READ

RECOGNITION & ENRICHMENT
OF ARCHIVAL DOCUMENTS

D6.9
Table and Form Analysis Tool P3

Florian Kleber, Markus Diem and Stefan Fiel
CVL

Distribution: http://read.transkribus.eu/

READ
H2020 Project 674943

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943

Project ref no.	H2020 674943
Project acronym	READ
Project full title	Recognition and Enrichment of Archival Documents
Instrument	H2020-EINFRA-2015-1
Thematic priority	EINFRA-9-2015 - e-Infrastructures for virtual research environments (VRE)
Start date/duration	01 January 2016 / 42 Months

Distribution	Public
Contract. date of delivery	31.12.2018
Actual date of delivery	28.12.2018
Date of last update	18.12.2018
Deliverable number	D6.9
Deliverable title	Table and Form Analysis Tool P3
Type	Report, Demonstrator
Status & version	Final report
Contributing WP(s)	WP4, WP6
Responsible beneficiary	CVL
Other contributors	CVL, ABP, NaverLabs, UPVLC
Internal reviewers	Naverlabs, NCSR
Author(s)	Florian Kleber, Markus Diem and Stefan Fiel
EC project officer	Christopher DOIN
Keywords	table analysis, table matching, forms analysis

Contents

1 Executive Summary **4**

2 Annotated Datasets and Planned Table Recognition Competition **4**

3 Table Matching **5**

4 Final Workflow for Table Analysis **6**

5 Towards Layout-agnostic Information Extraction from Untranscribed Handwritten Table Images **7**

6 Future Work **8**

1 Executive Summary

Due to the presence of structured documents in archives (tables) task 6.3 analyzes tables. Based on the GT definition presented in Deliverable D6.7 a dataset consisting of the documents of the Passau Diocesan Archives has been created, which is also used for the evaluation and is also a basis for the Large Scale Demonstrator (LSD). Since the dataset mainly consists of hand-drawn tables a high variation of the column width and rows height is present. To be capable to deal with this variation a new approach based on association graphs has been developed in D6.8, see [1]. The method detects the table region, the table columns and the header based on the line information using a specified template. Currently, the template is selected manually. This method is combined with the approach of Naverlabs to detect the rows (since separators are manually drawn or are missing, the baselines of the text are analyzed to detect the rows). A metric has also been defined for the evaluation of the detected table structure. The input is a page xml defining the table/form structure and the output is the alignment of the template to the current document image. Within the last year the main focus was the improvement of the presented methodology regarding table variation and noise robustness. Additionally, the annotation guidelines for annotating archival documents have been defined. Based on the newly created dataset a competition proposal for ICDAR 2019 has been drafted. Also the interface for Transkribus has been realized.

Additionally, UPVLC proposed a layout-agnostic information extraction from untranscribed handwritten table images.

Section 2 describes the GT dataset. Section 3 gives an overview of the methodology and the results of the table matching, while Section 4 shows the general workflow for table analysis. The layout agnostic information extraction from UPVLC is presented in Section 5. The future work is presented in Section 6. All modules including the Transkribus API are part of the CVL READ Framework. It is Open Source under LGPLv3 and available at github: `https://github.com/TUWien/ReadFramework`.

2 Annotated Datasets and Planned Table Recognition Competition

For the evaluation of Clinchant et al. [2] and Kleber et al. [1] (subset) the READ ABP Table Dataset has been annotated and published on Zenodo, see `https://doi.org/10.5281/zenodo.1226878`. The dataset holds a total of 26,579 scanned pages. On 4,578 pages, the requested information was recorded in manually drawn tables or manually extended table prints. For a detailed description see *D5.10 ScriptNet Large Scale Dataset P3*.

To organize a competition on table detection and recognition the ScriptNet Table Dataset was created. The entire dataset consists of 1142 images and has annotations of the table region, cells, table header, etc. A detailed description of the dataset is given again in *D5.10 ScriptNet Large Scale Dataset P3*. It is planned to use the Dataset for a competition planned in conjunction with ICDAR 2019. Thus, the dataset will be published in 2019.

