Visual analytics of Hebrew manuscripts codicological metadata

Tiago Pateiro
ISCTE-Instituto Universitário de Lisboa
Av. Forças Armadas, 1649-026
Lisboa, Portugal
tiago_miguel_pateiro@iscte-iul.pt

Debora Marques de Matos
Westfälische Wilhelms-Universität
Münster
Johannisstr. 1, 48143 Münster, Germany
debora.matos@uni-muenster.de

Elsa Cardoso
ISCTE-Instituto Universitário de Lisboa and
INESC-ID
Av. Forças Armadas, 1649-026
Lisboa, Portugal
elsa.cardoso@iscte-iul.pt

ABSTRACT

This paper presents the CodicoDaViz research project, developed with the goal of applying data visualisation techniques to the field of codicology. Adding to the multidisciplinary nature of digital humanities (DH), this project brings together a group of experts of DH, business intelligence and computer science. Using Hebrew manuscript data as a starting point, CodicoDaViz proposes an environment for exploratory analysis to be used by Humanities experts to deepen their understanding of codicological data, and to formulate new research hypotheses. In this paper we demonstrate how data visualisation was instrumental in understanding and structuring the dataset. Examples of the dashboards that have been designed (in Tableau) to enable an interactive and ad-hoc exploration of data are also discussed.

CCS CONCEPTS

• **Computer Applications → Arts and Humanities→** Linguistics; • **Information Systems → Information Systems Applications → Types of Systems →** Decision Support

ADDITIONAL KEYWORDS AND PHRASES

Visual Analytics, Digital Humanities, Jewish Book History, Codicology, Material Culture

ACM Reference format:
XXX

1 INTRODUCTION

Visual analytics (VA) 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. VA is a recent research area that combines the skills and knowledge from different disciplines, and is deeply related to decision support and business intelligence (BI) systems. It has been applied to a diverse set of contexts, for instance, in precision agriculture to improve decisions about crops [7], in healthcare to compare drug information enabling a faster integration into practice of new drugs [8], or in software engineering [9].

In the Humanities, albeit still in its early stages, data visualisation 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 [16], 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' [16].

The inter- and multidisciplinary nature of digital humanities (DH), we might argue, is the perfect background for a collaboration between BI and VA with humanities disciplines, and indeed there is a promising partnership already in place. However, and reminding us of the phases of DH, textual materials (specifically, literary and linguistic analyses) are the principal stage of said partnership. Therefore, the project henceforth described, entitled CodicoDaViz, may be a timely exception to that.

CodicoDaViz was developed with the goal of applying data visualisation techniques to the field of codicology. A field of inquiry in the humanities, codicology deals with books as material objects whilst considering the history of each artefact.

This project was developed by a three-person team with backgrounds in BI and data visualisation, computer science, and Jewish and DH studies. Such a partnership arose from a specific research need concerning the transition from manuscript to print of Jewish, or Hebrew, books. More specifically, the substantial amounts of metadata already available regarding the material aspects of books, but also the heterogeneity and dispersion of said data.

A central goal of this project was to categorise, clean, analyse, and visualise the raw metadata that already exists in relevant data sources. As such, in this paper we present the initial results and visual analytics that allow us to gain a deeper insight into the codicological details of Hebrew manuscripts. It focuses on the implementation of an analytical environment that can be used by experts to explore the existent information. The research method applied is of some importance since it demonstrates the high data management standards used in the project, and can also be replicated to other DH projects in book history, with different corpora.

Thus, the remainder of this paper is organized in four sections, as follows: section 2 provides an overview on related work done in DH, and how computational methods such as visualisation tools and/or digital data sources are helping researchers to provide richer content analysis through visualisation and storytelling. Section 3 proposes a framework in the context of

Hebrew manuscripts to address the modelling of metadata for analysis and visualisation, as well as the needed work to achieve an explorable dataset. This is done by bringing a special focus to data visualisation, showing how it can help not only to engage the audience through rich storytelling about these manuscripts, but also how it can help through the entire process of data cleaning (spotting incoherencies in data) as well as providing useful visual analytics to researchers. Finally, we conclude and summarize future work.

