Retro-digitizing and Automatically Structuring a Large Bibliography Collection

David Lindemann, Mohamed Khemakhem, Laurent Romary

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),[1] 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[2] (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

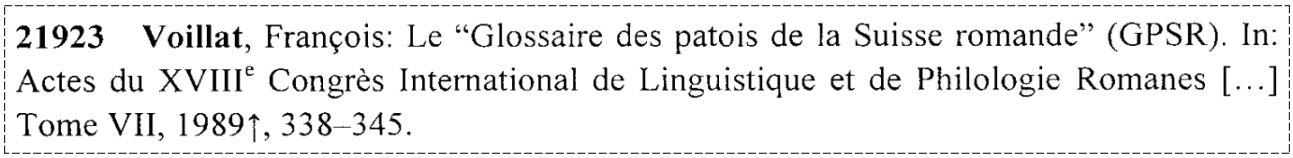
21923 Voillat, François: Le "Glossaire des patois de la Suisse romande" (GPSR). In: Actes du XVIII[e] Congrès International de Linguistique et de Philologie Romanes [...] Tome VII, 1989↑, 338–345.

Fig. 1: Wiegand Bibliography entry #21923, 300 DPI image

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

Original PDF Text Layer	21923·····Voillat,···F·r·a·n·c·i·s·:···Le···"Glossaire···des···patois···de···la···Suisse···r·o·m·a·n·d·e·"···(GPSR).···In:· Actes···du···X·v·î·î·Γ···Congres···International···de···Linguistique···et···de···Philologie·· ·R·o·m·a·n·e·s····[...]· T·o·m·e··VII,···1·9·8·9·|·,···3·3·8·-·3·4·5·.
Transkribus OCR	21923·Voillat,·François:·Le·“Glossaire·des·patois·de·la·Suisse·romande”·(GPSR).·In: Actes·du·XVIIIe·Congrès·International·de·Linguistique·et·de·Philologie·Romanes·[...] Tome·VII,·1989Î,·338-345.
Manual correction	21923·Voillat,·François:·Le·“Glossaire·des·patois·de·la·Suisse·romande”·(GPSR).·In: Actes·du·XVIII[e]·Congrès·International·de·Linguistique·et·de·Philologie·Romanes·[...] Tome·VII,·1989↑,·338-345.

Table 1: Wiegand Bibliography, entry #21923 transcriptions

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 Segmentation and representation as XML

GROBID-Dictionaries (Khemakhem, Foppiano and Romary, 2017; Khemakhem, Herold and Romary, 2018) is a machine-learning infrastructure, which has been initially conceived for extracting lexical information from digitised dictionaries. Given the interest raised in the field of Digital Humanities around the exploitation of entry-based retro-digitised text material, the system has been applied to new categories of documents, such as old manuscript auction catalogues, and legacy address directories (Khemakhem, Brando, *et al.*, 2018; Khemakhem, Romary, *et al.*, 2018). The outcome of these experiments has motivated us to adapt and integrate the information extraction tool in a digitization and segmentation pipeline for the presented entry-based bibliography resource. To this end, we kept the first two models in GROBID-Dictionaries for the recognition of page bodies and their entries and we relied on the pluggable nature of the GROBID models to integrate the bibliographic reference parsing model, a core model in GROBID[5], to structure the recognised entries (see table 2).

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).

