groups, there is dense interlinking with mean values averaging between 102.59 and 120.4. The least number of interlinks is for painters from the Serbian nationality group, with a mean value of 91.1. While painters from the Italian nationality group have a high number of interlinks with a mean value of 102.59, painters from other nationality groups are more densely interconnected.

The standard high deviation, as in other Wikipedia editions, reveals a small group of biographies that are very well linked, while the majority have a moderate degree of inter-connectivity.

From the perspective of bias, these findings reveal that the Italian Wikipedia does not heavily favor Italian artists in the case of interlinking, even if they are the most numerous. Rather, painters from the English and Spanish nationality groups are better integrated into the structure of the encyclopedia. As a result, Italian painters are at risk of being less findable compared to external peers within their own language edition, evidencing

the way editorial procedures can replicate wider cultural inequalities in the representation of knowledge.

Internal links in the Serbian Wikipedia edition

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
Serbian	715	99.92	77.0	77.40	30	739
English	264	147.05	112.0	104.17	30	825
German	250	133.51	115.0	69.19	30	500
French	237	136.18	116.0	77.43	30	722
Italian	115	119.54	99.0	60.28	30	386
Spanish	41	138.76	103.0	90.93	54	497

Table 5.15: Internal link counts for painter biographies across nationality groups in the Serbian Wikipedia edition.

The Table 5.15 shows that painters from the Serbian nationality group have a relatively high mean value with 99.92. This is a relatively high number, especially compared to their representation in other Wikipedia language editions such as German, Spanish and Italian. Compared to other nationality groups, it is a surprisingly low number as the mean values of other groups are between 119.54 and 147.05. This shows that despite being the most numerously represented group with a count of 715, they are less extensively interlinked than others.

This is further illustrated with the maximum values, with painters from the Serbian nationality group having a maximum of 739, while painters from the English and French nationality groups have maximums of 825 and 722. Additionally, the high standard deviation across nationalities further suggests uneven editorial practices.

From the perspective of bias, this indicated structural uneven representation. Even in a local language edition, foreign painters are better positioned within Wikipedia's network than domestic Serbian painters. While painters from the Serbian nationality group are dominant in numbers, they are less discoverable due to weaker interlinking.

5.3.4 Results and Conclusion on Internal Links

Internal links are one of the strongest indicators of article integration, because a more dense link structure improves navigability, discoverability, and algorithmic visibility within the encyclopedia. [Menghini et al., 2021] After analyzing the differences between the number of internal links in multiple Wikipedia language editions, an uneven distribution can be observed.

The French Wikipedia has the highest overall mean, varying between 145 and 153. Most biographies have over 150 links, some going into the thousands. An important point to note is that the editorial culture of using internal links is not only limited to French

painters but also covers other nationality groups. This shows a uniform editorial practice. Nevertheless, the highest mean value is reserved for the French nationality group, with the maximum number of links for a single biography being 3592 links. This shows that, while there is a uniform linking practice, there is still bias present based on the nationality of the painters.

The English edition also shows a strong number of links per biography, with a mean between 70 and 99. The highest number of internal links is still reserved for the painters of the English nationality group, with a mean of 99.8 and a maximum count of 3718. The painters from the Italian nationality group have the smallest mean value with 71 links, while other language editions have much higher values, ranging from 84 to 91.

In contrast to that, the German Wikipedia edition shows a lower mean value across all nationality groups, ranging from 61 to 72, suggesting that articles are less interconnected and that less editorial attention is given. In the Spanish Wikipedia edition, a higher number of internal links is present for English, French, and German painters, with a mean value ranging between 95 to 101, than for painters from the Spanish nationality group (mean=74.10). Similarly, for the Italian Wikipedia, where Italian artists are well integrated (mean=103), while English and Spanish artists are associated even more strongly (mean=120 and mean=112). This can suggest that while the nationality group itself is well interlinked, more global figures are better connected in Wikipedia's knowledge graph.

The Serbian Wikipedia language edition has a large difference between the mean number of links between nationality groups. The painters from the English nationality group have the highest number of internal links with 147, the German, French, Italian, and Spanish all have a mean value that spans between 119 and 138. This is a big difference from the Serbian nationality group, which has the lowest mean value with just 99. While this is a high number compared to how they are interlinked in some other Wikipedia language editions, it is still showing how less globally known artists are less interconnected even in their own language editions.

These results show that more dominant cultural traditions, such as English, French, and German, are better interlinked, even outside their "home" editions. Additionally, editorial attention gravitates to more globally well-known figures, further amplifying their network centrality and reinforcing visibility.

These findings imply that there are several layers of bias present. Larger language editions, such as English and French, have more extensive linking, which strengthens the visibility of painters. Smaller language editions, such as Serbian, Spanish and Italian, show a higher number of links for foreign figures over local ones, which can stem from the availability of data. Lastly, the existence of outliers that have a far higher count than the majority can create an imbalance and heighten the prominence of those painters. These patterns are indications of Wikipedia's network structure, which, while appearing to be neutral, simply reiterates the hierarchy of culture while granting greater discoverable space and navigation to the painters of the dominant tradition.

5.3.5 Categories across language editions

Categories in Wikipedia are used to link articles under a common topic and can be found at the bottom of a Wikipedia article [4]. They can be used as navigational tools within Wikipedia and can also contextualize articles within broader domains.

In the context of the dataset examined, they show how painters are grouped within each Wikipedia language edition based on their nationality group. The number of categories can indicate a different editorial focus and can show how well integrated biographies are. This difference can affect the ease with which new articles are found and the way readers understand the position of artists in the broader landscape. Although the number of categories does not indicate the quality or granularity of the categorization, it can serve as an indicator of editorial differences.

Categories in the English Wikipedia edition

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
English	17816	9.78	9.0	6.56	0	101
French	6861	8.84	8.0	5.25	0	65
German	5132	8.77	8.0	5.03	0	76
Italian	4385	8.28	8.0	3.30	0	51
Spanish	2922	8.02	8.0	4.95	0	46
Serbian	252	8.82	8.0	4.28	0	29

Table 5.16: Category counts for painter biographies across nationality groups in the English Wikipedia edition.

Table 5.16 shows the distribution of category counts for painter biographies across six nationality groups in the English Wikipedia edition. For painters from the English nationality group, the categorization is most extensive, with the mean value being 9.78. Articles about painters from the French, German, and Serbian nationality groups also show relatively high categorization with values ranging from 8.77 to 8.84. For the Spanish and Italian groups, the range is lower, with the mean being 8.02 and 8.82, respectively.

The median values are consistent across all groups, with a value of 8, with the English being the only one with a median of 9.

The standard deviation is the highest with painters from the English nationality group due to the high maximum count of 101. This demonstrates the highly networked classification practices of the English Wikipedia, supporting strong discoverability.

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
French	15,034	8.74	8.0	4.74	0	60
English	3,311	8.22	8.0	4.72	0	60
German	2,897	8.82	9.0	4.52	0	49
Italian	2,406	7.89	8.0	3.13	0	29
Spanish	877	8.00	7.0	4.96	0	45
Serbian	76	9.51	9.0	4.84	0	26

Table 5.17: Category counts for painter biographies across nationality groups in the French Wikipedia edition.

Categories in the French Wikipedia edition

Table 5.17 shows the distribution of category counts for painter biographies across six nationality groups in the French Wikipedia edition. Across all nationality groups, the average category count is relatively consistent, with the mean value ranging between 7.89 and 9.51. Painters from the German and Serbian nationality groups have the highest mean value, respectively 8.82 and 9.51, while painters from the Italian nationality group have the lowest mean with 7.89. The surprisingly high category count of the Serbian nationality group could be the work of a small group of editors creating the articles, as the count of painters is only 76. As there are no guidelines for the category counts, special editorial attention could have been given.

While the means are relatively consistent, the maximum values highlight that certain painters receive a richer categorization. For painters from the French and English nationality group, the maximum value is 60, which, in contrast to other groups, is a higher value. These outliers tend to be well-known figures, amplifying their visibility across multiple thematic entry points.

From the perspective of bias, these results suggest that categorization bias is less visible on a cross-national level but within national groups. It is important to note that the nationality group with the highest maximum values corresponds to the language of the Wikipedia edition.

Categories in the German Wikipedia edition

Table 5.18 shows the distribution of category counts for painter biographies across six nationality groups in the German Wikipedia edition. The categorization is consistent across nationality groups, with the mean value ranging between 6.56 and 7.78. English and Serbian biographies are slightly more richly categorized, while Italian and Spanish biographies have the lowest categorization. The standard deviations are also relatively close in value, with a range between 2.87 and 3.93. The biggest indicator of bias is as in

[4]Help:Categories — Wikipedia, https://en.wikipedia.org/w/index.php?title=Help:Categories&oldid=1265981411, Accessed on: 2025-07-23

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
German	19152	7.03	7.0	3.70	0	61
French	3208	7.13	7.0	3.69	0	36
English	2493	7.38	7.0	3.93	0	46
Italian	1421	6.56	6.0	2.87	0	27
Spanish	643	6.84	7.0	3.38	0	27
Serbian	54	7.78	7.0	3.18	0	15

Table 5.18: Category counts for painter biographies across nationality groups in the German Wikipedia edition.

the English and French Wikipedia editions the maximum count. The similarity in medians (6–7) confirms that categorization is a stable editorial norm in the German Wikipedia, helping to ensure the discoverability of painter biographies.

Categories in the Spanish Wikipedia edition

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
Spanish	6,146	7.44	7.0	3.88	0	48
French	1,641	8.35	8.0	4.03	0	33
English	1,560	7.97	7.0	4.24	0	49
Italian	1,477	7.58	7.0	2.67	0	27
German	1,190	8.08	8.0	3.77	0	40
Serbian	38	7.11	8.0	3.92	0	13

Table 5.19: Category counts for painter biographies across nationality groups in the Spanish Wikipedia edition.

Table 5.19 shows relatively uniform categorization with means ranging between 7.11–8.35. Painters from the French and German nationality groups have a slightly higher categorization. Across all nationality groups, the median values remain uniform with values of 7 and 8. As seen in the other tables, the maximum is the most apparent indicator of which Wikipedia language edition is the "local" one, with the Spanish nationality group having the second highest value with 48.

Categories in the Italian Wikipedia edition

Category counts in Table 5.20 reveal that painter biographies in the Italian edition are categorized in a similar value range as the other Wikipedia language editions. The mean values range between 7.24 and 9.72. Painters from the Serbian nationality group have the lowest count with 39, but the highest median value with 10 and a mean value of 9.72. This suggests that even though a very small subset of painters is documented, they are richly categorized. The reason for that could be the same as in the French

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
Italian	6793	7.24	7.0	3.15	0	42
French	1855	7.93	8.0	3.46	0	26
English	1280	8.51	8.0	3.95	0	51
German	1269	8.00	8.0	3.39	0	31
Spanish	412	8.39	8.0	3.87	0	29
Serbian	39	9.72	10.0	3.24	0	17

Table 5.20: Category counts for painter biographies across nationality groups in the Italian Wikipedia edition.

Wikipedia edition. Because of the low number of painters, it is possible that a small subset of editors gave special attention to documenting these figures. Although the painters from the Italian nationality group have the largest count, they have the lowest mean and median values (mean=7.24 and median=3.15). However, the medians (7–8) indicate that categorization is a well-established editorial practice, even if its intensity varies. As in the other Wikipedia language editions, the maximum value is highest for painters from the "local" nationality group.

Categories in the Serbian Wikipedia edition

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
Serbian	715	6.12	6.0	3.01	1	31
English	264	4.78	4.0	2.55	0	21
German	250	5.05	4.0	2.11	1	15
French	237	4.71	4.0	1.87	0	13
Italian	115	4.63	4.0	1.77	0	14
Spanish	41	4.12	4.0	1.38	0	7

Table 5.21: Category counts for painter biographies across nationality groups in the Serbian Wikipedia edition.

Table 5.21 indicates that painters from the Serbian nationality group are categorized more richly (mean=6.12 and median=6) compared to other groups, with mean values ranging between 4.12 and 5.05. This is also apparent with the median values, with the value being 4 in all groups except Serbian.

In the other Wikipedia language editions, the categorization was relatively uniform across nationality groups, with only the maximum value being the highest for the "local" nationality groups. While in the Serbian Wikipedia language edition, the "local" nationality group is most richly categorized by a margin. This reflects an editorial emphasis on categorizing local figures.

5.3.6 Results and Conclusion on Categories

After examining all language editions, a more uniform distribution of category counts can be observed, unlike references and internal links, where the disparities are more apparent. The median count ranges between 6 and 9 categories per painter biography across all language editions, which would suggest that categorization is a relatively stable editorial practice across Wikipedia language editions.

Nonetheless, some systematic biases can be observed. In larger language editions such as English, French and German, painter biographies have a higher maximum count. In contrast, smaller language editions such as Serbian and Italian have narrower categorization count ranges, which may hinder the discoverability of painter biographies.

The standard deviation varies between 2.87 and 6.56 across all language editions except the Serbian one, which has a smaller standard deviation, indicating that smaller language editions can be less informed in the topic of categorization.

Another aspect is the bias shown on a national basis. Painters who are from the nationality group of the spoken Wikipedia language editions tend to have richer categorizations. This is most apparent with the Italian and Serbian editions. Serbian painters, for example, tend to receive more categories on average in the Serbian Wikipedia than English, French, or Spanish painters. In contrast, Italian painters have a less rich categorization in their own edition. The bias is less apparent than for references and internal links, where the highest mean value often corresponded directly to the "local" nationality group. In categorization, the possible indicator of the pattern that the highest maximum count is from the "local" nationality group, but no definitive conclusions can be made.

Categories are important as they function as a navigational tool and influence how readers interact and discover painter biographies within Wikipedia. A higher number of categories can imply a higher importance of the biography and increase visibility. The bias present in categorization practices thus reinforces inequalities and makes foreign or minority groups less discoverable while giving greater importance to dominant groups.