2 RELATED WORK

Similar to other fields, digital approaches to Hebrew manuscripts has been primarily in terms of text (authorship identification, linguistic patterns, digital editions, text encoding, and so on) [24]. The most relevant exception is the work on automatic identification of join fragments developed by the Genizah projects [25]. 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 others around the world have made their materials available in an online platform for digital access to manuscripts in many collections of Hebrew materials around the world, known as Ktiv [22], 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 [1, 2]. In many senses, Sfardata [1, 2] 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 [23]. That being said, these tools still lack an 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 palaeographic descriptions). Particularly, the visualisation of uncertainty is still in discussion, and a much pertinent one within the humanities [4]. 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 [12].

Visualisation can follow a similar principle. If "mapping data to visual representations has been used for centuries to reveal patterns, to communicate complex ideas, and to tell stories" [21], current tools bring to the table this new aspect of interaction and iteration with the experts. Whilst visualisation can be a discovery tool, it is primarily a means to refine arguments and illustrate conclusions already drawn [20]. That is, graphic representations such as charts are not the actual data, but an Given the unstructured nature of our data, as further discussed, data cleaning and validation was a crucial stage in this method. interpretation of it to answer a specific research query, even if visualisation allows complex findings to be presented in an informative and engaging way [10]. For instance, radial trees and parallel coordinates seem to have a wide use when exploring high dimensional data [19]. 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. [3] 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 visualise the information in a radial tree or parallel coordinate plot.

Ali et al. [17], explore commercial solutions for big data visualisation such as Tableau, Microsoft Power BI and propose the usage of link/network analysis techniques as useful visualisation 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 visualisation and engaging with an audience should not be overlooked. Windhager F., et. al [18], try to go beyond the traditional approaches to visualisation on grid-based interfaces, and instead explore them as complex and comprehensive information spaces by the means of interactive visualisations in the scope of cultural heritage collections.

3 RESEARCH METHOD

Our approach uses an adaptation of the big data lifecycle methodology proposed by Erl et al. [11]. As shown in Figure 1, the proposed method was adapted to include data visualisation in the lifecycle. This allowed us to explore and explain the data, but to also further clean it and validate it. 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.

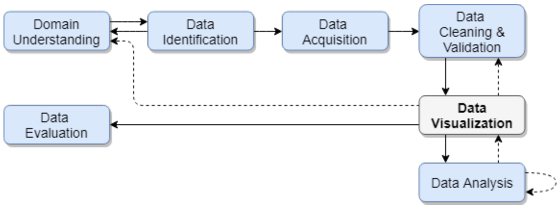


Figure 1: *Research method adapted from big data lifecycle, Erl et al. [11].*

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 refer to the extraction, transformation, and loading (*ETL*) process of gathering data, from one or multiple sources.

It is where we applied transformations, standardised categorical values, and obtained the first insights via data visualisation on

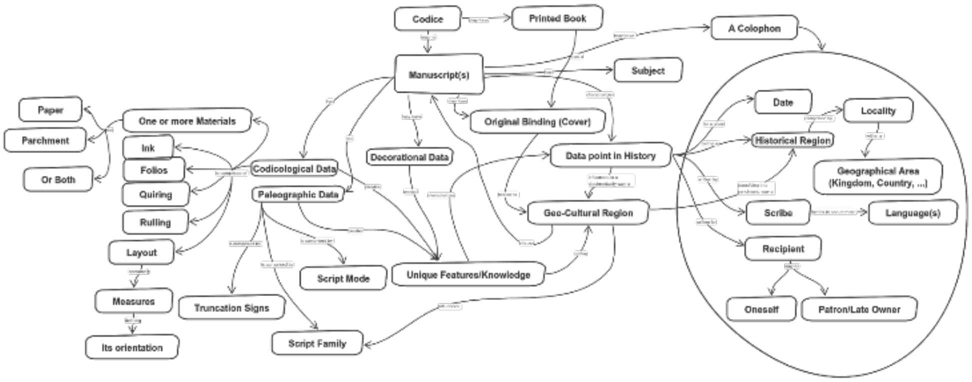


Figure 2 *Concept map covering the relations that describes a manuscript.*

the quality of the dataset. Iteratively, including several instances of data visualisation, it was possible to visualise each state of the dataset, and spot potential domain-specific concerns or data-specific concerns (such as missing values).

