From digitised sources to digital data: Behind the scenes of (critically) enriching a digital heritage collection[*]

Lorella Viola[1] and Antonio Maria Fiscarelli[1]

Luxembourg Centre for Contemporary and Digital History (C^{2}DH - University of Luxembourg) Maison des Sciences Humaines, 11, Porte des Sciences, Esch-sur-Alzette, L-4366, Luxembourg
lorella.viola@uni.lu
antonio.fiscarelli@uni.lu

Abstract. Digitally available repositories are becoming not only more and more widespread but also larger and larger. Although there are both digitally-born collections and digitised material, the digital heritage scholar is typically confronted with the latter. This immediately presents new challenges, one of the most urgent being how to find the meaningful elements that are hidden underneath such unprecedented mass of digital data. One way to respond to this challenge is to contextually enrich the digital material, for example through deep learning. Using the enrichment of the digital heritage collection *ChroniclItaly 3.0* [10] as a concrete example, this article discusses the complexities of this process. Specifically, combining statistical and critical evaluation, it describes the gains and losses resulting from the decisions made by the researcher at each step and it shows how in the passage from digitised sources to enriched material, most is gained (e.g., preservation, wider and enhanced access, more material) but some is also lost (e.g., original layout and composition, loss of information due to pre-processing steps). The article concludes that it is only through a critical approach that the digital heritage scholar can successfully meet the interpretive challenges presented by the digital and the digital heritage sector fulfil the second most important purpose of digitisation, that is to enhance access.

Keywords: Digital heritage · Contextual enrichment · Deep learning · Digital humanities

1 Introduction

As digitally available repositories are becoming larger and larger, finding the meaningful elements that are hidden within such an unprecedented mass of digital data is increasingly challenging. Moreover, if the collection is digitised (as

[*] This research is supported by the Luxembourg Centre for Contemporary and Digital History (C^{2}DH - University of Luxembourg) - *Thinkering Grant* and the Luxembourg National Research Fund (13307816).

opposed to be born digital), then further complexity is added to the picture, typically due to the often low quality of the digitised source (e.g., OCR mistakes, markings on the pages, low readability of unusual fonts, or a general poor condition of the original text). One way to respond to the pressing need for enhancing access and making the collections retrievable in valuable ways is to enrich the digital material with contextual information, for example by using deep learning. This process, however promising and continually advancing, is not free from challenges of its own which, to be effectively tackled, require technical expertise but more importantly, full critical engagement.

This article uses the enrichment of the digital heritage collection *ChroniclItaly 3.0* [10] as a concrete example to discuss such challenges. *ChroniclItaly 3.0* is a corpus of Italian American immigrant newspapers published between 1898 to 1936. Specifically, combining statistical and critical evaluation, it describes the gains and losses resulting from the decisions and interventions made by the researcher at every step of the process of enrichment, from the so-called pre-processing steps (tokenization, lowercasing, stemming, lemmatization, removing stopwords, removing *noise*) to the enrichment itself (Named Entity Recognition - NER, Geo-coding, Entity Sentiment Analysis - ESA). The enrichment process presented in this article was carried out in the context of a larger project, *DeepTextMiner* (DeXTER) which was supported by the Luxembourg Centre for Contemporary and Digital History (C^{2}DH - University of Luxembourg) *Thinkering Grant*. DeXTER aims to offer a methodological contribution to digital humanities (DH) by exploring the value of reusing and advancing existing knowledge. Specifically, it intends to experiment with different state-of-the-art NLP and deep learning techniques for both enriching digital data with contextual information (e.g., referential, geographical, emotional, topical, relational) and visualising it. The long-term goal is to devise a generalisable, interoperable workflow that could assist researchers in both these tasks.

The discussion will show how the passage from digitised sources to enriched material, while aiming to make collections more engaging and valuable on the whole through preservation and wider and enhanced access, is not free from disadvantages such as loss of the original layout and structure, loss of information due to pre-processing steps, introduction of new errors, etc. The article ultimately shows that it is only through an active, critical engagement with the digital sources that the digital heritage scholar can successfully meet the interpretive challenges presented by the digital and the digital heritage sector fulfil the second most important purpose of digitisation after preservation, that is to enhance access.

2 Digital source and methodology

This article describes the technical interventions and critical choices made towards enriching the digital heritage collection *ChroniclItaly 3.0* with contextual information (i.e., referential entities, geo-coding information, sentiment). By means of statistical and critical evaluation, it aims to show that digitally enabled research is highly dependent on the critical engagement of the scholar who is never 'completely removed' from the computational. From pre-processing the digital material to the enrichment itself, the paper documents the decisions and interventions made by the researcher at every step of the process and how these affected the material and inevitably impacted the final output, in turn impacting access, analysis and reuse of the collection itself.

