Vital Records: Uncover the past from historical handwritten records

Hervé Déjean
Naver Labs Europe
6 chemin de Maupertuis
38240 Meylan, France
herve.dejean@naverlabs.com

Jean-Luc Meunier
Naver Labs Europe
6 chemin de Maupertuis
38240 Meylan, France
jean-luc.meunier@naverlabs.com

Abstract

We present *Vital Records*, a demonstrator based on deep-learning approaches to handwritten-text recognition, table processing and information extraction, which enables data from century-old documents to be parsed and analysed, making it possible to explore death records in space and time. This demonstrator provides a user interface for browsing and visualising data extracted from 80,000 handwritten pages of tabular data.

1 Introduction

A great deal of human history is detailed in hand-written documents that have yet to be analyzed. Extracting information from such documents has, until recently, represented an extremely time-consuming and labour-intensive task. For this reason, there remains a treasure trove of untapped information that could help provide insight into, for example, the impact of industrialization on populations. Indeed, some of these records—such as those documenting epidemics—may provide useful context for our understanding of modern problems.

The recent advances in Handwritten Text Recognition (Mühlberger et al., 2019), and in Deep Learning at large, allow now us to automatically process large volumes of documents.

The *Vital Records* demonstrator, based on a collection of 80,000 pages, brings to life records that were handwritten by more than 700 different priests from 200 parishes in Germany from 1848 to 1878. The demo illustrates how state-of-the-art deep-learning methods—handwritten text recognition (HTR), table recognition (TR), and information extraction (IE)—can be used to transform these records into a digital format that can be queried and visualized in different ways to enrich our knowledge from previously unexplored sources of information.

The online demonstrator is available under this URL[1]

2 Using *Vital Records* to trace trends through history

Vital Records allows users to browse and visualize the dataset extracted from these German records using spatio-temporal criteria, in addition to the usual textual search queries (Figure 1).

Death records provide a range of useful information. As well as the name, age, date, and cause of death, the records include the profession of the deceased (Figure 3). With this information in hand, we are able to visualize trends over a given period. For example, we used these Death records to trace the evolution of professions between 1847 and 1877. The resulting graph, shown in Figure 4, shows the number of deaths recorded for which the field of profession (Stand) contains weaver (Weber), shoemaker (Schumacher), and miller (Müller). The graph may be considered a proxy for estimating the evolution of these professions during this period. It's possible, too, to spot locally relevant characteristics. For example, data extracted from these records shows a high number of glassmakers (Glasmacher) around

[1] https://europe.naverlabs.com/about/global-ai-rd-belt/eu-and-government-projects/vital-records

Proceedings of LaTeCH-CLfL 2020, pages 69–73
Barcelona, Spain (Online), December 12, 2020.

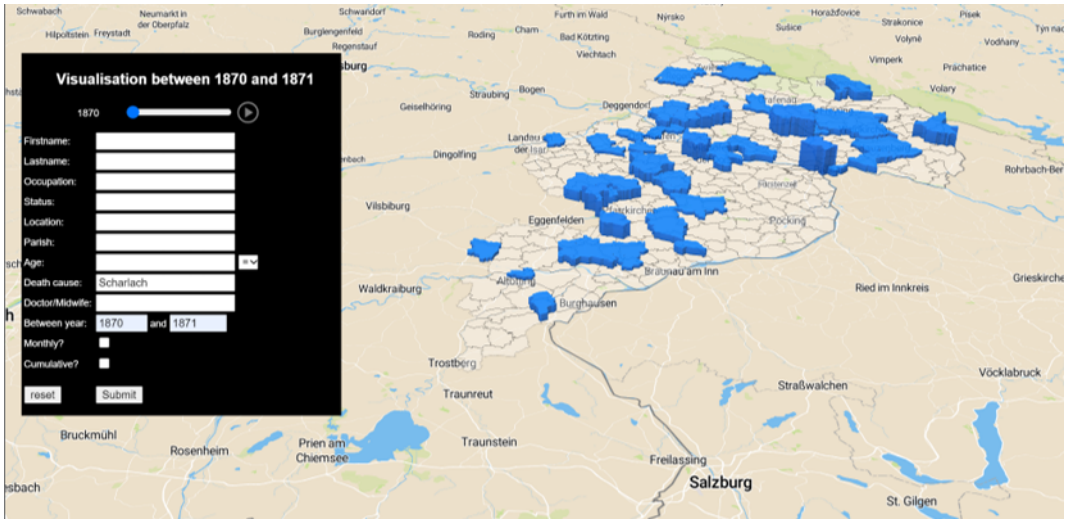


Figure 1: The user interface allows you to specify temporal and spatial criteria as well as to specify a value for each data field.

