

D6.8 Table and Form Analysis Tool P2

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. | 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 re- search environments (VRE) | |
| Start date/duration | 01 January 2016 / 42 Months | |

| Distribution | Public | |
|--------------------------|--|--|
| Contract. date of deliv- | 31.12.2017 | |
| ery | | |
| Actual date of delivery | 28.12.2017 | |
| Date of last update | 16.11.2017 | |
| Deliverable number | D6.8 | |
| Deliverable title | Table and Form Analysis Tool P2 | |
| Type | Report, Demonstrator | |
| Status & version | in progress | |
| Contributing WP(s) | WP4, WP6 | |
| Responsible beneficiary | CVL | |
| Other contributors | CVL, ABP, NaverLabs | |
| Internal reviewers | Naverlabs, NCSR | |
| Author(s) | Florian Kleber, Markus Diem and Stefan Fiel | |
| EC project officer | oject officer Martin MAJEK | |
| Keywords | table analysis, table matching, forms analysis | |

Contents

| 1 | 1 Executive Summary | | | | | |
|---|--|--------------------|--|--|--|--|
| 2 | 2 Ground Truth Dataset of Handwritten Tables 3 Evaluation | | | | | |
| 3 | | | | | | |
| 4 | Table Matching Methodology and Results4.1Matching Table Structure Using Association Graphs | 6 6 9 | | | | |
| 5 | Future Work | 9 | | | | |

1 Executive Summary

Due to the presence of structured documents in archives (forms, tables) task 6.3 analyzes tables and forms. 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 D6.7 has been developed. 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.

Section 2 describes the GT dataset, while the evaluation metric is presented in Section 3. Section 4 gives an overview of the methodology and the results of the table matching. The future work is presented in Section 5. All modules are part of the CVL READ Framework. It is Open Source under LGPLv3 and available at github: https://github.com/TUWien/ReadFramework.

2 Ground Truth Dataset of Handwritten Tables

The structure of tables/forms is represented as extended PAGE XML. The extension of the PAGE XML was defined in Deliverable D6.7 by CVL and Naverlabs and is also used by the table editor of Transkribus developed by UIBK. To evaluate the table analysis and for the Large Scale Demonstrator (LSD) a GT dataset has been created together with ABP and Naverlabs. The following paragraph describes the new dataset of the Passau Diocesan Archives.

The $ABP_S_1847 - 1878$ dataset contains information about the parishioners who died within the geographic boundaries of the various parishes of the Diocese of Passau between the years 1847 and 1878. The dataset holds a total of 26,579 scanned pages. According to the official order of the Catholic Church, the parish scribes had to record name, profession, religion, court, address, marital status, reason of death, dates of death and burial, age, names of doctor and priest as well as additional information in written form. The images display the records mainly in tabular format referring to one person per row. A thorough analysis of the dataset shows that for 22,001 images 88 different table prints were used. These unique layouts were further categorized into eleven template categories. The vast majority of scans (15,147 images) even fall into one single template category. Based on this collection a manual annotation of the table information has been done for 210 pages which is used as GT dataset.

Figure 1 shows an annotated table of the ABP dataset in Transkribus. The annotation has been done with the table editor.

