FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Professional Migration Analysis through Wikidata Linked Open Data

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Bc. Oana Gabriela Bălăceanu
Matrikelnummer 1328731

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Dr. Dieter Merkl

Wien, 2. September 2019

Oana Gabriela Bălăceanu	Dieter Merkl

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Professional Migration Analysis through Wikidata Linked Open Data

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Business Informatics

by

Bc. Oana Gabriela Bălăceanu
Registration Number 1328731

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Dr. Dieter Merkl

Vienna, 2nd September, 2019

Oana Gabriela Bălăceanu	Dieter Merkl

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

Erklärung zur Verfassung der Arbeit

Bc. Oana Gabriela Bălăceanu
Guglgasse 8, 3311A, 1110 Wien

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

Wien, 2. September 2019

Oana Gabriela Bălăceanu

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Dieter Merkl, who helped me find a research topic that is of great interest for me and with skillful guidance, innovative ideas and stoic patience oversaw my entire progress.

Secondly, I would also like to thank my family, friends, and coworkers for the support I received to make this thesis a reality. Special thanks go to my friend, Herwig Hollauf, who has given me essential feedback.

Kurzfassung

Der Schwerpunkt dieser Forschung liegt auf der Analyse der Migrationsmuster in Bezug auf Beruf, Geschlecht und Herkunftsland. Als Datensatz wurden die Linked Open Data angebotenen Quellen verwendet. Ziel ist es, 15 verschiedene Länder, innerhalb und außerhalb Europas, zu vergleichen. Dabei werden migrierte Entfernungen (kurze, mittlere und lange Entfernungen), Geschlechterverhältnisse, Tendenzen in das selbe Land zu migrieren sowie die Verbreitung und Internationalisierung als Kriterien genauer betrachtet. Der Zeitraum zwischen 500 v. Chr. und der Gegenwart wird gemeinsam mit dem Zeitraum zwischen 1945 und der Gegenwart analysiert, mit dem Ziel um zu ob die technologischen Fortschritte der Mobilität sowie der Kommunikation die Migrationsströme beeinflusst haben. Ich habe Wikidata als Datenquelle verwendet, das es auch Quelle für Wikipedia ist und über eine sehr hilfreiche Dokumentation und Unterstützung über Mitglieder verfügt.

Als Analysetechnik wurden gängige statistische Verfahren angewandt, bei denen bestimmte Daten gezielt abgerufen werden um Tendenzen durch Visualisierungsmethoden darzustellen und Ähnlichkeiten hervorzuheben.Die herangezogenen Daten umfassen die Gesamtzahlen der Entitäten, Migrationsdauern, Geburts-/ und Todes-Orte, Quell-/ und Ziel-Orte, Häufigkeiten pro Jahr sowie individuelle Informationen die eine Erstellung eines “Personalität” ermöglichen. Die als Ergebnis dieser Arbeit entwickelte Anwendung generiert auch zusätzliche Visualisierungsmittel, um den Einfluss von Diagrammen auf das Verständnis und die Gewinnung wertvoller Erkenntnisse aus Rohdaten hervorzuheben.

Das Ergebnis der Forschung hat bestätigt, dass es viele Ähnlichkeiten im Trend zwischen verschiedenen Ländern und Berufstätigkeiten bei der Migration gibt (z. B. Österreich zeigt Ähnlichkeiten zu Italien und Spanien, Deutschland zu Frankreich). Auch das Geschlecht weist eine starken Einfluss bei der Wahl einer großen Entfernung gegenüber der kleinen Entfernung auf. Eine überraschende Erkenntnis ist, dass die Menschen in den letzten Jahrzehnten deutlich sesshafter geworden sind. Weiters ist zu erkennen, dass es eine starke länderspezifische Tendenz besteht und dass in allen Fällen die Anzahl der Geburtsorte im Vergleich zur Anzahle der Todesart zurückgegangen ist.

Abstract

The focus of this research is analyzing the patterns of migration depending on the profession, gender, and country of origin while using the sources offered by Linked Open Data. The aim is to compare 15 different countries from both within and outside Europe in order to understand their distribution according to the distance transcended during migration (short, medium, medium-long, and long-distance), the gender ratios, the level of same-country migration, the spreading and degree internationalization. Time-wise, the interval between 500 BC and present has been analyzed together with the range between 1945 and present, in order to understand if recent technological advancements in traveling and communication have in any way impacted the migration flows. Due to its position as a source to Wikipedia and due to its growing community and documentation support, Wikidata was the Linked Open Data repository of choice as a data source.

The processing techniques applied to the analyzed data include typical statistical methods in which specific data is fetched, distributions are calculated, and trends are analyzed through visualization methods in order to find similarities. The fetched data includes data describing the total number of entities, total number per length of migration, total number of cities of birth and death, data on the source and destination locations, data count for each year and individual information in order to populate a "personality" profile. The application developed as a result of this thesis also generates additional visualization means in order to emphasize the impact that maps, graphs, charts, and diagrams have in understanding and extracting valuable insight from raw data.

The result of the research has confirmed that there are many similarities in trend between various countries and professions when migrating (e.g., Austria with Italy and Spain, Germany with France), that gender plays an active role when choosing long-distance over short-distance, that the last decades have, against intuition, made notable people more settled, that there is a strong country bias when fetching linked data, and that in all the cases, the number of places of birth is smaller than the number of places of death.

Contents

CHAPTER 1

Introduction

1.1 Motivation

Although many see it as a recent phenomenon, migration has been a part of human existence since its very beginning. Nowadays, people are easily impressed by figures like one million migrants coming to Europe in one year, but there have been many historical migrations that have far surpassed this number. For example, the largest voluntary migration has concerned Italy and involved a total of 13 million people leaving the country between 1880 and 1915 alone [Cho08]. When it comes to involuntary movement, a good example is given by the Second World War. During the Second World War, 90 million civilians moved to and around China [Buc19] and after the Second World War, around 8 million people left towards Germany from German-speaking areas in the Balkans, Czechoslovakia and Poland [Tsu09].

Humans have always moved in groups or as individuals for a wide range of reasons. Seeking freedom from war or persecution, escaping poverty or scarcity of resources, looking for new economic opportunities or being motivated by wanderlust and the desire of discovering new places, these have all been migration drivers during thousands of years, profoundly shaping the nature of the world we live in.

Currently, it is estimated that the number of people who have migrated to live and work in other countries has doubled from 99.8 million in 1980 to 200 million in 2005 [ES06]. When concentrating on professional migration only, it is hard to know the total number of immigrants at any given time, but some estimates place the numbers at around 1.5 million professionals from developing countries to the industrial ones alone [Sta00].

This situation is significantly affecting the regional and global dynamics, from economic to social, cultural, and political consequences. Currently, professional migration is one of the most active components across the world, leading to impactful phenomenons like

the “human capital flight” which brings together concepts like the “brain drain”, “brain circulation” and “brain gain” [DL05].

1.2 Problem Statement

Migration has always represented an important component in the global dynamics and history. Humans have crossed boundaries and territories ever since their early beginning, and this has had effects on the surroundings, first affecting the nature and the landscape, and afterward affecting the cultural, social, religious, political and economic aspects of the world.

Periodic mass migrations represent only the tip of the iceberg in migration studies. These large-scale movements are usually formed by an accumulation of smaller movements that fit into larger patterns with time [Isa07]. As a result, when exploring such broad processes, it is important to keep an eye on the individual experience in similar ways as on the overall pattern. Even when the number of migrants is small, their movement can have both negative and positive effects on a regional or global level. The exchange of languages, customs, and technology is an engine for innovation, as migration traces and connects different people and ideas. In addition, the movement of people accelerates the circulation of plants, animals, and infectious agents, which can eventually have impactful consequences.

As the population of the world increases at an alarming pace, and as people face a new level of freedom when communicating, traveling or changing careers, it is important to understand the potential and the effects of migration. If we can gather and conclude patterns based on previous migration data, we can bring some insight regarding the future of communities across the world. Migration leads to the development of innovations, but it also spreads past innovations, which sometimes results in the disappearance of languages and ways of life, as people come into contact. While there are some solutions in attempting to understand if the patterns differ in relation to being male or female, artist or engineer, born in Russia in 1750 or in Austria in 1970, it would usually require extensive literature study, as there isn’t a direct way on easily analyzing big sets of data in order to provide a general and clear conclusion.

In this field, data can offer a solid foundation, especially if used to gather insights through visualization. The production of information is exponentially increasing, and every day we are generating more, either from people going online or through sensors and devices. Recent estimates suggest that in 2020, the data production will be 44 times greater than in 2009 [KG11]. This can come with both advantages and disadvantages since the large quantity of data originated from different sources and presented under various formats and degrees of structure is affecting the way we gather, connect, filter, explore and use the data. Data creation is no longer one of the main problems, but instead, the identification of methods and techniques that we can use to turn the data into reliable and valuable knowledge, understandable not only to humans but also to machines [KSFN07][KAF+08].

There are many different types of data sources, as well as many possible methods and approaches for its processing. It is not completely clear if one data source is more suitable than others when analyzing dynamic processes like human migration, but they each have advantages and disadvantages. Highly structured data like Linked Data, one of the core concepts and pillars of the Web of Data, is also facing rapid growth and it is currently not possible to centralize and compute all the content into a single global repository. Even the largest existing public repository, Sindice, used to hold around 80 billion triples, which is just a fraction of the Linked Open Data Cloud [Bar13]. Richard Cyganiak and Anja Jentzsch have created the Linked Open Data cloud diagram[1] which describes the size magnitude of the current published open datasets on the Web [ABK+07]. Within the cloud, Wikidata constitutes one of the most important nodes, with 5,800,000,000 triples in Jan 2019[2].

The idea of a Semantic Web was introduced to a wider audience by Berners-Lee in 2001 [BLHL01]. According to his vision, the traditional Web as a Web of Documents should be extended to a Web of Data where not only documents and links between documents but any entity and any relation between entities can be represented on the Web.

While the Web of Data and the Linked Data approach offers a great set of benefits, according to Barbera [Bar13] we need a profound shift in the way data is produced, managed and disseminated in order for Linked Data to be consolidated and exploited at its full potential. Also, not all Linked Data is available for anyone to use and share, only Linked Open Data is data that can be freely used and distributed by anyone, subject only to the requirement to attribute and share-alike. One notable example for a Linked Open Data source is Wikidata, a crowd-sourced community effort that "acts as central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wikisource, and others" [3].

Currently, Wikidata offers Linked Open Data that covers a diverse range of aspects, from basic data on the birth and death of people to data on their profession, political and social circumstances (romantic relationships, family, influencers, associates, patrons, political views, etc.). This data, once collected and sanitized, can be used as a base for statistical analysis with the goal of revealing new insights. The result has potential in supplying important additional information for sociologists, governments, economists that can decide on applying laws and policies in order to better accommodate future population and professional demands and necessities.

This also comes as an aid to the usual approach on human migration research, which mostly relies on static sources consisting of files provided by a country's Government or Statistic Bureaus. Not only this but usually, as a migration trend becomes more prominent, these types of reports increase in number without offering a dynamic solution. For example,

[1] The Linked Open Data Cloud (2019): `http://lod-cloud.net/` accessed on 07.06.2019

[2] The Linked Open Data Cloud: Wikidata (2019): `https://lod-cloud.net/dataset/wikidata` accessed on 07.06.2019

[3] Wikidata Main Page (2018): `https://www.wikidata.org/wiki/Wikidata:Main_Page` accessed on 06.07.2019

even though the phenomenon on teachers migrating is not new, the number of reports have increased once this has become more pronounced in recent years [CLLVM14].

1.3 Aim of the Work

The present thesis is from several points of view a short study in world history. It addresses an extended period of time, it explores experiences drawn from many regions across the earth, and it emphasizes human interactions. Migration encourages one to think of connections, at least due to the fact that every individual that has died in a different place than the place of birth connects a point of origin and a point of destination. The extent in which these points vary come hand in hand with many factors, from the individual ones to the type and the internationalization of professions or the nature of the market and the involvement of the state.

The expected outcome of this study is an indication of the mutual influence between time period, profession, gender, and geographic dispersion. The thesis will answer the following research questions:

- Can we use Linked Open Data to analyze professional human migration patterns? What are the advantages, drawbacks, and challenges when using dynamic digital crowd-sourced data when analyzing social science topics such as migration?
- What is the degree of disparity when it comes to places of birth versus places of death? Can we offer a general comparative analysis on how migration is unfolding according to country, profession and gender?

For the analysis part, the thesis will mainly target the history of migrations for economic and professional reasons. The scope of the Master thesis will include:

- The time period starting with 500 BC until the present day interval with an emphasis on the interval beginning with 1500 AD. The migrations that start from this period may be seen as the first steps towards the creation of a world market for labor [Fis13];
- Profiles from people with a certain level of recognition and 'notability'. This includes persons that are the topic of a biographical article on Wikidata and are "worthy of notice" and "significant, interesting, or unusual enough to deserve attention or to be recorded "[4];
- Both female and male profiles that have a complete set of data (name, birth and death place and date, profession) compared while using an identical set of criteria (profession, region, time period);

[4]Notability (people) (2019): `https://en.wikipedia.org/wiki/Wikipedia:Notability_(people)` accessed on 07.06.2019

- Countries, professions, time periods and combination of these criteria that will return a result set of at least 100 entries;
- Various main theories and debates surrounding the subject of migration;
- Main historical events and periods chosen as a reference for criteria selection (e.g., early Silk Road and maritime trade in the Mediterranean and the Indian Ocean, the emergence of industry between 1700 and 1900, movement to cities, refugees, and diaspora);
- Possibility of looking beyond mass migrations, linking large and small migrations by offering information about the individual perspective.

Among the main components that will not be covered or analyzed in this thesis, we can enumerate:

- The earliest human migrations, including the earliest hominids, their development, and spread, due to the fact that data from that period is not offered in a consistent manner in Wikidata;
- A profound analysis of the historical events behind human migration and the theoretical background surrounding it.

The result will be provided in the form of a data visualization application that provides comparative study focusing on the various criteria mentioned before (profession, gender, time period, and geographical region). The outcome will show how and why some professionals are more prone to migrating on longer distance than others.

The conclusions will be based on the outcomes obtained from a program developed in the JavaScript programming language while using SPARQL as a method of getting the needed data for the statistical analysis. The program can fetch data published in Wikidata, with the collected information serving as an input for creating different diagrams and visual methods of analyzing data using the d3.js library. This approach is motivated by the large amount of available data and the advantages of using these mechanisms to easily extract, read, and draw conclusions [NSS16].

CHAPTER 2

State of the Art

The usage of technology in analyzing different humanistic aspects of the world has been a common practice, but using it for the analysis and prediction of migration specifically has not been very common and is lacking in literature. The majority of research in the field is mostly focusing on theoretical aspects and is gravitating around a specific topic without offering something general and universal.

From a macro perspective, the current situation in the migration field research shows the existence of two separate areas of literature.

- The first area lays the foundations for the theoretical work and knowledge with regards to the history and the patterns of migration. Since this area is currently the most extensive, I will follow the main research and concepts available on the history of human migration, and afterward, the literature on the patterns and typologies that characterizes it. Usually, studies are very specific and include topics like temporarily migration or targeting a particular country, period, type of migration, various data sources, or their benefits and shortcomings.
- The second part deals with the information technology area intersecting the theoretical and sociological aspects of human movements. When referring to information technology, I take into consideration the data sources used for the statistical analysis and the tools used to analyze and visualize the input data.

2.1 Theoretical Aspects on Human Migration

Global migration history and theory represent an extensive and diverse domain and constitute a considerable percentage of the research and literature available at the moment on the topic. The topic is so frequently studied that literature that analyses the research on migration is quite common, with both paperwork on very early migration[GH03]

or on very short recent periods of time[Man15]. Besides scientific papers, there are also numerous books and encyclopedias describing the phenomenon itself from its early beginnings.

A substantial ratio in the state-of-the-art revolves around the definition and the history of migration, while others treat topics like migration typologies, patterns, phenomenons, or consequences. In the present study, all the previous mentioned components will be investigated and presented. Among the types of research not covered are topics related to the "immobility paradox" (in which the emphasis is put on why people stay in their country of birth given the social and financial inequalities they are facing [Kin12]), the dispersion and the status of the 1.5 generations (which are formed by people who migrated as children and adolescents [EGW06]), auxiliary components of migrations like the migration network (represented by chains of people who facilitate the migrant's movement and their settlement) or the migration consequences, its costs and returns [Sja62].

While defining migration can be self-explanatory and usually literature doesn't necessarily include or extend this part, the best way to provide a complete definition is consulting major institutions and organizations that deal with this phenomenon on a continuous basis. According to the United Nations (UN), "While there is no formal legal definition of an international migrant, most experts agree that an international migrant is someone who changes his or her country of usual residence, irrespective of the reason for migration or legal status. Generally, a distinction is made between short-term or temporary migration, covering movements with a duration between three and 12 months, and long-term or permanent migration, referring to a change of country of residence for a duration of one year or more." [1].

Since the focus of the thesis is put on professional migration, and since skilled migration represents an increasingly significant component of global migration streams, a mention on the definition of highly skilled workers should be added. According to [Ire01] "having a university degree or extensive/equivalent experience in a given field" represents one of the main criteria.

Also in connection to the definition of migration, the subject of temporary versus permanent migration has been particularly studied, including in [Ire01] where the author analyses how governments respond to different types of migrant workers in order to protect the local jobs. With the exception of a few job sectors, there aren't many countries taking highly skilled professionals on a permanent basis, as the majority seek them on a temporary basis, in order to meet skills shortages until they can train the local skilled workers. Other countries utilize skilled migrants to generally improve the "stock" of brains.

[Ire01] adds that although there are still inequities in the treatment of permanent migrants, and there is a significant global move towards temporary skilled migration due to the

[1] UN Refugees and Migrants Definitions (2019): `https://refugeesmigrants.un.org/definitions` accessed on 07.06.2019

low cost of acquiring highly skilled workers, some professional labor markets can be described as becoming truly international as training, accreditation, ethics, and standards start to be recognized across the globe. This includes especially occupations in the IT industry as the universality of skills make issues like culture and social aspects not so important. Nursing is another example that is also becoming a more universal occupation as shortages in countries like Japan, US, and Canada lead to Filipino and other Asian nurses being offered work visas.

When it comes to presenting the history of migration, a notable example in this regards is given by [Man12], who in his book "Migration in World History", offers a detailed overview of both the history and the patterns surrounding the phenomenon.

Regarding the historical component, [Man15] refers to the migration timeline as formed by "remarkable stories of migration in every era". While the analysis covers much more distant times, the emphasis is put on the more recent time periods: in the ancient period, the author concentrates mostly on the Greek sailors (that created thriving commercial colonies along the edges of the Mediterranean and Black Seas) and on the Roman Empire (which was able to absorb the surrounding lands into a vast empire). Until that point, "land-based migrations gradually changed the culture of two major subcontinents: people speaking Indo-European languages spread from Central Asia into Iran and northern India, and people speaking Bantu languages moved from what is now Nigeria and Cameroon to many regions of central, eastern, and southern Africa".

[Man12] continues with Columbus' voyage, which was the critical event in the history of migration, since in the three centuries after that, "around two million settlers crossed the Atlantic from Europe to settle in the Americas. In the same three centuries, nearly eight million Africans were brought to the Americas, most of them in slavery". In the nineteenth and early twentieth centuries, improved transportation systems and new economic incentives enabled around 50 million European migrants to move across their home continent and afterward to both North and South America and beyond. In the same time period, another eighty million migrants moved across East and South Asia, repopulating regions from the Indian Ocean to Central Asia.

Besides a short historical presentation, [Man12] also brings into discussions several challenges faced when it comes to studying immigration. An important aspect when defining and tracking migration is understanding that people are usually accustomed to identifying communities as nations and ethnic groups. The issue with this classification is that the nations in which people claim citizenship are generally at most, a few hundred years old, similar to the ethnic groups, which although sometimes older, they also change often. Oppositely, tracking migration through language can lead through several thousand years back. Language is also important from the immigrant point of view as they must go through the effort and process of learning a new language and new customs, on the road and especially in the new community. This process, known as "seasoning," "socialization," or "acculturation," is an essential step in the successful completion of any act of migration.

Similar migration history studies have been done by [Bac18] who, in his book, has created a global overview of the whole global phenomenon or by [Bad08] who limits his analysis to Europe only. Emphasize is put on what many historians refer to as an "age of mass migration" that has occurred between 1850 and 1913 [ABE12] in which long distance migration happened at an unprecedented high rate. Some factors that led to this phenomenon were the cost of migration decreasing substantially, benefits of migration like wages increasing or the open border regime. This regarded especially the voluntary transatlantic migration of Europeans towards the Americas, but it is argued that the term should also include the other mass migrations that occurred in the same period in Asia.

Also surrounding the people's main reasons behind migrations, [Man12] classifies, and enumerates the following:

- Hope on improving personal situation by escaping an unhappy situation brought by social oppression and economic deprivation or by achieving a higher status;
- Bringing benefit to community or family by retrieving needed resources, learning new skills, or bringing back help. In recent times this might have meant sending cash home, but in earlier times it would be hunting or retrieving objects and food;
- Being a Samaritan by benefiting outside communities. This includes religious missionaries who move to new communities with the desire of spreading their faith;
- Voyaging for the pleasure of learning.

One paper also concentrating on Europe migration in the postmodern societies (1500-1900) [LL09], follows Patrick Manning's [Man12] theories, and while it focuses on illustrating the regional variation across time periods, it also classifies the migration forms into six categories, with Europe as a point of reference:

- Emigration - migration from European to non-European destinations, including colonial migration;
- Immigration - migration from non-European to European destinations;
- Colonization - settlement in 'empty' or sparsely populated spaces within Europe;
- migration to cities - movements to cities of over 10,000 inhabitants, mainly from the countryside;
- Migratory labor - the seasonal migration;
- Labour migration - migration of sailors and soldiers.

Gender and family in migration has also been a frequent subject in today's literature with examples like [Moc07] offering a general perspective on family patterns and demography in the global context or [AA00] and [Haw99] who concentrate on the typologies and patterns of female migration specific to a particular country (in these cases Bangladesh and Australia, respectively). [KTFW04] presents migration from the point of view and the consequences it has on a variety range of social groups - from a gender perspective to how families and the future generations deal with the process.

A more labor-oriented approach is presented in [Kje05], where Kjeldstadli offers a brief categorization of immigrants in a timeline-like manner. According to him, the first form of labor-related migration that arose was voluntary migration in which colonization from different centers to the peripheries took place. Examples in this sense are the Portuguese trading stations in East Africa, India and Southeast Asia, the so-called old immigration in the US or settler colonies in Africa. Kjeldstadli continues with the other forms of labor mobility represented by slavery, indentured or contract, and finally, economic migration.

In the case of slavery, the laborer and his or her labor power were the property of other men. Slavery was common between the 16th and the 19th century with a peak period in the 18th century, and the regions in which it occurred gravitated around western parts of the Arab areas, South America, the Caribbean and the southern states in the US. In contract labor, labor mobility got closer to a legal status. Although there was no room for negotiations, the system mobilized more work than the slave trade and was a common practice in the then British Commonwealth (especially with India as a labor source). Labor migrants are closely connected to industrial capitalism with the main difference compared to the previous two types being the fact that people are free to sell their labor power after negotiations and agreements.

Other types of bound workers are pointed by sociologist Robin Cohen [Coh16] as serfdom, debt bondage, apprentice labor, child labor, contract, and penal labor, forms of domestic service and concentration camp labor.

[KTFW04] introduces several time-related concepts that have great implications in migration. The first concept, 'time geography' is mainly associated with Hägerstrand who believed that the criteria for good social science 'are not to be found along the spatial cross-section but along the time-axis and in the particular sequence of events which makes up the life of each individual human being.' [Häg75]. [Pre77] added to these, the concept of "time-path", which is 'a weaving dance through time-space' from birth till death, although many of Hägerstrand's plottings of these life-paths were micro-temporal – a day, a week, a year. Fixed points (home, workplace, community center) are where individuals meet to form groups for a particular purpose, which might be related to a longer-term project such as creating a family, sustaining a livelihood, building a house or educating children. Such projects are dependent on the time, place, and individuals' relations with each other and with the structures of authority. Next, Hägerstrand draws attention to three kinds of time-geographical constraints that condition people's abilities and opportunities to carry out various activities and projects, including migration.

- Capability constraints, where the individual lacks the physical, financial, and social means to realize certain acts. For migrants, distance and travel costs are some of the more obvious examples;
- Coupling constraints, where the individual cannot move abroad because of personal or family obligations;
- Steering constraints where the individual cannot move because of mechanisms created with the intention of blocking access to migration (e.g., immigration laws).

Another interesting concept in the temporal aspect of migration is presented by [Cwe01], who shows a categorization for a long-term temporal outlook of the migrant experience:

- Strange times. Immigrants arrive with their own temporal baggage and some of which have to conform to the socio-temporal organization of life in the host society;
- Asynchronous times. In the past, distance and the slow speed of travel and communication created 'time rifts' between immigrants and their homeland. Immigrants engage a series of strategies to keep in touch with their country of origin via newspapers and magazines, videos, letters, telephone calls, satellite TV, email, and the Internet;
- Liminal times. For many migrants, the nature of their migration is seen as temporary and transitional; they are in a constant state of indecision;
- Diasporic times represent the times of long-term settlement; they thrive when immigrant communities recreate, to some extent, the rhythms of the social life of the homeland in the host society;
- Nomadic times. Between the liminal and the diasporic, migrants are seen as the bearers of new time conceptualizations and practices; as 'time pioneers' who are able to problematize and challenge dominant temporal constructs and devise new ways of thinking about and using time.