Even though the overall uniform medians of categorization suggest a less unequal spread than in the case of references, they still reflect bias in the organization of knowledge.

5.3.7 Sections across language editions

Sections subdivide articles into logical areas and can serve as an indicator of how comprehensively a topic is covered [5]. Although they cannot directly indicate the quality of an article, a larger number of sections can reflect a higher level of editorial interest in the topic.

In this thesis, the number of sections is taken as one of the indicators of the quality of the article. By analyzing the number of sections in each Wikipedia language edition and grouping them by the nationality group of the painters, it can suggest a higher degree of organization and editorial effort. This can indicate a bias in the quality of the articles based on the nationality groups and the Wikipedia language edition. A disparity in the section count can result in shorter articles, making them less credible and visible in search rankings.

Sections in the English Wikipedia edition

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
English	17816	6.38	6.0	4.52	0	67
French	6861	5.68	5.0	4.14	0	49
German	5132	5.55	5.0	3.83	0	46
Italian	4385	4.26	3.0	3.08	0	34
Spanish	2922	5.34	5.0	3.80	0	34
Serbian	252	5.95	5.0	3.20	0	18

Table 5.22: Section counts for painter biographies across nationality groups in the English Wikipedia edition.

In the Table 5.22 we can see the number of sections and how they vary in the English Wikipedia edition across multiple nationalities. The English nationality group has the highest median and mean value, with 6 and 6.34, respectively. This can be attributed to the large size of the Wikipedia edition, as well as to the large editorial community. The French, German and Spanish nationality groups have a similar mean varying between 5.34 to 5.68, indicating a fairly consistent pattern in structuring biographical content.

Painters from the Italian nationality group stand out within the comparably low mean value of 4.26 and median of 3. While Italian artists are important in art history, their biographies are described in less detail in terms of structure in the English Wikipedia edition. In contrast to that, painters from the Serbian Wikipedia edition have a high mean value of 5.95, comparable to the English nationality group. This suggests that even though they are smaller in numbers, the entries maintain structural and editorial consistency.

[5] Help:Section — Wikipedia, https://en.wikipedia.org/w/index.php?title=Help:Section&oldid=1300056387, Accessed on: 2025-07-23

From the perspective of bias, the results suggest that there is more editorial attention given to the structure of "local" painters.

Sections in the French Wikipedia edition

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
French	15,034	8.11	7.0	5.24	0	71
English	3,311	7.07	6.0	4.77	0	50
German	2,897	6.72	6.0	4.76	0	50
Italian	2,406	7.22	7.0	3.82	0	44
Spanish	877	7.25	6.0	5.46	0	50
Serbian	76	6.80	6.0	3.76	0	20

Table 5.23: Section counts for painter biographies across nationality groups in the French Wikipedia edition.

Table 5.23 shows the number of sections for painter biographies across six nationality groups in the French Wikipedia. French biographies again dominate the dataset ($n = 15{,}034$) and exhibit the highest average section count (mean = 8.11, median = 7.0). Spanish (mean = 7.25) and Italian painters (mean = 7.22) are close behind, while English (mean = 7.07), German (mean = 6.72), and Serbian painters (mean = 6.80) are described with slightly fewer sections on average.

Although the variation is less extreme than for reference counts, standard deviations indicate considerable diversity in editorial depth (e.g., $\sigma = 5.24$ for French painters, with a maximum of 71 sections). This suggests that while many biographies receive relatively balanced coverage, prominent figures may benefit from disproportionately detailed structures. These differences in organizational depth provide further evidence of editorial bias: painters from dominant cultural traditions (e.g., French, Italian, Spanish) tend to receive longer and more subdivided articles, while smaller groups remain comparatively less elaborated.

Sections in the German Wikipedia edition

In the Table 5.24 the results on section counts in the German Wikipedia edition are relatively consistent, averaging a mean value between 5.28 to 6.03 across nationality groups. English painters' articles are the most subdivided (mean=6.03), while French and Italian are slightly less structured with mean values of 5.56 and 5.51, respectively. Painters from the German nationality group have the second highest mean with 5.93, with the painters from the English nationality group having the highest mean value with 6.03.

Painters from the Spanish, French and Italian nationality groups have a lower mean and median value, indicating a relatively consistent editorial treatment but lower than the "local" nationality group.

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
German	19152	5.93	6.0	3.76	0	49
French	3208	5.56	5.0	4.05	0	48
English	2493	6.03	6.0	4.17	0	38
Italian	1421	5.51	5.0	2.98	0	32
Spanish	643	5.28	5.0	4.20	0	47
Serbian	54	5.72	5.0	3.10	0	13

Table 5.24: Section counts for painter biographies across nationality groups in the German Wikipedia edition.

Painters from the Serbian nationality group have a relatively high mean value of 5.72 despite the low number of painters (n = 54). This suggests that despite the low numbers, they receive a high level of editorial attention and structural completeness. The low maximum count of 13 reflects reduced editorial engagement relative to larger nationality groups.

From the perspective of bias, the overall similarity indicates that the German Wikipedia has adopted a common editorial template for structuring biographies, ensuring comparable article organization across different national traditions. Still, there is a slight home-language bias noticeable.

Sections in the Spanish Wikipedia edition

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
Spanish	6,146	5.66	5.0	3.76	0	67
French	1,641	5.44	5.0	4.10	0	67
English	1,560	6.25	5.0	4.34	0	47
Italian	1,477	5.21	5.0	2.88	0	30
German	1,190	5.37	5.0	3.55	0	34
Serbian	38	4.13	4.0	2.92	0	13

Table 5.25: Section counts for painter biographies across nationality groups in the Spanish Wikipedia edition.

Table 5.25 indicates broadly similar section counts across groups (means 4–6.25). English painters have the most sections on average (mean=6.25), while painters from the Spanish nationality group have the second highest mean value with 5.66.

Painters from the French, Italian and German nationality groups have a similar mean value ranging between 5.21 to 5.44 and the same median value of 5, indicating a similar structure.

The Serbian nationality group painters have the lowest mean and median values of 4.13 and 4. This can indicate a scarcity of editors familiar with the topics or reduced editorial attention.

From the perspective of bias, the Spanish language edition shows a moderate degree of national editorial preference. The moderate standard deviations and occasional extremes (max up to 67) again point to uneven editorial depth focused on a subset of biographies, but the general consistency between groups and the moderate standard deviations suggest a stable editorial standard.

Sections in the Italian Wikipedia edition

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
Italian	6793	6.53	6.0	3.73	0	49
French	1855	7.05	7.0	3.98	0	30
English	1280	7.33	6.0	4.78	0	44
German	1269	6.47	6.0	4.17	0	44
Spanish	412	6.70	6.0	4.22	0	27
Serbian	39	6.67	6.0	3.72	0	21

Table 5.26: Section counts for painter biographies across nationality groups in the Italian Wikipedia edition.

In the Table 5.26) are fairly stable between all nationality groups, with most biographies containing between 6–7 subsections. English and French painters are slightly more subdivided, with a mean value of 7.05 and 7.33. This reflects a tendency toward more detailed structural organization, which can also be seen in their own Wikipedia language editions. The Italian nationality group has the highest count (n = 6793), but compared to other nationality groups, a relatively low mean value of 6.53. This shows that although there is high coverage (many painter articles), it typically lacks structural depth. German, Spanish, and Serbian painters follow with means between 6.47 and 6.70, showing a relatively consistent pattern and limited disparity in article structure across nationality groups.

In the context of bias, these results show no "home-language" favoritism; other nationality groups on average have more structured articles. The only criterion that shows "home-language" bias is the maximum count of 49.

Sections in the Serbian Wikipedia edition

Section counts (Table 5.27) are relatively balanced, with most nationality groups averaging a median between 6 and 7 sections and a mean value between 6.12 and 7.34. The painters from the Serbian nationality group have the highest median value and the lowest standard deviation (median=7 and standard deviation=3.11). This implies there is a more uniform section count across articles.

Nationality	Count	Mean	Median	Std. Dev.	Min	Max
Serbian	715	6.80	7.0	3.11	0	22
English	264	6.89	6.0	4.54	0	39
German	250	6.22	6.0	3.75	0	32
French	237	6.12	6.0	3.46	0	24
Italian	115	5.49	5.0	3.37	0	19
Spanish	41	7.34	6.0	5.02	2	24

Table 5.27: Section counts for painter biographies across nationality groups in the Serbian Wikipedia edition.

The painters from the English nationality group have the highest mean value, suggesting editorial attention to more "internationally well-known" painters. The painters from the French and German nationality groups are very similar in both mean values(6.22 and 6.12) and median values (both are 6).

The painters from the Spanish nationality group have the highest mean value with 7.34. This could be the result of editorial attention from a small subset of contributors, as there is a very limited number of entries (n=41).

Painters from the Italian nationality group have the lowest mean and median values. A similar pattern can also be observed in other Wikipedia language editions.

5.3.8 Results and Conclusion on Sections

The analysis of sections shows a subtle bias in article structure across Wikipedia's language editions. In general, the section count is much more equally distributed than the reference count, with, on average, the section count being between 5 and 7 subsections per painter. This suggests that the editors tend to stick to a relatively coherent template with subsections of painters' biographies often including "Life", "Death", "Work" and etc.

Still, in larger language editions such as English and French, the median is above that average, with 6 and 8 sections, implying that the biographies are more subdivided. This indicates more structural depth, which likely stems from a larger editorial base. That is in particular noticeable for painters from the French nationality group in the French Wikipedia, with a mean of 8.11, which is the highest across all language editions and implies bias towards their home language edition.

In contrast, in smaller language editions such as Spanish and Serbian, a lower range can be seen, with Spanish averaging between 4 and 6 sections. Biographies of painters from the Italian nationality group have one of the lowest section counts across all Wikipedia language editions, with the lowest being in the English Wikipedia language edition, with a mean of 4.26. This pattern can also be seen in references and indicates that certain groups, outside of their "home" edition, receive less editorial attention. Editorial investment is not distributed uniformly: dominant linguistic and cultural groups (e.g., French in French,

English in English Wikipedia) receive advanced subdivisions, while others are maintained in shorter formats.

In conclusion, while the results show that the section counts are more evenly distributed than references or categories, there is still the presence of regional and language biases. A higher number of sections improves readability and the perceived completeness of biographies. This leads to painters with a higher section count being perceived as more detailed and reinforces presentational inequalities across Wikipedia's multilingual landscape.

CHAPTER 6

Visualization Techniques Assessment

In this section, three different visualization techniques are introduced and analyzed; we use these visualization methods as a foundation for the user study. The visualization techniques explored are heatmaps, flow maps and Sankey diagrams and we apply them to art history data. The main goal of this user study is to assess the interpretability, clarity and usability of each visualization type in the context of cultural and historical data.

The dataset that is going to be used for the study consists of painters from the "Italian nationality group" who are featured in the Italian Wikipedia language edition. This data was selected as it is sufficiently large while at the same time being not too large, i.e. the data is diverse enough, but will not overwhelm the participants in the user study. Although the nationality focus was suggested in part by supervisory guidance, it also provides a useful case study to examine the potential language and regional biases inherent in open-source databases such as Wikipedia.

Each visualization is constructed using the date of birth, date of death and coordinates for birth and death locations. These attributes are critical for mapping spatial movement and enabling the representation of migration patterns and concentrations of artistic activity.

In order to balance comprehensiveness with participant manageability, only three centuries were selected (15th, 16th and 18th). These centuries showed sufficient data availability and we avoid participant fatigue during the user study from too many visualizations. The 17th century was omitted because its patterns were less distinctive compared to the other centuries. For each of these centuries, visualizations were created. This produced 6 heatmaps (birth and death locations per century), 3 flow maps (flows from birth-to-death locations), and 3 Sankey diagrams (illustrating aggregated flows by city).

Each technique has advantages and limitations. While heatmaps emphasize spatial density and clustering, flow maps focus on directional movement, while Sankey diagrams highlight proportional flows between locations.

In the following sections, each visualization is discussed in more detail, their advantages and limitations and the rationale for inclusion is provided. This supports the structured evaluation conducted in the user study and contributes to a better understanding of how different formats influence comprehension, especially for historical and cultural datasets.

6.1 Heatmap

Heatmaps are graphical representations of data that use color coding to represent different values [1]. They most commonly consist of a grid in which rows and columns represent different variables and the intensity of the color of each cell corresponds to the volume of data points in that region[Gu, 2022]. The most common color palette uses cool colors, such as blue and green, for low values and warm colors, such as red and orange, for high values. Clustering is often applied to rows and columns to group similar data points in order to improve the interpretability and pattern recognition [2].

In the context of this thesis, by visualizing the distribution of painters' birth and death locations throughout history, heatmaps are able to highlight temporal shifts and show the formation of emerging cultural centers. They provide an intuitive interpretation of spatial densities and can be useful when working with larger datasets. This is especially the case with large datasets where individual data points may overlap or become visually clustered. In addition, they facilitate the identification of cultural centers and locations with a high density of artists. This aligns directly with the aim of the study of evaluating the degree to which different visualizations convey migration patterns.

Through the use of different data mining techniques, data such as place of birth, place of death, date of birth and date of death of painters are extracted. The data is then aggregated by century and mapped using an intensity-based color scheme on the geographic map of Europe. Areas with higher data point concentrations are rendered in red and orange, while areas with lower concentrations are shown in blue or green as can be seenin Figure 6.1.

The advantage of heatmaps in interdisciplinary research lies in the fact that they are intuitive visual summaries and enable researchers and the general public to interact with large datasets. Heatmaps can reveal insights that are less evident in textual representations of the same data by conveying dense spatial data without requiring the viewer to interact with raw numbers or vast amount of tabular data.