| ccert document cetter document cetter document cetter document cetter construct constru | Important Important | Kensacat Jose Jo | Up - Tu - Tu - Tu - Tu - Tu - Tu - Tu - Tu |
|---|--|--|--|
| Cecent document Lections Los / 36 Los / | Courses C | Note Constrainty Constrainty Pages Upleader Toges Upleader Toges Upleader Toges Upleader Toges Upleader Toges exalang@bit. | Up - Tu - Tu - Tu - Tu - Tu - Tu - Tu - Tu |
| ccert document 136 / 36 10 136 / 36 10 10 136 / 36 10 10 110 12031 12033 1110 1 | Netross Add Mach of 12000. Rule Add Mach of 12000. Rule Rest N Add Mach of 12000. Rule Rest N Add Mach of 12000. Rule Rest Rest School Addressing Machine Rest Re | Buter schriety Pages Upleader To calling@bits. enaling@bits. | Up Tu Tu Tu Tu Tu Tu Tu Tu Tu Tu |
| cccrt document lections: 1-16 / 36 H ID Take 20203 HTR.1 20203 HTR.2 20203 HTR.2 20203 HTR.2 20203 HTR.2 19565 ABP.2 19586 ABP.2 19587 ABP.2 19577 ABP.3 19577 ABP.3 19577 ABP.2 19573 ABP.2 19574 ABP.2 19575 ABP.2 19574 ABP.2 19575 ABP.2 19574 ABP.2 19575 ABP.2 19574 ABP.3 19575 ABP.2 19574 ABP.2 19575 ABP.2 19574 ABP.3 19575 ABP.4 19576 ABP.4 19577 ABP.4 19578 ABP.5 19574 ABP.4 | ADP HALD CT (2010, FM ADP HALD CT (2010, FM A B A B | rder) Pages Upplender Pages Upplender 15 ers.hog@bit. 16 ers.hog@bit. 16 ers.hog@bit. 20 ers.hog@bit. | Up - Tu - Wi - Tu - Tu |
| Initial and the second secon | ABP READ OT COSIN, Ru All H LIN H M OLIVIE H H R Testist, Stoletis R Testist, Stoletis R Testist, Cosletis R Stalid, Bulkensburge S Stalid, S Stalid P Stalid, S St | Image Upleader 1 collarg@bbt. 60 collarg@bbt. 61 collarg@bbt. 62 collarg@bbt. 63 collarg@bbt. 64 collarg@bbt. 65 collarg@bbt. 66 collarg@bbt. 67 collarg@bbt. 68 collarg@bbt. 69 collarg@bbt. 60 collarg@bbt. 61 collarg@bbt. 62 collarg@bbt. 63 collarg@bbt. 64 collarg@bbt. 65 collarg@bbt. 66 collarg@bbt. 67 collarg@bbt. 68 collarg@bbt. 69 collarg@bbt. 60 collarg@bt. 61 collarg@bt. | Up - Tu - Wi - Tu - Tu |
| 1-36 / 36 14 10 Title 2003 HTR 1 2003 HTR 1 2003 HTR 1 2003 HTR 1 2003 HTR 1 1906 ABP 2 1907 | A BP FEAD GT (250), Rui Rui (11) (11) (12) (12) (12) (12) Ruiss (2004) Ruiss (2004) | reder) Pages Uploader Pages Uploader s coshargabet. Pages uploader s coshargabet. s coshargabet. c cosh | Up Tu We Tu Tu Tu Tu Tu Tu Tu Tu Tu Tu |
| 10 / 36 / 14 17 10 Take 2000 HTR_1 2003 HTR_1 2003 HTR_1 19585 ABP_5 19584 ABP_5 19585 ABP_5 19586 ABP_5 19580 ABP_5 19581 ABP_5 19575 ABP_6 19575 ABP_6 19575 ABP_6 19575 ABP_6 19575 ABP_6 19577 ABP_6 19578 ABP_6 19577 ABP_6 19578 ABP_6 19577 | K K K | Poges Upleader 15 exalseggbbia; 16 exalseggbbia; 17 exalseggbbia; 18 exalseggbbia; 19 exalseggbbia; 20 exalseggbbia; | Up - Tu - Wi - Tu - Tu |
