Instituto Universitário de Lisboa

Department of Information Science and Technology

Visualization and Analytics of Codicological data of Hebrew books

TIAGO MIGUEL GARCEZ PATEIRO

A Dissertation presented in partial fulfillment of the Requirements for the Degree of

Master in Computer Engineering

Supervisor

Elsa Alexandra Cabral da Rocha Cardoso, Assistant Professor, PhD

ISCTE-IUL

Co-Supervisor

Débora Marques de Matos, PhD

Westfälische Wilhelms-Universität Münster

December 2018

Resumo

A presente dissertação tem como objetivo obter conhecimento estruturado de manuscritos hebraicos coletados por codicologistas. Estes manuscritos refletem a produção de livros de uma região específica, nomeadamente a região "Sefarad", no período entre os séculos X e XVI. O objetivo é fornecer um modelo de dados apropriado, usando um vocabulário comum, para diminuir a natureza heterogénea desses conjuntos de dados, bem como sua incerteza inerente causada pela natureza descritiva no campo da Codicologia. Este projeto de investigação foi desenvolvido com o objetivo de aplicar técnicas de visualização de dados e *data mining* no campo da Codicologia e Humanidades Digitais. Usando os dados de manuscritos hebraicos como ponto de partida, esta dissertação propõe um ambiente para análise exploratória a ser utilizado por especialistas em Humanidades Digitais e Codicologia para aprofundar a compreensão dos dados codicológicos, formular novas hipóteses de pesquisa, ou verificar existentes, e comunicar as suas descobertas de uma forma mais rica. Para melhorar as visualizações e descoberta de conhecimento, tentaremos usar métodos de *data mining*, como a *Association Rule Mining* e *Formal Concept Analysis*.

Palavras-Chave: *Visual Analytics*, Ciência de Dados, *Big Data Analytics*, *Data Mining*, *Machine Learning*, *Knowledge Discovery*, Humanidades Digitais, História de Livros Judaicos, Codicologia, *Material Culture*

Abstract

century. The goal is to provide a proper data model, using a common vocabulary, to decrease the heterogenous nature of these datasets as well as its inherent uncertainty caused by the descriptive nature of the field of Codicology. This research project was developed with the goal of applying data visualization and data mining techniques to the field of Codicology and Digital Humanities. Using Hebrew manuscript data as a starting point, this dissertation proposes an environment for exploratory analysis to be used by Humanities experts to deepen their understanding of codicological data, to formulate new, or verify existing, research hypotheses, and to communicate their findings in a richer way. To improve the scope of visualizations and knowledge discovery we will try to use data mining methods such as Association Rule Mining and Formal Concept Analysis. The present dissertation aims to retrieve information and structure from Hebrew manuscripts collected by codicologists. These manuscripts reflect the production of books of a specific region, namely "Sefarad" region, within the period between 10th and 16th

Keywords: Visual Analytics, Data Science, Big Data Analytics, Data Mining, Machine Learning, Knowledge Discovery, Digital Humanities, Jewish Book History, Codicology, Material Culture

Acknowledgments

I would like to express my sincere thanks to Professor Elsa Cardoso and Débora Matos for encouraging me throughout this path, for allowing me to have this opportunity and for sharing all the knowledge that was, without doubt, valuable. Thank you again for keeping me on track and sharing your experiences in such a patient and gracious way.

A special thank you to my girlfriend, for encouraging me on this journey, for giving up on several weekends and for traveling with me, thinking that I was going on vacation when in reality it was because of a conference. Finally, but not last, a big thank you to my family, especially my mother and sister, for all the encouragement given to finish this step.

... This work is dedicated to my grandfather.

CONTENTS

LIST OF FIGURES

ACRONYMS

ACM	Association for Computing Machinery
ARI	Adjusted Random Index
BI	Business Intelligence
BL	British Library
BLO	Bodleian Library in Oxford
CRISP-DM	Cross Industry Standard Process for Data Mining
CSV	Comma-Separated Values
D3	Data-Driven Documents
DH	Digital Humanities
DM	Data Mining
DSDAH	Data Science for Digital Art History
DSRM	Design Science Research
DSRM	Design Science Research Methodology
DW	Data Warehouse
EAJS	European Association of Jewish Studies
ETL	Extract, Transform and Load
FCA	Formal Concept Analysis
HTML	HyperText Markup Language
KDD	Knowledge Discover in Databases
MSS	Manuscript
NLI	National Library of Israel
PCA	Principal Component Analysis
RQ	Research Question
VA	Visual Analytics

SIGKDD Special Interest Group on Knowledge Discovery and Data Mining

Chapter 1

Introduction

The following dissertation proposes to address the lack of proper knowledge extraction tools by analytical means of Hebrew manuscripts. Manuscripts are the object of study of, primarily, experts in the field of Humanities, in particular from disciplines such as codicology and palaeography. Codicology can be referred to as the "archaeological study of the book" (Clark, 2014). It aims at tracing the manuscript's history by registering its producer, materials, decorations and techniques used. As for Palaeography, it is the study of ancient handwriting which also characterizes a manuscript and can answer the questions of who, when, and where a given manuscript was written. The interdisciplinary nature of the technical examination of books means that when someone has a specific research question, an outside collaboration is almost invariably required (Clark, 2014).

Codicology and palaeography, similarly to other disciplines in the Humanities, have increasingly benefitted from the introduction of computational approaches, the development of new tools and new means of knowledge (Schreibman, Siemens, & Unsworth, 2004, 2008). One positive aspect of the application of computational methods and tools in the Humanities is overcoming the limitations of a manual analysis (Schreibman et al., 2004). However, this often requires the adaptation of methods and tools to new objects of study. The exploration of domain-specific problems in art history, for instance, has led to significant developments in conservation sciences, and tools that are normally used by physicists and chemists have been adapted by conservators, librarians and art historians (Clark, 2014). Another example is how relational databases have been used to store information about book prints, and how visualization tools can be employed to explore the data and analytics, to provide insights on data based on statistics (Schreibman et al., 2004).

This methodological shift has fostered a new field of inquiry, commonly known as Digital Humanities (DH), which can be broadly characterized by the application of computational approaches to Humanities research. A fundamental aspect of Digital Humanities is the creation of new artifacts, which are digitally born and require a rigorous

study and understanding (Schreibman et al., 2004). However, most of these approaches, as initially are envisaged, require considerable technical expertise, and therefore are normally unavailable to Humanities scholars. This often results in the collaboration with technologists who, in turn, have little training in the Humanities disciplines (Schreibman et al., 2004). Nevertheless, experts from both sides must have enough comprehension of available technologies and grasp of the research needs, in order to judge the suitability and possibilities of the resources to use (Schreibman et al., 2004). Still, there is an ongoing discussion as to what the extent humanists (and Digital Humanists) should be proficient in programming languages, in order to facilitate collaboration and understand the possibilities of research of certain tools (Svensson & Goldberg, 2015). Still, collaboration in the Digital Humanities is, currently, more important than ever, to the extent that Kaplan, (2015) questions if we should still distinguish between computer scientists and humanists, in Digital Humanities communities.

However, instead of the tailoring of tools for research questions, the analysis of artifacts is frequently conditioned by those that are available. Schreibman et al., (2004) identifies this concern with visualization tools, where instead of being used for presenting information, they are used to create information. This matter is of importance, for quantitative exploration is gaining importance within the analysis of digitized manuscripts. While codicologists can collect massive amounts of heterogeneous datasets, often accompanied by digitized in high-resolution images of the artifacts, they still lack efficient and intuitive means to explore data and answer domain-specific research questions. That is, a new approach is needed to enable codicologists and other book experts with the quantitative exploration of large data (Chandna, Rindone, Dachsbacher, & Stotzka, 2016). Moreover, the collation of high-quality metadata data is still considered a fundamental step in the process. Spending time and effort expended at the creation stage, recording high-quality metadata, is likely to save users from future problems, and will result in a well-formed digital object, which will survive for the long term (Schreibman et al., 2004).

As we have seen, Digital Humanities have been using computation for a long time, to store information, collecting data, and analyze it with the goal of creating knowledge. Nowadays, information is intrinsically scattered and yet available in an immediate way. Huge amounts of information are capable of being stored in any device and the computational power has increased a lot. As such, Big Data has, inevitably, become a trend to be considered. However, can Digital Humanities embrace Big Data? The

challenges are posed by the term itself, ‘Big’ Data. To what extent does this apply to the Digital Humanities and to the Humanities? Understanding what is meant by ‘Big’ Data in this context will help to address the concerns on how the concept can be adapted, and even become a structured research field (Kaplan, 2015). The definition of Big Data is not always clear, but it can be interpreted as the field dedicated to the analysis of large amounts of information, that most of times are available in different sources, without any linkage between them (Erl, Khattak, & Buhler, 2016). It is, therefore, necessary to link these sources and process them to verify existing studies, that made the same analysis manually, and encounter potential new patterns.

Given the growing importance of large and networked cultural datasets, it is likely that Big Data will become an integral part of the Digital Humanities field (Erl et al., 2016; Kaplan, 2015). The challenge is how to make these cultural datasets ready to be networked. Given the heterogeneity of sources and lack of common standards, the gathered data is often unstructured. That is, it lacks a pre-defined model, that could be easily disseminated among different disciplines (Clark, 2014). In the specific context of manuscripts, their digitisation has long been under way and there is already a large corpus available. However, they still often lack metadata regarding the categorization of each object, or the semantic linkage between the different collections. This means that the existing data is devoid of intuitive means to be explored or to let the researcher explore domain-specific research questions. Moreover, the descriptive nature of data gathered by ‘traditional’ humanists lacks consistency in the process of gathering and registering data. Making this data devoid of intuitive means to be used as analytical sources. This is a current obstacle between Digital Humanities and classical approaches, since computing methods are gradually causing profound transformations that call for a reconsideration of its fundamental concepts (Gold, 2012).

One of the main challenges of Big Data lies in capturing, storing, analyzing, sharing, searching, and visualizing data. Yet one of the major aspects of Big Data analysis is that we can find interesting patterns in huge data sets, even though the results of the analysis are usually raw numbers, and by those numbers alone it is difficult to interpret anything. But if those numbers are represented visually, then it becomes much easier for our brain to find meaningful patterns and take decision accordingly (Ali, Gupta, Nayak, & Lenka, 2016). Data insights through visualization within the Big Data trend is a hot topic and the problems are cross-domain rather than domain-specific. The huge amounts of data, plus the high dimensionality linkage in terms of features, represent a challenge in terms of

exploratory analysis (Xie, Chenna, He, Le, & Planteen, 2016). Data visualization has had a critical role within the Digital Humanities, not only for knowledge extraction, but also for sharing experiences and to detect outliers. However, the same issues apply, and handling such massive amounts of information and characteristics makes it harder to perform proper visualizations, if the dimensionality goes beyond three dimensions. Nonetheless, it should be considered an area of study on its own (Kaplan, 2015).

This data science field, which is intertwined with other important concepts also of growing importance, such as Big Data and data-driven decision making (Provost & Fawcett, 2013), is itself an area of expertise which includes techniques for data cleaning, structuring, storage and analysis/visualization. Furthermore, most of the techniques derive from statistics, algebra and mathematics, and there is still a lack of a unified theorem that could be applied to all kinds of data sources. As stated before, one of the concerns within Digital Humanities is the lack of proper expertise regarding computer science. If the domain-specific nature of data science is still a reality, how can we provide the proper tools for Humanities' specific domains, such as codicology, and make proper quantitative analysis and exploration?

As stated before, the uncertainty of these domain-specific sources makes it hard to retrieve knowledge and properly communicate it. Therefore, Digital Humanities could benefit from the rising of visual analytics methods to fulfill that gap. Just as information visualization has changed our view on databases, the goal of Visual Analytics (VA) is to make our way of processing data and information transparent for an analytic discourse. The visualization of these processes will provide the means of communicating about them, instead of being left with the results (Keim et al., 2008). Visual Analytics can be seen as the science of analytical reasoning facilitated by visual representations of data. This implies the use of different types of analysis, data and a systematic research method to provide new and deeper insights about a certain problem or domain improving the knowledge extraction. Moreover, as a research field it combines the skills and knowledge from different disciplines and is deeply related to decision support and Business Intelligence (BI) systems.

1.1 Motivation and Research Context

In the Humanities/Digital Humanities, although still in an early stage, data visualization is increasingly shaping the field. Its potential is shown in the fostering of

new means of data exploration, often heterogeneous and uncertain in nature, and by opening new research questions across its various fields. Due to the heterogeneity in the high volume of the information collected by experts of different areas and its intrinsic uncertainty, as stated before, the data lacks standards and patterns, making the computing analysis harder to process, and restricting the exploration of manuscripts (Chandna et al., 2016). What can we do with huge amount of heterogeneous data? If we consider that each manuscript's metadata is information, if that information is not properly structured, we cannot extract knowledge. This is because knowledge appears from the linkage between information sources, and from the patterns it might have. Furthermore, exploration and visualization are critical to extract knowledge. Challenges, therefore, are still seen at a methodological level, particularly in the emphasis on quantitative analysis (Graham, 2017), but also in terms of the acceptance of results by the experts (although gradually less so). Some arguments must be considered, particularly the distinction that 'while the scientist's methods can be paraphrased without any loss, in the Humanities the description itself is understood to be part of the method' (Graham, 2017). The inter- and multidisciplinary nature of Digital Humanities is, therefore, the perfect background for a collaboration between BI and VA with Humanities disciplines.

Together with this dissertation, the project *CodicoDaViz* was created, by establishing a partnership between ISCTE-IUL and the University of Münster, in Germany. *CodicoDaViz* was developed with the goal of applying data visualization techniques to the field of codicology. Furthermore, this dissertation discusses the implementation of all the technical artifacts of the *CodicoDaViz* project. The initial motivation of this project concerns a specific research need regarding the transition from manuscript format to print, of books in Hebrew script and type. The substantial amounts of metadata already available regarding the material aspects of books, but also the heterogeneity and dispersion of availabledata, opened up the possibility of adopting data visualization techniques. Therefore, a central goal of this project is to categorize, clean, analyze, and visualize the raw metadata that already exists in relevant data sources. This research project also focuses on the implementation of an analytical environment that can be used by Humanities experts to explore the existent information. Leveraging experts' capabilities from the analysis of a single historical object at each time to a bigger picture of, perhaps, cultural insights of several manuscripts.

1.2 Research Methodology

The proposed approach uses the Design Science Research Methodology (DSRM) principles proposed by (Hevner & Chatterjee, 2010) as seen in Figure 1.1 which helps to define the scope of the proposed work. To contextualize, as stated in the Environment from Figure 1.1, the domain of this dissertation is within the Humanities field, more specifically, Codicology and Palaeography of Hebrew Manuscripts. Therefore, it is intended to help experts of this area with their studies providing a proper exploration model (knowledge base) solving the uncertainty within the currently available data sources.

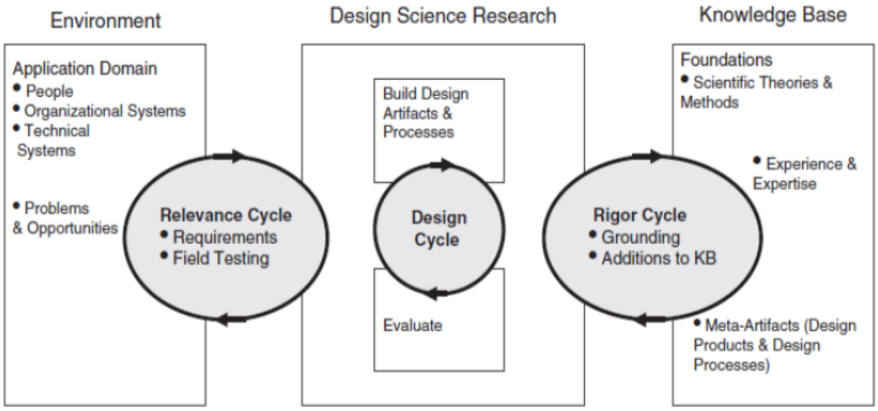

Figure 1.1 - Design science research cycles. (Retrieved from (Hevner & Chatterjee, 2010)*, p. 16)*

We will focus on the design, development, demonstration and evaluation of artifacts using the nominal process sequence proposed by Peffers, Tuunanen, Rothenberger, & Chatterjee, (2007) as displayed in Figure 1.2. These artifacts are expected to, when applied, help us achieve a solution for our perceived problem.

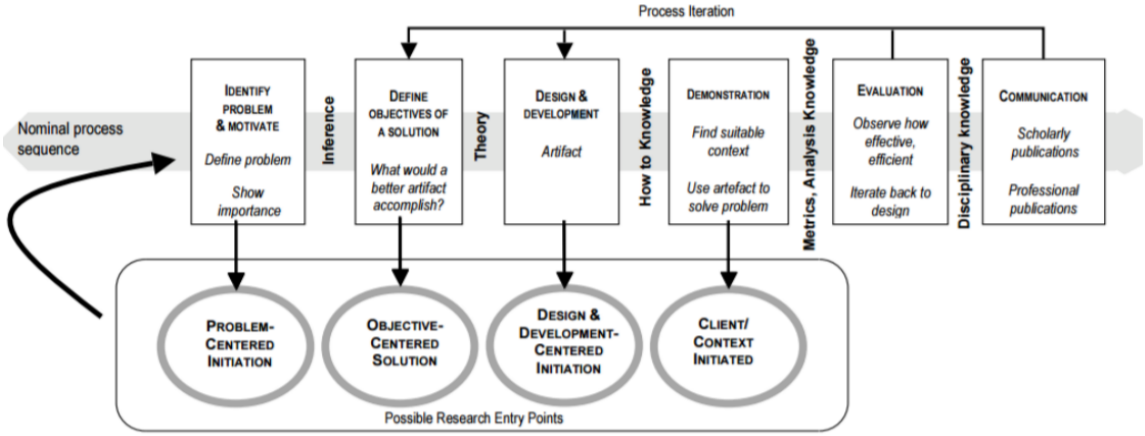

Figure 1.2 - Design Science Research Methodology in nominal process sequence form. (Retrieved from (Peffers et al., 2007)*, p. 14)*

Hevner & Chatterjee, (2010) maps the seven DSRM guidelines into a checklist grouped by Relevance, Design and Rigor cycles. In Figure 1.3 below we describe and schematize these questions in the form of a nominal process sequence to provide an easier understanding of the proposed artifacts, solutions, audience and planned entry points.

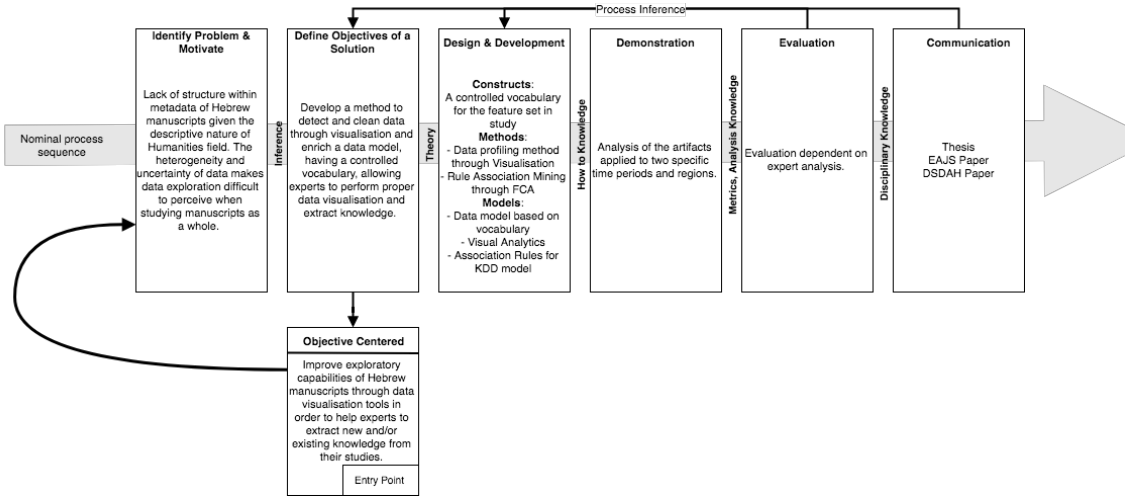

Figure 1.3 - DSRM adopted in this thesis. (Adapted from Figure 1.2)

We have an objective centered solution entry point since our research started based on the goal of, through visual analytics, extract knowledge to confirm existing studies or to raise new evidences.

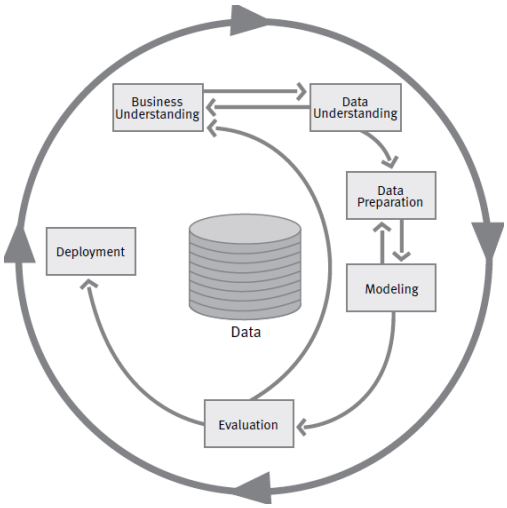

Figure 1.4 - CRISP-DM reference model. (Retrieved from (Chapman et al., 2000)*, p. 10)*

Additionally, in the design and development phase, the Big Data analytics lifecycle proposed by (Erl et al., 2016) is used to structure and plan the implementation phase. The Big Data analytics lifecycle, as proposed, has nine stages, which are very similar to the flow of CRISP-DM from Figure 1.4 and may be seen as a higher order model which includes data visualization, acquisition and extraction (Erl et al., 2016).

1.2.1 Identify Problem and Motivate

Currently, experts from the Humanities analyze each manuscript one by one tracing its characteristics and attributes manually. The descriptive nature of this field of study often leads to the lack of a controlled vocabulary that makes quantitative analysis impossible to achieve. Consequently, retrieving knowledge from a set of manuscripts describing trends and cultural patterns is very difficult to achieve. Moreover, the heterogeneity and uncertainty within data (caused by the lack of a proper data collecting standard) makes the application of computational methods rising within Digital Humanities, as well as linking sources from interdisciplinary study groups, nearly impossible to accomplish.

1.2.2 Define Objectives for Solution

Within the Codicological studies, more specifically the Hebrew manuscripts around the 15th century, an approach is proposed to address the problematic of heterogeneous metadata and inherent uncertainty providing a toolset to extract, process and store data from several sources into a structured model capable of storing this information in such way it is able of being explored and properly visualized, giving the experts a bigger picture of the available collections. This multidimensional model aims to reduce the uncertainty and heterogeneous concerns by storing the processed data with a controlled vocabulary that can be explored and visualized. To achieve this, a method for data profiling through the use of visualization is proposed, in order to achieve a controlled vocabulary.

Data visualization with exploratory analysis should, afterwards, be available online for experts to easily use. Being able to computationally analyze the entire corpus of manuscripts to discover cultural trends, by tracing behaviors along the centuries through visual analytics could provide new insights that could verify known assumptions as well as raise new ones. All of this could be presented in a richer way across study groups within the Humanities, since scholars would be able to showcase insights to their colleagues.

However the amount of metadata available even if we categorize and standardize it, the feature space is too big to manually infer meaningful visualizations and/or relations. Without previous input in terms of key metrics and/or indicators, the experts should have available a tool that could suggest associations for further exploration and analysis of

specific feature relations. To help the exploratory and analytical analysis, concept analysis and association methods are used, aiming to find trends and relevant patterns and/or associations in data that would suggest visualizations of most relevant features.

Based on the proposed approach, during the present dissertation, we will try to answer the following questions:

RQ1. Is it possible to easily harvest data from several sources and build a common data model to hold the information for codicological analysis of Hebrew books?

RQ2. Does the data model enable the discovery of new and relevant patterns to art history researchers?

1.2.3 Design and Development

To achieve our goals, we will create six artifacts. Each artifact addresses a part of our problem:

1. **(Construct) A controlled vocabulary of specific terms based on relevant features** – The goal of this artifact is to address the uncertainty and lack of structure that currently exists in order to build a categorized corpus capable of being explored.
2. **(Method) A methodology for data cleaning and integration** – design a set of rules and standards to clean the data based on the built vocabulary doing several iterations through data visualization to spot inconsistencies within the acquired corpus.
3. **(Method) Rule Association Mining through Concept Analysis** – since we don't have requirements in terms of indicators to perform visualizations, it is hard to foresee which visualizations, which analysis or which dashboards are relevant to answer the expert's questions. Moreover, the feature space of available metadata within each manuscript makes it impossible to achieve, by manually designing meaningful visualizations. The proposed method applies a mathematical algorithm called Formal Concept Analysis to the corpus to build a context tree of concepts.
4. **(Model) A multidimensional model based on vocabulary** – A model to store the data standardized with the applied methods and based on the designed vocabulary in order to be available for analytical explorations.

5. **(Model) Visual Analytics** – Build templates that could be applied to different collections of manuscripts sharing the same vocabulary to allow experts to have a richer perception and insights through visualization that could raise new or verify existing knowledge.
6. **(Model) Association Rules for Knowledge Discovery** – from the method of applying Concept Analysis this model stores the association rules in such way that experts can have insights of data relations and even enhance the visualization templates with new insights.

1.2.4 Demonstration

To demonstrate our artifacts, our initial corpus was taken from the online platform Sfardata (Beit-Arié, 2017) and afterwards enriched through the method of data profiling by using personal catalogues. This platform "*locates all the medieval codices written in the Hebrew script, which contained explicit production dates or at least scribe names; to study and document all their visual and measurable material features and scribal practices in situ, i.e. in the libraries in which they were kept; and classify these features and practices in order to expose a historical typology of the hand-produced Hebrew book and provide users of Hebrew manuscripts with a tool for identifying the production region and assessing the period of the studied manuscripts. Indeed, since the initiation of the project, almost all the dated manuscripts that were located have been studied and documented in some two hundred and fifty libraries and private collections*"[1].

We will use two different datasets, all of them produced in the Sephardi region, from different periods. Our initial dataset was retrieved from the codicological descriptions of almost all dated Hebrew manuscripts that have been collected and stored in Sfardata (Beit-Arié, 2017) between periods 1400 and 1500 for Sephardi scripts. This dataset was the baseline of work where we applied and built the methodology of data cleaning and visualization. The second dataset increases the period from 900 to 1400 and from 1500 to 1540 were used to evaluate the effectiveness of the data cleaning process through visualization measured in terms of effort and to apply association rule mining and concept analysis to extract knowledge and enhance our visualizations through dashboards. Moreover, experts are expected to include printed books from this extended period for the same region.

[1] Definition taken from Sfardata (Beit-Arié, 2017) website, http://sfardata.nli.org.il

1.2.5 Evaluation

The evaluation of our proposed framework is still undergoing and will not be completed in the timeframe of this dissertation. However, in the context of this research work, an assessment will be made by our expert within the Humanities to see if the proposed work can enhance the experience of knowledge extraction through visualization.