2.1 ChroniclItaly 3.0

ChroniclItaly 3.0 [10] is a corpus of Italian immigrant newspapers published in the United States between 1898 and 1936. The corpus includes the front pages of the following titles: *L'Italia*, *Cronaca Sovversiva*, *Il Patriota*, *La Libera Parola*, *La Rassegna*, *La Ragione*, *L'Indipendente*, *La Sentinella*, *La Sentinella del West*, and *La Tribuna del Connecticut* for a total of 8,653 issues. The collection includes further issues as well as three additional titles from its two previous versions, *ChroniclItaly* [8] and *ChroniclItaly 2.0* [9]. As the previous versions, *ChroniclItaly 3.0* has been machine-harvested from *Chronicling America*, an Internet-based U.S. directory[1] of digitised historical newspapers published in the United States from 1789 to 1963. The corpus features *prominenti* (mainstream), *sovversivi* (anarchic), and independent newspapers,[2] thus providing a very nuanced picture of the Italian immigrant community in the United States at the turn of the twentieth century.

Immigrant newspapers were continually fighting against the risk of bankruptcy and owners were often forced to discontinue the titles; for this reason, some titles such as *L'Italia* - one of the most mainstream Italian immigrant publications in the U.S. at the time - managed to last for years, while others like *La Rassegna* or *La Ragione* could survive only for a few months. This is reflected in the composition of the collection which therefore presents gaps across titles (Figure 1). Also due to the newspapers' economic struggles, the number of issues vary greatly across titles, with some titles publishing thousands of issues and others only a few hundreds. The overall coverage of issues is nonetheless relatively evenly distributed across the whole period and titles of different orientation co-exist at different points in time thus ensuring that a degree of balance is kept throughout the collection. Users should however take into account factors such as over- or under-representation of some titles, potential polarisation of topics, etc. when engaging with the resource.

2.2 Methodology

This paper describes the process of contextually enriching the digital heritage collection *ChroniclItaly 3.0* from the so-called pre-processing steps to the enrichment itself. Combining statistics and critical reflection, the aim is to provide an

[1] http://https://chroniclingamerica.loc.gov/

[2] For further information about the classification of the titles based on their political orientation, please refer to [11]

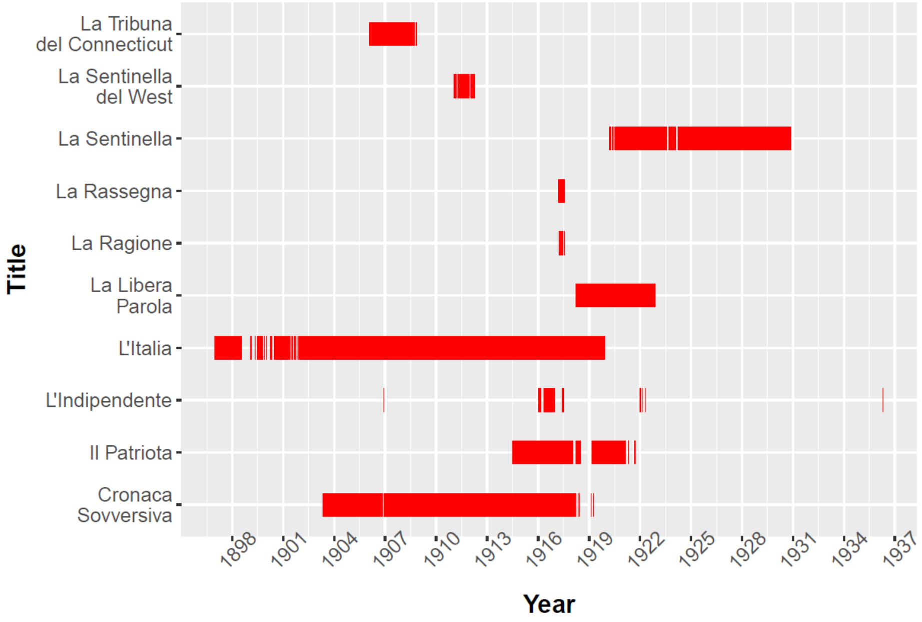


Fig. 1. Distribution of issues within *ChroniclItaly 3.0* per title. Red lines indicate at least one issue in a three month period.

insider perspective on the technical and analytical challenges posed by the task of enriching digital heritage material. The discussion is divided in two parts: the first part discusses the pre-processing steps, which include tokenization, removing words with less than two characters, removing numbers, dates and special characters); the second part is concerned with the enrichment itself (NER, geocoding, ESA).

