informatics

MDPI

Article

Genealogical Data Mining from Historical Archives: The Case of the Jewish Community in Pisa

Angelica Lo Duca [1,*], Andrea Marchetti [1], Manuela Moretti [1], Francesca Diana [2], Mafalda Toniazzi [2] and Andrea D'Errico [1]

[1] Institute of Informatics and Telematics, National Research Council, 56124 Pisa, Italy; andrea.marchetti@iit.cnr.it (A.M.); manuela.moretti@iit.cnr.it (M.M.); andrea.derrico@iit.cnr.it (A.D.)
[2] Department of Civilisation and Forms of Knowledge, University of Pisa, 56100 Pisa, Italy; dianafrancescavalentina@gmail.com (F.D.); mafalda.toniazzi@cise.unipi.it (M.T.)
* Correspondence: angelica.loduca@iit.cnr.it

Abstract: The Jewish community archive in Pisa owns a vast collection of documents and manuscripts that date back centuries. These documents contain valuable genealogical information, including birth, marriage, and death records. This paper aims to describe the preliminary results of the Archivio Storico della Comunita Ebraica di Pisa (ASCEPI) project, with a focus on the extraction of data from the Nati, Morti e Ballottati (NMB) Registry document in the archive. The NMB Registry contains about 1900 records of births, deaths, and balloted individuals within the Jewish community in Pisa. The study uses a semiautomatic pipeline of digitization, transcription, and Natural Language Processing (NLP) techniques to extract personal data such as names, surnames, birth and death dates, and parental names from each record. The extracted data are then used to build a knowledge base and a genealogical tree for a representative family, Supino. This study demonstrates the potential of using NLP and rule-based techniques to extract valuable information from historical documents and to construct genealogical trees.

Keywords: entity extraction; digital manuscripts; digital humanities; genealogical tree

Citation: Lo Duca, A.; Marchetti, A.; Moretti, M.; Diana, F.; Toniazzi, M.; D'Errico, A. Genealogical Data Mining from Historical Archives: The Case of the Jewish Community in Pisa. *Informatics* **2023**, *10*, 42. https://doi.org/10.3390/informatics10020042

Academic Editor: Roberto Theron

Received: 21 March 2023
Revised: 27 April 2023
Accepted: 9 May 2023
Published: 11 May 2023

1. Introduction

The Jewish presence in Pisa, already strong in the Middle Ages, saw incredible growth thanks to the Privilegi de' mercanti levantini e ponentini issued by the Grand Duke of Tuscany Ferdinando I in 1591 and 1593. In the following two centuries, it became a multifaceted reality, composed of Italian, Sephardi, Ashkenazi Jews, and conversos. The local community was characterized by the absence of a ghetto and was connected, from an economic and cultural point of view, with the major commercial centers of the Mediterranean, the Levant, Northern Europe, and South America. The archive of the local Jewish community constitutes a very rich heritage for the history of Judaism and a bright example of building a community identity through internal documentation. The archive of the Jewish community in Pisa owns a wealth of documents and manuscripts dating back centuries. These documents and manuscripts contain various genealogical information, including birth, marriage, and death records. For many members of the Jewish community, genealogy is an important part of their identity and culture. Not only does it provide information about where a person's ancestors came from, but it also helps to preserve the culture by providing a tangible link between generations [1].

This paper describes the preliminary results of the Historical Archive of the Jewish Community in Pisa (ASCEPI) project [2], which aims at enhancing the documents contained in the archive of the Jewish community in Pisa. This paper describes the ASCEPI project from two points of view: humanistic and IT. From a humanistic point of view, the paper describes the historical context and the organization of the ASCEPI archive. From an IT point of view, the paper describes a case study with a procedure followed to extract the

entities from one of the archive documents and organize them in a genealogical tree. The IT procedure focuses on (Figure 1): (a) digitizing and transcribing the archive manuscripts, (b) mining information from the manuscripts and organizing it in a knowledge base, (c) making the digital copy of the manuscripts and the extracted information available on the web, and (d) building a genealogical tree with the information contained in the knowledge base.

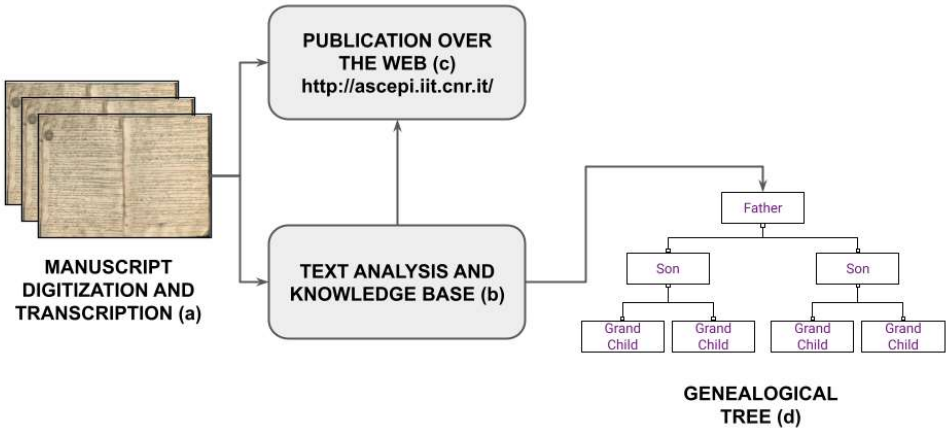


Figure 1. The workflow described in this paper.

This paper focuses on the procedure used to mine information from one of the archive's documents, Registro Nati, Morti e Ballottati (NMB Registry, for short), extract entities, and organize them to build a genealogical tree. The NMB Registry contains about 1900 records about births, deaths, and balloted people of the Jewish community in Pisa. It was compiled from 1749 to 1854 and is written partially in Italian and partially in Hebrew.

The extracted information includes personal data about people in that registry, such as name, surname, birth and death dates, the father's name, and so on. A knowledge base related to the people in the NMB Registry was built, and some statistics were calculated. In addition, a genealogical tree of a representative family, Supino, was built. The next step will move toward constructing a genealogic tree of the whole Jewish community in Pisa.

To extract data from the NMB Registry, a semiautomatic pipeline was defined, which includes a preliminary phase of digitization and transcription of the manuscript. In the second phase, a combination of Natural Language Processing (NLP) techniques and rule-based techniques were applied to extract information from each record of the archive. In the final phase, all the results were published on the web and were made available on the website of the Jewish community archive.

This paper mainly has three objectives. Firstly, it aims to enhance the cultural heritage of the Jewish community in Pisa, making it available on the web. Secondly, the paper defines a semiautomatic procedure for extracting entities from manuscript transcriptions involving domain experts to achieve 100% accuracy. Finally, the paper aims to demonstrate that it is possible to construct a family tree starting from the entities extracted from the archive.

