Fueling Time Machine: Information Extraction from Retro-Digitised Address Directories

Mohamed Khemakhem, Carmen Brando, Laurent Romary, Frédérique Mélanie-Becquet, Jean-Luc Pinol

Fueling Time Machine: Information Extraction from Retro-Digitised Address Directories

Mohamed Khemakhem[1,2,3]
mohamed.khemakhem@inria.fr

Carmen Brando[4]
carmen.brando@ehess.fr

Laurent Romary[1,2,5]
laurent.romary@inria.fr

Frédérique Mélanie-Becquet[6]
frederique.melanie@ens.fr

Jean-Luc Pinol[7]
jean-luc.pinol@ens-lyon.fr

[1] Inria ALMAnaCH, Paris
[2] Centre Marc Bloch, Berlin
[3] Paris Diderot University, Paris
[4] CRH (EHESS / CNRS UMR 8558), Paris
[5] BBAW - Berlin-Brandenburgische Akademie der Wissenschaften, Berlin
[6] LATTICE, ENS / Paris 3 / CNRS UMR 8094, Paris
[7] LARHRA CNRS UMR 5180, Lyon

Whereas mapping systems, such as Google Maps or Bing, have become nowadays the common tools to geocode addresses or to browse neighborhoods on modern maps, browsing a legacy map representing a geographical snapshot of historical cities is far from being accomplished. The issue is related in the first place to the lack of data allowing a system to map a given address to a throwback location. Such information are abundantly available in dedicated paper resources, such as legacy address directories[1]. But even digitised, mining the content of these resources remains limited due to the ad-hoc employed information extraction techniques.

Time machine[2] is a major large scale project aiming to bridge this gap, among many others, by analysing and valorising the content of legacy documents for the ultimate purpose of redrawing the historical, social and economical heritage of Europe. In this context, we present our approach and first results of a state-of-the-art technique for extracting information from digitised address directories.

[1] Historical maps are evidently an important source of geolocalised information, our proposed approach aims to be complementary to well-known methods for georeferencing old maps and thus deals with a new kind of historical source.
[2] http://timemachineproject.eu

Our labour has been motivated by two emerging factors. First, the public release of several digitised versions in high-definition from the legacy address directories “Annuaires-almanach” of Paris, made available by the French National Library[3]. The directory series, which had been edited since the 18th century, carry a joint description of the commercial activities and postal information of the french capital. Second, a recently implemented approach by Khemakhem et al. 2017 and Khemakhem et al. 2018 has given an information extraction system, GROBID-Dictionaries, which has been designed to structure digitised dictionaric resources by using machine learning models. We have been struck by the similarities in the structures of dictionaries and address directories, where both resources share a semasiological representation. In fact, the latters could be perceived as encyclopedic resource where locations are described as unique concepts.

We have used Text Encoding Initiative (TEI) as a common modeling standard and proposed a first encoding of entries in an address directory. We distinguish between two categories of entries (see table 1). The first is reserved for each entry describing a single occupant in a unique or a shared address. In other terms, to each number in a street, one or many occupants could be assigned and for each one of them corresponds an entry. The second category of entries gathers the description blocs of a common street. An entry in this case encapsulates information like the name of the street, length, neighbouring street, etc.