Finally, the last step concerned the analysis of the data via visualisation, which allowed us 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.

4 CODICODAVIZ DATASET

4.1 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. Visualisation of codicological data is a useful means of study of said inter-cultural exchanges, with a mutual benefit for Jewish and adjacent book cultures. If by the fifteenth century many codicological practices had already crystalised, the introduction of the printing press posed a new challenge that may also have shaped manuscript production. Our dataset was retrieved from the codicological descriptions of almost all dated Hebrew manuscripts copied until 1540 that have been collected and stored in Sfardata [1, 2]. Our dataset comprises only fifteenth century manuscripts copied between 1400 and 1500, written in Sefardi script. This is a geo-cultural area that corresponds to the Iberian Peninsula and North Africa, but with historical and cultural ties to other regions in the Mediterranean. As further explained, the information in Sfardata [2] is incomplete in many instances, likely due to difficulties experienced during the consultation of the artefacts. Therefore, in several cases data was enriched with access to catalogues. Moreover, manuscripts are listed according to scribal hands and not codicological units, thus enlarging the corpus. This meant an additional step to determine which hand is the most representative, which was far from straightforward.

As part of our interdisciplinary process it was necessary to acquire a basic knowledge of the main concepts of Hebrew codicology and palaeography. This revealed to be crucial in the the next phases of what was a typical framework analysis in data science. Without a clear domain-specific knowledge, the data cleaning, transformation, and processing would have not been possible. Furthermore, the data visualisation and exploration would not be adequate to answer many of the research questions. The following concept map in Figure 2 summarises 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. With this we hope to provide a better understanding of the story that a manuscript tells us.

The proposed work aims to provide a framework for data acquisition, treatment, and visualisation in order to increase its quality and make it suitable for computational tools, thus reducing the communication bottleneck, and foster a mutual understanding across scholarly disciplines.

4.2 Data identification

Regarding the map presented in Figure 2, it is necessary to describe each attribute in use, and its meaning for our datasource. This dataset, in its raw state (from Sfardata [2] database), has 40 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 artefact. 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 multihand copies, as previously mentioned, established a series of rules to identify the most characterising hand.

The defined feature set was grouped in three major sections, all concerning the overall description of the artefacts. These are: history, codicology and a palaeography. As seen in the concept map (Figure 2), 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*" [13]. 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*" [13].

4.2.1 Historical Data. This category answers to questions regarding the who, when, where, what and how of each artefact. These attributes 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 artefact.

4.2.2 Codicological Data. This category comprises the attributes regarding the physical composition of the artefact. 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. Partly quantifiable, some features such as ink are entirely descriptive. Other features such as pricking were not considered due to the lack of substantial information. With regard to format, it was necessary to divide the corpus in sub-groups according to its size and orientation. Finally, here was also included the information on the presence of decoration.

4.2.3 Palaeography Description. Although palaeography is by itself a field of study, the data source includes a general description of the family and mode of writing. 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.

4.3 Data acquisition

Since no public APIs were available to collect the data from the Sfardata, there was an additional process of acquisition. An ETL process was developed to extract the information, applying minimal transformations and process to the data, as seen below in Figure 3.