With respect to the state of the art (Section 6), the work described in this paper proposes a semiautomatic procedure that combines an automatic model and the control of a domain expert to achieve a precision, recall, and accuracy of 100%. In practice, the model is fed with the output of a domain expert and is run many times until it converges, i.e., all the evaluated metrics are equal to 100%. The novelty of the proposed model consists in evaluating the model's goodness not in terms of precision, recall, and accuracy but in terms of rounds required by the model to converge. In addition, this paper is a case study of a single archive that adds nuance to the understanding of the challenges confronting the automated extraction of genealogical records.

The paper is organized as follows. To understand the cultural heritage owned by the Jewish community in Pisa better, first, the paper describes the historical background

related to the Jewish presence in Pisa (Section 2). This section summarizes the historical period of reference relating to the manuscripts considered in this paper. Section 3 gives an overview of the content of the archive of the Jewish community in Pisa. The article's main content starts in Section 4, where the data extraction procedure from the NMB Registry, the construction of the database, and the publication of the manuscripts and their transcription on the web are described. Section 5 describes the case study of the Supino family tree, built starting with the data contained in the knowledge base. Section 6 describes the related literature. Related works are moved toward the end of the paper to anticipate the content of this study immediately after the historical background. Finally, Section 7 describes conclusions and future work.

2. The Jewish Presence in Pisa

One of the first pieces of evidence of the Jewish presence in Pisa, dating back to 1160, comes from the traveler Benjamin of Tudela, who in his report tells of twenty coreligionists without mentioning their origin or their economic activities [3]. The arrival of the new settlers in the sixteenth century and, in particular, the invitation from the government authorities in 1354 [4] to settle in the city (to which not only Roman Jews and Italian Jews in general responded, but also Jews from other Mediterranean countries) show that the testimonies begin to follow one another with some continuity [5].

At the beginning of the 15th century, Vitale di Matassia de Synagoga arrived in the city from Rome. He was the progenitor of one of the most important families of the Italian Jewish Renaissance, which later adopted the surname "da Pisa" [5]. Having obtained the monopoly of charitable activity, in 1407, he rented from the Opera del Duomo a building located in the chapel of S. Margherita, next to the Campano Tower, formerly owned by the Christian banker Parasone Grasso and where, before 1395, Daniele di magister Melli da Bertinoro, again representing a society of coreligionists, established a bank. The family from Pisa bought the building in 1466, in which there was also a large room used as a synagogue (at least until 1570–1571) and known as casa delli ebrei until the end the 16th century [5].

From 1548, at the invitation of Cosimo de' Medici, some "new Christians" of Portuguese origin arrived in the city, while in the 1560s, Iberian and Levantine presences are attested. The Jewish bank, however, remained in the hands of the da Pisa family, which was linked to the da Rieti family. In 1550–1570, the Jewish population amounted to about 100 units (compared to 100,000 inhabitants) [6].

The opening of the ghettos of Florence and Siena and the related prohibitions (1570–1571) forced the Rieti family to close the credit bank; the Levantines continued to enjoy the privileges granted them by the Italian group in 1551, but there was a temporary decline. At about the same time, the Tridentine Creed was imposed as a condition for admission to the university, but in the 1590s, thanks to the privileges granted by Ferdinand I, Jews were again able to apply to study medicine.

In 1591 and 1593, Ferdinand I invited merchants of all origins to settle in Pisa (and Livorno), especially Levantines and Marrans. In 1595, the invitation was extended to Jews expelled from the Duchy of Milan [6]. This led to the flourishing of cloth, leather, soap, and wax factories and the introduction of cotton processing.

The Levantine presence led to the establishment of the first ritual Sephardic synagogue, housed in Palazzo Lanfranchi, which was probably opened for worship in 1591, when the ancient Sefer Torah, ceded to Florence twenty years earlier, was returned to Pisa. In 1594, however, a more modest building was rented in Via Palestro, later embellished with expensive carpentry, leather upholstery, and gilding, and finally purchased in 1647.

The Grand Ducal Privileges, published in 1593 and known as the Livornine, provided that the Pisan Jews, and perhaps other groups of Levantines and Marrans, received more specific guarantees regarding the right to leave Tuscany for Islamic lands, permission to export Jewish books, exemption from customs duties on all movable property, and recognition of the Massari of the Jewish community as judges in civil and criminal cases

involving only Jews. The Massari were now also given the right to "vote," that is, to admit new members to the community by secret ballot. As for conversions, it was declared forbidden to baptize minors without the consent of their families, while the latter were allowed to visit the baptized in the house of the catechumens.

The figure of Massari, whose office lasted a year, had been created shortly before the privileges of 1591 came into force. From a document, they animated a governing body (Ma'amad) [6]. They were probably elected by the plenary assembly of the Levantine merchants already present in Pisa. Among the duties of the Massari was the election of treasurers, who in turn were responsible for the collection and management of funds intended for a whole range of activities: from social assistance to the maintenance of religious structures and religious education to visiting the sick, ransoming Jewish slaves, and helping communities in the Holy Land.

The seventeenth century saw a demographic growth of the Pisan settlement: a census of 1613 showed 93 Jewish families for a total of 441 people, which decreased in 1622 to 394 (probably due to immigration to Livorno); however, in the census of 1643, the registration of 75 families (for a total of 348 people) shows that 75% of them were Sephardic; 24% Italian, mostly Roman origin; and 1% percent of origin uncertain [7].

The eighteenth century seemed to pass without major upheavals, except for 1787, when there was an episode of intolerance against three Jews and two Muslims from North Africa. On their way from Livorno to Pisa, they stopped to look at the sculptures on the doors of the Duomo, and their behavior was interpreted by some bystanders as disrespectful toward the sacred images. For this reason, the victims were surrounded, insulted, and beaten while the crowd cheered the violence against the "Turks" and "Jews", but they managed to escape [8].

The Jewish presence revitalized Pisa not only in economic terms: throughout history, the number of scholars and scientists multiplied and contributed to enriching the cultural heritage. One thinks, for example, of the relationship between the da Pisa family and the Abravanel family or the figure of Vitale (Yehyel Nissim) by Simone di Vitale da Pisa (b. 1493?—d. before 1572), a profound connoisseur of the Sacred Scriptures, the Qabbalah, philosophy, and astronomy [9].

After the seventeenth and eighteenth centuries then followed the relevant testimonies about the stay of important rabbis in Pisa, including Yitzhaq Uziel (1602–1613), Azaryah Picho (1610–1627), Benyamin Babli and his disciple Abraham Sulema, Yehudah Sabibi (1657), Natan Shapiro, and Rafael Meldola [6,10]. Finally, of particular interest is that in 1785, Gad di Samuele Foa, a member of the Sabbioneta family of printers, founded a Jewish printing house there, the Fua (Foa) printing house [11].

3. Overview of the Archive of the Jewish Community

The archive of the Jewish community in Pisa is one of those hidden cultural assets that few people can consult and study directly. Thanks to the Internet and modern technologies, it is possible to consult at least part of it online. In addition, thanks to modern artificial intelligence techniques, it is possible to facilitate the scholar's work in extracting information from texts.