| 1:36 / 36 14 10 Table 10 Table 20038 HTR, 1 20038 HTR, 1 20031 HTR, 1 19564 ABP, 2 19584 ABP, 2 19584 ABP, 2 19585 ABP, 2 19581 ABP, 2 19575 ABP, 2 19575 ABP, 3 19575 ABP, 3 19575 ABP, 3 19576 ABP, 3 19577 ABP, 1 19577 ABP, 3 19577 ABP, 4 19577 ABP, 3 19577 ABP, 4 19577 ABP, 1 19577 ABP, 1 19577 ABP, 2 19577 ABP, 1 19577 ABP, 2 19577 ABP, 2 19577 ABP, 2 19577 ABP, 3 19577 ABP, 4 19577 ABP, 4 <t< td=""><td>A lange of the second s</td><td>Pages Upleader 1 exalsagübez, 60 exalsagübez, 61 exalsagübez, 80 exalsagübez,</td><td>Up - Tu - Wi - Wi - Tu - Tu - Tu - Tu - Tu - Tu - Tu - Tu - Tu - Tu</td></t<> | A lange of the second s | Pages Upleader 1 exalsagübez, 60 exalsagübez, 61 exalsagübez, 80 exalsagübez, | Up - Tu - Wi - Wi - Tu - Tu - Tu - Tu - Tu - Tu - Tu - Tu - Tu - Tu |
| ID Table 20003 HTR_1 20003 HTR_1 20031 HTR_1 19585 ABP_2 19586 ABP_2 19581 ABP_2 19583 ABP_3 19584 ABP_3 19585 ABP_3 19581 ABP_3 19575 ABP_3 19576 ABP_3 19577 ABP_3 19574 ABP_4 19574 ABP_4 19575 ABP_4 19576 ABP_4 19577 ABP_4 19574 ABP_5 19575 ABP_5 19576 ABP_6 19577 ABP_6 19578 ABP_6 19579 | In Franker, Southers R., Tankar, J., Kimorgu, J., Hanna, Baptist B., Stindar, J., Shannaka, Parkar, S., Santa- B., Shannakar, S., Santa- B., Santa-Barg, S., Santa-Barg, S., Santa- Bar, S., Santa-Barg, S., Santa- B., Santa-Barg, S., Santa- Santa-Barg, S., Santa- B., Santa-Barg, S., Santa- B., Santa-Bard, Santa- Santa-Bard, Santa-Bard, Santa- Bard, Santa-Bard, Santa-Bard, Santa- Bard, Santa-Bard, Santa-Bard, Santa- Bard, Santa-Bard, Santa-Bard, Santa- Santa-Bard, Santa-Bard, Santa- Santa-Bard, Santa-Bard, Santa- Santa-Bard, Santa-Bard, Santa-Bard, Santa- Bard, Santa-Bard, Santa-Bard, Santa- Bard, Santa-Bard, Santa-Bard, Santa- Bard, Santa-Bard, Santa-Bard, Santa- Bard, Santa-Bard, Santa-Bard, Santa-Bard, Santa- Bard, Santa-Bard, Santa-Bard, Santa-Bard, Santa-Bard, Santa- Bard, Santa-Bard, Santa- Bard, Santa-Bard, Santa-Bard, Santa-Bard, Santa-Bard, S | Pages Upleader 15 exalang@bist. 16 exalang@bist. 10 exalang@bist. 20 exalang@bist. 21 exalang@bist. 22 exalang@bist. 23 exalang@bist. 24 exalang@bist. 25 exalang@bist. 26 exalang@bist. 27 exalang@bist. 28 exalang@bist. 29 exalang@bist. 29 exalang@bist. 20 exalang@bis | Up - Tu - Wi - Tu - Tu - Tu - Tu - Tu - Tu - Tu - Tu - Tu - Tu |
| 20003 HTR_1 20033 HTR_1 20033 HTR_1 19585 ABP_5 19504 ABP_5 19505 ABP_5 19506 ABP_5 19507 ABP_6 19571 ABP_6 19575 ABP_6 19576 ABP_6 19577 ABP_6 19573 ABP_6 19574 ABP_6 19575 ABP_6 19576 ABP_6 19577 ABP_6 19578 ABP_6 19579 ABP_6 19570 ABP_6 19571 ABP_6 19572 ABP_6 19573 ABP_6 19574 ABP_6 19575 ABP_6 19576 <td>R. Tested, Kinoya, Johana, Bapist R. Tested, Kinoya, Johana, Bapist R. Tested, Complete P. Poppinger, Mohael, S. P. Johnes, Comp. S. P. Schniedger, Joseph S. P. Schniedger, J. Schnied, S. P. Mach, McNool, S. P. Holmer, Kurl, S. P. Schniedger, Smars, S. P. Facher, McNael, S.