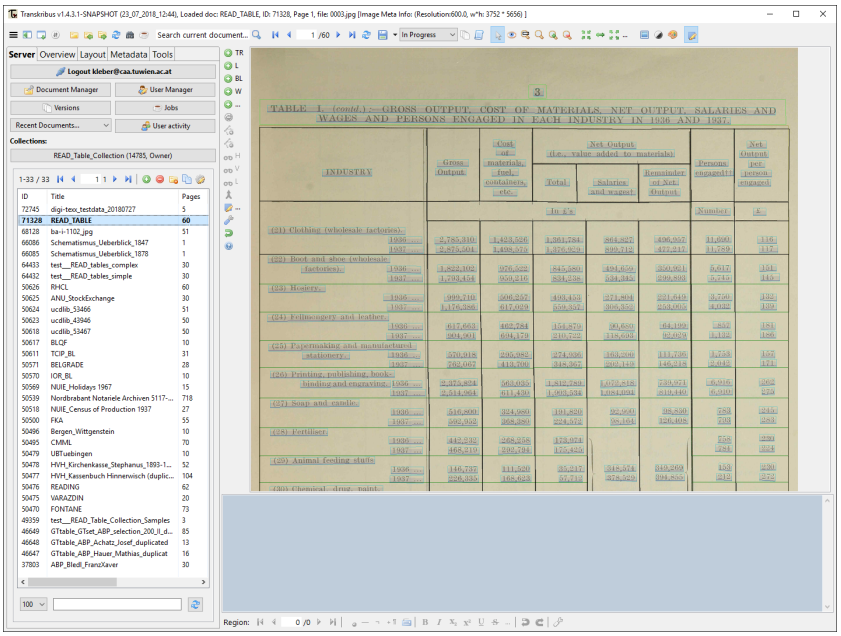

Figure 1: Sample page of table dataset with table caption/table headers/cells and baselines annotated.

Figure 1 shows an example image of the ScriptNet dataset. For both datasets the GT is stored as extended PAGE XML. A minimal sample of a PAGE XML is shown in Listing 1.

The *TableRegion* region contains a unique table id and the coordinates of the table region (quadrilateral). A *TableRegion* can contain any number of *TableCell*s. Each cell has a *row* and *col* attribute defining the row and the column. Each row and each column represents a set of related data. A cell can span several rows and several columns which is defined with the attributes *rowSpan* and *colSpan*. Four additional attributes (*leftBorderVisible, rightBorderVisible, topBorderVisible, bottomBorderVisible* define if border separators are present. Each cell can contain several *Textline*s where each TextLine consists of *Baseline*s. Additionally to the contents of each cell, each text region containing baselines outside the table are annotated as well.

3 Table Matching

The methodology is described in D6.8. The improvements done in 2018 improves the robustness against noise and variation of table layouts. The results of the table matching are presented in Kleber et al. [1] (ICFHR 2018) and are summarized in table 1:

In overall 142 documents with 5 different table layouts have been used. The GT was annotated manually. It can be seen that the MCM is 88.28% which shows a reliable table matching. The errors results from the outermost borders, which can be detected