3 Towards enrichment: Pre-processing

Before working towards enriching a digital (textual) collection, and indeed before performing any Natural Language Processing (NLP) task, the digital material needs to be *prepared*. It is all too often assumed that this operation, also ambiguously referred to as *cleaning*, is not part of the enrichment process and can therefore be fully automatic, unsupervised and launched as a one-step pipeline. In this section, we show how it is on the contrary paramount that this phase is tackled critically as each one of the taken actions will have consequences on the material, on how the algorithms will process such material and therefore on the output and finally, on how the collection will be accessed and the information retrieved and interpreted.

Deciding which ones of the pre-processing operations should be performed and how depends on many factors such as the language of the data-set, the type of material, the specific enrichment task, to name but a few. Typical operations that are considered part of this step are tokenization, lowercasing, stemming, lemmatization, removing stopwords, removing *noise* (e.g., numbers, punctuation marks, special characters). In principle, all these interventions are optional as the algorithms will process whichever version of the data-set is used. In reality, however, pre-processing the digital material is key to the subsequent operations for several reasons. First and foremost, pre-processing the data will remove most OCR mistakes which are always present in digital textual collections to various degrees. This is especially true for corpora such as historical collections, repositories of under-documented languages, or digitised archives from handwritten texts. Second, it will reduce the size of the collection thus decreasing the required processing power and time. Third, it is *de facto* a data exploration step which allows the digital heritage scholar to look more closely at the material.

It is important to remember that each one of these steps is an additional layer of manipulation and has direct, heavy consequences on the material and therefore on the following operations. It is critical that digital scholars assess carefully to what degree they want to intervene on the material and how. For this reason, this part of the process of contextual enrichment should not be considered as separate from the enrichment itself, on the contrary, it is an integral part of the entire process.

The specific pre-processing actions taken towards enriching *ChroniclItaly 3.0* were: tokenization, removing numbers, dates, removing words with less than two characters and special characters. Numbers and dates were removed because they are not only typically considered irrelevant to NER and ESA, but sometimes they may even interfere with the algorithm performance, thus potentially worsening the quality of the output. The last two operations were performed because it was found that special and isolated characters (characters delimited by spaces) were in fact OCR mistakes.

Other actions included merging words wrongfully separated by a newline, a white space or punctuation. Once this step was performed, the collection totalled up to 21,454,455 words. In Figure 2, we show how this step impacted each title per year in terms of the percentage of material removed and preserved while Figure 3 displays such percentages aggregated for each title.

As the figures show - particularly Figure 3 - on average, after this step 30% of the material was removed from each title. This means that, if on the one hand the step was likely to result in more reliable and interpretable material, on the other it came at the expenses of potentially important information. In this respect, the challenge does not lie so much in performing the pre-processing itself, rather in assessing which operations minimise the loss of potentially useful information and maximise the enhancement of the resource. As the technology is still not perfect, digital heritage scholars and institutions must respond to this challenge by carefully pondering the pros and cons of enriching digital collections and duly warn the users about the performed interventions.

As for the specific pre-processing operations to perform, we chose to not remove stopwords and punctuation, even though they are typically considered

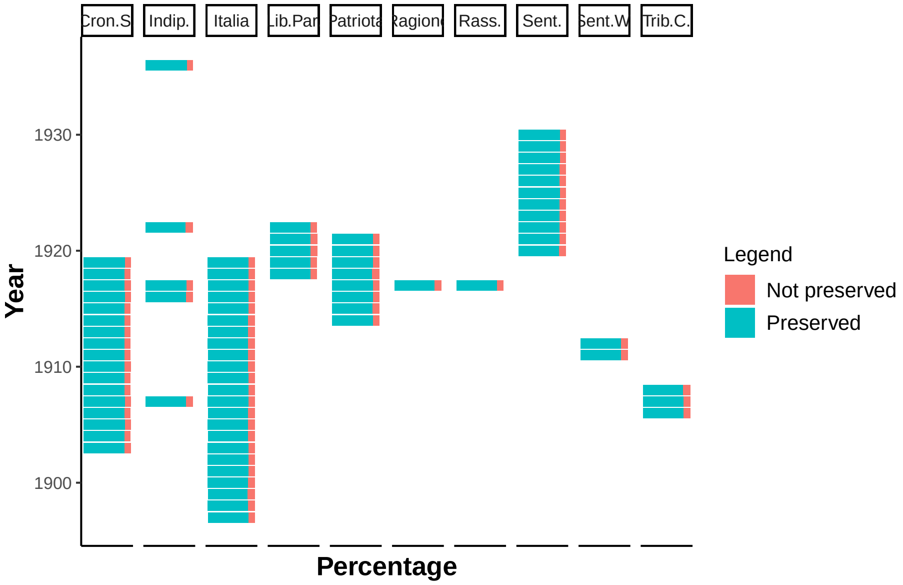


Fig. 2. Impact of pre-processing *ChroniclItaly 3.0* per title per year.