Continuing with the second component of the theoretical migration, the patterns of human migration represent the underlying logic, the recurring choices, and the interacting factors characterizing the phenomenon [Man12]. Although each migratory movement has its unique conditions and experiences, the underlying habits of human behavior make it possible to generalize migrations while at the same time emphasizing their distinctiveness. Also, while these changes remain based on certain fundamental habits, shared for all human history, the character of human history in every age differs from that in the period before, which also gives distinct aspects to each period.

Manning offers a categorization of migration while assessing their similarities with the migration patterns for other mammals and connects each category of migrants with a demographic profile.

The first category represented by the home-community migration involves the movement of individuals from one place to another within the home community. In other words, the offspring of one family moves to another family to find mates. Home-community migration is necessary for the reproduction of the species in order to maintain a sufficiently large genetic pool. Most humans experience home-community migration, as they start new families of their own. Home-community migrants are mostly female, as males are more likely to stay attached to the households of their parents.

Colonization is the departure of individuals from one community to establish a new community that replicates the home community. This type of migration is the primary means by which an animal species extends its geographic range: it involves moving into unoccupied territory or expelling previous occupants. Usually, the colonists settle in an environment very similar to that of their home community, and thereby maintain the same style of life. Colonists are more often male than female and are dominantly young adults.

Whole-community migration is the displacement of all the members of a community. Some species migrate habitually, usually in an annual cycle enabling them to complete their life cycle. Humans do not have a universal pattern of whole-migration, but some nomadic communities do migrate by adopting the habits of the animals they have come to dominate. Another trend is that whole communities may migrate in order to flee a natural disaster such as famine or a human disaster such as expulsion from their homeland.

Cross-community migration consists of selected individuals and groups, leaving one community and moving to join another community. This pattern is followed universally by humans, and rarely by other species since language provides the basic reason for this distinct pattern of migration. The migrants leave their home communities for various reasons – to benefit the home community, to benefit themselves, to escape the home community, or because they have been forcibly removed. Cross-community migrants are generally small in number and mostly consisting of young male adults. While the reality of cross-community migration is often complex, there is a simple typology of four commonly used terms to summarize this type of migration:

- Settlers are people who move in order to join an existent community that is different from their own, with the intention of staying at their destination;
- Sojourners are those moving to a new community, usually for a specific purpose, with the intention of returning to their home community;
- Itinerants move from community to community, but who have no single home to which they expect to return;
- Invaders arrive as a group in a community with the objective of seizing control rather than joining.

Cross-community migration influences every aspect of human migration because it creates and spreads changes. While home-community migration serves biological reproduction, cross-community migration enables a division of labor, spreads new technology, languages and encourages whole-community migration, by building connections between nomadic and settled populations. [Man12] also defines development as the complex process of transformation in human society. Ultimately the community might encounter limits on its resources, either as its population expands or as its resources shrink. This challenge might lead to out-migration of some of the deprived, taking their knowledge and skills to other communities. Another consequence is that the scarcity of resources might influence remaining members of the community to create innovations in technology and social organization. These innovations, if successful, can be implemented at home, and eventually be passed on to people in other communities.

Another set of theories concerning professional migration is presented by [Ire01] and starts by explaining the human capital theory according to which people move to find employment and remuneration more appropriate to their formal education and training. At this micro-level approach, there is no room for informal education, discrimination, or other factors that lead to imperfections in the labor market. The second body of theory is the structuralist neo-Marxist, which enables the impact of gender, race, class, but does not allow for institutional factors such as ethnic, professional, or industry unions. The third body of theory, the "structuration" approach incorporates structural and institutional elements where besides the private capital and the state being engaged in active recruitment to fill labor needs, there are also individual and organizational agents who produce context that motivate migration and set qualifications for hiring.

[Ire01] continues with the categorization of professional migrants by presenting five types of skilled migration:

- By motivation. Some cases within this typology include "forced exodus", "ethical emigration", "brain drain", "government induced" or "industry-led". The post-war "brain drain" occurred when the winner countries led by the US and the Soviet Union, entering a new stage of scientific and technological development caused the loss of valuable skilled personnel from developing to more developed countries. As counter-measured to this, we can mention the "reverse brain drain". "Government induced" reffers to government recruitment as in the instance of German specialists taken to the US to work in rocket and spacecraft engineering or to the Soviet Union as part of the Soviet's missile programme. "Industry-led" applies to situations where employers and businesses are the major drivers behind the selection and migration of skilled immigrants with the emergence which with the development of the internet has become an important instrument in this process;
- By nature of source and destination. The largest movement of skilled labor is from less developed countries to post-industrialized countries. Lack of economic opportunities, inadequate working, and intellectual environments are major factors when deciding to leave. Oil-rich states in the Middle East and the US, Canada, UK,

Japan, Singapore, Hong Kong, Taiwan, and Australia represent the most attractive and common destinations for skilled workers;

- By channel or mechanism. This category includes the following major channels:
  - The internal labor markets of multinational corporations;
  - Companies with international contracts that move staff to service their offshore work;
  - International recruitment agencies that handle large numbers of self-generated flows;
  - Small recruitment agents or ethnic networks;
  - Recruitment through other mechanisms, such as the internet.
- By the length of stay. As previously mentioned, length of stay has been a common mean of describing migration flows, especially with the distinction between permanent and temporary. Many countries show a willingness to admit temporaries while they attempt to close their doors to permanent workers;
- By type and level of regulatory mechanisms, the level of internationalization and the global labor market demand/supply situation;
- By mode of incorporation. Skilled flows may also be divided by the nature of the integration and reception of skillful migrants into destination economies:
  - "Disadvantaged" reception is one in which skilled immigrants face an unfavorable or hostile official reception, discrimination, or lack of legal status. More likely to occur in the past situation in countries of permanent immigration, and it has been notable in describing and understanding how and why skillful migrants choose their destinations if and when they have a choice;
  - "Neutral" context in which migrants become incorporated into the primary market at an appropriate level;
  - "Advantaged" situation, where due to political, social or economic factors, migrants experience mobility upward to positions of professional and civic leadership. More accepting, adaptable countries (e.g., Canada) and countries with large migrant populations, were often able to provide excellent opportunities for foreign professionals.

2.2 Information Technology and Human Migration

In combination with Linked Open Data and data visualization, migration data can provide a powerful instrument that can bring insight into both past historical events and future expectations. As digitization is taking over the world, we must question how it can or will impact some less digitized components of our life and history. To discover the

unknown information hidden in the data and address complex real-world problems, it is critical to be able to identify multivariate relations of migration flows while examining the spatial distribution.

The literature regarding the combination of information technology and human migration includes methods, techniques, and general solutions that deal with creating tools based on structured data layers that offer a way to analyze geographical clustering. In this case, we refer to Linked Open Data as being used in various research projects covering both data visualization and migration research.

When it comes to extracting, analyzing, connecting, and enriching Linked Open Data repositories, proLOD++ is a representative example [BNA+10]. This web-based profiling tool, helps users gain a deeper understanding of the underlying structure and semantics by analyzing N-Triple files containing all information of a data set. Another frequent entry in this category is represented by tools and articles that connect and publish data to the Linked Data Cloud. Among these, we can mention, for example, the Smithsonian museum [SKY+13] working on publishing its Linked Data.

Another accessible and explored category, is represented by the combination between Linked Open Data, together with different visualization approaches. This section includes mostly tools developed within Galleries, Libraries, Archives, and Museums (GLAM) institutions or independent developers that publish their code Open Source. An example in this regards is LODView [VAB+15], one of the most stable and concrete application that deals with both Linked Open Data (LOD) gathering and visualization and includes main features like multiple language capabilities, a widget area that contains multimedia elements and other data published in the LOD cloud regarding the same topic. eCultureMap [ZV13] is another remarkable example of a tool supported by Europeana and its partners in order to gather cultural and educational content in one single geographical knowledge map. In other words, it is a very efficient re-use of the Europeana data published under the Creative Commons Zero license. Using locations as the primary type of information, eCultureMap also shows cultural heritage objects in an international context (through linking to the Europeana portal) and national framework (through connecting to the national portals). Visualization wise, eCultureMap puts focus only on a type of visualization (map view). The last tool mentioned within this category is LODStories, a web application that offers users the possibility of exploring art linked data in a story-like manner. Users start with a topic of their choice and create a path formed by subjects and their connecting properties. After the path is established, the user can create videos using the Smithsonian database, Google, and the YouTube API. The generated videos can then be shared with other users on the platform and social media. The goal of LODStories is to create materials that can be used in a variety of educational settings.

When it comes to bringing together the research on human mobility and Linked Open Data, the literature offers quite seldom information. Instead, the more common approach, in this case, is using data provided by Governments and Statistical Offices. A concrete example in this sense is presented in [dSFGdS14] where the individual motivations and

environmental factors that determined skill labor loss in the case of Brazil is analyzed. The data layer in this research is provided by the Ministry of Labor and Employment and the Brazilian Statistics Bureau and contains recent entries from the 1995–2006 interval. One example that does use a Linked Data repository offered by the Shenzhen Special Economic Zone in China is described in [LC04], and it presents the relationship between migration and gender in China.

Moving further and including the visual-oriented approaches alongside human migration and structured data, we find a field with a lot of potential left for exploration. An example that focuses on data presentation and user interaction while attempting to enable a more in-depth understanding of global human migration patterns is The Monarch Room [CPP+18], a museum application described as 'An Interactive System for Visualization of Global Migration Data'. The paper describes the first design and research efforts, together with the methods used to capture visitors' attention in order to support them in their learning goals. The migration flows are displayed in an unified view, and it combines criteria such as time scales, location, and reasons behind the movement. While the paper gives no information about the used data source, we do get important insights about the most significant challenges when visualizing flow maps for spatial interaction data. One limitation is that geographically embedded networks are usually very large and often involve thousands of locations and connections between nodes. In addition, another constraint is the amount of variables associated with data flow (e.g., number of nomads, personal information such as age or occupation). As a solution, the Monarch uses contextual filters and represents people movements through lines, which can be distinguished by color and thickness. Colors represent in and outflows. Thickness, in turn, offers a quantitative representation of the amount of people, where dashed lines represent less, and thicker lines represent more people moving.

The challenges of visually representing large data sets are also presented by [SJ17] where the issue of overflooding and overlapping through design is analyzed. It is indeed a challenging problem when having to analyze and understand patterns in massive spatial interactions, which can easily have thousands of nodes (locations) and millions of connections. Described as location-to-location movements, these dynamic flow processes are often an important factor in a wide range of fields such as business or government, which requires taking the right decision based on the available data.

Expanding on the subject is [Guo09] who proposes an approach in which the solution is to aggregate locations into regions based on the flow structure/topology (as opposed to grouping edges into bundles based on geometric adjacency). Due to its repercussion, the use of spatial data is often employed in cultural-based settings, such as museums, as it enables the understanding of many historical processes that drive social, behavioral, or economic phenomena. The paper concludes that existing flow mapping approaches have three major limitations:

- flow mapping is only effective in portraying small datasets and will quickly become cluttered as the data set size increases.

- flow mapping often uses the default geographic unit of the observational data (e.g., counties or states), which may not be the best unit to represent and uncover underlying patterns due to the dramatic differences among the units.
- multivariate information cannot be visualized simultaneously with the flow patterns, and thus, it is difficult to perceive patterns that involve both flow structures and multivariate relations.

Regarding the types of data sets being used to demonstrate the developed approach, the 2000 U.S. county-to-county migration represents the foundation in this example. As an exploratory approach, flow maps are commonly used to visualize spatial interactions, since the origin and the destination of a flow are connected with a straight or artificially curved line. However, as introduced above, traditional flow maps are only effective in visualizing small datasets. [Guo09] gives a mathematical representation for the spatial interactions which naturally form a network/graph, where each node is a location (or area) and each link is an interaction between two nodes (locations). Such spatial interaction networks (e.g., county-to-county migrations) normally consist of:

- S: a set of locations (nodes), e.g., counties or states in the U.S., can range from dozens (e.g., U.S. states), through thousands (e.g., U.S. counties), to millions (e.g., mobile phones or individual household locations);
- F: a set of flows (links) among locations, directed or undirected; may vary from thousands (e.g., migration flows among 48 states), through millions (e.g., migration flows among >3000 counties), to billions (e.g., packages delivered by the post annually).
- Vf: a set of variables for each flow, e.g., the number of migrants for different age groups, income levels, and occupations, which normally ranges from a dozen to several hundred.

Another proposal made by [Guo09] in order to cope with the large number of flow lines and reduce the cluttering in a flow map is using sampling (in order to select and show only a small subset of data at a time, with the disadvantage of not supporting an overview of general patterns in the data) or derive and visualize line densities or group edges into bundles, which can effectively resolve the cluttering problem in the visual display.

Also, as previously mentioned by both [Guo09] and [CPP+18], spatial interaction data often contains multivariate information. For example, each migration flow at the county level contains the origin county, destination county, the number of migrants, and dozens of other flow variables such as counts of migrants for different age groups, income levels, etc. To fully understand spatial interactions and its driving processes, it is important to examine information across different perspectives and obtain an aggregated understanding of the overall patterns. This understanding should include information on the age category of the migrants or how different occupation tend to have different effects.

In a more extensive visual approach, [GCML06] reports and integrates computational, visual, and cartographic methods to develop a geovisual analytic approach for exploring and understanding spatio-temporal and multivariate patterns. The developed methodology and tools can help analysts investigate complex patterns across multivariate, spatial, and temporal dimensions via clustering, sorting, and visualization. Specifically, the approach involves a self-organizing map, a parallel coordinate plot, several forms of reorderable matrices (including several ordering methods) and a 2-dimensional cartographic color design method. The approach makes it possible to derive complex patterns and gain insights from spatio-temporal and multivariate data sets. The implemented visualization system has at least two important advantages:

1. its effectiveness in detecting and visualizing geographic, temporal, and multivariate patterns in multiple ways;
2. its component-based design that provides flexibility in addressing a range of analysis questions or a variety of different data sets by allowing easy connection to other visual and computational methods.

An interesting approach on migration research is made by [WLM+14] who tracks and visualizes the movement within China during the Spring Festival travel season, also known as Chunyun period. With the most massive annual flow of people, hundreds of millions return to their towns and villages for the holiday phenomenon happening since the late 1980s. The movement of rural Chinese to the cities has contributed a lot to the extraordinary Chinese growth. While various migration models were employed to make an analysis based on census data in almost all previous migration studies it is nearly impossible to make in-depth analysis about the migration routes, patterns, and temporal and spatial trend. The novelty of the paper is using Location-Based Service (LBS) provided by smartphones in order to collect people's geographical position data both from a spatial and temporal perspective, making it the largest global seasonal migration on earth.

The pattern discovered shows that from the temporal perspective, the migration population is remarkably regular. From the spatial aspect, the migration direction between regions before and after Chinese New Year's Day is the opposite.

Also worthy of mentioning here as the last reference is [SAPJ18] who offer an example on how the computer has mediated ways of researching and identifying historical networks on the formation of the Norwegian-American communities by using the open Application Programming Interface (API) of the National Archives of Norway. The result is an interactive statistical temporal and spatial visualization of the Norwegian migration within the USA that occurred between 1870 and 1920.

CHAPTER 3

Tools and Technologies

In order to successfully implement the proposed requirements for the thesis, the following structure and system have been created:

- The analysis component is represented by a series of data visualization and statistical methods which will be described in both "Tools and Technologies" and "Implementation and Analysis" chapters;
- The foundation blocks of the application is formed by HTML, CSS, and Javascript (together with its libraries - E.g., D3, jQuery, Ajax, d3sparql, noUiSlider, Leaflet);
- The data source is represented by the Wikidata Knowledge Base (KB);
- The data retrieval method is represented by the SPARQL Resource Description Framework (RDF) query language.

3.1 Data Visualization

[RPMK19] defines data visualization as being a general term used to describe the effort that helps understand the importance of data by placing it in a visual context. In this manner, patterns and correlations that go undetected in large and complex text-based datasets can be recognized easier with data visualization.

The explanation behind this lays in the fact that humans are capable of distinguishing differences in line size, shape, orientation, and color without investing significant processing effort. According to [Few04b], these are referred to as "pre-attentive attributes", and while it requires time and effort to identify the number of times a digit appears in a series of numbers("attentive processing"), it is easier to observe it when it is different in size or color. Moreover [HAARV17] claims that 2/3 of the brain's neurons can be

involved in the visual processing, which can show a different set of potential connections and relationships, not as evident as in non-visualized quantitative data.

An important early example of an information graphic is represented by Charles Joseph Minard's diagram of Napoleonic France's invasion of Russia from 1869 [Fri02]. The diagram (displayed in Figure 3.1) shows the losses suffered by Napoleon's army in the 1812–1813 period with the following variables plotted:

- the size of the army
- its location on a two-dimensional surface (x and y)
- time
- direction of movement
- temperature

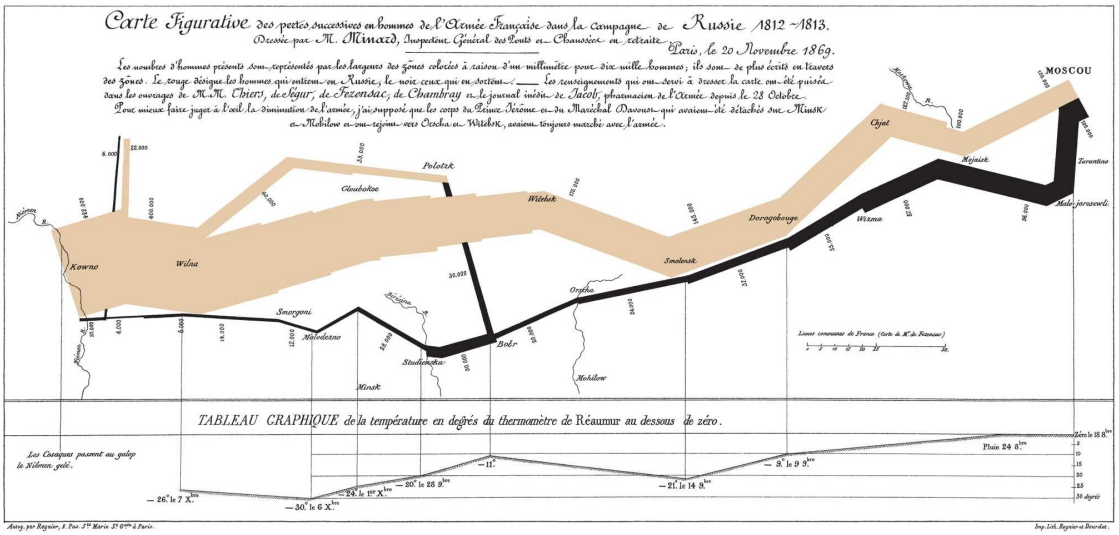

Figure 3.1: Minard's diagram of the Napoleonic invasion of Russia

The line width illustrates a comparison of the army size at points in time, while the temperature axis suggests a cause of the change in army size.

As mentioned in the previous chapters, the vast amounts of data created by Internet activity and an increasing number of sensors in the environment bring new challenges when processing and analyzing this data. These challenges are propagated to data visualization, where data scientists and engineers are helping to address the issue.

[Tuf01] defines a series of principles for effective graphical display in order to have "complex ideas communicated with clarity, precision, and efficiency":

- show the data
- induce the viewer to think about the substance rather than about methodology
- avoid distorting what the data has to say
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation or decoration
- be closely integrated with the statistical and verbal descriptions of a data set.

Not applying these principles may result in "chartjunk" or misleading graphs, which distort the message or support an erroneous conclusion.

In [Few04a], Few described eight types of quantitative messages and how they can be better communicated depending on the corresponding type of graphics:

- Time-series: A single variable is captured over a period of time in order to demonstrate a trend. E.g., line charts;
- Ranking: Categorical subdivisions are ranked in ascending or descending order in order to demonstrate comparison. E.g., bar charts;
- Part-to-whole: Categorical subdivisions are measured as a ratio to the whole. E.g., pie charts or bar charts;
- Deviation: Categorical subdivisions are compared against a reference. E.g., bar charts;
- Frequency distribution: Shows the number of observations of a particular variable for a given interval. E.g., histograms, bar chart or boxplot to visualize key statistics about the distribution;
- Correlation: Comparison between observations represented by two variables to determine if they tend to move in the same or opposite directions. E.g., scatter plots;
- Nominal comparison: Comparing categorical subdivisions in no particular order. E.g., bar charts;
- Geographic or geospatial: Comparison of a variable across a map or layout. E.g., cartograms.

Another interesting mention within this chapter is represented by [LE07] where the development of a "Periodic Table of Visualization Methods" has been addressed. According to the authors, the table is a functional metaphoric homage to the chemistry periodic table, that although not exhaustive, "it does provide an overview over more than hundred useful visualization methods of great variety" which can assist "researchers and practitioners alike in choosing adequate visualization methods for their needs". The interactive chart includes six types of data visualization methods and categories: data, information, concept, strategy, metaphor and compound[1].

In the next subsections, the primary data visualization formats used within the application will be presented.

3.1.1 Pie Chart

A pie chart (or a circle chart or pie graph) is a circular statistical graphic, which is divided into slices to illustrate different proportions. As a result, the arc length, central angle, and area of each slice are proportional to the number it represents. The earliest known pie chart is usually credited to William Playfair's Statistical Breviary, who more than 200 years ago, in1801, presented a series of pie charts (among which, one regarded the proportions of the Turkish and German empires before 1789) [Spe05].

While it is named due to its similarity to a pie which has been sliced, pie charts present variations on the way it can be presented [BEW16].

- the 3D pie chart (or perspective pie chart) is often used for aesthetic reasons since it gives the chart a 3D aspect. This is due to the fact that the third dimension does not enhance the reading of the data, but it has a distorted effect of perspective. The use of unnecessary dimensions that don't display the data of interest is discouraged in general for other charts, not only for pie charts;

- the doughnut chart is a variant of the pie chart, with a blank center that allows the addition of information about the data as a whole. Doughnut charts are comparable to pie charts in the sense that their scope is to illustrate proportions, but they can support multiple statistics at once;

- the exploded pie chart is a chart with one or more sectors separated from the rest of the circle. This effect is used to emphasize a sector or smaller segments of the chart with very small proportions and does not change the angle, area, or arc length;

- the ring chart (or the sunburst/multilevel pie chart), is used to visualize hierarchical data displayed in concentric circles. The inner circle constitutes the root node, with the rest of the hierarchy moving outwards;

[1] A periodic table of visualization methods(2007): `http://www.visual-literacy.org/periodic_table/periodic_table.html` accessed on 15.07.2019

- the elliptical chart, compressed horizontally into an ellipse, by removing the unneeded parts of a full circle. The ellipse strongly distorts area and arc length, but not the angle;

- the square pie chart (or the waffle chart), demonstrates how smaller percentages are more easily displayed than on circular charts. Just like in the case of the ellipse chart, the square pie has a nonlinear effect on area and arc length.

3.1.2 Bar Chart

A bar chart is a type of statistical graph that displays grouped data with rectangular bars. The bars can be plotted either vertically or horizontally, and the lengths of bars are proportional to the values they represent [HLD+07]. A bar graph shows comparisons between discrete categories with one axis of the chart showing the types being analyzed, and the other axis representing the measured value.

Just as in the case of the pie chart, many sources consider William Playfair to be the inventor of the bar chart with the creation of The Commercial and Political Atlas where he presented the "Exports and Imports of Scotland to and from different parts for one Year from Christmas 1780 to Christmas 1781" [Wai97].

Bar charts hold a discrete domain of categories and are sized so that all the data can fit on the chart. When there is no intentional ordering of the categories, bars can be arrayed in any order. Bar charts arranged from highest to lowest frequency are called Pareto charts. Bar graphs can also be employed for more complex comparisons of data with grouped and stacked bar charts. In the case of grouped bar charts, for each categorical group, there are at least two bars that show the values of two or more measured variable. As for the stacked bar charts, the grouping is made on top of each other, with the height of the resulting bar exposing the combined result of the groups. However, stacked bar charts are not the solution to data sets where some groups have negative values [SG14].

3.1.3 Chord Diagram

A chord diagram (also referred to as radial networks or dependency diagrams) is a graphical method for displaying connections or flows between several entities or nodes. Chord diagrams get their name from the terminology used in geometry where a chord represents a geometric line segment whose endpoints lie on a circle.

The data are organized radially with the relationships between the data points typically represented as arcs or Bézier curves. This makes Chord Diagrams ideal for comparing the similarities within a dataset or between different groups of data as values are assigned and represented proportionally by the size of each connection and color can be used to group the data into different categories.

Interactivity is another advantage when using this type of diagram because it can make the analysis easier by hovering and highlighting a specific group and all its connections.

On the other hand, over-cluttering represents an issue with Chord Diagrams when there are too many connections displayed [2].

3.1.4 Geovisualization

Geovisualization is a discipline containing a set of tools and techniques that support the analysis of geospatial data through the use of visualization. Geovisualization developed as a research field in the early 1980s, mainly as a result of the cartographic design and information visualization work of Jacques Bertin, a graphic theorist [RMD06].

Traditional, static maps have limited and poor exploratory capabilities since the graphical representations are entirely linked to the geographical data beneath. Geographic Information System (GIS) and geovisualization take advantage of the ability of modern microprocessors and provide interactive maps with the ability to render changes in real time, adjust the data on the fly, zoom in or out, change the visual appearance and explore different layers of the map on a digital display [MGP+04].