Listing 1: Minimal sample of a PAGE XML.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<PcGts
  xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15
  http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15/pagecontent.xsd">
    <Metadata>
      <Creator>CVL</Creator>
      <Created>2017-07-24T08:36:56+07:00</Created>
      <LastChange>2018-10-24T07:40:33Z</LastChange>
    </Metadata>
<Page imageFilename="document.tif" imageWidth="4284" imageHeight="3390">
      <TableRegion id="Table_1496312825286_86" custom="readingOrder {index:0;}">
            <Coords points="145,286 4115,286 4115,3227 145,3227"/>
            <TableCell id="TableCell_1496312851662_91" row="0" col="0" rowSpan="1" colSpan="1"
                                                leftBorderVisible="true" rightBorderVisible="true"
                                                topBorderVisible="true" bottomBorderVisible="true">
                <Coords points="145,286 148,514 511,522 500,282"/>
                <CornerPts>0 1 2 3 </CornerPts>
                <TextLine id="line_1500446726051_1" custom="readingOrder {index:0;}">
                    <Coords points="207,368 447,365 448,415 208,418"/>
                    <Baseline points="208,413 448,410"/>
                    <TextEquiv>
                        <Unicode>Name.</Unicode>
                    </TextEquiv>
                </TextLine>
            </TableCell>
                                                ...
      </TableRegion>
</Page>
</PcGts>
```

Table 1: Evaluation of the proposed table matching.

	ABP_GT dataset
MTM	0.9785
JI (Table)	0.9305
MCM	0.8828
JI (Cell)	0.8374
USeg	0.0545
Miss	0.0761

as the documents borders, which lead to a wider first and last column. For a detailed description see Kleber et al. [1].

4 Final Workflow for Table Analysis

Figure 2 shows the final workflow for the table processing.

The first part is the definition of a table template which is applied to a document image using the proposed methodology. Afterwards, the method of Clinchant et al. [2] is applied to detect the table rows which leads to the final table structure which can be used for information extraction. The table matching has been integrated in the Transkribus API.

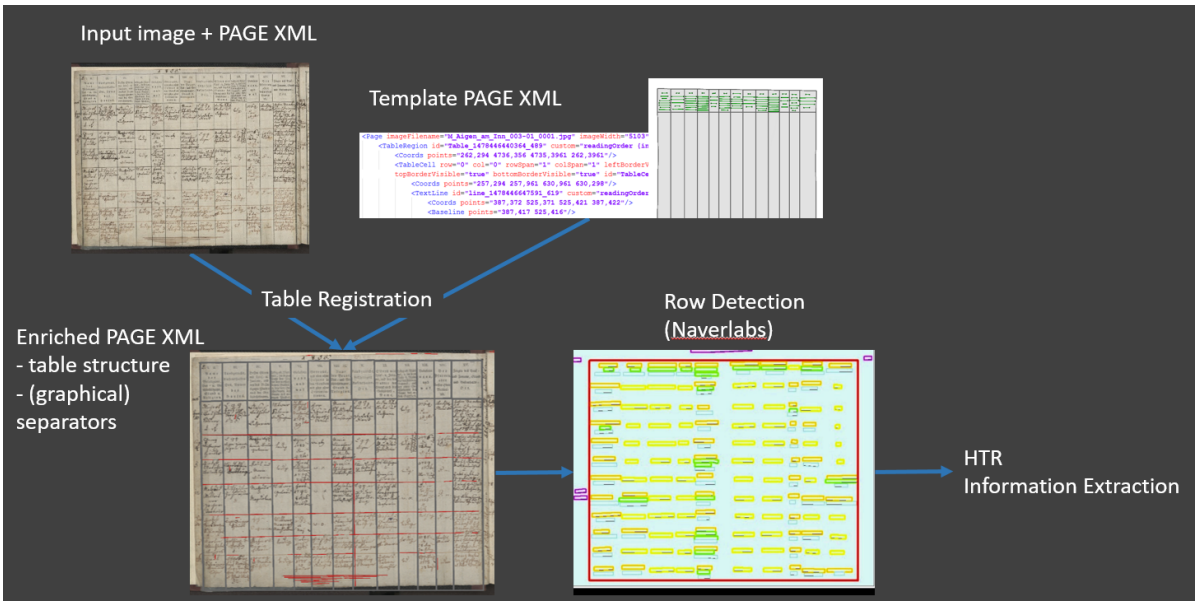

Figure 2: Workflow of table processing.

5 Towards Layout-agnostic Information Extraction from Untranscribed Handwritten Table Images

Traditional approaches to Information Extraction from text documents assume the document layout is known, or they try first to somehow obtain a correct layout. In addition, these approaches assume the document text is also exactly known. This paradigm is also the traditional one when the document contain tabular information even though table layouts may be more complex, variable and challenging than other simpler types of document layout.

This reasoning seems adequate when tables in the collection of interest exhibit sufficiently clear and homogeneous layout. In this case, one can somehow build *table templates* for the (few) types of table layouts expected in order to allow extracting subimages of columns, rows and/or individual cells. This may actually be the right way to go for table collections with *preprinted table layouts* and *printed or typed text contents*.

However, the situation encountered in millions of historic archival documents is quite different. In these documents, *handwritten tables*, with casual, *hand-drawn layouts* (or even "virtual" layouts, where no column or raw guide lines have been physically drawn) must be considered. Moreover, the text in the table cells is often unclear, abbreviated and rather inconsistently follow the expected cell confinement. Under these circumstances the above paradigm becomes elusive. First, *no* reliable table templates can be built and second, the textual contents in the table cells is uncertain (stochastic, technically speaking), rather than exactly known.