Although heatmaps offer an overview of the density of points and geographic context, they also have limitations. For example, individual values and data points are not portrayed. Despite the limitations of heatmaps, they remain a valuable visualization tool for exploratory data analysis and for communicating findings to both expert and

[1] What is a heatmap? benefits, limitations and examples — Optimizely, https://www.optimizely.com/optimization-glossary/heatmap/, Accessed on: 2025-07-17

[2] A Complete Guide to Heatmaps, https://www.atlassian.com/data/charts/heatmap-complete-guide, Accessed on: 2025-04-06

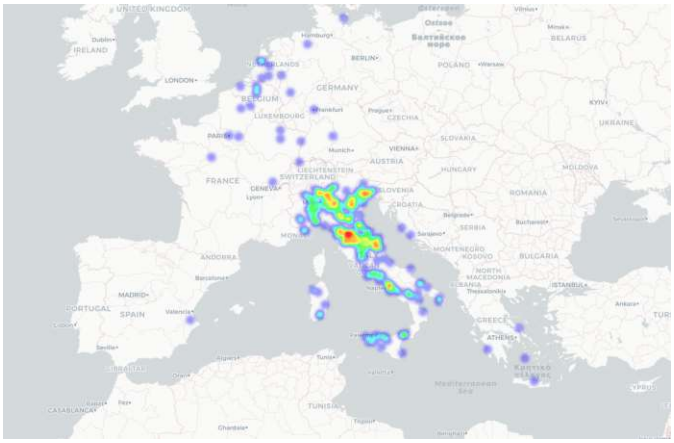


Figure 6.1: Heatmap

non-expert users. Although heatmaps give readily interpretable visualizations of spatial concentration, their shortcomings make it desirable to complement them with other visualization techniques.

From a technical standpoint, a custom Python script was developed using the pandas, folium and folium.plugins.HeatMap libraries. The geographic coordinates used for the heatmap rendering were extracted from the WKT (Well-Known Text)[3] formatted Point() strings in the dataset. A regular expression parser was implemented to extract latitude and longitude values for both birth and death locations.

The historical data was divided into CSV files, where each file represented a century and was subdivided into birth-and-death subsets. In the case of missing or malformed coordinates, the data was omitted to ensure integrity. The Python Folium library was used for data mapping and was chosen because it is a powerful Python library that allows users to create interactive maps and geo-spatial visualizations that can be shared as standalone HTML files or integrated into web applications [4].

In order to improve the readability of the map and minimize over-plotting in densely populated areas, the radius, blur and min_opacity parameters were fine-tuned. The maps were saved as HTML files for later use in the user study.

In the scope of this thesis, heatmaps serve as one of the three visualization methods that are employed alongside Sankey diagrams and flow maps to visualize the spatial and temporal movement of painters across Europe between the 15th and 20th centuries. They are an integral part of the user study, in which users are asked to assess the clarity, usability and interpretability of each visualization technique when applied to the same underlying data set.

[3] Well-Known Text (WKT) — GEOS, https://libgeos.org/specifications/wkt/, Accessed on: 2025-07-27

[4] Folium Documentation, https://folium.readthedocs.io/en/latest/, Accessed on: 2025-07-27

6.2 Flow Maps

Flow maps are a form of visualization in which the movement of entities is represented across a geographical map [5]. In such maps, directional lines are typically drawn between the origin and destination points with the width of the lines often representing the intensity of the flow[Briney, 2014]. They are commonly used as a visualization method in the fields of transportation, logistics and epidemiology and are used to depict spatial mobility[Phan et al., 2005].

In the context of this thesis, flow maps are used to depict the geographical movement of painters from their place of birth to their place of death. Each individual flow represents a painter, with the starting point being the place of birth and the ending point the place of death; these data points are accordingly shown with green and red dots.

By encoding the individual trajectories, flow maps offer insight both on a micro- (individual painters) and macro-level (aggregated view). We can provide a visualization of how artistic centers evolved over time by mapping these trajectories across centuries.

One of the key benefits of using flow maps is the clear origin-to-destination structure. Unlike maps such as heatmaps and Sankey diagrams, where the presence of data is revealed and shown on a less granular level, flow maps can retain and show individual data points.

In the user study conducted as part of this thesis, participants have the option for more exploratory analysis, which allows them to evaluate the readability of such maps, but also delves into more interpretive tasks. Flow maps are particularly suited for this study as they highlight the directional nature of migration and show not only where painters were born and died but also the trajectories between these data points.

Flow maps also come with limitations. One of the main challenges of flow maps is the over-plotting in cases where a significant amount of data is present. In such cases, over-plotting leads to an overlap of flow lines, making the visualization cluttered. This can hide more granular details and reduce the overall readability of the map. In addition to that, short-distance flows can be difficult to visualize and distinguish. Moreover, missing data, such as the place of birth and the place of death, can affect visualization and overemphasize regions that have better coverage of historical data. This issue is particularly prevalent in historical data on Wikipedia, as editorial attention varies between articles and language editions.

In this thesis, to create the flow maps, the Python library "Folium" was used. The data were divided into individual CSV files, where each file represented a century based on the birth dates of the painters. A custom Python script was developed to extract latitude and longitude values for each painter's birth and death locations, then generate individual flows to connect these coordinates. To make the visualization more user-friendly, tool tips were assigned to each flow containing information such as the artist's name, place of birth

[5] Flow maps — data.europa.eu, https://data.europa.eu/apps/data-visualization-guide/flow-maps, Accessed on: 2025-04-07

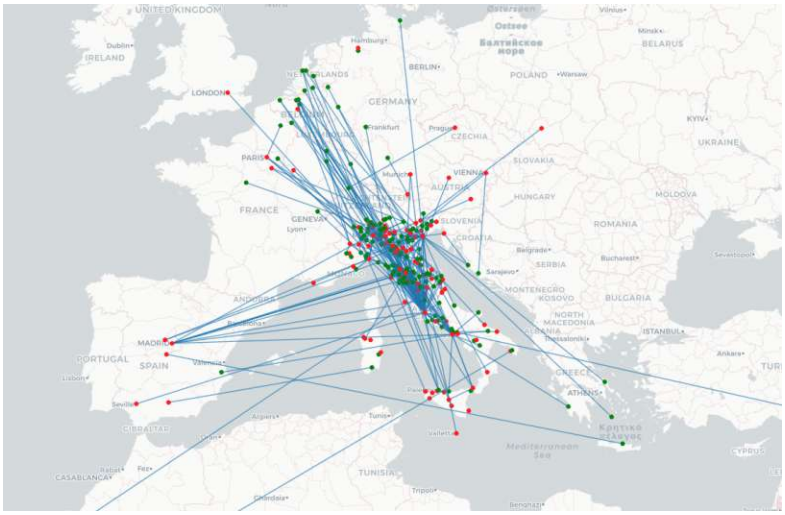


Figure 6.2: Flowmap

and place of death. This made it possible for the user to explore the data and get more information on the individual painters. In order to maintain readability, a consistent color scheme was used, blue lines representing the flows, green dots representing birth locations and red dots representing death locations, as can be seen in Figure 6.2.

Despite the limitations that flow maps pose, they are a powerful tool for visualizing trajectories and showing individual and aggregated patterns. For this thesis, they are part of the user study and are assessed for their clarity, informativeness and readability.

6.3 Sankey Diagram

Sankey diagrams are a specific type of flow diagram used to depict a flow from one set of values to another with the width of the flow being proportional to the quantity it represents[6]. It was developed in 1898 by Irish captain and engineer Riall Sankey to visualize the efficiency of steam engines by illustrating the flow of energy from the heat source to various parts of the engine [KENNEDY and SANKEY, 1898][7]. Since then, they have been widely used in a wide range of domains.

In the context of this thesis, Sankey diagrams are used to illustrate migration patterns of painters from their place of birth to their place of death. Sankey diagrams, in contrast to heatmaps and flow maps, do not use geographical maps, but rather abstract geographical locations into nodes and show the movement between them. This can be useful in the

[6] Sankey Diagram | Charts — Google for Developers, https://developers.google.com/chart/interactive/docs/gallery/sankey, Accessed on: 2025-07-25

[7] SankeyArt | Turn financial statements into beautiful Sankey diagrams — SankeyArt, https://www.sankeyart.com/content/blog/what-is-a-sankey-diagram/, Accessed on: 2025-07-25

context of recognizing patterns and general historical trends, such as cultural centers and hubs.

The data used in these visualizations stems from the data extracted from painters' biographies sourced from Wikipedia and Wikidata. Different data mining techniques were used and place of birth, place of death, date of birth and date of death were extracted. The data was aggregated by century and by quantity to provide a manageable overview. Special consideration was given to the uneven distribution of locations; many painters were born or died in places represented by only a small number of individuals. Including all of these locations would have led to visual clutter and would have significantly reduced the readability of the diagrams. To address this, locations with fewer than three associated individuals were grouped under a common category labeled "Other," thereby preserving overall structure while improving clarity.

Sankey diagrams were generated using a custom Python script using the pandas Python library and plotly. Plotly is an interactive, open-source and browser-based graphing library for Python built on top of the Python programming language [8].

For each century-specific dataset, the birth and death place names were first normalized (trimmed, lowercased and capitalized) to ensure consistency. As each node represents a location, in order to avoid visual clutter from sparsely occurring locations, places appearing fewer than a set threshold (FREQ_THRESHOLD = 3) were grouped into a common "Other" category. This abstraction helps to emphasize proportional relationships, though at the cost of losing geographic specificity. By grouping smaller categories under "Other", these diagrams also reveal how editorial biases in Wikipedia coverage may lead to the over-representation of major cultural centers.

For the layout, the nodes were defined by separating birth locations on the left and death locations on the right. Flows between grouped birth and death locations were then calculated using a group-by-aggregation, with each unique pair forming a directed flow. In order to differentiate between the flows different colors were used, an example of a Sankey diagram used in the user study can be seen in 6.3.

To ensure readability, interactive hover labels were implemented, containing the numerical values of the quantities represented by the flows.

Sankey diagrams have the advantage of numerical values, which can be presented and interpreted, whereas heatmaps and flow maps lack this ability. While Sankey diagrams have a strong advantage also in emphasizing comparative flows and proportions, they have their limitations. They lack the geographical map on which heatmaps and flow maps are built, which could be of great importance for understanding the geographical aspect of the migrations of painters. While flow maps show the trajectories of movement on a geographical map, Sankey diagrams render those locations to abstract nodes. This abstraction could prove difficult for users unfamiliar with the underlying geography. In addition to that, if there are large numbers of unique locations, the diagram can become

[8] plotly/plotly.py, https://github.com/plotly/plotly.py, Accessed on: 2025-11-20

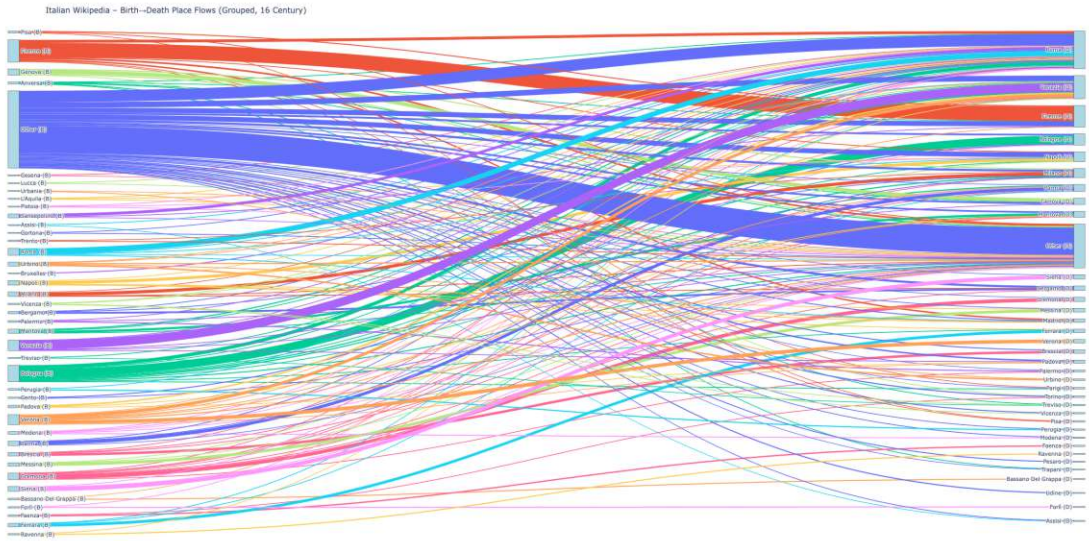


Figure 6.3: Sankey Diagram

visually overcrowded with the number of nodes and the number of thin flows that are present between them. To combat that, grouping can be done, but that makes the visualization lose the more granular information.

Nevertheless, in the context of this thesis, Sankey diagrams offer a valuable complementary perspective to heatmaps and flow maps. While heatmaps provide a static density-perception and flow maps highlight individual or multiple flows on a map, Sankey diagrams highlight directed relations, structure and amount, independent of map constraints.

CHAPTER 7

Results and Analysis of the User Study

7.1 User Study: Comprehension and Usability of Visualizations

After analyzing the individual visualization techniques from a theoretical standpoint and their advantages and limitations, we turn our attention to experimentally evaluating the interpretability, clarity and usability of these visualization methods for art history.

In order to evaluate this a user study was conducted. The aim of this study was to analyze how effectively each visualization converts data on artistic migration and which one would be best suited for this and similar datasets.

The user study was conducted in September of 2025 using Google Forms[1]. Each participant was presented with the visualizations of the same dataset and asked follow-up questions. Those questions can be divided into three overarching categories: factual questions, Likert-scale evaluation and open-ended questions. Each visualization technique used the same dataset that represented data from three centuries (15th, 16th and 18th), yielding a total of twelve visualizations (six heatmaps, three flow maps and three Sankey diagrams).