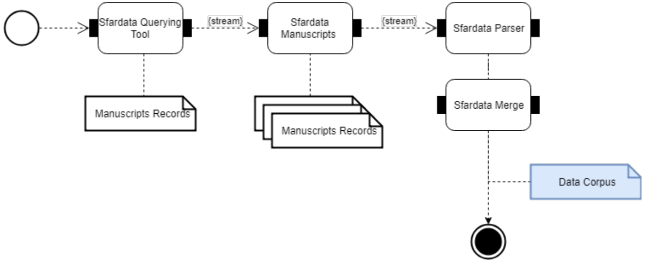
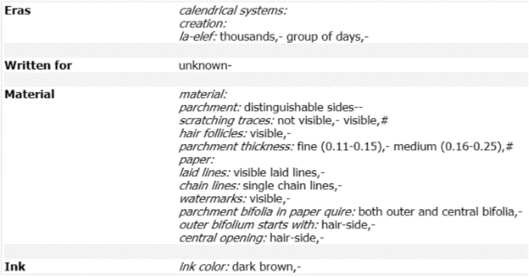


Figure 3: *ETL process fed by Sfardata [1].*

Therefore, applying the query described from the previous section (manuscripts in Sephardi script from 1400 to 1500), we extracted all information concerning the historical, codicological and palaeographic information for each artefact. This information was subsequently merged into a CSV file containing all information. The first step of the ETL 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. Since the platform has no standards regarding the data registered a manual intervention is still required, as described in the following sections.

4.4 Data cleaning and validation

Given the nature of the features of manuscripts, the data held by Sfardata [2] 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 Figure 4.



Eras	*calendrical systems:* *creation:* *la-elef:* thousands,- group of days,-
Written for	unknown-
Material	*material:* *parchment:* distinguishable sides-- *scratching traces:* not visible,- visible,# *hair follicles:* visible,- *parchment thickness:* fine (0.11-0.15),- medium (0.16-0.25),# *paper:* *laid lines:* visible laid lines,- *chain lines:* single chain lines,- *watermarks:* visible,- *parchment bifolia in paper quire:* both outer and central bifolia,- *outer bifolium starts with:* hair-side,- *central opening:* hair-side,-
Ink	*ink color:* dark brown,-

Figure 4: *Sample from Sfardata [2] describing Manuscript's material.*

Consequently, a more in-depth data profiling was limited due to the fact that the available raw data could not be statistically analysed. 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 the next 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.

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 Figure 4). Transformations such as 'some,-decorated,-' within the illumination feature were transformed into 'Partly Decorated', for instance. 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 5 displays the missing values problem within our corpus, in which the more blank a plot bar is the more missing values exist.

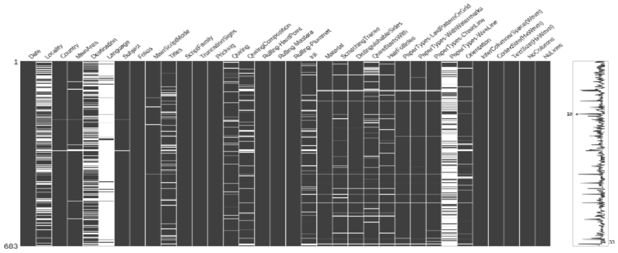


Figure 5: *Missing values plot from several attributes.*

Using visualisation at this stage allowed us to have faster insights on the data and its incoherencies. In fact, after analysing the geographical data, as seen in Figure 6, 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).