1.2.6 Communication

The communication of this research was done through the publication of two papers within the Digital Humanities field:

- Marques de Matos, D., Pateiro, T. & Cardoso, E. (2018). Crossing the Line: Data visualization of Codicological Data of Hebrew Manuscripts and Incunabula. 11th Congress of the European Association of Jewish Studies (EAJS, 2018)
- Pateiro, T., Cardoso, E. & Marques de Matos, D. (2018). Visual analytics of Hebrew manuscripts codicological metadata. In 1st KDD Workshop on Data Science for Digital Art History: tackling big data Challenges, Algorithms, and Systems, organized in the ACM SIGKDD 2018. London

The first publication was presented at the conference from the European Association of Jewish Studies (EAJS), where we demonstrated how Humanities experts and computer scientists could work together to provide richer visualizations of Hebrew manuscripts collections. The second publication is a full paper presented at the Workshop of Data Science for Digital Art History (DSDAH), organized as part of the ACM KDD2018 conference. The proceedings of this workshop will be published as a special issue at the International Journal of Digital Art History (due in November 2018). This paper presents our methodology of data preparation and discusses how visualization can provide richer insights and even spot incoherencies during the data cleaning process.

1.3 Document Structure

This document is structured as follows:

- Chapter 2 presents the Related Work, introducing concepts as well as similar work on other DH fields and how they could help to address our problem;

- Chapter 3 presents our Case Study and the Big Data Lifecycle method used to describe how the problem was tackled and how the DSRM artifacts were built;
- Chapter 4 applies the Visual Analytics templates in the perspective of Humanities experts to demonstrate the impact of these tools in the story telling of the data;
- Chapter 5 summarizes our research made so far, the limitations raised and how we intend to address them in future work.

Chapter 2

Related Work

In this chapter we will present some of the research about the subjects that will be needed for the Design and Development chapter (Chapter 3). The research includes some works within the Digital Humanities field using computational methods for knowledge discovery as well as data structuring which we believe will serve as an input, alongside with other gained knowledge, to the next research steps.

Similar to other fields, digital approaches to Hebrew manuscripts have focused primarily on text, i.e., authorship identification, linguistic patterns, digital editions, text encoding, and so on (Gold & Klein, 2016). The most relevant exception is the work on automatic identification of join fragments developed by the Genizah projects (Wolf, Potikha, Dershowitz, Shweka, & Choueka, 2011). In the context of Hebrew books, the main collections of manuscripts such as those at the National Library of Israel (NLI), British Library (BL), or the Bodleian Library in Oxford, (BLO) to name but a few, have made a substantial part of their materials available online. These and many other collections around the world have made their materials available in an online platform for digital access to manuscripts around the world, known as *Ktiv* (NLI, n.d.), hosted by the NLI. However, metadata provided by most collections does not go beyond catalogue descriptions, often lacking codicological metadata. In contrast, the materiality of Hebrew manuscripts is thoroughly described and available in a database known as Sfardata (Beit-Arié, 2017; Sfardata, n.d.). In many senses, Sfardata (Beit-Arié, 2017; Sfardata, n.d.) is a unique tool. It has no counterpart in other book cultures, it hosts substantial amounts of descriptions of dated Hebrew manuscripts until 1540, and has drawn methodologies with impact in material culture studies (Bausi et al., 2015). That being said, these tools still lack intuitive means to explore domain-specific research questions dealing with codicological metadata.

To some extent, this can be understood by the very nature of codicological data, which is intrinsically descriptive and heterogenous. In other words, it can be quantitative (measurements), and simultaneously subjective and qualitative (for instance, in terms of palaeographical descriptions). Particularly the visualization of uncertainty is still in

discussion, and a much pertinent one within the Humanities (Jänicke & Wrisley, 2012). Although codicological metadata still lacks a systematic set of rules, other adjacent fields such as palaeography are already setting a broad frame of work where Big Data can be processed by computers, but experts are as necessary, particularly to deal with ambiguous and complex datasets. As such, the process flow is semi-automatic, interactive and iterative, and results can be re-used (Hassner, Rehbein, Stokes, & Wolf, 2013). The potentiality of quantitative analysis and/or exploration of codicological units can be seen in platforms such as *eCodicology* (Chandna et al., 2016) for medieval manuscripts where we can see an effort to solve the heterogeneity and uncertainty using computational methods and visualization tools. Similar to companies, the Humanities face the same challenges when talking about Big Data. Although the volume might not be the same, the variety and veracity are key topics to consider when working with these inter-disciplinary collections of metadata.

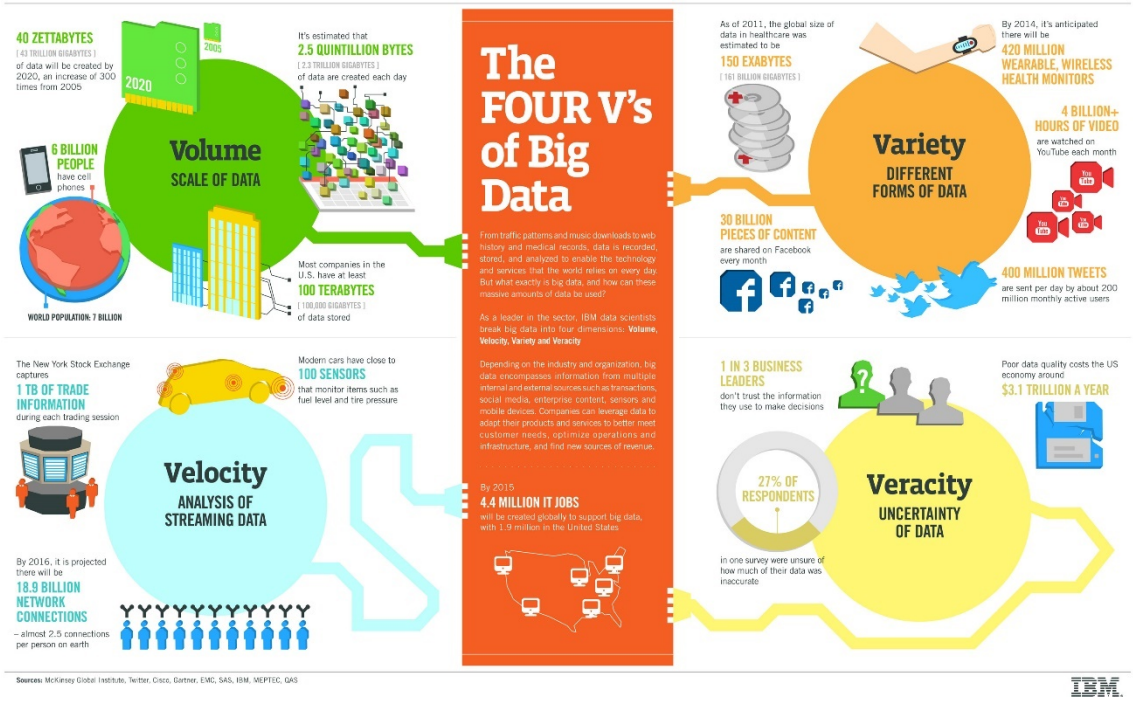

Figure 2.5 Four V's of Big Data. Taken from IBM (2018)[2]

The challenges, as seen in Figure 2.5, are clearly the same given the descriptive and uncertain nature of the available data from Hebrew manuscripts. The variety can be seen as the number of collections available, the dispersion of information, from Sfardata to Ktiv, and the problematic veracity of information, given the collaborative nature lacking a common method (Graham, 2017).

[2] Infographic available in https://www.ibmbigdatahub.com/infographic/four-vs-big-data

Visual Analytics can play an important role when talking about veracity, complementing the Big Data Analytics nowadays. The driver for this is increasingly higher volume and, most of all, complexity of the data (L. Zhang, Stoffe, & Behrisch, 2012).

Visual analytics is the science of analytical reasoning facilitated by visual representations of data. This implies the use of different types of analysis, data and a systematic research method to provide new and deeper insights about a certain problem or domain. Visual Analytics is a recent research area that combines the skills and knowledge from different disciplines and is deeply related to decision support and business intelligence systems. It has been applied to a diverse set of contexts, for instance, in precision agriculture to improve decisions about crops (Wachowiak, Walters, Kovacs, Wachowiak-Smolkov, & James, 2017), in healthcare to compare drug information enabling a faster integration into practice of new drugs (Lamy et al., 2017), or in software engineering (Staron, Sahraoui, & Telea, 2018). In the Humanities, although still in its early stages, data visualization is increasingly marking the field. Its potential is shown in the fostering of new means of data exploration, often heterogeneous in nature, and by opening up new research questions across its various fields. Challenges are still seen at a methodological level, particularly in the emphasis on quantitative analysis (Graham, 2017), but also as in terms of acceptance of results by the experts (although gradually less so). Some arguments must be taken into account, particularly the distinction that “while the scientist’s methods can be paraphrased without any loss, in the Humanities the description itself is understood to be part of the method” (Graham, 2017). If “mapping data to visual representations has been used for centuries to reveal patterns, to communicate complex ideas, and to tell stories” (Bailey & Pregill, 2014), current tools bring to the table this new aspect of interaction and iteration with the experts. Whilst visualization can be a discovery tool, it is primarily a means to refine arguments and illustrate conclusions already drawn (Wulfman, 2014). That is, graphic representations such as charts are not the actual data but an interpretation of it, to answer a specific research query, even if visualization allows complex findings to be presented in an informative and engaging way (Radich, 2017). For instance, radial trees and parallel coordinates seem to have a wide use when exploring high dimensional data (Hinrichs, Forlini, & Moynihan, 2016). This is a useful solution when data is categorical, but standard plotting based on numerical axis are harder to use in this case. Chandna et al., (2016) propose a framework for visual analysis of medieval manuscripts, where the system uses image segmentation and feature extraction from digitised manuscripts to

create measurements that are combined with other metadata to visualize the information in a radial tree or parallel coordinate plot.

Ali et al., (2016), explore commercial solutions for Big Data visualization such as Tableau, Microsoft Power BI, and propose the usage of link/network analysis techniques as useful visualization tools for high dimensional data (i.e., a dataset with a high number of features).

These proposed techniques (i.e., link/network analysis) still require manual tuning and do not always allow the development of a storytelling type of narrative. Communicating results through data visualization and engaging with an audience should not be overlooked. Windhager et al., (2018), try to go beyond the traditional approaches to visualization on grid-based interfaces, and instead explore them as complex and comprehensive information spaces by the means of interactive visualizations in the scope of cultural heritage collections.

The remainder of this chapter is organized in the following sections:

- Section 2.1 introduces the Big Data Lifecycle and how it relates to the CRISP-DM model;
- Section 2.2 covers the major task of data cleaning and preparation for source integration;
- Section 2.3 describes some Data Visualization techniques for visualizing high dimensional sources;
- Section 2.4 introduces knowledge discover and some computational techniques for pattern analysis.

2.1 Big Data Lifecycle

The CRISP-DM (Cross Industry Standard Process for Data Mining) methodology is described in terms of a hierarchical process model, consisting of sets of tasks described at four levels of abstraction (from general to specific, as in Figure 2.6): phase, generic task, specialized task, and process instance. As a standard process model it was created to address data mining concerns (Chapman et al., 2000).

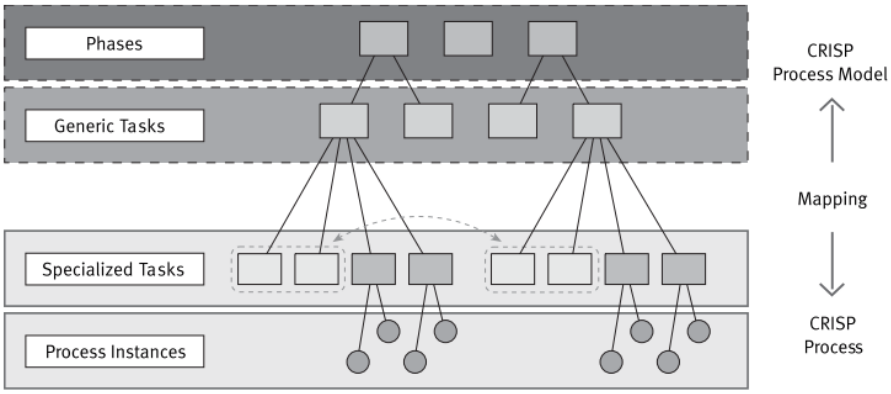

Figure 2.6 - Four level breakdown of the CRISP-DM methodology. Retrieved from (Chapman et al., 2000)*, p. 6*

The data mining context drives mapping between the generic and the specialized level in CRISP-DM. Currently, we distinguish between four different dimensions of data mining contexts as seen in Figure 2.7.

	Data Mining Context			
Dimension	**Application Domain**	**Data Mining Problem Type**	**Technical Aspect**	**Tool and Technique**
Examples	Response Modeling	Description and Summarization	Missing Values	Clementine
	Churn Prediction	Segmentation	Outliers	MineSet
	...	Concept Description	...	Decision Tree
		Classification		...
		Prediction		
		Dependency Analysis		

Figure 2.7 - Dimensions of data mining contexts and examples. Taken from (Chapman et al., 2000)*, p.7*

With this model we can foresee the contexts that will need to be addressed in future work, namely dependency analysis, missing values, and decision trees (association rules).

Making the bridge from this method, as already shown in Figure 1.4, to Big Data Lifecycle is possible and makes a strategic approach to the problem, where the data scientist is required to follow a step-by-step approach to guarantee the data is accurate providing truthful insights, meaning, achieving veracity (Berwind, Bornschlegl, Kaufmann, & Hemmje, 2016), as seen in Figure 2.8.

Big Data analysis differs from traditional data analysis primarily due to the volume, velocity and variety characteristics of the data being processes. To address the distinct requirements for performing analysis on Big Data, a step-by-step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing and repurposing data (Erl et al., 2016).

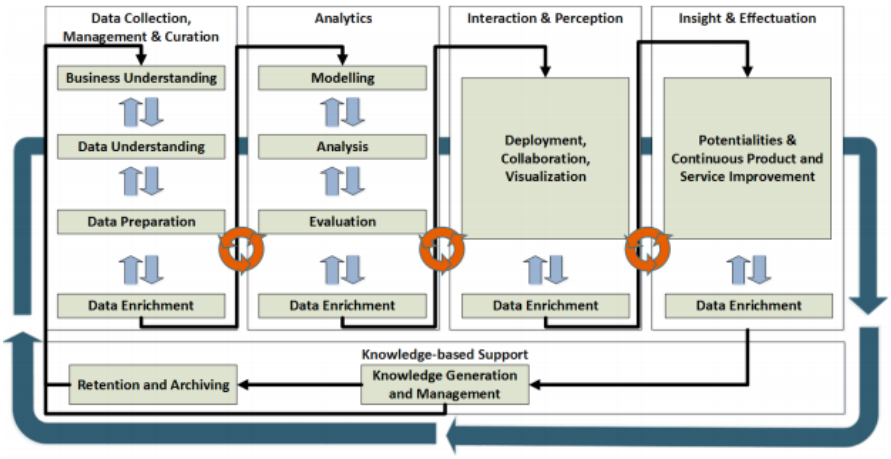

Figure 2.8 - CRISP4BigData Reference Model Version 1.0 from Kaufman 2016. Taken from (Berwind et al., 2016)*, p.5*

2.2 Data Cleaning and Preparation

Data can be unstructured without any indication of validity, making the data cleaning one of the most important steps in the lifecycle since it can provide wrong decisions (Erl et al., 2016). It is estimated that due to the importance of this step, it should usually take 80 percent of the time invested in any data project (S. Zhang, Zhang, & Yang, 2003).

Several techniques exist to address the problems regarding this critical stage. However, a good context knowledge is of utmost importance to determine which techniques to apply. Moreover, as seen in Figure 2.8, data understanding plays an important role in such way that the analyst should look at the features and correct the values (Abbot, 2004). Variable Cleaning (transforming variables to standardize the vocabulary) and Feature Creation (adding value to our dataset) are some of the techniques available (Abbot, 2004).

2.2.1 Missing Values

One of the most common concerns when handling unstructured sources is the presence missing values. Taking the Humanities as example, its descriptive and its collaborative nature may have multiple representations for the same value, as we will demonstrate in section 3.3.4 (Chapter 3). Moreover, the presence of null values and or unknown values is a possibility.

Abbot, (2004) provides several approaches to deal with missing values, such as: (1) deletion, which is the simplest solution. A previous knowledge of the scope is required in order to decide whether or not to delete entire columns or entire rows, (2) mean and

median imputation, where we analyze the mean/median of a feature and based on these statistical calculations we perform the filling of the calculated value, (3) imputation with a constant, where we assign a specific value as a constant, for example, assuming "unknown" as a valid and analyzable value, (4) business specific rules imputations, where an empty cell might be caused for a specific value in another feature.

2.3 Data Visualization of high dimensional sources

Visualizing high-dimensional data is a challenging process, and the visualization of uncertainty is still a topic for discussion within the community (Jänicke & Wrisley, 2012). Visualization allows us to take complex findings and present them in a way that is informative and engaging (Radich, 2017). However, it becomes more complex on larger amounts of information, and as it grows the effective communication must remain easy to the users (Radich, 2017). Not only for the users, understanding the data at this stage of the lifecycle through data visualization might help to have a glance of the state of those sources (Keim et al., 2008).

One of the biggest challenges in data visualization is to find general representations of data that can display the multivariate structure of more than two variables. Several graphic types like mosaic plots or parallel coordinate plots, and the grand tour, have been developed over the course of the last three decades. But all of them lack interactivity, the ability to engage the user in such way slicing the information should be easy and available (Theus, Chen, & Unwin, 2008).

Among the most classical plots, Theus et al., (2008) identifies the ones that are more useful (although lacking interactivity):

1. Mosaic Plots (purely for categorical data);
2. Parallel Coordinate Plots (only for continuous data);
3. Projection Pursuit and Grand Tour (only for continuous data);
4. Trellis Displays (for mixed scales);

However, as show in Figure 2.9, for example the parallel coordinate plots, the problem of over plotting is much more serious as it is with scatterplots.

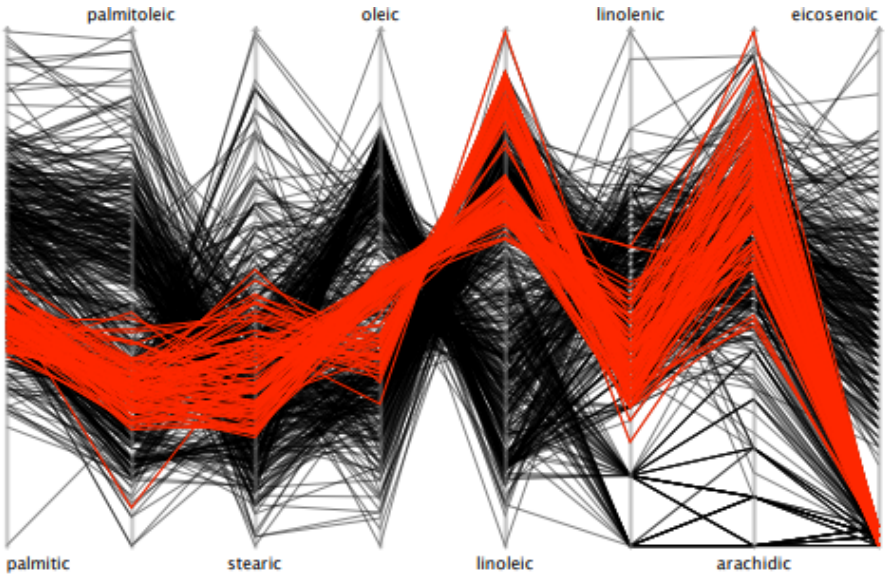

Figure 2.9 - Example of over plotting through parallel coordinates, taken from (Theus et al., 2008)

Since features are only visible at adjacent axes, a flexible permutation mechanism is needed to get a comprehensive view on all data (Theus et al., 2008).

Decision support systems are often built on top on known indicators and metrics, however, when no previous input is well defined, building dashboards with random features might cause noise instead of building understandable knowledge.

Users should be able, therefore, to explore data associations and have quick insights. Data-Driven Documents (D3) is a novel representation approach to visualization for the web. Rather than hide the underlying scene graph within a toolkit-specific abstraction, D3 enables direct inspection and manipulation (Bostock, Ogievetsky, & Heer, 2011). Allowing these designed tools to interactively engage the user to explore data associations, as seen in Figure 2.10 which could help finding patterns in the data.

Figure 2.10 Example of Geographic D3 and Radial Tree in D3 for content exploration. Taken from (Bostock et al., 2011)

2.4 Knowledge Extraction

The history of Data Mining (DM) and Knowledge Discovery (KD) is not much different. In the early 1990s, when the KDD (Knowledge Discovery in Databases) processing term was first coined, there was a rush to develop DM algorithms that were capable of solving all the problems of searching for knowledge in data. Apart from developing algorithms, tools were also developed to simplify the application of DM algorithms. From the viewpoint of DM and KD process models, the year 2000 marked the most important milestone: CRISP-DM was published (Chapman et al., 2000). CRISP-DM is the most used methodology for developing DM and KD projects (Marbán, Mariscal, & Segovi, 2009).

KDD is defined as the non-trivial extraction of valid, implicit, potentially useful and ultimately understandable information in large databases (Han, Kamber, & Pei, 2011)

Knowledge based support as addressed before in Figure 2.5 from Chapter 2 Related Work (page 13) is the ability to make decisions based on the gained knowledge, whether from visual analytics tools, or from automatic computational methods such as pattern analysis and data mining. Knowledge, unlike information, is useful, whereas information is just refined data. Extracting knowledge enables the observer to clearly understand the outcome and make decisions on top of it.

Most of the times, the researcher might not have the necessary knowledge to know for what and for whom he wants to produce a specific set of dashboards over specific set of features (Keim et al., 2008). Having machine learning or data mining tools to find patterns and associations without having such input might become useful to extract unknown information. Machine learning and data mining are research areas of computer science whose quick development is due to the advances in data analysis research, growth in the database industry, and the resulting market needs for methods that are capable of extracting valuable knowledge from large data stores (Fürnkranz, Gamberger, & Lavrač, 2012). Moreover, the complexity of the problem is exponential in the size of all of the possible combinations within the dataset feature/value relation, and since this relation has to be scanned several times during the process, efficient algorithms for mining frequent item sets are required (Lakhal & Stumme, 2005).

2.4.1 Data Mining and Unsupervised Learning Algorithms

Unsupervised learning algorithms concerns the analysis of unclassified examples. Moreover, it is used when the goal is not to predict but yet to find relationships within the unlabeled dataset (Fürnkranz et al., 2012). Data mining methods are typically unsupervised. They are used to induce interesting patterns (such as association rules) from unlabeled data. The induced patterns are useful in exploratory data analysis (Fürnkranz et al., 2012).

Frequent itemset and pattern mining leads to the discovery of associations and correlations among items in large transactional or relational data sets. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining such patterns from their databases. The discovery of interesting correlation relationships among huge amounts of business transaction records can help in many business decision-making processes such as catalog design, cross-marketing, and customer shopping behavior analysis (Han et al., 2011).

2.4.1.1 K-Means Algorithm

K-Means are considered as one of the simplest algorithms in machine learning to solve the clustering problem, in which the number of clusters k is chosen in advance, after which the goal is to partition the inputs into sets in $S_1, \dots, S_k$ a way that minimizes the total sum of squared distances from each point to the mean of its assigned cluster. In this algorithm the number of clusters are predefined as 'K', which refers to the number of centroids. The algorithm iteratively calculates the closeness between points to each centroid and as a result compact groups of items are formed as clusters. K-Means uses a squared error function as the objective function, and it minimizes the squared error distance between items and chosen centroid (Grus, 2015). Partition based algorithms are iterative and use a distance formula to measure similarity between items (Bhargav & Pawar, 2016).

The accuracy of a clustering algorithm is dependent of the shape of the produced cluster in space. It is known that K-Means only handles circular convex shapes and performs what is called hard clustering. This type of clustering indicates that one point in space either belongs or not to a cluster. Hierarchical clustering is still a hard-clustering algorithm but allows non-convex shapes (Bhargav & Pawar, 2016).

2.4.1.2 Hierarchical Clustering

An alternative approach to clustering is to "grow" clusters from the bottom up. At the end, we will have one giant cluster containing all the inputs. If we keep track of the merge order, we can recreate any number of clusters by unmerging. For example, if we want three clusters, we can just undo the last two merges. Unlike K-Means, Hierarchical Clustering requires a measure of similarity between groups of data points (Grus, 2015).

In Hierarchical clustering, data is organized according to hierarchy of proximity which are generated by the intermediate nodes. As such, the entire dataset can be represented by a dendrogram in which the leaf nodes represent the data itself. Like a tree, a data-class can have a subclass and this division can continue to the leaves.

Hierarchical Clustering strategies can be of two types (Bhargav & Pawar, 2016) , which in turn might be suitable for the previously proposed visualizations:

- Agglomerative ("bottom-up") – each row starts as a cluster and then pairs of clusters are merged;
- Divisive ("top-down") – every row starts as one cluster and then splits are performed in a recursive way.

2.4.1.3 Association Rule Mining

Association rule learning is an unsupervised learning method, with no class labels assigned to the examples (Fürnkranz et al., 2012). It finds all the rules existing in the database that satisfy some minimum support and minimum confidence constraints. For association rule mining, the target of mining is not predetermined (Liu, Hsu, & Ma, 1998). The problem of association rule mining has first been stated in 1993. Five years later, several research groups discovered that this problem has a strong connection to Formal Concept Analysis (FCA) (Lakhal & Stumme, 2005).

This type of algorithm is often used for recommendation systems, from e-commerce platforms to marketing segmentation (Lin, Alvarez, & Ruiz, 2002). Cases like "*If a customer buys bread, he's 70% likely of buying milk.*"[3] might help researchers finding unlikely association within data.

[3] Example taken from https://www.upgrad.com/blog/association-rule-mining-an-overview-and-its-applications/

2.4.2 Formal Concept Analysis

Since determining the frequent item sets is the computationally most expensive part, most research has focused on this aspect. Most algorithms follow the way of the well-known *Apriori* algorithm (Han et al., 2011). Other algorithms are based on the extraction of maximal frequent item sets. They combine a level wise bottom-up traversal with a top-down traversal in order to quickly find the maximal frequent item sets (Han et al., 2011). However, all of these algorithms have to determine the supports of all frequent item sets becoming computationally expensive. As such, some infrequent item sets could be discarded from these computation to increase performance (Han et al., 2011).

Using basic results from Formal Concept Analysis (FCA), it is possible to derive a concept hierarchy (ontologies) from an item set of object and their attributes. The use of FCA allows not only an efficient computation, but also to drastically reduce the number of rules that have to be presented to the user, without any information loss (Han et al., 2011). Formal concept analysis can be considered as one of the best formalization of the Data Mining field (Buzmakov, 2015).