7.1.1 Structure of the User Questionnaire

The questionnaire was structured into three main parts, each focusing on one visualization type. In order to prevent participants from changing their answers based on other visualizations, participants were asked to focus on the indicated specific maps and not refer to the visualizations in other sections. Within each part, participants were first asked in the same sequence factual questions (for example, "Which city has the highest number of births in the 15th century?") to test comprehension and accuracy. Subsequently, the participants evaluated the visualizations using 5-point Likert scales, ranging from Strongly

[1] Google Forms — Google Workspace, https://workspace.google.com/products/forms/, Accessed on: 2025-11-12

Disagree (1) to Strongly Agree (5). Likert scale questions are questions used in surveys to assess the level of agreement or disagreement of participants with a series of statements[2]. The Likert scale questions were divided into several criteria in order to assess the overall feedback of the users:

- Ease of understanding
- Clarity of migration
- Pattern recognition
- Information load
- Engagement
- Further engagement
- Added value vs. text

After completing these questions, participants were asked open-ended questions in order to gain qualitative feedback on the strengths, weaknesses and potential improvements of the visualizations. Example questions included:

- "How would you improve the visualizations?"
- "Did any of the visualizations confuse or overwhelm you?"
- "Did the visualizations raise new questions for you?"

The full questionnaire is provided in the Appendix of this thesis.

7.2 Comparative Evaluation Based on KPIs

In order to have a structured comparison of the visualization methods, a set of Key Performance Indicators (KPIs) was defined. They serve as the basis for the evaluation and capture the quantitative and qualitative aspects of these visualizations.

The following KPIs were applied:

- Accuracy of interpretation: this was measured by the percentage of correct answers to factual questions for each visualization

[2]Likert Scales: Definition Questions, https://www.surveymonkey.com/mp/likert-scale/, Accessed on: 2025-10-21

- Clarity: this was measured by Likert-scale ratings on questions pertaining to ease of understanding and clarity of migration
- Pattern recognition: this was measured by Likert-scale ratings on questions pertaining to the ability to extract meaningful migration patterns
- Information load: this was measured by Likert-scale ratings on questions pertaining to information load (too little or too much information).
- Engagement: this was measured by Likert-scale ratings on questions for further exploration of the data
- Added value vs. text: this was measured by Likert-scale ratings on questions where participants assessed whether the visualization revealed insights not visible in tables or text.
- Overall preference and appeal: this was measured by comparative questions at the end of the survey.

The analysis using quantitative and qualitative methods. A quantitative analysis on factual accuracy expressed in percentages and Likert-scale evaluation for each visualization method and across multiple questions was used.The qualitative analysis was done on the open-ended questions in order to find potential areas of improvement, strengths and recurring challenges. This structured evaluation is the basis for determining the suitability of each visualization for representing this dataset and identifying the strengths and weaknesses of each visualization technique. These insights can be relevant for researchers in the field of art history and digital humanities when selecting appropriate visualization tools.

This analysis provides a basis for determining which visualization method is the most suitable for representing migration data and for identifying contexts in which each method excels. Such insights are particularly relevant for scholars in art history and the digital humanities, who must balance clarity, interpretability and analytical depth when selecting appropriate visualization tools.

7.3 Participant Overview

7.3.1 Sample and Consent

The questionnaire was completed by 80 participants. At the beginning of the questionnaire, participants were required to give their informed consent via a checkbox. Only after agreeing to the consent form could they participate in the study and fill out the questionnaire. The consent statement was:

"By checking this box, I confirm that I agree and understand that my responses will be used for the purpose of this user study. I acknowledge that my answers will be analyzed to gain insights and inform research findings. I am aware that my participation is voluntary,

my participation may be withdrawn at any point and all data collected will be handled responsibly in a way that protects my privacy."

Optionally, participants could provide their email address, and 62 participants did so. Email addresses were not used for analysis and were stored separately; they will be deleted after the completion of the thesis.

Age was reported by all 80 participants. Table 7.1 summarizes the distribution.

Table 7.1: Age distribution (n=80).

Age group	n	%
20-30 years	41	51.3%
30-40 years	33	41.3%
40-50 years	1	1.3%
50+ years	5	6.3%

7.3.2 Background in Art History

To support the analysis and the research question, the participants were asked if they had studied art history or taken university courses related to art history. This was done to distinguish between the two groups later. By doing so, we can analyze whether prior knowledge had an effect on the perception of the visualizations.

This item had 80 responses: 60 reported "No" and 20 reported "Yes". This split was used to define the two analysis groups used throughout the evaluation:

- Group A (art-history background): $n = 20$
- Group B (no art-history background): $n = 60$

7.4 Quantitative Results

7.4.1 Accuracy of Comprehension

The goal of the questions in the user study was to assess whether the user could answer the questions correctly after examining the individual visualizations. Heatmaps consistently outperformed Sankey diagrams and flow maps across all questions and participant groups. In the tables 7.2 and 7.3 the accuracy scores for each visualization technique per question are shown across the two participant groups.

As heatmaps outperformed both flow maps and Sankey diagrams, the conclusion is that heatmaps provided the shortest perceptual path. Users were able to map color/density directly to quantity without decoding edge widths (Sankey) or disentangling overplotted routes (flow).

Question	Flow Map	Sankey Diagram	Heatmap
Which city has the highest number of births in the 15th century?	78.95%	84.21%	100%
Which city appears most prominently in death locations but not in birth locations during the 16th century?	68.42%	84.21%	100%
Which city shows the highest number of deaths in the 18th century?	73.68%	78.95%	89.47%
Which century shows the widest geographic spread of birthplaces?	42.1%	42.1%	42.1%

Table 7.2: Group A : Accuracy rates across questions and visualization techniques

Question	Flow Map	Sankey Diagram	Heatmap
Which city has the highest number of births in the 15th century?	77.04%	85.24%	91.8%
Which city appears most prominently in death locations but not in birth locations during the 16th century?	77.04%	85.24%	90.16%
Which city shows the highest number of deaths in the 18th century?	78.7%	73.77%	90.16%
Which century shows the widest geographic spread of birthplaces?	40.9%	36.06%	34.43%

Table 7.3: Group B : Accuracy rates across questions and visualization techniques

- Question 1: Which city has the highest number of births in the 15th century?

  For group A: Heatmaps scored highest with an accuracy score of 100% , with Sankey diagrams following with 84.21% and flow map with a score of 78.95%.

  For group B: Heatmaps scored highest with an accuracy score of 91.8% with Sankey diagrams following with 85.24% and flow map with a score of 77.04%.

  Here we can see that Group A had higher scores, which can be attributed to having prior knowledge of the topic. The highest difference in accuracy can be seen with heatmaps, with a difference of 8.2%, while for Sankey diagrams and flow maps the difference is less than 2%.

- Question 2: Which city appears most prominently in death locations but not in birth locations during the 16th century?

  For group A: Heatmaps scored highest with an accuracy score of 100% , with Sankey diagrams following with 84.21% and flow map with a score of 68.42%.

For group B: Heatmaps scored highest with an accuracy score of 90.16% , with Sankey diagrams following with 85.24% and flow map with a score of 77.04%.

On this question, group B scored higher for the visualization flow map and Sankey Diagram, while group A scored higher with heatmaps.

- Question 3: Which city shows the highest number of deaths in the 18th century?

  For group A: Heatmaps scored highest with an accuracy score of 89.47% with Sankey diagrams following with 78.95% and flow map with a score of 73.68%.

  For group B: Heatmaps scored highest with an accuracy score of 90.16% with Sankey diagrams following with 73.77% and flow map with a score of 78.7%.

  For this question, group B scored higher on flow maps and heatmaps, while for Sankey diagrams, group A scored higher.

- Question 4: Which century shows the widest geographic spread of birthplaces?

  For group A: Heatmaps scored highest with an accuracy score of 42.1% with Sankey diagrams following with 42.1% and Flow map with a score of 42.1%.

  For group B: Heatmaps scored highest with an accuracy score of 40.9% with Sankey diagrams following with 36.06% and Flow map with a score of 34.43%.

  For this question, both group A and group B scored low scores. While group A scored higher, these results are indicative that none of these visualizations are optimal for showing geographical distribution.

Among both question 1 and question 2 heatmaps, the most accurate indicator that mapping color/density directly to quantity supports fast and correct lookup without decoding edge widths (Sankey diagrams) or disentangling overlapping routes (flow maps). For question 3, heatmaps had again the highest accuracy score, with close second-place variation between Sankey and flow maps across groups. For question 4, all three visualization techniques produced low accuracy, indicating that none adequately communicated the spatial distribution. While group A scored slightly better, the relative ordering of visualization effectiveness remained the same, which strengthens the generality of the finding.

7.4.2 Likert-Scale Evaluations

In the table 7.4 the specific statements and their corresponding criteria can be seen. The rating scale is from 1 to 5, with 1 being the lowest rating ("Strongly Disagree") and 5 the highest rating ("Strongly Agree"). For the question "How would you rate the amount of information presented?" the values were 1 ("too little information") and 5 ("too much information").

Again, as for the task-based accuracy questions, the analysis was done across group A and group B. An overview of the results can be seen in Figure 7.1 and 7.2.

Question	Criteria
This type of visualization was easy to understand.	Ease of understanding
It clearly represented migrations.	Clarity of migration
I was able to extract meaningful patterns of migration.	Pattern recognition
How would you rate the amount of information presented?	Information load
I would recommend this visualization for similar data.	Engagement
It made me curious to explore the data further.	Further engagement
It revealed information I wouldn't notice in a table or text.	Added value vs. text

Table 7.4: Question and criteria

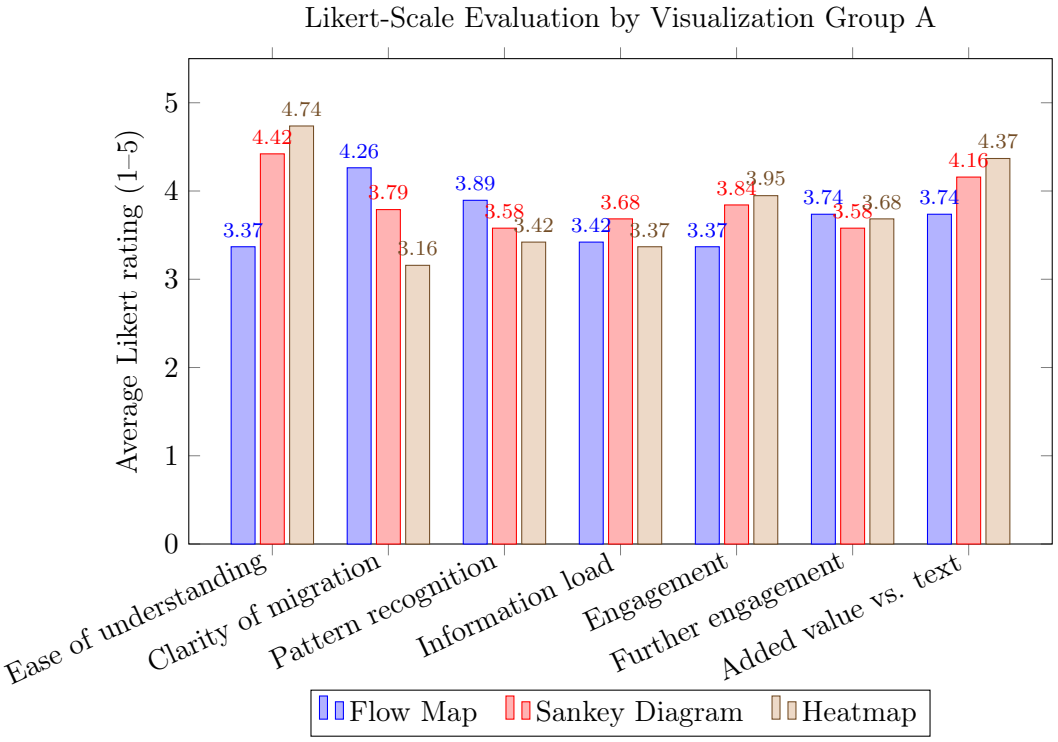


Figure 7.1: Group A : Average Likert ratings across the three visualizations .

Across both groups, heatmaps received the highest ratings for "Ease of understanding" (Group A: 4.74; Group B: 4.49) and "Engagement" (A: 3.95; B: 3.87). They also topped "Added value vs. text" in Group A (4.37), while flow maps scored slightly higher in Group B (3.74). However for "Clarity of migration", flow maps were rated clearest (A: 4.26; B: 3.92), exceeding both Sankey diagrams and heatmaps. The Sankey diagram scores were generally in the middle, with a slightly higher perceived "Information load" for Group A

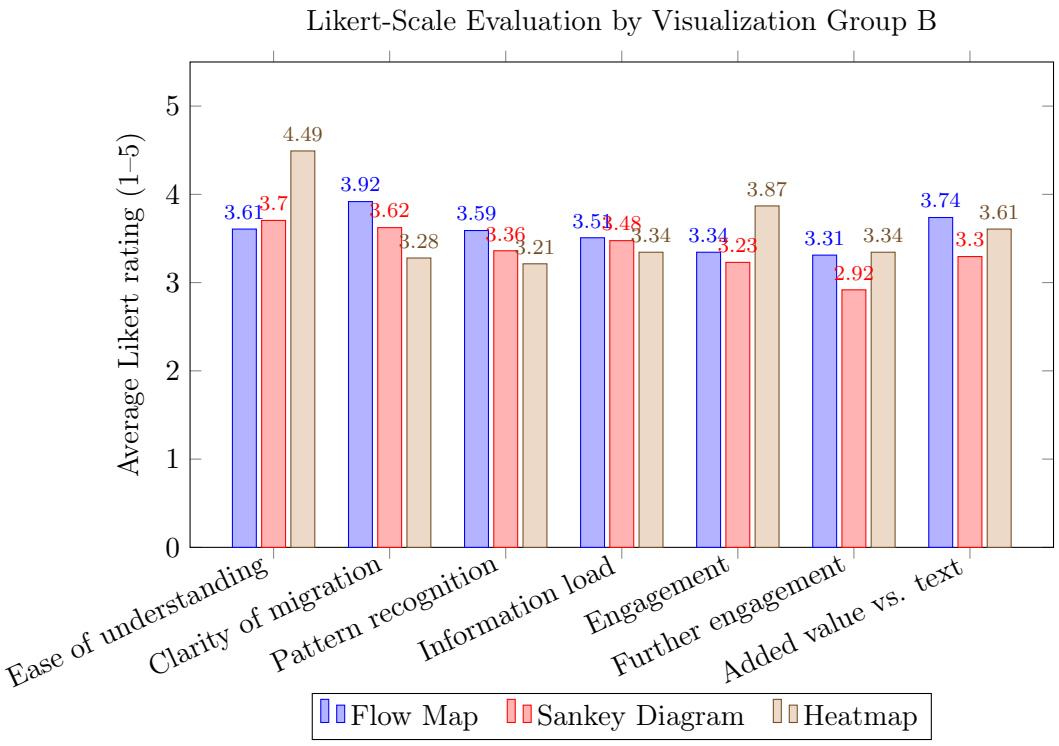


Figure 7.2: Group B: Average Likert ratings across the three visualizations

(3.68), thus indicating more visual effort than the alternatives.