as not semantically salient and in fact even detrimental to the models' performance. The choice was motivated by the enrichment actions to follow, namely NER, geo-coding, and ESA: prepositions and articles are often part of locations, organisations and names while punctuation is a typical sentence delimiter, that is it sets the sentences' boundaries, indispensable to perform sentiment analysis. Therefore, removing such material from the collection could have had a negative impact on the enrichment quality. We also decided to not lowercase the material. Lowercasing can be a double-edged sword; for instance, if lowercasing is not performed, the algorithm will treat 'USA', 'Usa', 'usa', 'UsA', 'uSA', etc. as distinct tokens, even though they may all refer to the same entity. On the other hand, once lowercased, it may become difficult for the algorithm to recognise entities, thus outputting many false negatives, and for the scholar to distinguish between homonyms, thus potentially skewing the output. As entities such as persons, locations and organisations are typically capitalised, we decided to not perform lowercasing in preparation for NER and geo-coding. Once these steps were completed, however, the entities were lowercased so that the issue of multiple items referring to the same entity (e.g., 'USA' and 'Usa') could be overcome (*cfr.* Section 5).

4 Enrichment: Named Entity Recognition

The growing digitisation of material and immaterial cultural heritage has proportionally increased the relevance of Artificial Intelligence (AI) for researchers

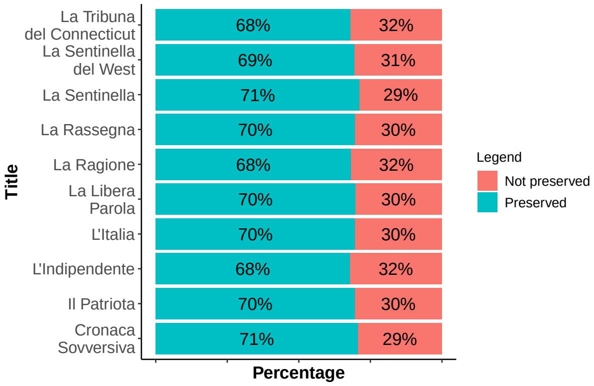


Fig. 3. Impact of pre-processing on *ChroniclItaly 3.0* per title.

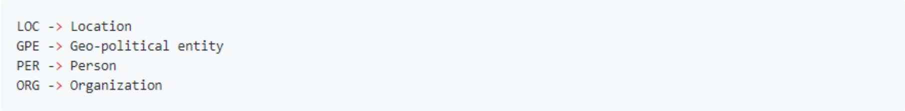
in the humanities as well as libraries and cultural heritage institutions [3, 13]. AI offers users and researchers methods to analyse, explore and make sense of the different layers of information sometimes hidden in large quantities of material. For instance, one effective application of AI is the possibility to enrich the digital material with data that could allow for in-depth cultural analyses. One example of such text enrichment is Named Entity Recognition (NER), that is using contextual information to identify referential entities such as names of persons, locations and organisations.

NER tasks have over the years undergone major changes, however it has been repeatedly proven that machine learning algorithms based on neural networks outperform all previous methods. For this reason, the NER enrichment of *ChroniclItaly 3.0* was performed by using a deep learning sequence tagging tool that implements Tensorflow [5]. The algorithm combines Bidirectional Long Short-Term Memory (BiLSTM) and Conditional Random Fields (CRF) as top layer with character embeddings which was found to outperform CRFs with hand-coded features. Methodologically, the character embeddings are trained with pre-trained word embeddings while training the model itself. The character and subword based word embeddings are computed with FastText [1] as it was found that, by retrieving embeddings for unknown words through the incorporation of subword information, issues with out-of-vocabulary words are significantly alleviated.

The sequence tagging model for the Italian language was trained on I-CAB[3] (Italian Content Annotation Bank), an open access corpus annotated for entities (i.e. persons-PER, organisations-ORG, locations-LOC, and geopolitical entities-GPE), temporal expressions, and relations between entities. I-CAB contains 525 news articles taken from the Italian newspaper *L'Adige* and totals up around 180,000 words. Embeddings were computed using Italian Wikipedia and trained using Fastext with 300 dimensions. The results of the F1 score for the Italian

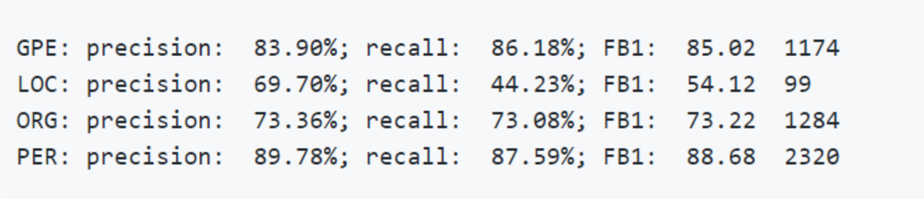
[3] http://ontotext.fbk.eu/icab.html

models are: accuracy: 98.15%; precision: 83.64%; recall: 82.14%; F1: 82.88. Figure 4 shows the tags of the NER algorithm, Figure 5 shows the accuracy scores per entity, while Figure 6 shows the final output of the sequence tagger on *ChroniclItaly 3.0*.