Formally, FCA is a method based on mathematical theories oriented at applications in knowledge representation, knowledge acquisition, data analysis and visualization (known as conceptual clustering) (Obiedkov, 2018) where each concept intent is exactly the largest itemset of the equivalence class of θ it belongs to. For any itemset $X \subseteq M$, the concept intent of its equivalence class is the set X''. The concept intents can hence be considered as 'normal forms' of the (frequent) item sets. In particular, the concept lattice[4] contains all information to derive the support of all (frequent) item sets (Han et al., 2011).

Formal Concept Analysis operates by encoding the data set transforming the feature space as true or false values where the categorical values are transposed to new features using the Feature Creation method proposed by (Abbot, 2004) and presented in section 2.2.

[4] A concept lattice is an abstract structure as part of order theory and algebra. It can be seen as an ordered (or partially ordered) set based on mathematical functions to determine its order.

	H	CAD	OH	P	AAE	F	O=
12	x	x			x	x	x
13	x	x		x		x	x
10		x	x				x
4	x			x	x	x	

(a) Formal Context.

Molecule	Binding Mode
12	DFG-out
13	DFG-out
10	Type-1
4	Type-1

(b) Molecule Binding Modes.

Figure 2.11 - Set of formal concepts using binary encoding of dataset. Example taken from (Buzmakov, 2015)*, p. 27*

As seen in Figure 2.11, the formal context transposes the categorical values for each preselected object (in the present case, the molecule) to new features and assigns a 'X' if that category is present or not for the given object.

A formal concept corresponds to a pair of maximal sets of objects and attributes, i.e. it is not possible to add an object or an attribute to the concept without violating the maximality property (Buzmakov, 2015).

Concept	A Set of Molecules (Extent)	A Set of Substructures (Intent)
C_0		H, CAD, OH, P, AAE, F, O=
C_1	4	H, P, AAE, F
C_2	12	H, CAD, P, F, O=
C_3	13	H, CAD, AAE, F, O=
C_4	10	CAD, OH, O=
C_5	12, 13	H, CAD, F, O=
C_6	13, 4	H, P, F
C_7	12, 4	H, AAE, F
C_8	12, 13, 10	CAD, O=
C_9	12, 13, 4	H, F
C_{10}	12, 13, 10, 4	

Figure 2.12 - A set of formal concepts w.r.t context on Figure 2.11. Example taken from (Buzmakov, 2015)*, p. 27*

Figure 2.12 shows all concepts that can be found for the context. Formal concepts can be partially ordered with respect to the extent inclusion (dually, intent inclusion). For example, $(\{13\};\{CAD, O =, OH\}) \leq (\{13,10,12\},\{CAD, O =\})$. This partial order of concepts is shown in Figure 1.2. The number of formal concepts for a given context can be exponential with respect to the cardinality of the set of objects or the set of attributes. Moreover, even the problem of computing the size of the concept lattice is *#P-complete*[5] (Buzmakov, 2015).

[5] A problem is P-Complete if it is in P and every problem in P can be reduced in polynomial time.

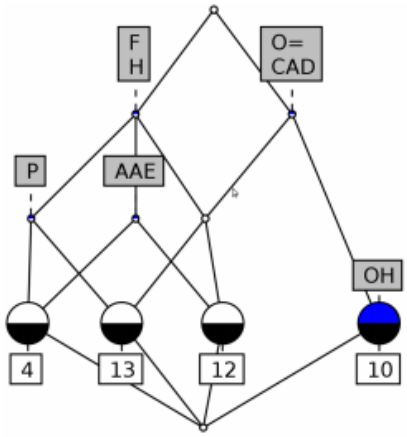

Figure 2.13 - The FCA-lattice for the context on Figure 2.12. Example taken from (Buzmakov, 2015)*, p. 28*

This network-based representation, as seen in Figure 2.13, demonstrate a potential use for richer visualizations and exploration of associations and patterns. In fact, in the Humanities fields there are already applications of FCA to Museum Collections (Cole, Fritjov, Ducrou, Eklund, & Wray, 2018).

FCA can, therefore, be used to think with the data as it allows the researchers to explore the feature space following a philosophical logic of human thought (Cole et al., 2018). Cole et al., (2018), starts the process by defining the scope of the phenomenon, the objects of study and their features by asking the following questions:

- *What is the scope of the work? The entire collection available or a subset based on specific categories or periods?*
- *What are the objects in study? Should we consider all the available metadata from all fields around an object or just a subset of them?*
- *What characteristics of the objects do we aim to study? And how do we define and extract those characteristics? What features of our objects interest us?.*

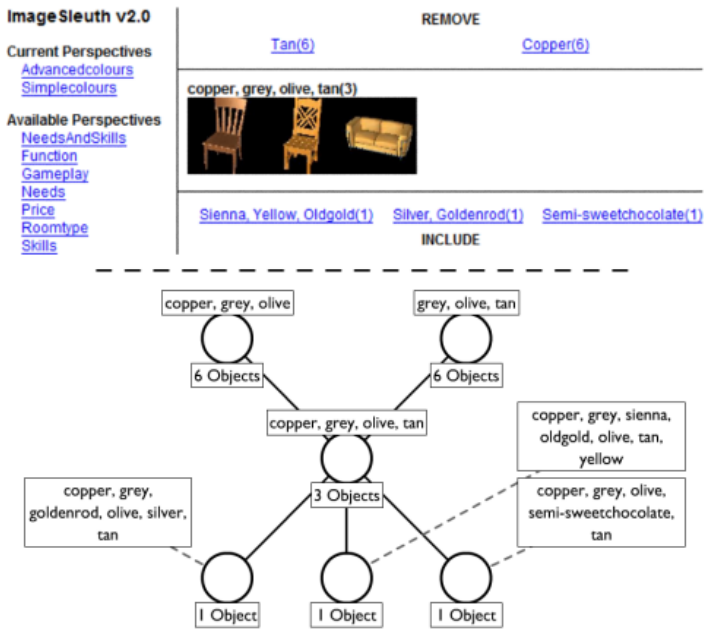

Figure 2.14 - Central pane showing the current concept: intent as thumb-nails. Example taken from (Cole et al., 2018)

Within Formal Concept Analysis, we define the scope of our objects and their features as attributes. This conceptual approach allows a semantic analysis of the data set which, consequently, leads to better information retrieval using visualization tools as seen in Figure 2.14. The obtained conceptual graph, more than allowing an exploratory visualization, allows the researchers to extract ontologies based on associative rules (as mentioned in section 2.4.1.3 Association Rule Mining) and which we can verify in Figure 2.15.

Formal concept, in natural language	#obj
paintings that depict the Illawarra	8
works that evoke identity issues and social critique	6
surreal works that depict animal imagery	6
vibrant and abstract paintings	11
intricate works that depict nature	6
vibrant works that evoke a sense of calm	6
works that depict heavy industry and the Illawarra	7

Figure 2.15 - Example of associative rules in natural language from "A Place of Art". Example taken from (Cole et al., 2018)

A similar approach could be taken for the Hebrew manuscripts, mapping such questions to our work to define the scope. Moreover, retrieving ontologies could leverage the semantical analysis made by the experts. As we will describe in the next sections, the use of this concept clustering technique was applied to the Hebrew manuscript in order to extract relevant ontologies and to visualize such hierarchies.

Chapter 3

Design and Development

In this chapter we will describe the design and development of the built artifacts and their use. The case study will be a subset of Sephardi (individual handwriting) Hebrew manuscripts, between year ≈ 900 and 1540. The rationale taken for these specific periods will be described in the next sections, where we describe our case study. Furthermore, this was the initial set of study that will support the development of our environment as well as to assess its effectiveness. This chapter is organized as follows:

- Section 3.1 introduces the technologies and programming languages used within this research;
- Section 3.2 briefly describes the scope of the collected data retrieved from Sfardata (Beit-Arié, 2017) and address initial concerns of the data;
- Section 3.3 presents the proposed framework, *CodicoDaViz*, along with its artifacts built to achieve the goals of this research following Big Data Lifecycle;
- Section 3.4, although still part of *CodicoDaViz*, presents the visualization artifacts as templates to be used by the experts and how this research modelled them;
- Section 3.5, addressing the high dimensionality and categorical nature of our data, this section explores the approaches taken to perform knowledge retrieving using both unsupervised learning and data mining techniques to enrich the visualizations.

3.1 Development Languages

The nature and scope of this thesis inserted within the Digital Humanities field led us to use tools from the Data Science field. No unified language could be found that would fit all the needs, therefore, several languages and tools were used. For data extraction and

HTML processing from Sfardata (Beit-Arié, 2012) we used C#.NET[6] framework given the professional experience in using this language. For the Big Data Analytics scope of this thesis we considered R[7] language and Python[8]. Although R is a reference by academics in terms of visualization, robustness and performance of some algorithms (available through maintained packages), our choice has fallen to Python which is rapidly increasing in terms of popularity and has a very active community in both academic and enterprise areas. Again, previous experience in Python language as well as handling Data Science packages led us to choose it alongside with Jupyter Notebooks[9] that would allow us to easily share the knowledge.

Visualization packages within Python provide us a quick and easy way to explore data but since the goal is to achieve a Visual Analytics tool, Tableau was the choice given its flexibility, easy integration and available documentation. As we will see in later sections, association maps and relational trees were built to explore content association. These were developed using JavaScript's d3.js[10] framework which is a widely used visual tool for data-driven documents providing understandable ways of exploring data and retrieving valuable insights.

For ETL purposes the BI suite provided by Pentaho[11] (Kettle) was used to load our data-mart (using MySQL[12]) which is stored in Azure[13].

3.2 Case Study

Our corpus includes codicological descriptions from two distinct periods in history, both from manuscripts in Sephardi handwriting. The initial dataset was retrieved from all dated manuscripts produced in the 15th century (from 1400 to 1500) which was a continuation of previous research (the PhD thesis of (Matos, 2017)) in late-15th century Sephardi manuscripts. This previous research was a descriptive analysis of the manuscripts within this period and no structure was given to the data, in fact, it was the driver for this work given the hard task of analyzing such units from different sources. The mentioned period produced a corpus with 580 unique manuscripts from Sfardata.

[6] https://www.microsoft.com/net
[7] https://www.r-project.org/
[8] https://www.python.org/
[9] http://jupyter.org/
[10] https://d3js.org/
[11] https://www.hitachivantara.com/go/pentaho.html
[12] https://www.mysql.com/
[13] https://azure.microsoft.com/

However, after some initial explorations and analysis, the data variation was considered insufficient for pattern discovery. To increase the research possibilities and scope of analysis, we extended the period to include codicological units produced from ≈ 10th century to late-14th century. Increasing the available corpus from 580 manuscripts to 846 (a contribution of 31% to our object of study). Additionally, manuscripts produced from early-16th century (from 1501 to 1540) were also included with a contribution of 160 codicological units.

For the development phase, the given periods and region resulted in $n = 1006$ unique manuscripts composed by 51 features. This results in a feature space of $n * 51$, being n the number of manuscripts.

3.2.1 Sfardata

As said earlier in section 1.2.4 Demonstration, our main source of data is originally from Sfardata (Beit-Arié, 2017) online database holding information of Hebrew manuscripts, which was afterwards enriched with catalogues, upon their consultation by our experts. Although we are interested in manuscripts as objects, in Sfardata (Beit-Arié, 2017), manuscripts are listed according to scribal hands and not as codicological units, thus enlarging the corpus. This meant an additional step to determine which hand is the most representative, which was far from straightforward.

Colophon Details	
Eras	*calendrical systems:* *creation:* *li-briat olam:* thousands,- month,-
Written for	a patron-
Material	*material:* *parchment:* distinguishable sides-- *scratching traces:* not visible,- *hair follicles:* visible,- corresponding sides in openings- hair side v. hair side flesh v. flesh,- *quire starts with:* hair-side,- *parchment thickness:* fine (0.11-0.15),-
Ink	*ink color:* dark brown,- brown,- light brown,-

Figure 3.16 - An example taken from Sfardata (Beit-Arié, 2017) *demonstrating its descriptive nature*

As shows, the metadata available in Sfardata (Beit-Arié, 2017) is far from being prepared, at its raw state, for application of computational methods of knowledge discovery, pattern analysis and visual analytics tools. The same illustrates the descriptive nature of the collected data which leads to uncertainty given the lack of a proper

vocabulary shared across experts who studied and registered these codicological units. Consequently, the corpus available in Sfardata (Beit-Arié, 2017) is incomplete in many instances, likely due to difficulties experienced during the consultation of the artifacts. Therefore, in several cases data was manually enriched with access to catalogues.

Since no public APIs were available to collect the metadata from the Sfardata (Beit-Arié, 2017), there was an additional process, described in the next sections, of document acquisition to parse the information available in this platform.

3.3 *CodicoDaViz*

CodicoDaViz acts as a toolset for our proposed framework including the six artifacts built to provide the proper tools for experts within the Humanities field to do exploratory analysis and have quick insights of their studies.

Our development methodology uses an adaptation of the Big Data lifecycle methodology proposed by (Erl et al., 2016). As shown in Figure 3.17, the proposed method was adapted to include data visualization in the life cycle. This allowed us to explore and explain the data, but to also further clean it and validate it through an iterative process. Consequently, this led to the re-definition of the scope of our study. As demonstrated throughout this work, this step helped to increase the ability to spot erroneous or missing values, and to evaluate confidence levels from what the data showed.

Furthermore, much of the artifacts built within our proposed framework arose from identified needs on each step of the work method.

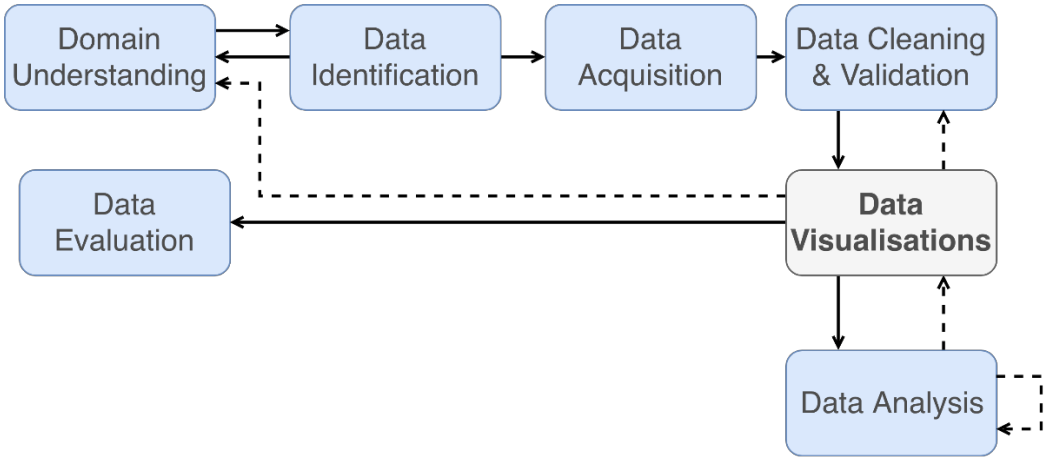

Figure 3.17 - Adapted lifecycle to use visualization as a central data inspector step

Domain understanding refers to the scope of the work, the research questions and the relationships inherent to manuscripts. Having a domain-specific understanding is essential to understand the data, its composition, and what to expect, in order to perform a clear analysis within it. Next, it is necessary to identify what data is available and its sources, which is followed by a subsequent step of data acquisition. This refers to the extraction and loading process of gathering data, from one or multiple sources.

Given the unstructured nature of our data, as further discussed, data cleaning and validation was a crucial stage in this method. It is where we applied transformations, standardized categorical values by applying the developed common vocabulary artifact (which was a result of several iterations through the available data alongside with expert's input domain specific decision as described later in this document in section 3.3.3), and obtained the first insights via data visualization on the quality of the dataset. Iteratively, including several instances of data visualization, it was possible to visualize each state of the dataset, and spot potential domain-specific concerns or data-specific concerns (such as missing values). Once the method design for data cleaning was concluded, and the data cleaned, the step of Data Representation refers to the final model and how it will be stored for further use.

Finally, the last step concerned the analysis of the data via visualization, which provided an iterative process and a gradual construction of narratives that convey the results and conclusions obtained. The next sections are devoted to these stages, and particularly to how the iterative nature of this method allowed us to determine a final corpus to be explored in further studies.

The overall architecture and connection of each artifact can be seen in the following Figure 3.18:

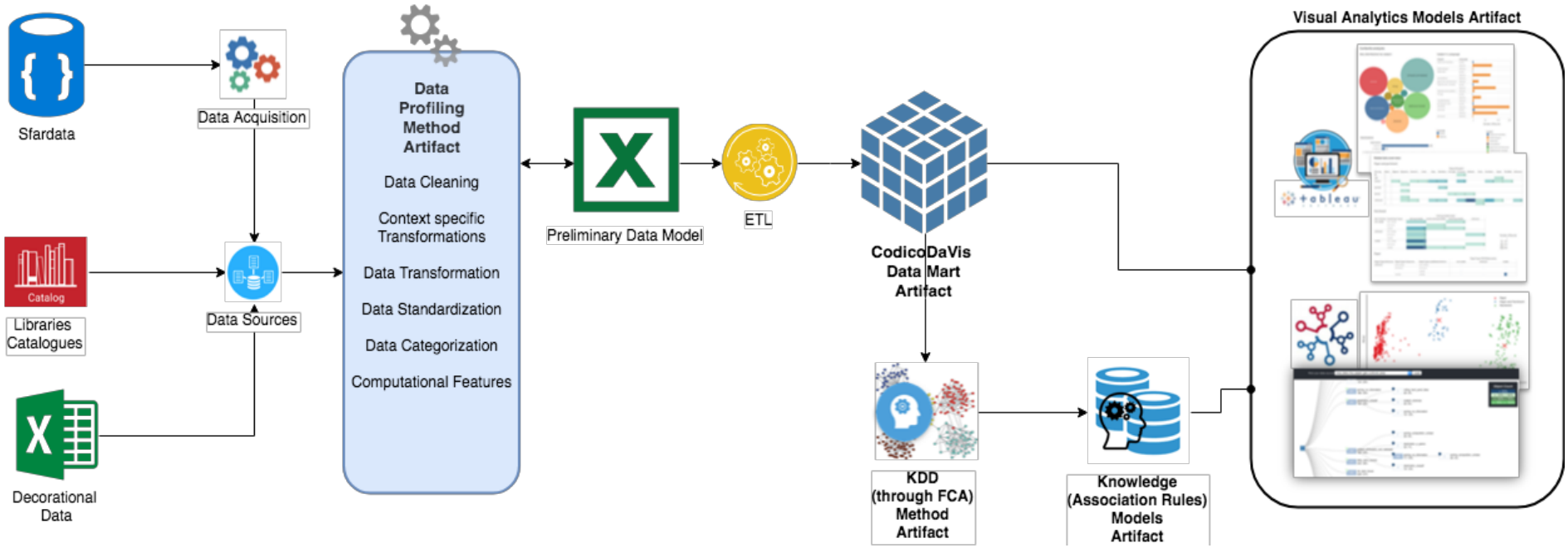

Figure 3.18 - CodicoDaViz framework architecture.

3.3.1 Domain Understanding

The interaction of medieval Jewish communities settled in Europe and in the Mediterranean basin with other cultures often resulted in the assimilation of local practices and influences from other book cultures, in terms of material and artistic aspects. This is particularly relevant for books produced in the western Mediterranean, where three book cultures, Latin, Arabic and Hebrew, coexisted. If by the fifteenth century many codicological practices had already crystalized, the introduction of the printing press posed a new challenge, that may also have shaped manuscript production.

The concept map in Figure 3.19 below summarizes the main attributes and respective relations, in order to explore and draw conclusions. This conceptual map shows how a feature contributes to or inherits from other features, having the manuscript as a central entity. The understanding of such domain specific field was only possible with the brainstorming and conceptualization of such terms. In fact, the illustrated concept map was a result of some iterations and sessions to manually achieve such scheme. With this we hope to provide a better understanding of the story that a manuscript tells us, for example, what materials were used for its codicological data and how it reflects the given date of a manuscript. Figure 3.19 illustrates the available metadata of a manuscript collected by experts along several years. Conceptually, the given map provides a glance of the high dimensional space in study in terms of available features and their dependencies.

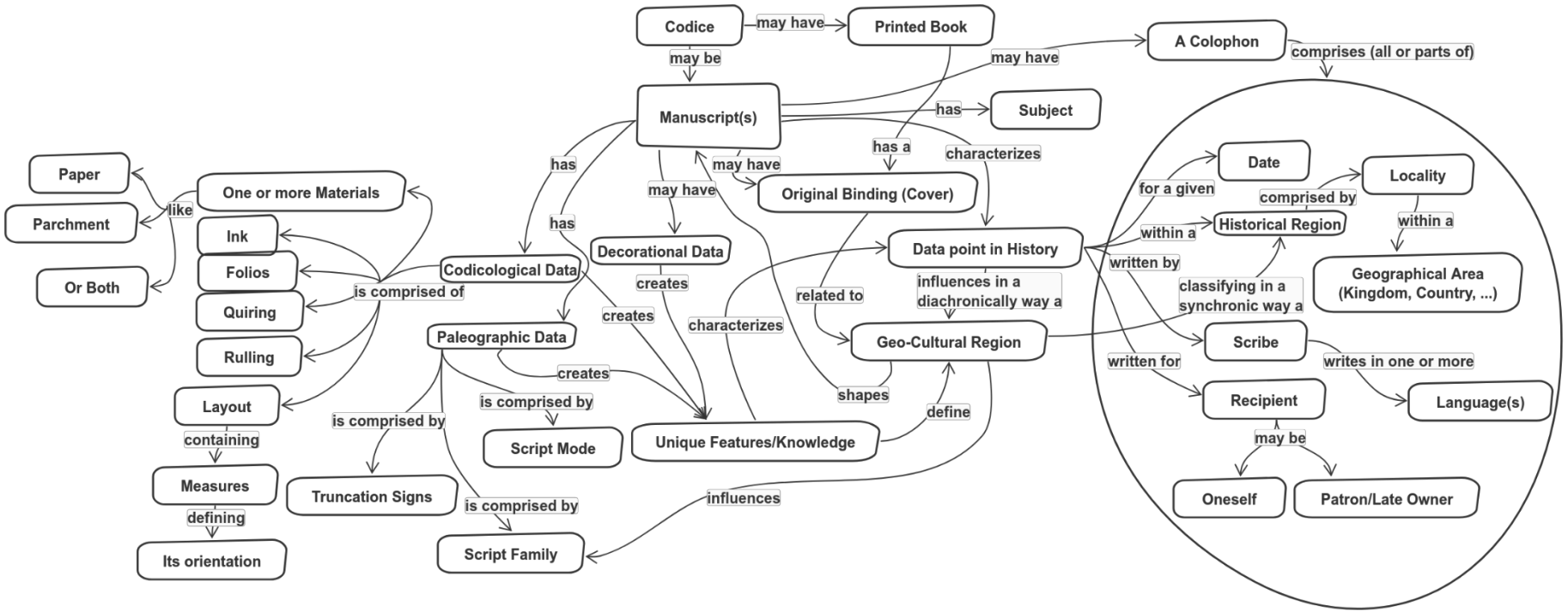

Figure 3.19 - Concept map covering the relations that describes a manuscript.

3.3.2 Data Acquisition

The main source of information used in the scope of this thesis, Sfardata (Beit-Arié, 2017), does not provide a proper way of sharing its data. Therefore, an additional step, specific for this database, was needed. A parsing process was developed to extract the information, applying minimal transformations and process to the data, as seen below in Figure 3.20.

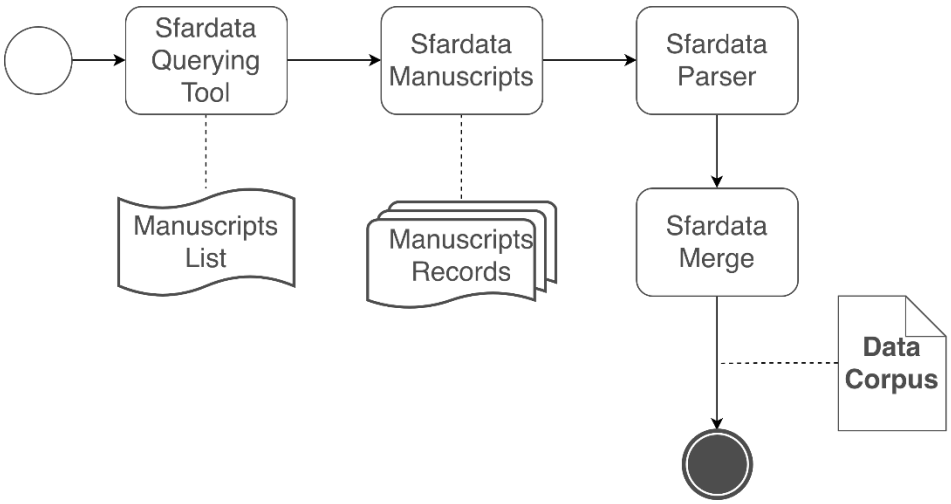

Figure 3.20 - Sfardata specific data acquisition process

Therefore, querying Sfardata (Beit-Arié, 2017) for all the Sephardi manuscripts between the scoped periods (from ≈ 10th century to early-16th century), we extracted all information concerning the historical, codicological and palaeographical information for each manuscript. Such metadata retrieved can be seen in the earlier presented concept map in Figure 3.19 in page 35. The concept map shows the considered features. This information was subsequently merged into a CSV file containing all information. The first step of the acquisition process is to request the HTML with the resulting manuscripts for a given query. With those, for each manuscript we requested the data for each feature and stored it in a text file. In this stage the retrieved documents were parsed, and the desired features processed and cleaned. Finally, we merged that information in one single CSV file.

3.3.3 Data Identification (Controlled Vocabulary artifact)

This step was the driver taken to build the identified construct artifact for our work. As shown below we identified the possible values for each feature by mind mapping our

acquired data. Identifying the possible range of values for each feature allowed us to define a controlled vocabulary for the feature set to be applied in the whole corpus.

In order to understand the scope and to complement our domain presented in Figure 3.19, it is necessary to describe each attribute in use, and its meaning for our data source. This dataset, in its raw state (from Sfardata (Beit-Arié, 2017) database), has 51 features distributed by codicological, historical and palaeographical categories, which create specific relationships within a manuscript. This metadata was created by the team behind Sfardata since the 1960s, upon the consultation of each artifact.

Each manuscript description also includes historical details, such as the identification of the scribe, area of production (often based on script), and subject. Additionally, we have computed some fields such as orientation and format, and, in multi-hand copies, as previously mentioned, established a series of rules to identify the most characterizing hand.