For the criteria "Information load" across all three techniques, the results were the midpoint (Group A: 3.37-3.68 and Group B: 3.34-3.51), indicating that participants did not perceive substantial under- or overloading with information.

Across the two groups, the patterns of Likert-scale evaluation were stable. Group A ranked slightly higher "Added value vs. text" on heatmaps, whereas Group B assigned a narrow advantage to flow maps on that criterion. These differences are small and do not overturn the overall pattern.

The Likert scale evaluation complements the task accuracy results. Heatmaps were voted as both the most accurate and easiest to understand. This supports the task-based accuracy findings and reinforces their suitability for quantitative lookup and hot-spot detection. While flow maps scored highest across both groups for "clarity of migration" and "pattern recognition," they underperformed on accuracy. This can be attributed to overlapping routes and clustering of edges.

7.5 Comparative Evaluation

In this section, we focus on which visualization technique participants would choose in case they are faced with similar tasks. In the previous sections on accuracy and evaluation of the Likert-scale, the objective was to analyze task performance and perceived qualities. With this in mind, these results are complementary to each other to derive actionable guidance on visualization selection. The results can be seen in Tables 7.5 and 7.6

Question	Flow Map	Sankey Diagram	Heatmap
Which visualization did you find most informative?	15.79%	31.58%	52.63%
Which was most visually appealing?	26.31%	5.26%	68.42%
Which gave the clearest overview of the data?	31.58%	42.10%	26.32%
Which would you use to explain this data to a non-expert?	26.32%	10.53%	63.16%
Which visualization best reveals long-term patterns of painters' movement?	42.1%	26.32%	31.58%
Which visualization do you feel best highlights anomalies or surprising trends?	36.84%	21.05%	42.1%

Table 7.5: Group A: Comparative evaluation

Question	Flow Map	Sankey Diagram	Heatmap
Which visualization did you find most informative?	27.87%	31.15%	40.9%
Which was most visually appealing?	11.47%	19.67%	67.21%
Which gave the clearest overview of the data?	9.83%	42.62%	47.54%
Which would you use to explain this data to a non-expert?	21.31%	27.87%	50.82%
Which visualization best reveals long-term patterns of painters' movement?	49.18%	26.2%	24.59%
Which visualization do you feel best highlights anomalies or surprising trends?	19.67%	36.06%	44.26%

Table 7.6: Group B: Comparative evaluation

Heatmaps scored highest across both groups for the questions concerning informativeness, visual appeal and explaining the data to a non-expert. For the question regarding the evaluation of long-term patterns, flow maps were preferred by both groups (Group A: 42.10%; Group B: 49.18%). This aligns with the Likert finding that flow maps scored highest on "clarity of migration", even though they scored lower than heatmaps on accuracy. The result indicates that the paths provide a compelling narrative cue for directionality that users prefer to communicate about movement.

For the question on the clearest overview, the results differ across the two groups. Group A favors Sankey diagrams (42.10%), whereas Group B favors heatmaps (47.54%) with Sankey diagrams close behind (42.62%). This split is consistent with Sankey diagram's strength in summarizing aggregate relationships.

7.6 Qualitative Feedback

At the end of the questionnaire, there were 4 open-ended questions:

- How would you improve the visualizations?
- Did any of the visualizations confuse or overwhelm you? Please explain.
- Did the visualizations raise any new questions for you?
- Did you find that some visualizations helped you better answer specific questions? If yes, please elaborate.

The goal of these questions was to have a deeper understanding of the opinion of the participants and to gather possible suggestions for improvement. The answers were coded into subgroups based on recurrent motifs and a differentiation was made between the two participant groups. The answers were divided into two main groups are: "No remarks" and "Actionable improvements/General remarks". The group "No remarks" consisted of answers such as "No", "/" and "-". From this group, no further insights could be taken. The group "Actionable improvements/General remarks" was further analyzed and answers were thematically coded. Special consideration was given to the common topics and differences between the two participant groups.

7.6.1 How would you improve the visualizations?

Overall comments for all maps - Design, color, legends

Both participant groups suggested clearer legends and the use of fewer or different colors. Some examples of the comments "In the heatmap would choose different color palettes; Sankey diagram and Flowmap need to be simpler.","different selection of the colors" and "By making legends clearer, keeping colors consistent, and simplifying the design to emphasize main patterns.". Some noted difficulty of reading city names at small scales, but there could be an issue here present in case the user did not realize that the maps could be viewed in a separate window and zoomed in and out of.

Flow map suggested improvements

The most common comments were concerning overplotting and the difficulty in perceiving directions of flows and counts. Users suggested counts next to densely plotted areas and grouping of regions. Some examples of responses are: " In the flow diagram, I would

show the number of lines that merge into a dot at all times", "The flow diagram is useful with clear paths and colors but can feel cluttered; it could be improved by simplifying paths and grouping regions." and "Flow map: perhaps arrows indicating movement. It is not possible to tell if a given painter was born and died in the same location due to the overlapping dots.".

Sankey Diagram suggested improvements

Users appreciated Sankey diagrams for assessing volumes of painters but struggled with a quick overview in cases of many cities present. Some suggested ordering cities by volume, ranging from larger cities to smaller and highlighting matching nodes in case of hovering over a flow. Some noted that the lack of a geographical map made the reasoning more difficult and clustering by major and minor cities would have helped. Some examples of the comments received are: "Sankey diagram: perhaps ordering data from largest to smallest on either side.", " Using the same colors for the same places in the second visualization." "The sun was a lot of information in one spot and maybe too much to understand at once, so maybe split the data into major and minor cities or further cluster it."

Heatmaps suggested improvements

The heatmaps were most often praised for clarity, but users suggested counts in case of zooming in on the maps and the use of a different color palette. A subset of users suggested combining birth and death maps in a single view or creating animations. Some examples of the comments given: "Adding a toggle to switch between births and deaths would improve visibility. Also, providing a side-by-side view comparison would be a good idea." and "in the heatmap, would choose different color palettes". Combining several maps and adding context. Several users suggested combining different maps, for example, combining the heatmap and flow to pair spatial density with routes. In addition, short annotations and legends were suggested to help users not familiar with the data interpret key findings better. Some examples of the suggestions given: "add a one-line key insight per map" and "I would combine several different ways. And made it visually more beautiful, design more beautiful."

The suggestions given were clear and consistent and give a roadmap for possible improvements for the next iteration. The qualitative feedback aligned with the quantitative findings of prior questions. Incorporating the improvements is expected to improve the accuracy and perceived information load and strengthen the overall communicability for non-expert audiences.

Differences between the participant groups

7.6.2 Did any of the visualizations confuse or overwhelm you? Please explain.

The responses to this question were separated into two subgroups:

- "Yes, there was confusion"
- "No"

Additionally, the "Yes, there was confusion" group was coded based on the visualization techniques named and the cause of the confusion. Most frequently named as overwhelming and/or confusing was the Sankey diagram, followed by flow maps and in last place heatmaps which were rarely named.

Sankey diagram confusion

The Sankey diagram was mentioned the most in both group A and B as the most confusing visualization. Points that were mentioned several times were a lack of geographical context, visual density and unfamiliarity with the visualization technique. Example of the responses given are: "Sankey diagram, too much everything", "Yes, the Sankey diagram felt a bit overwhelming because of too many lines and colors." and "I'm not used to Sankey; it got busy fast.".

Flow map confusion

Both groups mentioned flow maps as confusing but in smaller numbers than Sankey diagrams. Most answers were regarding the overlapping of lines and nodes in densely populated areas, the direction of the lines not being obvious and the lack of numerical values. Examples of answers that illustrate these points are: "Flow map... not possible to answer without counting and tracking lines.", "Overlapping dots—needed a counter next to the city." and "Lines overlapped in northern Italy; without arrows direction wasn't obvious.".

Heatmaps confusion

Heatmaps were the least mentioned as confusing for both groups. For group A heatmaps, unlike the other maps, were never directly mentioned as the most confusing, only one answer was mentioning heatmaps: "Yes... All of these three approaches lack visual identity, and graphic beauty that would better communicate with the viewer.".

For group B, the main point mentioned was not regarding heatmaps as a visualization but rather of an "organizational" nature as in the heatmaps should be placed next to each other to better show the differences. Some of the examples of responses given are: "Heat map... not able to determine how migrations were flowing.", "Two heatmaps next to each other were a bit confusing if colors weren't aligned." and "Heatmap had me going back and forth a lot".

7.6.3 Did the visualizations raise any new questions for you?

The answers were classified into two groups :

- "Yes"
- "No"

The responses from the group that answered "Yes" were then further subdivided by visualization name and question. In the sections below, specific responses were shown and patterns were summarized.

For group A, approximately 78% of participants answered that they had no new questions, compared to group B were around 53% of participants of the user study who answered that the visualizations did not raise any further questions for them.

For group A, there were only 4 entries that answered yes, so the answers can also be shown individually:

- "Yes, I was trying to remember historical data about art history."
- "Yes - I would be very curious to see correlations with other socio-economic conditions of the prominent death places. Was it a highly developed economy? Access to artistic education or ease of finding "sponsors" that drove artists to these places? "
- "If the death places were also possibly the places where the artists were the most active during their life and work (I'm guessing yes) "
- "Yes, I was trying to remember historical data about art history."

For group B, some overarching patterns could be observed. The first point of interest was concerning the geographic and temporal migration patterns, with specific interest given to cities such as Rome, Florence and St. Petersburg. The second point of interest was concerned with the underlying socioeconomic and historical context behind those migrations. A smaller subgroup of participants questioned the data collection and visualization, an example response is: "Yes, the visualizations made me wonder more about the reasons behind certain flow patterns and how the data was collected."

Overall, these responses indicate that the visualizations communicated the data effectively but also stimulated contextual thinking, prompting participants to connect visual patterns with historical, geographi, and social dimensions of art history.

7.6.4 Did you find that some visualizations helped you better answer specific questions? If yes, please elaborate.

The answers were classified into two groups :

- "Yes, a visualization was named and/or question"
- "No, no visualization was named"

The responses from the group that answered "Yes" were then further subdivided by visualization name and question. In the sections below, specific responses were shown and patterns were summarized.

Heatmaps for a quick overview

Many participants wrote that heatmaps were most helpful for summarizing information at a glance and for answering questions such as "Which city has the most births/deaths?". Some of the responses given are: "Yes, the Heatmap seemed the easiest for finding specific answers.", "Yes. Migration better explained by flowmap however, heatmap showed better static info." and "Easiest to find the city and see the geographic spread; less clutter than the flow map.".

Flowmaps for migration patterns and overall movement

Participants responded that, for migrational patterns and understanding of directional movements, flow maps were the easiest to understand. Some examples of the exact responses are: "Migration better explained by flow map; heatmap showed better static info." , "the Sankey diagram was good for showing city trends, the heat map was good for showing regional trends". and "It was the easiest to answer questions related to the heatmap, although I think the flow map would be the best with some improvements.".

Sankey diagrams for exact counts

A subset of participants used Sankey diagrams for answering questions because of their exact counts and explicit birth-to-death pairings. Some examples of responses are: "Absolutely. Sankey and heatmaps made it very easy to see where the births and deaths happened." , "Sankey Diagram was the only one that gave exact numbers to the listed questions. . . ." and "Yes, Sankey Diagram helped to answer on questions with the most accuracy and Heatmap would be helpful if I had limited amount of time to provide answers and to notice a migration.".

Depending on the question

Multiple participants noted that, depending on the question asked, they would prefer different visualizations, highlighting the strengths and weaknesses of each. Examples of responses given: "Yes. Migration patterns -> flow map and Sankey. Heatmap -> snapshot information at a glance ", "The flow map is the most inconvenient for answering questions related to the number of places of birth and death, while for questions regarding migration I prefer flow and heat, and the sankey is the most useless to me. Considering everything, my 1st choice would be heat map and least preferred the sankey " and "Each visualization has its strengths; it all comes down to the question.".

Some participants found Sankey diagrams "easiest overall," while others rated Sankey "least useful". These differences typically referenced numbers on hover (favoring Sankey) versus geographic reasoning (favoring heatmap/flow).

These results are consistent with the quantitative results. Heatmaps were chosen as best for summarizing information, which is consistent with the highest accuracy on tasks and the highest rating "ease of understanding". Flow maps are preferred for understanding movement patterns that align with Likert scores on "clarity of migration" and "pattern recognition". Sankey diagrams were chosen as the most effective tehcnique for exact counts, but the lack of geographic context lowered their effectiveness.

Differences based on participant groups

Overall, the findings between groups were similar but some differences could be observed between the groups. Group A focused more on contextualizing the visualizations to historical knowledge, while group B focused on ease of use, visual legibility.

CHAPTER 8

Limitations

8.1 Bias Analysis

In chapter 5, the analysis of language and regional bias across Wikipedia Language editions was discussed. Through the use of KPIs such as Language Match Rate, Local Hero Score and representation in their native language bias was detected. Additionally, through qualitative indicators such as number of references, number of sections, number of categories and number of internal links, bias was detected in the structure of articles.