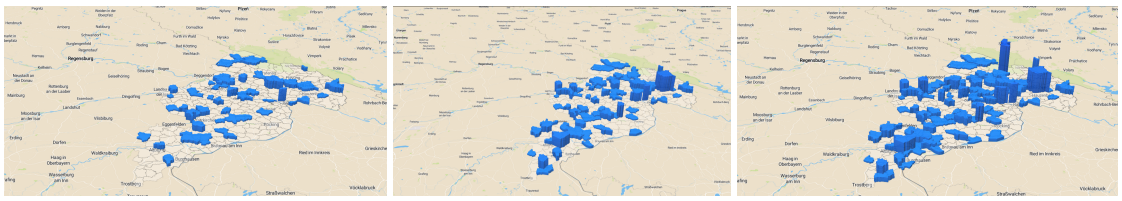


Figure 2: Example of temporal visualisation: number of deaths per parish, death cause:scarlet fever (Scharlach) in January, July and December 1871.

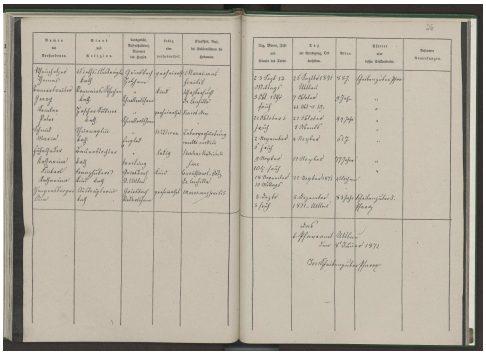


Figure 3: Example from the German records in which deaths are recorded in table format.

Zwiesel, a region that remains well-known for its production of glassware to this day. In addition to the visualization of data using graphs, *Vital Records* makes possible the development of temporal animations for which the timestep can be adjusted by year, month, or day. By combining both spatial and temporal information in this way, we are able to visualize the evolution of a given query over space and time. Our favourite query is the spread of scarlet fever in 1871.

It's also possible to track specific events, such as the opening of a hospital in a parish, or the arrival of a train line. Searching for professions containing the string *bahn* will provide you with a map of the three Bahnstrecke, or train lines: Passau-Regensburg, the Bavarian Forest Railway, and the München-Simbach.

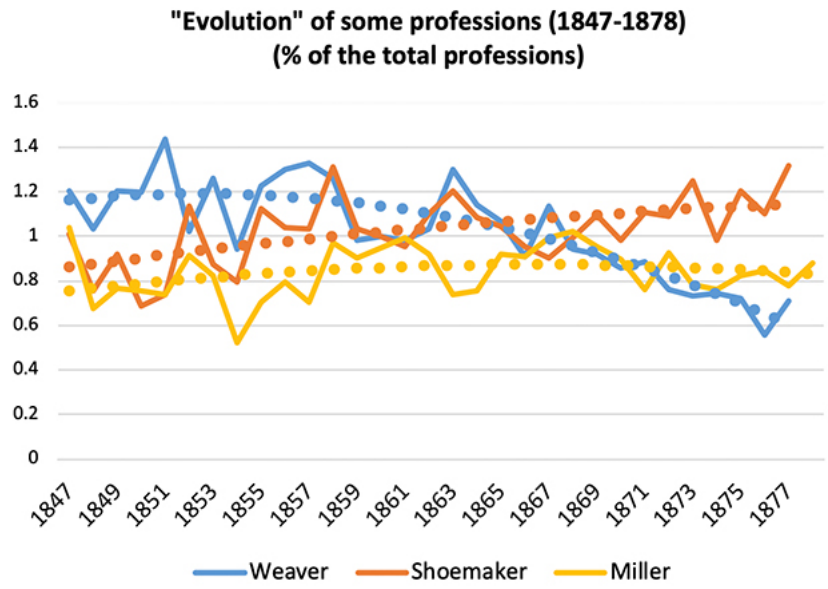


Figure 4: Visualization and data obtained from the *Vital Records* demonstrator, based on handwritten records from 200 parishes in Germany. This graph shows the evolution of the professions of weaver (blue), shoemaker (orange), and miller (yellow) between 1847 and 1877. Dotted lines represent the regression lines.