</td> <td>15 evalang@bist. 60 evalang@bist. 11 evalang@bist. 30 evalang@bist. 31 evalang@bist. 32 evalang@bist. 33 evalang@bist. 34 evalang@bist. 35 evalang@bist.</td> <td>- Tu - Wo - Tu - Tu - Tu - Tu - Tu - Tu - Tu - Tu</td> | R. Tested, Kinoya, Johana, Bapist R. Tested, Kinoya, Johana, Bapist R. Tested, Complete P. Poppinger, Mohael, S. P. Johnes, Comp. S. P. Schniedger, Joseph S. P. Schniedger, J. Schnied, S. P. Mach, McNool, S. P. Holmer, Kurl, S. P. Schniedger, Smars, S. P. Facher, McNael, S. | 15 evalang@bist. 60 evalang@bist. 11 evalang@bist. 30 evalang@bist. 31 evalang@bist. 32 evalang@bist. 33 evalang@bist. 34 evalang@bist. 35 evalang@bist. | - Tu - Wo - Tu - Tu - Tu - Tu - Tu - Tu - Tu - Tu |
| 20033 HTR_1 19564 ABP_5 19564 ABP_5 19564 ABP_5 19584 ABP_5 19584 ABP_5 19584 ABP_5 19584 ABP_5 19587 ABP_5 19587 ABP_5 19576 ABP_5 19577 ABP_6 19573 ABP_6 19574 ABP_6 19577 ABP_8 19574 ABP_6 19574 ABP_6 19576 ABP_6 19577 ABP_8 19574 ABP_6 19576 ABP_6 19577 ABP_6 19578 ABP_7 19579 ABP_6 19570 Thee, ABP_6 | R: Institute, Nationages, Johann, Bayrist R; Tectorisk, complete IP: Strind, Johannabeptint S. P: Poppinger, Michael S. P: Weber, Genegy, S. P: Schminger, Joseph, S. P: Schminger, Joseph, S. P: Schalpringer, Johan, S. P: Johanger, Michael S. P: Johanger, Joseph, S. P: Johanger, Joseph, S. P: Johanger, Joseph, S. P: Johang, Joseph, Johan, S. P: Johang, Joseph, Johan, S. P: Johang, Joseph, Johan, S. P: Johang, Joseph, Johan, S. P: Joseph, Michael, S. | evalang@bint. | - Wo - Tu - Tu - Tu - Tu - Tu - Tu - Tu - Tu |
| 2003 HTR_1 19564 ABP_2 19585 ABP_2 19585 ABP_2 19585 ABP_2 19585 ABP_2 19585 ABP_2 19585 ABP_2 19586 ABP_2 19586 ABP_2 19586 ABP_2 19576 ABP_2 19577 ABP_3 19575 ABP_2 19575 ABP_2 19574 ABP_3 19575 ABP_2 19574 ABP_3 19575 ABP_2 19574 ABP_3 19575 ABP_3 19572 ABP_3 19573 ABP_4 19574 ABP_3 19575 ABP_4 19576 ABP_4 19577 ABP_6 19578 ABP_6 17699 ABP_6 17609 ABP_6 | R. TetSet (complete P popprez (Michael S Weber, Greeg S 9, Samberger, Joseph S 9, Samberger, Joseph S 9, Schmidseler, Joseph S 9, Schmidseler, Joseph S 9, Schmidseler, Joseph S 9, Patienty, Schult, S 9, Patienty, Schult, S 9, Patienty, Schult, S 9, Patienty, Schult, S 9, Faster, Michael S 9, Faster, Michael S | evalang@biat. | - Wo - To - To - To - To - To - To - To - T |
| Harrison ABP 19554 ABP 19554 ABP 19554 ABP 19554 ABP 19554 ABP 19554 ABP 19550 ABP 19551 ABP 19557 ABP 19575 ABP 19577 ABP 19573 ABP 19574 ABP 19575 ABP 19576 ABP 19577 ABP 19578 ABP 19579 ABP 19574 ABP 19575 ABP 19576 ABP 19577 ABP 19578 ABP 19579 ABP 19570 ABP 19571 ABP 19572 ABP 17669 ABP 1768 ABP | re_zerenzy animarkepüts P. Oppinger_Unknels P. Whene_Encergs P. Samberger_Useph,S P. Scheipringer_Antan,S P. Pieringer_Antan,S P. Pieringer_Antan,S P. Pieringer_Antan,S P. Nacher,S P. Nacher,S P. Nacher,S P. Scheipringer_Smith,S P. Frankerburger_Smith,S P. Frankerburger_Smith,S P. Frankerburger_Smith,S P. Frankerburger_Smith,S P. Frankerburger_Smith,S P. Schetz,Michnel,S | evalang@bist. | - Tu - Tu - Tu - Tu - Tu - Tu - Tu - Tu |