While the findings of this chapter show clear patterns of bias in Wikipedia's coverage of artists, several limitations must be acknowledged.

First, we discuss limitations regarding the scope and diversity of the data analyzed. The analysis was done on six Wikipedia editions (English, German, French, Spanish, Italian and Serbian). The editions were chosen to ensure we analyze both larger and smaller Wikipedia versions, but they are a small subset of the more than 300 language editions in Wikipedia. Additionally, the analysis was done on a subset of data available in those Wikipedia language editions, namely painters born between the 15th and 20th centuries. Other domains then art history may reveal different dynamics and bias.

Limitations regarding the data completeness, consistency and accuracy need to be taken into consideration as well. The extraction of biographical information, such as nationality, relied on often inconsistent, historical or inconsistently formatted data. Missing values were obtained through scraping of Wikipedia articles, which introduced another layer of possible inconsistency. The factual accuracy of the data was not analyzed and is not guaranteed in the scope of this thesis. The analysis of points such as reference counts and internal links was also done from a quantitative aspect; the quality of references was not considered.

These limitations indicate that while the results of this analysis of bias show that in this dataset, there is language and regional bias, they are not definitive or exhaustive

measurements of bias. Nevertheless, they provide a robust point for understanding how imbalances affect the visibility of artists and the construction of art-historical narratives in Wikipedia.

8.2 Visualization Assessment

The conducted user study has several limitations. The first limitation is the overall sample size of 80 participants and the uneven group distribution (20 participants with an art history background and 60 without). This uneven distribution limited the cross-group comparison, and the overall sample size limited the generalization of the findings.

The usage of Google Forms imposed constraints on how the visualizations were viewed, as Google Forms did not allow interaction and zooming. In order to mitigate this, a GitHub repository was used where each visualization was hosted individually and the links to each visualization were incorporated into the Google Form. An explanation on how to use the link and visual cues was implemented, but still some participants mentioned that they had issues reading city names or overlapping elements, which would suggest skill issues rather than flaws in the specific visualizations.

Another limitation is the number of different visualizations shown. As there are many different types of visualization, a subset of 3 visualization techniques was chosen based on how compatible they would be with the dataset and in agreement with the thesis supervisor. We focused on this limited number of visualizations as to keep the cognitive strain on the participants manageable and the time needed to answer the questionnaire reasonable. Including more visualization techniques would have induced more strain due to the prolonged duration needed to complete the questionnaire which could in turn negatively affect the results.

Lastly, the content focus was limited to painters from the Italian nationality group and to the 15th, 16th and 18th centuries. This was done to make the amount visualizations manageable as more data in combination with having to assess multiple visualizations could overload the participants.

CHAPTER 9

Conclusion and Future Research

The analysis of painter biographies across six Wikipedia language editions shows that regional and language biases significantly affect how artists and art history are represented in open-source databases. Through the use and analysis of several KPIs such as Local Hero Score, Language Match Rate (LMR) and Artist Coverage Across Languages in Wikipedia, the presence of painters in their "native" and "non-native" Wikipedia editions was explored. These KPIs were chosen as the Local Hero Score (LHS) captures the completeness of national coverage, the Language Match Rate (LMR) reflects the share of local versus foreign painters and Artist Coverage Across Languages in Wikipedia focuses on how widely painters appear across different Wikipedia languages. These KPIs are complementary to each other and show that across all editions painters were represented in higher numbers and more completely in their "native" Wikipedia edition. For example, in the German Wikipedia, 19152 (or 94.29%) of painters from the German nationality group are represented (compared across the six Wikipedia editions). This number significantly drops in other Wikipedia language editions, such as the French one, where only 3208 German painters are documented. The comparison with the Getty ULAN further reinforces that, while internationally recognized painters are represented, less prominent figures are underrepresented.

Representation bias is not limited to whether an artist is included or not, but also extends to how they are documented. Article quality was examined as a potential second layer of bias. In terms of article quality, indicators such as the number of references, internal links, categories and sections indicate how depth of content and visibility differ. The analysis of the metadata and the structure of the articles revealed significant differences between the language editions and the nationalities. Larger language editions such as English, French and German exhibit more extensive referencing, dense interlinking and a higher categorization count, which improves article visibility and discoverability. In contrast, smaller editions such as Italian, Serbian and Spanish often had, for example, lower referencing counts and less dense interlinking. This reflects local editorial choices

rather than standardized criteria. These differences reduce the comparability of painters' biographies across language editions and affect the use of Wikipedia as a reliable, cross-lingual dataset.

These findings indicate that language and regional bias exist. Imbalances in the editor community are also reflected in the content it produces. Often the editorial focus is shifted to documenting the own cultural heritage and internationally prominent figures. While the open structure of Wikipedia allows for the generation of large amounts of knowledge and global participation, it also perpetuates greater inequities in representation by privileging communities with stronger editorial communities and engagement. Such inequity influences the discoverability and the construction of art historical narratives in the digital age. Ultimately, language and regional bias in Wikipedia contribute to reinforcing the dominance of certain cultural traditions while limiting the visibility of others.

For the second research question, we looked at different visualization techniques and measured how effective they are in representing the distribution of artistic activity across Europe leveraging a user study. Through the use of qualitative and quantitative questions, the result of the user study offered a comprehensive overview of the visualization techniques (heatmaps, flow maps and Sankey diagrams) in the context of the specific dataset. Several findings emerged that show the strengths and weaknesses of the individual visualization methods and their usefulness and interpretability for the specific dataset.

Heatmaps outperformed other visualization techniques across several different criteria, most notably "accuracy of comprehension" and "perceived ease of understanding". Participants were able to correctly answer the factual questions, which were then underscored with the Likert scale question results. Through the use of color-density encoding, a direct mapping between location and quantity was able to be done. These findings underscore the use of color-based spatial representations for summarizing large-scale datasets on a geographical map.

Flow maps received the highest ratings for "clarity of migration" and "pattern recognition," which showed that participants found them most suitable for visualizing migrational patterns and movement. Nevertheless, flow maps also had limitations in the form of visual clutter and overlapping of paths due to the size of the dataset. Participants noted that there were issues due to overplotting, lack of directionality and suggested the implementation of additional numerical annotations to better visually represent the dataset.

Sankey diagrams yielded the most mixed results and showed the difference between the two participant groups most clearly. The clear depiction of the flow between birth and death places was praised alongside the clear numerical values. However, they were most frequently described in the qualitative feedback as confusing, particularly with participants which had no prior art historical knowledge. In summary, participants viewed Sankey diagrams as best for analytical tasks where specific numerical comparison needed to be done, rather than for providing a quick overview and intuitive exploration.

Across the two participant groups, the results revealed that prior knowledge influenced moderate preferences but did not change the overall ranking of the visualizations. The qualitative analysis underscores the quantitative results. In the answers to open-ended questions, participants frequently mentioned the need for improved legends, a different color schema and interactive features such as toggles. Some participants proposed hybrid approaches such as integrating elements of the flow map into heatmaps to create multi-layered visualization designs that support both overview and detail. Important to note is that when participants noted confusion, they reflected on the visualizations and not on the data itself.

9.1 Future Research

In this thesis, the regional and language bias in Wikipedia within the sphere of art history was examined. With the growing use of artificial intelligence (AI) an interesting direction for future work could be to measure how the biases that are present in databases affect the outputs of AI. Large language models (LLMs) and other AI systems are commonly trained on data from different databases such as Wikipedia. This has the consequence that biases that are present in databases are carried over and reproduced in the models. Future research could measure how these biases are reproduced and evaluate whether strategies for mitigating bias have a positive impact on reducing bias in such models. With AI increasingly being used for information gathering, the accurate and fair representation of data becomes not only a research challenge but an imperative.

Future research on the topic of visualization evaluation could focus on more interactive visualization techniques which could support a deeper understanding and user engagement. Based on the feedback from the user study, participants recommended the use of more interactive maps for exploring art history data. This idea can be expanded into an observational study on how participants respond to such visualizations and how the data is received and understood.

This thesis focused on databases such as Wikipedia, Wikidata and Getty ULAN but expanding this analysis to other databases could create a broader overview of bias present in art historical data and its effects. This could provide a deeper understanding of how art historical narratives are shaped and how they influence AI output.

CHAPTER 10

Appendices

10.1 Survey Questionnaire

In this section, the questions and images used in the survey questionnaire will be presented. As the questions are repetitive for each visualization, only once will the questions be shown, and the visualizations will be shown in the subsequent section. The questions were repeated for each visualization technique. At the end of the questionnaire, the comparative feedback section was asked just once.

10.1.1 Information about the Participant

Please provide your e-mail address: _____

Age (select one):
20–30 years
30–40 years
40–50 years
50+ years

Did you study art history or take any courses related to art history during your university studies?
Yes No

10.1.2 Comprehension Questions

1. Which city has the highest number of births in the 15th century? (select one)
Venice
Florence
Siena
Verona

2. Which city appears most prominently in death locations but not in birth locations during the 16th century? (select one)
Rome
Venice
Bologna
Naples

3. Which city shows the highest number of deaths in the 18th century? (select one)
Rome
Venice
Naples
Florence

4. Which century shows the widest geographic spread of birthplaces? (select one)
15th century
16th century
18th century

5. Is there a noticeable shift in birth places between the 15th and 18th centuries? If yes, please give a short explanation.

10.1.3 Perception & Usability (Likert Scale)

Please indicate your agreement with the following statements (1 = Strongly Disagree, 5 = Strongly Agree).

Statement	1	2	3	4	5
This type of visualization was easy to understand.					
It clearly represented migrations.					
I was able to extract meaningful patterns of migration.					
I would recommend this visualization for similar data.					
It made me curious to explore the data further.					
It revealed information I wouldn't notice in a table or text.					

Amount of Information (Likert Scale)

How would you rate the amount of information presented? (1 = Too little information, 5 = Too much information).

Amount of Information	1	2	3	4	5
Overall amount of information					

Comparative Feedback

The visualizations presented were one of the following types: Heatmap, Sankey Diagram, and Flow map. The aim is to evaluate the visual clarity and interpretability of each visualization technique.

1. Which visualization did you find most informative? (select one)
Heatmap
Sankey Diagram
Flow map

2. Which was most visually appealing? (select one)
Heatmap
Sankey Diagram
Flow map

3. Which gave the clearest overview of the data? (select one)
Heatmap
Sankey Diagram
Flow map

4. Which would you use to explain this data to a non-expert? (select one)
Heatmap
Sankey Diagram
Flow map

5. Which visualization best reveals long-term patterns of painters' movement? (select one)
Heatmap
Sankey Diagram
Flow map

6. Which visualization do you feel best highlights anomalies or surprising trends? (select one)
Heatmap
Sankey Diagram
Flow map