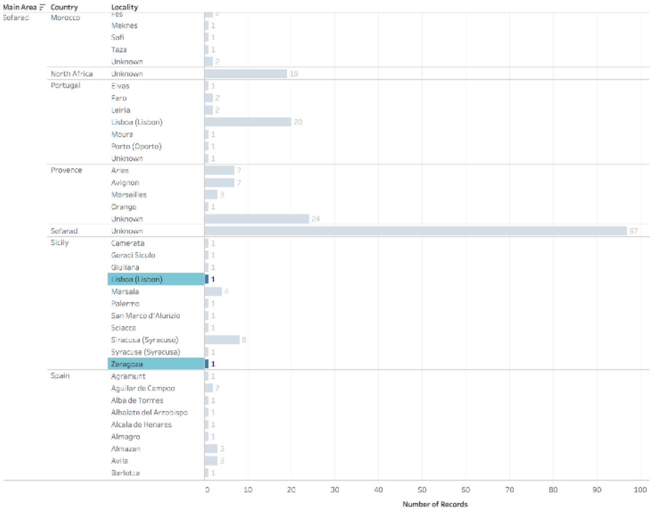


Figure 6: *Data visualization as a key tool to spot wrong geographies.*

Therefore, the second step consisted in reviewing the entire corpus and analyse 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, in order to reduce noise within the corpus for future computational methods such as unsupervised algorithms. This led us to immediately spot data differences within the same manuscript. Therefore, the corpus was splitted into two subsets: unique manuscripts as a whole object, and another with all the multi-hands in each manuscript. 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 characterises a manuscript was built on the following conditions for different scenarios:

If the datasource 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 multihand copies), this indicates that the other is the main hand;

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

In the absence of the two 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, 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.

To enrich the corpus with meaningful features capable of being analysed and visualised, some attributes were computed. Measurement attributes, which are discrete but not categorical, do not provide useful information when visualised. 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:

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

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

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

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

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 7, 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

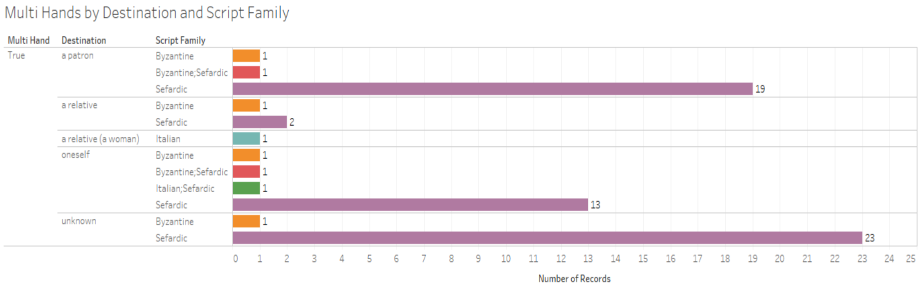


Figure 7 *Visualization used to spot unexpected information requiring another iteration.*

would be more likely to be the case for Italy. This inconsistency challenged us to verify the origin of the manuscript. In Sfardata [2] 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.

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

4.5 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 visualisations 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 119 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 this is 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 [1], 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.

5 Data visualization of Hebrew manuscripts

Data visualisation 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 [6]. In contrast, only in recent years has the potential of data visualisation been fully acknowledged in Humanities. That being said, 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' [4]. Yet current methods and tools for data visualisation also open new research questions and unprecedented amounts of data [14].

Although still scattered in a variety of sources, there are substantial amounts of codicological metadata on Hebrew manuscripts, such as the database Sfardata [1], 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, CodicoDaViz proposes an environment for exploratory analysis to be used by Humanities experts to deepen their understanding of codicological data, and to formulate new research hypotheses. 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 [5] puts it, "it's like hunting for pearls in oysters" [5]. As such, hypotheses must be tested, analysed and visual displays explored in order to achieve an effective visualisation. 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 paper 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, was the chosen method to provide an interactive and intuitive data exploration platform for Humanities experts. Additionally, Tableau Stories can potentially also be used for an explanatory 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. Five perspectives of analysis were defined to address the visualisation of codicological data of manuscripts: (1) material aspects, including writing material, format, layout, quiring and ruling methods; (2) contents and purpose, that is the main subject and destination (for a patron or oneself), including the presence of decoration; (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; (4) geographic analysis, within a geo-cultural region, kingdom and locality; and (5) historical analysis, which takes into account the 'biography' of each artefact, their scribes and commissioners, as well as their incorporation in book collections.

5.1 Codicological dashboards

In this paper two examples are provided. As shown in the first dashboard (figure 8), 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. Following Sirat [15], 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 was employed more frequently, an indication of the sources and, to some extent, the origins of the scribes. As for destination, one observes that commission was the main reason of copy, including one book copied by a groom for his bride, although one should bear in mind that our corpus is only composed of dated manuscripts.

