Retro-digitizing and Automatically Structuring a Large Bibliography Collection
David Lindemann, Mohamed Khemakhem, Laurent Romary

To cite this version:
David Lindemann, Mohamed Khemakhem, Laurent Romary. Retro-digitizing and Automatically Structuring a Large Bibliography Collection. European Association for Digital Humanities (EADH) Conference, Dec 2018, Galway, Ireland. <hal-01941534>

HAL Id: hal-01941534
https://hal.archives-ouvertes.fr/hal-01941534
Submitted on 1 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Retro-digitizing and Automatically Structuring a Large Bibliography Collection

David Lindemann, Universität Hildesheim, UPV/EHU University of the Basque Country
Mohamed Khemakhem, Inria ALMAnaCH, Centre Marc Bloch, Paris Diderot University
Laurent Romary, Inria ALMAnaCH, Centre Marc Bloch, Berlin-Brandenburgische Akademie der Wissenschaften

1 Introduction

In this paper, we present a generic workflow for retro-digitizing and structuring large entry-based documents, using the 33,000 entries of Internationale Bibliographie der Lexikographie, by Herbert Ernst Wiegand, as an example (published in four volumes (Wiegand 2006-2014)). The goal is to convert the large bibliography, at present available as collection of images, to TEI compliant XML, a structured format that enables enhanced interoperability and search functionalities (Lindemann, Kliche and Heid, 2018). Images of the printed publication are first processed with Optical Character Recognition (OCR) tools which are part of the Transkribus software application (Mühlberger and Terbul, 2018), the output of which is used for creating manually validated Hand-Written Text Recognition (HTR) training material. The retro-digitised output is the used to train and create dedicated machine learning models in GROBID-Dictionaries (Khemakhem, Foppiano and Romary, 2017), a tool for automatic segmentation of entry-based text documents and representation as TEI-compliant XML. Both Transkribus and GROBID-Dictionaries are tools freely available to the community. Preliminary results suggest that the proposed workflow yields good precision in retro-digitisation and segmentation.

2 Creating a text layer for PDF images

The PDF version of the original resource available to us contains (a) images of the 2,704 pages of the printed publication with a resolution of 300 DPI (cf. example in Fig. 1), and (b), an additional text layer, presumably the output of an OCR engine, that presents a massive amount of encoding errors. The images have been processed using the OCR engine built into Transkribus; as it can be observed in the example shown in Table 1, this version contains considerably less errors: On the one hand, the set of properly recognized characters is much larger, and on the other hand, no mistakes are found regarding whitespaces.

---

1 Available at: http://transkribus.eu
2 Available at: https://github.com/MedKhem/GROBID-Dictionaries
The manually corrected OCR output has been used as training material for the HTR engine available through Transkribus. We point out that we make use of HTR technology for the recognition of printed text despite the fact that OCR engines can also be adapted to special fonts and character sets by training (See e.g. Clausner, Pletschacher and Antonacopoulos, 2014; Springmann, Fink and Schulz, 2016). The specialty of the workflow using Transkribus lies in the combination of OCR software\(^3\) to create automatically a draft transcription, which after manual correction using the Transkribus interface becomes a ground truth version used for training the HTR engine from scratch.\(^4\) In a training set of 104 pages (around 36,600 words), a set of 158 different glyphs is recognized; the character recognition error rate has been reduced to around 0.5%.

3 The OCR engine built in Transkribus, as for July 2018, is ABBYY FineReader 11.
4 The HTR engine built in Transkribus is developed at Computational Intelligence Technology Lab (CITlab) of the University of Rostock (Leifert et al., 2016).

3 The OCR engine built in Transkribus, as for July 2018, is ABBYY FineReader 11.
4 The HTR engine built in Transkribus is developed at Computational Intelligence Technology Lab (CITlab) of the University of Rostock (Leifert et al., 2016).
Table 2: Integrated GROBID models (yellow) in GROBID-Dictionaries architecture

To guarantee TEI compliance of the structures extracted by the hybrid chain of models, we created a second serialization for the Dictionary Body Segmentation model by the introduction of `<bibliStruct>` element (see example in Figure 1).

```xml
<bibliStruct xml:id="21923"
  <analytic>
    <author>
      <persName xmlns="http://www.tei-c.org/ns/1.0">
        <forename type="first">François</forename>
        <surname>Voillat</surname>
      </persName>
    </author>
    <title level="a">Le “Glossaire des patois de la Suisse romande” (GPSR)</title>
  </analytic>
  <monogr>
    <title level="m">Actes du XVIIIe Congrès International de Linguistique et de Philologie Romanes [...] Tome VII</title>
    <imprint>
      <date type="published" when="1989"/>
      <biblScope unit="pp" from="338" to="345">338-345</biblScope>
    </imprint>
  </monogr>
</bibliStruct>
```

Figure 1: Example of a segmented bibliographic entry using the hybrid architecture

Given the fact that all GROBID models rely on text and markup features for information extraction, the aforementioned OCRisation workflow has led to a better input for the segmentation machine learning models, where they are catered with more exact and meaningful text, especially in the case of field separators or structure indicators, such as bullets, arrows, etc.

4 Conclusions and Outlook

The presented workflow has worked out very well for this segmentation and representation as XML of a single entry-based publication available initially as a collection of images. Using the described tool pipeline, the manual effort needed for the production of training material for character recognition and segmentation has been kept considerably low.
The same workflow could be adopted to retro-digitize resources with similar features, i.e. (a) an availability as image of the printed version, (b) a print that contains non-standard fonts and/or characters so that out-of-the-box OCR approaches may lead to noisy results, and (c) an entry-like document structure, as in dictionaries or bibliographies.

5 Bibliography