10.2 Visualizations

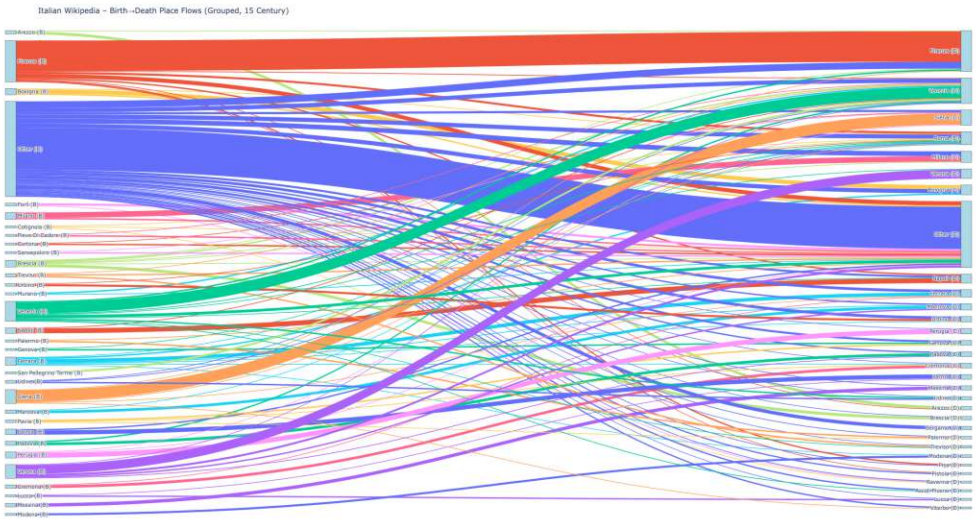


Figure 10.1: Sankey Diagram 15th Century

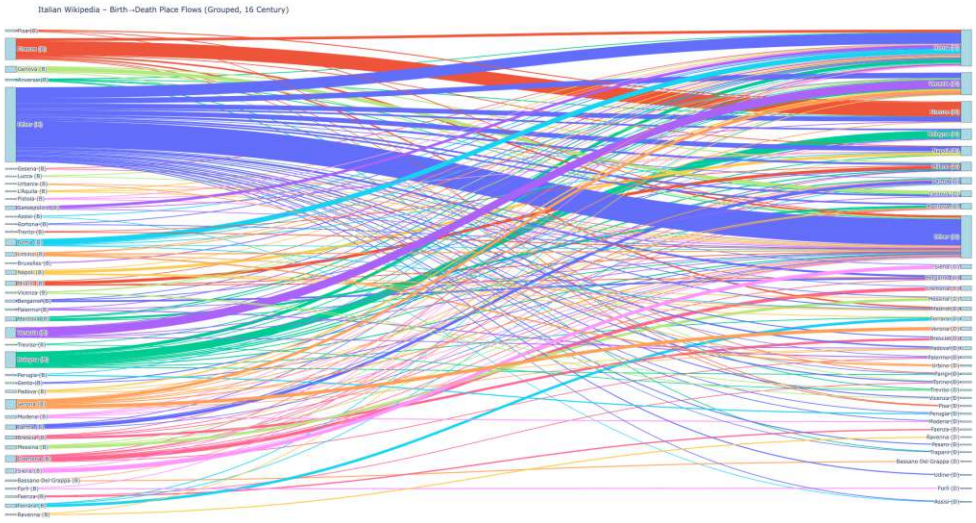


Figure 10.2: Sankey Diagram 16th Century

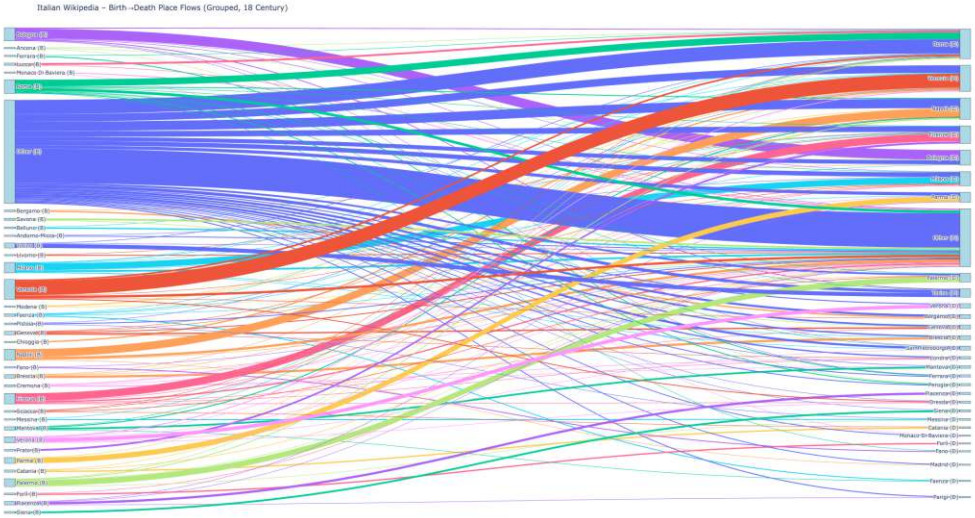


Figure 10.3: Sankey Diagram 18th Century

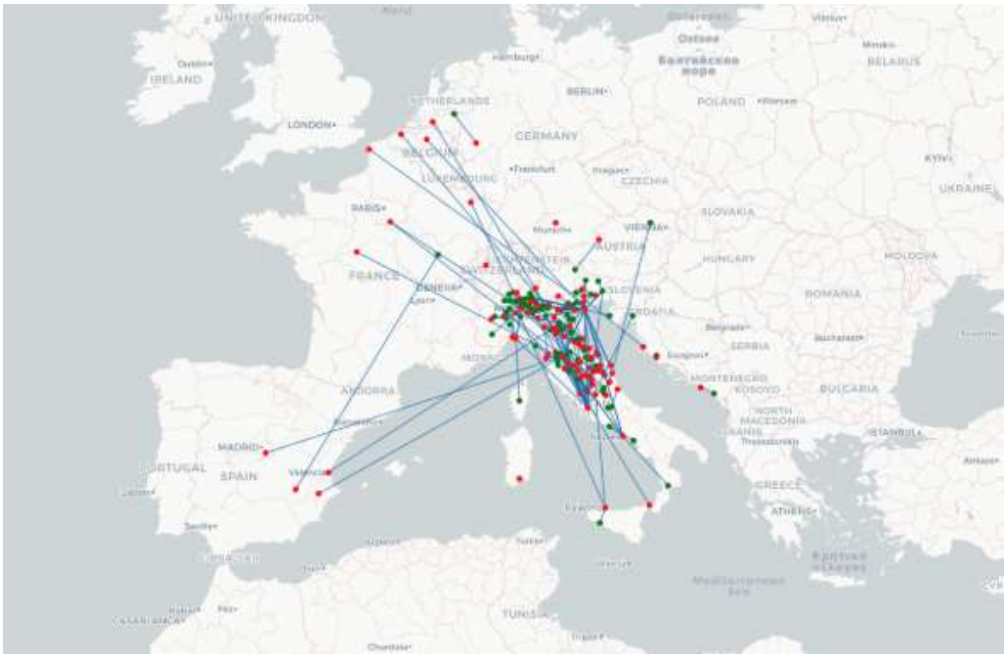


Figure 10.4: Flow Map 15th Century

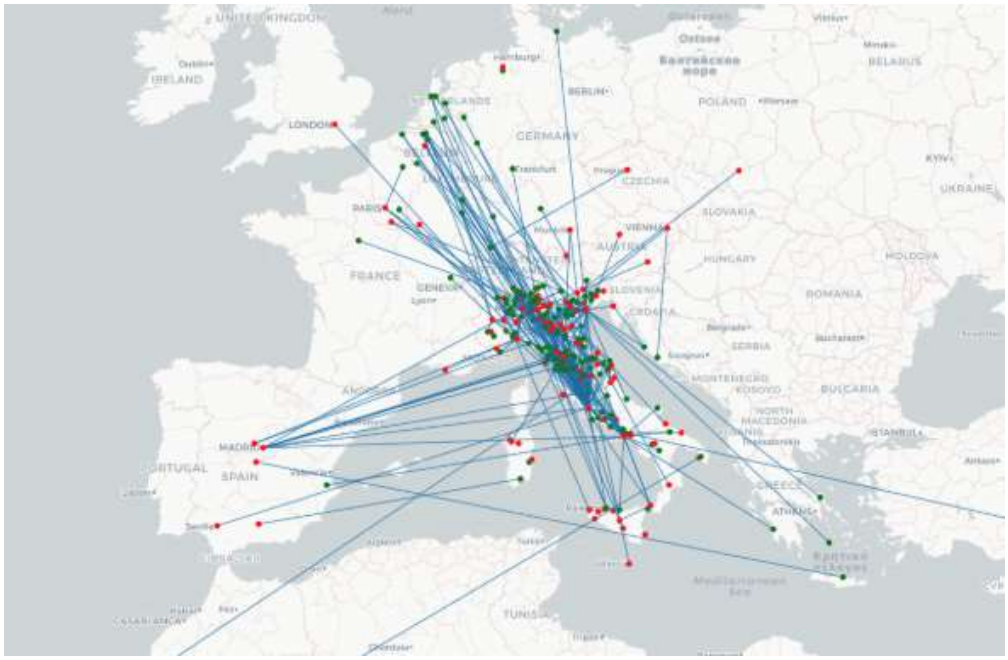


Figure 10.5: Flow Map 16th Century

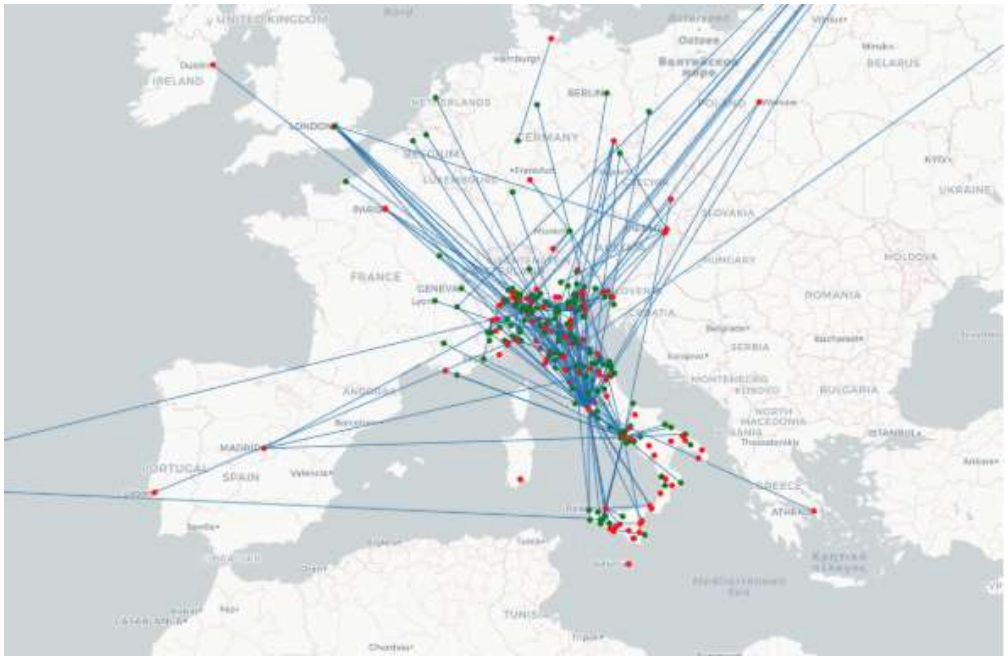
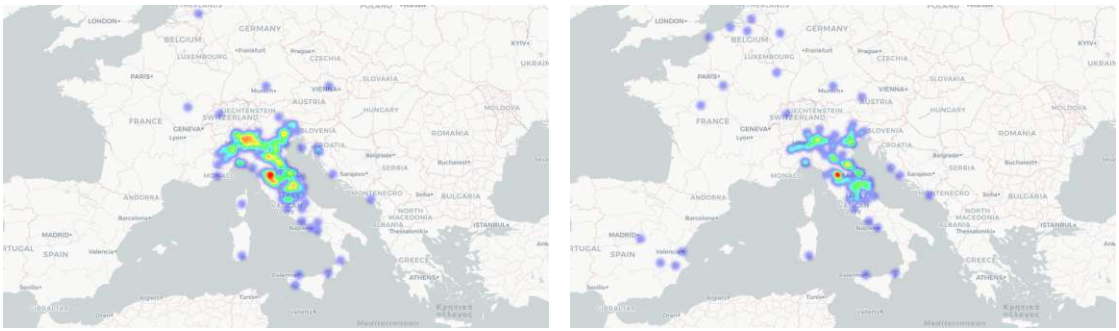
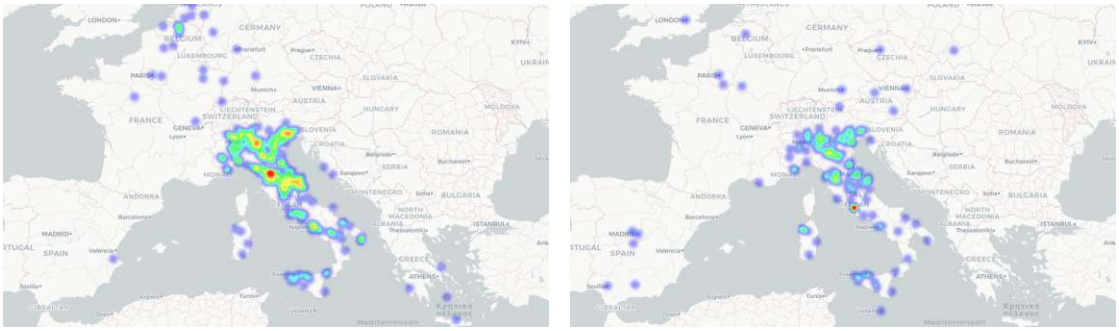


Figure 10.6: Flow Map 18th Century



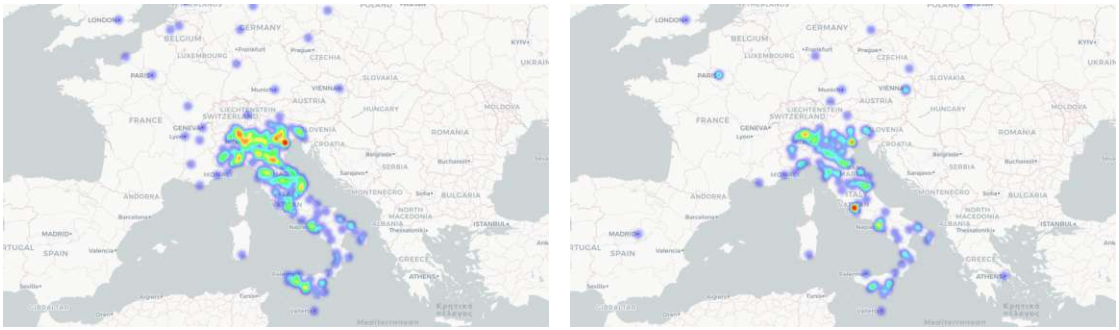
(a) 15th Century Birthplaces (b) 15th Century Death Places

Figure 10.7: Heatmaps 15th Century



(a) 16th Century Birthplaces (b) 16th Century Death Places

Figure 10.8: Heatmaps 16th Century



(a) 18th Century Birthplaces (b) 18th Century Death Places

Figure 10.9: Heatmaps 18th Century

Overview of Generative AI Tools Used

In the course of writing this thesis, several AI tools were used. All results and text that were received from such tools were reviewed and edited to ensure correctness and academic integrity. Grammarly[1] and ChatGPT[2] were used for grammar correction, refining academic tone, and formatting references. ChatGPT was also used as help in formatting tables and figures in LaTeX. For the theoretical part of the thesis, ChatGPT and perplexity.ai[3] were used for discovering new research papers; the content and validity of the papers were read and reviewed by the author of the thesis. In the data collection process and for the creation of the visualizations, ChatGPT was leveraged to assist in the correction of the code if needed, with an overview of the author. For the translation of the abstract from the english to the german language DeepL Translator[4] was used and subsequently checked by the author for any translation errors.

[1] Grammarly, https://www.grammarly.com/, Accessed on: 2025-11-13
[2] ChatGPT, https://chatgpt.com/, Accessed on: 2025-11-13
[3] Perplexity AI, https://www.perplexity.ai/, Accessed on: 2025-11-13
[4] DeepL Translator, https://www.deepl.com/en/translator, Accessed on: 2025-11-20

List of Figures

List of Tables

Bibliography

[Addlesee, 2019] Addlesee, A. (2019). Constructing SPARQL Queries. Wallscope.

[Ahmad et al., 2020] Ahmad, M. A., Teredesai, A., and Eckert, C. (2020). Fairness, accountability, transparency in AI at scale: lessons from national programs. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, page 690, New York, NY, USA. Association for Computing Machinery.