3 Deep-learning and Recognition: the Technology behind *Vital Records*

3.1 Handwritten Text Recognition

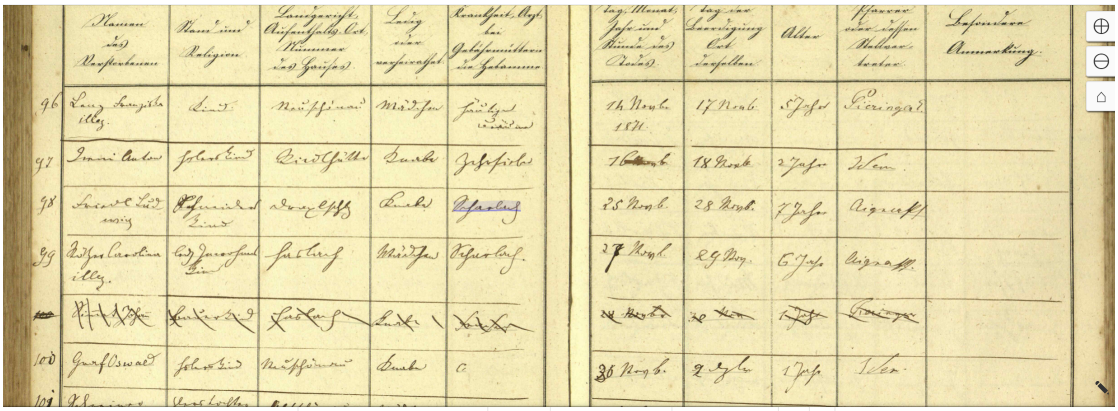
To develop this technology and create the Vital Records demo, we began by transforming handwritten tables—like the one shown in Figure 3—into a digital format. The field of handwritten document processing has improved significantly over the past few years, thanks to the neural network paradigm. First, we used the automated recognition and transcription platform Transkribus (Mühlberger et al., 2019) to transcribe 1000 pages from the German records. We then used these transcriptions to train an HTR model that could automatically digitise the information from the Death, Birth, and Wedding records of the archive. We found the character error rate (CER) to be around 10%, meaning that on average one letter out of every ten is wrongly recognised.

3.2 Table Recognition

Next, we focused on converting an image of a table into a spreadsheet (Figure 5). Understanding a page means grasping the layout, the relations between textual elements within the page, and so on. Although such 'table understanding' remains a challenge, it's one that can now be addressed with deep neural network technology (Prasad et al., 2019). We developed a neural network model that would learn how to organize a set of lines such that the information could be arranged into table rows. To achieve this, we enlisted the help of a graph convolutional network. By using the annotated collection (1000 pages containing tables that had been annotated to train the models), our system was able to recognise nine out of ten table rows. Figure 5 shows the original image and the reconstructed table using the Web interface provided by the Transkribus Plateform[2].

Building a user interface that enables experts to search the collection through space and time could provide an answer to the question 'Are the results sufficiently good to be useful to users of the archive or social historians?'

[2]https://transkribus.eu/r/read/projects/



Namen des Verstorbenen	Stand um Religion	Landgericht, Aufenthalts=Ort, Nummer des Hauses	Ledig oder verheirathet	Krankheit, Arzt bei Gebährmüttern die Hebamme	Tag, Monat, Jahr und Stunde des Todes.	Tag der Beerdigung Ort derselben.	Alter	Pfarrer oder dessen Stellver¬treter.	Besondere Anmerkung.
Franz Kangiter illeg.	Kind.	Neuschönau	Mädchen	häuligen Dirn im	14 Novbr 1871.	17 Novb.	5 Jahr	Peringart.	
deni Anton	Hehlers Kind	Aedlhütte	Knabe	Zehrfieber	1 Oktemb.	18 Novb	3 Jahre	Idem	
Friedl Lud wig	Schneider kind	Draplschl	Knabe	Scharlach	25 Novb.	28 Novb.	7 Jahre	Aigenff	
Röther Carolina illeg.	led Inwohners Kind	haslach	Mädchen	Scharlach.	27 Novb.	29 Nov.	6 Jahr	Aigraß.	
Ssath Phl	/Bauerkd	EusbachKnabe	Naden	24 Novbr	10 Jen.	12 Jahr	Einger		
Grafsswal	holerskind	Neuschönau	Knabe	6	26 Novb.	2 Jahre	1 Jahr	Jeher	

Figure 5: The original image and the corresponding extracted table.