5 See https://grobid.readthedocs.io/en/latest/grobid-04-2015.pdf.

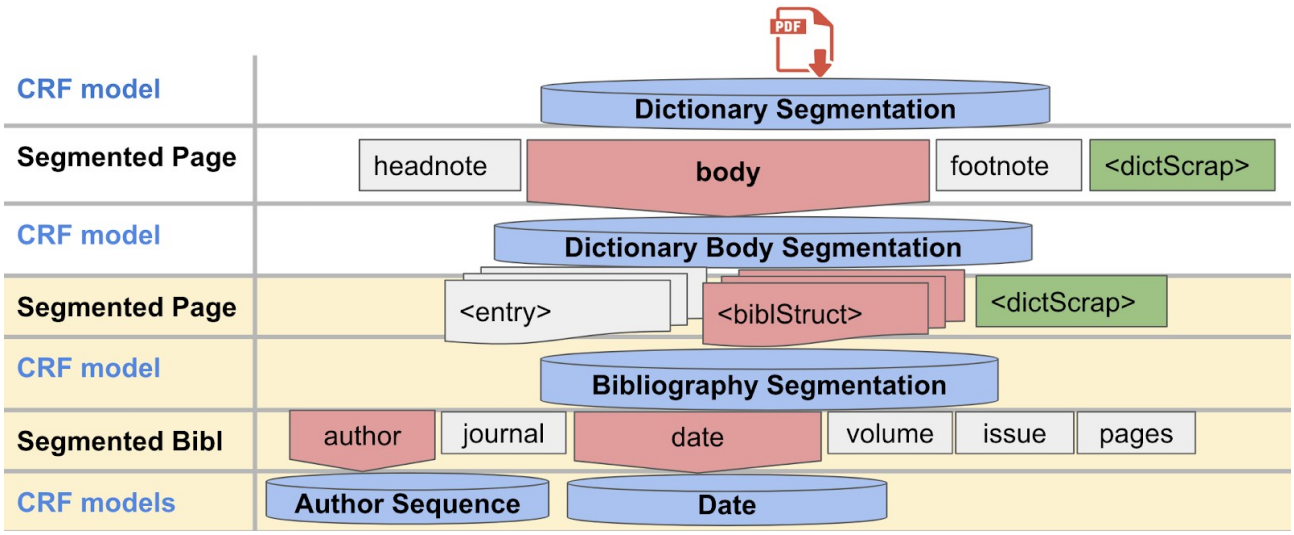

Table 2: Integrated GROBID models (yellow) in GROBID-Dictionaries architecture

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 <biblStruct> element (see example in Figure 1).

```
<biblStruct 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>
 </biblStruct>
```

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

Clausner, C., Pletschacher, S. and Antonacopoulos, A. (2014) 'Efficient OCR Training Data Generation with Aletheia', in *Proceedings of the International Association for Pattern Recognition (IAPR)*. Tours (France), pp. 7–10.

Khemakhem, M., Romary, L., *et al.* (2018) 'Automatically Encoding Encyclopedic-like Resources in TEI', in *Proceedings of The 18th annual Conference and Members Meeting of the Text Encoding Initiative Consortium*. Tokyo.

Khemakhem, M., Brando, C., *et al.* (2018) 'Fueling Time Machine: Information Extraction from Retro-Digitised Address Directories', in *Proceedings of The 8th International Conference of Japanese Association for Digital Humanities*. Tokyo.

Khemakhem, M., Foppiano, L. and Romary, L. (2017) 'Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields', in Kosem, I. et al. (eds) *Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017. eLex 2017 conference, 19-21 September 2017, Leiden, The Netherlands*, Leiden: Lexical Computing. Available at: https://elex.link/elex2017/wp-content/uploads/2017/09/paper37.pdf.

Khemakhem, M., Herold, A. and Romary, L. (2018) 'Enhancing Usability for Automatically Structuring Digitised Dictionaries', in *GLOBALEX workshop at LREC 2018*. Miyazaki, Japan. Available at: https://hal.archives-ouvertes.fr/hal-01708137 (Accessed: 20 April 2018).

Leifert, G. *et al.* (2016) 'CITlab ARGUS for historical handwritten documents', *Computing Research Repository*. Available at: http://arxiv.org/abs/1605.08412.

Lindemann, D., Kliche, F. and Heid, U. (2018) 'Lexbib: A Corpus and Bibliography of Metalexicographcal Publications', in *Proceedings of EURALEX 2018*. Ljubljana, pp. 699–712. Available at: http://euralex.org/publications/lexbib-a-corpus-and-bibliography-of-metalexicographical-publications/.

Mühlberger, G. and Terbul, T. (2018) 'Handschriftenerkennung für historische Schriften. Die Transkribus Plattform', *b.i.t.*, 21(3), pp. 218–222.

Springmann, U., Fink, F. and Schulz, K. U. (2016) 'Automatic quality evaluation and (semi-) automatic improvement of OCR models for historical printings', *Computing Research Repository*. Available at: http://arxiv.org/abs/1606.05157 (Accessed: 12 June 2018).

Wiegand, H. E. (2006a) *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung, Band 1: A-H*. Berlin, Boston: De Gruyter. doi: 10.1515/9783110892215.

Wiegand, H. E. (2006b) *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung, Band 2: I-R*. Berlin, Boston: De Gruyter. doi: 10.1515/9783110892918.

Wiegand, H. E. (2007) *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung, Band 3: S-Z*. Berlin, Boston: De Gruyter. doi: 10.1515/9783110892925.

Wiegand, H. E. (2014) *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung, Band 4: Nachträge*. Includes a print version and an ebook. Berlin, Boston: De Gruyter. doi: 10.1515/9783110403145.