| ABP_U 19583 ABP_X 19583 ABP_X 19584 ABP_X 19585 ABP_X 19586 ABP_X 19587 ABP_X 19578 ABP_X 19579 ABP_X 19576 ABP_X 19577 ABP_X 19573 ABP_X 19574 ABP_X 19575 ABP_X 19570 GTset, 17766 ABP_C 17669 ABP_L | yoyana (Jimude) Weber, Georg S Ramberger, Joseph S Samberger, Joseph S Schniegkerd, Joseph S Schlappinger, Anton S Parkel Jahob S Pathole S Model (McNed S Frankraterger Simon S Frankraterge Simon S Frankraterge Simon S | calangbis. evalangbis. | - To - To - To - To - To - To - To - To |
| Hall Hall 19502 ABP 19503 ABP 19504 ABP 19505 ABP 19506 ABP 19507 ABP 19578 ABP 19575 ABP 19575 ABP 19576 ABP 19577 ABP 19578 ABP 19579 ABP 19570 ABP 19571 ABP 19572 ABP 19570 GTset, 17769 ABP 17669 ABP 17668 ABP | P. Stamborger (Joseph, S P. Schnipforger, Adam, S P. Sintiger (Joseph, S P. Pieringer, Anton, S P. Markov, Keth, S P. Holmer, Keth, S P. Holmer, Keth, S P. Holmer, Keth, S P. Franker, McNael, S P. Scheler, McNael, S | 10 evalang@bia. | - Tu - Tu - Tu - Tu - Tu - Tu - Tu - Tu |
| 19581 ABP_S 19581 ABP_S 19570 ABP_P 19576 ABP_P 19577 ABP_C 19577 ABP_C 19575 ABP_C 19573 ABP_C 19573 ABP_C 19573 ABP_C 19570 G1set, 17769 ABP_C 17669 ABP_L | P. Schmidteder Joseph, S P. Schleppinger, Adam, S P. Pininger, Anton, S P. Paneh, U. Jacob, S P. Made, Michael, S P. Helmer, Karl, S P. Groub, Upnaz, S P. Jenher, Michael, S P. Jocher, Michael, S | 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. | - Tu - Tu - Tu - Tu - Tu - Tu - Tu |
| 19580 ABP,5 19579 ABP,9 19578 ABP,9 19576 ABP,7 19576 ABP,7 19576 ABP,7 19577 ABP,6 19577 ABP,7 19573 ABP,7 19572 ABP,8 19170 GTset, 17769 ABP,0 17669 ABP,1 | Ø, Schlappinger, Adam, S B, Pieringer, Anton, S Ø, Pench Ljakob, S B, Mader, Micchael, S B, Grouet J, ganz, S B, Grouet J, ganz, S B, Frankenberger, Simon, S B, Flocher, Michael, S | evalang@bist. evalang@bist. evalang@bist. evalang@bist. evalang@bist. evalang@bist. evalang@bist. evalang@bist. evalang@bist. | - Tu - Tu - Tu - Tu - Tu - Tu |
| 19579 ABP_P 19576 ABP_R 19577 ABP_N 19576 ABP_F 19577 ABP_F 19573 ABP_F 19574 ABP_R 19575 ABP_R 19576 ABP_R 19577 ABP_R 19573 ABP_R 19170 GTset, 17769 ABP_L 17680 ABP_L 17688 ABP_L | P. Pieringer, Anton, S P. Peschi, Lakob, S P. Mador, Michoel, S P. Holmer, Karl, S P. Gruek, Uppaz, S P. Frankerberger, Simon, S P. Fischer, Michael, S | 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. | - Tu - Tu - Tu - Tu - Tu |