The defined feature set was grouped in three major sections, all concerning the overall description of the artifacts. These are: history, codicology and a palaeography. As seen in the concept map (Figure 3.19), these three sections are intrinsically dependent on each other. Such a holistic approach is based on Beit-Arié, who states that "*identifying the provenance of a manuscript cannot rely on the script type alone, but on the correlation between it and the codicological profile, which reflects the production zone; and, if the script type does not match the codicological profile, it can then testify as to the copyist's origin*" (Beit-Arié, 2012). That is, single feature analysis may be insufficient to draw conclusions, but the comparison of different types of data is a lot more promising. As put by Beit-Arié, "*Similar practices in different circumstances would prove that they were not conditioned by social, economic, or cultural context, but were universally inherent in making a codex. Similar practices in similar circumstances would prove that they are conditioned by those circumstances, as in the case of the introduction of the plummet. Different practices may be the consequence of factors other than technological, such as aesthetic conventions, economic or scholarly needs*" (Beit-Arié, 2012).

3.3.3.1 Historical Data

This category answers to questions regarding the who, when, where, what and how of each artifact. These attributes, as shown in Figure 3.21, are indicated by the original scribe, or inferred from additional information, including secondary sources. Data obtained include the scribe's name, number of hands, probable geographic location,

subject and language, as well as destination. This can be used to place a given book within a historical timeline and consider how a certain period (including historical events) and region have influenced the physical aspects of the artifact. These independent codicological entities are defined by geographical and cultural boundaries, not necessarily political ones.

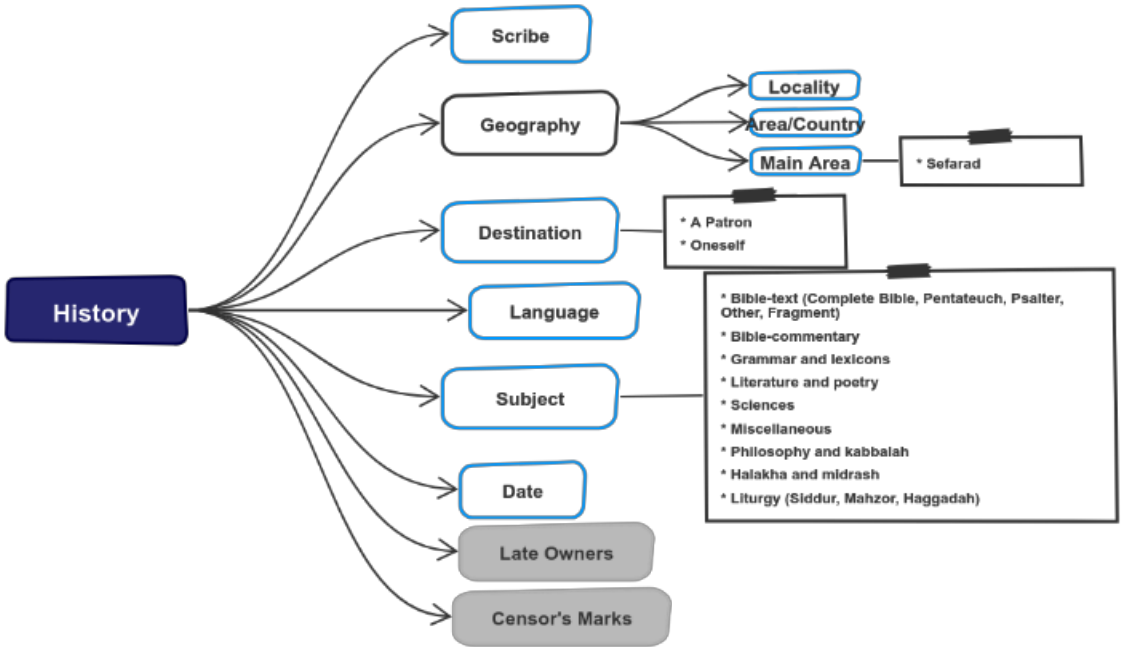

Figure 3.21 - Part 1 of 3 of mind map with the Historical attributes

We shall formulate their definitions in geo-cultural terms which will be used to designate chief production zones, while the deriving adjectives will be used to designate the script type and codicological type. This typological division is synchronic and reflects the appearance of the types after their consolidation and diffusion; it is thus suitable for describing periods that have yielded an abundance of surviving witnesses. Unavoidably, because of its complexity and stage-by-stage development, the dynamic ‘Sefardic’ type must be presented individually and diachronically, and this is also the case in respect to the ‘Italian’ codicological type.

3.3.3.2 Codicological Data

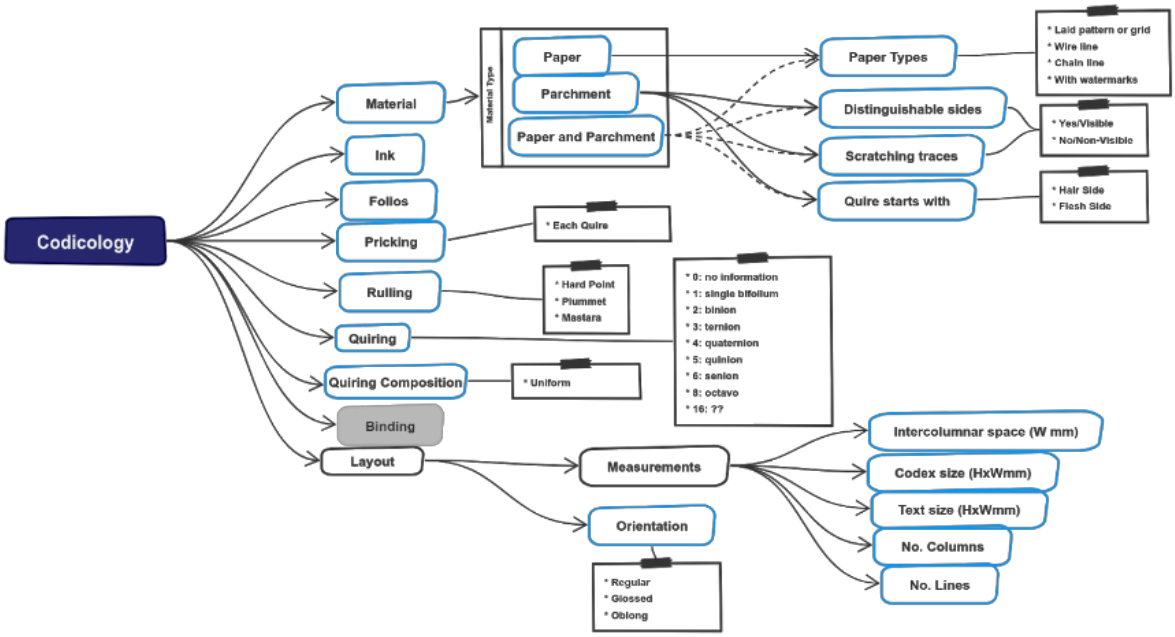

Figure 3.22 - Part 2 of 3 of mind map with the Codicological attributes

This category comprises the attributes regarding the physical composition of the artifact. More specifically, the type of writing material and its quality, ink, number of folios, quiring system, type of ruling, page layout (number of columns and lines), and format. Partially quantifiable, some features, shown in Figure 3.22, such as ink are entirely descriptive. Other features such as pricking were not considered due to the lack of substantial information. Regarding format, it was necessary to divide the corpus in sub-groups according to its size and orientation. Through data identification we know, beforehand, that for Material there are features intrinsically dependent on the type used to produce the manuscript. Finally, here was also included the information on the presence of decoration.

3.3.3.3 Palaeography Description

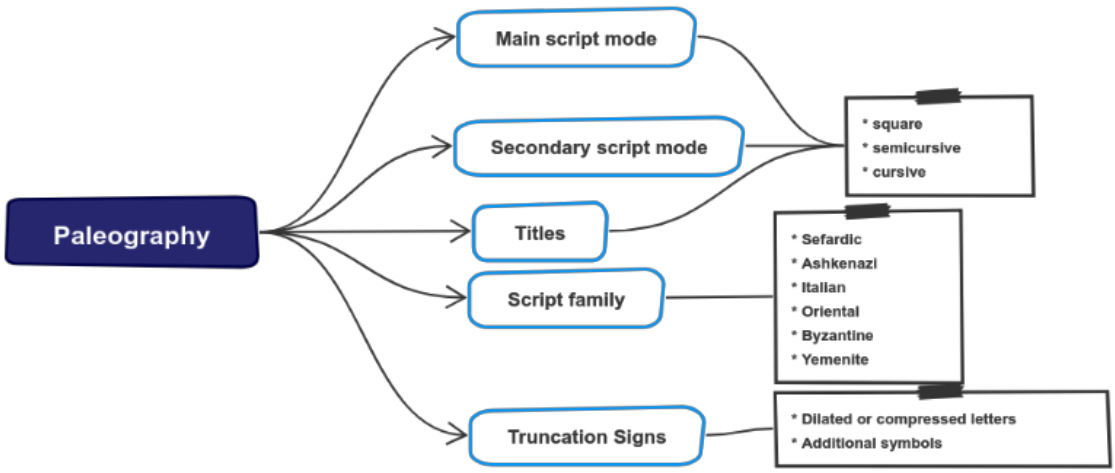

Figure 3.23 - Part 3 of 3 of mind map with the Palaeography attributes

Although palaeography is by itself a field of study, the data source includes a general description of the family and mode of writing, as Figure 3.23 shows. Often this is the main criteria behind the association of a manuscript to a specific geo-cultural region. For instance, at the lack of more information, a manuscript in Sefardi script will be ascribed to Sefarad, even though other elements can eventually further determine a specific region (for instance, plummet ruling and Sefardi writing will probably indicate an Italian origin for the manuscript). In this category are included features such as script mode, titles, and script family.

3.3.4 Data Profiling, Preparation and Cleaning Method

Given the nature of the features of manuscripts, the data held by Sfardata (Beit-Arié, 2017) was not easily extracted nor ready for computational analysis. The high level of uncertainty, the lack of structure and the descriptive nature of the features can be seen in the sample from Sfardata, displayed in, previously presented, .

Consequently, a more in-depth data exploration was limited because the available raw data could not be statistically analyzed. However, it provided good insights on the quality of the data, its structure, and the challenges ahead, namely the unstructured nature of the data provided by our source, as discussed in this section.

Although partially automated, the process of data acquisition required an extensive manual intervention due to the inherent lack of consistency and structure of the corpus.

This manual cleaning was only achieved due to the team's expertise, and data transformation rules have been carefully annotated. Annotating each step of data transformation contributed for the proposed method to be applied to subsequent data sources to be integrated in our corpus.

The first step was to profile the data obtained, and since acquisition was based on HTML markup, each feature value required to be fixed by hand, as well as the conversion of similar values (see). Transformations such as "*some,- decorated,-*" within the illumination feature were transformed into "*Partly Decorated*", for instance. So, for decoration metadata the following rules were applied:

- Unknown ⇒ the field is empty;
- None ⇒ the field explicitly states "*none*";
- Partially Decorated ⇒ when we have "*some,- decorated,-*";
- Partially Decorated and Illuminated ⇒ for values stating only "*some,-*";
- Decorated ⇒ when we only state "*decorated,-*";
- Illuminated ⇒ when we only state "*illuminated,-*";
- Decorated and Illuminated ⇒ "*decorated,- illuminated,-*"

Similar approach was taken to the remaining feature set to achieve a categorical set of metadata with a controlled vocabulary within our corpus. After that, an analysis on the missing values was performed, including how to semantically distinguish an unknown value from a missing value (blank value). For instance, in the manuscripts' watermarks, should a blank value have the same meaning as "not visible"? Therefore, the rule applied was to mark these entries as unknown values rather than perform some value inference. Figure 3.24 displays the missing values problem within our corpus, in which the more blank a plot bar is the more missing values exist. It confirmed what we already knew from the data identification step, that the high level of missing values for features such as "Hair Follicles" or "Scratching Traces" are justified by its dependency on the type of material used.

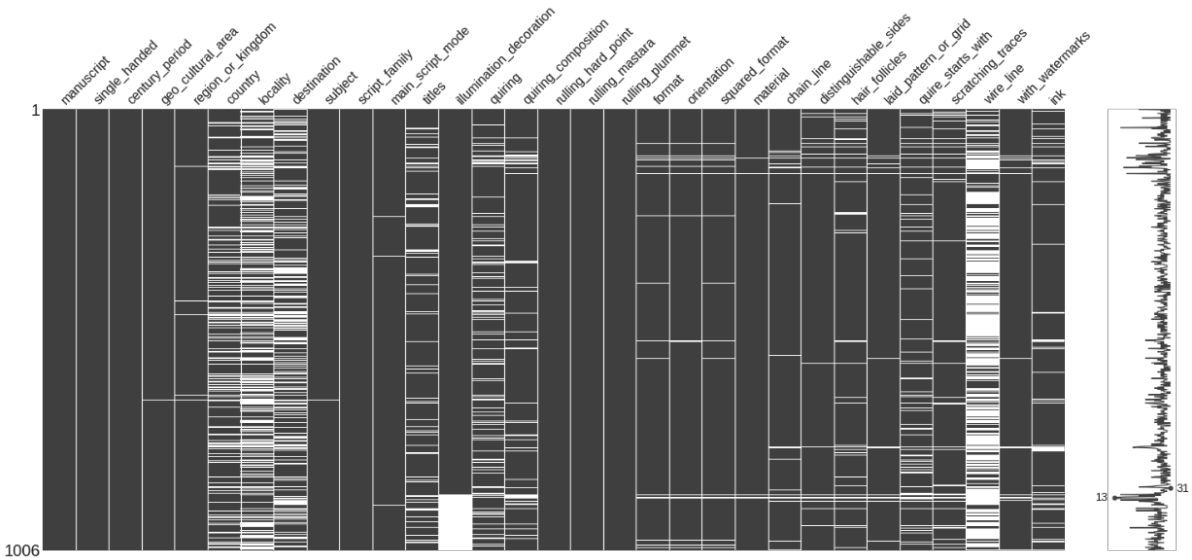

Figure 3.24 - Missing values plot from several attributes.

Therefore, one of the steps was to introduce the category of "*n/a*" for features that were dependent on values within our data and shouldn't be marked as unknown but as "non-applicable".

Using visualization at this stage allowed us to have faster insights on the data and its incoherencies. In fact, after analyzing the geographical data, as seen in Figure 3.25, it was possible to detect cases where the manuscript was geographically misplaced, stating cities that did not belong to the collected region (e.g., Zaragoza in Sicily instead of Spain).

Therefore, the second step consisted in reviewing the entire corpus and analyze its geographical data to fix additional mistaken records. Some manuscripts were ascribed to "*Sefarad*", others had the localities in different notations (for instance, using the Arabic word: *Ashbilia* for Seville, *Ashbona* to Lisbon, and so on). However, instead of changing the original information, new categories were created and annotated with the correct values. With this in place, documentation of each case allowed us to perform an overall analysis, as well as comparisons and/or profiling on data quality between original and processed values.

The next step was dedicated to the identification of manuscripts with more than one scribal hand. This led us to immediately spot data differences within the same manuscript.

Consequently, the corpus was divided into two subsets: unique manuscripts as a whole codicological unit, and another with all the multi-hands in each manuscript.

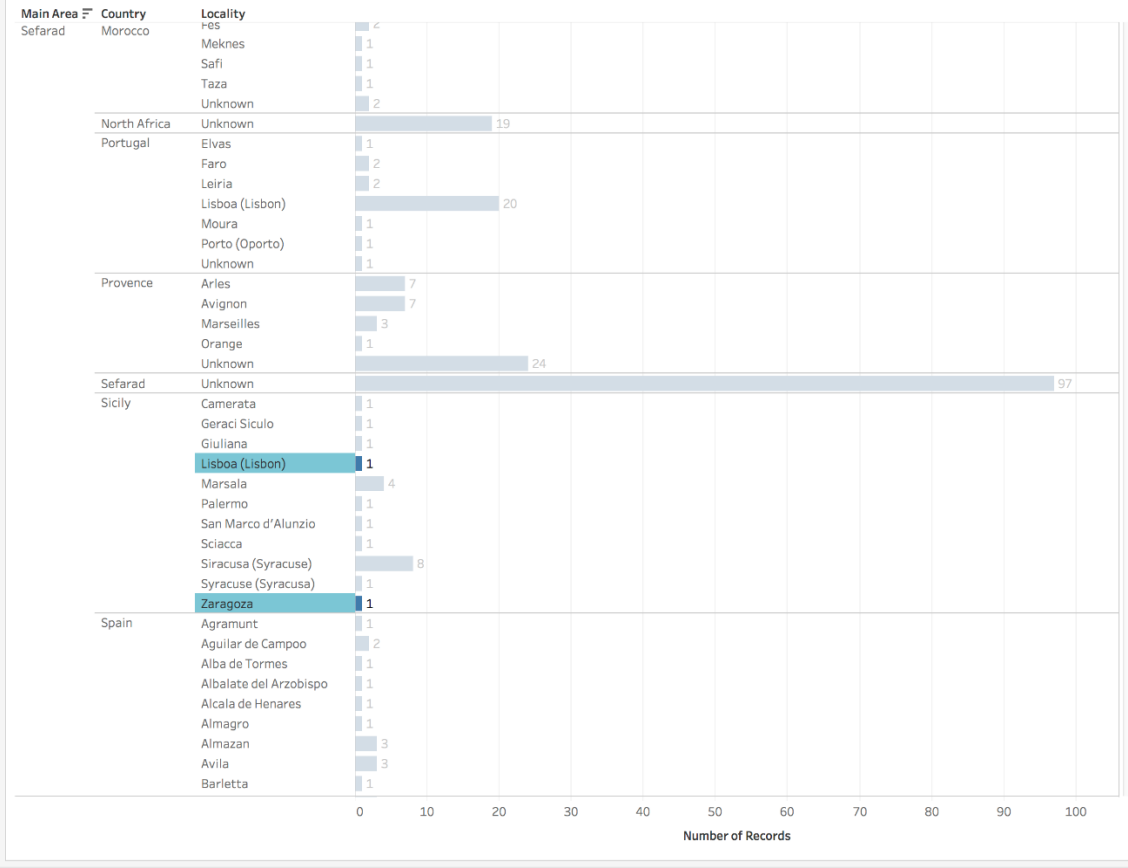

Figure 3.25 - Data visualization as a key tool to spot wrong geographies.

Again, documenting each case and preserving the originals allowed us to perform further comparisons and enrich the domain understanding step with these evidences. The rule to determine which hand better characterizes a manuscript was built on the following conditions for different scenarios:

Rule 1. If the data source indicates that a scribe is "participant" (wrote less than 10%), it is singled out. In the case of two hands (the most common case in multi-hand copies), this indicates that the other is the main hand;

Rule 2. If the colophon is written by a specific scribe and he does not refer to another scribe;

Rule 3. In the absence of the two previous conditions, the number of folios will determine which is the most significant hand.

Some manuscripts required additional enrichment from external catalogues because the amount of missing values prevented the application of the above rule. Particularly significant was the fact that some of the selected hands were not initially included in the query made on Sfardata (Beit-Arié, 2017), due to using a different script family. This means that Sefardi script was in many instances a secondary script, and the codicological

features that define the manuscript are found in the data concerning non-Sefardi hands. An evidence appeared when handling these multi handed manuscripts regarding the metadata within each hand, that sometimes was different. A decision was made to consider the metadata set for the determined main hand according to the previous rules.

To enrich the corpus with meaningful features capable of being analyzed and visualized, some attributes were computed. Measurement attributes, which are discrete but not categorical, do not provide useful information when visualized. Therefore, orientation, and format were added and computed based on codex size. Considering Codex Height as Ch and Codex Width as Cw and Codex Proportion, P the following formulas were applied to obtain these calculated attributes:

$$O(x) = \begin{cases} Oblong, & Ch(x) < Cw(x) \\ Regular, & Ch(x) \geq Cw(x) \\ Unknown, & Ch(x) = Cw(x) \end{cases} \quad (1)$$

$$F(x) = \begin{cases} Pocket, & Ch(x) \leq 100 \\ Small, & 100 < Ch(x) \leq 200 \\ Medium, & 200 < Ch(x) \leq 300 \\ Large, & 300 < Ch(x) \leq 400 \\ Oversized, & Ch(x) > 400 \\ Unknown, & Ch(x) = 0 \end{cases} \quad (2)$$

$$P(x) = \begin{cases} 100 - ((\frac{\left(Ch(x) - \left|\frac{Ch(x)-Cw(x)}{2}\right|\right) \times \left(Cw(x) - \left|\frac{Ch(x)-Cw(x)}{2}\right|\right)}{Ch(x) \times Cw(x)}) \times 100), & Ch(x) \neq 0 \ \vee Cw \neq 0 \\ 0, & Ch(x) = 0 \ \wedge Cw = 0 \end{cases} \quad (3)$$

$$SF(x) = \begin{cases} Yes, & P(x) \leq 10 \\ No, & P(x) > 10 \\ Unknown, & P(x) = 10 \end{cases} \quad (4)$$

Additional context specific rules were applied to the transformation process that are specific to codicological studies of Hebrew manuscripts. For instance, the language in which the manuscript was written, only 3.79% of the artifacts had this attribute filled, but from the context it was possible to assume that when not specified it should be assumed as Hebrew. Another example is the visualization presented in Figure 3.26, showing “a woman” as destination. Based on the knowledge of the expert on Hebrew book culture, the initial inclusion of a female destination in the corpus was identified as an

inconsistency, since this was not expected in the context of the Iberian Peninsula or in Sephardi manuscripts in general but would be more likely to be the case for Italy. This inconsistency challenged us to verify the origin of the manuscript. In Sfardata (Beit-Arié, 2017) we could see that it was a multi- hand manuscript, copied by a groom for his bride, and the colophon further confirmed that the manuscript was copied in Italy.

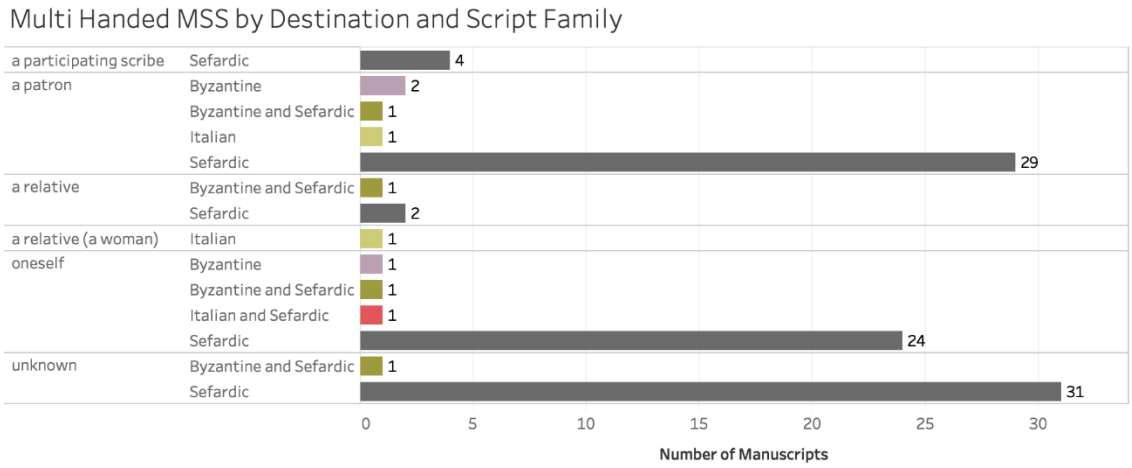

Figure 3.26 - Visualization used to spot unexpected information requiring another iteration.

Having a single entry for such a destination could have been missed otherwise, however the data visualization analysis triggered another iteration to review the multi hands step to confirm this evidence.

3.3.5 Data Aggregation and Representation

Since our data cleaning and preparation was fundamental, the acquired data was stored in an Excel format for further treatment. As said before, in Sfardata (Beit-Arié, 2017) the results displayed are listed according to scribal hands and not as codicological units. Therefore, the acquired data had duplicated units depending on the number of contributing scribes.

Based on the method of data cleaning and preparation, one of the decisions was to aggregate these scribal hands by codicological unit following the decision rule presented in section 3.3.4. Hence, the acquired manuscripts needed a preliminary model of work to apply the cleaning process as we will present next.

3.3.5.1 Preliminary Data Model

This model was the basis of the manual preparation taken to achieve a final corpus based in our vocabulary construct. From the acquired data exported to Excel format we

divided the dataset in three worksheets: one containing all the unique codicological units, one with single handed units and other with the remaining multi handed manuscripts. Moreover, the decision made to use this Excel format was due to the manual step needed and a collaborative tool was required based on expert input for context specific rules.

This way we could apply the aggregation rules for the multi handed manuscripts in a much more contained way where we decided which hand was to be considered as the main hand for codicological unit representation.

	E	F	G	H	I	J	K	L
1	ScribeNo	Date	LocalityDBSfardata	AreaDBSfardata	MainAreaDBSfardata	GeoCulturalArea	GeoCulturalArea-Notes	RegionOrKingdom
196	1	1456	Vieste	Southern Italy	Italy	Italy		Southern Italy
197	1	1413	Unknown	Sefarad/Italy	Unknown	Sefarad	No codicological reason to asso	Unknown
198	1	1470	Mantova (Mantua)	Northern Italy	Italy	Italy		Northern Italy
199	1	1490	Unknown	Spain	Sefarad	Sefarad		Spain
200	1	1468	Sevilla (Seville)	Spain	Sefarad	Sefarad		Spain
201	2	1446	Unknown	Spain	Sefarad	Sefarad		Spain
202	1	1468	Unknown	Sefarad	Sefarad	Sefarad		Sefarad
203	1	1402	Unknown	Provence	Sefarad	Sefarad		Provence
204	1	1462	Salamanca	Spain	Sefarad	Sefarad		Spain
205	1	1481	Unknown	Sefarad	Sefarad	Sefarad	Probably from Portugal, needs (	Sefarad
206	1	1414	Constantine (Citra)	Algeria	Sefarad	Sefarad		Algeria
207	1	1490	Napoli (Naples)	Southern Italy	Italy	Italy		Southern Italy
208	1	1412	Unknown	Sefarad	Sefarad	Sefarad		Sefarad
209	1	1427	Unknown	Sefarad	Sefarad	Sefarad		Sefarad
210	1	1493	Iraklion (Candia)	Crete	Byzantium	Byzantium		Crete
211	1	1402	Padova	Northern Italy	Italy	Italy		Northern Italy
212	1	1424	Unknown	Provence	Sefarad	Sefarad		Provence
213	2	1406	Marseilles	Provence	Sefarad	Sefarad		Provence

MSS All | MSS SingleHanded | MSS MultiHanded

Figure 3.27 - Data sample taken from Excel to show the model structure

To document every step taken in the cleaning process additional columns were created as marked as annotations where each decision taken was annotated, as seen in above Figure 3.27, for geo-cultural area feature for example.