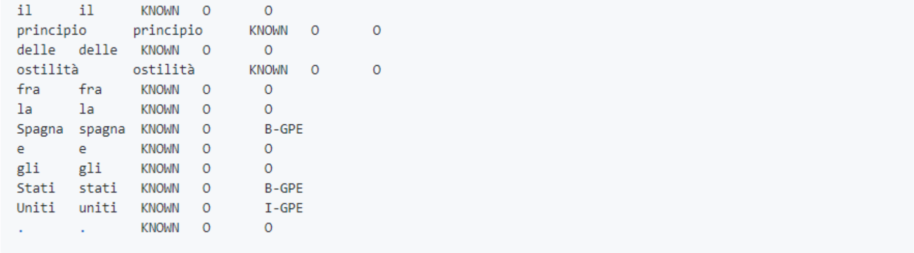
```
LOC -> Location
GPE -> Geo-political entity
PER -> Person
ORG -> Organization
```

Fig. 4. Output tags of the NER algorithm.



```
GPE: precision:  83.90%; recall:  86.18%; FB1:  85.02  1174
LOC: precision:  69.70%; recall:  44.23%; FB1:  54.12  99
ORG: precision:  73.36%; recall:  73.08%; FB1:  73.22  1284
PER: precision:  89.78%; recall:  87.59%; FB1:  88.68  2320
```

Fig. 5. Accuracy scores of the NER algorithm for Italian models per entity.



```
il      il      KNOWN   O       O
principio       principio       KNOWN   O       O
delle   delle   KNOWN   O       O
ostilità        ostilità        KNOWN   O       O
fra     fra     KNOWN   O       O
la      la      KNOWN   O       O
Spagna  spagna  KNOWN   O       B-GPE
e       e       KNOWN   O       O
gli     gli     KNOWN   O       O
Stati   stati   KNOWN   O       B-GPE
Uniti   uniti   KNOWN   O       I-GPE
.       .       KNOWN   O       O
```

Fig. 6. Output of the NER algorithm on *ChroniclItaly 3.0*.

The NER algorithm retrieved 547,667 entities, which occurred 1,296,318 times across the ten titles. A close analysis of the entities, however, revealed a number of issues which required a manipulation of the output. These issues included: entities that had been assigned the wrong tag (e.g., New York - PER), multiple entities referring to the same entity (e.g., Woodraw Wilson, President Woodraw Wilson), elements wrongfully tagged as entities (e.g., venerdì 'Friday'

- ORG). Therefore, a list of these exceptions was compiled and the results adjusted accordingly. Once the data were modified, the data-set counted 521,954 unique entities which occurred 1,205,880 times. Figure 6 shows how the intervention affected the distribution of entities across the four categories - geopolitical, persons, locations, organisations - per title.

As it can be seen, the redistribution of entities varied across categories and titles, in some cases dramatically. For instance in *La Rassegna*, the number of entities in the LOC category significantly decreased whereas it increased in *L'Italia*. This action required a combination of expert knowledge and technical ability as the entities had to be carefully analysed and historically triangulated in order to make informed decisions on how to intervene on the output without introducing errors. Although time-consuming and in principle optional, this critical evaluation intervention nevertheless significantly improved the accuracy of the tags thus overall increasing the quality of the NER output in preparation for the following stages of the enrichment, namely geo-coding and ESA and ultimately, adding more value to the user's experience for access and reuse. It is therefore a highly recommended operation.