| 19578 ABP_P 19577 ABP_N 19575 ABP_C 19575 ABP_C 19575 ABP_F 19573 ABP_F 19573 ABP_B 19070 GTset, 1769 ABP_C 1769 ABP_L 17688 ABP_L | Ø.Peschi Jakob, S IP. Mader, Michael, S Ø. Holmer, Karl, S IP. Gruebil Jgnaz, S Ø. Frankærbærger, Simon, S Ø.Flischer, Michael, S | 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. | - Tu - Tu - Tu |
| 19577 ABP_M 19576 ABP_F 19577 ABP_G 19577 ABP_G 19577 ABP_G 19577 ABP_G 19577 ABP_G 19573 ABP_G 19572 ABP_G 19370 GTset, 17669 ABP_L 17668 ABP_L | P_Mader_Michael_S P_Holmer_Kail_S P_GruebUgnaz_S P_Frankenberger_Simon_S P_Flocher_Michael_S | 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. | - Tu - Tu |
| 19576 ABP_F 19575 ABP_C 19574 ABP_F 19573 ABP_F 19572 ABP_B 19370 GTset, 17769 ABP_C 17689 ABP_L | Ø_Holmer_Karl_S P_Grueb(Jgnaz_S Ø_Frankanberger_Simon_S P_Fischer_Michael_S | 30 evalang@bist. 30 evalang@bist. 30 evalang@bist. | - Tu |
| 19575 ABP_C 19574 ABP_F 19573 ABP_F 19572 ABP_B 19370 GTset, 17769 ABP_C 17688 ABP_L | P_GrueblJgnaz_S P_Frankanberger_Simon_S P_Fischer_Michael_S | 30 evalang@bist. 30 evalang@bist. | Tot |
| 19574 AB9_F 19573 AB9_F 19572 AB9_B 19370 GTset, 17769 AB9_C 17689 AB9_L 17688 AB9_L | P_Frankerberger_Simon_S P_Fischer_Michael_S | 30 evalang@bist. | - 19 |
| 19573 ABP_F 19572 ABP_B 19370 GTset, 17769 ABP_G 17689 ABP_C 17688 ABP_L | P_Fischer_Michael_S | 30 mm famou@hink | - Tu |
| 19572 ABP_8 19370 GTset, 17769 ABP_0 17689 ABP_0 17688 ABP_L | | ov evalanguoist. | - TU |
| 19370 GTset, 17769 ABP_G 17689 ABP_C 17688 ABP_L | IP_Bogner_Max_S | 30 eva.lang@bist. | - Tu |
| 17689 ABP_C 17688 ABP_L | set_AIP_Overview_50_pages | 50 eva.lang@bist. | - Th |
| 17688 ABP_L | P_Guinas_Joseph_S | 30 evalang@bist. | - Fn |
| TTOOD ADP_C | P_Dick_oseph_s | 30 evalanggeldt. | - 10 |
| 17697 ADD C | P_Cether_Worgang_s | 30 evaluargeoist. | - 10 |
| 17/86 APP C | P Schmidhuber Johann-Georg S | 30 evaluation hist | - Tu |
| 17685 ABP K | P Klaempfl Joseph S | 30 evalang@his | Tu |
| 17684 ABP R | P Rottmayr Joseph S | 30 evalang@hid. | - Tu |
| 17451 ABP | P Jungbauer, Ferdinand_S | 30 eva.lang@bist. | - Wi |
| 17447 ABP N | P_Muenich_Simon_S | 30 eva.lang@bist. | - we |
| 17446 ABP_R | P_Ritzinger_Johann_Baptist_S | 30 eva.lang@bist. | - We |
| 17411 ABP_V | P_Visino_Johann_Nepomuk_S | 30 eva.lang@bist. | - Tu |
| 14828 ABP_N | P_Malgersdorf_vor_1600 | 10 wolfgang.fron. | WA |
| 13521 ABP_F | P_Hauer_Mathias | 18 wolfgang.fron. | Wi |
| 13520 ABP_A | P_Achatz_Josef | 18 wolfgang.fron. | we |
| 7048 GTset | set_ABP_selection_200_I | 85 eva.lang@bist. | - Th |
| 7047 GTset | set_ABP_selection_200_1 | 115 eva.lang@bist. | - Th |
| SW2 Hands | indbuch_Bistum_Passau_1828 | 298 evalang@bist. | - M |
| Serr ABP_0 | er_overview_random_600 | eou pnilip | 0 |
| | | | |
| | | | |
| < | | | |