An important and complex question that arises in both the traditional cartography and geovisualisation is how to project maps of the 3-dimensional spherical Earth onto 2-dimensional surfaces. An answer to this is represented by map projections or in other words, a systematic transformation of the latitude and longitude coordinates of locations from the surface of a sphere or ellipsoid into areas on a plane[Sny89].

All projections distort the surface to some degree and depending on the purpose of the map, several projections exist in order to preserve specific properties of the sphere-like body at the expense of other characteristics. The 'orange peel problem' is possibly the most widely-cited analogy used to explain why a three-dimensional world cannot be represented in two dimensions without any kind of distortion. Distortions in terms of shape, distance, direction, or area are inevitable [3].

Some of the most common map projections are:

- Mercator, created by the Flemish cartographer Gerardus Mercator as a navigational tool for sailors and has been the most popular map projection in the world since 1569 when Antarctica was not yet discovered. On a Mercator projection, Greenland is approximately the same size as Africa, even though in reality, Africa is almost 14 times larger. Despite such distortions, Google Maps, Bing, Yahoo, or OpenStreetMaps continue supporting and using it to display the world.
- The Gall-Peters projection was created in 1973 by the German filmmaker and journalist Arno Peters as a response to the issues displayed by the skewed Mercator projection. Peters claimed that enlarging Europe and North America, was giving

[2] The Data Visualisation Catalogue:Chord Diagram `https://datavizcatalogue.com/methods/chord_diagram.html` accessed on 15.07.2019

[3] National Georgraphic: Investigating Map Projections `https://www.nationalgeographic.org/activity/investigating-map-projections` accessed on 15.07.2019

white nations a sense of supremacy over non-white nations. As a result, an equal-area projection was designed. On the other hand, in its quest of removing size distortions, the map stretched some places near the poles and Equator to an unexpected level [4].

- The Robinson projection was created in 1963 by the American geographer and cartographer Arthur H. Robinson who came up with a projection that concentrated more on the appearance of the map than precise measurement of places.

- The AuthaGraph projection was created in 1999 by the Japanese architect Hajime Narukawa by equally dividing a spherical surface into 96 triangles, and it's currently the most accurate one in existence. These triangles were projected onto a tetrahedron, which helped maintain the proportions of land and water and helped to unfold the map into a perfect, flat rectangle. The creator, however, intends to refine the process one step further in order to increase the number of subdivisions, and as a result, improve its accuracy [BN14].

3.2 Web Technologies

The application developed within the master thesis lays on four main Web pillars: Hypertext Markup Language (HTML), Scalable Vector Graphics (SVG), Cascading Style Sheets (CSS), and JavaScript. JavaScript together with its libraries (D3, jQuery, Ajax, d3sparql, noUiSlider, Leaflet) represents the main component within the implementation phase.

3.2.1 HTML

HTML is the standard markup language for creating Web pages[5]. HTML was developed by Tim Berners-Lee while he was working at CERN and it exploded during the 1990s together with the growth of the Web[6]. Web browsers receive HTML documents from web servers or local storage and render them into web pages. HTML provides a way to build documents by offering structural elements described by tags. "Tags" are a type of syntax that forms the skeletal structure of a web page written in the form of specific text enclosed between angular brackets (e.g., <html>,<head>,<body>).

The World Wide Web (Web) is a network of information resources made available through three mechanisms[7]:

[4] The Peters Projection and the Mercator Map `https://www.thoughtco.com/peters-projection-and-the-mercator-map-4068412` accessed on 15.07.2019

[5] w3schools: HTML Introduction `https://www.w3schools.com/html/html_intro.asp` accessed on 13.07.2019

[6] w3.org: A history of HTML `https://www.w3.org/People/Raggett/book4/ch02.html` accessed on 13.07.2019

[7] w3.org: HTML 4.01 `http://www.w3.org/2006/07/home/wire/20080229/html401-per` accessed on 13.07.2019

- A uniform naming scheme or address for the localization of resources encoded through Universal Resource Identifier (URI). In this category, URLs form a subset of the more general URI naming scheme;
- Protocols, for access to the named resources (e.g., HyperText Transfer Protocol (HTTP));
- Hypertext, for easy navigation between resources (e.g., HTML).

HTML documents work well across different browsers and platforms, and each new version of the language has attempted to reflect a greater consensus among the industry players. HTML was developed with the idea that all the types of devices should be able to use the information on the Web: PCs, cellular telephones, devices for speech output and input, and so on.

3.2.2 CSS

CSS is a style sheet language created for the styling and formatting of web pages through setting a variety of rules like position, color, formatting. CSS is a foundation technology of the World Wide Web, besides HTML and JavaScript.

CSS is designed to enable the disconnection of presentation and content, which can promote content accessibility and provide flexibility and control. The "cascading" term originates from the specified priority scheme that decides which style rule is applied if more than one rule matches an element.

The CSS specifications are maintained by the The World Wide Web Consortium (W3C), and besides HTML, other markup languages including XHTML, XML, SVG, and XUL support the use of CSS.

CSS has a simple syntax consisting of a list of rules formed by one or more selectors and uses several English keywords to specify the names of various style properties.

CSS can be added to HTML elements in 3 different ways:

- Inline - by using the style attribute to apply format to a single HTML element;
- Internal - by using a <style> element in the <head> section;
- External - by using an external CSS file to define the style for several HTML pages[8].

[8] w3schools: HTML Styles - CSS `https://www.w3schools.com/html/html_css.asp` accessed on 13.07.2019

3.2.3 SVG

SVG is a format used to draw two-dimensional XML-based vector graphics and animations that do not lose quality when zoomed or resized. SVG is a W3C Recommendation and integrates with other W3C standards such as the DOM and XSL.

Some of the advantages that SVG has compared to other image formats (e.g., JPEG, GIF, or Bitmap) are:

- SVG components can be created and edited with any text editor;
- SVG components can be searched, indexed, scripted, and compressed;
- SVG components are scalable and zoomable;
- SVG is an open standard;
- SVG files are pure XML.

SVG is supported by all the major modern web browsers and provides some basic shapes like lines, rectangles, circle, ellipse, polygon, or custom shapes by combining or modifying the basic ones. SVG allows three types of graphic objects which can be grouped, styled, transformed, and combined into:

- vector graphic shapes such as paths;
- outlines like as of straight lines and curves, bitmap images;
- text.

One other aspect to mention is that within HTML, SVG is created using the <svg></svg> tags and has the origin(0,0) at the "top-left" corner as opposed to conventional "bottom-left" [9].

3.2.4 JavaScript

JavaScript is a high-level, interpreted programming language that enables interactivity and adds functionality to a basic HTML web page. Together with HTML and CSS, JavaScript is one of the core technologies of the World Wide Web. As previously mentioned, while HTML defines the content of web pages and CSS is used to specify the layout, JavaScript is used to program the behavior.

As of May 2017, 94.5% of 10 million most popular web pages used JavaScript[10]. As a multi-paradigm language, JavaScript supports event-driven, functional, and imperative

[9] w3schools: SVG Tutorial `https://www.w3schools.com/graphics/svg_intro.asp` accessed on 13.07.2019

[10] w3techs: Usage statistics of JavaScript as client-side programming language on websites `https://w3techs.com/technologies/details/cp-javascript/all/all` accessed on 13.07.2019

programming styles. It has capabilities for working with text, arrays, dates, regular expressions, and the DOM, but the language itself relies on the host environment (usually the Web browser) to include I/O operations.

JavaScript offers a series of additional libraries and frameworks, which leads to the term "Vanilla JavaScript" which refer to plain JavaScript, not extended in any way. Web pages are not the only place where JavaScript can be employed. Many server applications (Node.js) or databases (e.g., MongoDB) use JavaScript as their programming language. Also, the JSON serialization format is based on JavaScript [11].

3.2.5 D3.js

"A picture is worth a thousand words". Without visualization, statistics statements are just a set of statistics, and as the Web is becoming more accessible and complex every day and with progress in browser technology, it is now possible to visualize big sets of data on the fly across a variety of devices.

A popular way of doing this is represented by D3.js (or Data-Driven Documents), a JavaScript library that produces dynamic, interactive data visualizations in web browsers by using the previously mentioned technologies and standards: SVG, HTML, and CSS. Compared to many other libraries, D3.js offers since 2011 an excellent control and customization capabilities, performance, and level of interactivity over the final visual result.

The main idea at the base of D3 was bridging the gap between static display of data, and interactive and animated data visualizations. The primary usage is binding arbitrary data (JSON, CSV, TSV, etc.) to a Document Object Model (DOM) and through the use of JavaScript, CSS, HTML, and SVG, apply transformations to the document that are driven by that data. The result can be an HTML output or interactive SVG charts with dynamic behavior like animations or interactions. All the data transformations and renderings are done client-side, in the browser.

Implementation wise, including the D3's library will offer access to a global object that can be used to call various functions through chain syntax or method chaining. This way, large datasets can be easily bound to SVG objects to generate rich text/graphic charts and diagrams [BOH11].

3.2.6 Leaflet.js

Today, the most widely used mapping libraries are Google Maps and Leaflet. Known as "slippy maps", they are designed to display fast and easy maps on the web, while allowing zooming and panning around the map.

[11] w3schools: JavaScript Tutorial `https://www.w3schools.com/js/default.asp` accessed on 13.07.2019

Leaflet has become one of the most popular JavaScript mapping libraries, supporting most mobile and desktop platforms and being used by major web sites such as FourSquare, Pinterest, Facebook and Flickr. [12]

Being Open Source, Leaflet is an accessible alternative to Google Maps that was designed with simplicity, performance and usability in mind. Leaflet can be easily used when combined with D3.js's data manipulation features, and for using D3.js for vector based graphics. On the other hand, Google Maps can be more difficult to combine with D3.js, as deeper integration is not really possible without significant tweaking and hacking.

Regarding map projections all mapping libraries support the Spherical Mercator projection (or a variant of the Mercator projection that is the de facto standard for Web mapping applications out of the box). On the other hand, if other projections are needed, the Proj4js library must be used in order to do transformation from one coordinate system to another. Here comes another advantage of Leaflet, since this JavaScript library has a Proj4Leaflet plugin while Google Maps, does not.

D3.js can be very powerful when it comes to handling geographical information, and this is also due to the fact that map building works very well with geographic data formatted in JSON formats like the GeoJSON and TopoJSON specifications. GeoJSON is "a format for encoding a variety of geographic data structures" designed to represent discrete geometry objects grouped into feature sets of name/value pairs. [13]

TopoJSON extends GeoJSON and eliminates redundancies by storing relational information between geographic features, not just spatial data (e.g., given a map with several countries bordering each other, the shared parts of the borders will be stored twice in GeoJSON while in TopoJSON, it will be stored just once). As a result, geometry is much more compact and combined, and the file is typically 80% smaller [14]. D3.js also offers built-in support for many different geographic projections, through full geometric transformations.

A common use of Leaflet involves binding a Leaflet "map" element to an HTML element (e.g. div). Like for other web map libraries, the primary display model implemented by Leaflet is one base map, plus zero or more translucent overlays and zero or more vector objects displayed on top.

3.3 Wikidata

As previously mentioned, Semantic Web was first introduced to a broader audience by Berners-Lee in 2001 [BLHL01]. The starting point to the concept stands around the idea that the traditional Web of Documents should be extended to a Web of Data where we

[12] Leaflet an open-source JavaScript library for mobile-friendly interactive maps - `https://leafletjs.com/` accessed on 13.07.2019

[13] GeoJSON Main Page `https://geojson.org/` accessed on 13.07.2019

[14] Building Great Web Maps: A D3.js Tutorial | Toptal. `https://www.toptal.com/javascript/a-map-to-perfection-using-d3-js-to-make-beautiful-web-maps` accessed on 13.07.2019

don't have documents and links between documents, but entities and relations between entities.

At the foundation of Semantic Web stands Knowledge Graphs (KGs), defined as knowledge bases (KB) or combinations of an ontology and instances of the classes that belong to that particular ontology, which also consist of a large number of facts about entities [FEMR15].

DBpedia and Wikidata are two such online KB projects that focus on offering structured data from Wikipedia in order to ease its exploitation on the Linked Data Web. As a fundamental difference between the projects, we can highlight that Wikidata has an open centralized nature, whereas DBpedia is more popular in the Semantic Web and the Linked Open Data communities and depends on the different linguistic editions of Wikipedia [AGMRTL17].

Wikidata was launched in October 2012 with the aim of creating a free knowledge base about the world that can be read and edited by humans and machines alike. Wikidata is a collaboratively edited knowledge base hosted by the Wikimedia Foundation, and it is a common source of open data that Wikimedia projects such as Wikipedia can use under a public domain license [15].

Later, in September 2015, the Wikimedia Foundation announced the release of the Wikidata Query Service, which lets users run SPARQL queries on the data contained in Wikidata[16]. As of June 2019, Wikidata information is used in 65.53% of all Wikipedia articles [17].

DBpedia represented the 2nd data source option analyzed in the context of the thesis. Defined as a project aiming to extract structured content from the information created in the Wikipedia project, DBpedia allows users to semantically query relationships and properties of Wikipedia resources, including links to other related datasets [18].

The project was started by the researchers from the Free University of Berlin and Leipzig University, and the first publicly available dataset was published in 2007. As Wikipedia articles consist of both free text and structured information (categorization information, images, geo-coordinates and links to external Web pages) embedded in the articles' "infobox" tables, the data is extracted and put in a uniform dataset which can be afterward queried. The knowledge base was made available under free licenses, allowing others to reuse the dataset. [ABK+07]

[15] Wikimedia page on Wikidata (2018): `https://meta.wikimedia.org/wiki/Wikidata` accessed on 14.07.2019

[16] Announcing the release of the Wikidata Query Service (2015): `https://lists.wikimedia.org/pipermail/wikidata/2015-September/007042.html` accessed on 14.07.2019

[17] Percentage of articles making use of data from Wikidata (2019): `http://wdcm.wmflabs.org/WD_percentUsageDashboard/` accessed on 14.07.2019

[18] DBpedia Interlinking: `https://wiki.dbpedia.org/services-resources/interlinking` accessed on 14.07.2019

The 2016 release of the DBpedia data set describes 6.0 million entities, with 5.2 million classified in a consistent ontology. This included 1.5M person, 810k places, 135k music albums, 106k films, 20k video games, 275k organizations, 301k species, and 5k diseases. In total, it consisted of 9.5 billion RDF triples, from which 1.3 billion were extracted from the English edition of Wikipedia and 5.0 billion from other language editions. [19]

The relationship between DBpedia and Wikidata has been presented clearly by [AGMRTL17] where it can be observed how Wikidata represents one of the sources for different DBpedia projects with having Wikipedia as a central component, and crowdsourced data as external sources. This structure can be seen in Figure 3.2.

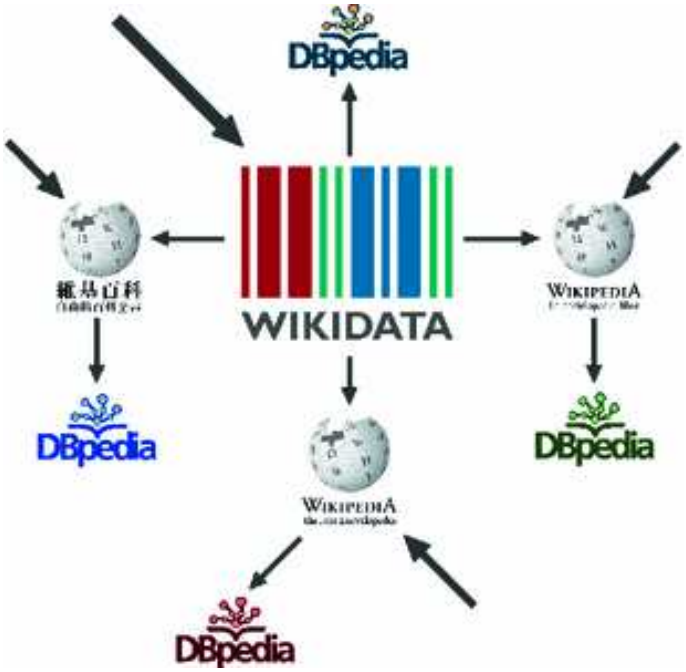

Figure 3.2: The Wikipedia, DBpedia and Wikidata interaction

While freely available knowledge graphs (KGs) have not been in the focus of an extensive comparative study, a comprehensive comparison is given by [FEMR15] for both DBpedia and Wikidata alongside Freebase, OpenCyc, and YAGO.

The following conclusions are important concerning general information of the comparison between Wikidata and DBpedia:

- Language support: Most KGs either support only English or other languages included additionally. The "Main" DBpedia is English (properties etc.), but linked

[19] DBpedia Blog Post YEAH! We did it again ;) – New 2016-04 DBpedia release `https://blog.dbpedia.org/2016/10/19/yeah-we-did-it-again-new-2016-04-dbpedia-release/` accessed on 14.07.2019

localized versions are available in 125 languages. Remarkable in this setting is Wikidata in terms of number of languages supported. (almost every language, even dialects);

- Covered domains: Both DBpedia and Wikidata contain general knowledge scopes and domains;
- Fact representation: Both DBpedia and Wikidata use the Resource Description Framework (RDF) to represent extracted information;
- Dynamicity: Many KGs are static in the sense that they are not continuously updated. DBpedia is created by computationally-expensive information extraction processes, which makes it mainly static with the exception of DBpedia live – a derived version of DBpedia that is continuously updated through Wikipedia analyzing around 84 articles per minute. On the other hand, Wikidata is dynamic since a user community maintains data and even extends the schema;
- The provenance of data: For covering knowledge about general domain entities in DBpedia, Wikipedia content is exploited to some degree with the help of information extraction tools. In this case, no quality assurance check is implemented. For Wikidata, data is maintained by users and bots and references can be attached which reveal the source – and therefore indirectly the trustfulness of the statement.

Next, the main data elements and concepts related to Wikidata will be explained[20]:

- Items - being a document-oriented database, Wikidata focused on items which include topics, concepts, and objects, each identified by a unique number, prefixed with the letter Q, known as a "QID". Within the thesis, examples of items include:
  - countries (e.g. Q414 for Argentina);
  - professions (e.g. Q42973 for architect);
  - gender (e.g. Q6581072 for female).
- Statements represent the way known information about an item is recorded in Wikidata. Statements may map a property to one or more values like in the example of the "occupation" property which for Marie Curie could be linked to the values "physicist" and "chemist", to indicate the fact that she engaged in both occupations. The values may be of several types, including other Wikidata items, strings, numbers, or media files. Another element, properties, appoint what types of values they may be paired with and may establish more complex rules about their intended usage and constraints. Optionally, qualifiers can be utilized to improve and refine the meaning of a statement by giving additional information that applies

[20]Wikidata Help:Contents `https://www.wikidata.org/wiki/Help:Contents` accessed on 14.07.2019

to the scope of the statement. Statements may also be annotated with references that point to a source supporting the content.

- Lexemes - In linguistics, a lexeme is a unit of lexical meaning. Likewise, Wikidata's lexemes are items suitable to store lexicographical data like the language or the form to which the lexeme refers.

3.4 SPARQL

SPARQL is an RDF query language or in other words, a semantic query language for databases that can retrieve and manipulate data stored in RDF format. It was made a standard by the World Wide Web Consortium and is recognized as one of the key technologies of the semantic web [21].

SPARQL allows for a query to consist of triple patterns, conjunctions, disjunctions, and optional patterns. Implementations for multiple programming languages exist, and there are tools that can translate SPARQL queries to other query languages like SQL [22] and XQuery [23].

The main advantage of SPARQL is that it enables users to write queries against "key-value" data or "subject-predicate-object" triples that follow the RDF specification of the W3C. The schema information or the ontology is often provided externally, to enable joining of different datasets unambiguously. In addition, SPARQL provides a specific graph traversal syntax for data that can be considered as a graph. SPARQL provides a comprehensive set of analytic query operations such as JOIN, SORT, AGGREGATE for data whose schema is intrinsically part of the data rather than needing a separate schema definition.

Queries can be distributed to various SPARQL endpoints where they can be computed, and results gathered in a procedure known as federated query.

As for the query forms, the SPARQL language specifies four distinct query variations for different purposes:

- SELECT query used to extract raw values from a SPARQL endpoint with the results being returned in a table format;
- CONSTRUCT query used to extract information from the SPARQL endpoint and convert the results into valid RDF;

[21] w3c SPARQL Query Language for RDF `https://www.w3.org/TR/rdf-sparql-query/` accessed on 15.07.2019

[22] D2R Server: Accessing databases with SPARQL and as Linked Data `http://d2rq.org/d2r-server` accessed on 15.07.2019

[23] The SPARQL2XQuery framework `http://www.dblab.ntua.gr/~bikakis/SPARQL2XQuery.html` accessed on 15.07.2019

- ASK query used to provide a true/false result for a query on a SPARQL endpoint;
- DESCRIBE query used to extract an RDF graph from the SPARQL endpoint.

Each of these query variations takes a WHERE block in order to restrict the query.

CHAPTER 4

Implementation and Analysis

4.1 Data Retrieval

When selecting the data sources, the focus gravitated around those Knowledge Graphs (KGs) that are freely accessible, freely usable, that incorporate the Semantic Web standards and that cover general cross-domain knowledge. In the scope of the Data Source analysis, KGs which are not openly available such as the Google Knowledge Graph or the Facebook Graph were not included. In the same category also entered the KGs which are only accessible via an API (e.g., WolframAlpha) and unstructured or weakly structured knowledge collections.

Since the source of the data can be easily changed with the adaptation of the SPARQL endpoint and potential small changes on the queries, we will concentrate in this thesis on Wikidata and its capabilities. As previously mentioned, Wikidata constitutes the central structured data storage for the unstructured Wikimedia projects (including Wikipedia, Wikivoyage, and Wikisource). Besides offering primary data like the place and date of birth/death, Wikidata also provides data on citizenship, cause of death, marriage/partners, awards, membership in organizations, and political parties. Another reason for choosing Wikidata is the consistent database with new content being added every day by bots and the community. On the other hand, due to the fact that the platform was created through crowd-sourcing, it is prone to errors, and an essential part of the Data Retrieval stage also includes the sanitizing process where erroneous or incomplete data will be removed or skipped.

In the resulted data pool, only English data from both female and male notable people that have assigned at least one profession and are distributed to all the geographical areas and time eras will be selected.

In order to retrieve the data, SPARQL queries have been used. Some of the limitations here firstly regarded fetching large quantities of data and secondly dealing with unsanitized

data entries. Currently, the SPARQL queries allow a certain length and number of queries per interval. Another issue here is also time-outs or more precisely how far is it possible to broaden narrow queries before they fail.

In some cases, there may then be ways to re-write the queries or change their optimization to allow broader queries to be run, pushing the limits further. Types of queries like non-restrictive (with very many results, e.g. "list of all humans") or negative searches (e.g., "all humans without images" that inspects records for every human) will be slow or will not pass/return results. In order to bypass such issues, the previously described criteria were applied in such a way that a consistent number of results containing valuable data that verified all the requested information.

The second big challenge when retrieving the data concerns Wikidata being a crowdsourcing platform. This means that the platform is susceptible to erroneous or incomplete entries. For example, duplicates are bypassed in the queries by using the "DISTINCT" statement, but this will not filter the results that have:

- one or more different result component (E.g., the geographical coordinates of the birth or death place even though the rest of the data is identical like shown in Figure 4.1);
- have one different/additional character which leads to them being seen as separate entries (see Figure 4.2)

Jean Baptista von Schweitzer	Birth: Frankfurt am Main, Germany	Death: Giessbach, Switzerland	1833 - 1875	383km
Jean Baptista von Schweitzer	Birth: Frankfurt am Main, Germany	Death: Giessbach, Switzerland	1833 - 1875	379km

Figure 4.1: Two duplicates illustrating the same entity but fetched as different due to a separate component, in this case, the exact geographical coordinates which in the figure, resulted in different birth-death distances.

James Kirkwood, Jr.	Birth: Hollywood, United States of America	Death: Manhattan, United States of America	1924 - 1989	4923km
James Kirkwood Jr.	Birth: Hollywood, United States of America	Death: Manhattan, United States of America	1924 - 1989	4923km

Figure 4.2: Two duplicates illustrating the same entity but fetched as different due to the "," character.

Only data that was passing the previously mentioned criteria were taken into consideration. In other words, only items that had:

- date of birth, with the property P569 (Wikidata description: date on which the subject was born);
- date of death, with the property P570 (Wikidata description: date on which the subject died);

- place of birth, with the property P19 (Wikidata description: most specific known (e.g. city instead of country, or hospital instead of city) birth location of a person, animal or fictional character);
- place of death, with the property P20 (Wikidata description: most specific known (e.g. city instead of country, or hospital instead of city) death location of a person, animal or fictional character);
- country corresponding the place of birth or death, with the property P17 (Wikidata description: sovereign state of this item)
- coordinate location corresponding the place of birth or death, with the property P625 (Wikidata description: geocoordinates of the subject);
- gender with the property P21 (Wikidata description: sexual identity of subject);
- profession or occupation with the property P106 (Wikidata description: occupation of a person);
- label in order to avoid the "QID"-style naming and fetch the text in a human-readable format (E.g., Pablo Picasso instead of Q5593).

Optionally, the sexual orientation (Wikidata property P91), religion (Wikidata property P140), and ethnicity (Wikidata property P172) can be set, but due to the fact that the result set is small, these criteria were not applied for the analysis performed within the present thesis.

The country of birth and death depends on the city of death and birth. This is caused by the fact that while a country has changed significantly across history, cities have remained more constant. As an example, Munich (https://www.wikidata.org/wiki/Q1726) will appear as part of:

- Germany;
- Nazi Germany;
- Weimar Republic;
- German Empire;
- Kingdom of Bavaria;
- Electorate of Bavaria;
- West Germany;
- Allied-occupied Germany.