This archive contains about 6000 documents, organized in a three-level hierarchy: 121 fonds, 1634 titles, and 4269 objects. For example, in the archive, there is the collection "Statutes and Regulations", where there is a correspondence between Mr. De Pas and Chancellor Galligo Isache (see Figure 2).

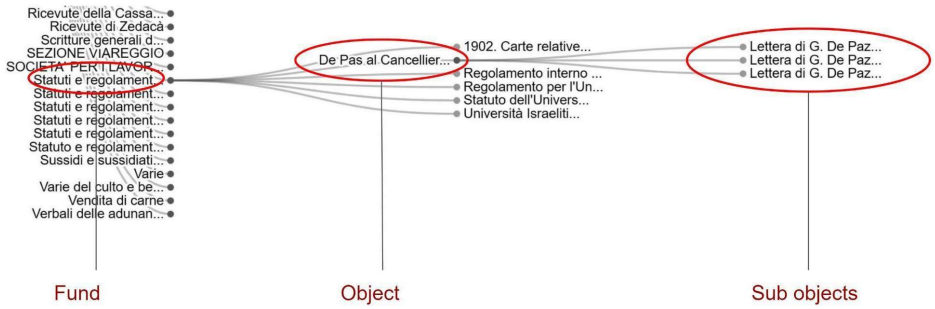


Figure 2. Three letters, present in the archive, are part of an epistolary between Mr. De Pas and Chancellor Galligo Isache.

There is an inventory of the entire archive that was created between 1997 and 2007. This was digitized from 2006 until 2012 using proprietary software called Sesamo [12], which was very popular among Italian archivists at that time. Unfortunately, as often happens, the software ceased to be maintained in 2012, and the electronic use of the inventory was lost. For this reason, at one point, those who frequented the archive could only use a paper version of the inventory. Only in 2020, because of the ASCEPI project, was the inventory recovered and converted into an open format such as CSV. On the ASCEPI website [2], it is possible to browse the inventory and download items as a CSV file. Figure 3 resumes the history of the inventory of the Jewish Archive.

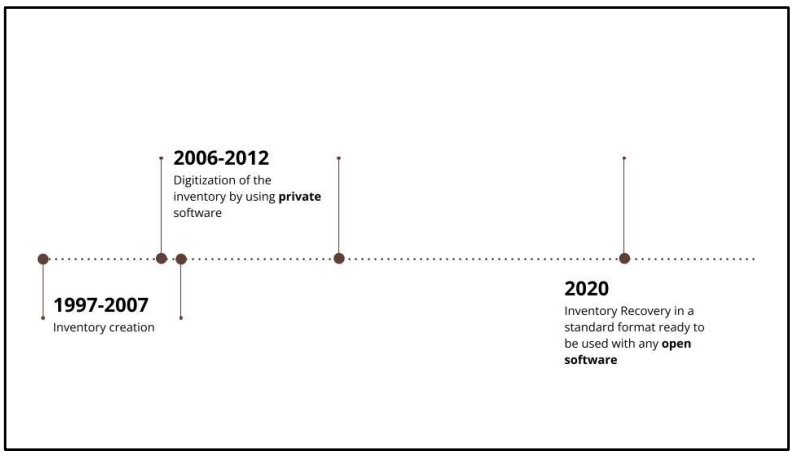


Figure 3. History of the inventory of the archive of the Jewish community in Pisa.

Also, because of the ASCEPI project, the digitization campaign of the 6000 documents contained in the archive has begun. Currently, only a few dozen documents have been digitized, among which, however, is the digitization of the NMB Registry, which consists of 83 pages, all of them transcribed and translated. In the near future, new documents that contain personal information that can increase knowledge of the Jewish community in Pisa will be digitized and transcribed.

4. Entity Extraction from the NMB Registry

Figure 4 shows the workflow of extracting data from the NMB Registry and publishing them over the web. The workflow comprises four phases: digitization, transcription and translation, text analysis and entity extraction, and, finally, publication on the web.

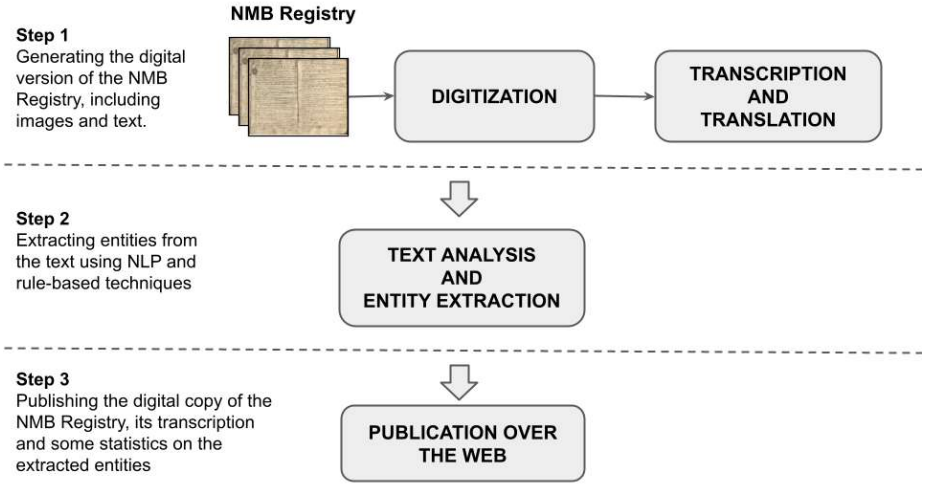


Figure 4. Workflow followed to extract data from the NMB Registry.

4.1. Digitization

The NMB Registry was digitized using a camera with a 50 mm lens, an aperture of 5.6, a shutter speed of about 1/250 s, and a sensitivity of ISO 320. Figure 5 shows an example of a digitized manuscript.

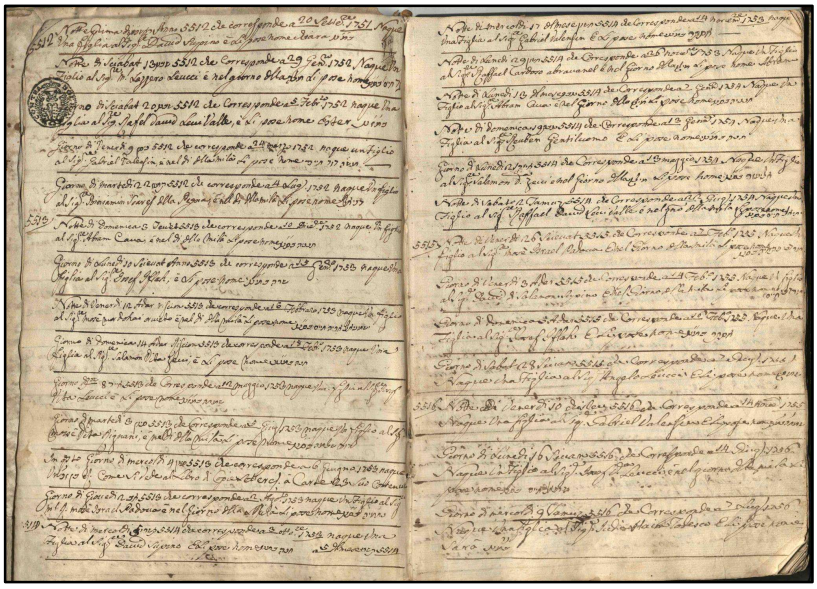


Figure 5. An example of a digitized manuscript of the NMB Registry.