Figure 1: Table editor of the Transkribus software showing an annotated image of the ABP GT dataset.

3 Evaluation

The evaluation of the table matching is based on Shahab et al. [1] and Burie et al. [2]. Shahab et al. encode the table information directly in the image format which is similar to document image segmentation (each pixel belongs to a certain cell/row/column/table) [1]. Note that the encoding of the table as an image can be generated by using the table description in the PAGE xml. Based on these description established methods for evaluating image segmentation methods can be applied. Shahab et al. define the following measures: Correct Detections, Partial Detections, Over-Segmentations, Under-Segmentations, Missed Segments, and False Positive Detections (see [1] for a detailed description). Additionally, the Jaccard Index (JI) to measure the overlapping of a detected document region (quadrilateral) with the annotated document region in the image is used. The JI is introduced by the ICDAR2015 Competition on Smartphone Document Capture and OCR (SmartDoc) [2]. Due to the proposed methodology (table structure matching based on a *template*, see 4) the following measures of Shahab et al. and Burie et al. are used.

- Mean Cell Match (MCM): $\frac{1}{N} \sum_{i=1:N} \frac{|G_i \cap S_i|}{|G_i|}$ where G_i corresponds to the area of each cell of the GT segmentation and S_i is the corresponding cell area detected by the proposed methodology (one-to-one correspondence). N is the number of cells of the table. This corresponds to the value *Correct Detections* (the value is thresholded in [1]).
- Mean Table Match (MTM): same as MCM but only for the entire table region.
- Under-Segmentation (USeg) defines the number of cells that have a major overlap with more than one GT segment: overlap of the corresponding cell S_i with all

 $G_{j\neq i} > T.$

- Missed Segments (Miss) defines the number of cells that do not have a major overlap with the corresponding detected segment (number of segments with MCM < T.
- Jaccard Index $JI = \frac{1}{N} \sum_{i=1:N} \frac{area(G_i \cap S_i)}{area(G_i \cup S_i)}$. The JI has a range from 0 to 1, where 1 is the best segmentation possible. For the table matching, the JI is calculated for the detected table region as well as for all table cells (mean value for all cells for one table is calculated).

4 Table Matching Methodology and Results

In the following Sections the current implementation of the form/table analysis is shortly summarized. Furthermore, the planned future work for D6.9 is presented. Additionally, the first evaluation on the presented dataset in Section 2 is shown.

4.1 Matching Table Structure Using Association Graphs

In D6.7 a method based on Beveridge and Riseman [3] has been developed to align tables based on line models to a corresponding template. Experiments have shown that especially for hand-drawn tables (e.g. documents of ABP, see Section 2) variations of the column width and the rows height (header) are present, which cannot be solved with the method of D6.7. Thus, a new table template matching has been developed. It matches the hierarchical structure of the table using an association graph (see Pelillo et al. [4] and Ishitano [5]) by finding a maximum clique [6]. Consider the matching of the table template T to a given document D. The template defines all table cells, and thus all visible (horizontal and vertical) cell borders TL_i where *i* is the number of cell borders. A line detection in D gives all (horizontal and vertical) lines DL_j where *j* is the number of detected lines in D. Each node N_i in the association graph *G* corresponds to a pair of horizontal/vertical lines $N_k = (TL_i, DL_j)$. To create edges between nodes, the compatibility of every combination of two pairs of nodes is examined. Two nodes $N_1 = (TL_1, DL_1)$ and $N_2 = (TL_2, DL_2)$ are compatible if they fulfill the following criteria:

- if TL_1 is left/above from TL_2 then also DL_1 must be left/above from DL_2
- if m=dist(TL_1, TL_2) and n=dist(DL_1, DL_2), then $m \times (1-T) \le n < m \times (1+T)$ for a certain threshold T.

The largest maximal clique on G defines the best matching of document D to a given template T. The line detection in D6.7 and [7] is based on Zheng et al. [8] and also described in Diem et al. [9] is also used here.

Figure 2 shows a table image of a document from the Passau Diocesan archive. The second image shows the rough alignment of the table template with the current document. It can be seen that there are variations of the width of the columns. The rough



Figure 2: Example of the table matching based on a line model. The first image shows the document image, the second image shows the rough alignment of the table template to the document (red lines), and the last image shows the resulting maximal clique (blue horizontal lines, green vertical lines.