The selection of the country among the previously mentioned ones is made according to the selected time period for the birth/death dates.

4.2 Implementation

In order to reach the expected results, an application with Linked Open Data querying and visualization capabilities had to be implemented. Using the outcomes from the previous steps, a prototype that offers visualization and statistical analysis capabilities will be applied. While the most popular way of displaying Linked Open Data is currently under a graph form [MTG14] or as a text-based interface [LKS+14], the prototype presented in the thesis will explore a combination of geographical and temporal visualization. In other words, the expected results will be reached by constructing a map view that reacts to the selection of a temporal slider.

As presented in the Technology chapter, D3.js is the main library used for the visualization components, while the Wikidata repository, together with SPARQL were the choices for the data source and data retrieval. For data retrieval itself, Asynchronous JavaScript and XML (AJAX), which is a set of web development techniques used to create asynchronous web applications, was used.

Regarding the structure of the application, the graphics interface can be divided into three distinctive components:

1. The controller section - displays a set of basic configuration options which reflect the criteria used to fetch the result set:

   - time period - a double-sided slider employed to select the time interval for the period which will be analyzed. The time span extends from 500 B.C to the present day. Several time points which are regarded as triggers in the history of migration are distributed below the slider;
   - country of birth and death - 2 separate drop-downs used to select the source country and the destination country for the analysis. Each set was populated with data from Wikidata and can contain countries that are not existing anymore (e.g., Ottoman Empire);
   - profession - displayed by using a simple dropdown with data fetched from Wikidata. The professions displayed are grouped by domain and only include occupations that returned at least 1000 hits (people labeled with the professions on Wikidata);
   - gender - displayed through a simple selector containing the options "male"(Wikidata item Q6581097), "female" (Wikidata item Q6581072) and "all".

   As previously mentioned, the user interface offers the possibility to extend the analysis to further criteria by providing the options "religion", "sexual orientation" and "ethnic group". Since the result set was not large enough, the previous three criteria were only exploratory, added to understand the dynamics better. Figure 4.3 shows both the basic configuration, as well as the extended one.

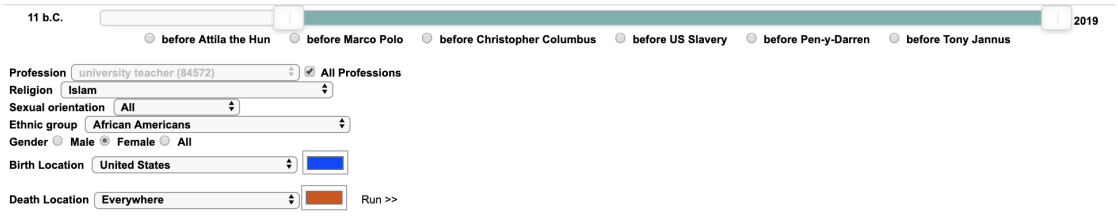

Figure 4.3: The controller section and parts of the visualisation section

2. The analysis section - once the criteria are selected, the user will proceed into fetching the needed data and generate a series of statistics and diagrams:

   - Set of statistics on the number of results and the distribution according to different distance intervals (see Figure 4.4a):
     - total number of results;
     - 0km where the birth place coordinates are the same as the death place coordinates. This implies that migration did not occur;
     - 100km where the distance between the birth place and the death place coordinates is between 0 and 100km. This implies that migration is local, approximately within a region;
     - 1000km where the distance between the birth place and the death place coordinates is between 100km and 1000km. This implies that migration is approximately within a country;
     - 5000km where the distance between the birth place and the death place coordinates is between 1000 and 5000km. This implies that migration is approximately within the continent;
     - 5000km+ where the distance between the birth place and the death place coordinates is over 5000km. This implies that migration has occurred on an intercontinental level.

     Each set of statistics is reset at every application run.
   - Generated pie chart describing the previously mentioned distance distribution. Every run will append an additional pie chart (see Figure 4.4b);
   - Map displaying the coordinates of the places of birth and death. Each run will create an additional layer containing the results of the selected criteria alongside a controller segment where the information that will be visualized can be set (birth markers, death markers and the path/segment between them). The colors of the markers on the map can be configured in the user interface through two separate color pickers. Also, as seen in Figure 4.5, marks are clickable and will display the individual information set together with a link to the Wikidata page of the selected entry. Also, in order to avoid overlap, a clustering mechanism has been developed (also represented in Figure 4.5);

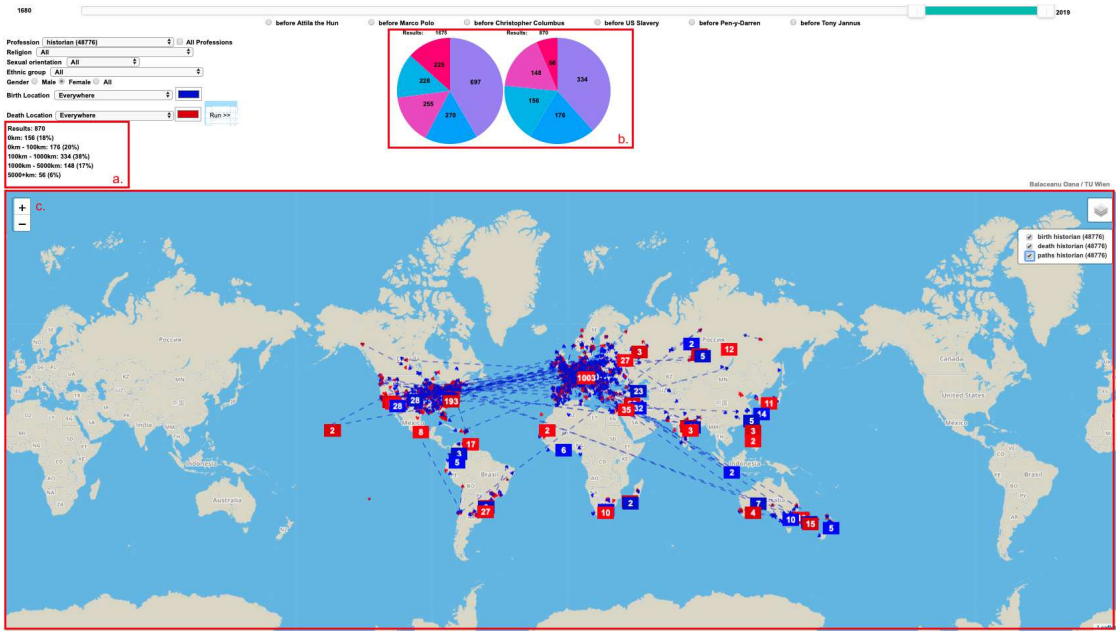

Figure 4.4: The analysis section containing the statistics summary, pie charts and the map.

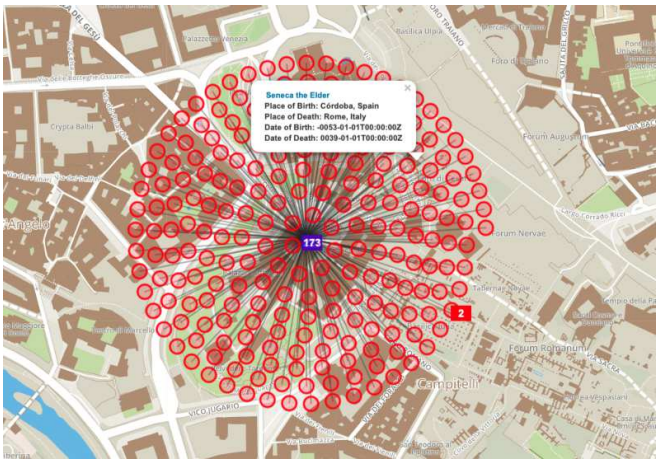

Figure 4.5: Solution for the cluster overlapping and the information label displayed on click events.

- A quantitative timeline displayed through a bar chart with the x-axis representing the Year values and the y-axis representing the Number of returned entities/people. In other words, the visualization shows the distribution of people across the previously selected time period (see Figure 4.6a). While in other cases Ajax was used to send and receive the SPARQL requests, in the

case of the bar chart diagram, the d3sparql.js library was employed in order to fetch and create the visualization in a more straightforward manner;

- Chord diagram (Figure 4.6b) showing the general flow from source countries (displayed through the thick end of the arch) to the destination country (presented through the thin end of the arch). There is no weight applied to the connections, and the diagram responds to hover actions, just as displayed in Figure 4.7 where the user hovered on the connection corresponding to the flow moving from Austria to Russia;

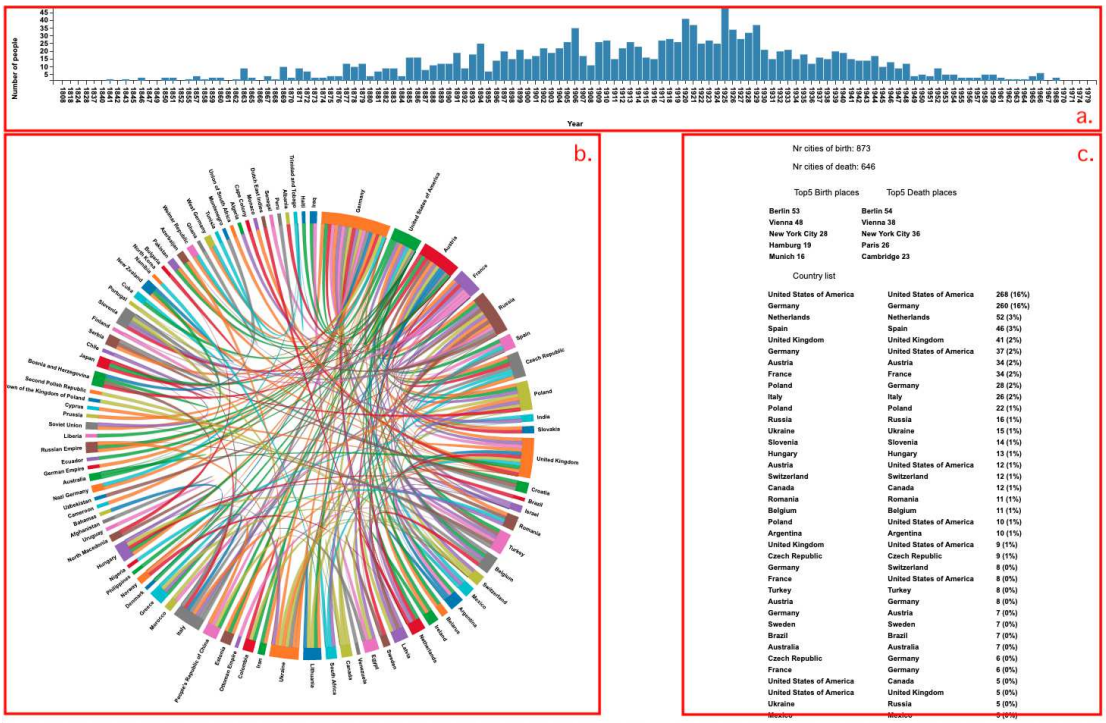

Figure 4.6: The analysis section containing the bar chart, the chord diagram and the statistics on movement.

- A section on the movement statistics, where various results from the previous diagrams are fetched and classified according to the geographical position and flow (Figure 4.6c):
  - the total number of cities of birth vs. the total number of cities of death - these numbers will help demonstrate the hypothesis according to which, the cluster formed by places of birth is larger than the one of places of death, implying that there exist cultural and economical hotspots across the world that attract people to move;
  - top 5 lists for the most popular cities of birth and cities of death. These lists show a flow on the population dynamics for important cultural and

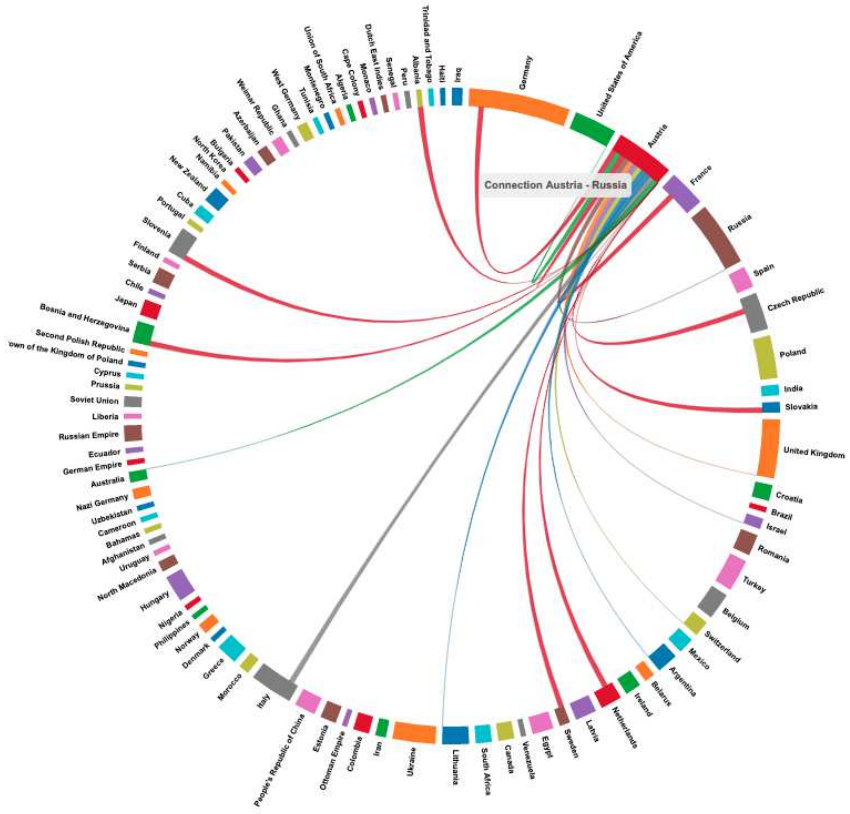

Figure 4.7: The hovering effect present on the chord diagram.

economical hotspots while also emphasizing the previously mentioned hypothesis where the cluster formed by places of birth is larger than the one of places of death. The top entries count can also be easily changed according to the analysis needs;

– top country flow used to display the international movement for the current analysis (numeric and percentage display).

3. The reference section which displays a list of all the people included in the results (Figure). Used in order to offer an individual view of the results, the list displays information on:

   - complete name of the person (label);
   - link to the corresponding WikiData page;
   - place of birth and death containing both city and country. These correspond to the time era when the person has lived (e.g., Luise Kimme was born in 1939 in Bremen, Nazi Germany [1]);

[1] Luise Kimme page `https://www.wikidata.org/wiki/Q15444301` accessed on 20.07.2019

- year of birth and death;
- distance between the locations of birth and death.

Sophie Menter	Birth: Munich, Germany	Death: Stockdorf, Germany	1846 - 1918	20km
Annemarie von Gabain	Birth: Morhange, France	Death: Berlin, Germany	1901 - 1993	848km
Ellen Swallow Richards	Birth: Dunstable, United States of America	Death: Boston, United States of America	1842 - 1911	49km
Renate Riemeck	Birth: Wrocław, Poland	Death: Alsbach-Hähnlein, Germany	1920 - 2003	948km
Edith Stein	Birth: Wrocław, Poland	Death: Auschwitz II-Birkenau concentration camp, Poland	1891 - 1942	264km
Anni Albers	Birth: Berlin, Germany	Death: Orange, United States of America	1899 - 1994	9643km
Marietta Blau	Birth: Vienna, Austria	Death: Vienna, Austria	1894 - 1970	0km
Elisabeth Grümmer	Birth: Yutz, France	Death: Warendorf, Germany	1911 - 1986	350km
Marie Curie	Birth: Warsaw, Poland	Death: Sancellemoz, France	1867 - 1934	1731km
Mary Kenneth Keller	Birth: Ohio, United States of America	Death: Dubuque, United States of America	1913 - 1985	911km
Liselotte Welskopf-Henrich	Birth: Munich, Germany	Death: Garmisch-Partenkirchen, Germany	1901 - 1979	89km
Winifred Asprey	Birth: Sioux City, United States of America	Death: Poughkeepsie, United States of America	1917 - 2007	2501km
Nina Bari	Birth: Moscow, Russia	Death: Moscow, Russia	1901 - 1961	0km
Carolina Michaëlis de Vasconcellos	Birth: Berlin, Germany	Death: Porto, Portugal	1851 - 1925	2749km

Figure 4.8: The reference section.

4.3 Results and Analyses

In this chapter, the visualization and the statistical outcome from the previous step will be analyzed, and the conclusions will be presented. The results will be compared and interpreted in accordance with the factors and dynamics that governed the selected time period.

Besides verifying that there is a decrease of the number of places of death compared to places of birth, the expected result of the master thesis is the understanding of the distribution and movement of humans while taking into account their profession, gender, country of birth and time period. This comprises and uses all the data available in the Wikidata repository, with the segment of analyzed people will including personalities that have at least data assigned to birth, death, and professional situation with labels in the English language.

The central hypothesis presented in the thesis revolves around the idea that while people are born dispersed across the world, they will be a propensity towards gathering in specific hotspots later in life. These hotspots or clusters differ according to the time era in which we are analyzing the migration dynamics and according to different migration pull/push factors and forces. An example on this matter would be Florence, an expected hotspot for artists that were active during the Renaissance period due to its political structure (also influenced by the patronage of its dominant family, the Medici), and as a consequence of the migration of Greek scholars to Italy following the Fall of Constantinople to the Ottoman Turks.[2]

There is a unique situation pertaining to each professional arena, and this thesis will, therefore, differentiate and examine it. On the same point of view, most skilled migration streams are heavily male, so as a result, women might determine different patterns and typologies separately.

[2]Encyclopaedia Britannica - Renaissance `https://www.britannica.com/event/Renaissance` accessed on 20.07.2019

Migration is activated by many various triggers, each different according to region and time era. In the modern-day world, national policies, bilateral and multilateral agreements are becoming important in facilitating the flow of skilled labor. Another novel trigger is represented by universities, that allow migration to occur in a number of ways: growth in numbers of international students studying abroad and increase in numbers of foreign students studying at home for qualifications granted by higher education institutions from developed countries. Internationalization of professions is increasingly driving people to move due to more accessible qualifications and accreditations. The training of IT professionals, in particular, falls into this category and the certificates of some companies are now recognized as more valuable than university degrees. Nursing is another example of a profession undergoing internationalization due to inadequate training of nurses in various countries and the rapidly aging populations. When the global labor market experiences shortages, there is a different move towards less control and more flexibility. The focus of the thesis is not to analyze historical reasons for why people moved within a region, but to find general patterns that can give some insights on the general behavior.

The size of the data set when all the countries, time interval, professions, and genders are taken into account is 706.769 notable people as of June 2019. Due to large data sets (when more than 20.000 entries fetched) some queries are timed out by the Wikidata's SPARQL service. Due to this limitation, the time interval can be at times divided into smaller subintervals and the results are computed accordingly. In the same way, a dataset is considered within the thesis too small and as a result, inconsistent, if the count per country is less than 100 results.

The first step of the analysis will comprise of gathering general data for all the professions with a clear differentiation per gender (male vs. female) and country. A short vs. long distance research will be done together with verifying the spreading and internationalization degrees with the mention that for a clearer analysis, the 0km and the 0-100km percentages have been merged, although the implementation is presenting them separately:

- The spreading degree is comparing the number of birth cities and death cities while having a particular country, profession, and gender in mind. A limitation to be mentioned regarding the spreading is that within Wikidata, a location might be represented as two different cities. E.g., 2nd arrondissement of Paris can appear as 2nd arrondissement or as Paris according to the returned time period or notable person. Also depending on the time period is the country to which the city belongs to. As it can be seen in Figure 4.9 , when selecting a country that had different states across ages, can return different names depending on the era. This leads to having a less aggregate result when counting the number of people born and dead in each city or country and a larger count when computing the spreading degree. When undergoing the analysis, I will display either "y", "n" or "-" as it follows:
  - "y" as in "yes", there is a spread of cities when it comes to comparing the number of cities of birth versus cities of death (or in other words, the number of cities of birth is smaller than the number of cities of death);

– "n" as in "no", there is no spread of cities when it comes to comparing the number of cities of birth versus cities of death (or in other words, the number of cities of birth is greater than the number of cities of death);

– "-" as in no changes found when comparing the number of cities of birth versus cities of death (or in other words, the number of cities of birth is equal to the number of cities of death).

Country list

China	People's Republic of China	32 (24%)
China	United States of America	5 (4%)
China	Jin dynasty	2 (1%)
China	Han dynasty	2 (1%)
China	Sui dynasty	2 (1%)
China	Song dynasty	2 (1%)
China	Tang Empire	2 (1%)
China	Cao Wei	2 (1%)
China	Xia	2 (1%)
China	Northern Zhou	2 (1%)
China	Later Liang dynasty	2 (1%)
China	Northern Wei	2 (1%)
China	Later Han dynasty	2 (1%)
China	Later Tang Dynasty	2 (1%)
China	Western Wei	2 (1%)
China	Western Yan	2 (1%)
China	Later Jin Dynasty	2 (1%)
China	Later Zhou dynasty	2 (1%)
China	Yan	2 (1%)
China	Former Qin	2 (1%)
China	Han Zhao	2 (1%)
China	Later Qin	2 (1%)
China	Later Zhao	2 (1%)
China	Liu Qi	2 (1%)
China	Qi (Huang Chao)	2 (1%)
China	Taiwan	2 (1%)
China	Japan	1 (1%)
China	Mongolia	1 (1%)
China	North Korea	1 (1%)
China	United Kingdom	1 (1%)
China	France	1 (1%)
China	Canada	1 (1%)
China	India	1 (1%)

Figure 4.9: Different returned state forms when querying for China

Another aspect to be noted when analyzing spreadability is that:

– when the analysis contains one country as a source and the whole world as a destination, the number of places of birth is smaller than the number of places of death (since the whole world contains a bigger pool of possibilities when selecting cities). Some exceptions in this case are:

1. small countries that contain an important hotspot. This usually includes countries like France (with Paris as a hotspot) or Italy (with Rome as a hotspot);
2. countries with a very large surface (e.g., Russia, China).

  - when the analysis contains one country as a source and one country as a destination, the number of places of birth is greater than the number of places of death.

- The internationalization degree is verifying if the top destination cities contain cities that are not present in the birth country. When undergoing the analysis, I will enumerate the foreign cities present in the top five destination cities or "-" if there are none.

For each analysis, also the distribution of the last generation will be considered when the dataset is consistent enough in order to see the changes that occurred in the last decades and how technology and ease of communication and transportation have enabled people to migrate. By last generation, I am referring to the notable people born starting with the year 1945 and have a date of death associated.

Within the analysis, two different time intervals will be taken into consideration:

- main time interval, expanding from 500BC until nowadays;
- the last generation, a time interval beginning from 1945, used in order to verify if the introduction of transportation and communication means within the general usage has in any way, impacted migration (especially the very long distance migration).

A series of countries as birth location from different regions of the world and of different sizes will be taken into account (for better readability, the European and Non-European results are separated into two different sections):

- Europe countries:
  - Germany;
  - France;
  - Italy;
  - Spain;
  - Poland;
  - United Kingdom (UK);
  - Ireland;
- Non-European countries:
  - United States of America (USA);
  - Russia;
  - China;

  - Brazil
  - South Africa;
  - Australia;
  - Japan.

Once the general situation for male and female is analyzed, a separate analysis will be done for a set of three different professions:

- painters for both men and women;
- engineers for both men and women;
- politicians for both men and women;

In short, for each of the previously mentioned countries and professional categories, I will provide for the main time period and for the last generation, the following results:

- total result count together with the ratio of the last generation within the main time period (or in other words, have the last decades influence the ease of becoming notable for a particular country, gender or profession?);
- same-country movements flows or the percentages of people who were born and died in the same country;
- distance distribution (with bolded values for the maximum and minimum values) as it follows:
  - 0 - 100km for local movement;
  - 100 - 1000km for regional movement;
  - 1000km - 5000km for intracontinental movement;
  - 5000km or more for intercontinental movement.
- top international destinations. The following abbreviations for major cities will be used in tables:
  - LA - Los Angeles
  - NYC - New York City
  - NJ - New Jersey
  - LND - London
  - CDMX - Mexico City
  - SF - San Francisco
  - HK - Hong Kong

  - Phil. - Philadelphia
  - Strasb. - Strasbourg
  - Edinb. - Edinburgh
  - P. Escondido - Puerto Escondido

- spreading indicator for country-to-world;
- same-country spreading together with the difference in percentage between the number of cities of birth and cities of death. The general and main hypothesis of the thesis is that these percentage show a decrease.
- for the female datasets only, percentage of the female segment in the total population pool;

4.3.1 Male General Overview

The analysis begins with fetching the males distribution according to the difference in distance length when comparing the birth and death places (data as of June 2019). This will represent a reference dataset. Some general observations include:

- the total number of male professionals for the entire time period and all regions with all the attributes present is 637.112;
- all the queries have returned consistent datasets, with the largest represented by the US with 93.681 results and the lowest represented by South Africa with 1.464 results;
- the numbers change when querying for the last generation (people born starting with 1945) with the largest dataset represented by the US with 4.619 and the smallest represented by Ireland with 91 results;
- this brings the need to evaluate how the dataset size has changed when comparing the result count for the entire time interval and the last generation (or in other words, what is the proportion of notable people in the last generation). The result count for Germany is the smallest (with 1,7%), while the result count for South Africa is the greatest (with 9,6%). This means that South Africa was greatly influenced by the latest decades accomplishments when offering the world notable people.