4.2. Transcription and Translation

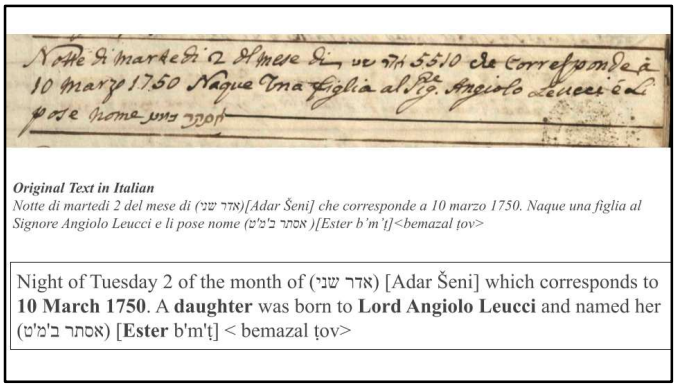
As far as the treatment of the manuscripts to be digitized is concerned, the first thing to determine is whether it is a regestation or a transcription. Regestation in its broadest form was chosen for documents such as the series of Atti Criminali, many of which had gaps or damage that would have made transcription unnecessarily complicated but which, at the same time, did not stand in the way of a correct interpretation of the contents. The NMB Registry, on the other hand, which is in an almost perfect state of preservation, was transcribed in its entirety. The transcription, both of the Italian and Hebrew parts of the text, was first carried out according to the usual standards of documentary editing, resolving acronyms and abbreviations where possible. In the second step, the transcription was revised so that the usual diacritical marks did not interfere with the data extraction by the algorithm. Of course, to preserve the correctness of the transcription as an output of a

historical source, no diacritical marks were changed, added, or removed in the version that can be viewed on the website along with the images.

4.3. Text Analysis and Entity Extraction

The goal of this article is to automatically extract entities for all records contained in the NMB Registry. For this purpose, a model that combines NLP techniques and regular expressions was developed. There are some cases where the applied model cannot extract the information correctly. These cases were handled manually.

The basic idea behind the developed model is that all records in the NMB Registry almost always have the same sentence construction. Figure 6 shows the typical structure of a record.



Original Text in Italian

Notte di martedì 2 del mese di (אדר שני)[Adar Šeni] che corresponde a 10 marzo 1750. Naque una figlia al Signore Angiolo Leucci e li pose nome (אסתר ב'מ'ט)[Ester b'm'ṭ]<bemazal ṭov>

Night of Tuesday 2 of the month of (אדר שני) [Adar Šeni] which corresponds to **10 March 1750**. A **daughter** was born to **Lord Angiolo Leucci** and named her (אסתר ב'מ'ט) [**Ester** b'm'ṭ] < bemazal ṭov>

Figure 6. The typical structure of a record in the NMB Registry. At the top, there is the original record extracted from the manuscript, and in the middle, there is the original transcription in Italian. At the bottom is the English translation.

The sample record contains the following entities: 10 March 1750 (child's birth date), Ester (daughter's name), Lord Angiolo Leucci (father's name), and daughter (child's gender). Extracted information could be used to build family trees, provided that extracted people appear as children and parents.

Figure 7 describes the workflow used to extract entities from the NMB Registry.

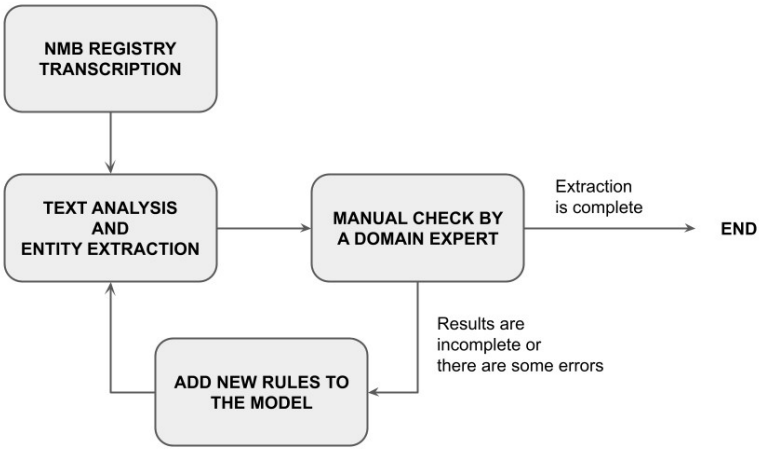


Figure 7. The workflow for entity extraction from the NMB Registry.

In the bottom left part of the figure, there is the NMB Registry. Text analysis and entity extraction (model) are applied to the NMB Registry to mine a table containing the possible candidate entities (Table 1). A domain expert controls the produced table manually. If they find some errors, the model is fed with these errors and then is applied again on the NMB Registry to produce a new table of possible candidates. The domain expert controls the

produced table again. The procedure continues until the table produced by the model does not contain any error.

Table 1. An extract of the table extracted by the model for text analysis and entity extraction.

Son	Date	Gender	Surname	Grandfather	Father
Ribqa	12 October 1749	F	Sezzi		Salamon
Rachel	26 June 1750	F	Supino	Salamon	David
Jacobbe Vita	15 July 1814	M	Soria		Aron

A semiautomatic procedure with a domain expert in the middle was chosen because of the requirement of an accuracy of 100%. In addition, this strategy was feasible since the number of records was reduced. A semiautomatic procedure speeds up the extraction process compared to a manual procedure.

Entity extraction was performed as follows. First, the text was split into records. For each record, part of speech (POS) analysis was performed. Then, the model started with an initial set of mapping rules to extract entities. The mapping rules exploited the sentence structure, which was almost always the same. After running the model, the domain expert identified the errors. Based on the identified errors, new rules were defined to handle those errors. If no rule could be associated with an error, an exception was added to the model to handle that error. This process was executed until the domain expert did not recognize any error.

Extracting entities described in Figure 6 requires a certain number of rounds. A round is a complete run of the model and the domain expert control. For each round, first, the model is run, and then an expert controls the results manually and marks errors. Errors are incorporated in the next round, and the model is run again. Error incorporation involves adding new rules or exceptions to the model to handle errors.

To demonstrate how the model works, a simplified procedure to extract the father entity is illustrated.