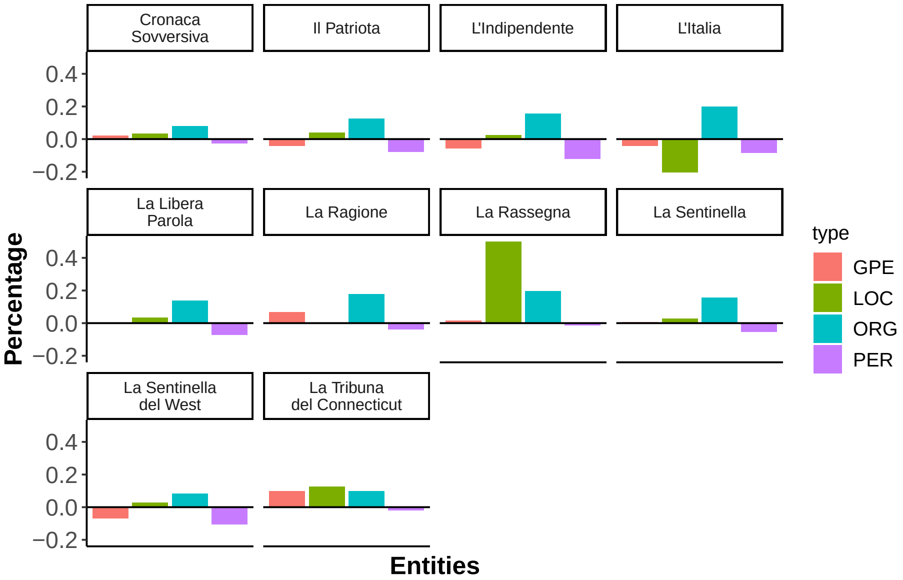


Fig. 7. Distribution of entities per title after intervention. Positive bars indicate a decreased number of entities after the process, negative bars indicate an increased number.

5 Enrichment: Geo-coding

The relevance of geo-coding for digital heritage collections lies in what has been referred to as *Spatial turn*, the study of *space* and *place* [4] as distinct entities in that it sees place as created through social experiences and can therefore be both real and imagined whereas space is essentially geographic. Spatial humanities is a recently emerged interdisciplinary field within digital humanities which, following the *Spatial turn*, focuses on geographic and conceptual space, particularly from a historical perspective. Based on Geographic Information Systems (GIS), locations in data-sets are geo-coded and displayed on a map. Especially in the case of large collections with hundreds of thousands of geographical entities, the visualisation is believed to help scholars access the different layers of information that may be behind geo-references. Indeed, such process of cross-referencing geo-locations with other type of data (e.g., historical, social, temporal) has provided researchers working in fields such as environmental history, historical demography, and economic, urban and medieval history with new insights leading them to propose alternatives to traditional positions and even explore new research avenues [12]. The theoretical relevance of using NER to enrich digital heritage collections lies precisely in its great potential for discovering the cultural significance underneath referential units and how that may have changed over time.

Another challenge posed to digital heritage scholars is that of the language of the collection. In the case of *ChroniclItaly 3.0*, for example, almost all choices made by the researcher towards enriching, including NER and geo-coding, were conditioned by the fact that the language of the collection is not English. The relative lack of appropriate computational resources available for languages other than English often dictates which tools and platforms can be used for specific tasks. For geo-coding, for instance, it was found that setting the API language as the language of the data-set improves the accuracy of the geo-coding results [12]. For this reason, the geographical entities in *ChroniclItaly 3.0* have been geo-coded using the *Google Cloud Natural Language API* within the *Google Cloud Platform Console* which provides a range of NLP technologies in a wide range of languages, including Italian.

Moreover, when enriching collections in languages other than English, digital heritage scholars often find themselves confronted with the additional challenge of having to choose between training their own models or using existing models. While the former scenario may sometimes not be an option due for instance to the lack of specific training of the scholar, time or money limits, etc., the latter alternative may also not be ideal. Even when trained in the target language (like in the case of *ChroniclItaly 3.0*), typically the training would have occurred within the context of another project, for different purposes, possibly using data-sets with very different specifics from the one the scholar is enriching. Researchers are then usually forced to sacrifice the possibility to achieve a potentially much higher quality and/or accuracy of their results in the interest of time, money or both. For example, as shown in Figure 5, although the overall F1 score of the NER algorithm for Italian models was satisfactory (82.88), the individual performance for the entity LOC was rather poor (54.19). This may depend on several factors (e.g., lack of this type of locations in the data-set used for training the model, different locations tagged as LOC) which are related to the wider challenge of accessing already available trained models in the desired language. Because of the low score of the category LOC, we decided to geo-code only GPE entities. Though not optimal, the decision was made also considering that GPE entities are generally more informative as they would typically refer to countries and cities (though it was found to retrieve also counties and States) while LOC entities are typically rivers, lakes, and geographical areas (e.g., the Pacific Ocean). Future work on the collection could focus on performing NER using a more fine-tuned algorithm and geo-code the LOC-type entities.

In total, 2,160 GPE-type entities were geo-coded, these are referred to 283,879 times throughout the whole corpus. Figure 8 shows the distribution of unique GPE entities per title.