[Ashrafimoghari, 2023] Ashrafimoghari, V. (2023). Detecting Cross-Lingual Information Gaps in Wikipedia. In Companion Proceedings of the ACM Web Conference 2023, WWW '23 Companion, pages 581–585, New York, NY, USA. Association for Computing Machinery.

[Bishop, 2018] Bishop, C. (2018). Against digital art history. International Journal for Digital Art History.

[Bresciani and Eppler, 2015] Bresciani, S. and Eppler, M. (2015). Extending Tam to Information Visualization: A Framework for Evaluation. Electronic Journal of Information Systems Evaluation, 18(1):pp46-58–pp46-58.

[Briney, 2014] Briney, A. (2014). Overview of flow mapping. GIS Lounge.

[Brown, 2015] Brown, W. (2015). Undoing the Demos : Neoliberalism's Stealth Revolution. pages 1–296. Publisher: Princeton University Press.

[Callahan and Herring, 2011] Callahan, E. S. and Herring, S. C. (2011). Cultural bias in Wikipedia content on famous persons. J. Am. Soc. Inf. Sci. Technol., 62(10):1899–1915.

[Card et al., 1999] Card, S. K., Mackinlay, J., and Shneiderman, B. (1999). Readings in information visualization: using vision to think. Morgan Kaufmann.

[Chen, 2023] Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. Humanities and Social Sciences Communications, 10(1):567. Publisher: Palgrave.

[Collier and Bear, 2012] Collier, B. and Bear, J. (2012). Conflict, criticism, or confidence: an empirical examination of the gender gap in wikipedia contributions. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, CSCW '12, pages 383–392, New York, NY, USA. Association for Computing Machinery.

[Drucker, 2014] Drucker, J. (2014). Graphesis : visual forms of knowledge production. Cambridge, Massachusetts : Harvard University Press.

[Eom and Shepelyansky, 2013] Eom, Y.-H. and Shepelyansky, D. L. (2013). Highlighting Entanglement of Cultures via Ranking of Multilingual Wikipedia Articles. PLoS ONE, 8(10):e74554.

[Erxleben et al., 2014] Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., and Vrandečić, D. (2014). Introducing Wikidata to the Linked Data Web. In Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., and Goble, C., editors, The Semantic Web – ISWC 2014, pages 50–65, Cham. Springer International Publishing.

[Graells-Garrido et al., 2015] Graells-Garrido, E., Lalmas, M., and Menczer, F. (2015). First women, second sex: Gender bias in wikipedia. In Proceedings of the 26th ACM conference on hypertext & social media, pages 165–174.

[Gu, 2022] Gu, Z. (2022). Complex heatmap visualization. Imeta, 1(3):e43.

[Hall, 2013] Hall, G. (2013). Toward a Postdigital Humanities: Cultural Analytics and the Computational Turn to Data-Driven Scholarship. American Literature, 85(4):781–809.

[Harpring, 2010a] Harpring, P. (2010a). Development of the getty vocabularies: Aat, tgn, ulan, and cona. Art Documentation: Journal of the Art Libraries Society of North America, 29(1):67–72.

[Harpring, 2010b] Harpring, P. (2010b). Development of the Getty Vocabularies: AAT, TGN, ULAN, and CONA. Art Documentation: Journal of the Art Libraries Society of North America, 29(1):67–72. Publisher: The University of Chicago Press.

[Hern, 2015] Hern, A. (2015). Wikipedia's view of the world is written by the west. The Guardian.

[Hube, 2017] Hube, C. (2017). Bias in wikipedia. In Proceedings of the 26th International Conference on World Wide Web Companion, pages 717–721.

[KENNEDY and SANKEY, 1898] KENNEDY, A. B. W. and SANKEY, H. R. (1898). THE THERMAL EFFICIENCY OF STEAM ENGINES. REPORT OF THE COMMITTEE APPOINTED TO THE COUNCIL UPON THE SUBJECT OF THE DEFINITION OF A STANDARD OR STANDARDS OF THERMAL EFFICIENCY FOR STEAM ENGINES: WITH AN INTRODUCTORY NOTE. (INCLUDING APPENDIXES AND PLATE AT BACK OF VOLUME). Minutes of the Proceedings of the Institution of Civil Engineers, 134(1898):278–312.

[Kim et al., 2014] Kim, D., Son, S.-W., and Jeong, H. (2014). Large-Scale Quantitative Analysis of Painting Arts. Scientific Reports, 4(1):7370. Publisher: Nature Publishing Group.

[Lam et al., 2011] Lam, H., Bertini, E., Isenberg, P., Plaisant, C., and Carpendale, S. (2011). Empirical studies in information visualization: Seven scenarios. IEEE transactions on visualization and computer graphics, 18(9):1520–1536.

[Lamqaddam et al., 2018] Lamqaddam, H., Brosens, K., Truyen, F., Beerens, R. J., De Prekel, I., and Verbert, K. (2018). When the tech kids are running too fast: Data visualisation through the lens of art history research. IEEE Transactions on Visualization and Computer Graphics.

[Lee et al., 2020] Lee, B., Seo, M. K., Kim, D., Shin, I.-s., Schich, M., Jeong, H., and Han, S. K. (2020). Dissecting landscape art history with information theory. Proceedings of the National Academy of Sciences, 117(43):26580–26590. Publisher: Proceedings of the National Academy of Sciences.

[Lee et al., 2017] Lee, S., Kim, S.-H., and Kwon, B. C. (2017). VLAT: Development of a Visualization Literacy Assessment Test. IEEE Transactions on Visualization and Computer Graphics, 23(1):551–560.

[Lewoniewski, 2023] Lewoniewski, W. (2023). The Most Cited Scientific Information Sources in Wikipedia Articles Across Various Languages. Biblioteka, 36(27):269–294. Publisher: Uniwersytet Adama Mickiewicza.

[Lewoniewski et al., 2017a] Lewoniewski, W., Węcel, K., and Abramowicz, W. (2017a). Relative quality and popularity evaluation of multilingual wikipedia articles. In Informatics, volume 4, page 43. MDPI.

[Lewoniewski et al., 2017b] Lewoniewski, W., Węcel, K., and Abramowicz, W. (2017b). Relative Quality and Popularity Evaluation of Multilingual Wikipedia Articles. Informatics, 4(4):43. Publisher: Multidisciplinary Digital Publishing Institute.

[Lezcano Airaldi et al., 2025] Lezcano Airaldi, A., Irrazábal, E., and Diaz Pace, A. (2025). Best Practices and Evaluation Methods for Narrative Information Visualizations: A Systematic Review:. In Proceedings of the 20th International Conference on Evaluation of Novel Approaches to Software Engineering, pages 378–389, Porto, Portugal. SCITEPRESS - Science and Technology Publications.

[Lopez, 2021] Lopez, P. (2021). Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. Internet Policy Review, 10(4):1–29.

[Menghini et al., 2019] Menghini, C., Anagnostopoulos, A., and Upfal, E. (2019). Wikipedia Polarization and Its Effects on Navigation Paths. In 2019 IEEE International Conference on Big Data (Big Data), pages 6154–6156, Los Angeles, CA, USA. IEEE.

[Menghini et al., 2021] Menghini, C., Anagnostopoulos, A., and Upfal, E. (2021). How Inclusive Are Wikipedia's Hyperlinks in Articles Covering Polarizing Topics? In 2021 IEEE International Conference on Big Data (Big Data), pages 1300–1307. arXiv:2007.08197 [cs].

[Miquel-Ribé and Laniado, 2018] Miquel-Ribé, M. and Laniado, D. (2018). Wikipedia Culture Gap: Quantifying Content Imbalances Across 40 Language Editions. Frontiers in Physics, 6. Publisher: Frontiers.

[Mora-Cantallops et al., 2019] Mora-Cantallops, M., Sánchez-Alonso, S., and García-Barriocanal, E. (2019). A systematic literature review on Wikidata. Data Technologies and Applications, 53(3):250–268. Publisher: Emerald Publishing Limited.

[Munzner, 2025] Munzner, T. (2025). Visualization Analysis and Design. In Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Courses, SIGGRAPH Courses '25, pages 1–2, New York, NY, USA. Association for Computing Machinery.

[Nicholson et al., 2021] Nicholson, J. M., Uppala, A., Sieber, M., Grabitz, P., Mordaunt, M., and Rife, S. C. (2021). Measuring the quality of scientific references in Wikipedia: an analysis of more than 115M citations to over 800,000 scientific articles. The FEBS journal, 288(14):4242–4248.

[Odell et al., 2022] Odell, J., Lemus-Rojas, M., and Brys, L. (2022). Wikidata Data Model. Book Title: Wikidata for Scholarly Communication Librarianship Publisher: IUPUI University Library.

[Onwuegbuzie and Dickinson, 2008] Onwuegbuzie, A. and Dickinson, W. (2008). Mixed Methods Analysis and Information Visualization: Graphical Display for Effective Communication of Research Results. The Qualitative Report, 13(2):204–225.

[Perez, 2019] Perez, C. C. (2019). Invisible women: Data bias in a world designed for men. Abrams.

[Phan et al., 2005] Phan, D., Xiao, L., Yeh, R., Hanrahan, P., and Winograd, T. (2005). Flow map layout. In IEEE Symposium on Information Visualization, 2005. INFOVIS 2005., pages 219–224. ISSN: 1522-404X.

[Rapp, 2023] Rapp, A. (2023). Human–computer interaction. In Oxford Research Encyclopedia of Psychology.

[Roy et al., 2020] Roy, D., Bhatia, S., and Jain, P. (2020). A Topic-Aligned Multilingual Corpus of Wikipedia Articles for Studying Information Asymmetry in Low Resource Languages. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 2373–2380, Marseille, France. European Language Resources Association.

[Roy et al., 2022] Roy, D., Bhatia, S., and Jain, P. (2022). Information asymmetry in wikipedia across different languages: A statistical analysis. Journal of the Association for Information Science and Technology, 73(3):347–361.

[Schloss, 2025] Schloss, K. B. (2025). Perceptual and cognitive foundations of information visualization. Annual Review of Vision Science, 11(1):303–330.

[Schröer et al., 2021] Schröer, C., Kruse, F., and Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Computer Science, 181:526–534.

[Schwalbe, 2016] Schwalbe, M., editor (2016). Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop. National Academies Press, Washington, D.C.

[Sigaki et al., 2018] Sigaki, H. Y. D., Perc, M., and Ribeiro, H. V. (2018). History of art paintings through the lens of entropy and complexity. Proceedings of the National Academy of Sciences, 115(37):E8585–E8594. Publisher: Proceedings of the National Academy of Sciences.

[Srinivasan et al., 2015] Srinivasan, R., Rudolph, C., and Roy-Chowdhury, A. K. (2015). Computerized Face Recognition in Renaissance Portrait Art: A quantitative measure for identifying uncertain subjects in ancient portraits. IEEE Signal Processing Magazine, 32(4):85–94.

[Sun et al., 2021] Sun, X., Yin, H., Liu, B., Chen, H., Meng, Q., Han, W., and Cao, J. (2021). Multi-level Hyperedge Distillation for Social Linking Prediction on Sparsely Observed Networks. In Proceedings of the Web Conference 2021, WWW '21, pages 2934–2945, New York, NY, USA. Association for Computing Machinery.

[Surkemper et al., 2019] Surkemper, L., Klinke, H., and Underhill, J. (2019). International Journal for Digital Art History: Issue 3, 2018: Digital Space and Architecture. BoD – Books on Demand. Google-Books-ID: 74CIDwAAQBAJ.

[Sweller, 1988] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive science, 12(2):257–285.

[Tan et al., 1999] Tan, A.-H. et al. (1999). Text mining: The state of the art and the challenges. In Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases, volume 8, pages 65–70.

[Thelwall et al., 2019] Thelwall, M., Bailey, C., Tobin, C., and Bradshaw, N.-A. (2019). Gender differences in research areas, methods and topics: Can people and thing orientations explain the results? Journal of Informetrics, 13(1):149–169.

[Tuscher et al., 2025] Tuscher, M., Filipov, V., Kamencek, T., Rosenberg, R., and Miksch, S. (2025). Nodes, edges, and artistic wedges: A survey on network visualization in art history. In Computer Graphics Forum, page e70154. Wiley Online Library.

[Ugail et al., 2023] Ugail, H., Stork, D. G., Edwards, H., Seward, S. C., and Brooke, C. (2023). Deep transfer learning for visual analysis and attribution of paintings by Raphael. Heritage Science, 11(1):268. Publisher: Nature Publishing Group.

[Vincent and Hecht, 2021] Vincent, N. and Hecht, B. (2021). A Deeper Investigation of the Importance of Wikipedia Links to Search Engine Results. Proc. ACM Hum.-Comput. Interact., 5(CSCW1):4:1–4:15.

[Vrandečić et al., 2023] Vrandečić, D., Pintscher, L., and Krötzsch, M. (2023). Wikidata: The Making Of. In Companion Proceedings of the ACM Web Conference 2023, WWW '23 Companion, pages 615–624, New York, NY, USA. Association for Computing Machinery.

[Wagner et al., 2015] Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. (2015). It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. Proceedings of the International AAAI Conference on Web and Social Media, 9(1):454–463.

[Wagner et al., 2016] Wagner, C., Graells-Garrido, E., Garcia, D., and Menczer, F. (2016). Women through the glass ceiling: gender asymmetries in Wikipedia. EPJ Data Science, 5(1):1–24. Publisher: SpringerOpen.

[Windhager et al., 2019] Windhager, F., Federico, P., Schreder, G., Glinka, K., Dörk, M., Miksch, S., and Mayr, E. (2019). Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges. IEEE Transactions on Visualization and Computer Graphics, 25(6):2311–2330.

[Wirth and Hipp, 2000] Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, volume 1, pages 29–39. Manchester.

[Zhang and Terveen, 2021] Zhang, C. C. and Terveen, L. (2021). Quantifying the Gap: A Case Study of Wikidata Gender Disparities. In Proceedings of the 17th International Symposium on Open Collaboration, OpenSym '21, pages 1–12, New York, NY, USA. Association for Computing Machinery.