1. A manual POS tagging of the first records is conducted, as shown in Figure 8.

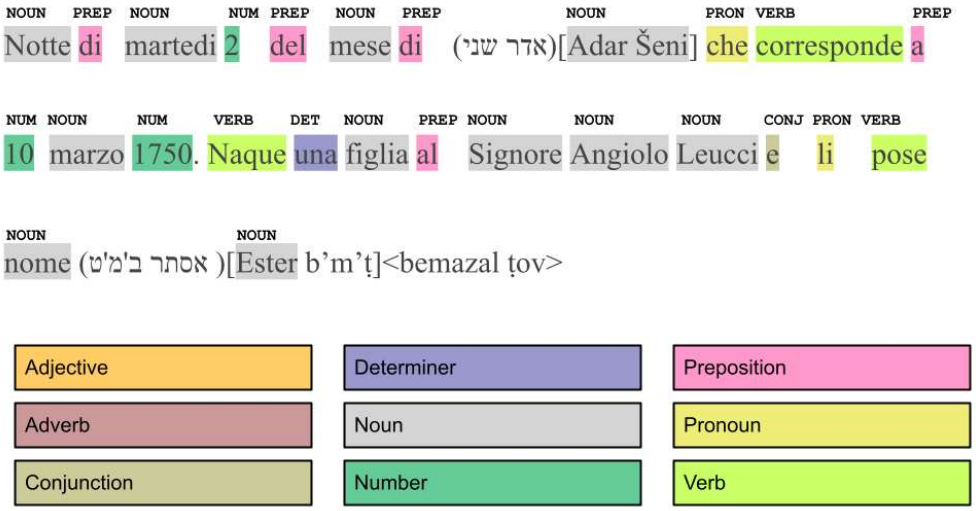


Figure 8. The result of a manual POS tagging of a record in the NMB Registry.

2. Based on the POS tagging, an initial set of rules is defined to extract the father entity, as shown by the following code.

```
first_tokens = ['Signore']

last_tokens = ['e']


pattern_father = [[{'LOWER' : {'IN' : first_tokens}},

                   {'POS':'NOUN', 'OP' : '?'},

                   {'POS':'NOUN', 'OP' : '?'},

                   {'LOWER': {'IN' : last_tokens}} ]]
```

The rule specifies that the string identifying the father's name and surname starts with one of the words contained in the first_tokens list, then it contains some matching part of speech objects, and finally, it ends with one of the words containing in the last_tokens list.

3 The model is run with this initial rule. Then, the domain expert reviews the output and finds some errors. For example, the father contained in the following entity is not extracted because the list of first_tokens does not contain the word Signor:

Il giorno di venerdì primo del mese di Adar primo (sic) 5575 relativo alli 10 febbraio 1815 alle ore 6 mattina cessò di vivere la Signora Ester del Signor Samuel Cardoso Abrabanel moglie del Signore Abramo Cava, e fu sepolta nel solito campo. (תנצב"ה) [t'n'ṣ'b'h'] <t'hay nafsho/ah ṣrurah b'ṣror ha-ḥayyim>

4. In the next round, the rule is updated. In the previous example, this would be accomplished by including the word Signor in the first_token list. Then, the algorithm is run.
5. The process terminates when the father entity is extracted from all the records.

To extract the father, the following final mapping rule was used:

```
first_tokens = ['signore', 'signor', 'a', 'Signore']
last_tokens = ['e', 'nominata', ',', 'nel', 'alle', 'che', 'abitante', 'una', 'nominato',
'povero', 'le', 'algerino', 'al', 'un' ]
pattern_father = [[{'LOWER' : {'IN' : first_tokens}},
                {'POS':'NOUN', 'OP' : '?'},
                {'POS':'PROPN', 'OP' : '+'},
                {'POS':'NOUN', 'OP' : '?'},
                {'POS':'ADP', 'OP' : '?'},
                {'POS':'AUX', 'OP' : '?'},
                {'POS':'PROPN', 'OP' : '*'},
                {'POS':'ADP', 'OP' : '?'},
                {'POS':'PROPN', 'OP' : '*'},
                {'LOWER': {'IN' : last_tokens}} ]]
```

To extract the son, a similar rule was defined, as shown in the following piece of code:

```
first_tokens = ['nome', 'nominata', 'nominato', 'nominò', 'nominolla','circuncisione',
'milà', 'detto']
last_tokens = ['b'', 'nel', 'figlia', ',']
pattern_son = [[{'LOWER' : {'IN' : first_tokens }},
          {'POS':'PROPN', 'OP' : '*'} ,
              {'POS' : 'NOUN' , 'OP': '*'},
           {'LOWER' : {'IN' : last_tokens }}  ]]
```

The code to extract entities was written in Python and used the SpaCy library [13].

After extracting entities from the registry, some statistics on the most frequently used names for both male and female newborns were calculated (Figure 9). The analysis revealed that the top three most frequently used male names were Isache, Josef, and David. For female names, the top three were Ester, Rosa, and Sarà.

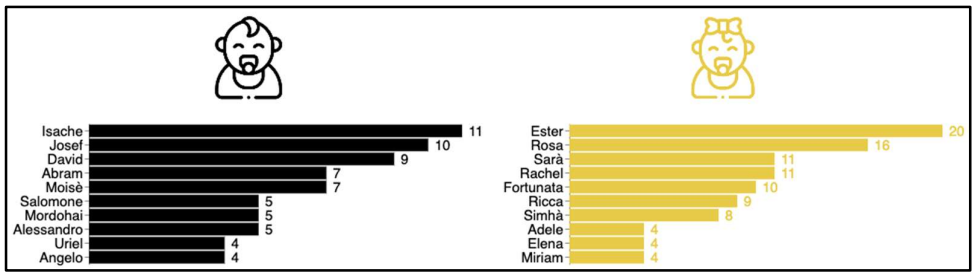


Figure 9. The top 10 most frequent names for newborns.

4.4. Model Evaluation

To evaluate the model, the confusion matrix shown in Figure 10 was used. The confusion matrix uses the following concepts:

- True Positive (TP): The record contains an entity, and the algorithm extracts it correctly.
- True Negative (TN): The record contains no entity, and the algorithm does not extract any entities.
- False Positive (FP): The record contains an entity, but the algorithm does not extract it, or it extracts a wrong entity.
- False Negative (FN): The record contains no entity, but the algorithm extracts an entity.

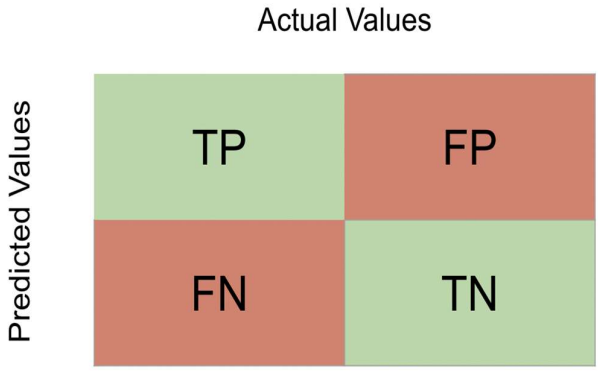


Figure 10. The confusion matrix.

Starting from the confusion matrix, the following metrics were measured for each extracted entity after each round:

- Precision is calculated as TP/(TP + FP) and measures how many predicted positives are true positives.
- Recall is calculated as TP/(TP + FN) and measures how many actual positives are correctly predicted as positives.
- Accuracy is calculated as (TP + TN)/(TP + TN + FP + FN) and measures how many predictions are correct out of all predictions made.