3.3 Information Extraction and Spatio-Temporal Indexing

Once the information in the table had been extracted (see Figure 5), we used a state-of-the-art named-entity recognition tool trained with synthetic data. The textual data in this kind of records are regular in terms of data type (e.g. names, dates, location, and family situation), but show some irregularities: the first column contains the person name, but sometimes her religion as well. Depending on the writer (priest) the name occurs as first name(s) - family name, or the reverse. Dates are also subject to many variations (presence of the year or not, use of Arabic or Latin numbers, abbreviation for the month). In order to cope with these variations, and also to deal with HTR errors, we wrote a text generator that allowed us to generate a large quantity of training data with these variations and errors and then train named-entity recognition tool with it to recognize each word category (first name, last name, death date,. . .).

Finally, we focused on spatio-temporal indexing. Associating each record to a geographical location and a temporal point enables navigation of the data through space and time. We achieved spatial indexing simply, through metadata at the parish level (since we know the parish associated with each book). The second indexing step, i.e. temporal indexing, requires information to be extracted from each record. Even if there is a dedicated column for this in every type of record, its extraction and normalization requires some processing. This is because the string extracted from the image by the HTR must be normalized and converted to a timestamp that a computer can handle. Furthermore, although the month and day are usually written in the record itself, the year may be ‘factorized’ at the page level, might only exist in the first record of the page, or could even occur a few pages before the one being processed (something a human would not have an issue inferring). Detecting this is still noisy and we’re currently working on our model to improve its ability to appropriately extract this data.

4 Future work

We have almost finished processing the Birth and Wedding records, and will be adding them to the demo soon, along with the ability to search across the three types of records. We hope that one outcome of this

upgrade will be the automatic generation of family trees.

Some issues in terms of Information Extraction are still pending and seem to require a specific step to solve them: for instance the idem/ditto sign occurring regularly in columns in order to avoid to repeat the same value (figure 3, last column).

Additionally, a major milestone will be to infer demographic information from these data, such as the population size, and some demographic statistics (birth and death rate, for instance). This can only be computed/estimated when the three types of records are linked, enabling identification of the same person across the three records (Sylvester and Hacker, 2020; Özgür Akgün et al., 2020).

5 Conclusion

In summary, *Vital Records* uses information extracted (via HTR and TR machine-learning technology) from archival documents to enable users to visualize trends through history, and even to track specific events. The demonstrator that we have developed, based on data obtained from German parish records between 1848 and 1878, showcases the spatio-temporal capabilities of the demonstrator.

Acknowledgements

We would like to thanks the reviewers for their interesting feedback. Many thanks to the Archive of the diocese of Passau, Germany for having provided the data and their valuable expertise. Part of this work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943 (project READ). We would also like to thank all archives who contributed to the historical dataset.

References

Günter Mühlberger, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, Hervé Déjean, Markus Diem, Stefan Fiel, Basilis Gatos, Albert Greinoecker, Tobias Grüning, Günter Hackl, Vili Haukkovaara, Gerhard Heyer, Lauri Hirvonen, Tobias Hodel, Matti Jokinen, Philip Kahle, Mario Kallio, Frédéric Kaplan, Florian Kleber, Roger Labahn, Eva Maria Lang, Sören Laube, Gundram Leifert, Georgios Louloudis, Rory McNicholl, Jean-Luc Meunier, Johannes Michael, Elena Mühlbauer, Nathanael Philipp, Ioannis Pratikakis, Joan Puigcerver Pérez, Hannelore Putz, George Retsinas, Verónica Romero, Robert Sablatnig, Joan-Andreu Sánchez, Philip Schofield, Giorgos Sfikas, Christian Sieber, Nikolaos Stamatopoulos, Tobias Strauß, Tamara Terbul, Alejandro H. Toselli, Berthold Ulreich, Mauricio Villegas, Enrique Vidal, Johanna Walcher, Max Weidemann, Herbert Wurster, and Konstantinos Zagoris. 2019. Transforming scholarship in the archives through handwritten text recognition. *Journal of Documentation*, 75(5):954–976.

Animesh Prasad, Hervé Déjean, and Jean-Luc Meunier. 2019. Versatile layout understanding via conjugate graph. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 287–294. IEEE.

Kenneth M. Sylvester and J. David Hacker. 2020. Introduction to special issues on historical record linking. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(2):77–79.

Özgür Akgün, Alan Dearle, Graham Kirby, Eilidh Garrett, Tom Dalton, Peter Christen, Chris Dibben, and Lee Williamson. 2020. Linking scottish vital event records using family groups. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(2):130–146.