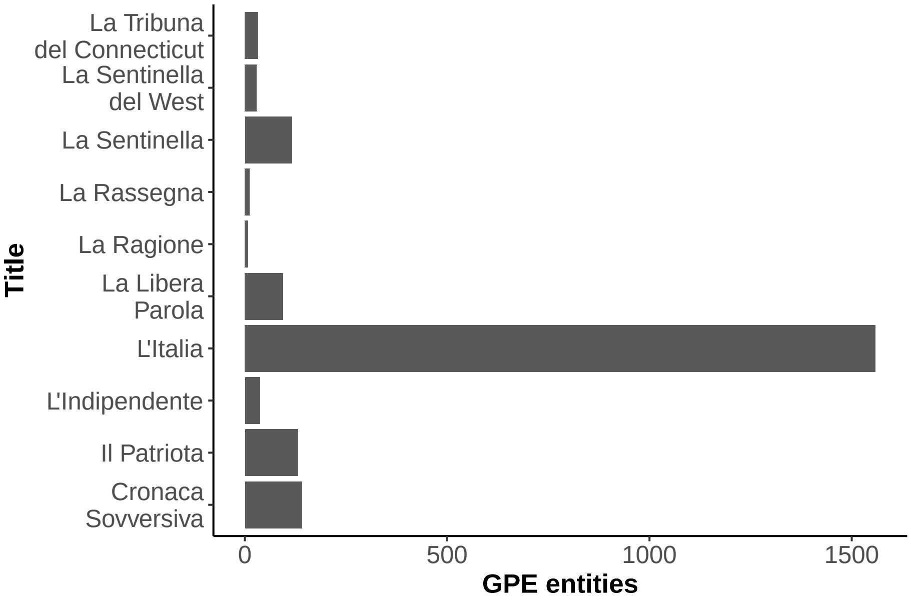


Fig. 8. Distribution of unique GPE entities per title.

6 Enrichment: Entity Sentiment Analysis

Another enriching technique used to add value to digital heritage collections is Entity Sentiment Analysis (ESA), a way to identify the prevailing emotional attitude of the writer towards referential entities. While Sentiment Analysis (SA) identifies the prevailing emotional opinion within a given text, ESA adds a more targeted and sophisticated quality to the enrichment as it allows users to identify the layers of meaning humans attached historically to people, organisations, geographical spaces. Understanding the meaning humans invested in such entities and, in the case of historical collections such as *ChroniclItaly 3.0*, how that meaning may have changed over time, provides digital heritage scholars with powerful means to access part of collective narratives of fear, pride, longing, loss. ESA enables us to connect these subjective connotations to the named entities extracted from digitised texts [12, 2, 6, 7] thus offering a more engaging and complete fruition of information as well as tackling new questions or revising old assumptions. For example, the ESA of *ChroniclItaly 3.0* may allow researchers to investigate how immigrants made sense of their diasporic identities within the host community and how their relationship with the homeland may have changed over time.

The process of performing ESA on the collection required several steps:

- Identify the sentence delimiters (i.e., full stop, semicolon, colon, exclamation mark, question mark) and divide the textual material accordingly. At the end of this step, 677,030 sentences were obtained;
- Select the most frequent entities for each category and in each title. As each title differs in size, we used a logarithmic function to obtain a more representative number of entities per title (2*log2 function used). At the end of this step, 228 entities were obtained distributed across titles as shown in Figure 9;
- Select only the sentences that contained the entities identified in the previous step. This step was done to limit the number of API requests and reduce processing time and costs. The selection returned 133,281 sentences distributed across titles as shown in Figure 10;
- Perform ESA on the sentences so selected. When the study was carried out, no suitable SA models trained on Italian were found, therefore this step was performed once again using the *Google Cloud Platform Console.*

7 Conclusions

This article discussed the technical and analytical challenges posed by the task of enriching a digital heritage collection and used the enrichment of *ChroniclItaly 3.0* [10] as a concrete example. Specifically, the article combined statistical analysis and critical evaluation to describe the impact that the decisions and interventions made at every step of the enrichment process had on the collection. Through such discussion, the aim was to show that, paradoxically, enriching a digital heritage collection means sacrificing content. Therefore, an active, critical engagement of the digital heritage scholar is an absolutely necessary requirement to minimise such losses and ensure that the enrichment actions truly increase the value of the collections, making the users' experience more engaging and thus widening and enhancing access.