[3] http://gallica.bnf.fr/ark:/12148/cb32695639f/date

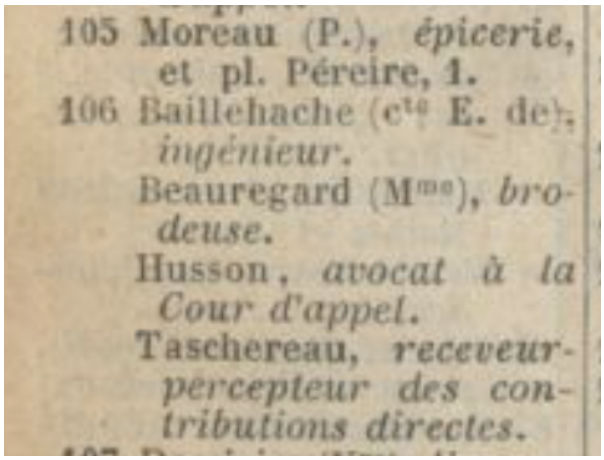
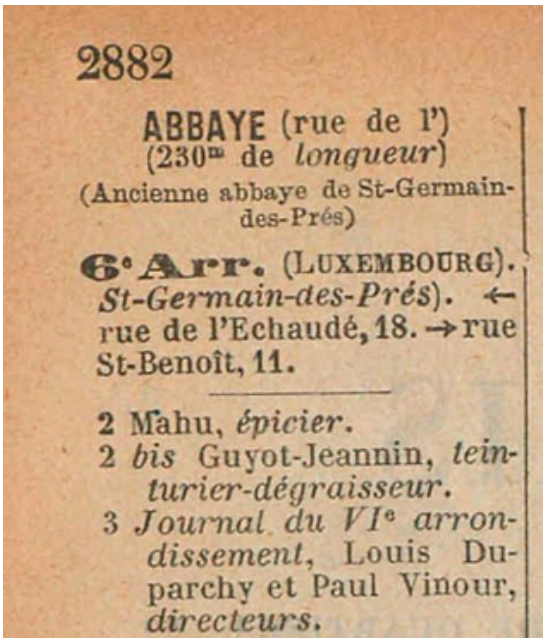
Digitised Sample	TEI Encoding
	<entry> <num>105</num> <form> <persName> <surname>Moreau</surname> <addName>(P.)</addName> </persName> </form> <pc>,</pc> <sense> <def>épicerie</def> <pc>,</pc> <lbl>et</lbl> <address>pl. Péreire, 1.</address> </sense> </entry>
	<entry> <form>ABBAYE (rue de V)</form> <sense>(230” de longueur) (Ancienne abbaye de St-Germain-des-Prés) 6* Arr. (Luxembodrg). St-Germain-des-Prés). <-rue de l'Echaudé, 18.-►rue St-Beno!t, 11</sense> </entry>

Table 1: Both images in lines 2 and 3 correspond respectively to excerpts of pages 3500 and 2882 of the 1901 release[4] of the annuaires-almanach

[4] http://gallica.bnf.fr/ark:/12148/bpt6k9763088f

The current architecture of GROBID-Dictionaries, based on cascading machine learning models, has been to a large extent able to support the presented encoding of the textual information and extract the macro structures. In fact, the first level of segmentation has the mission to differentiate between the different parts of a digitised page. The second level relies on a model for segmenting a page body to entries which will be further segmented in the third level to main semantic blocks.
Despite sometimes the noisy OCRised data (see table 1), till the third level, the only required adaptation of the system has been the implementation of a new label to mark the numbering of entries <num>. After this minor adaptation, a first experimentation of the system has shown interesting results for the first 3 segmentation levels, which we report in table 2.

Model	**Annotated Data**	**F1-Score**	
Dictionary Segmentation	10 Pages 7 training, 3 evaluation	**Micro Average**	**Macro Average**
		99.61	72.12
Dictionary Body Segmentation	319 Entries 270 training, 49 evaluation	**Micro Average**	**Macro Average**
		98.61	95.7
Lexical Entry	208 Entries 160 training, 48 evaluation	**Micro Average**	**Macro Average**
		90.31	91.36

Table2: Evaluation of the first three segmentation models

Although the models had given better results with dictionaries in previous experimentations, the current results are still considered impressive given the different nature of the address directories and the noise in the OCRs, especially for the first model. The outcome should be improved as soon as we annotate more data and further strengthen the selected features, if needed. To reach the complete encoding presented in table 1, we are investigating the creation of new models to be integrated in the existing architecture for processing the clusters of texts labels. Before considering building new models trained from scratch, the integration of models used for the same purpose in the GROBID[5] family projects is likely to be the most efficient solution, such

[5] https://github.com/kermitt2

for the parsing of addresses and person names. We are considering also to improve the OCRs for known entries such as the majority of street names, which could be checked against existing defined lists.

In conclusion, fueling a Time machine with structured information extracted from legacy address directories seems not to be an issue anymore thanks to the availability of the target digitised material and the advanced extraction techniques embedded in GROBID-Dictionaries. The existing architecture of the tool could be further improved by annotating more data, plugging in existing models or creating new ones to be applied in larger scale or on similar documents in other languages. Finally, our aim is to further retrodigitise releases of the Annuaires-almanach and geocode historical postal addresses listed there thereby to analyse commercial activity in old Paris taken from large amounts of historical sources as introduced by Kaplan et al. 2017.

References

1. Kaplan, Frédéric and Isabella di Lenardo. *Big data of the past*. Frontiers in Digital Humanities 4 2017.
2. Khemakhem, Mohamed, Luca Foppiano, and Laurent Romary. *Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields*. eLex 2017.
3. Khemakhem, Mohamed, Axel Herold, and Laurent Romary. *Enhancing Usability for Automatically Structuring Digitised Dictionaries.* GLOBALEX workshop at LREC 2018.