The objective of the evaluation was to calculate the number of rounds required for the model to converge, i.e., achieve a precision, recall, and accuracy of 1. The following entities were considered: father, son, date, and gender. Grandfather was not considered because it was extracted from the father's entity.

For date and gender, the model converged after just one round. For the father entity, the model converged after four rounds. Table 2 shows the measured metrics for the father entity after each round.

Table 2. Measured metrics for the father entity.

Round	Precision	Recall	Accuracy
1	0.9433962264	1	0.9442379182
2	0.9962264151	1	0.9962825279
3	0.9962264151	1	0.9962825279
4	1	1	1

Table 3 shows the measured metrics for the son entity after each round. The model converged after five rounds. It is interesting to note that precision and accuracy decreased from round 2 to round 3. This was probably due to the incorrect incorporation of errors in the model after round 2.

Table 3. Measured metrics for the son entity.

Round	Precision	Recall	Accuracy
1	0.9341085271	1	0.936802974
2	0.969348659	1	0.970260223
3	0.9652509653	1	0.9665427509
4	0.9543726236	1	0.9553903346
5	1	1	1

4.5. Publication over the Web

The publication of the texts and images on the web was made with a WordPress-based application. A web site for the project ASCEPI was created and can be reached at the address http://ascepi.iit.cnr.it/ (accessed on 10 May 2023). The iPages Flipbook WordPress plug-in [14] was used to manage images, which can be shown as carousels or thumbnails, with the possibility of seeing them on a full page, enlarging them, and navigating through them with the mouse. The use of the site is mediated by a drop-down menu, as shown in Figure 11.

Figure 11. The dropdown menu of the project website.

The digitized documents appear under the "DIGITALIZATION" menu item, while the "INVENTORY" item allows you to consult the archive inventory in various ways. The images are owned by the Jewish community in Pisa, and to access them, it is necessary to register by creating an account with a password, which will maintain the right of access also for subsequent times.

5. Case Study: The Supino Family Tree

Starting from the extracted entities, relationships between people can be extracted to generate a family tree. By identifying names, birth dates, and other relevant information, a system can determine the connections between individuals and create a visual representation of their family history.

Table 4 shows the extracted people belonging to the Supino family. Starting from the relationships extracted in Table 2, the genealogical tree shown in Figure 12 was built.

Table 4. Extracted people belonging to the Supino family.

Grand Father	Father	Son	Son Birth Date	Son Gender
Salamon	David	Rachel	26 June 1750	F
-	David	Chiara	20 September 1751	F
-	David	Channah	3 October 1753	F
Salamon	David	Mordechai Haim	14 February 1755	M
-	David	Mosè Haim	10 January 1758	M
Salomon	Isache	Sarà	31 December 1761	F
Salamone	Isache	Scielomo	9 March 1767	M
-	Mordehai Haim	David Haim	4 February 1781	M
Salamon	Moisè	Salamon David	5 May 1786	M
David	Moisè	David	10 October 1786	M
Salamon	Moise	Luna	1 June 1793	F

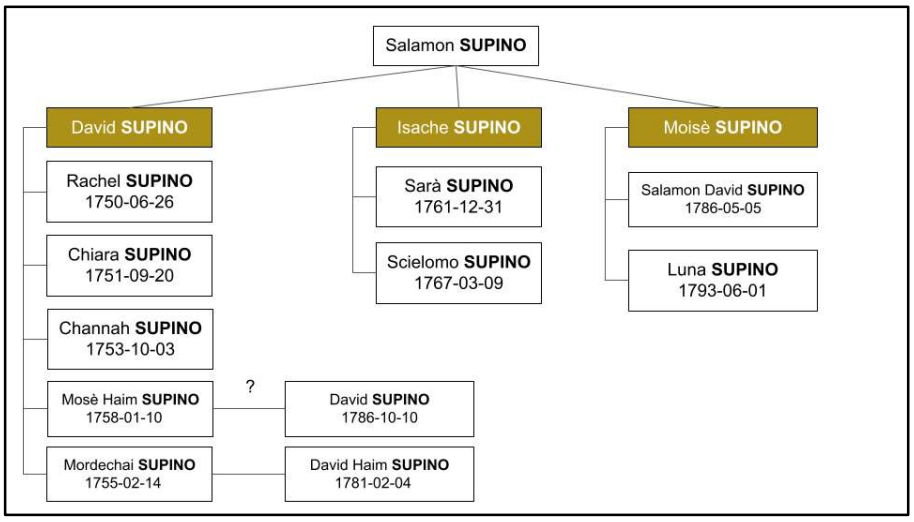


Figure 12. The genealogical tree of the Supino family was extracted from the register NMB.

In reconstructed parental relationships, the only uncertainty that arises regards the relationship between Mosè Haim Supino and David Supino. This is because the name Mosè Haim Supino appears in two forms in Table 4.

Figure 11 illustrates the genealogy of the paternal line only. It depicts the lineage of male ancestors, highlighting the connections between the father and son, grandfathers and grandsons, and so on. The figure does not include any information about the maternal lineage or other relatives beyond the direct paternal line. While the paternal line can provide valuable information about family history, it is just one piece of the puzzle. Exploring the maternal lineage can also uncover interesting details and shed light on ancestors who may have been overlooked in previous research.

6. Related Work

A vast subset of the literature exists on the topic described in this paper, including many efforts to digitize manuscripts and publish them on the web [15–18]. This paper does not pretend to include a detailed review of the current literature on similar work. Instead, the focus of this paper is to compare this paper with other existing Jewish studies.

In general, with respect to the current literature on Jewish studies, this paper proposes a complete case study, which starts from the digitization of manuscripts and concludes with the construction of a family tree. Instead, the existing efforts end with the publication of

digitized manuscripts and their transcriptions on the web. In addition, this paper proposes a methodology that integrates the evaluation of a domain expert in the model workflow to make the model achieve a performance of 100%.

Jewish studies have been a research subject for many years, with scholars focusing on various aspects of Jewish history, culture, and religion [19]. In addition, many works have emerged on Jewish studies and their representation on the web. The use of digital humanities in Jewish studies can be classified into three main categories: (a) digitization of documents and subsequent publication on the web, (b) extraction of entities contained in documents and creation of a knowledge base related to the extracted entities, and (c) construction of genealogical trees related to the extracted entities. The remainder of this section will review works in each category, examining the different approaches to representing Jewish studies in the digital age.

6.1. Digitization of Documents and Website Creation

In recent years, there has been a significant effort to digitize manuscripts and other historical documents in all cultural fields, including Jewish culture. This has involved a concerted effort to preserve and make valuable materials accessible that might otherwise be lost or inaccessible. In particular, many Jewish cultural institutions have invested significantly in digitization efforts, recognizing the importance of preserving and sharing Jewish communities' rich history and heritage worldwide. This has included digitizing ancient manuscripts, historical texts, and other important documents that shed light on the Jewish experience over the centuries.