In a more detailed approach, Table 4.1 shows the distribution for the European male professional population, with Germany as the country with the most results (90.926 men) and Ireland with the smallest data set (3.032 men).

As it can be noticed, the country with the most locally settled male population among the ones analyzed within Europe is Austria (54%) while the smallest local community is

	Austria	Germany	France	Italy	Spain	Poland	UK	Ireland
Total	14.110	90.926	62.349	49.494	21.298	31.502	40.439	3.032
Total '45	304	1.549	1.120	1.544	1.020	1.167	1.362	91
%	2,1%	1,7%	1,8%	3,1%	4,8%	3,7%	3,3%	3%
Same-country	63,1%	77,4%	83%	82,8%	79,9%	26,7%	67%	26,5%
Same-country '45	59,8%	70,6%	78,6%	81,6%	86,6%	44%	64,6%	53,8%
0-100km	**54%**	41%	43%	53%	46%	22%	35%	**18%**
0-100km '45	50%	42%	**36%**	53%	**57%**	43%	37%	38%
100-1000km	35%	50%	48%	38%	38%	55%	41%	35%
100-1000km '45	34%	49%	44%	35%	33%	39%	34%	35%
1000-5000km	5%	5%	5%	5%	6%	18%	5%	10%
1000-5000km '45	7%	4%	7%	5%	3%	11%	7%	8%
5000+km	6%	4%	4%	**3%**	10%	5%	20%	**36%**
5000+km '45	9%	**5%**	8%	6%	7%	8%	**22%**	19%

Table 4.1: Relative distribution for the male professionals in Europe

represented by Ireland (18%). These percentages have become in general more uniform starting with the generation from 1945, with the least local being represented by France and the most local being in Spain. Regarding long-distance movements, Italian men have generally avoided 5000 or more kilometers travels, with only 3% leaving the continent, while in Ireland, 36% of notable men preferred very long distances migration. Again these percentages have in general increased since 1945, except for Spain and Ireland.

As for the same-country migration, Poland and Ireland display a predisposition of not settling within the birth country (22% and 18%), while Austria is the country with the largest percentage (54%). The case of Austria is even more evident considering it is the second smallest sized country after Ireland.

Table 4.2 displays the previously enounced spreading and internationalization degrees for each European country chosen for the analysis. Regarding the popularity of international destinations, only Poland and Ireland displayed preferability for top destinations outside their countries (Poland with Berlin and Ireland with New York City, London, Philadelphia, Toronto and only Dublin as an Irish destination). When looking at migration preferences starting with year 1945, the United Kingdom has had a significant number of male professionals moving in the United States and Belfast, while Ireland kept London as the main destination and changed the other United States destinations with Miami.

Regarding spreadability, all the European countries except Austria, United Kingdom, and Ireland had the number of birth places smaller than the number of death places. Spreadability can be also observed in the "Same country spread" row, where the thesis hypothesis is confirmed as in every analyzed country, there is a decrease ranging from 60.6% to 10%. In the eventuality of source and destination represented by the entire world, the numbers vary as it follows:

	Austria	Germany	France	Italy	Spain	Poland	UK	Ireland
Top International Destinations	-	-	-	-	-	Berlin	-	NYC LND Phil. Toronto
Top International Destinations '45	-	-	-	-	-	-	LA Belfast	LND Miami
Spread	y	n	n	n	n	n	y	y
Spread '45	y	n	y	n	n	n	y	y
Same country	1.335	10.557	9.314	5.235	2.966	2.565	5.208	294
spread	935	6.601	7.082	3.609	1.671	1.009	4.685	230
Change	-30%	-37,5%	-24%	-31%	-43,6%	-60,6%	-10%	-21,7%

Table 4.2: Spreading and internationalization degree for European male professionals

- the total number for places of birth: 114.998;
- the total number for places of death: 70.907;
- decrease: 38,34%.

When analyzing non-European countries (see Table 4.3), we can observe that with the exception of Brazil, the tendency to migrate on long distance is generally higher than in Europe. Among those analyzed, China is the country with the lowest predisposition to stay in the same location, while the greatest is represented by Brazil. Brazil is also the country that leaves the continent the least. When it comes to medium to long-distance movement, South Africa and Australia display a more limited predisposition which coincides with the countries being isolated geographically or surrounded by less developed nations.

Moving towards the period starting with the year 1945, we can observe that all the percentages for local migrations have increased with the exception of the US and Australia. Same applies for long-distance migration, but with the exception of South Africa. This, in general, suggests that medium migration has decreased in order to compensate for the changes.

The same-country migration is less scattered than in Europe, with percentages ranging from 47% for China and 82% for Brazil. These percentages have generally increased, starting with the last generation with the exception of Australia and Japan, where the decrease was approximately 1%.

When it comes to major cities outside the country of origin (see Table 4.4), only Russia and South Africa has cases where people moved and died in such places, and all these

	US	Russia	China	Brazil	South Africa	Australia	Japan
Total	93.681	17.265	3.178	5.982	1.464	8.300	4.396
Total '45	4.619	1.572	188	518	151	289	294
%	4,9%	9,1%	5,9%	8,6%	10,3%	3,4%	6,7%
Same-country	77,3%	65,3%	47%	82,2%	51,5%	81,3%	71,7%
Same-country '45	90,6%	78,8%	56,3%	86,4%	76,1%	80,2%	70,7%
0-100km	33%	28%	**18%**	38%	26%	**42%**	39%
0-100km '45	**31%**	40%	34%	38%	**47%**	33%	45%
100-1000km	31%	31%	50%	36%	36%	34%	27%
100-1000km '45	29%	22%	38%	37%	29%	34%	36%
1000-5000km	28%	31%	14%	20%	14%	13%	43%
1000-5000km '45	30%	26%	10%	19%	11%	18%	8%
5000+km	8%	10%	18%	**5%**	**25%**	11%	10%
5000+km '45	10%	12%	**18%**	**6%**	14%	15%	11%

Table 4.3: Relative distribution for the male professionals outside Europe

locations are in Europe. The situation changes as of 1945, where there is no international destination in the top five death places for any of the non-European countries included in the analysis.

	US	Russia	China	Brazil	South Africa	Australia	Japan
Top International Destinations	-	Paris Kiev	-	-	LND	-	-
Top International Destinations '45	-	-	-	-	-	-	-
Spread	n	n	n	n	y	n	n
Spread '45	y	n	n	n	y	y	n
Same country	14.134	2.605	620	1.060	237	1.365	762
spread	9.704	866	256	498	185	1.311	522
Change	-31,3%	-66,7%	-58,7%	-53%	-21,9%	-3,9%	-31,5%

Table 4.4: Spreading and internationalization degree for Non-European male professionals

The hypothesis of the thesis is again confirmed for non-European countries, where in all the cases, the country-to-country spreading suggests a decrease ranging from approximately 4% for Australia to 67% for Russia.

4.3.2 Female General Overview

The analysis continues with fetching the female distributions according to the difference in distance length when comparing the birth and death places (data as of June 2019). This will represent a reference dataset. Some general observations include:

- the total number of female professionals for the entire time period and all regions with all the attributes present is 69.657;
- the previous observation leads to a fast conclusion - the male-female ratio is gravitating around the value range 9:1 (total datasets contains 637.112 for men and 69.657 for women). In other words, for every one notable woman that has been included in Wikidata, there are nine notable men. For exact country on each female analysis, I will mention the proportion within the entire dataset, but as a small mention, the closest male:female ratio is present in the case of South Africa, with five notable men for every notable woman, and the furthest ratio is in the case of Italy, with 14 men for every one woman.
- all the queries have returned consistent datasets, with the largest represented by the US with 14.277 results and the lowest represented by South Africa with 271 results;
- similarly to the case of male professionals, the numbers change when querying for the last generation (people born starting with 1945) with the largest dataset represented by the US with 1.232 results and the smallest represented by Ireland with 20 results;
- this brings again the need to evaluate how the dataset size has changed when comparing the result count for the entire time interval and the last generation. In the case of female professionals, the smallest result appear to be for Austria (with 3,6%), while the result count for Japan is the greatest (with 2,3%). Besides this, in general, the last generation percentage is a lot higher outside Europe, with rates beginning at 8,6% for the United States.

When it comes to the distribution related to the length of travel, women, in general, are more predisposed to long-distance migration, with a smaller number of persons staying locally, and more significant amounts traveling on longer distances.

Some significant differences in percentages are shown in Table 4.5 in the case of Austria where the number of women migrating on long distances is almost double the number of men (12% vs. 6%) and Ireland where the number of men migrating on long distances is virtually double the number of women (36% vs. 19%).

Starting with 1945, the most latent migration occurred in Spain, while the least targeted France, but with similar values with the United Kingdom and Ireland. Similarly, for long distance, the highest predisposition happened in the case of Ireland, while the smallest targeted Austria.

The same-country migration varies from 34,8% in the case of Poland and 77,9% in the case of Italy. These percentages generally increase for the last generation, with the exception of France and Ireland. In the eventuality of source and destination represented by the entire world, the numbers vary as it follows:

	Austria	Germany	France	Italy	Spain	Poland	UK	Ireland
Total	1.870	7.760	4.957	3.494	2.272	3.190	4.959	434
Total '45	67	315	197	221	212	199	265	20
%	3,6%	4%	4%	6,3%	9,3%	6,2%	5,3%	4,6%
Same-country	46,4%	67,9%	77,4%	77,9%	77,1%	34,8%	66,8%	38,9%
Same-country '45	59,7%	72%	77,1%	84,6%	86,3%	48,7%	66,8%	35%
0-100km	31%	32%	40%	**47%**	45%	**24%**	32%	29%
0-100km '45	32%	28%	35%	53%	**58%**	36%	36%	**25%**
100-1000km	37%	55%	48%	41%	38%	52%	44%	42%
100-1000km '45	48%	61%	49%	35%	33%	46%	40%	35%
1000-5000km	11%	7%	6%	7%	6%	17%	6%	10%
1000-5000km '45	7%	6%	7%	6%	2%	13%	4%	10%
5000+km	12%	**4%**	6%	5%	10%	7%	**19%**	**19%**
5000+km '45	**3%**	5%	10%	6%	6%	6%	20%	**30%**
% female	11,7	7,8	7,4	6,6	9,6	10,1	11	12,5

Table 4.5: Relative distribution for the female professionals in Europe

- the total number for places of birth: 19.127;
- the total number for places of death: 13.573;
- decrease: 39%.

	Austria	Germany	France	Italy	Spain	Poland	UK	Ireland
Top International Destinations	Munich Berlin NYC	Vienna	-	-	CDMX Buenos Aires	Berlin Paris	NYC LA	London Paris NYC
Top International Destinations '45	Munich	Vienna	-	-	-	-	NYC	NYC Geneva Rome
Spread	y	y	y	y	n	y	y	y
Spread '45	n	n	y	n	n	n	y	y
Same country	182	1.386	1.149	727	575	371	1.209	86
spread	133	1.191	1.025	582	371	181	1.134	70
Change	-26,9%	-14%	-10,7%	-19,9%	-35,5%	-51,2%	-6,2%	-18,6%

Table 4.6: Spreading and internationalization degree for European female professionals

Women are also showing a more clear degree of internationalization and spreading, as displayed in Table 4.6. In the case of country-to-world spreading, only for Spain the

number of places of birth was higher in comparison with the number of places of death (this is usually not the case for smaller countries when analyzing a specific country vs. the whole world situations). Generally, all the countries with the exception of France and Italy had at least one international location in their top 5 places of death. Spain and the United Kingdom only had destinations in South America and the United States, while Austria and Ireland had destinations in both Europe and across the ocean in the States. Germany and Poland had destinations only in Europe.

When it comes to the last generation, the international presence in the top five lists decreases with Germany and Austria exchanging migrants in Vienna and Munich, and United Kingdom having people leaving the country for New York City. Ireland is again present with three locations: New York City, Geneva, and Rome.

As for the general spreading degree, all the country face a decrease when counting country-to-country places of birth vs. places of death with percentages ranging from 6,2% for the United Kingdom and 51,2% for Poland.

Moving forward to female professionals outside Europe, we can observe again that the trend is for women to travel proportionally more on longer distances. Table 4.7 shows a higher difference of proportions when it comes to South African migrations when differentiating it according to gender.

	US	Russia	China	Brazil	South Africa	Australia	Japan
Total	14.277	2.588	506	672	271	837	530
Total '45	1.232	235	45	106	26	92	108
%	8,6%	9%	8,9%	15,7%	11,5%	11%	20,3%
Same-country	78,5%	60,3%	30,2%	82,3%	44,6%	67,4%	70%
Same-country '45	78%	72,7%	57,7%	82%	57,7%	76,2%	72%
0-100km	26%	28%	**16%**	40%	25%	32%	**44%**
0-100km '45	26%	35%	**21%**	36%	27%	36%	**45%**
100-1000km	29%	30%	45%	40%	27%	30%	36%
100-1000km '45	29%	17%	43%	44%	20%	21%	36%
1000-5000km	35%	32%	10%	14%	17%	15%	5%
1000-5000km '45	35%	34%	0%	18%	0%	22%	3%
5000+km	9%	10%	29%	**6%**	**31%**	22%	14%
5000+km 45	11%	15%	36%	**3%**	**53%**	20%	16%
% female	13,2	13	13,7	10	15,6	9,2	10,7

Table 4.7: Relative distribution for the female professionals outside Europe

While 25% of men travel on intercontinental distances, about 31% of the women do the same. On the same subject, 11% of men chose to remain in the same place, while 13% of women do that. This suggests that South African men are inclined to travel short to medium distances, which means that their travels remain within the country or the African continent more than women do.

China also faces a similar situation, with similar percentages for same locations, and a high increase for very long distance migration (18% for men vs. 29% for women). Similar increases in long-distance migration in the case of the female population can be seen for the US (9% vs. 8%), China (29% vs. 18%), Brazil (6% vs. 5%).

Starting with 1945, the extremes for short and long distances remain the same. China has the least local female population, while Japan has the most stable one. Brazil is the country least predisposed for very long distance migration, while South Africa is the most.

Same-country migration displays percentages ranging from 44,6% for South Africa to 78,5% for the United States. The last generation brings increases, with the exception of the United States and Brazil, where the decrease is under 0,5%.

	US	Russia	China	Brazil	South Africa	Australia	Japan
Top International Destinations	-	Paris Kiev Helsinki	NYC	-	LND	LND	-
Top International Destinations '45	-	Paris	Washington DC LA	-	LND Sydney	LA	Paris
Spread	n	n	n	n	y	y	n
Spread '45	y	n	-	n	n	y	n
Same country	3.503	433	75	217	69	262	169
spread	2.712	199	32	90	39	217	114
Change	-22,6%	-54%	-57,3%	-58,5%	-43,5%	-17,2%	-32,5%

Table 4.8: Spreading and internationalization degree for female professionals outside Europe

When it comes to the spreadability degree, all the countries confirm the hypothesis according to which the number of cities of birth is higher than the number of cities of death with decreases ranging from 17% to approximately 59%.

Russia shows to have three non-Russian cities in its top five international destinations (Paris, Kiev, Helsinki), while China has New York City as a top destination, and South Africa has London. The United States of America and Brazil have all their main destinations inland.

When it comes to the generation born from 1945 onwards, the top international destinations are present for the same countries as for the entire timeframe and contain Paris for Russia, Washington D.C. and Los Angeles for China, London, and Sydney for South Africa and Los Angeles for Australia and newly introduced is Paris for Japan.

4.3.3 Male Painter General Overview

Moving away from the general overview, we start the introduction of profession analysis. Since the profession of being a painter represents a small part of the entire dataset, there will be some disadvantages when analyzing each portion (i.e., a smaller dataset which will lead to not so exact and detailed results and percentages. The trend, however, should be in general respected). Some general observations include:

- the total dataset for the entire male painter population is 43.796;
- all the queries have returned consistent datasets, with the largest represented by Germany with 7.585 results and the lowest represented by South Africa with 21 results;
- the numbers change when querying for the last generation with the largest dataset represented by the US with 72 results and the smallest represented by Ireland with no results. This leads to labeling the dataset for female painters born and dead after 1945 as being inconsistent.
- a significant difference in the case of painters is represented by the ratio between the entire time period and the latest generation. While in the case of analysis for all the professions we had percentages ranging between 1,7% and 4,8%, in the case of painters specifically, the percentages range between 0,4% for Italy (without considering Ireland's empty dataset) and 6% for Australia.

	Austria	Germany	France	Italy	Spain	Poland	UK	Ireland
Total	1.619	7.585	5.849	5.785	1.950	1.962	2.272	138
Total '45	9	65	35	24	50	24	15	0
%	0,5%	0,8%	0,6%	0,4%	2,5%	1,2%	0,6%	0%
Same-country	62%	71%	82%	80%	81%	26%	68%	26%
Same-country '45	78%	83%	85%	92%	96%	54%	53%	0%
0-100km	52%	46%	47%	**62%**	52%	22%	42%	**21%**
0-100km '45	44%	40%	31%	62%	**70%**	67%	20%	**0%**
100-1000km	31%	47%	45%	30%	36%	54%	40%	49%
100-1000km '45	33%	51%	54%	29%	28%	21%	40%	0%
1000-5000km	5%	5%	4%	5%	7%	20%	6%	10%
1000-5000km '45	0%	5%	9%	8%	0%	4%	7%	0%
5000+km	**3%**	**3%**	**3%**	**3%**	6%	4%	13%	**20%**
5000+km '45	22%	5%	6%	**0%**	2%	8%	**33%**	**0%**

Table 4.9: Relative distribution for the male painters in Europe

Analyzing the percentages for the most local and the most international countries (Table 4.9), we can conclude that Italy is the country where people preferred to stay in the same

region for the entire time period, while since 1945, the place was taken by Spain. When it comes to the country that remains the least locally, the place is taken by Ireland.

On the other extreme, the most global countries are Ireland for the entire time frame and the United Kingdom for the period starting with 1945. The least international spot is shared by Austria, Germany, France, and Italy and specifically by Italy since 1945 where nobody notable has decided to migrate for more than 5000 kilometers.

	Austria	Germany	France	Italy	Spain	Poland	UK	Ireland
Top International Destinations	Munich	-	Rome	-	Paris	Berlin Paris Munich	NYC Paris	LND NYC Phil. Rome
Top International Destinations '45	Lima	-	-	-	-	NYC	NYC	-
Spread	y	n	y	n	n	n	y	y
Spread '45	n	n	y	n	-	-	n	-
Same country spread Change	224 163 -27,2%	1.535 1.126 -26,6%	1.232 1.062 -13,8%	1.086 706 -35%	456 262 -42,5%	217 116 -46,5%	521 430 -16%	12 7 -42%

Table 4.10: Spreading and internationalization degree for European male painters

Looking at the degree of internationalization, Table 4.10 denotes more destinations than in the general analysis. Only Germany and Italy don't have an international destination in its top five, while Ireland remains the country with only one local place (Dublin) in its top - the only difference is Toronto being replaced by Rome.

The last generation has few international destinations, only with Austria (Lima), Poland (New York City) and United Kingdom (New York City). This is also because the dataset for the time period starting with 1945 is very small or empty.

Again, the country to country migration faced a decrease when comparing the number of cities of birth with the cities of death, ranging from approximately 14% to 45%.

Outside Europe (Table 4.11), we can observe that the least local nation from the ones analyzed it represented by China, while the most local is again Brazil. The situation changes for the last generation, where both Brazil and China are taking the highest spot (besides Australia, which has only one notable person that died in the same country of birth, making the ratio 100%).

In the same manner, the most global non-European country is South Africa, while the least ones are Russia and China. The situation changes again for the last generation,

	US	Russia	China	Brazil	South Africa	Australia	Japan
Total	2.400	1.337	219	233	21	133	295
Total '45	72	70	5	10	1	8	6
%	3%	5,2%	2,2%	4,3%	4,7%	6%	2%
Same-country	70%	64%	36%	85%	41%	68%	69%
Same-country '45	77%	79%	40%	100%	100%	88%	33%
0-100km	29%	30%	**18%**	**50%**	24%	41%	47%
0-100km '45	31%	59%	60%	60%	**100%**	38%	**17%**
100-1000km	32%	27%	55%	29%	10%	23%	35%
100-1000km '45	33%	14%	0%	30%	0%	13%	33%
1000-5000km	24%	33%	13%	10%	14%	10%	5%
1000-5000km '45	28%	16%	20%	10%	0%	38%	0%
5000+km	14%	**10%**	14%	**10%**	**52%**	26%	14%
5000+km '45	8%	11%	20%	**0%**	**0%**	13%	**50%**

Table 4.11: Relative distribution for the male painters outside Europe

where half of the notable painter from Japan moved for more than 5000 kilometers while none of the South African and Brazilian artists did the same.

The same-country migration has gravitated around the values of 36% for China and 85% for Brazil with increases for every country starting with the last generation, except for Japan.

	US	Russia	China	Brazil	South Africa	Australia	Japan
Top International Destinations	Paris	Paris Helsinki	-	Paris	-	London	Paris
Top International Destinations '45	Paris	-	Rennes Singapore	-	-	-	Paris NYC Berlin
Spread	n	n	n	n	y	y	n
Spread '45	n	y	-	-	-	-	-
Same country	785	327	47	88	6	51	106
spread	646	129	21	35	6	40	78
Change	-17,7%	-60,5%	-55,3%	-60,2%	0%	-21,6%	-26,2%

Table 4.12: Spreading and internationalization degree for Non-European male painters

Looking at the internationalization and spreading degree, Paris represents a top destination for four different countries (the United States, Russia, Brazil, and Japan), with additionally Helsinki for Russia and London for Australia. Paris remains in the top also for the last generation (for the United States and Japan), while China has Rennes and Singapore as its main destinations. Japan brings additionally New York City and Berlin.

Regarding the change in the number of places of birth vs. places of death, all the countries are affected by a decrease ranging from 60,5% to 0% (in the case of South Africa, due to its very small dataset.

4.3.4 Female Painter General Overview

For female painters we have:

- the total dataset regardless or location for the entire time period of 6.532. This leads to a ratio of approximately one woman to every seven men;
- all the queries have returned relatively consistent datasets, with the largest represented by the US with 812 results and the lowest represented by South Africa with ten results;
- the last generation, on the other hand, has returned very small datasets with numbers ranging from 23 results for the US and 0 results for Brazil, South Africa, Japan and Ireland.
- the dataset change for the last generation shows percentages from 0,5% for the United Kingdom to 10% for China (except the countries that returned empty datasets for the last generation).

	Austria	Germany	France	Italy	Spain	Poland	UK	Ireland
Total	196	732	574	242	129	238	384	33
Total '45	5	23	7	3	11	5	2	0
%	2,6%	3,1%	1,2%	1,2%	8,6%	2,1%	0,5%	0%
Same-country	57%	68%	80%	74%	75%	25%	62%	55%
Same-country '45	60%	83%	29%	100%	82%	60%	100%	0%
0-100km	48%	36%	48%	52%	49%	**19%**	33%	**63%**
0-100km '45	**80%**	34%	**0%**	67%	72%	40%	50%	**0%**
100-1000km	31%	51%	43%	37%	33%	41%	43%	24%
100-1000km '45	0%	52%	43%	33%	27%	40%	50%	0%
1000-5000km	11%	8%	4%	8%	8%	32%	9%	6%
1000-5000km '45	20%	13%	43%	0%	0%	0%	0%	0%
5000+km	10%	5%	5%	**3%**	11%	8%	**14%**	6%
5000+km '45	**0%**	**0%**	14%	**0%**	**0%**	**20%**	**0%**	**0%**
% female	10,8	8,8	8,9	4	6,2	10,8	14.4	19,3

Table 4.13: Relative distribution for the female painters in Europe

Table 4.13 shows that the most local place within the entire time period is represented by Ireland, while the least one is represented by Poland. The case of Ireland here is also due to the fact that the dataset if very small, skewing the results greatly. If Ireland

is not taken into account, the next most local position is Italy, with around 52% of people choosing to stay and eventually dying in the immediate nearby location of their city of birth. The percentages change drastically with the last generation, where 80% of Austrian women remained close to their place of birth, while none of the women in France and Ireland did the same (Ireland again skews the results through the fact that its dataset is empty).

The situation changes for the most global locations, with the United Kingdom having the most significant percentage (14%) of women traveling on distance longer than 5.000km, and Italy again being the least global with about 3%. Due to the small or inexistent datasets, many countries (Austria, Germany, Italy, Spain, the United Kingdom, and Ireland) have no women traveling for more than 5.000km in the last generation.

The same-country migration remains above 50% with the exception of Poland (25%), and in all the cases except France and Ireland, this percentage has increased in the last generation.

Also to be mentioned is that the most significant ratio of female painters within Europe appears in Ireland with almost 20% while the smallest is in Italy, with only 4%.