Figure 3: Another example of the table matching based on a line model and an association graph.

alignment is done by a correlation of the template image and the document image. Afterwards, the association graph is calculated and the maximum clique of the graph is visualized in the last image. Figure 3 shows a second example.

Based on the columns and the table region, the table rows are detected based on the detected baselines of the Basic Layout Analysis (Task 6.2). The detection of the rows is done in Task 6.5 (Document Understanding, D6.14).

4.2 Results

The evaluation of the table matching is done on the dataset described in Section 2. The metric is explained in Section 3 and the results are shown in Table 1.

| | ABP_GT |
|------------|---------|
| | dataset |
| MTM | 0.9785 |
| JI (Table) | 0.9305 |
| MCM | 0.8936 |
| JI (Cell) | 0.8754 |
| USeg | 0.0796 |
| Miss | 0.0566 |

Table 1: Evaluation of the proposed table matching.

5 Future Work

To avoid a binarization of the image and the resulting problems, a line detector using the gradient information of gray value images will be used. Thus, a state of the art line detection based on von Gioi [10] will be implemented in the READ framework for D6.9. Additionally, the current methodology will be enhanced to avoid errors for ambigous cell borders (can occur if two lines are detected at the table borders). A weighted maximum clique method will be tested to achieve better results. Also an automated template selection will be tested in D6.9.

Currently, also a command line interface for the table module exists, which will be the basis for the integration task into Transkribus. The integration to Transkribus will be realized in the next deliverable.

References

 A. Shahab, F. Shafait, T. Kieninger, and A. Dengel, "An open approach towards the benchmarking of table structure recognition systems," in *Proceedings* of the 9th IAPR International Workshop on Document Analysis Systems, ser. DAS '10. New York, NY, USA: ACM, 2010, pp. 113–120. [Online]. Available: http://doi.acm.org/10.1145/1815330.1815345

- [2] J. Burie, J. Chazalon, M. Coustaty, S. Eskenazi, M. M. Luqman, M. Mehri, N. Nayef, J. Ogier, S. Prum, and M. Rusinol, "Icdar2015 competition on smartphone document capture and ocr (smartdoc)," in 13th International Conference on Document Analysis and Recognition (ICDAR), Aug 2015, pp. 1161–1165.
- [3] J. R. Beveridge and E. M. Riseman, "How easy is matching 2d line models using local search?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 6, pp. 564–579, Jun 1997.
- [4] M. Pelillo, K. Siddiqi, and S. W. Zucker, "Matching hierarchical structures using association graphs," *IEEE Transactions on Pattern Analysis and Machine Intelli*gence, vol. 21, no. 11, pp. 1105–1120, Nov 1999.
- [5] Y. Ishitani, "Model matching based on association graph for form image understanding," in *Proceedings of 3rd International Conference on Document Analysis* and Recognition, vol. 1, Aug 1995, pp. 287–292 vol.1.
- [6] J. Konc and D. Janezic, "An improved branch and bound algorithm for the maximum clique problem," MATCH Commun. Math. Comput. Chem., vol. 58, pp. 569– 590, 2007.
- [7] F. Kleber, M. Diem, and R. Sablatnig, "Form Classification and Retrieval using Bag of Words with Shape Features of Line Structures," in *Document Recognition* and Retrieval XXI, 2014.
- [8] Y. Zheng, C. Liu, X. Ding, and S. Pan, "Form frame line detection with directional single-connected chain," in *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, 2001, pp. 699–703.
- [9] M. Diem, F. Kleber, and R. Sablatnig, "Document Analysis Applied to Fragments: Feature Set for the Reconstruction of Torn Documents," in *Proceedings of the International Workshop on Document Analysis Systems (DAS)*, D. Doermann, V. Govindaraju, D. Lopresti, and P. Natarajan, Eds., Boston, USA, June 2010, pp. 393–400.
- [10] R. G. von Gioi, J. Jakubowicz, J. M. Morel, and G. Randall, "Lsd: A fast line segment detector with a false detection control," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 722–732, April 2010.