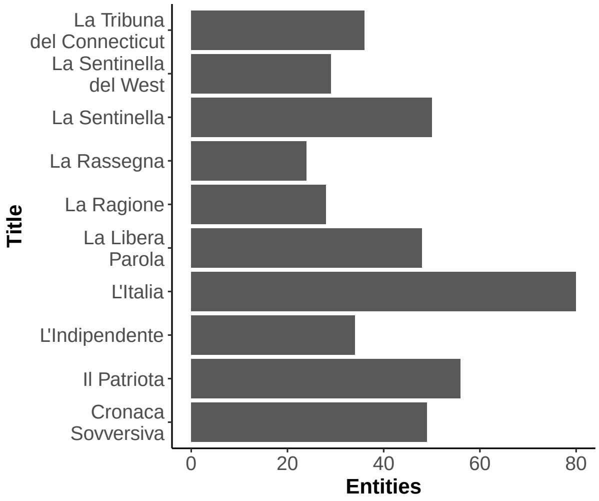


Fig. 9. Distribution of selected entities for ESA across titles.

The enrichment actions discussed here are part of a wider process of enrichment within DeXTER. Additional enrichment operations include topic modelling and network analysis, which were not discussed here due to word count limitations. Though not exhaustive, we nevertheless hoped to have shown that it is only through a continuous, critical engagement with the digital sources that the digital heritage scholar can successfully meet the challenges presented by the digital and fulfil the main purposes of digitisation: preservation and knowledge access.

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information (2017)
2. Donaldson, C., Gregory, I.N., Taylor, J.E.: Locating the beautiful, picturesque, sublime and majestic: spatially analysing the application of aesthetic terminology in descriptions of the english lake district. Journal of Historical Geography **56**, 43–60 (2017), doi: 10.1016/j.jhg.2017.01.006
3. Fiorucci, M., Khoroshiltseva, M., Pontil, M., Traviglia, A., Del Bue, A., James, S.: Machine learning for cultural heritage: A survey. Pattern Recognition Letters **133**, 102 – 108 (2020). https://doi.org/https://doi.org/10.1016/j.patrec.2020.02.017, http://www.sciencedirect.com/science/article/pii/S0167865520300532
4. Murrieta-Flores, P., Martins, B.: The geospatial humanities: past, present and future. International Journal of Geographical Information Science **33**(12), 2424–2429 (2019). https://doi.org/10.1080/13658816.2019.1645336, https://doi.org/10.1080/13658816.2019.1645336
5. Riedl, M., Padó, S.: A named entity recognition shootout for german. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 120–125 (2018), doi: 10.18653/v1/P18-2020

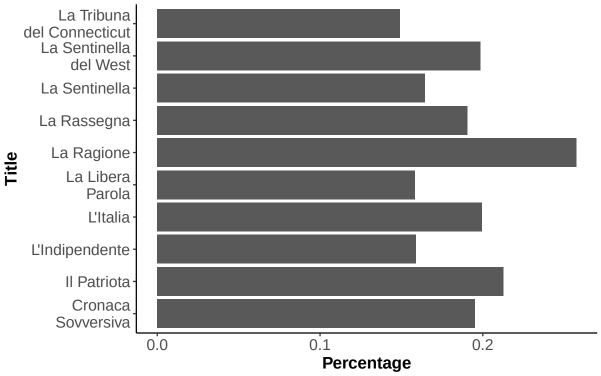


Fig. 10. Ratio of sentences containing selected entities across titles.

6. Tally Jr, R.T.: Geocritical explorations: Space, place, and mapping in literary and cultural studies. Springer (2011), doi: 10.1057/9780230337930
7. Taylor, J.E., Donaldson, C.E., Gregory, I.N., Butler, J.O.: Mapping digitally, mapping deep: exploring digital literary geographies. Literary Geographies **4**(1), 10–19 (2018)
8. Viola, L.: ChroniclItaly. A corpus of Italian language newspapers published in the United States between 1898 and 1922. Utrecht University (2018), doi: 10.24416/UU01-T4YMOW
9. Viola, L.: ChroniclItaly 2.0. A corpus of Italian American newspapers annotated for entities, 1898-1920. Utrecht University (2019), doi: 10.24416/UU01-4MECRO
10. Viola, L.: ChroniclItaly 3.0. A contextually enriched digital heritage collection of Italian immigrant newspapers published in the USA, 1898-1936 (In press)
11. Viola, L., Verheul, J.: Mining ethnicity: Discourse-driven topic modelling of immigrant discourses in the usa, 1898–1920. Digital Scholarship in the Humanities **35**(4), 921–943 (2019), doi: 10.1093/llc/fqz068
12. Viola, L., Verheul, J.: Machine learning to geographically enrich understudied sources: A conceptual approach. In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence-Volume 1: ARTIDIGH. pp. 469–475. SCITEPRESS (2020), doi: 10.5220/0009094204690475
13. Weber, A., Ameryan, M., Wolstencroft, K., Stork, L., Heerlien, M., Schomaker, L.: Towards a digital infrastructure for illustrated handwritten archives. In: Digital Cultural Heritage, pp. 155–166. Springer (2018)