The Friedberg Jewish Manuscript Society [20] project is one of the most representative projects in this category. It aims to digitize and make Jewish manuscripts from around the world accessible through an online database. The Hebrew Manuscripts Digitisation [21] Project is a joint effort by the British Library and the National Library of Israel to digitize and make Hebrew manuscripts accessible in their collections. The Cairo Genizah Collection [22] is a collection of thousands of Jewish documents and fragments discovered in an attic in Cairo in the late 19th century. The documents from the 9th to the 19th century include everything from religious texts to personal letters and business documents. The collection has been digitized by several institutions, including the University of Cambridge, and is now freely available online. Last but not least is The Judaica Collection [23] at the National Library of Israel, one of the most extensive collections of Jewish literature, history, and culture in the world. The collection includes over five million items, including manuscripts, books, periodicals, photographs, maps, and music recordings.

The Jewish Atlantic World [24] is a project led by Prof. Laura Leibam and was inspired from her research for her book *Messianism, Secrecy, and Mysticism: A New Interpretation of Early American Jewish Life*. At present, the website contains a queryable database, collecting images of everyday objects as well as graves and various artifacts, and testifying to the Jewish presence in many of the key ports in North America and the Caribbean during the Modern Era. The collection is very important in highlighting the history of Jewish families from a global perspective.

6.2. Entity Extraction

While digitizing manuscripts is an important first step in preserving cultural heritage, it is only the beginning of the process. Once digitized, it is essential to extract information from these manuscripts and make them available to researchers and the public. Historical and ancient texts pose many challenges for entity extraction, such as spelling variation, language change, and lack of large corpora. The main techniques to extract entities from historical texts include temporal entity extraction, event extraction, and named entity recognition (NER) [25,26]. In many cases, entity extraction is performed manually [27]. In other cases, automatic techniques for entity extraction are used [28–30]. The main techniques for NER include rule-based approaches, machine-learning-based approaches, and deep-learning approaches [25]. However, using NER on historical documents can be challenging

because the available pretrained models were trained on contemporary datasets [31,32]. With respect to the current literature, this paper combines Natural Language Processing, rule-based techniques, and manual inspection to reach an accuracy of 100%. The trained model could be used to extract entities from similar documents.

6.3. Genealogical Tree

Building family trees is a niche research area. Interesting works can be found in the literature, but they do not focus on Jewish studies specifically. Many works focus on how to represent a genealogical tree visually, including through GeneaQuilts [33], VisFCAC [34], Enhanced Family Tree [35], GenealogyVis [36], Family Metro Map [37], and K-Graphs [38].

Other works focus on other problems related to genealogical trees. Folkman et al. described identifying entities in different genealogical trees through machine-learning techniques [39,40]. Wang et al. proposed DKR, which builds kinship ties (for example, father–child) starting from the people in a photo [41]. Koylu et al. used 92,832 user-contributed family trees to extract social information about the population [42].

7. Conclusions

This study has shown the potential of the ASCEPI project to extract data from historical documents and construct genealogical trees. The NMB Registry document contains valuable information about births, deaths, and balloted individuals within the Jewish community in Pisa. Using a semiautomatic pipeline of digitization, transcription, and NLP techniques, personal data such as names, surnames, birth and death dates, and parental names were extracted from each record. This information was used to build a knowledge base and genealogical tree for a representative family in Pisa by the name of Supino. This demonstrates how powerful tools such as NLP can be used to unlock the wealth of knowledge that is hidden in these ancient documents.

In future work, the data contained in inventory documents, such as marriage records, could be combined to extract the maternal line of a genealogical tree. However, a semi-automatic procedure should be implemented to select the documents of interest from the archive based on the entities extracted. In addition, since the Jewish community is the custodian of the ancient Jewish cemetery in Pisa, information extracted from the graves in the cemetery could be combined with those in the archive. As a result, a web application could be built containing a map of the graves in the cemetery, and for each grave, the family tree of the people contained could be shown [43].

Author Contributions: Introduction, A.L.D. and M.T.; The Jewish Presence in Pisa, M.T. and F.D.; Overview of the Archive of the Jewish Community, A.M.; Entity Extraction from the NMB Registry, A.L.D., A.M., M.T., F.D. and A.D.; Case Study: The Supino Family Tree, A.L.D.; Related Work, A.L.D. and M.M.; Conclusions and Future Work, A.L.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the ASCEPI project, which is funded by CNR.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the Jewish community in Pisa for making the manuscripts in the archive available, and Federico Prosperi and Alessandro Prosperi for their support in implementing the ASCEPI project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ball, R. Visualizing genealogy through a family-centric perspective. *Inf. Vis.* **2017**, *16*, 74–89. [CrossRef]
2. The ASCEPI Project. Available online: http://ascepi.iit.cnr.it/ (accessed on 16 March 2023).
3. Tudelensis, B. *Itinerarium*; Colorni, V., Ed.; Bologna, Italy, (anastatic reprint) 1967; p. 18.