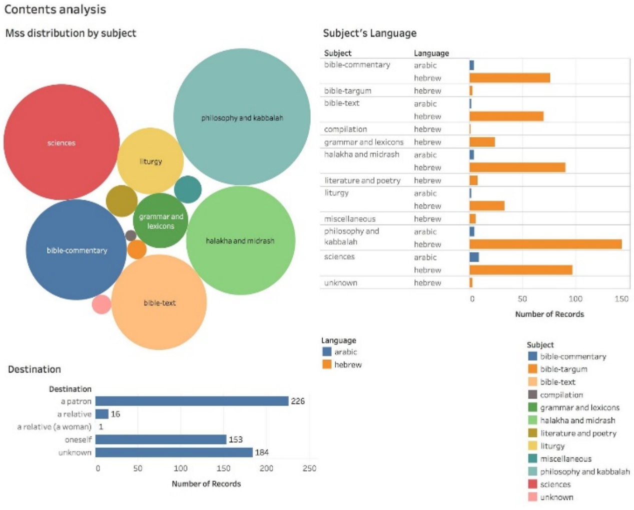
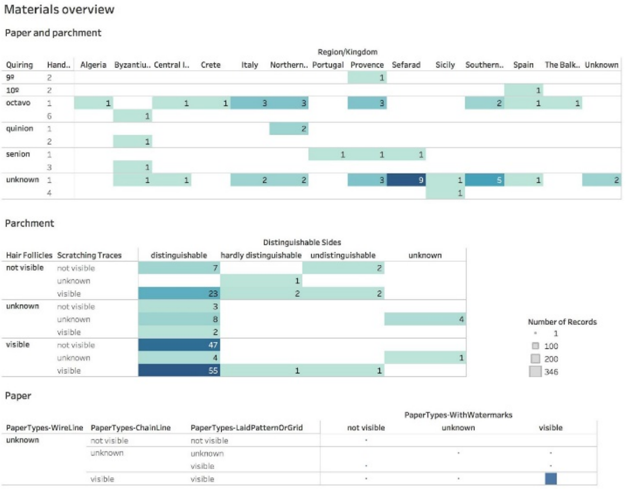


Figure 8: *Dashboard for Content analysis allowing filtering by subject and language.*

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 9). These mixed quires usually include an outer and central bifolium of parchment and the remaining bifolia are in paper. According to Beit-Arié [13], one fifth of all mixed quires appear in Byzantium, early on. Following the data shown in Figure 9 it is possible to conclude that in the fifteenth century this type of quire appears spread into several regions, but particularly into regions adjacent to Byzantium.



Materials overview

Paper and parchment

		Region/Kingdom													
Quiring	Hand..	Algeria	Byzantiu..	Central I..	Crete	Italy	Northern..	Portugal	Provence	Sefarad	Sicily	Southern..	Spain	The Balk..	Unknown
9º	2								1						
10º	2												1		
octavo	1	1		1	1	3	3		3			2	1	1	
	6		1												
quinion	1						2								
	2		1												
senion	1							1	1	1					
	3		1												
unknown	1		1	1		2	2		3	9	1	5	1		2
	4										1				

Parchment

		Distinguishable Sides			
Hair Follicles	Scratching Traces	distinguishable	hardly distinguishable	undistinguishable	unknown
not visible	not visible	7		2	
	unknown		1		
	visible	23	2	2	
unknown	not visible	3			
	unknown	8			4
	visible	2			
visible	not visible	47			
	unknown	4			1
	visible	55	1	1	

Number of Records
· 1
□ 100
□ 200
□ 346

Paper

			PaperTypes-WithWatermarks		
PaperTypes-WireLine	PaperTypes-ChainLine	PaperTypes-LaidPatternOrGrid	not visible	unknown	visible
unknown	not visible	not visible	·		
	unknown	unknown		·	·
		visible	·		·
	visible	visible	·	·	■