	Austria	Germany	France	Italy	Spain	Poland	UK	Ireland
Top International Destinations	Paris LND	-	NYC	Paris	CDMX Paris	Paris Berlin LND	NYC Toronto	LND Rome Edinb. Toronto
Top International Destinations '45	Cologne	Strasb. Vienna	Amsterdam Madrid Fallbrook Jerusalem	-	-	CDMX	-	-
Spread	y	y	y	y	y	y	y	y
Spread '45	-	n	y	-	-	-	-	-
Same country	33	227	194	73	54	34	144	10
spread	29	221	200	63	47	25	129	11
Change	-12,1%	-2,6%	+3%	-13,7%	-13%	-26,5%	-10,4%	+10%

Table 4.14: Spreading and internationalization degree for European female painters

Looking at the level of internationalization and spreading (Table 4.14), we can see three unique aspects:

- There is a great selection of international destinations, including for countries where this is usually not the case (e.g., France or Italy). These destinations become quite obscure for the last generation, mainly due to the small datasets (e.g., Fallbrook for France);

- all the European countries display a positive spreading degree when comparing the country pool to the whole world pool for the entire time period, and inexistent spreading for the last generation;

- France and Ireland display an increase when comparing the country-to-country spread. This disproves the thesis statement, but after a closer look, we can understand why this phenomenon is happening:

  – France: the increases of 3% happens due to the fact that while Paris remains the major hotspot of the country with 176 people being born there and 196 people dying there, the next main cities do not have such results: 20 people were born in Lyon, and 5 died there, 8 people were born in Strasbourg, and 6 people died there and so forth. A snippet of the top cities of birth and death can be seen in Figure 4.10, where to the previously enounced causes we can add the fact that the number of people being born and dying in Paris is actually higher the initially computed due to the fact that specific Parisian neighborhoods and districts tend to be taken separately for certain period of time (e.g., 6th arrondissement of Paris);

Top5 Birth places	Top5 Death places
Paris 176	Paris 196
Lyon 20	Neuilly-sur-Seine 6
Strasbourg 8	Strasbourg 6
Bordeaux 7	Sèvres 5
10th arrondissement of Paris 6	Lyon 5
Marseille 5	14th arrondissement of Paris 4
Dijon 4	6th arrondissement of Paris 4
Versailles 4	9th arrondissement of Paris 4
Saint-Mandé 3	15th arrondissement of Paris 4
Cherbourg 3	Nantes 4
4th arrondissement of Paris 3	16th arrondissement of Paris 3
2nd arrondissement of Paris 3	Tours 3
8th arrondissement of Paris 3	10th arrondissement of Paris 3

Figure 4.10: Explanation of the number of cities increase percentages in the case of France for the same-country spread

  – Ireland: the increase of 10% happens due to the fact that similarly to France, there are many cities returned as standalone locations when they actually are part or suburbs of bigger cities. After analyzing Figure 4.11, we can observe that the towns that are not common in both places of birth and places of death are Ballitore, Glenageary, and Navan (for places of birth) and Rathgar, Milltown, Edgeworthstown and Ranelagh (for places of death). Ballitore, Rathgar, and Ranelagh are all parts of Dublin, canceling the initial increase.

Table 4.15 emphasizes the fact that at least 60% of the female painters remained in the country of birth (with the highest percentage being represented by Russia with 95%),

Figure 4.11: Explanation of the number of cities increase percentages in the case of Ireland for the same-country spread

while for the last generation, wherever there is a dataset, at least 94% of them remained in their mother country.

	US	Russia	China	Brazil	South Africa	Australia	Japan
Result count	812	194	20	35	10	70	17
Result count '45	23	7	2	0	0	3	0
%	2,8%	3,6%	10%	0%	0%	4,3%	0%
Same-country	76%	95%	60%	84%	88%	82%	83%
Same-country '45	94%	100%	100%	0%	0%	100%	0%
0-100km	38%	29%	**10%**	43%	40%	31%	**53%**
0-100km '45	22%	**43%**	**0%**	**0%**	**0%**	**0%**	**0%**
100-1000km	26%	25%	10%	40%	30%	33%	18%
100-1000km '45	22%	0%	0%	0%	0%	67%	0%
1000-5000km	25%	39%	15%	9%	20%	16%	0%
1000-5000km '45	35%	43%	100%	0%	0%	33%	0%
5000+km	11%	**8%**	**65%**	9%	10%	20%	29%
5000+km '45	**22%**	14%	**0%**	**0%**	**0%**	**0%**	**0%**
% female	25,3	14,5	8,4	13	32,2	34,5	5,4

Table 4.15: Relative distribution for the female painters outside Europe

The most local country is represented by Japan with 53%, while the least one is in the case of China with 10%. Long-distance wise, 65% of the Chinese women traveled for 5000km or more, while only 8% of the Russian women did the same thing.

Similarly to the case of European female painters, we find in the analysis that female painters outside Europe are also going to more cities outside their country of birth, with the last generation choosing unexpected locations (e.g., Pinneberg). With the exception of Brazil which is one of the most local countries among all the ones analyzed, all of

	US	Russia	China	Brazil	South Africa	Australia	Japan
Top International Destinations	-	Paris Berlin Helsinki	NYC Paris Montreal	-	Zimbabwe	London Paris	São Paulo
Top International Destinations '45	P. Escondido Lincolnshire	Berlin Paris Pinneberg	Manila	-	-	-	-
Spread	y	y	y	n	y	-	y
Spread after '45	-	y	-	-	-	-	-
Same country	312	39	3	20	7	31	10
spread	294	22	2	8	5	26	9
Change	-5,7%	-43,6%	-3,33%	-60%	-28,6%	-16,1%	-10%

Table 4.16: Spreading and internationalization degree for Non-European female painters

them are displaying a degree of spreading or no degree whatsoever, and all of them are showing a decrease when comparing the number of places of birth with the number of places of death in a country setting.

4.3.5 Male Engineer General Overview

Some introductory observations regarding the male engineers are:

- the total dataset regardless or location for the entire male population and time period is 9.067;
- the largest dataset is represented by the US with 1.530 results, and the smallest is represented by South Africa with 18 results;
- the last generation is again relatively inconsistent, with values ranging from 1 entry in China and South Africa to 59 entries in the United States. This leads to a ratio ranging from 0,4% in Italy to 9,8% in Japan when compared to the entire time period.

In Table 4.17, we can notice that again, the most local country for both the entire time frame and the last generation is represented by Italy while the least one is represented by Ireland. Similarly, the least global ones are France and Italy, while the most global one is again Ireland.

The same-country spreading ranges from the value of 40% for Ireland and 83% for France, while in the case of the last generation, all the percentages decreased with the exception of Spain.

Also similarly to the situation of painters, the international destination set is more consistent, including for the generation born after 1945, with places that are not obviously

%	Austria	Germany	France	Italy	Spain	Poland	UK	Ireland
Total	160	1.340	1.226	998	253	588	1.006	53
Total '45	8	9	15	4	8	21	18	3
%	5%	0,7%	1,2%	0,4%	3,2%	3,6%	1,8%	5,7%
Same-country	45%	65%	83%	81%	81%	37%	68%	40%
Same-country '45	13%	33%	15%	75%	100%	33%	39%	0%
0-100km	33%	30%	31%	**50%**	38%	19%	26%	**6%**
0-100km '45	13%	22%	20%	**75%**	38%	48%	11%	**0%**
100-1000km	42%	57%	60%	42%	49%	56%	49%	45%
100-1000km '45	0%	56%	60%	0%	63%	33%	39%	0%
1000-5000km	9%	3%	5%	5%	5%	16%	7%	11%
1000-5000km '45	0%	0%	13%	0%	0%	10%	11%	0%
5000+km	16%	10%	**3%**	**3%**	8%	9%	18%	**38%**
5000+km '45	88%	22%	7%	25%	**0%**	10%	39%	**100%**

Table 4.17: Relative distribution for the male engineers in Europe

	Austria	Germany	France	Italy	Spain	Poland	UK	Ireland
Top International Destinations	Berlin	-	-	-	-	Berlin	-	London NYC Düsseldorf
Top International Destinations '45	Tonlé San New York	SF HK Como	Moscow	NJ	-	Smolensk Pakistan	Israel	China India Pakistan
Spread	y	n	n	n	n	n	y	y
Spread '45	y	-	-	n	n	n	y	-
Same country	29	478	522	344	98	125	416	3
spread	22	304	371	210	55	45	383	0
Change	-24,1%	-35,4%	-28,9%	-38,9%	-43,9%	-64%	-7,9%	-100%

Table 4.18: Spreading and internationalization degree for European male engineers

seen as hotspots for engineering. In a very different manner from the painter case, the spreadability degree is mainly negative, with the exception of Austria, the United Kingdom, and Ireland. The same country spread shows a decrease for all the European countries ranging from -100% for Ireland (a sign that all the people moved outside the country) to almost 8% for the United Kingdom.

When analyzing the situation outside Europe, we can observe in Table 4.19 that the least local country among the ones analyzed is China while the most local one is Brazil. On the other hand, the most global country is South Africa, while the least global is again

Brazil.

	US	Russia	China	Brazil	South Africa	Australia	Japan
Total	1.530	386	70	110	18	57	51
Total '45	59	15	1	4	1	4	5
%	3,8%	3,9%	1,4%	3,6%	5,5%	7%	9,8%
Same-country	45%	77%	28%	81%	50%	74%	71%
Same-country '45	73%	73%	100%	75%	100%	25%	100%
0-100km	23%	22%	**3%**	**37%**	17%	32%	28%
0-100km (1945)	22%	27%	**0%**	**75%**	**0%**	**0%**	**0%**
100-1000km	35%	30%	69%	37%	28%	28%	47%
100-1000km (1945)	27%	13%	100%	0%	100%	0%	100%
1000km-5000km	33%	37%	10%	18%	6%	16%	10%
1000km-5000km(1945)	34%	40%	0%	0%	0%	25%	0%
5000+km	9%	11%	19%	**8%**	**50%**	25%	16%
5000+km (1945)	17%	20%	**0%**	25%	**0%**	**75%**	**0%**

Table 4.19: Relative distribution for the male engineers outside Europe

	US	Russia	China	Brazil	South Africa	Australia	Japan
Top International Destinations	-	Paris Warsaw	Lincoln Portland	-	NYC	London	-
Top International Destinations '45	-	SF	-	Paris	-	-	-
Spread	n	n	n	n	y	y	n
Spread after 1945	y	y	-	-	-	y	-
Same country	690	146	15	47	6	32	29
spread	598	58	6	23	6	29	21
Change	-13,3%	-60,2%	-60%	-51%	0%	-9,4%	-27,6%

Table 4.20: Spreading and internationalization degree for Non-European male engineers

Again, the thesis hypothesis has been confirmed with all the analyzed countries displaying a decrease when it comes to comparing the country-to-country number for places of birth vs. places of death, with percentages ranging from 0% to 60,2%. Compared to the European situation, the countries from outside Europe have a smaller number of international destinations, but with all of them being located in Europe or North America. As for the last generation, only Russia with San Francisco and Brazil with Paris presented external hotspots.

4.3.6 Female Engineer General Overview

Looking at women engineer, we confront with the first dataset that is too small to be analyzed:

- the total number of females that have the attribute "engineer" associated with their profession on Wikidata is 163. This leads to a ratio of 1 to 55, or for every one female engineer, we will have 55 male engineers.
- the largest dataset is contained by the United States with 50 results, while the smallest one is represented by China and Japan with no results;

	Austria	Germany	France	Italy	Spain	Poland	UK	Ireland
Total	1	8	9	1	1	4	22	2
Total '45	0	0	2	0	0	0	0	0
%	0%	0%	22,2%	0%	0%	0%	0%	0%
Same-country	100%	63%	67%	100%	100%	0%	88%	0%
Same-country '45	0%	0%	50%	0%	0%	0%	0%	0%
0-100km	0%	13%	22%	100%	100%	0%	41%	0%
0-100km '45	0%	0%	0%	0%	0%	0%	0%	0%
100-1000km	100%	50%	56%	0%	0%	100%	45%	50%
100-1000km '45	0%	0%	50%	0%	0%	0%	0%	0%
1000-5000km	0%	0%	22%	0%	0%	0%	0%	0%
1000-5000km '45	0%	0%	50%	0%	0%	0%	0%	0%
5000+km	0%	38%	0%	**0%**	0%	0%	14%	**50%**
5000+km '45	0%	0%	0%	**0%**	0%	0%	0%	**0%**
% female	0,6	0,6	0,7	0,1	0,4	0,7	2,1	3,6

Table 4.21: Relative distribution for the female engineers in Europe

In the case of Europe, the most extensive result set is given by the United Kingdom with 22 females, and only one result for Austria, Italy, and Spain. Looking at the dataset for the last generation, only France had two female engineers with a birth and a death date assigned. The size of the datasets creates a lot of very big or very small percentages, with countries only having one individual in total. Also, among the small percentages, many people travel abroad, leading to the same country spreading to be equal to 0%.

On the other hand, the international destinations we find in the case of European female engineers resemble the same style for other professions with smaller datasets that include locations which cannot be directly regarded as an industry or technological hotspot. Despite this, after a short analysis, we can see that obscure locations like Straßkirchen actually have a significant meaning to technically focused people (Straßkirchen contains the second-largest Photovoltaic Power Plant).

	Austria	Germany	France	Italy	Spain	Poland	UK	Ireland
Top International Destinations	-	Standford Toronto	Krakow	-	-	Berlin Straßkirchen	-	London Boston
Top International Destinations '45	-	-	Krakow	-	-	-	-	-
Spread	-	-	n	-	-	-	y	-
Spread '45	-	-	-	-	-	-	-	-
Same country	1	5	6	1	1	1	19	1
spread	1	4	4	1	1	1	19	1
Change	0%	-20%	-33,3%	0%	0%	0%	0%	0%

Table 4.22: Spreading and internationalization degree for European female engineers

	US	Russia	China	Brazil	South Africa	Australia	Japan
Total	50	7	0	1	1	1	0
Total '45	8	0	0	0	0	0	0
%	16%	0%	0%	0%	0%	0%	0%
Same-country	100%	71%	0%	100%	0%	100%	0%
Same-country '45	100%	0%	0%	0%	0%	0%	0%
0-100km	26%	28%	0%	0%	0%	0%	0%
0-100km '45	13%	0%	0%	0%	0%	0%	0%
100-1000km	32%	14%	0%	0%	0%	100%	0%
100-1000km '45	38%	0%	0%	0%	0%	0%	0%
1000-5000km	42%	57%	0%	100%	0%	0%	0%
1000-5000km '45	50%	0%	0%	0%	0%	0%	0%
5000+km	0%	0%	0%	0%	**100%**	0%	0%
5000+km '45	0%	0%	0%	0%	0%	0%	0%
% female	3,2	1,8	0	0,9	5,3	1,7	0

Table 4.23: Relative distribution for the female engineers outside Europe

In a similar way with the analyzed countries from within Europe, the ones outside Europe are going through the same lack of dataset entries for female engineers. Table 4.23 shows that with the exception of the United States and Russia for the entire time interval, each state has at most one result. For the last generation, only the United States has any entries. This leads to an almost empty table for the generation that was born and died after 1945 and a severely skewed trend in the graphic, with almost only peaks and lows.

Looking at the degree of internationalization and spreading in Table 4.24, we can notice that highest probability of women to succeed in Engineering when born outside Europe is to be born in the United States and remain there. While we cannot regularly observe many international destinations within the Top Five places to die, we can notice that Russian female engineers have chosen Almaty and Istanbul, and South African females have chosen Manchester as a destination.

The hypothesis of the thesis is still confirmed, and despite the fact that most of the city count change was zero, Russia has met a decrease as expected.

While this will further likely affect the final analysis and results of the present master thesis, we can still conclude that while the trend of the women engineers is not following the directions of the other professions presented so far, we can observe that indeed until the last decades, being a female in a technical field was in not among the more desirable professions. This thesis is not trying to explain the causes or the effects behind these results, but only to present similarities wherever they are found.

	US	Russia	China	Brazil	South Africa	Australia	Japan
Top International Destinations	-	Almaty Istanbul	-	-	Manchester	-	-
Top International Destinations '45	-	-	-	-	-	-	-
Spread	-	n	-	-	-	-	-
Spread '45	n	-	-	-	-	-	-
Same country	37	5	0	1	0	1	0
spread	37	3	0	1	0	1	0
Change	0%	-40%	0%	0%	0%	0%	0%

Table 4.24: Spreading and internationalization degree for Non-European female engineers

4.3.7 Male Politician General Overview

The last profession analyzed brings the following key facts:

- the male politician dataset is more consisting, with around 112.334 entries having the attribute "politician" associated with their profession. A reason for such a significant difference between the total count for engineers and the total count for politicians is that the profession of being a politician can represent an umbrella to different professional positions and sub-professions while being an engineer is more specific.
- all the queries have returned consistent datasets, with the largest represented by the US with 20.460 results and the lowest represented by South Africa with 229 results;

- the numbers remain the same when querying for the last generation with the largest dataset represented by the US with 323 and the smallest represented by South Africa with ten results;
- when evaluating how the dataset size has changed when comparing the result count for the entire time interval and the last generation, the smallest proportion appears in France with 0,5%, while the largest appears in Russia with 8,4%. These percentages are, however smaller than the general overview, showing that the profession of being a politician is not as attractive as before compared to other professions.

	Austria	Germany	France	Italy	Spain	Poland	UK	Ireland
Total	2.712	16.943	17.940	9.133	4.595	4.359	4.901	735
Total '45	19	125	95	150	134	159	57	12
%	0,7%	0,7%	0,5%	1,6%	2,9%	3,6%	1,2%	1,6%
Same-country	78%	79%	89%	87%	78%	24%	49%	30%
Same-country '45	84%	84%	87%	87%	89%	40%	75%	75%
0-100km	**69%**	53%	53%	58%	44%	25%	24%	**21%**
0-100km '45	**68%**	63%	58%	63%	65%	**38%**	50%	59%
100-1000km	28%	43%	42%	38%	41%	55%	31%	25%
100-1000km '45	26%	34%	50%	27%	29%	40%	32%	25%
1000-5000km	2%	3%	3%	4%	5%	18%	4%	6%
1000-5000km '45	0%	1%	5%	4%	0%	23%	2%	8%
5000+km	2%	2%	2%	**1%**	10%	2%	40%	**48%**
5000+km '45	**1%**	**1%**	3%	**1%**	3%	10%	**18%**	8%

Table 4.25: Relative distribution for the male politicians in Europe

In Table 4.25 we can see that with the exception of Poland, the United Kingdom, and Ireland, male politicians are generally predisposed to remaining close to their place of birth and avoiding very long distance travels. Also, the same-country percentages range from 44% for Spain to 69% for Austria if we do not take into consideration the previously mentioned countries. All these percentages increase for the last generation with the exception of France, where it marginally decreases and Italy, where it remains the same. We can see a significant change in the last generation for Poland, the United Kingdom and Ireland, where the same-country percentages increase drastically, sometimes even more than double (Ireland faces an increase from 30% to 75%).

Another interesting aspect when it comes to male politicians is that with the exception of Poland, the United Kingdom, Ireland, and Spain, there are no international destinations in the Top Five places to die. Spain represents a separate case, since also in the last generation, it joins the other countries that focused on more domestic locations. As the hypothesis of the thesis states, in all the case, there is a decrease in places of birth vs. places of death that ranges from 11,1% to 47,5%.

	Austria	Germany	France	Italy	Spain	Poland	UK	Ireland
Top International Destinations	-	-	-	-	CDMX	Berlin Israel	Toronto Wellington	LND NYC Brisbane Toronto
Top International Destinations '45	-	-	-	-	-	Smolensk	Belfast	Moscow Washington DC
Spread	n	n	n	n	n	n	y	y
Spread '45	-	n	n	n	n	n	y	y
Same country	785	4.431	4.808	2.116	1.115	650	1.071	112
spread	498	2.800	3.807	1.257	585	367	952	89
Change	-36,6%	-36,8%	-20,8%	-40,6%	-47,5%	-43,5%	-11,1%	-20,5%

Table 4.26: Spreading and internationalization degree for European male politicians

	US	Russia	China	Brazil	South Africa	Australia	Japan
Total	20.460	1.774	1.256	2.081	229	1.981	566
Total '45	323	150	31	75	10	31	47
%	1,6%	8,4%	2,5%	3,6%	4,3%	1,6%	8,3%
Same-country	79%	62%	47%	83%	64%	87%	75%
Same-country '45	80%	75%	48%	91%	80%	94%	74%
0-100km	**45%**	17%	**10%**	36%	21%	40%	44%
0-100km '45	48%	31%	**22%**	37%	40%	29%	**62%**
100-1000km	31%	43%	71%	37%	48%	38%	43%
100-1000km '45	35%	26%	61%	43%	50%	42%	34%
1000-5000km	20%	32%	14%	24%	11%	18%	11%
1000-5000km '45	13%	29%	3%	19%	0%	26%	4%
5000+km	3%	8%	4%	3%	**20%**	4%	**2%**
5000+km '45	4%	**13%**	**13%**	1%	10%	3%	**0%**

Table 4.27: Relative distribution for the male politicians outside Europe

Looking at non-European countries in Table 4.27, we can notice that the most local nations are the United States and Japan for both the entire analyzed time period and the last generation, while the least local is China. On the other hand, the most global one is South Africa, while the least one is again Japan.

When analyzing Table 4.28, we can see that the predisposition of staying within the country is also shown in the Top Five international cities. Only Russia and South Africa have external locations during the entire analyzed period, and only China has San

	US	Russia	China	Brazil	South Africa	Australia	Japan
Top International Destinations	-	Paris Helsinki Berlin	-	-	LND	-	-
Top International Destinations '45	-	-	SF	-	-	-	-
Spread	n	n	n	n	n	n	n
Spread '45	y	n	n	n	n	y	n
Same country	6.507	681	370	656	96	711	265
spread	4.528	197	125	263	55	585	146
Change	-30,4%	-71%	-66,2%	-59,9%	-42,7%	-17,7%	-44,9%

Table 4.28: Spreading and internationalization degree for Non-European male politicians

Francisco as an outside location for the last generation. The hypothesis of the thesis is again confirmed, with decreases ranging from 17,7% to 71%.

4.3.8 Female Politician General Overview

Although significantly smaller than the male counterpart, the female politician dataset is still consistent enough to undergo analysis (compared to the engineer dataset):

- The total dataset is of 6.282 entries, leading to a female:male ratio of 1 to 18. In other words for every female politician present on Wikidata, we have 18 men;
- The largest and smallest datasets among the one analysed are represented by Germany with 991 results and Ireland returning 17 results. As a ratio, the last generation displays more significant percentages than in the case of male politicians with values ranging from 0% in China to 21,3% in Brazil, showing that the degree of easiness or the demand of becoming a politician as a female as increased since 1945.

Among the least local countries, we can enumerate Poland, but due to the very small dataset, the place is taken by Ireland for the last generation (Table 4.29). The most local country is represented by Austria while starting with 1945, the Austrian percentage drastically decreases, and the place is taken by Spain. Globally, we have at the top Ireland with the most significant predisposition of traveling on distances larger than 5000km, while the lowest inclination is shown in the case of Italy.

Similarly to other professions, the degree of internationalization is relatively high, with external cities in all the countries except Germany, France, and Italy. The last generation had outside locations everywhere with the exception of Italy, Spain, and Ireland. The thesis hypothesis is once again confirmed with decreases ranging from approximately 4% to 44%.

	Austria	Germany	France	Italy	Spain	Poland	UK	Ireland
Total	129	991	296	236	194	249	230	17
Total '45	7	31	11	19	37	21	12	1
%	5,4%	3,1%	3,7%	8%	19%	8,4%	5,2%	5,9%
Same-country	67%	81%	73%	78%	81%	39%	69%	53%
Same-country '45	71%	81%	73%	100%	95%	52%	83%	100%
0-100km	**55%**	48%	37%	49%	48%	**12%**	35%	30%
0-100km '45	28%	48%	27%	58%	**70%**	29%	25%	**0%**
100-1000km	29%	49%	53%	41%	39%	62%	48%	47%
100-1000km '45	57%	48%	64%	42%	30%	29%	58%	100%
1000-5000km	12%	2%	7%	9%	5%	24%	4%	6%
1000-5000km '45	0%	3%	9%	0%	0%	43%	8%	0%
5000+km	5%	2%	2%	**0%**	9%	2%	13%	**18%**
5000+km '45	**14%**	0%	0%	0%	0%	0%	8%	0%
% female	4,6	5,8	1,6	2,5	4,2	5,4	4,5	2,3

Table 4.29: Relative distribution for the female politicians in Europe

	Austria	Germany	France	Italy	Spain	Poland	UK	Ireland
Top International Destinations	Berlin Munich	-	-	-	CDMX	Berlin Smolensk Israel	Dublin	LND Belfast
Top International Destinations '45	Kenya Pfäffikon	Strasburg	Barcelona	-	-	Smolensk	Koźmice Wielkie	-
Spread	y	n	y	n	n	y	y	n
Spread '45	n	n	-	n	n	n	y	-
Same country	37	414	162	118	97	34	117	9
spread	31	292	141	86	71	19	112	7
Change	-16,2%	-29,5%	-13%	-27,1%	-26,8%	-44,1%	-4,3%	-22,2%

Table 4.30: Spreading and internationalization degree for European female politicians

Outside Europe (Table 4.31), we can observe that the largest dataset is again represented in the case of the United States, while the smallest appears for Japan. China has no female politician born and dead after 1945. The ratio of the last generation within the entire time period is relatively high, ranging from 5,3% in Brazil to 21,3% which tells us that one-fifth of all the female politicians in that location were born and died in the last few decades.

Within the entire time period, the most local country is represented by Australia and

	US	Russia	China	Brazil	South Africa	Australia	Japan
Total	842	133	134	61	41	83	28
Total '45	63	7	0	13	5	9	2
%	7,5%	5,3%	0%	21,3%	12,2%	10,8%	7,1%
Same-country	81%	64%	24%	82%	76%	84%	61%
Same-country '45	86%	71%	0%	92%	100%	0%	0%
0-100km	36%	20%	**11%**	19%	40%	**46%**	**46%**
0-100km '45	43%	43%	**0%**	38%	60%	**89%**	0%
100-1000km	36%	44%	78%	44%	46%	31%	32%
100-1000km '45	33%	14%	0%	31%	40%	0%	0%
1000-5000km	23%	29%	7%	25%	10%	19%	7%
1000-5000km '45	21%	43%	0%	31%	0%	11%	0%
5000+km	**4%**	7%	5%	11%	5%	**4%**	**14%**
5000+km '45	3%	0%	0%	0%	0%	0%	**100%**
% female	3,9	7	9,6	2,8	15,2	4,2	4,9

Table 4.31: Relative distribution for the female politicians outside Europe

Japan, while the least one is represented by China. On the other hand, the most global country is Japan, while the least global for female politicians are represented by the United States and Australia.