3.3.5.2 ETL

On the previously presented model, once the data cleaning method was achieved and all of the data within our corpus was compliant with the built vocabulary. This step was needed to feed our data mart artifact. The process of extracting data from multiple sources, transforming it based on our vocabulary (i.e. business rules) and loading it into a data warehouse is called the ETL (Extract, Transform and Load) process.

Although we have currently a single source of input the idea behind this step is to make it easily adjusted to new inputs in the future. However, the manual cleaning process is currently a blocker for a total automated extraction and feeding process.

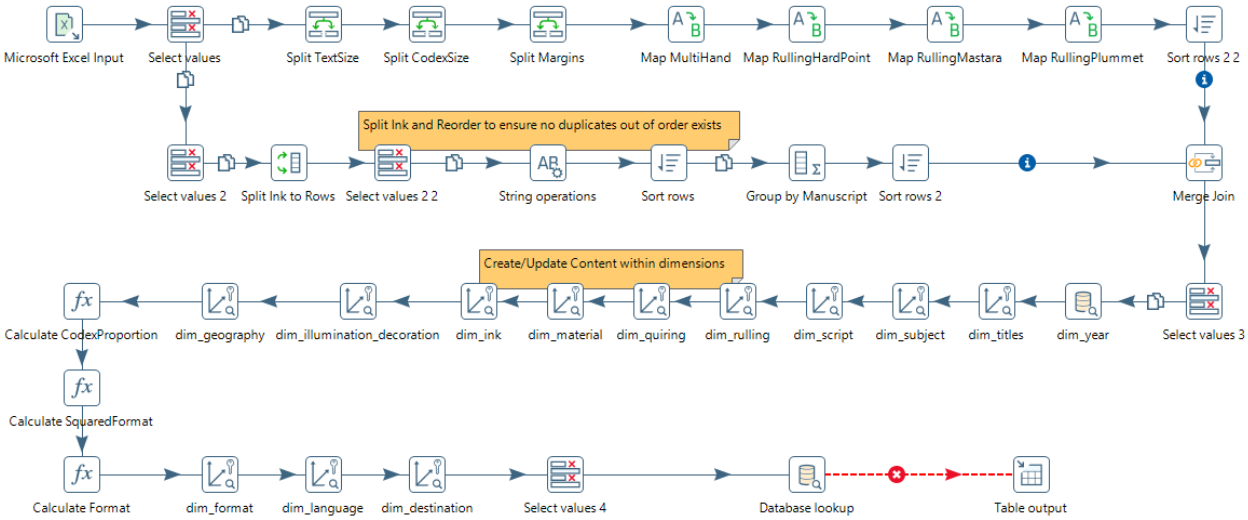

Figure 3.28 - ETL workflow to feed data mart

As seen in above Figure 3.28, the initial step loads the preliminary data model in Excel format where we select the "MSS All" worksheet which contains all the processed codicological units. These manuscripts are streamed to the various steps of the transformation workflow where the computed features are calculated (orientation, format, etc.). To avoid duplicated manuscripts in our data mart, a check by "manuscript" (considered the key identifier) is performed to allow us to insert new data based on delta differences by this key. The same is applied for the remaining dimensions to be presented in the next section.

3.3.5.3 Multidimensional Model

A new model artifact – a multidimensional model – was designed to store the corpus based on the controlled vocabulary construct. This model artifact enables an exploratory analysis of the corpus applying visual analytics techniques.

The proposed model follows the commonly used conceptual schema from Data Warehousing field defined as star schema. The decision to use a Data Warehousing approach is reasoned by the context of Big Data approach of the present dissertation.

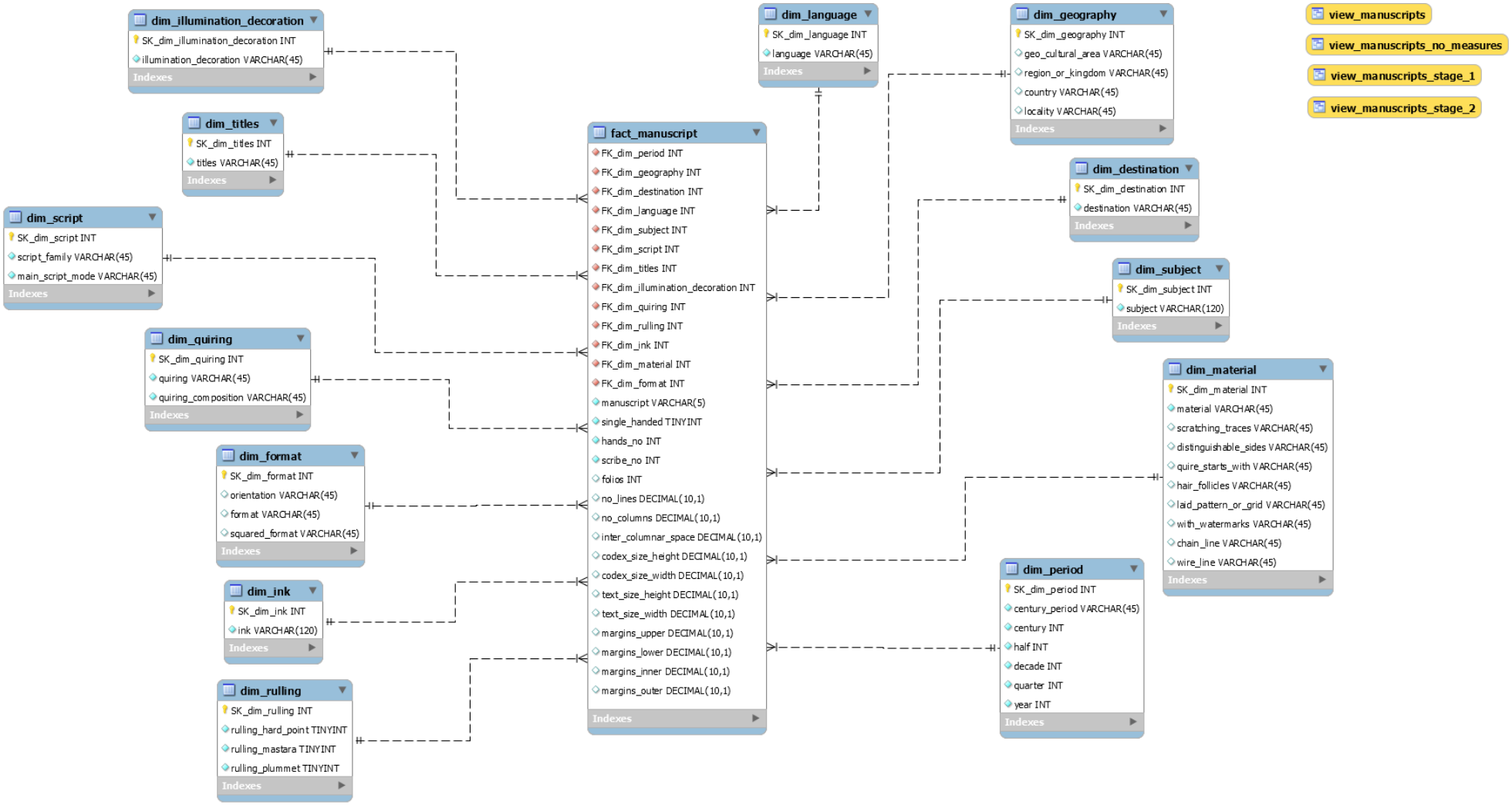

Figure 3.29 - CodicoDaViz data mart artifact.

We can identify our operational schemas being at this moment the preliminary model presented in the previous section integrated via an ETL process. Data Warehouses are at the core of decision support systems. They store integrated information extracted from various and heterogeneous data sources, making it available in a multidimensional form for analyses aimed at improving the users' knowledge of their business (Golfarelli & Rizzi, 2018).

This data mart schema has in its very essence a fact table fed by several, dimensions, as shown in Figure 3.29. The star schema separates our process data into facts and dimensions. Facts hold the measurable, quantitative data about the codicological units, while dimensions are descriptive attributes that give context to facts. Our fact table "fact_manuscript" has a granularity of a single codicological unit which was previously aggregated from the list of potential scribal hands.

The dimensions tables have hierarchies, which have been identified from previous explorations that we believe to provide the expert what they expect to explore in terms of lookups. The "period" dimension has a time hierarchy allowing the slicing by century, half a century, a quarter, a decade and year. The same applies to "geography" dimension which by the expert's feedback was decided to consider the cultural area as the higher-level entry of a manuscript followed by historical region or kingdom, the present country for modern geography and the lowest geographical point, locality. "Material" dimension on the other hand, represents the possible clusters of depending features based on the material type as stated before.

The design of this artifact followed the concept map presented in Figure 3.19 where we define our grammar and vocabulary construct in terms of relations that involve our fact, a manuscript.

3.4 Visual Analytics

For initial exploratory analysis of data, static plotting resources, such as bar plots, scatter plots and/or parallel-coordinates, play an important role to get in touch with the data in analysis. However, reasoning over these methods rapidly showed how difficult it becomes when the feature space is too big, and the data is not by nature nominal.

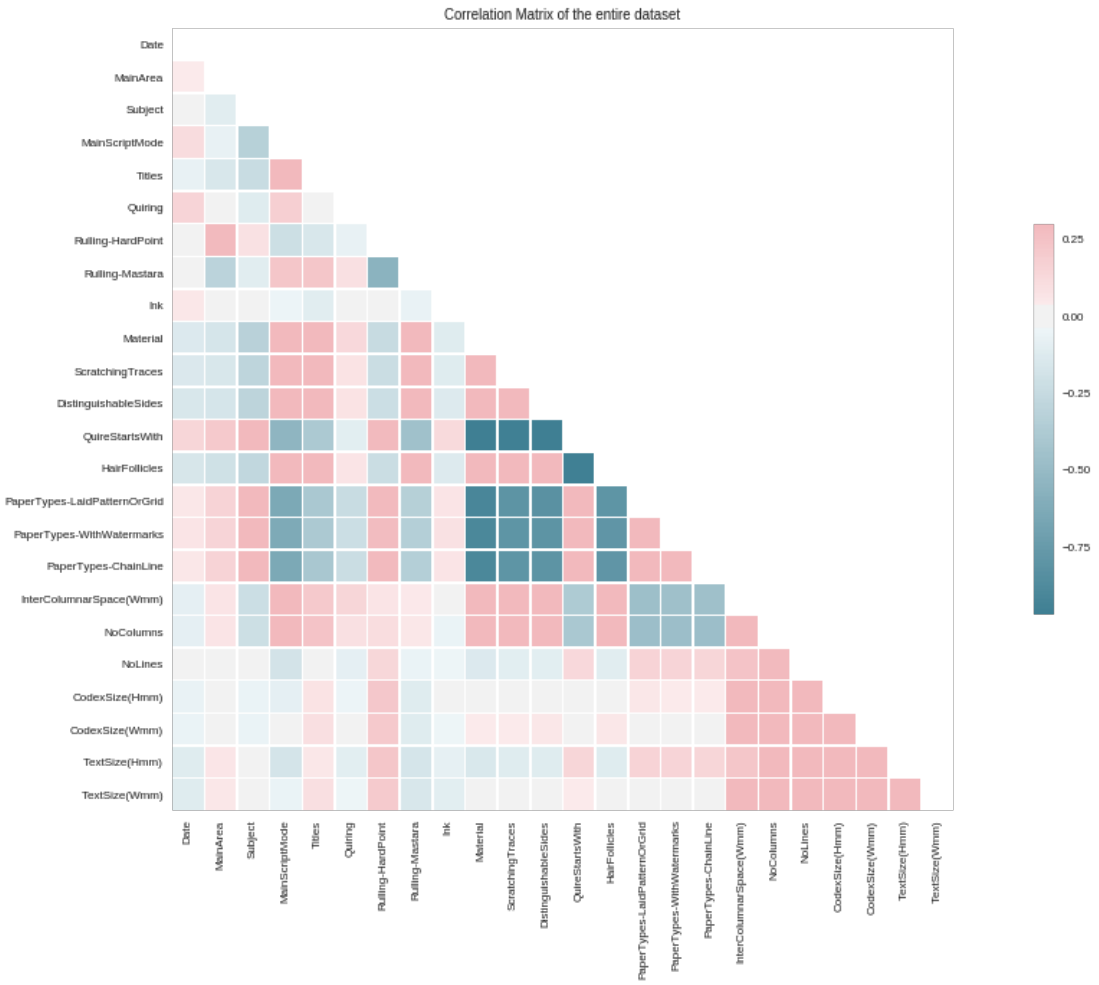

Figure 3.30 - Correlation Heatmap for feature space

Handling categorical data does not allow us to apply the most common methods of quantitative analysis as well as statistical techniques. As seen in Figure 3.30 the outcome does not give the expert any meaningful information or even correct one, since these techniques are based on distance metrics and we cannot assume that, for example, a "paper" produced manuscript labeled as one is lower than a "parchment" manuscript labeled as two, because for these methods two is naturally higher than one and therefore "more important".

Another problem faced in initial exploration was the number of features available in our corpus and how to represent it in a two-dimensional space, without losing the ability to extract knowledge. Moreover, we were not being able to answer the question "Who or what defines the 'relevance of information' for a given task?" (Keim et al., 2008). Meaning that we do not know from the expert, who is the decision-maker, which visualizations should be available without biasing the outcome. Statistical analysis and other quantitative approaches have long been part of methodologies for historical,

linguistic and other forms of inquiry, of which one of the best known is Franco Moretti's concepts of 'distant' and, by extension, 'close reading' (Jänicke & Wrisley, 2012). Yet current methods and tools for data visualization also open new research questions and unprecedented amounts of data (Kaplan, 2015).

Therefore, the visualization of these processes should provide the means of communicating about them, instead of being left with the results. Using Visual Analytics will foster the constructive evaluation, correction and rapid improvement of our processes and models and the improvement of our knowledge and our decisions (Keim et al., 2008). Ultimately, the goal is to provide to the experts an artifact that is interactive and capable of meet the expert needs for data exploration and reasoning about. Data visualization plays an important role in data science projects. It can be used to increase the understanding on data, to highlight properties that were not anticipated, to identify problems that need to be corrected, or to facilitate research hypothesis formulations (Ware, 2004). In contrast, only in recent years has the potential of data visualization been fully acknowledged in the Humanities.

Dashboards have been designed to enable an interactive and ad-hoc exploration of data. This approach enables both exploratory and explanatory analyses. The purpose of an exploratory analysis is, first, to understand the data and identify key aspects that can be communicated. As Knaflic, (2015) puts it, "it's like hunting for pearls in oysters" (Knaflic, 2015). As such, hypotheses must be tested, analyzed and visual displays explored in order to achieve an effective visualization. An explanatory analysis, conversely, places its emphasis on the message that needs to be conveyed. In other words, focus must be on the "pearls" rather than the (opened) "oysters". Despite the importance of both, in this thesis the focus is on the exploratory analysis of the dataset. As such, we have applied a visual analytics perspective to analyze the data. We defined dimensions and metrics of analysis, as well as dimension hierarchies to enhance the data exploration capabilities. The design of dashboards, using Tableau Desktop[22], was the chosen method to provide an interactive and intuitive data exploration platform for Humanities experts. Additionally, Tableau Stories[23] can potentially be used for an explanatory analysis.

The result was a template considered as a model artifact that could be enhanced and apply to further data sources. This Visual Analytics template model was built as a result

[22] https://www.tableau.com/products/desktop
[23] In Tableau, a story is a sequence of visualizations that work together to convey information. (https://onlinehelp.tableau.com/current/pro/desktop/en-us/stories).

of the previously presented artifacts. Conceptually, it is a model based in our concept map that defines our controlled vocabulary and is expected to work on top of our multi-dimensional model fed by processed data using the data profiling method. These templates are Tableau workbooks that can be applied to any data source as long as the base structure is compliant with our conceptual schema presented earlier.

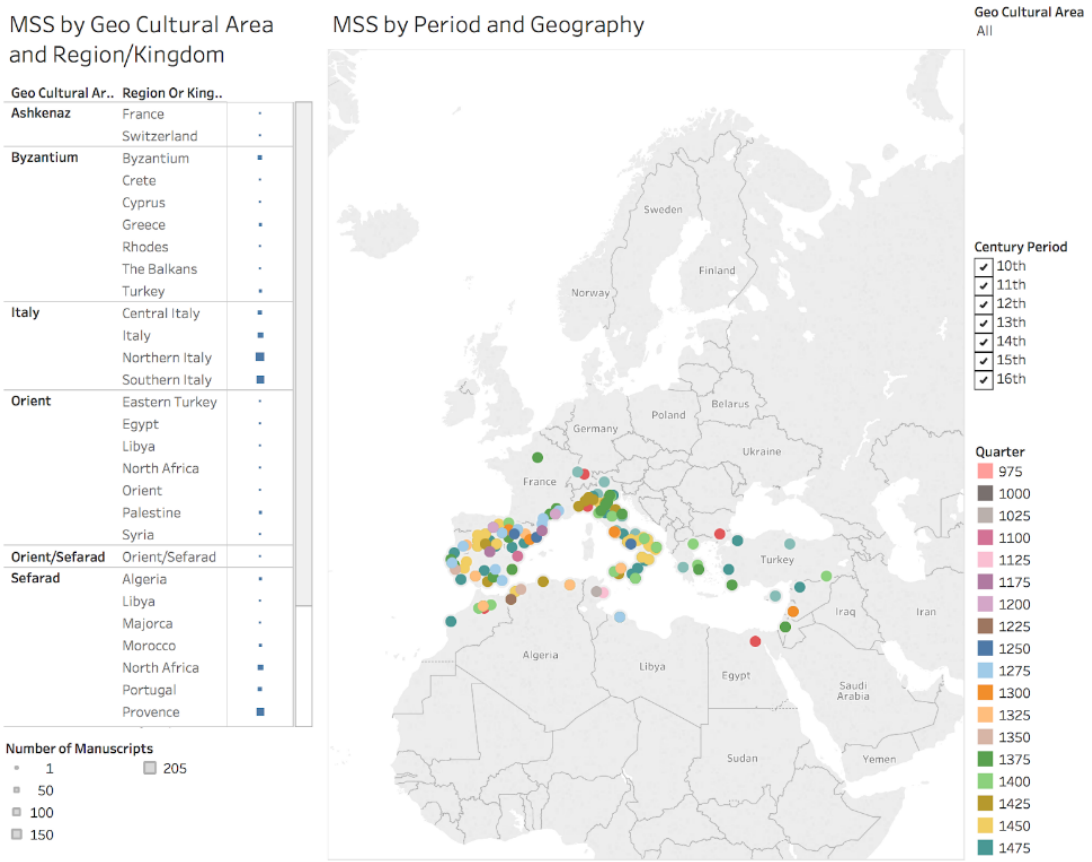

Figure 3.31 - Example of a dashboard for geographic information

As result, we provide to experts an interactive tool where they observe what they want when they want without needing to request a computer scientist. Furthermore, they are able to communicate their data stories in a much richer way, as seen in Figure 3.31.

The proposed model for visual analytics follows some standards to make the look and feel of the presentation consistent. Each feature in analysis has a common color scheme in order to keep consistency across dashboards and to keep the story straight when presenting multiple dashboards. This is also applied to quantifiable measures, such as counts and or sums which follow the same coloring scheme. Such detail is important when communicating our findings to audiences. Therefore, a special attention should be taken styling these templates (Knaflic, 2015).

Finally, alongside with the expert these visualizations suffered several iterations between the cleaning process as mentioned earlier when data inconsistencies were captured.

3.4.1 Conceptual Areas of Analysis

The identification of main areas of analysis was undertaken using a typical business intelligence reasoning, that is, defining the big picture (or overview) and then assess hierarchical levels of information. These perspectives conceptually model our templates reasoning over five perspectives and groups the available worksheets/dashboards labeled by a defined color scheme.

These perspectives of analysis were defined to address the visualization of codicological data of manuscripts:

Perspective 1. material aspects, including writing material, format, layout, quiring and ruling methods;

Perspective 2. contents and purpose, that is the main subject and destination (for a patron or oneself), including the presence of decoration;

Perspective 3. scribe and palaeography, that is, all aspects dealing with the writing of the codex, including number of hands and type and mode of script;

Perspective 4. geographic analysis, within a geo-cultural region, kingdom and locality;

Perspective 5. historical analysis, which considers the ‘biography’ of each artifact, their scribes and commissioners, as well as their incorporation in book collections.

3.5 Unsupervised Learning and Knowledge Extraction

Typical Business Intelligence designs are built with predefined performance indicators that lead the business onwards on the decision-making process. Moreover, visual analytics are built to display these indicators in a clear way to the interested parties within the organization. However, when no previous knowledge on which indicators and which features we should visualize is known it becomes hard to build the right visualization tools in a meaningful way.

For a feature space of 51 features, assuming a two-dimensional space, we would end up with 2601 relational matrix among all features. Plus, the drill down and slicing of information, no meaningful knowledge would be capable of being extracted.

The need for computational methods for automatic data mining processes came due to some expert bias, since the building of exploratory templates were dependent on the expert's input. More important than making exploratory analysis on what we already expected to know and confirm, the goal was to find the 'outliers' in the data that could extract new knowledge and raise new questions. However, deciding which relationships would be interesting to see revealed to be a time-consuming task with no direction given the matrix of combinations.

Therefore, to tackle this problem unsupervised learning techniques to discover hidden patterns within our data and knowledge extraction techniques were performed in such way it would lead us to spot significant associations to build new plots and/or dashboards.

Based on the loaded data, we performed additional data processing steps to make the data more suitable for applying such machine learning methods. Analysis on the variation of the possible values were applied to drop features where most of the possible values were the same, marked unknown values as missing values and dropped those who represented more than 50% of the corpus.

3.5.1 Cluster Analysis

To find hidden patterns within our data the first approach was to apply unsupervised learning algorithms to see if we could find any clusters. Unlike supervised learning, these algorithms do not require us to inform the algorithm of the expected output in terms of classification. Clustering has been discussed extensively in many areas such as similarity search, customer segmentation, pattern recognition and trend analysis (Wang & Gu, 2010). This section is part of a previous work in machine learning field (Pateiro, 2018) and we will present the overall results that were obtained, a more detailed analysis was performed in the referred work.

Moreover, based on initial explorations we knew that the data had more than 70% of nominal and ordinal features and, therefore, algorithms based on similarity metrics rather than distance metrics were expected to perform better. The cosine similarity metric, for example, is explained as follow:

$$sim(x, y) = \frac{1}{\frac{x \cdot y}{\|x\| \times \|y\|}} \qquad (5)$$

Since our data is nominal, we did not want to put weight on the features' values. Similarities based on the angle of two vectors rather than their length are expected to behave better. Furthermore, the use of traditional K-Means type algorithm is limited to numeric data (Ahmad, 2007). We compared both k-methods algorithms as well as hierarchical algorithms to see the accuracy of each one. Furthermore, a benchmark was performed with other algorithms to retrieve some conclusions. To benchmark the clustering algorithms the sample had n = 683 manuscripts and p = 34 features.

As said before, given the categorical properties of our data set and after applying one-hot encoding technique, the number of features increased a lot depending on the unique set of nominal values that we have which causes the curse of dimensionality. Moreover, high dimensionality data caused the visualization to be harder and less meaningful. The approach was to drop the target column and the remaining nominal features to be exploded with one-hot encoding which will transform their possible values into columns increasing dimensionality.

In this section we will present some initial findings applied to clustering based on the material type.

3.5.1.1 K-Means Algorithm

Before applying the K-Means algorithm we used PCA (Principal Component Analysis) as a feature reduction technique against our encoded dataset. From the PCA reduction, the features available can be explained with an average of six components, for a minimum variance of 70% (which ideally should be $\geq 80\%$). Moreover, the scaling performed had a noticeable impact on the results. But those results can be caused from the fact that more than 70% of our features are binary (since we applied PCA based on a dataset with one-hot encoding). Since the algorithms always try to minimize the optimal distance for a binary feature that might result in undesired outcomes. The accuracy of a clustering algorithm is dependent on the shape of the produced cluster in space. It is known that K-Means only handles circular convex shapes and performs what is called hard clustering. This type of clustering indicates that one point in space either belongs or not to a cluster. Hierarchical clustering is still a hard-clustering algorithm but allows non-convex shapes (Bhargav & Pawar, 2016).

To determine the number of k we calculated the Elbow method using distance functions as well as the silhouette score. However, sometimes it is ambiguous and does not allow us to clearly identify the elbow point. So, silhouette score was also used to measure how close each point in a cluster is to the points in its neighboring clusters.

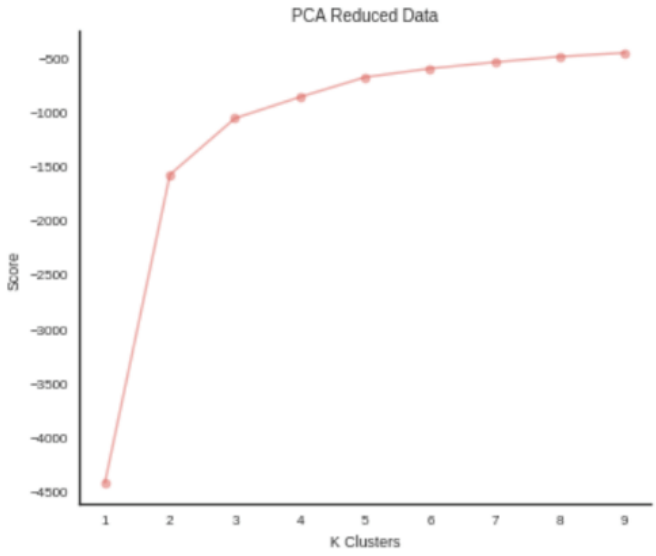

Figure 3.32 - Elbow measure based on K-Means score

The plot in Figure 3.32 above shows the application of Elbow method to determine the ideal number of *k* clusters. However, it is not clear if the determined number is two or three (although we know it should be three, considering the existing material types).

Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. Silhouette coefficients near 1.0 indicate that the sample is far away from the neighboring clusters. A value of zero indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster [24].