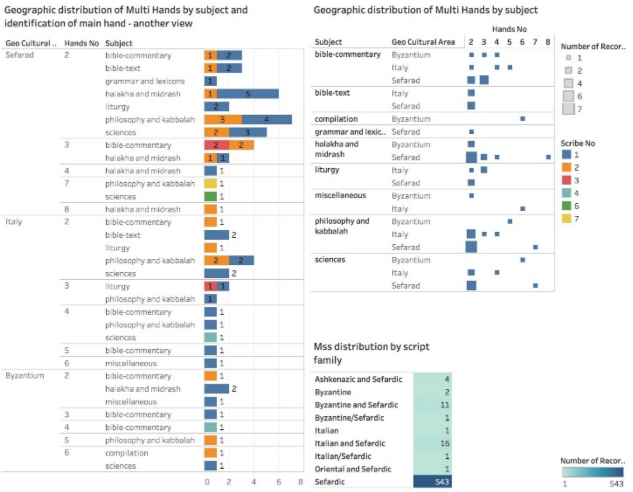
Figure 9: *Dashboard for Material Overview with a drill-down by material type.*

5.2 Data analysis

One of the most significant analysis concerns the multi-hands records. As shown in Figure 10, 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. Whilst 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 8. Finally, our process highlighted the collaboration between various script families, which in turn materialise the mobility of scribes.

6 Conclusions and further work

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 visualisation 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. This paper described the initial results from the research project CodicoDaVis, that aims to apply visual analytics techniques to codicological metadata. The pilot study described focuses on a corpus of Hebrew manuscripts written in Sefardi script between 1400 and 1500. The multidisciplinarity of the team was one of key factors of this project, covering areas from BI and data visualisation, machine learning, to Jewish and DH studies.



Ashkenazic and Sefardic	4
Byzantine	2
Byzantine and Sefardic	11
Byzantine/Sefardic	1
Italian	1
Italian and Sefardic	16
Italian/Sefardic	1
Oriental and Sefardic	1
Sefardic	543

Figure 10: *Visualization with geographic information of multi-handed manuscripts analyzing the subject, script and the number of hands.*

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), with different levels of data aggregation: geo cultural area | region/kingdom | current country | locality.

The BI and VA reasoning, with clear dimension of analysis and metrics, enabled the rapid development of data visualisations 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 visualisations. 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. Application of data mining and machine learning algorithms will be done in the near future to further enhance the structure of the data in the corpus and to identify hidden patterns in dimensions that may be of interest to experts. Unsupervised learning algorithms, cluster and feature analysis can be used to identify unknown patterns in data. 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.

REFERENCES

[1] B.A. Malachi. 1994. SFARDATA. The Henri Schiller codicological database of the Hebrew palaeography project. Gazette du livre médiéval, 25 (Autumn 1994), 24-29. DOI: 10.3406/galim.1994.1280

[2] Sfardata. The Codicological Data-Base of the Hebrew Palaeography Project The Israel Academy of Sciences and Humanities. Retrieved from http://sfardata.nli.org.il

[3] S. Chandna, F. Rindone and C. Dachsbacher. 2016. Quantitative exploration of large medieval manuscripts data for the codicological research. In Proceedings of 2016 IEEE 6th Symposium on Large Data Analysis and Visualization, Baltimore, MD, USA.

[4] S. Jänicke and D. Wrisley. 2013. Visualizing Uncertainty: How to Use the Fuzzy Data of 550 Medieval Texts? In Proceedings of 2013 Digital Humanities conference, Lincoln, Nebraska, USA.

[5] C. N. Knaflic. 2015. Storytelling with data. Wiley, New Jersey

[6] C. Ware. 2004. Information visualization: perception for design (2nd ed.). Morgan Kaufmann Publishers, Elsevier, San Francisco

[7] M.P. Wachowiak, D.F. Walters, J.M. Kovacs, R. Wachowiak-Smolíková, A.L. James. 2017. Visual analytics and remote sensing imagery to support community-based research for precision agriculture in emerging areas. Computers and Electronics in Agriculture, vol. 143, 149-164