	US	Russia	China	Brazil	South Africa	Australia	Japan
Top International Destinations	-	-	NYC	Berlin	-	-	Paris
Top International Destinations '45	-	Riga Kaunas	-	-	-	-	-
Spread	-	n	n	n	n	n	n
Spread '45	n	-	-	n	-	-	-
Same country	509	52	28	32	21	54	14
spread	437	26	11	17	14	43	10
Change	-14,1%	-50%	-60,7%	-46,9%	-33,3%	-20,4%	-28,6%

Table 4.32: Spreading and internationalization degree for Non-European female politicians

Compared to male politicians, female politicians display a smaller number of international locations in their Top Five lists, with only Latvian cities associated with Russia for the last generation. The thesis hypothesis is once more confirmed, with decreases ranging from 14% to almost 61%.

4.3.9 Conclusions

General Overview

Within the general conclusion section, I will illustrate a series of trends and statistics per each country analyzed as it follows:

a) general ratio of women within the total population as presented in the "Relative Distribution" tables for the entire period of time, for both within and outside Europe;

b) general migration distribution according to distance separated for each gender as presented in the "Relative Distribution" tables for the entire period of time, for both within and outside Europe;

c) same-country migration levels as presented in the "Relative Distribution" tables for the entire period of time, for both within and outside Europe;

d) spreading decrease for female and male for the whole period of time and for both within and outside Europe, which also represents the main hypothesis of the thesis as presented in the "Spreading and Internationalization Degree" tables.

After analyzing all the previously stated information, these are the conclusions for the entire male and female professional population:

1. when analyzing the proportion of female vs.male, we can observe that some professions show a higher rate of becoming notable than other among the male population (Figure 4.12). While considered a more modern job, being an engineer as a female displays extremely low percentages of becoming a personality, with ratios ranging from 0% for China and Japan, 0.1% for Italy to the other end of the spectrum - 5,3% for South Africa. On the other hand, among the ones analyzed, the profession of being a painter offers a more significant probability of becoming a personality, with percentages ranging from 4% in Italy (most likely due to the already increased number of male painters) to 34.5% in Australia.

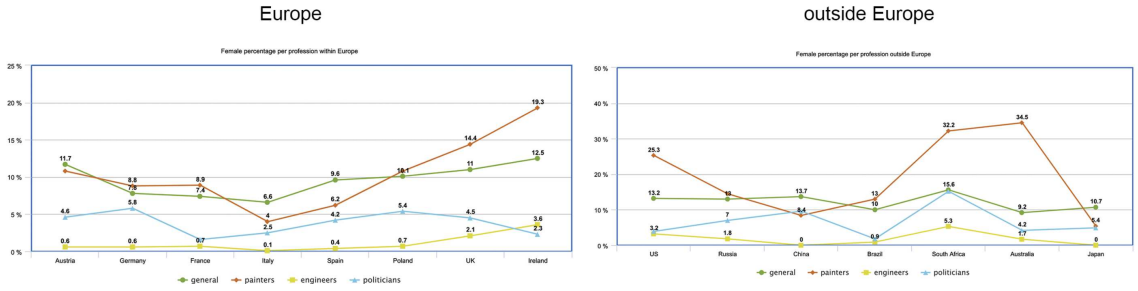

Figure 4.12: Female distribution within the whole dataset according to profession

Although rated as a location with low chances for female professionals in general, Italy, on the other hand, can be seen as an average location to be a politician. The "worst" location for this profession is represented by France, while the best location is again South Africa. Looking at which lines are following the most similar trend, we can observe how the "general" and the "engineer" ones are facing very similar increases and decreases, almost with an identical multiplier (regardless of the fact that indeed, the female engineer dataset is among the smallest and as a result, the least consistent). Very similar trends for all the four lines can also be observed between Italy, Spain and to some extent, with Austria. With the exception of the politician profession, also Germany and France display strong similarities. Outside Europe, the most similar pairs are Russia and Brazil, besides South Africa and Australia;

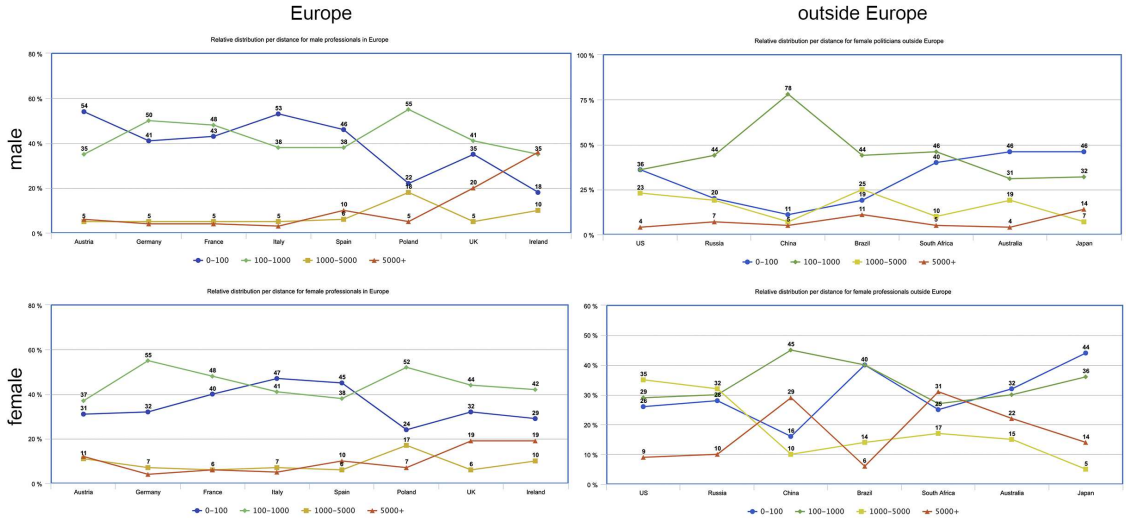

Figure 4.13: Comparison of migration distribution according to distance

2. comparing the distribution according to distance helps us see a clear trend between the male and the female population. Surprisingly, for the male population, Austria and Italy follow very similar patterns with Australia, while Spain is relatively similar to Japan (all the five countries present similarities, with each pair exchanging the portion of long and very long-distance migration in the ranking. In the same manner, Germany and France are very similar for both male and female segments. In general, gender-wise, with the exception of Ireland and China, countries that display an unexpected increase for the very long-distance migration, we can observe similarities when comparing the male dataset with the female dataset (see Figure 4.13);

3. when comparing the same country migration, we can notice that again, all countries and all genders present a clear trend for Germany, France, Italy, and Spain following

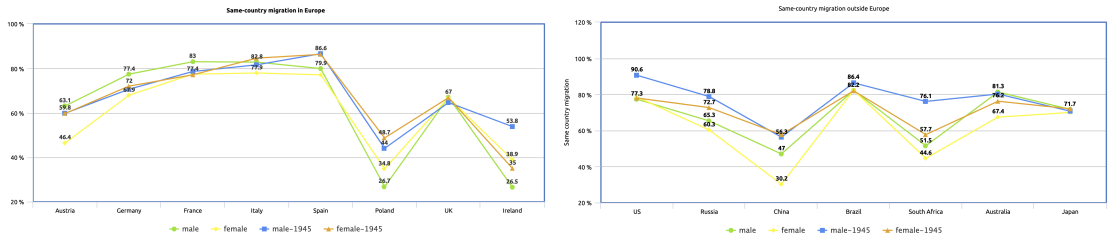

Figure 4.14: Same-country migration for males and females

a very similar path. Poland and Ireland are also displaying strong similarities. Outside Europe, the trend is again less obvious, but all the lines do follow one direction, with China and South Africa showing clear lows; (see Figure 4.14).

4. looking at the decrease of places of birth in comparison with places of death, we can observe from Figure 4.15 that in general, the decrease for male migrants is higher than the female one with the exception of several non-European countries (Brazil, South Africa, and Australia). The most significant decrease difference between the male and the female segments can be seen in the case of Germany, South Africa, and Australia, while the smallest appears the United Kingdom, China, and Japan.

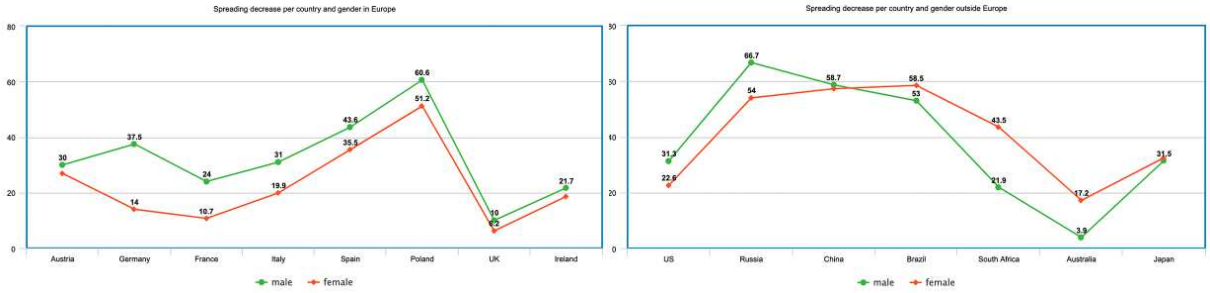

Figure 4.15: Spreading decrease for males and females

Introducing Painters

The introduction of the painter profession brings a series of more or less consistent datasets that either follow the trends set in the general overview or break it completely.

1. the painter distribution (Figure 4.16) tends to follow the general distribution, with the exception of the last generation's 0-100km component and the 5000+km component outside Europe. Again France and Germany tend to follow the same trend for male painters, and while they do not follow the same one for female painters, the values seem to be identical with the general overview. Austria, Italy, and Spain can be again clustered for both female and male painters, while for Poland, the United Kingdom, and Ireland, the values are yet more dispersed

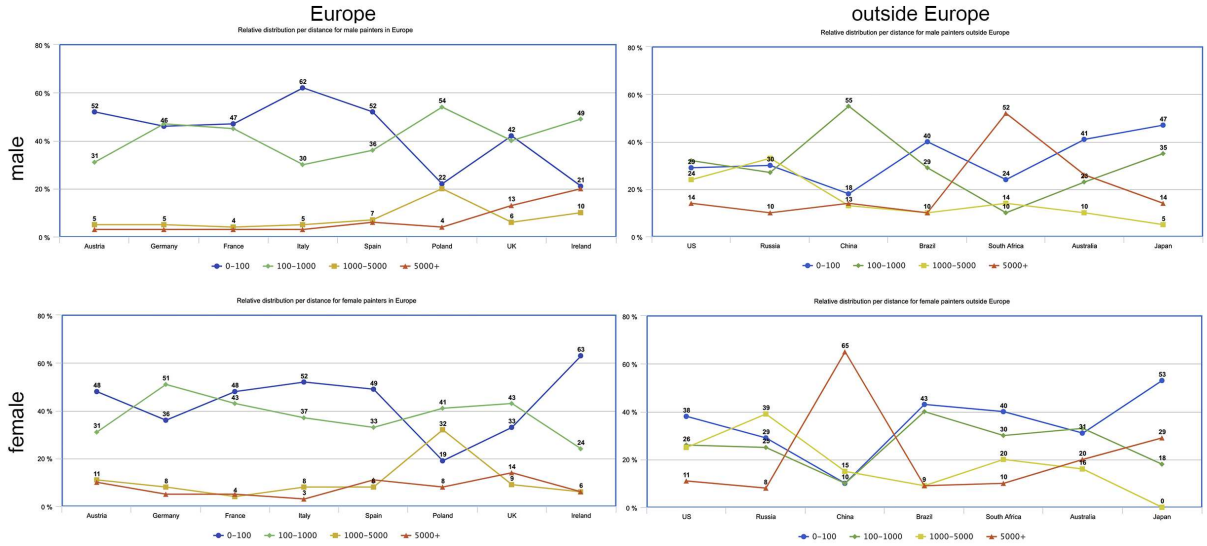

Figure 4.16: Comparison of painter migration distribution according to distance

but still following a trend. Outside Europe, it is harder to see a precise pattern match at certain countries like China, Brazil, and South Africa, where we can notice unexpected peaks and distributions. On the other hand, the peak China experienced for its 100-1000km migration, is again observed for painters only.

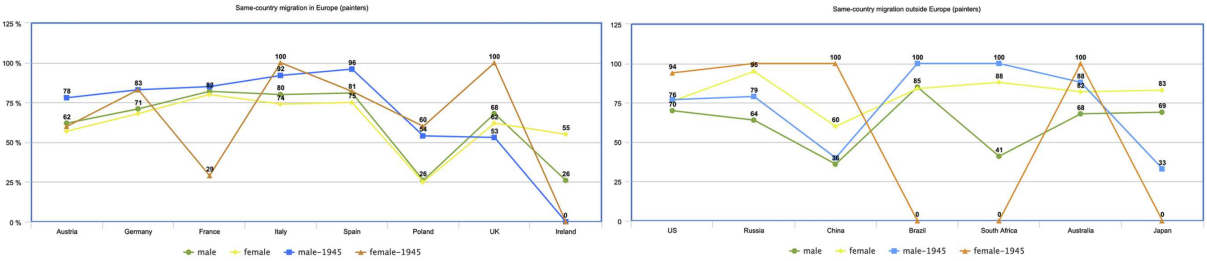

Figure 4.17: Same-country migration for painter males and females

2. analyzing the same-country migration brings interesting insight and precise pattern when it comes to the entire time frame for the male and female painters. On the other hand, the last generation, due also to the smaller dataset, breaks the initial general analysis. (see blue and red lines in Figure 4.17). Again, Poland, Ireland, China, and South Africa display "lows" within the graph, showing that also male and female painters born in these countries choose, in general, not to migrate to the same state.

3. in Figure 4.18 can be noticed the spreading decreases affecting the painters in both Europe and outside Europe. As previously analyzed in the Results and Analysis

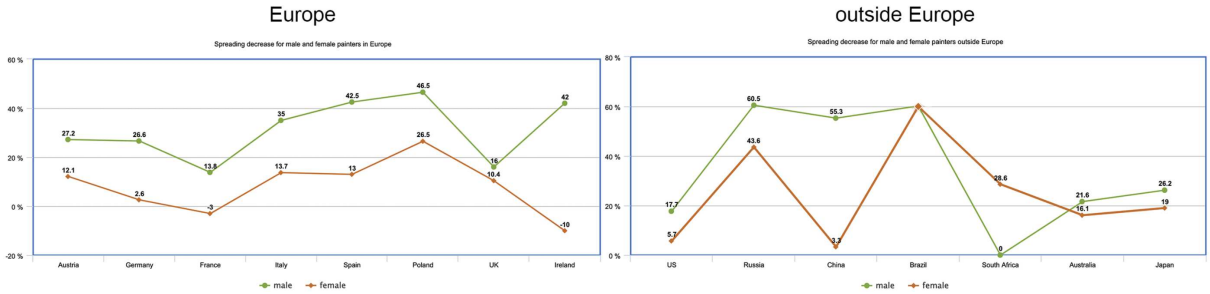

Figure 4.18: Spreading decrease for painter males and females

chapter, the female in France and Ireland are experiencing an increase due to small and inconsistent/duplicate data entries. In general, the decrease for men is higher than the decline for women with the exception of Australia and Japan. Also to be mentioned is that when compared to the general spreading decrease, European women and men follow a very similar trend, while in general, female painters outside Europe do not show very close similarities with the exception of the United States, Russia, and Brazil).

Introducing Engineers

The introduction of painters offers the smallest dataset from the entre analysis. Comparing it to the general and the painter analysis brings the following conclusions:

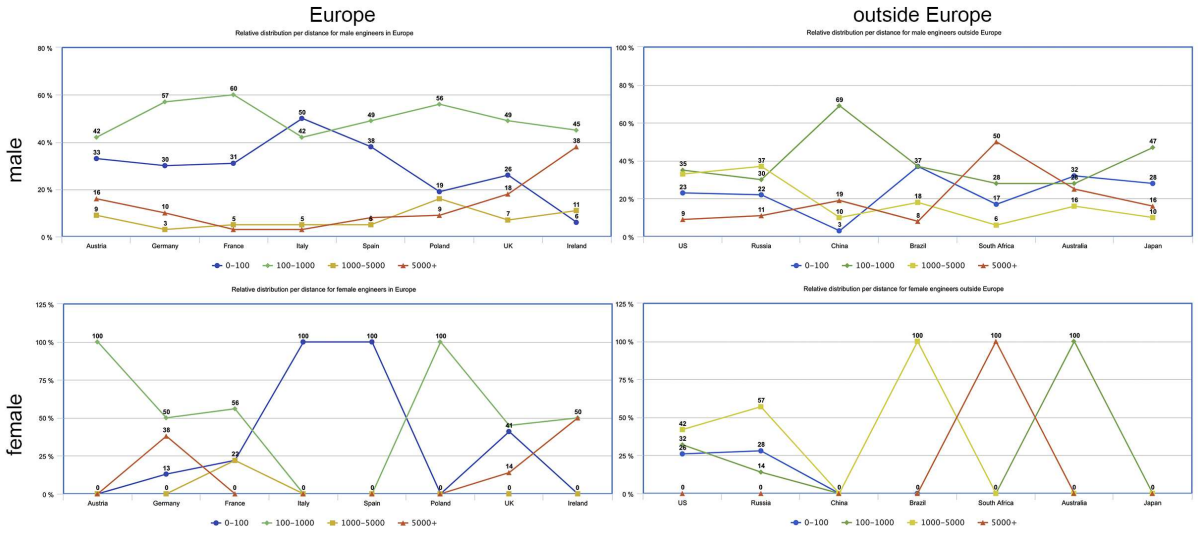

Figure 4.19: Comparison of engineer migration distribution according to distance

1. In Figure 4.19, we can observe how the engineer migration distribution presents more significant changes compared to the painter dataset. The female dataset is

too small to be in any way consistently analyzed and compared to the previous distribution analyses, so the focus, in this case, will gravitate around the male dataset. Again, there are similarities between the behavior of male engineers in France and Germany, but not so apparent similarities between the ones in Austria, Spain, and Italy like previously observed. Also, in the case of Austria, the engineer trends compared to the general trends are not correlated, with the 100-1000km component taking the place of the previously highest 0-100km. Italy closely follows the direction from the entire dataset while the others present consistent changes. Outside Europe, one of the most noticeable difference is the fact that China persists the peak for its 100-1000km component while having the other components closely aggregated. Japan and Russia also present jointly correlated values, with medium distance migration, although switched still very close as values.

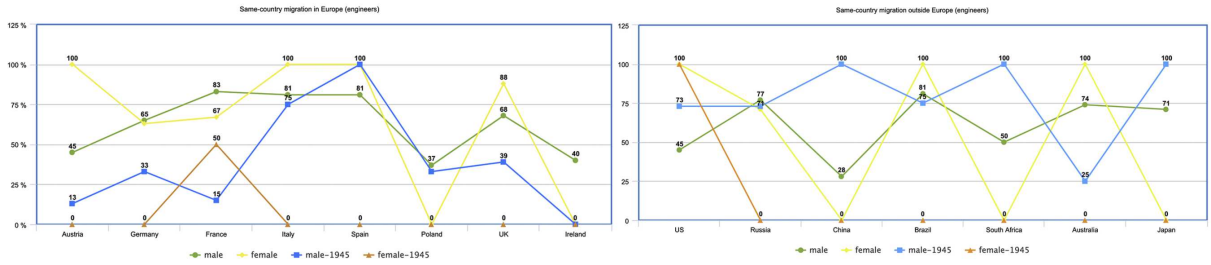

Figure 4.20: Same-country migration for engineer males and females

2. In the case of same-country migration, we can again notice a solid trend correlation when it comes to male engineers for the entire analyzed time period. Due to small or empty datasets, the male and female engineers of the last generation can not bring much insight to the conclusion. There are, however, substantial similarities for the engineer women datasets compared to the entire dataset for Germany and France (Figure 4.20).

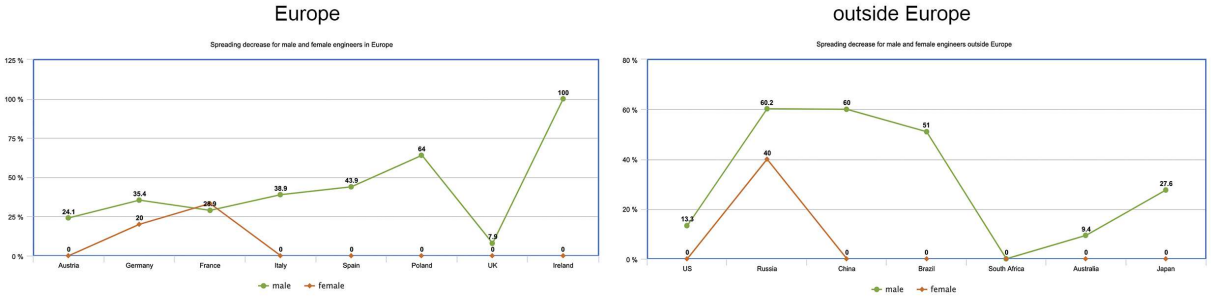

Figure 4.21: Same-country migration for engineer males and females

3. Figure 4.21 confirms what has been previously mentioned in the Engineer conclusion, with the male population being the only group following the global dataset. For

women, on the other hand, in all the countries except France and Russia the decrease was 0.

Introducing Politicians

The last analyzed profession also displays one of the most substantial datasets. The conclusions are the following:

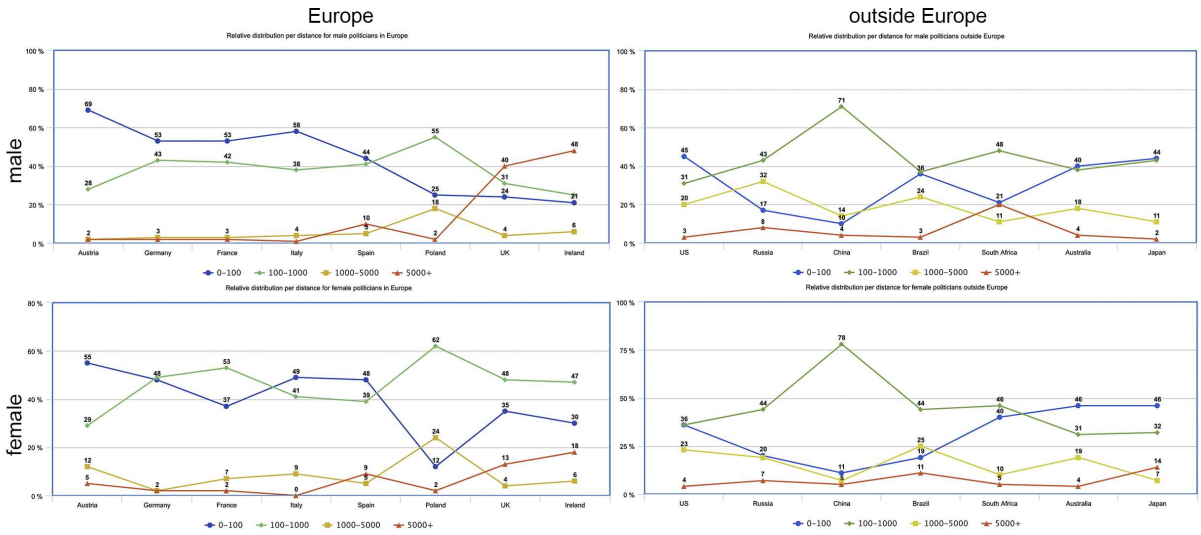

Figure 4.22: Comparison of politician migration distribution according to distance

1. Similarly to painters, there is a clear correlation to the general dataset for male politicians (Figure 4.22). Again, Germany and France move closely together, while Austria, Italy, and to some extent, Spain. Ireland and the United Kingdom are still displaying similar behavior with Poland reverting all the lines. Outside Europe, we have a more similar action when analyzing the United States and Japan and Australia, with China peaking again for the 100-1000km component. China has a similar behavior for female politicians, and again, the United States shows apparent similarities with Australia, while Russia behaves to some degree similar to Australia. For European women, again Austria and Italy and separately the United Kingdom and Ireland have very similar distance ordering, but in this situation, Germany and France do not share so many correlations. Poland again displays a more random behavior compared to the other European countries.

2. Looking at the same-country migration degree, with the exception of the female politicians belonging to the last generation and from China and Australia, all the datasets show close trends to the each of their general counterparts.

3. the spreading decrease for politicians (Figure 4.24) show a more similar trend to the general population for the women segment than for men. While very similar

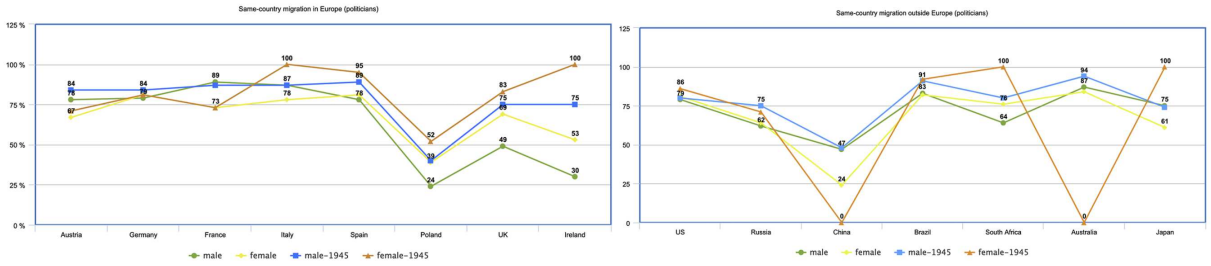

Figure 4.23: Same-country migration for politician males and females

as well, the non-European section differs from the general non-European from the ranking of female compared to the male population. Outside Europe, in all the countries there was a more significant decrease for the female population than the male one.