[24] As documented in http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

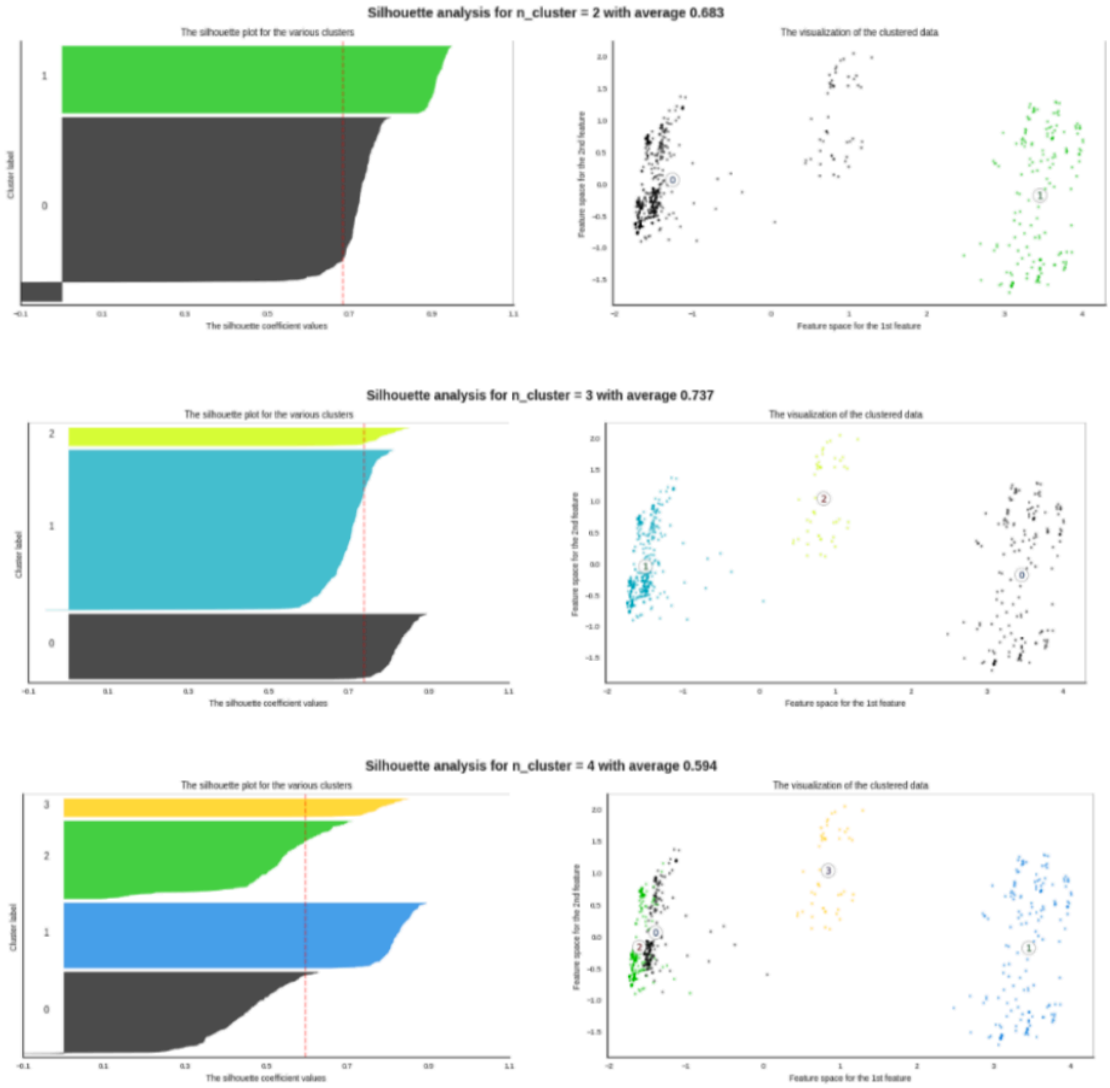

Figure 3.33 - Silhouette score and sampling for two, three and four clusters

The chosen *k*, as seen in Figure 3.33, has scored an average of 73% which, above 70%, is a good indicator. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation), so an average of 73% tells us that the assigned objects to their clusters are well assigned.

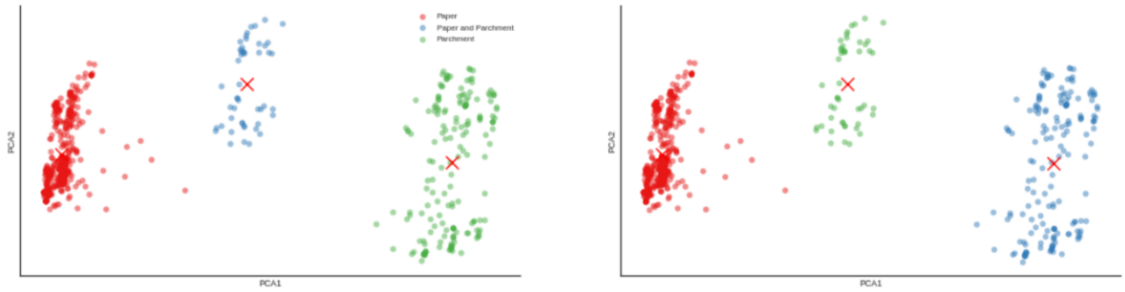

Figure 3.34 - K-Means labelled data (right) against original PCA components (left)

From Figure 3.34, we scored an average accuracy of 65% (based on 50 runs for the *PCA* reduced one-hot encoded dataset) with a silhouette score of 0.74 which refers to how similar an object is to its own cluster. And the obtained result is reasonable in terms of found structure. As for the similarity score (adjusted random score) we got a value of one which state that the clusters are identical. Moreover, the homogeneity and completeness score of one satisfies homogeneity stating that among all clusters data points are member of a single class and all those classes are members of the same cluster (completeness).

The confusion matrix result states that for “Paper” the algorithm correctly identified them, but the remaining classes got swapped. Incorrectly identified 49 “Paper and Parchment” as “Parchment” and incorrectly identified 174 “Parchment” as “Paper and Parchment”.

3.5.1.2 Hierarchical Clustering

After some research we stated that datasets with high volume of categorical nominal features would not perform well on distance-based algorithms, since the possible values would always be zero or one and too many features are generated from one hot encoding for this type of values. As alternative, similarity measures such as cosine (using the angle between two features), would perform better in our dataset. One advantage are the possible cluster shapes which allows a better granularity on the clusters' points, since gaussian mixtures and k-methods operate better on convex shapes (Bhargav & Pawar, 2016).

To determine which method and metric fits better in our dataset we used the Cophenetic Correlation Coefficient. This coefficient compares (correlates) the actual pairwise distances of all samples to those implied by the hierarchical clustering. The closer the value is to one, the better the clustering preserves the original distances. Given the type of dataset we were expecting to have better results with similarity measures. For the dataset in its pure state we got cosine as better metric with a cophenetic result of 0.998.

The accuracy obtained using hierarchical clustering was worse than with K-Means with an average percentage of 64% (against the same 50 runs). Even the homogeneity and completeness scores were very low and does not satisfies homogeneity stating that among all clusters data points are not member of a single class and all those classes might not be members of the same cluster (completeness). The adjusted random score (ARI) is close to zero which indicates random labeling independently of the number of clusters and

samples. The dendrogram plotted showed promising results but that does not translate into success once we see the labels calculated. With a silhouette score of 0.97 which refers to how similar an object is to its own cluster (cohesion) compared to other clusters (separation) it improved a lot.

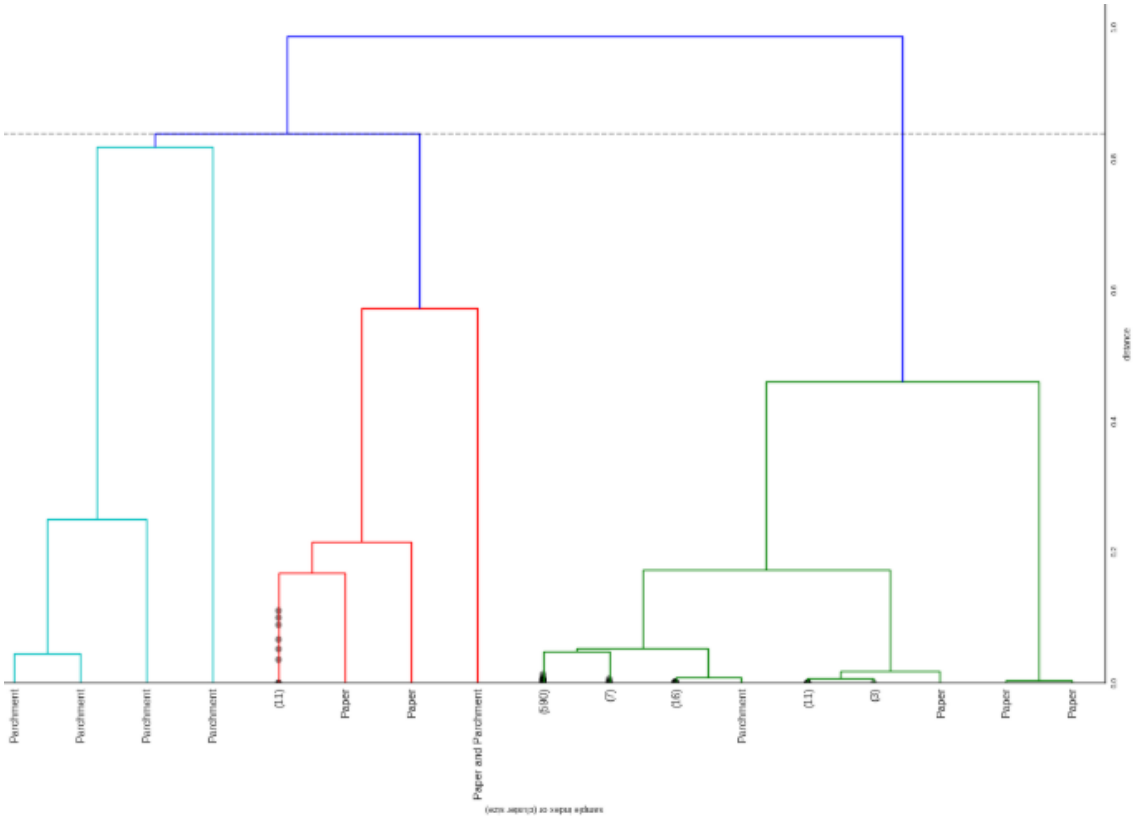

Figure 3.35 - H. Clustering Dendrogram for Material analysis

3.5.1.3 Results

After exploring K-Means and Hierarchical Clustering we performed benchmarks against other algorithms. The benchmark used 50 runs applied to different datasets (PCA based dataset, raw data set, *MinMax* and *MaxAbs* scaling methods) and different algorithms. The obtained average score was below 40%. The inconsistent nature of accuracy makes it hard to retrieve information. Beside the higher accuracy results in general the accuracy did not go above 65%-70% which was not the expected result.

Overall results had shown an accuracy of 65% with some variations, where K- Means and Gaussian Mixture had some runs with 100% scores. However, their deviations are high being Spectral Clustering and Hierarchical Clustering the ones with highest consistency but always below 70%, as shown in Figure 3.36.

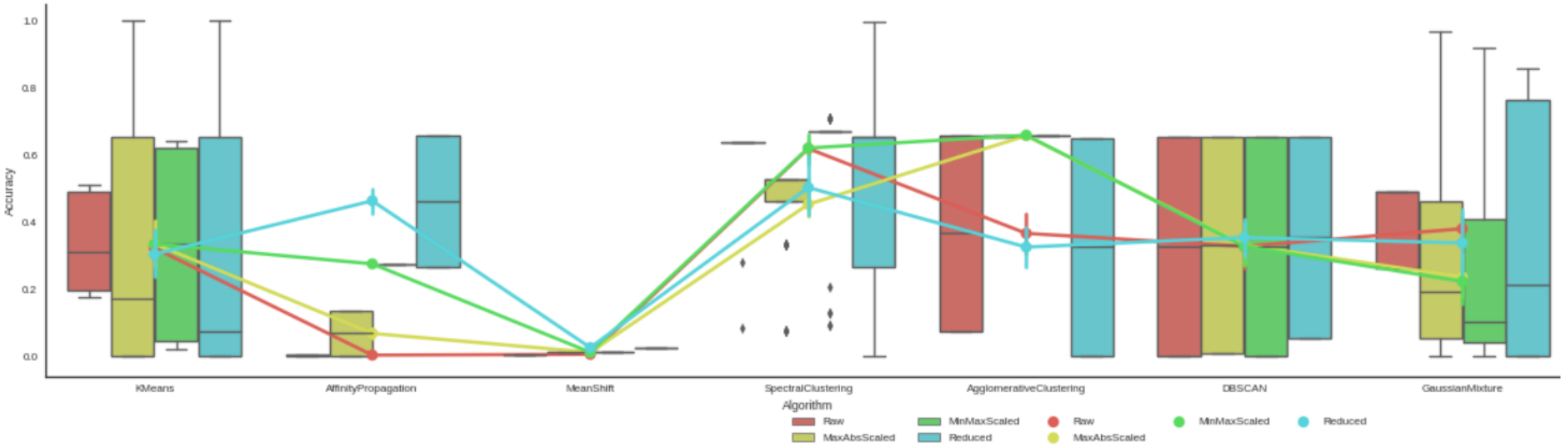

Figure 3.36 - Overall Accuracy per algorithm and dataset including mean and deviation

Ideally, storing the model state for K-Means that had the highest score as fitness function could help us keeping a good accuracy.

In high dimensional datasets the visualization of information and even relevant clusters is clearly a challenge that needed to be addressed. Model-based algorithms are not fitted for this data tracing and even the PCA technique used makes it harder to see which features contributes the most for each cluster.

Hierarchical clustering was the best approach in terms of data visualization given its graph-based nature and, therefore, can be improved to increase the quality of visualization. Although the dendrogram was a good plotting tool as well as the parallel coordinates for dense datasets visualization becomes harder.

Therefore, after some research we believe that instead of exploring cluster analysis to retrieve groups of manuscripts we should focus in feature clustering. To see how each specific categorical value relates with another in an associative manner.

In the next section we will present this data mining approach called Formal Concept Analysis.

3.5.2 Formal Concept Analysis

Formal concept analysis (FCA) is a method for data mining, knowledge representation and information retrieval. As explained in Chapter 2 section 2.4.1, it is based on concepts and concepts hierarchies extracted using theory of complete lattices.

Due to our needs to enhance the visual explorations, we defined a method to extract rule associations based on the FCA algorithm. In fact, given the corpus in study and an optional input on what feature should be our object key, the method then generates the encoded source to apply the FCA and retrieve the context files.

This data mining technique, unlike cluster methods from unsupervised learning algorithms, performs well on high dimensional datasets with categorical features. FCA models these concepts as units of reasoning, consisting in extents (all objects that belong to a concept) and intents (all attributes common to all of those objects). Moreover, the ability of extracting the concepts and feature relations revealed to be easier than for clustering methods.

Given the categorical nature of our data, clustering methods applied in the initial analysis revealed to be inaccurate, since these algorithms operate on distance measures such as Euclidean measures. The problem with this approach was that two categories within a feature ended up being measured in terms of weights when, in reality, they should have the same importance/weight. Therefore, similarity measures should be used to define how similar an object is based on true or false approach or based in the minimal angle performed between two objects regardless how far away they are in terms of vectorial space distance.

FCA uses similarity measures instead which we believe to be a better approach for knowledge extraction.

Like clustering methods, FCA also requires the definition of an object to be the key term for analysis. We needed to change our dataset in order to apply FCA, since it operates on what it is called a context of analysis. Meaning that for all the possible values within a feature, one-hot encoding method was applied to transform the dataset in a binary structure where an object has or not (true or false) and value of a specific feature. For categorical variables where no such ordinal relationship exists, the integer encoding is not enough. In fact, using this encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results as stated before. Therefore, one-hot encoding operates on a binary representation to infer whether it is true (one) or false (zero).

single_handed_false	single_handed_true	century_period_10th	century_period_11th	century_period_12th	century_period_13th
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
1	0	0	0	0	0

Figure 3.37 - Encoded data sample prepared for FCA

As shown in Figure 3.37, the nominal values of a feature are transposed to new features such that a given manuscript only has, for example, true or false "*single_handed*" feature and not both. Applying this method resulted in 188 features increasing our dimensional space by more than 50%.

Figure 3.38 shows the developed visualization tool based on the concept tree after the the application of FCA algorithm. We used the *InClose*[25] tool (as a black box approach) to apply the algorithm and retrieve the concept tree. The tree represents the extents and intents of each concept within the context. For the example above, a context was created for manuscripts where material type is "Paper" and our object key is the "Geo Cultural Area". Of course, from some previous knowledge of data, we knew beforehand that we could drop the features whose presence is dependent on the type of material.

From the results, we can observe that 34% of our objects has a subject "Philosophy and Kabbalah" and from those, 12% have no information about quiring. In fact, not having information about quiring makes the quiring composition unclear, which makes

[25] https://sourceforge.net/projects/inclose/

sense given the dependency between these two attributes. Moreover, the corpus dropped all the non-nominal features (for example, measurements) in order to only consider categorical values.

The scalability of this method depends on how deep we are interested to go in the depth of the context tree. The generated trees can be very large and harder to process on further processes or even for exploratory purposes.

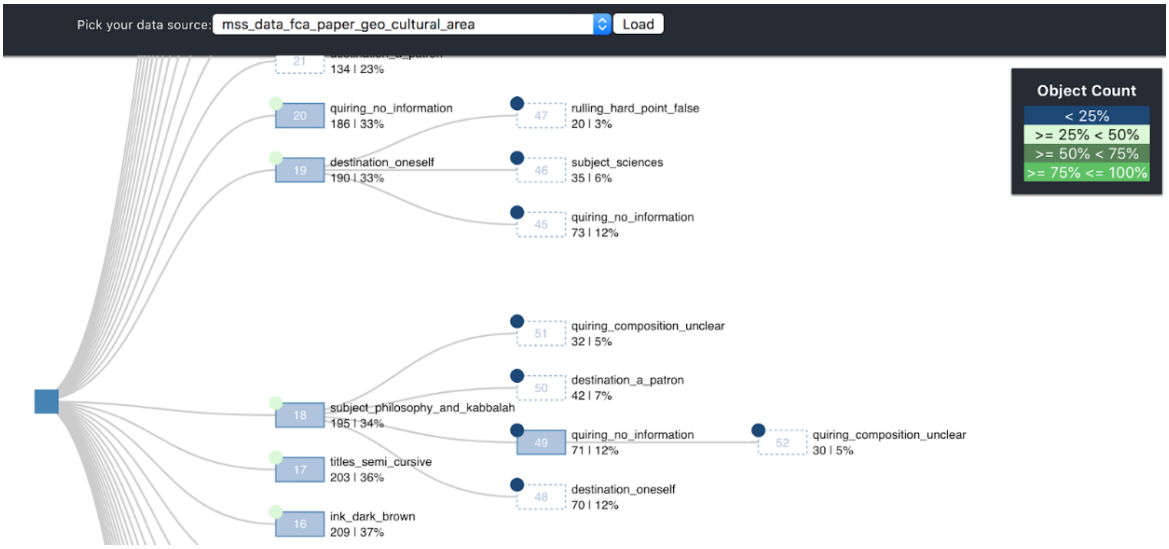

Figure 3.38 - Tree of FCA produced concepts[26]

The example above generates 157656 concepts for 145 codicological units plus 188 features. To reduce the depth size of the tree we can define what is called as minimum support. The approach was to focus on the size of the concepts, using the well-known idea of minimum support (Andrews & Orphanides, 2010; Zaki & Hsiao, 2005) to filter out relatively small concepts (noise) from the data. This was achieved by specifying a minimum number of objects and/or attributes for a concept (Andrews & Orphanides, 2010). Following the example above, applying the FCA algorithm using a minimum size of extent (number of objects) to 20, the number of produced concepts is reduced to 91813.

3.5.3 Association Rule Mining

Association rule mining finds all the rules existing in the database that satisfy some minimum support and minimum confidence constraints. For association rule mining, the target of discovery is not predetermined, while for classification rule mining there is one and only one predetermined target (Liu et al., 1998). This data mining technique is

[26] Available in https://codicodavis-kdd.herokuapp.com/

currently used on recommendation systems for many businesses, such as customer segmentation and/or product recommendations.

As for the context tree generated the amount of associations can be very large. Therefore, using FCA helped reducing the number of resulting rules without loss of information. The task of mining association rules is to determine all pairs $X \rightarrow Y$ of subsets of *M* attributes and *G* objects (forming a binary relation $I \subseteq G \times M$ where object *g* has attribute *m*) such that the support $supp(X \rightarrow Y) := supp(X \cup Y)$ is above the threshold *min(supp)* $\in [0, 1]$, and the confidence $conf(X \rightarrow Y) := \frac{supp(X \cup Y)}{supp(X)}$ is above a given threshold *min(conf)* $\in [0, 1]$. Association rules are for instance used in warehouse basket analysis, where the warehouse management is interested in learning about products frequently bought together (Lakhal & Stumme, 2005).

The association rules are a model artifact (from DSR) to be used by the experts to perform a more in-depth analysis of data. This model is an outcome of the previous method applied through FCA, presented in section 3.5.2 from page 61.

```
11211 < 177 > ink_dark_brown rulling_hard_point_true illumination_decoration_none =[99%]=> < 176 > laid_pattern_or_grid_visible;
11212 < 177 > ink_dark_brown rulling_hard_point_true illumination_decoration_none =[99%]=> < 176 > orientation_regular;
11213 < 177 > titles_square quiring_composition_uniform main_script_mode_semi_cursive single_handed_true illumination_decoration_none =[99%]=> < 176 >
laid_pattern_or_grid_visible;
11214 < 177 > subject_philosophy_and_kabbalah rulling_hard_point_true =[99%]=> < 176 > orientation_regular;
11215 < 353 > main_script_mode_semi_cursive single_handed_true illumination_decoration_none rulling_mastara_false orientation_regular =[99%]=> < 351 >
squared_format_no;
11216 < 353 > main_script_mode_semi_cursive single_handed_true rulling_hard_point_true rulling_mastara_false orientation_regular =[99%]=> < 351 >
squared_format_no;
11217 < 176 > titles_square format_medium main_script_mode_semi_cursive rulling_hard_point_true chain_line_visible script_family_sefardic squared_format_no
orientation_regular laid_pattern_or_grid_visible =[99%]=> < 175 > with_watermarks_visible;
```

Figure 3.39 - Association Rule model sample using ConExp[27]

Following the previously presented example for paper manuscripts by geo cultural area, Figure 3.39 illustrates the generated model for rule association. This model was generated using *ConExp*[27] tool which is a concept explorer based in FCA which generates association rules. The selected sample shows that a manuscript with a "*Philosophy and Kabbalah*" subject and ruling hard point in 99% of cases its orientation is regular. In fact, rather than analyzing facts with high amount of similarity (hence 99%) the experts could use this model to have insights on outliers, meaning that an evidence of unexpected facts could lead to new findings and raise new questions.

Furthermore, the modelled templates for visual analytics could be enhanced to slice the information based on these rules and visualize them in more meaningful ways.

[27] http://conexp.sourceforge.net/

Although in its early stage, the rule mining techniques and FCA provided good insights on the explored data and the potentialities such method can bring to enhance the visualization with further discoveries. The goal is to include these concept trees as part of the exploratory analysis and storytelling in such an easier way for experts as we will see in the following Chapter 4 Demonstration. There, we demonstrate the visualization templates and how they can enrich the communication of findings and how they enable the desired exploratory capabilities.

Chapter 4

Demonstration

This chapter shows the application of visual analytics onto our processed data source. With the data profiling method applied and the data model loaded with our corpus we will apply the visualization templates in the perspective of our expert in the Humanities/Digital Humanities.

Moreover, these templates are going to be applied to the entire corpus of manuscripts where the two datasets from different periods were merged.

This chapter is organized as follows:

- Section 4.1 provides a brief description on how the storytelling is going to be conducted as well as some context of the scope;
- Section 4.2 shows how an expert would use the dashboards to present their research and how they are able to interact with these dashboards;
- Section 4.3 closes the chapter with a few conclusions over the data based on the presented visualizations;

4.1 Overview

Although still scattered in a variety of sources, there are substantial amounts of codicological metadata on Hebrew manuscripts, such as the database Sfardata (Beit-Arié, 2017), entirely dedicated to the description of all dated Hebrew manuscripts copied until 1540, as well as library catalogues and a variety of expert publications that provide us with abundant, albeit heterogenous codicological descriptions.

Hence, using Hebrew manuscript data as a starting point, these templates provide an environment for exploratory analysis to be used by Humanities experts to deepen their understanding of codicological data, and to formulate new research hypotheses. As stated in Chapter 3, Section 3.4.1, the analysis and study of the corpus follows five perspectives of codicological data of manuscripts. We will present a demonstration covering the first

four perspectives, material aspects, contents and purpose, scribe and palaeography and geographic analysis. The available information presented in the following dashboards regards to Sephardi manuscripts produced between 10th to early-16th centuries. Such manuscripts were chosen based on the expert's need to verify ongoing work in her field that required such subset of analysis.

Our demonstration is based on the storytelling that is made possible to achieve following the dashboards, raising evidences and confirming theories based on the shown data.

4.2 Codicological Dahsboards

As shown in the first dashboard (Figure 4.40), where information on contents and purpose is presented, several conclusions can be drawn. The first concerns the distribution of subjects, specifically the fact that although one would expect bibles and related texts would predominate, one observes instead that philosophy and kabbalah and the sciences are more significantly numerical. However, these are not subjects with significant representation in the early Hebrew printed editions. Hebrew printing began in the early 1470s, meaning that for 30 years handwritten and printed Hebrew books coexisted. Therefore, it can be concluded that these subjects were preferred in handwritten formats. Moreover, these are the two main types of subjects where Arabic language was employed more frequently, an indication of what were the original sources of these texts, and, to some extent, the origins of the scribes, who likely lived in areas more influenced by Arabic scholarship, specifically Northern Africa and Spain. As for destination, one observes that commission was the main reason of copy. A curious observation is that one book was copied by a groom for his bride, which is quite an unusual event.