[8] J.B. Lamy, H. Berthelot, M. Favre, A. Ugon, C. Duclos, A. Venot. 2017. Using visual analytics for presenting comparative information on new drugs. Journal of Biomedical Informatics, vol. 71, 58-69

[9] M. Staron, H. Sahraoui, A. Telea (Guest editors). 2018. Special section on Visual Analytics in Software Engineering, Information and Software Technology, vol. 98, 117

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

[11] T. Erl, Wajid Khattak, and Paul Buhler. 2016. Big Data Fundamentals. 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

[12] T. Hassner, Malte Rehbein, Peter A. Stokes, and Lior Wolf. 2013. Computation and Palaeography : Potentials and Limits. 2, 14–35. DOI:https://doi.org/10.4230/DagMan.2.1.14

[13] M. Beit-Arié. 2018. Hebrew Codicology: Historical and Comparative Typology of Hebrew Medieval Codices based on the Documents of the Extant Dated Manuscripts in Quantitative Approach. (2018). Retrieved from http://web.nli.org.il/sites/NLI/Hebrew/collections/manuscripts/hebrewcodicology/Documents/Hebrew-Codicology-continuously-updated-online-version-ENG.pdf

[14] F. Kaplan. 2015. A Map for Big Data Research in Digital Humanities. Front. Digit. Humanit. 2, May (2015), 1–7. DOI:https://doi.org/10.3389/fdigh.2015.00001

[15] C. Sirat. 1983. L'édition des textes philosophiques médiévaux, Questions de méthodologie. Da'at 10, (1983), 3–13.

[16] E. Graham. 2017. Introduction: Data Visualisation and the Humanities. English Stud. 98, 5 (2017), 449–458. DOI:https://doi.org/10.1080/0013838X.2017.1332021

[17] Syed Mohd Ali, Noopur Gupta, Gopal Krishna Nayak, and Rakesh Kumar Lenka. 2016. Big data visualization: Tools and challenges. 2016 2nd Int. Conf. Contemp. Comput. Informatics (2016), 656–660. DOI:https://doi.org/10.1109/IC3I.2016.7918044

[18] Florian Windhager, Paolo Federico, Gunther Schreder, Katrin Glinka, Marian Dork, Silvia Miksch, and Eva Mayr. 2018. Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges. IEEE Trans. Vis. Comput. Graph. 14, 8 (2018), 1–1. DOI:https://doi.org/10.1109/TVCG.2018.2830759

[19] Uta Hinrichs, Stefania Forlini, and Bridget Moynihan. 2016. Speculative Practices: Utilizing InfoVis to Explore Untapped Literary Collections. IEEE Trans. Vis. Comput. Graph. 22, 1 (2016), 429–438. DOI:https://doi.org/10.1109/TVCG.2015.2467452

[20] Clifford E. Wulfman. 2014. The Plot of the Plot: Graphs and Visualizations. J. Mod. Period. Stud. 1, (2014), 94–109.

[21] Jefferson Bailey and Lily Pregill. 2014. Speak to the Eyes: The History and Practice of Information Visualization. Art Doc. J. Art Libr. Soc. North Am. 33, 2 (2014), 168–191.

[22] Ktiv. The International Collection of Digitized Hebrew Manuscripts. Retrieved from http://web.nli.org.il/sites/nlis/en/manuscript

[23] Alessandro Bausi, Pier Borbone, Françoise Briquel-Chatonnet, Paola Buzi, Jost Gippert, Caroline Macé, Marilena Maniaci, Zisis Melissakis, Laura Parodi, and Witold Witakowski (Eds.). 2015. Comparative Oriental Manuscript Studies. COMSt.

[24] M.K. Gold and L.F. Klein. *Debates in the Digital Humanities 2016.*, 2016. Print. http://www.worldcat.org/title/debates-in-the-digital-humanities-2016/oclc/928613280&referer=brief_results

[25] L. Wolf, L. Potikha, N. Dershowitz, R. Shweka and Y. Choueka. (2011) Computerized Paleography: Tools for Historical Manuscripts. 18th IEEE International Conference on Image Processing. 3545-3548.