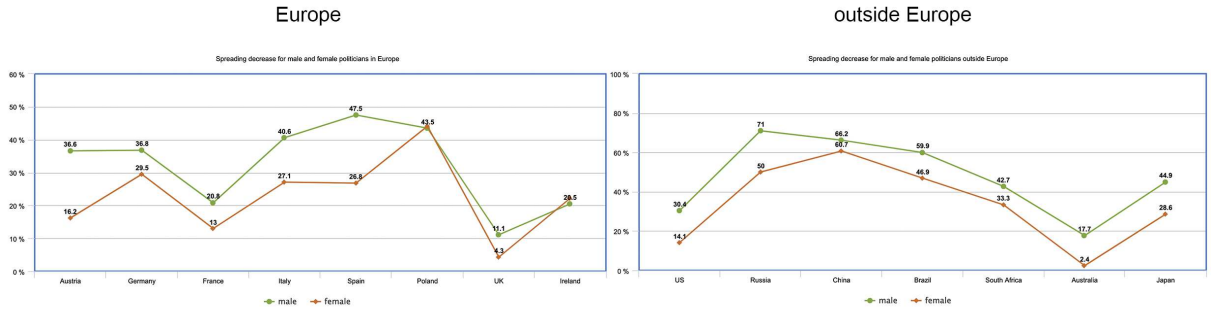

Figure 4.24: Spreading decrease for politician males and females

CHAPTER 5

Summary and Conclusions

5.1 Summary

In this study, the analysis of three different professions has been undergone together with an overview of the general situation which contains the entire population that has a profession assigned. Both men and women have been taken into account from a selection of countries from within and outside Europe. Also to be mentioned is that this thesis did not aspire to analyze the cultural, social, political, geographical or economical aspects that might trigger a nation to collectively behave the way it does, but only to find patterns where they exist.

While the primary analysis has revolved around the entire time frame, also the last generation (notable people born and dead after 1945) has been briefly mentioned to see if the proportions and results have significantly changed in the recent years, when communication and transportation have been exponentially improved and become more accessible to the people.

Besides answering the research questions directly, there have been several interesting insights that have been concluded at the end of the study:

- the thesis can be taken as a way of telling who, where, and when can be successful in a particular profession. Since the data only consists of notable people who made their way into the Wikidata repository, we can take their success as a sign that being born in certain circumstances can improve chances when pursuing a specific career. Within the study, it is clear that these chances are different when comparing females with males and different countries. Also, it is clear that for some circumstances, the opportunities have significantly increased in the last generation. Some examples in this regard are:

  - the profession with the greatest result count is the one of being a politician, while the smallest result count appears for engineers. As we will see next, the amounts are in general not proportional with different countries containing the largest or smallest datasets per profession.
  - **painters**:
    * Within Europe, Germany represents the country that gave the biggest number of painters for both male and female, while Ireland was the country with the smallest number. Looking at the last generation, the greatest chance of succeeding is in the case of Spain with 2,5% for male and 8,6% of the female population born after and dead after 1945. The lowest chance appears in Ireland with 0 painters in the dataset.
    * Outside Europe, the United States gives the biggest number of male and female painters, while South Africa gives the smallest. Last generation-wise, the best ratio appears in Australia for male painters (6%), and China for female (10%);
  - **engineers**:
    * Within Europe, similarly to the painter population, the largest male dataset is offered by Germany, while the smallest is given by Ireland. On the other hand, for female, the largest dataset is in the case of the United Kingdom while the smallest, formed by one individual appears in Austria, Italy, and Spain. Around one-fifth of the female engineers were born in France in the last generation, making it the highest percentage in comparison to the highest male percentage being in Ireland with 5,7%.
    * Outside Europe, the most significant dataset appears in the United States for both males and females, while the smallest are from South Africa for males and China and Japan for females.
  - **politicians**:
    * Within Europe, the largest dataset shows that the most consistent dataset for male European politicians is in the case of France, while the smallest is again from Ireland. For female politicians, the largest is in the case of Germany, while the smallest is again in Ireland.
    * Outside Europe, the largest is at a very far distance in the case of the United States, while the smallest is Australia. In the case of female politicians, the largest is again for the United States, but the smallest is in Japan.

- in the same way, knowing which cities in the world represent a hotspot for a certain job can eventually influence people to migrate there in order to pursue their plans. Some examples in this regards are:
  - Austria: Munich, Berlin, New York City, Paris, London;
  - Germany: Vienna;

  - France: New York City, Amsterdam;
  - Italy: Paris;
  - Spain: Mexico City, Buenos Aires, Paris;
  - Poland: Berlin, Paris, Munich, New York City;
  - United Kingdom: New York City, Los Angeles, Paris, Toronto, Dublin;
  - Ireland: London, Paris, New York City, Geneva, Rome, Toronto, Edinburgh;
  - United States: Paris;
  - Russia: Paris, Kiev, Helsinki, Berlin;
  - China: New York City, Los Angeles, Washington DC, Paris, Montreal;
  - Brazil: Paris;
  - South Africa: London, Sydney;
  - Australia: London, Los Angeles, Paris;
  - Japan: Paris, New York City.

  This leads to the conclusion that the most popular international hotspots are Paris, New York City, London, and Berlin.

- another conclusion that can be made at the end of the study is concerning the way in which the male vs. female ratio was affected from profession to profession and from country to country. Indeed the female population is underrepresented within the Wikidata repository and if or how this can change in the future, is a question not threated in this context. As far as this thesis analyzed, the last generation has not provided much improvement, with incomplete and very small datasets returned in almost all the professions, which eventually lead to patterns being unfollowed. Other than that, regarding the female distribution within the entire dataset, we have concluded from this study that:

  - Generally, the best percentages to become notable as a woman within the entire population are in South Africa, where 15,7% of all the individuals were female. The worst chances are in Italy, with only 6,6%.
  - **painters**: highest percentage in Australia with 34,5% (but closely followed by South Africa again with 32,2%) while the lowest is in Italy with 4%;
  - **engineers**: Highest percentage in South Africa 5,3% (followed by Ireland with 3,6%), lowest in China and Japan with 0% (followed by Italy with 0,1%);
  - **politicians**: Highest percentage is in South Africa with 15,2%, while the lowest is in France with 1,6%.

- country size was not a factor that has been taken into consideration when analyzing the migration patterns since the countries were only compared to each other once the relative distributions have been computed. It would be, however, interesting for future research to understand if this attribute can play a significant role when analyzing the subject.

- each country displayed some clear patterns standalone or in comparison with other countries. Looking at each separately, we have:
  - Austria: initially though as having similarities to Germany, Austria proved to be more similar to Italy and to some extent to Spain. However it did have some hotspot exchanges with Germany;
  - Germany and France: proved to show not only the closest migration patterns but also female distribution trends;
  - Italy: is in general, the least inter-continental country which contradicts the famous great Italian migration towards the US. To keep in mind is that the present thesis only analyses notable people. If notable people behave like the general population or not, is not a subject of the thesis;
  - Spain: besides the previously mentioned similarities with European countries, Spain has also displayed interestingly similar patterns compared to Japan. Also, together with Italy, the country displays the largest same-country migration in Europe;
  - Poland: a unique migration pattern, characterized with increases for regional (100-1000km) migration for engineers, politicians and male painters (the same behavior was observed in China). Also, in general, the smallest same-country migration among all the ones analyzed and due to its high emigration percentages, Poland also displays the largest spreading decrease (since few people end up within the country eventually);
  - The United Kingdom: also unique migration patterns, with some similarities when compared to Ireland;
  - Ireland: smallest dataset and the country with the highest rate of inter-continental migration;
  - The United States: having the largest intra-continental migration, it has a resemblance to the migration behavior of Russia;
  - Russia: a unique observation towards Russia is related to its very high spreading decrease, a fact most likely explained by its extensive surface;
  - Brazil; among the least predisposition to leave the country and as a result, the continent. Another notable fact here is that one-fifth of female politicians in Brazil were born after 1945;
  - South Africa: unique migration pattern that minimizes the intra-continental movement, which shows that indeed, the south African avoid the neighboring countries;
  - Australia: another unique patterned country that sometimes shares similarities with Japan.
- the Last generation has shown that against intuition, people will look to settling locally even in modern times. This might mean that the easiness of becoming

notable without moving on great distances has increased (sometimes even within the same country). The last generation has also shown that some otherwise more extreme countries (e.g., Ireland, and Poland) have eventually become more settled starting with 1945;

5.2 Answering the Research Questions

Returning to the research questions enounced at the beginning of the study:

- Can we use Linked Open Data to analyze professional human migration patterns? What are the advantages, drawbacks, and challenges when using dynamic digital crowd-sourced data when analyzing social science topics such as migration?

  The answer is yes, we can, but this does not come with ease as there are several drawbacks:

  - while the data set is extensive, it is still created and maintained by people, which generally makes it prone to being incomplete or incorrect. Duplicates or almost identical entries referring to the same entity are also a widespread issue within Linked Open Data;
  - biased predisposition of completeness depending on the country is still very common. Some countries seem to be more careful in maintaining their repositories, while others are still a bit behind with it, making datasets too small to be analyzed;
  - update rate - although this can also be an advantage, I will add this within the disadvantaged category as, during my analysis, I have observed how the numbers and data returned for each query can differ from a day to day basis. The research and the values included are valid as of May-July 2019;
  - although this can also be considered as an advantage, due to the large data sets (usually when more than 20.000 entries fetched) some queries are timed out by the Wikidata's SPARQL service. This has lead to trying different approaches, but in the end, the fastest way was to divide the queries into smaller time intervals and aggregate the results at the end;
  - lack of support when connecting the returned dataset with the visualization methods. While there are some libraries that are trying to mediate the interaction between these components (e.g., d3sparql.js), they were very inflexible in what regarded customization. At the end all the visualizations were implemented using only bare d3.js.

  Among the advantages detected while working on the study, I can enumerate the following:

  - dynamic data - even though mentioned as a drawback, the dynamic data opens a range of other opportunities when working on this subject, like analyzing

the rate in which new data is introduced and updated for certain countries, professions or general topics/fields. This can represent a good way of extending the research for the future;

- large data repository and access to a variety of data categories and attributes - Wikidata offers a great range of properties and objects that can be easily queried and afterward easily used;
- ease of implementation - although the learning curve is relatively steep, it still offers a reasonably fast learn path with good documentation and community support;
- being open-source - this has also positively influenced the previous aspect which has lead to a great community that learns and implements Open Linked Data (and to some degree in combination with data visualisation).

- What is the degree of disparity when it comes to places of birth versus places of death? Can we offer a general comparative analysis on how migration is unfolding according to country, profession, and gender?

  As it has been proven in the previous chapters, there is a degree of disparity observed according to the country, profession, and gender of the individuals in the sense that the number of places of birth has sustainably decreased compared to the number of places of death, as it follows:

  - Austria: 30% for men and 26,9% for women;
  - Germany: 37,5% for men and 14 % for women;
  - France: 24% for men and 10,7% for women;
  - Italy: 31% for men and 19,9% for women;
  - Spain: 43,6% for men and 35,5% for women;
  - Poland: 60,6% for men and 51,2% for women;
  - United Kingdom (UK): 10% for men and 6,2% for women;
  - Ireland: 21,7% for men and 18,2% for women;
  - United States of America (USA): 31,3% for men and 22,6% for women;
  - Russia: 66,7% for men and 54% for women;
  - China:58,7% for men and 57,3% for women;
  - Brazil: 53% for men and 58,5% for women;
  - South Africa: 21,9% for men and 43,5% for women
  - Australia: 3,9% for men and 17,2% for women;
  - Japan: 31,5% for men and 32,5% for women;
  - Worldwide: 38,3% for men and 29% for women.

These numbers have changed according to the profession, but as it has been previously stated, due to smaller and inconsistent datasets, in many cases, duplicates have altered the final results. Each case of abnormal behavior (Irish and French female painter) has been analyzed and at the end explained in the Results and Analysis chapter.

List of Figures

List of Tables

Acronyms

AJAX Asynchronous JavaScript and XML. 40

API Application Programming Interface. 19

CSS Cascading Style Sheets. 27, 28

GIS Geographic Information System. 26

GLAM Galleries, Libraries, Archives, and Museums. 16

HTML Hypertext Markup Language. 27, 28

HTTP HyperText Transfer Protocol. 28

KB Knowledge Base. 21

KGs Knowledge Graphs. 37

LBS Location-Based Service. 19

LOD Linked Open Data. 16

RDF Resource Description Framework. 21, 35

SVG Scalable Vector Graphics. 27, 29

UN United Nations. 8

URI Universal Resource Identifier. 28

W3C The World Wide Web Consortium. 28

Bibliography

[AA00] RM Ahsan and AS Ahmad. International female migration for work from Bangladesh, 1985–1997. *Oriental Geographer*, 44(1):17–33, 2000.

[ABE12] Ran Abramitzky, Leah Platt Boustan, and Katherine Eriksson. Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration. *American Economic Review*, 102(5):1832–56, 2012.

[ABK+07] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A nucleus for a web of open data. In *The Semantic Web: 6th International Semantic Web Conference*, pages 722–735. Springer, 2007.

[AGMRTL17] D Abián, F Guerra, J Martínez-Romanos, and Raquel Trillo-Lado. Wikidata and DBpedia: a comparative study. In *Semanitic Keyword-based Search on Structured Data Sources*, pages 142–154. Springer, 2017.

[Bac18] Massimo Livi Bacci. *A short history of migration*. John Wiley & Sons, 2018.

[Bad08] Klaus Bade. *Migration in European history*, volume 4. John Wiley & Sons, 2008.

[Bar13] Michele Barbera. Linked (open) data at web scale: research, social and engineering challenges in the digital humanities. *JLIS. it - Italian Journal of Library, Archives, and Information Science*, 4(1):91, 2013.

[BEW16] E Bertini, N Elmqvist, and T Wischgoll. Judgment error in pie chart variations. In *Proceedings of the Eurographics/IEEE VGTC conference on visualization: Short papers*, pages 91–95, 2016.

[BLHL01] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific american*, 284(5):28–37, 2001.

[BN14] Liora Bigon and Ambe J Njoh. The cartography of the unseen. *Material Culture Review/Revue de la culture matérielle*, 80, 2014.

[BNA+10] Christoph Böhm, Felix Naumann, Ziawasch Abedjan, Dandy Fenz, Toni Grütze, Daniel Hefenbrock, Matthias Pohl, and David Sonnabend. Profiling linked open data with proLOD. In *2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*, pages 175–178. IEEE, 2010.

[BOH11] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3: Data-driven documents. *IEEE Trans. Visualization & Comp. Graphics Journal (Proc. InfoVis)*, 2011.

[Buc19] Andrew Buchanan. *World War II in Global Perspective, 1931-1953: A Short History*. Wiley-Blackwell, 2019.

[Cho08] Mark I Choate. *Emigrant Nation: The Making of Italy Abroad*. Harvard University Press, 2008.

[CLLVM14] Marie-Louise Caravatti, S Mc Lederer, Allison Lupico, and Nancy Van Meter. Getting teacher migration & mobility right. *Education International Journal*, 1, 2014.

[Coh16] Robin Cohen. *Migration and its enemies: Global capital, migrant labour and the nation-state*. Routledge, 2016.

[CPP+18] Ana Caraban, Teresa Paulino, Ricardo Pereira, Pedro Spence, Campos, et al. The monarch room: an interactive system for visualization of global migration data. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference. BCS Learning & Development Ltd*, pages 1–5. BCS Learning & Development Ltd., 2018.

[Cwe01] Saulo B Cwerner. The times of migration. *Journal of Ethnic and Migration Studies*, 27(1):7–36, 2001.

[DL05] Sunita Dodani and Ronald E LaPorte. Brain drain from developing countries: how can brain drain be converted into wisdom gain? *Journal of the Royal Society of Medicine*, 98(11):487–491, 2005.

[dSFGdS14] Ricardo da Silva Freguglia, Eduardo Gonçalves, and Estefania Ribeiro da Silva. Composition and determinants of the skilled out-migration in the brazilian formal labor market: A panel data analysis from 1995 to 2006. *EconomiA*, 15(1):100–117, 2014.

[EGW06] Mark Ellis and Jamie Goodwin-White. 1.5 Generation Internal Migration in the US: Dispersion from States of Immigration? *International Migration Review*, 40(4):899–926, 2006.

[ES06] ECOWAS-SWAC/OECD. *Atlas on regional integration in West Africa*. OECD Paris Publisher, 2006.

[FEMR15] Michael Färber, Basil Ell, Carsten Menne, and Achim Rettinger. A comparative survey of DBpedia, Freebase, Opencyc, Wikidata, and Yago. *Semantic Web Journal*, 1(1):1–5, 2015.

[Few04a] Stephen Few. Eenie, meenie, minie, moe: selecting the right graph for your message. *Intelligent Enterprise*, 7:14–35, 2004.

[Few04b] Stephen Few. Tapping the power of visual perception. *Visual Business Intelligence Newsletter*, 39:41–42, 2004.

[Fis13] Michael H Fisher. *Migration: A world history*. Oxford University Press, 2013.

[Fri02] Michael Friendly. Visions and re-visions of Charles Joseph Minard. *Journal of Educational and Behavioral Statistics*, 27(1):31–51, 2002.

[GCML06] Diansheng Guo, Jin Chen, Alan M MacEachren, and Ke Liao. A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE transactions on visualization and computer graphics*, 12(6):1461–1474, 2006.

[GH03] Michael J Greenwood and Gary L Hunt. The early history of migration research. *International Regional Science Review*, 26(1):3–37, 2003.

[Guo09] Diansheng Guo. Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1041–1048, 2009.

[HAARV17] BP Harshitha, R Amith, S Abhishek, and C Rohit Vibhu. *Agricultural Data Visualization for Prescriptive Crop Planning*, volume 49. International Journal of Computer Trends and Technology (IJCTT), 2017.

[Häg75] Torsten Hägerstrand. *Space, time and human conditions*. Saxon House & Lexington Books, 1975.

[Haw99] L Hawthorne. Female, mobile and skilled: The migration process and professional integration of ESB and NESB nurses in Australia, 1986-1996. In *Fourth Annual International Metropolis Conference, Washington DC December*, pages 7–11, 1999.

[HLD+07] Ming C Hao, Julian Ladisch, Umeshwar Dayal, Meichun Hsu, and Daniel Keim. Method for visualizing large volumes of multiple-attribute data without aggregation using a pixel bar chart, May 22 2007. US Patent 7,221,474.

[Ire01] Robyn Iredale. The migration of professionals: theories and typologies. *International migration*, 39(5):7–26, 2001.

[Isa07] Ann Katherine Isaacs. *Immigration and emigration in historical perspective*, volume 1. Edizioni Plus, 2007.

[KAF+08] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual analytics: Definition, process, and challenges. In *Information visualization*, pages 154–175. Springer, 2008.

[KG11] W Koff and P Gustafson. CSC leading edge forum data revolution. *CSC's Leading Edge Forum*, page 68, 2011.

[Kin12] Russell King. Theories and typologies of migration: an overview and a primer. Willy Brandt series of working papers in international migration and ethnic relations 3/12. *Malmö Institute for Studies of Migration, Diversity and Welfare (MIM).–Malmö*, 2012.

[Kje05] K Kjeldstadli. Immigration and industrialisation, Norway c. 1840-1949. *Essays on industrialisation in France, Norway and Spain, eds. K Bruland & JM Olivier, Unipub-Oslo Academic Press, Oslo*, pages 169–182, 2005.

[KSFN07] Andreas Kerren, John T Stasko, Jean-Daniel Fekete, and Chris North. Workshop report: information visualization–human-centered issues in visual representation, interaction, and evaluation. *Information Visualization*, 6(3):189–196, 2007.

[KTFW04] Russell King, Mark Thomson, Tony Fielding, and Tony Warnes. Gender, age and generations, state of the art report Cluster C8. *Sussex Centre for Migration and Population Studies, University of Sussex*, 2004.

[LC04] Zai Liang and Yiu Por Chen. Migration and gender in China: An origin-destination linked approach. *Economic Development and Cultural Change*, 52(2):423–443, 2004.

[LE07] Ralph Lengler and Martin J Eppler. Towards a periodic table of visualization methods for management. In *IASTED Proceedings of the Conference on Graphics and Visualization in Engineering (GVE 2007), Clearwater, Florida, USA*, 2007.

[LKS+14] Denis Lukovnikov, Dimitris Kontokostas, Claus Stadler, Sebastian Hellmann, and Jens Lehmann. DBpedia viewer-an integrative interface for DBpedia leveraging the DBpedia service eco system. In *LDOW Linked Data on the Web*. CiteSeer Scientific Literature Digital Library, 2014.

[LL09] Jan Lucassen and Leo Lucassen. The mobility transition revisited, 1500–1900: what the case of europe can offer to global history. *Journal of Global History*, 4(3):347–377, 2009.

[Man12] Patrick Manning. *Migration in world history*. Routledge, 2012.

[Man15] Joseph Joseph Mangalam. *Human Migration: A Guide to Migration Literature in English 1955–1962*. University Press of Kentucky, 2015.

[MGP+04] Alan M MacEachren, Mark Gahegan, William Pike, Isaac Brewer, Guoray Cai, Eugene Lengerich, and F Hardistry. Geovisualization for knowledge construction and decision support. *IEEE computer graphics and applications*, 24(1):13–17, 2004.

[Moc07] Leslie Page Moch. Connecting migration and world history: demographic patterns, family systems and gender. *International Review of Social History*, 52(1):97–104, 2007.

[MTG14] András Micsik, Sándor Turbucz, and Attila Györök. *LODmilla: a Linked Data Browser for All*. CEUR-WS. org Central Europe Workshop Proceedings, 2014.

[NSS16] Lekha Nair, Sujala Shetty, and Siddhanth Shetty. Interactive visual analytics on Big Data: Tableau vs D3.js. *Journal of e-Learning and Knowledge Society*, 12(4), 2016.

[Pre77] Allan Pred. The choreography of existence: comments on Hägerstrand's time-geography and its usefulness. *Economic geography*, 53(2):207–221, 1977.

[RMD06] Theresa-Marie Rhyne, Alan M MacEachren, and Jason Dykes. Guest editors' introduction: Exploring geovisualization. *IEEE Computer Graphics and Applications*, 26(4):20–21, 2006.

[RPMK19] Beschi Raja, J Pamina, P Madhavan, and A Sampath Kumar. Market behavior analysis using descriptive approach. *SSRN - Social Science Research Network*, 2019.

[SAPJ18] Jana Sverdljuk, Lars Jynge Alvik, Anna Peterson, and Lars G Johnsen. Historical networks and identity formation: Digital representation of statistical and Geo-Data: Case study of norwegian migration to the USA (1870-1920). pages 10–24. CEUR - Central Europe Workshop Proceedings, 2018.

[SG14] Marc Streit and Nils Gehlenborg. *Points of view: bar charts and box plots*. Nature Publishing Group, 2014.

[SJ17] Daniel M Stephen and Bernhard Jenny. Automated layout of origin–destination flow maps: US county-to-county migration 2009–2013. *Journal of Maps*, 13(1):46–55, 2017.

[Sja62] Larry A Sjaastad. The costs and returns of human migration. *Journal of political Economy*, 70(5, Part 2):80–93, 1962.

[SKY+13] Pedro Szekely, Craig A Knoblock, Fengyu Yang, Xuming Zhu, Eleanor E Fink, Rachel Allen, and Georgina Goodlander. Connecting the Smithsonian American Art Museum to the Linked Data Cloud. In *Extended Semantic Web Conference*, pages 593–607. Springer, 2013.

[Sny89] John P Snyder. Album of map projections, United States Geological Survey Professional Paper. *U.S. Geological Survey Professional Paper*, 1395, 1989.

[Spe05] Ian Spence. No humble pie: The origins and usage of a statistical chart. *Journal of Educational and Behavioral Statistics*, 30(4):353–368, 2005.

[Sta00] Peter Stalker. *Workers without frontiers: the impact of globalization on international migration*. International Labour Organization, 2000.

[Tsu09] Takeyuki Tsuda. *Diasporic homecomings: Ethnic return migration in comparative perspective*. Stanford University Press, 2009.

[Tuf01] Edward R Tufte. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, 2001.

[VAB+15] Fabio Valsecchi, Matteo Abrate, Clara Bacciu, Maurizio Tesconi, and Andrea Marchetti. DBpedia Atlas: Mapping the Uncharted Lands of Linked Data. LDOW Linked Data on the Web, 2015.

[Wai97] Howard Wainer. Visual Revelations: Tom's Veggies and the American Way. *Chance*, 10(3):40–42, 1997.

[WLM+14] Xianwen Wang, Chen Liu, Wenli Mao, Zhigang Hu, and Li Gu. Tracing the largest seasonal migration on earth. *arXiv preprint arXiv:1411.0983*, 2014.

[ZV13] Franc J Zakrajšek and Vlasta Vodeb. eCultureMap–link to Europeana knowledge. In *International Conference on Theory and Practice of Digital Libraries*, pages 184–189. Springer, 2013.