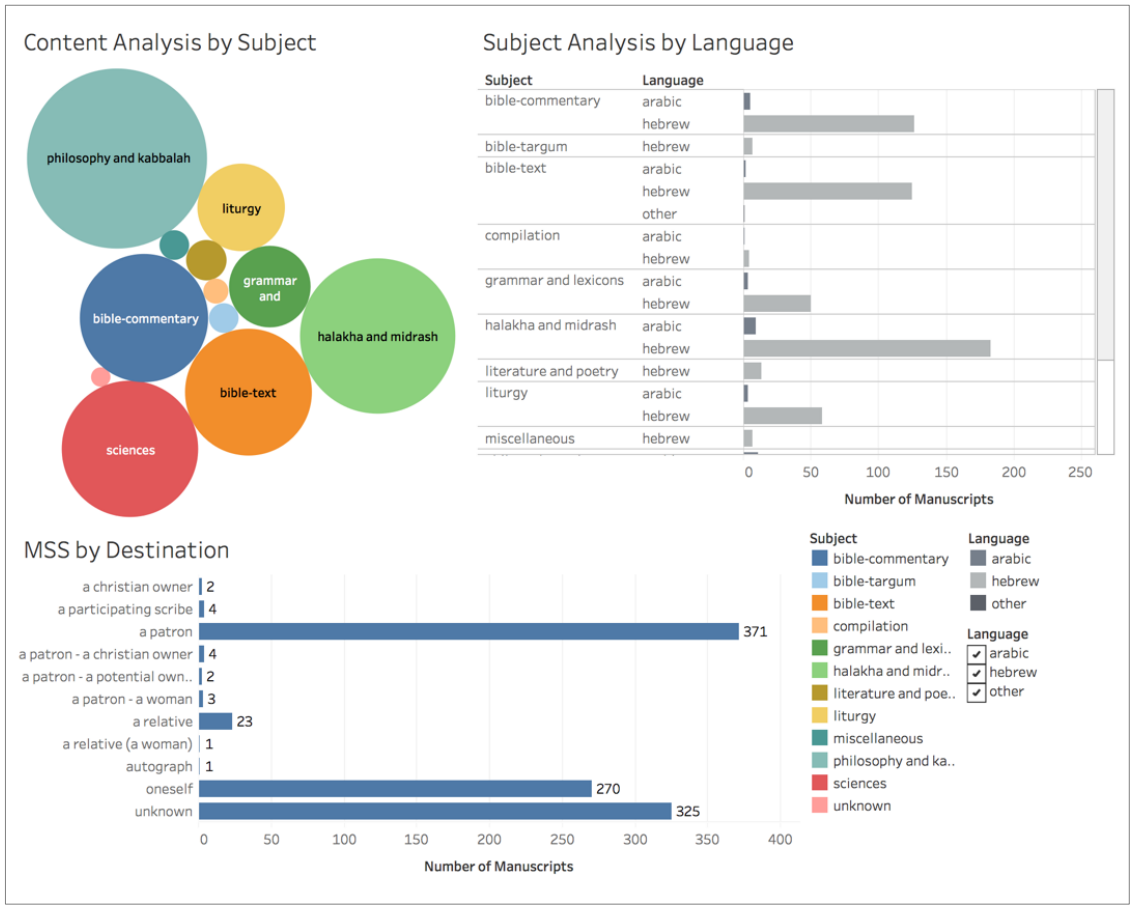

Figure 4.40 - Dashboard for Content Analysis on Subject and Destination

A second example concerns the materials employed in Hebrew manuscripts of the fifteenth century. Particularly important is the use of quires made of both parchment and paper (see dashboard in Figure 4.41). These mixed quires usually include an outer and central bifolium of parchment and the remaining bifolia are in paper. According to Beit-Arié, (2018), one fifth of all mixed quires appear in Byzantium, early on. Following the data shown in Figure 4.41 it is possible to conclude that by the fifteenth century, however, this type of quire appears spread into several regions, although, in particular, in regions adjacent to Byzantium.

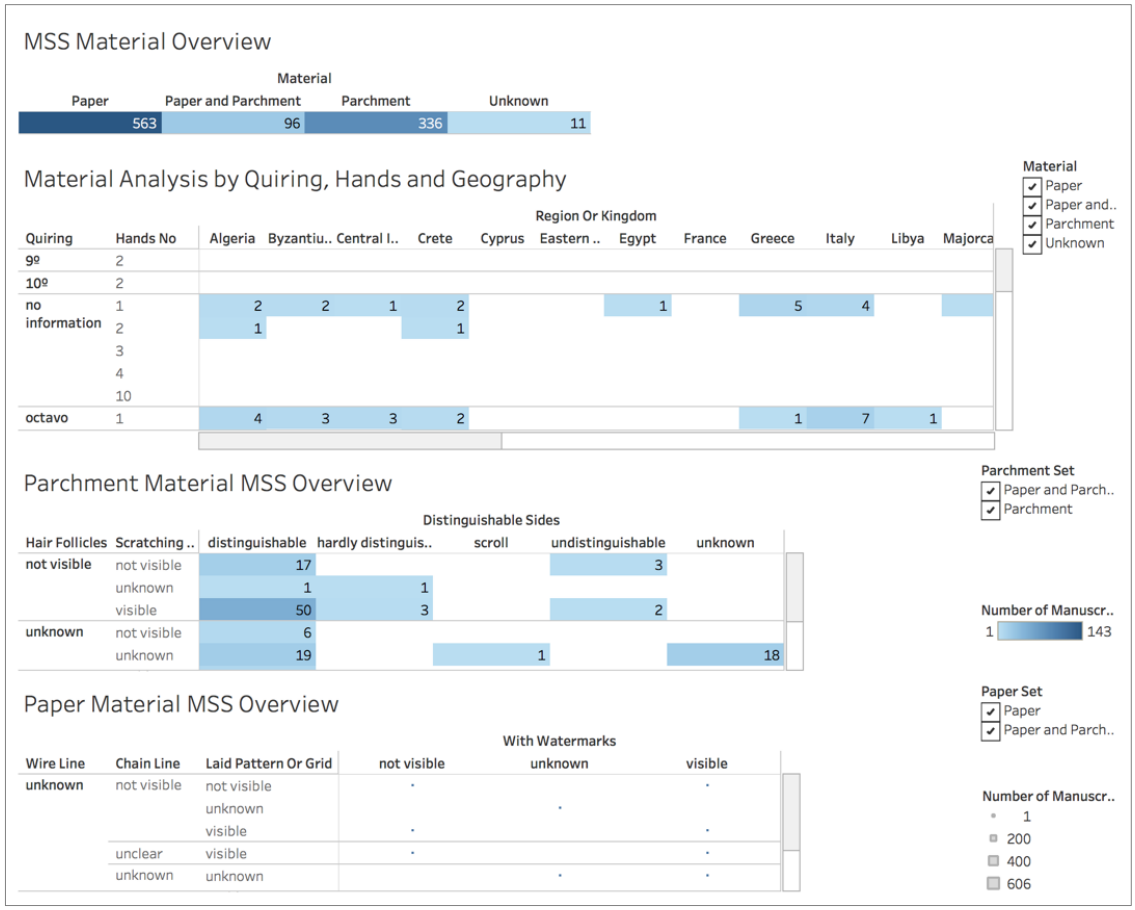

Figure 4.41 - Dashboard for Material Aspects Overview with a drill-down by material type

Another conclusion that stemmed from visualizing the data concerns the format of Hebrew books. As shown in Figure 4.42, there is a predominance of medium formats, followed by small-sized codices. In contrast, there are few oversized volumes, contrary to what happens in other geo-cultural areas of production of Hebrew books. This indicates that books were predominantly intended for personal use (oversized and large formats, conversely, were intended for communal reading). Another distinction, between medium and small formats, concerns the portability of the latter, whereas the former was the preferential format for study. The portability of small formats also explains the existence of pocket-sized codices. Even though their amount is not particularly telling, this is a format that only appears in the mid-fifteenth century. It was an easy format to carry, for instance, to the synagogue, and its usage can be predominantly associated with Italy, followed by Portugal. There are plenty of other examples of this format, but none possess date and therefore were not integrated in the corpus.

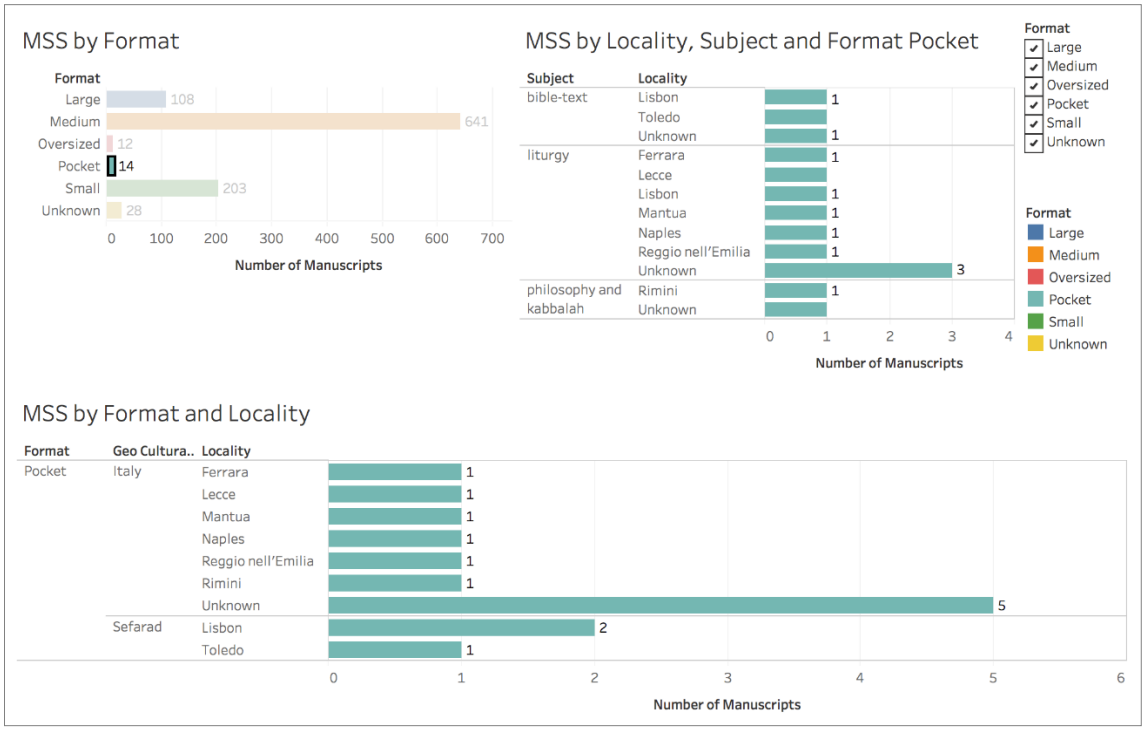

Figure 4.42 Dashboard for Material Aspects showing interactive format analysis

Books with decoration are the primarily surviving medieval artistic Jewish heritage. In general, these were luxury items, done with great care, and often reflect the artistic tendencies of the places where they were done. Decorated books are not the majority, and in some cases, they only include minor elements of decoration. As shown in Figure 4.43, the majority of books that include any means of decoration were prepared for a specific person, who commissioned it. In a few cases, it is also possible to conclude that books served as gifts for women, which indicates that Jewish women were literate, at least in Hebrew. As expected, the most common type of text to be decorated in the Bible, the most sacred text, followed by liturgical volumes. However, it was surprising to see that texts of philosophy and kabbalah were also commonly decorated.

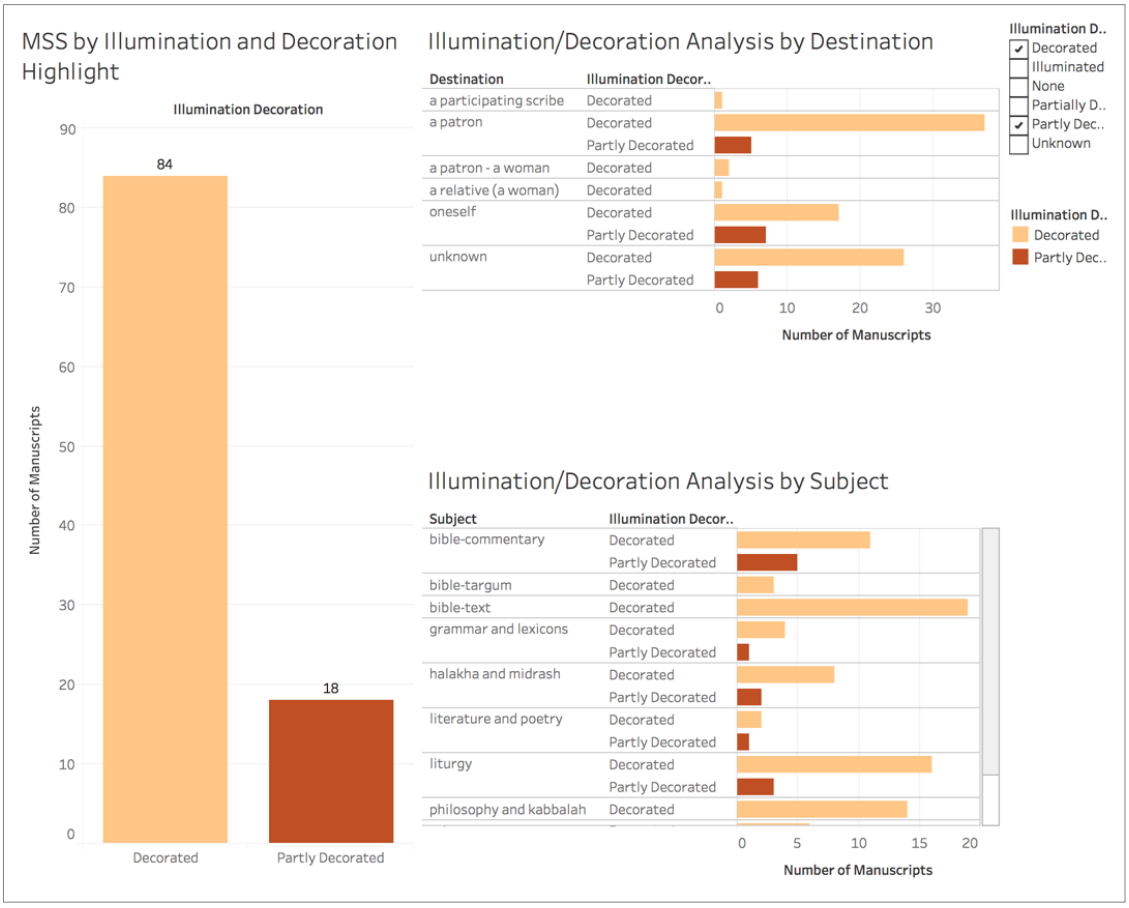

Figure 4.43 - Dashboard with Decoration info. based on subject

Still with regard to manuscripts with decoration, which are, as mentioned, considered as luxury items, it is possible to conclude that, first of all, semi-cursive script was gradually employed in these items, even though in early productions this script mode was preferred for texts of a more practical nature. Also, as shown in Figure 4.44, this was the most common mode for writing books in Sephardi script after the expulsion of Jews from the Iberian Peninsula. This is of great importance, since this would eventually influence the preparation of type for Hebrew printing in Italy, and in time it became the most common typeface across Europe and the Mediterranean basin, even in regions that are not typically associated with Sefarad.

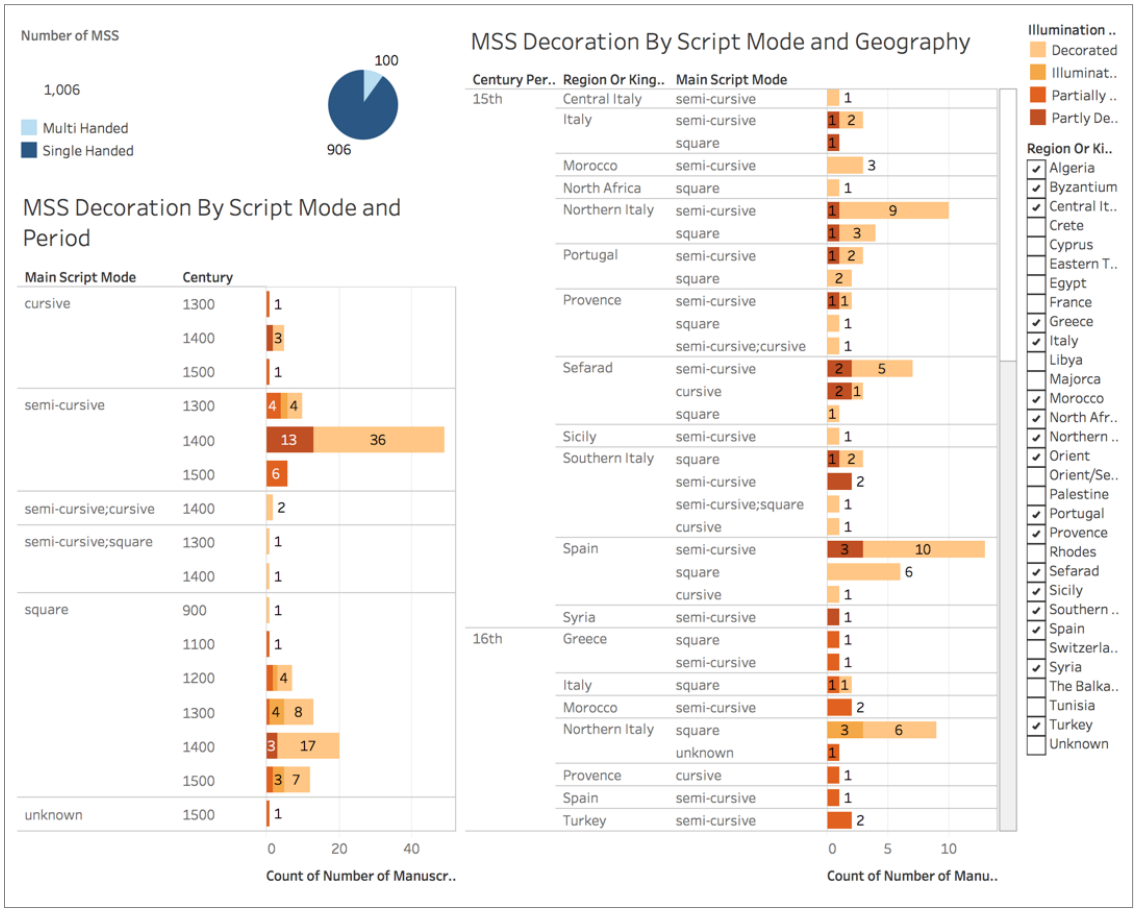

Figure 4.44 - Dashboard with Palaeographical analysis

Another conclusion that stemmed from visualization is the distribution of subjects in the different regions where books in Sephardi script were copied. As shown in Figure 4.45, contrary to what was expected, there are more copies of non-biblical texts than of biblical volumes. Some of the subjects, especially sciences, also imply the use of Arabic languages, which, as previously explained, was the main language for this type of texts. The predominance of non-biblical texts is seen in the main regions, specifically Sefarad and Italy. Also, important to notice is that, despite the higher number of volumes commissioned by a patron, the number of volumes for personal use is also outstanding, again demonstrating the high literacy among the medieval Jewish communities.

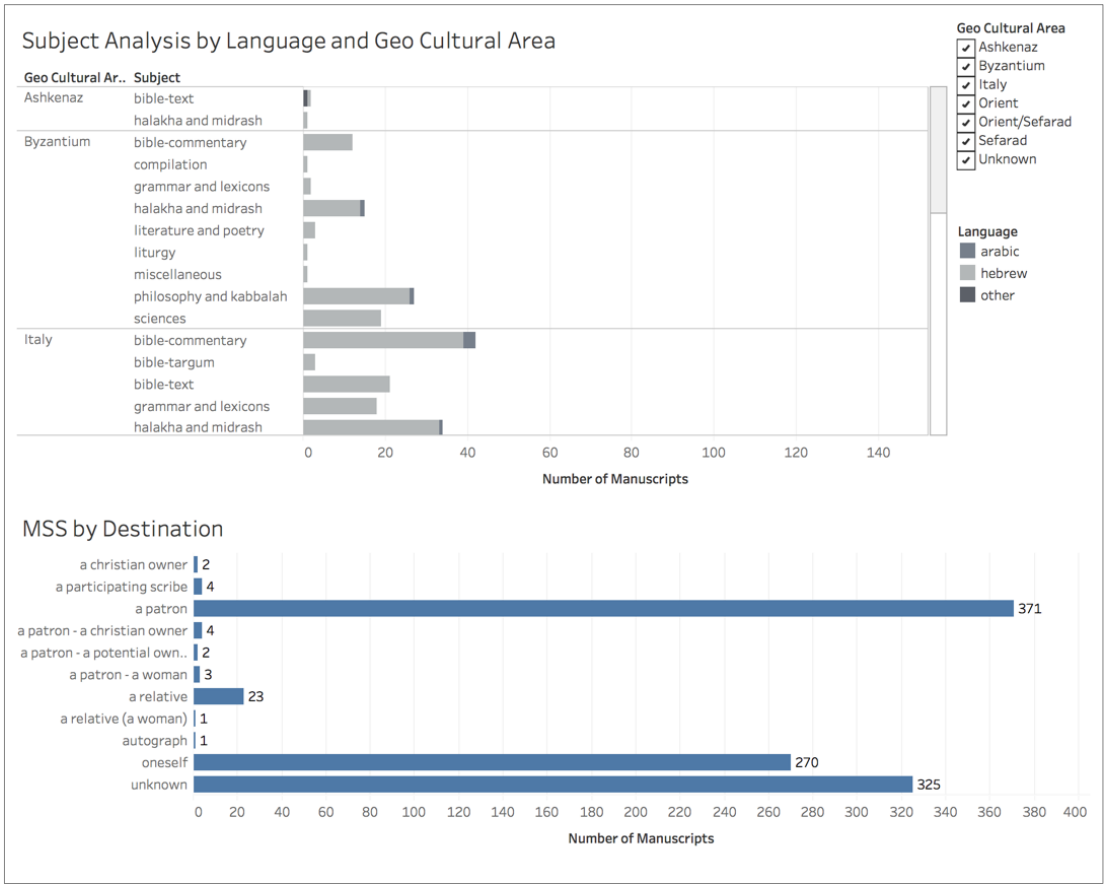

Figure 4.45 - Dashboard with contents and purpose analysis

Considering the corpus of manuscripts in Sefardi script from the fifteenth century, it is possible to conclude that paper is the predominant material, thus supplanting the use of parchment, as shown in Figure 4.46. As mentioned, this also meant the development of a hybrid type of quire, composed of paper and parchment (in the latter sides). This type of quire is especially found in the Iberian Peninsula, although for the same region it is possible to observe the increasing predominance of paper – especially in Spain.

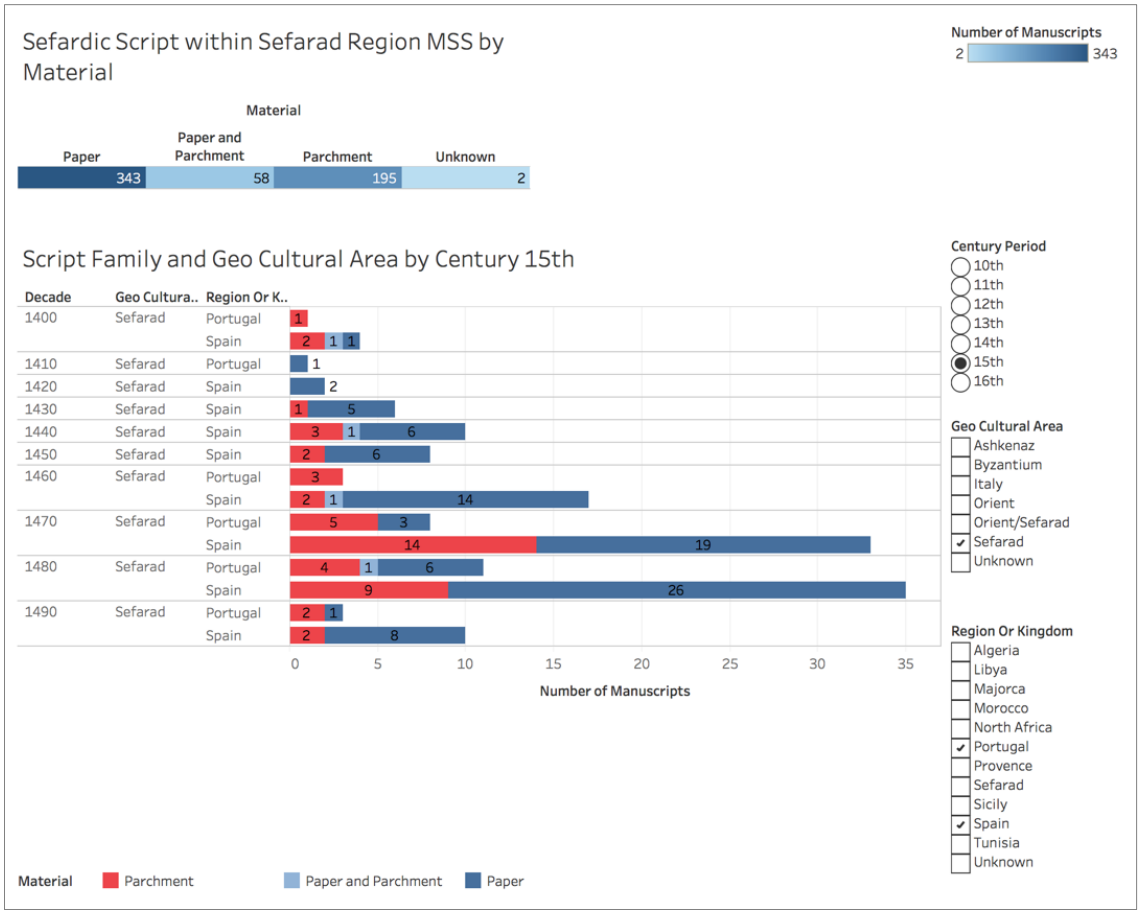

Figure 4.46 – Dashboard with Material overview

Finally, the geographic visualization (Figure 4.47) of the collated data allowed a broader view of where Sephardi script was employed. Even though it initially started in North Africa and southern Spain, it took over the entire Mediterranean basin, and it can also be found in other geo-cultural areas such as Ashkenaz (mainly Germany and France), which suggests the interest in Sephardi books, and the migration of scribes.

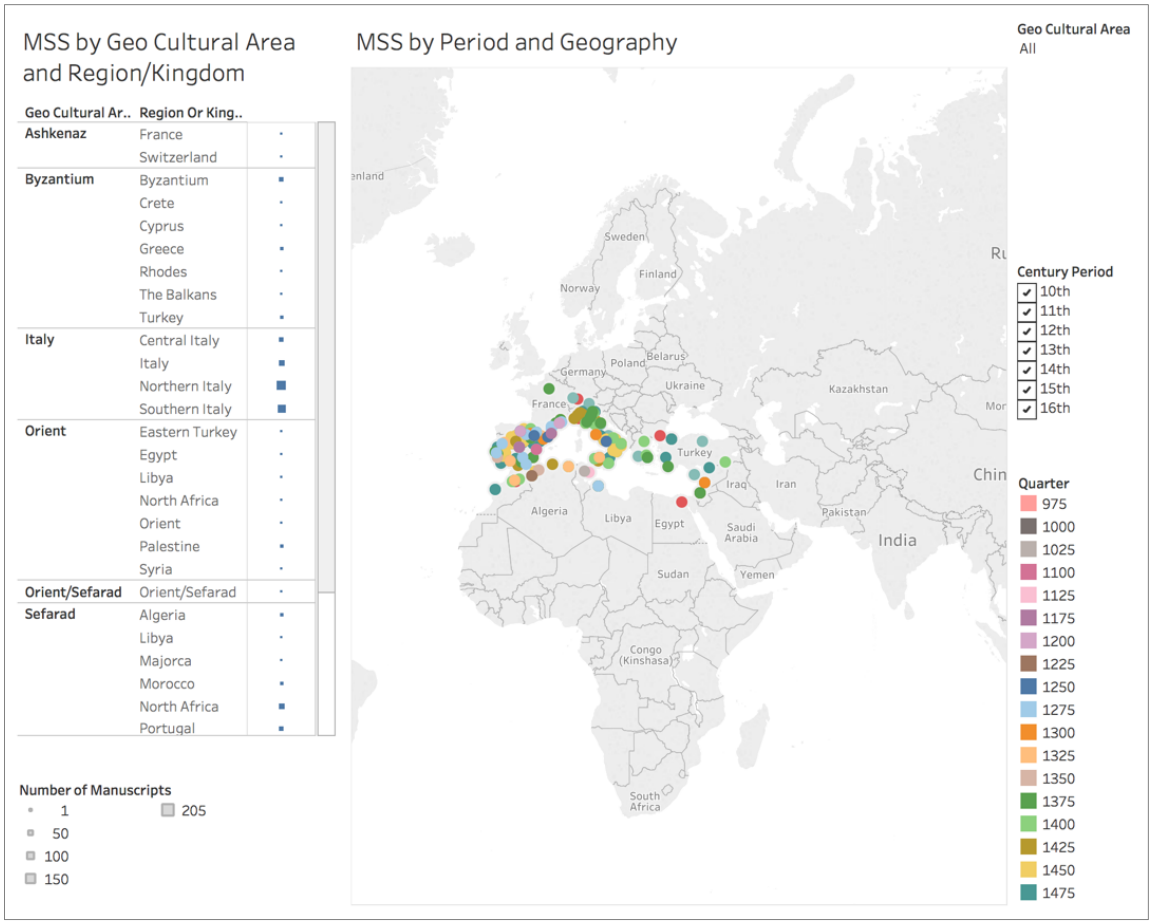

Figure 4.47 - Dashboard with Geographical information

Another look regarding the geography of these manuscripts is to see how the geographic area influences the subject when produced with more than one scribal hand. From Figure 4.48 we can observe that “bible-commentary” or “halakha and midrash” productions have a significant contribution from more than one scribe and, perhaps, it is due to the nature of these subjects that sometimes are compilations of texts.