In this work we explore a radically different approach to this problem. Rather than attempting to start determining layout information, first, we obtain a probability map of the possible words appearing in every image pixel (or every sufficiently small image region). Then, we make only shallow assumptions as to how textual information in a

table is generally organized.

In previous works we have developed the idea of using word probability maps, which we refer to as "word probabilistic indexes" (PIs), to support fast and accurate keyword spotting in large collections of handwritten text images. For instance, in [3] we carried out preparatory experiments aiming at large-scale indexing a historical German collection of manuscript parish records. The results of these experiments, in terms of the standard precision-recall metrics used in keyword spotting, were good and clearly supported the feasibility of indexing the large collection aimed at. The resulting PIs permit efficient and effective search for individual words as well as for conventional combined-word queries [3], as can be seen in the web on-line demonstrator at `http://transcriptorium.eu/demots/kws-Passau`.

In addition, and more important in this deliverable, in [3] we went one step further and explored the use of PIs to support structured queries for *information extraction* from untranscribed handwritten images containing tabular data. Rather than relying on previous accurate layout analysis or table template matching (rather hopeless aims for handwritten tables, as discussed above), we just assume that information in tables is roughly organized in *vertical columns*. Each column is assumed to have a *heading*, containing words which describe the type of information expected in this column. The information itself or *"content"* is arranged in *cells* located below the column heading and, in many cases, cells are further collocated along *horizontal rows*. Rows may provide relations between cells data; for instance, a row of cells may constitute a *record* which assembles all the data associated to a birth or a marriage, as described by the headings of the corresponding columns.

To assess the potential of this idea, preliminary experiments were carried out in [3] using a relatively small but complex and representative subset of documents from the collection aimed at (200 page images, about half of which are complex tables). The task was to honor information extraction queries of the form: "(`column-heading, column-content`)", where `column-heading` is an AND-combination of column heading words and `column-content` is a (single) keyword. For example, the query (NAMEN DER BRAUT, MARIA) (*name of the bride*, *Maria*, in English), should retrieve all the table columns which include the content word "Maria" below the three words "namen" "der" and "braut", clustered in the corresponding column headings. Interestingly, headings geometry need not be determined before hand, but become instead geometrically self-defined by the bounding boxes of the spotted heading words.

Results of these experiments, reported in [3] were very good, assessing the viability and adequateness of the proposed approach.

6 Future Work

The proposed methodology allows information extraction of a bunch of documents which use the same table layout. Currently, the template must be defined manually. As a future work machine learning will be used to automatically extract the table template from a collection. This will allow for a fully automated workflow.

References

[1] F. Kleber, M. Diem, H. Dejean, J.-L. Meunier, and E. Lang, "Matching table structures of historical register books using association graphs," in *Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018, pp. 217–222.

[2] S. Clinchant, H. Dejean, J. Meunier, E. M. Lang, and F. Kleber, "Comparing machine learning approaches for table recognition in historical register books," in *13th IAPR International Workshop on Document Analysis Systems (DAS)*, April 2018, pp. 133–138.

[3] E. Lang, J. Puigcerver, A. H. Toselli, and E. Vidal, "Probabilistic indexing and search for information extraction on handwritten german parish records," in *Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Niagara Falls, USA, 2018, pp. 44–49.