4. Lonardo, P.M. *Gli Ebrei a Pisa*; Doc. VII; Forni: Bologna, Italy, 1982; pp. 40–41.
5. Luzzati, M. *La Casa dell'Ebreo Saggi sugli Ebrei a Pisa e in Toscana nel Medioevo e nel Rinascimento*; Nistri Lischi: Pisa, Italy, 1985.
6. Toaff, T. *La Nazione Ebrea a Livorno e a Pisa (1591–1700)*; Olschki: Firenze, Italy, 1990.
7. Frattarelli Fischer, L. L'insediamento ebraico nella Pisa del '600. *Crit. Stor.* **1987**, *24*, 3–54.
8. Salvadori, R. *Breve Storia Degli Ebrei Toscani IX–XX Secolo*; Le Lettere: Firenze, Italy, 1995; p. 89.
9. Guetta, A. Vita Religiosa ed Erudizione Ebraica a Pisa: Yechiel Nissim da Pisa e la crisi Dell'aristotelismo. In *Gli Ebrei di Pisa (Secoli IX–XX)*; Luzzati, M., Ed.; Pacini Editore: Pisa, Italy, 1998; pp. 45–67.
10. Mortara, M. *Indice Alfabetico dei Rabbini e Scrittori Israeliti*; Sacchetto: Padova, Italy, 1886; p. 38.
11. Amram, D. *The Makers of Hebrew Books in Italy, London*; The Holland Press Limited: London, UK, 1963; p. 397.
12. Grassi, R. Sesamo 4. In *Archivi&Computer: Automazione e Beni Culturali*; Carocci: Rome, Italy, 2003; p. XIII/3. ISBN 884302597X.
13. The SpaCy library. Available online: https://spacy.io/ (accessed on 16 March 2023).
14. The iPages Flipbook Plugin. Available online: https://wordpress.org/plugins/ipages-flipbook/ (accessed on 16 March 2023).
15. Jayanthi, N.; Indu, S.; Hasija, S.; Tripathi, P. Digitization of ancient manuscripts and inscriptions-a review. In *Advances in Computing and Data Sciences: First International Conference, ICACDS 2016, Ghaziabad, India, 11–12 November 2016*; Revised Selected Papers 1; Springer: Singapore, 2017; pp. 605–612.
16. Abrate, M.; Del Grosso, A.M.; Giovannetti, E.; Duca, A.L.; Luzzi, D.; Mancini, L.; Marchetti, A.; Pedretti, I.; Piccini, S. Sharing Cultural Heritage: The Clavius on the Web Project. In Proceedings of the LREC 2014, Ninth International Conference on Language Resources and Evaluation, Reykjavik, Iceland, 26–31 May 2014; pp. 627–634.
17. The New York Public Library. Available online: https://www.nypl.org/collections/nypl-recommendations/guides/goodspeed-manuscript-collection (accessed on 12 April 2023).
18. The Schoenberg Database of Manuscripts. Available online: https://sdbm.library.upenn.edu/ (accessed on 12 April 2023).
19. Sicuro, M. *Una Piccola Comunità Ebraica al Confine Orientale Veneto-Asburgico in età Moderna: Ontagnano (1577–1797)*; EUT Edizioni Università di Trieste: Trieste, Italy, 2022.
20. The Friedberg Jewish Manuscript Society Project. Available online: https://fjms.genizah.org/ (accessed on 13 March 2023).
21. The Hebrew Manuscripts Digitisation Project. Available online: https://www.bl.uk/hebrew-manuscripts (accessed on 13 March 2023).
22. The Cairo Genizah Collection. Available online: https://cudl.lib.cam.ac.uk/collections/genizah/1 (accessed on 13 March 2023).
23. The Judaica Collection. Available online: https://www.nli.org.il/en/at-your-service/who-we-are/collections/judaism-collection (accessed on 13 March 2023).
24. The Jewish Atlantic World Project. Available online: https://rdc.reed.edu/c/jewishatl/home/ (accessed on 14 March 2023).
25. Ehrmann, M.; Hamdi, A.; Pontes, E.L.; Romanello, M.; Doucet, A. Named entity recognition and classification on historical documents: A survey. *arXiv* **2021**, arXiv:2109.11406.
26. Trias, F.; Wang, H.; Jaume, S.; Idreos, S. Named entity recognition in historic legal text: A transformer and state machine ensemble method. In Proceedings of the Natural Legal Language Processing Workshop 2021, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 172–179.
27. Aejas, B.; Bouras, A.; Belhi, A.; Gasmi, H. Named Entity Recognition for Cultural Heritage Preservation. In *Data Analytics for Cultural Heritage*; Belhi, A., Bouras, A., Al-Ali, A.K., Sadka, A.H., Eds.; Springer: Cham, Switzerland, 2021. [CrossRef]
28. van Hooland, S.; De Wilde, M.; Verborgh, R.; Steiner, T.; Van de Walle, R. Exploring entity recognition and disambiguation for cultural heritage collections. *Digit. Sch. Humanit.* **2013**, *30*, 262–279. [CrossRef]
29. Erdmann, A.; Wrisley, D.J.; Allen, B.; Brown, C.; Cohen-Bodénès, S.; Elsner, M.; Feng, Y.; Joseph, B.; Joyeux-Prunel, B.; De Marneffe, M.-C. Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 3–5 June 2019.
30. Pontes, E.L.; Cabrera-Diego, L.A.; Moreno, J.G.; Boros, E.; Hamdi, A.; Sidère, N.; Coustaty, M.; Doucet, A. Entity linking for historical documents: Challenges and solutions. In *Digital Libraries at Times of Massive Societal Transition: 22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020, Kyoto, Japan, 30 November–1 December 2020*; Springer International Publishing: Cham, Switzerland, 2020; pp. 215–231.
31. Manjavacas, E.; Fonteyn, L. Adapting vs. Pre-training Language Models for Historical Languages. *J. Data Min. Digit. Humanit.* **2022**, 1–19. [CrossRef]
32. Ehrmann, M.; Romanello, M.; Najem-Meyer, S.; Doucet, A.; Clematide, S.; Faggioli, G.; Potthast, M. Extended Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. In *CEUR Workshop Proceedings (No. 3180)*; CEUR-WS: Aachen, Germany, 2022; pp. 1038–1063.
33. Bezerianos, A.; Dragicevic, P.; Fekete, J.D.; Bae, J.; Watson, B. Geneaquilts: A system for exploring large genealogies. *IEEE Trans. Vis. Comput. Graph.* **2010**, *16*, 1073–1081. [CrossRef] [PubMed]
34. Gonzalez, J.; Nguyen, N.V.; Dang, T. VisFCAC: An Interactive Family Clinical Attribute Comparison. *arXiv* **2022**, arXiv:2208.11688.
35. Xiang, F.; Zhu, S.; Wang, Z.; Maher, K.; Liu, Y.; Zhu, Y.; Chen, K.; Liang, Z. Enhanced family tree: Evolving research and expression. In *ACM SIGGRAPH 2020 Art Gallery*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 367–373.
36. Liu, Y.; Dai, S.; Wang, C.; Zhou, Z.; Qu, H. GenealogyVis: A System for Visual Analysis of Multidimensional Genealogical Data. *IEEE Trans. Hum.-Mach. Syst.* **2017**, *47*, 873–885. [CrossRef]

37. Korst, J.; Pronk, V.; van Wijk, J.J. A visualization of family relations inspired by the london metro map. In Proceedings of the 13th International Symposium on Visual Information Communication and Interaction, Eindhoven, The Netherlands, 8–10 December 2020; pp. 1–8.
38. Mansueli, V.A.P.; Okano, M.T. Representations of genealogies in graph theory: K-Graphs. In Proceedings of the 27th International of Association for Management Technology, Birmingham, UK, 22–26 April 2018; pp. 1–10.
39. Folkman, T.; Furner, R.; Pearson, D. GenERes: A genealogical entity resolution system. In Proceedings of the 2018 IEEE International Conference on Data Mining Workshops (ICDMW), Singapore, 17–20 November 2018; pp. 495–501.
40. Leskinen, P.; Hyvönen, E. Reconciling and Using Historical Person Registers as Linked Open Data in the AcademySampo Portal and Data Service. In *The Semantic Web–ISWC 2021: 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24–28, 2021, Proceedings 20*; Springer International Publishing: Cham, Switzerland, 2021; pp. 714–730.
41. Wang, M.; Feng, J.; Shu, X.; Jie, Z.; Tang, J. Photo to family tree: Deep kinship understanding for nuclear family photos. In Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data, Seoul, Republic of Korea, 22–26 October 2018; pp. 41–46.
42. Koylu, C.; Guo, D.; Huang, Y.; Kasakoff, A.; Grieve, J. Connecting family trees to construct a population-scale and longitudinal geo-social network for the U.S. *Int. J. Geogr. Inf. Sci.* **2020**, *35*, 2380–2423. [CrossRef]
43. Lo Duca, A.; Bacciu, C.; Marchetti, A. Towards a smart navigation of cemeteries as cultural sites. In *Ancient Greek Art and European Funerary Art*; Cambridge Scholars Publishing: Newcastle upon Tyne, UK, 2019; p. 321.