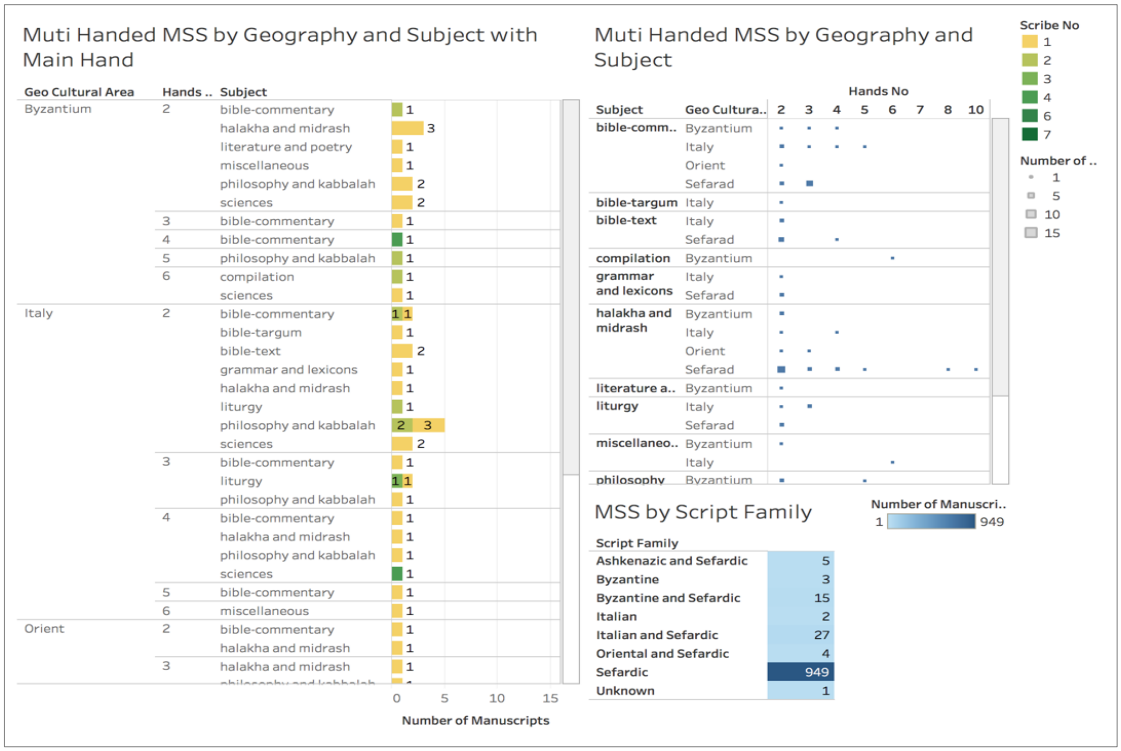

Figure 4.48 - Visualization with geographic information of multi-handed manuscripts analyzing the subject, script and the number of hands

4.3 Data Analysis

One of the most significant analysis concerns the multi-hands records. As shown in Figure 4.48, these vary from two participating hands, the majority of records, reaching up to eight. Hebrew books were not copied in the environment of a scriptorium, as were Latin manuscripts, therefore one must reason that in all probability these result from a learning process and, possibly, the environment of a school. Although the records with higher hand numbers occur in an unknown region, it is in Italy and Byzantium that one observes more variety in the number of hands. While this was expected for the latter, it was less so for Italy. With regard to subject, there is a correlation between multi-hand and predominant types of text, as referred to in Figure 4.40. Finally, our process highlighted the collaboration between various script families, which in turn materialize the mobility of scribes

Chapter 5

Conclusions

The complexity of the data analysis increases with the amount of available metadata gathered from these manuscripts. With the range of computational methods available today, experts are able to identify new evidences in data and deal with new research questions. Furthermore, storytelling through visualization of unexpected new patterns and feature interconnection based on data can be an engaging tool for new audiences and foster the sharing of information among experts. However, the descriptive nature of the gathered information is not necessarily compatible with these techniques, unless data cleaning and transformation is applied. The method built made these visualizations possible given the controlled vocabulary, although it took more than 80% of consumed time in this research.

Business intelligence solutions are applied to structured data, in which data is defined in terms of dimensions (context) and metrics (measurements). This "structured reasoning" was pivotal in bringing more structure to a highly unstructured dataset. One good example is the definition of a new geographic hierarchy (inexistent in Sfardata (Beit-Arié, 2017)), with different levels of data aggregation: geo cultural area | region/kingdom | current country | locality. Another example is the categorization of codex sizes, where we defined formulas to make format intervals (i.e., "small", "pocket", "oversized", …) as well as for orientation.

The BI and VA reasoning, with clear dimension of analysis and metrics, enabled the rapid development of data visualizations in Tableau that helped Jewish culture experts to expand their insight of the corpus. New DH research questions were raised due to the exploration of the designed visualizations. Several dashboards have been defined in order to provide the experts with an interactive and intuitive data exploration tool. In this paper we detailed only a few, as examples. The proposed research method can be replicated in other research contexts. In particular, this work has the potential to be applied to other codicological studies, namely with other adjacent book cultures such as Arabic and Latin manuscripts, from which interesting comparisons can be drawn. Moreover, it could now be applied to personal catalogues from experts to include Incunabula (early printed books)

in Sephardi type, 1470-1540, Sephardi region (emphasis on Italy). The rationale for this dataset is to bridge together the research of two traditionally separated disciplines, manuscripts and printed books, to explore commonalities and differences. For instance, to investigate whether early printed books have similarities with manuscripts in terms of material aspects. Although printed books raise new concerns in terms of analysis (given the fact that a manuscript is a single object unlike printed books that may hold several editions) we are confident that increasing our *corpus* keeping the controlled vocabulary is still perfectly feasible.

Finally, although in its early stages, the use of data mining techniques such as association rule mining through formal concept analysis provided good insights on knowledge discovery to complement and/or recommend further visualizations. This technique applied to the current *corpus* showed that having this associations between the immense feature space of categorical values could help experts to build ontologies, spotting outliers and unexpected relationships, which raised new visualizations. Without knowing which key metrics and indicators and to avoid expert bias, these initial recommendations and/or associations stand as a very useful manner to have a more in-depth knowledge of data raising new visualizations. Moreover, the unsupervised learning algorithms which we hoped to help finding unknown patterns within our data were far away from our expectations. For highly categorical datasets, these distance-based algorithms could not provide us the knowledge we were hoping. To avoid weighting the categorical values, the binary encoding technique used increased the dimensional space to more than 150 new features. Consequently, the extraction of knowledge from the clusters showed to be inefficient since no relevant features stand out.

5.1 Analysis of Research Questions

RQ1. As shown during the development phase in section 3.3.4 from Chapter 3, the data cleaning method has shown how a methodic cleaning process can make the available data, descriptive and uncertain in its nature, capable of being explored as shown in the Demonstration (Chapter 4) where the researchers were able to visualize and communicate the results. Therefore, it is possible to build a controlled vocabulary and model (as described in section 3.3.5.3) in such way that we achieved a store *corpus* to hold codicological information. However, the process of cleaning and transforming the external sources into

our controlled vocabulary is far from being easy given the nature of these collections, though, we have shown that it is achievable.

RQ2. As shown in section 4.2 from Chapter 4, the built *corpus* already confirms the scholarship. Researchers were able to verify what is already known which indicates a well-formed *corpus* so far. However, given this is still an ongoing work, the finding of new knowledge is expected to happen once a proper design, to visualize associations built with Formal Concept Analysis, is in place and available for researchers. With this tool and new corpora, it is expected to discover new knowledge.

5.2 Limitations

Although the performed steps for data cleaning allowed us to obtain a more structured and explorable corpus there are still some topics that will require further analysis. Most of them are context dependent on Hebrew manuscripts and the data source used. The geographic information, critical to rich visualizations and storytelling, is limited by the lack of collected information and the inherent uncertainty regarding region. This is due to the fact that there are still 177 artifacts with no information besides region/kingdom. Consequently, inferring the current country and locality is nearly impossible.

Furthermore, the unknown or missing values require additional enrichments from external sources. In fact, addressing these concerns within Digital Humanities is critical, to allow different groups to share standard information, and to contribute to good practices when collecting data on manuscripts.

Also, we have identified manuscripts copied by several hands, which produces several entries for the same manuscript. Although such duplicated manuscripts problem can be handled, the problems arise where the script is marked with different types (the same manuscript might be marked as Sefardic and Byzantine at the same time).

The previous statement led us to the discovery of missing information due to the query being done on Sfardata (Beit-Arié, 2017), but since the adopted methodology is iterative, it was resolved by an additional processing round. Furthermore, the method of annotating each case, keeping the original information (for instance, the partitions between multi-hand and single hand) provided us the tools to enrich the analysis to compare the original and new data.

In other instances, such as quiring, there are several inconsistencies among multi-hands, where not all hands are fully described. These situations limit our approach to infer the data since there is a high level of uncertainty. However, in some cases the information could be completed based on the data provided for the other hands.

That is, the corpus obtained showed us the type information collected and provided great insights to consider in the future when applying machine learning techniques. Most features are categorical, which needs to be addressed when applying models that are based on statistical measures. More technically, the libraries used for unsupervised learning had some limitations, such as the K-Means method where it was possible to use different distance metrics in order to adjust the accuracy.

Moreover, although still ongoing work, the visualization of concept trees and association rules built with Formal Concept Analysis did not show to experts the ideal design for an easy knowledge extraction. Some key design principles are needed to better represent truth, make it clearer and bottom line, understandable. Therefore, information should present clearly and free of expectations of how the data should be organized. It should be presented "as is". One raised difficulty from expert was the unclear way the concepts were presented, as seen in Figure 5.49 for 15th century, whereas the truth within the data, primarily its high level and mid-level associations, should be expressed with clarity.

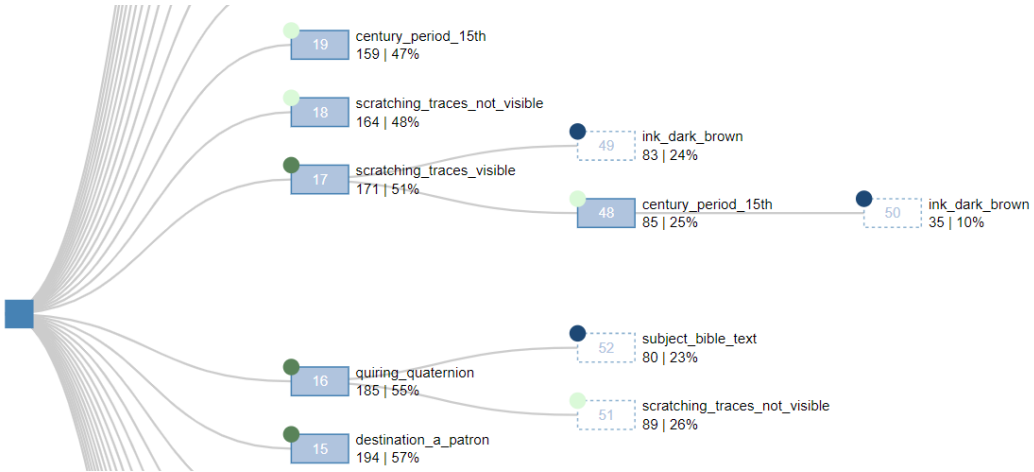

Figure 5.49 - Evidence raised on how unclear is having repeated nodes in different levels

Unlike Formal Concept Analysis, we were not able to improve the accuracy of unsupervised learning algorithms, even for hierarchical clustering, the accuracy did not achieve more than 60%. Nonetheless, further research to better fitting these algorithms to

perform well on categorical datasets should be possible and, more important, to be able to extract information from the clusters which we were not able to do.

5.3 Further Work

Now that we have shown how visualization can help the experts the necessary tools to have better insights of Jewish collections and how these manuscripts behave between each other, we should expand our *corpus* to include early printed books and manuscript decoration metadata. This would allow experts, as said earlier, to study the transition between manuscripts and printed books.

However, increasing the *corpus* in both available features and number of records we need to do a deeper study on Formal Concept Analysis algorithm to address some scalability concerns as well as to improve the designs of the concept trees visualizations. A prototype should be developed alongside with experts' expectations to make the exploration of associations clearer and more truthful. For bigger datasets, the concepts tree is very big, generating more than a million concepts. Therefore, the should be easily and clearly explored. This recent work using Formal Concept Analysis should, therefore, be better studied to build a recommendation system of interesting visualizations in order to avoid expert bias and raise new unexpected results.

The method of data cleaning and treatment, once stable in its rules, should be made automatically or partially manual. This step is still a largely time-consuming task.

BIBLIOGRAPHY

Abbot, D. (2004). *Applied predictive analytics: Principles and techniques for the professional data analyst* (Vol. 11). John Wiley & Sons, Inc.

Ahmad, A. (2007). A k -mean clustering algorithm for mixed numeric and categorical data. *Data and Knowledge Engineering*, *63*, 503–527. https://doi.org/10.1016/j.datak.2007.03.016

Ali, S. M., Gupta, N., Nayak, G. K., & Lenka, R. K. (2016). Big data visualization: Tools and challenges. *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, 656–660. https://doi.org/10.1109/IC3I.2016.7918044

Andrews, S., & Orphanides, C. (2010). Analysis of large data sets using formal concept lattices. *CEUR Workshop Proceedings*, *672*, 104–115.

Bailey, J., & Pregill, L. (2014). Speak to the Eyes: The History and Practice of Information Visualization. *Journal of the Art Libraries Society of North America*, *33*(2). Retrieved from http://www.jeffersonbailey.com/speak-to-the-eyes-the-history-and-practice-of-information-visualization/

Bausi, A., Borbone, P., Briquel-Chatonnet, F., Buzi, P., Gippert, J., Macé, C., … Witakowski, W. (Eds.). (2015). *Comparative Oriental Manuscript Studies*. *European Science Foundation*. COMSt.

Beit-Arié, M. (2012). Hebrew Codicology: Historical and Comparative Typology of Hebrew Medieval Codices based on the Documents of the Extant Dated Manuscripts in Quantitative Approach, 1–117. Retrieved from http://web.nli.org.il/sites/NLI/Hebrew/collections/manuscripts/hebrew_codicology/Pages/default.aspx

Beit-Arié, M. (2017). *SFARDATA The Henri Schiller codicological database of the Hebrew palaeography project , Jerusalem*.

Beit-Arié, M. (2018). Hebrew Codicology. Historical and Comparative Typology of Hebrew Medieval Codices based on the Documentation of the Extant Dated Manuscripts Using a Quantitative Approach. Retrieved from http://web.nli.org.il/sites/NLI/Hebrew/collections/manuscripts/hebrewcodicology/

Documents/Hebrew-Codicology-continuously-updated-online-version-ENG.pdf

Berwind, K., Bornschlegl, M. X., Kaufmann, M. A., & Hemmje, M. (2016). Towards a Cross Industry Standard Process to Support Big Data Applications in Virtual Research Environments (Abstract). *Collaborative European Research Conference*, (January), 56–59.

Bhargav, S., & Pawar, M. (2016). A Review of Clustering Methods forming Non- Convex clusters with , Missing and Noisy Data A Review of Clustering Methods forming Non-Convex clusters with , Missing and Noisy Data. *INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING*, (March). Retrieved from https://www.researchgate.net/publication/299804393%0AA

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D3data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2301–2309. https://doi.org/10.1109/TVCG.2011.185

Buzmakov, A. (2015). *Formal Concept Analysis and Pattern Structures for mining Structured Data*. Universit´e de Lorraine. Retrieved from https://hal.inria.fr/tel-01751818v2

Chandna, S., Rindone, F., Dachsbacher, C., & Stotzka, R. (2016). Quantitative exploration of large medieval manuscripts data for the codicological research. In *IEEE 6th Symposium on Large Data Analysis and Visualization* (pp. 20–28). https://doi.org/10.1109/LDAV.2016.7874306

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0*. *CRISP-DM Consortium*. https://doi.org/10.1109/ICETET.2008.239

Clark, M. (2014). The Archaeology of the Book: Formulating Analytical Research Questions. *E-Conservation Journal*, (2), 10–16. https://doi.org/10.18236/econs2.201403

Cole, R., Fritjov, D., Ducrou, J., Eklund, P., & Wray, T. (2018). Showing Context and Relationships in Museum Collections using Formal Concept Analysis. *Proceedings of the 1st SIGKDD Workshop on Data Science for Digital Art History*. Retrieved from http://dsdah2018.blogs.dsv.su.se/files/2018/08/DSDAH2018_paper_3.pdf

Erl, T., Khattak, W., & Buhler, P. (2016). *Big Data Fundamentals*. *Software Quality Professional* (Vol. 18). Prentice Hall. Retrieved from

https://search.proquest.com/docview/1817038432?accountid=12217%0Ahttp://link.periodicos.capes.gov.br/sfxlcl41?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&genre=unknown&sid=ProQ:ProQ%3Atechnology1&atitle=Big+Data+Fundamentals&title=Softwar

Fürnkranz, J., Gamberger, D., & Lavrač, N. (2012). Foundations of Rule Learning. https://doi.org/10.1007/978-3-540-75197-7

Gold, M. K. (2012). *Debates in Digital Humanities*. University of Minnesota Press. Retrieved from http://dhdebates.gc.cuny.edu/book/1

Gold, M. K., & Klein, L. F. (2016). *Debates in the Digital Humanities 2016*. Retrieved from http://dhdebates.gc.cuny.edu/debates?id=2

Golfarelli, M., & Rizzi, S. (2018). From Star Schemas to Big Data: 20+ Years of Data Warehouse Research. In *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years* (pp. 93–107). Springer. https://doi.org/10.1007/978-3-319-61893-7_6

Graham, E. (2017). Introduction: Data Visualisation and the Humanities. *English Studies*, *4217*, 1–10. https://doi.org/10.1080/0013838X.2017.1332021

Grus, J. (2015). *Data Science from Scratch*. O'Reilly.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining - Concepts and Techniques*. *Data Mining: Concepts and Techniques*. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

Hassner, T., Rehbein, M., Stokes, P. A., & Wolf, L. (2013). Computation and Palaeography : Potentials and Limits. https://doi.org/10.4230/DagMan.2.1.14

Hevner, A., & Chatterjee, S. (2010). Design Research in Information Systems. In *Design Research in Information Systems - Theory and Practice* (Vol. 22, pp. 9–23). Springer. https://doi.org/10.1007/978-1-4419-5653-8

Hinrichs, U., Forlini, S., & Moynihan, B. (2016). Speculative Practices: Utilizing InfoVis to Explore Untapped Literary Collections. *IEEE Transactions on Visualization and Computer Graphics*, *22*(1), 429–438. https://doi.org/10.1109/TVCG.2015.2467452

Jänicke, S., & Wrisley, D. (2012). Visualizing Uncertainty: How to Use the Fuzzy Data of 550 Medieval Texts?

Kaplan, F. (2015). A Map for Big Data Research in Digital Humanities. *Frontiers in*

Digital Humanities, *2*(May), 1–7. https://doi.org/10.3389/fdigh.2015.00001

Keim, D., Andrienko, G., Fekete, J., Görg, C., Melançon, G., Keim, D., … Stasko, J. T. (2008). Visual Analytics: Definition, Process and Challenges. In *Information Visualization - Human-Centered Issues and Perspectives* (pp. 154–175). Springer. Retrieved from https://hal-lirmm.ccsd.cnrs.fr/lirmm-00272779

Knaflic, C. N. (2015). *Storytelling with Data*. Wiley.

Lakhal, L., & Stumme, G. (2005). Efficient mining of association rules based on formal concept analysis. *Lecture Notes in Computer Science Volume 3626. Formal Concept Analysis*, 180–195. https://doi.org/10.1007/11528784_10

Lamy, J. B., Berthelot, H., Favre, M., Ugon, A., Duclos, C., & Venot, A. (2017). Using visual analytics for presenting comparative information on new drugs. *Journal of Biomedical Informatics*, *71*, 58–69. https://doi.org/10.1016/j.jbi.2017.04.019

Lin, W., Alvarez, S. A., & Ruiz, C. (2002). Efficient Adaptive-Support Association Rule Mining for Recommender Systems. In *Data Mining and Knowledge Discovery* (Vol. 6, pp. 83–105). https://doi.org/10.1023/A

Liu, B., Hsu, W., & Ma, Y. (1998). Integrating Classification and Association Rule Mining. *KDD-98 Proceedings*.

Marbán, S., Mariscal, G., & Segovi, J. (2009). A Data Mining & Knowledge Discovery Process Model. *Data Mining and Knowledge Discovery in Real Life Applications*, (May 2014). https://doi.org/10.5772/6438

Matos, D. (2017). *Script and Decoration of Late-Fifteenth Portuguese Hebrew Manuscripts: A Digital Approach*. King's College.

NLI. (n.d.). Ktiv. The International Collection of Digitized Hebrew Manuscripts. Retrieved from http://web.nli.org.il/sites/nlis/en/manuscript

Obiedkov, S. (2018). Introduction to Formal Concept Analysis. Retrieved September 1, 2018, from https://www.coursera.org/learn/formal-concept-analysis

Pateiro, T. (2018). *Hidden Codicological Metadata Patterns in Hebrew Books*.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24(3)*(3), 45–78. Retrieved from http://doi.org/10.2753/MIS0742-1222240302

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

Radich, R. (2017). Big Data for Humans: The Importance of Data Visualization. Retrieved January 25, 2017, from http://dataconomy.com/2017/05/big-data-data-visualization/

Schreibman, S., Siemens, R., & Unsworth, J. (2004). *A Companion to Digital Humanities*. Blackwell Publishing Ltd. https://doi.org/10.1111/b.9781405103213.2004.x

Schreibman, S., Siemens, R., & Unsworth, J. (2008). *A Companion to Digital Humanities*. Malden,MA: Wiley-Blackwell.

Sfardata. (n.d.). The Codicological Data-Base of the Hebrew Palaeography Project The Israel Academy of Sciences and Humanities. Retrieved from http://sfardata.nli.org.il

Staron, M., Sahraoui, H., & Telea, A. (2018). Special section on Visual Analytics in Software Engineering. *Information and Software Technology*, *98*, 117. https://doi.org/https://doi.org/10.1016/j.infsof.2018.03.001

Svensson, P., & Goldberg, D. T. (2015). *Between Humanities and the Digital*. MIT Press. Retrieved from https://mitpress.mit.edu/books/between-humanities-and-digital

Theus, M., Chen, C., & Unwin, A. (2008). Handbook of Data Visualization, (June 2014), 0–7. https://doi.org/10.1007/978-3-540-33037-0

Wachowiak, M., Walters, D., Kovacs, J., Wachowiak-Smolkov, R., & James, A. (2017). Visual analytics and remote sensing imagery to support community-based research for precision agriculture in emerging areas. *Computers and Electronics in Agriculture*, *143*(C), 149–164. https://doi.org/10.1016/j.compag.2017.09.035

Wang, J., & Gu, R. (2010). An extended fuzzy k-means algorithm for clustering categorical valued data. *Proceedings - International Conference on Artificial Intelligence and Computational Intelligence, AICI 2010*, *2*, 504–507. https://doi.org/10.1109/AICI.2010.225

Ware, C. (2004). *Information Visualization: Perception for Design: Second Edition*. *Information Visualization: Perception for Design: Second Edition*. Elsevier. https://doi.org/10.1016/B978-1-55860-819-1.X5000-6

Windhager, F., Federico, P., Schreder, G., Glinka, K., Dork, M., Miksch, S., & Mayr, E.

(2018). Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges. *IEEE Transactions on Visualization and Computer Graphics*, *2626*(c), 1–20. https://doi.org/10.1109/TVCG.2018.2830759

Wolf, L., Potikha, L., Dershowitz, N., Shweka, R., & Choueka, Y. (2011). COMPUTERIZED PALEOGRAPHY: TOOLS FOR HISTORICAL MANUSCRIPTS. *Ieee International Conference On Image Processing*, 3606–3609. https://doi.org/10.1109/ICIP.2011.6116481

Wulfman, C. E. (2014). The Plot of the Plot: Graphs and Visualizations. *Journal of Modern Periodical Studies*, *5*(1), 94–109. https://doi.org/10.1353/jmp.2014.0006

Xie, Y., Chenna, P., He, J. (Selena), Le, L., & Planteen, J. (2016). Visualization of big high dimensional data in a three dimensional space. *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies - BDCAT '16*, 61–66. https://doi.org/10.1145/3006299.3006340

Zaki, M. J., & Hsiao, C. J. (2005). Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE Transactions on Knowledge and Data Engineering*, *17*(4), 462–478. https://doi.org/10.1109/TKDE.2005.60

Zhang, L., Stoffe, A., & Behrisch, M. (2012). Visual Analytics for the Big Data Era – A Comparative Review of State-of-the-Art Commercial Systems. In *IEEE Conference on Visual Analytics Science and Technology*. https://doi.org/10.1021/ja0264549

Zhang, S., Zhang, C., & Yang, Q. (2003). Data Preparation for Data Mining. *Applied Artificial Intelligence*, 17:375–381. https://doi.org/10.1080/08839510390219264