

D5.9 ScriptNet Large Scale Dataset P2

Florian Kleber, Markus Diem, 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 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 | 11.12.2017 | |
| Deliverable number | D5.9 | |
| Deliverable title | ScriptNet Large Scale Dataset P2 | |
| Туре | report | |
| Status & version | in progress | |
| Contributing WP(s) | WP5 | |
| Responsible beneficiary | CVL | |
| Other contributors | UPVLC, DUTH, NCSR, ABP | |
| Internal reviewers | UPVLC,NCSR | |
| Author(s) | Florian Kleber, Markus Diem, Stefan Fiel | |
| EC project officer | Martin MAJEK | |
| Keywords | baseline, KWS, writer identification, HTR, table, Dataset | |

Contents

| 1 | Executive Summary | 4 | |
|---|---|-------------------------|--|
| 2 | Competition Datasets 2.1 ScriptNet: Dataset for Document Image Binarisation. ICDAR 2017 | | |
| 3 | Datasets 3.1 ABP_S_1847-1878 | 5 5 6 6 | |

1 Executive Summary

This task comprises the selection of the document images, the definition of the Ground Truth (GT) for the corresponding task, the management of the data production, the distribution of data to training and evaluation sets and the description of the datasets. Based on the planned datasets in D5.8 five competitions have been carried out at ICDAR 2017. All datasets have been made publicly available together with their GT (downloadable at the competition website (see ScriptNet) or from Zenodo.org, e.g. cBad). The competitions are summarized in Section 2. Section 3 describes new datasets and also further planned datasets for D5.10. For the creation of the datasets the presented benchmarking tool¹ in D5.8 is partly used.

2 Competition Datasets

In D5.8 datasets have been prepared for the planned competitions at ICDAR 2017. Finally, the following competitions have been carried out at ICDAR:

- ICDAR2017 Competition on Baseline Detection (cBAD) [1]
- ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI) [2]
- ICDAR2017 Competition on Handwritten Text Recognition on the READ Dataset (ICDAR2017 HTR)
- ICDAR2017 Competition on Document Image Binarization (DIBCO 2017) [3]

The planned datasets for the four competitions (1-4) have been described in D5.8 and are now publicly available together with the annotated ground-truth. Additionally a dataset for keyword spotting has been prepared². The datasets can be downloaded either from the scriptnet³ site or from Zenodo⁴. The results of all competitions have been published at ICDAR 2017 [1, 2, 4, 3].

2.1 ScriptNet: Dataset for Document Image Binarisation. ICDAR 2017

The DIBCO 2017 testing dataset consists of 10 machine-printed and 10 handwritten document images for which the associated ground truth was built manually for the evaluation. The selection of the images in the dataset was made so that representative degradations appear. The machine-printed documents of the dataset originate from collections that belong to the IMPACT project, while the handwritten document images originate from collections that belong to READ project. The testing dataset along with

¹https://github.com/TUWien/Benchmarking

 $^{^{2}} https://scriptnet.iit.demokritos.gr/competitions/7/$

³https://scriptnet.iit.demokritos.gr/competitions/

 $^{^4\}mathrm{e.g.}$ cBad, <code>https://zenodo.org/record/835441#.WhgH80cxlhE</code>

the associated ground truth as well as the evaluation software are publicly available at: http://vc.ee.duth.gr/dibco2017/benchmark.

3 Datasets

For the Large Scale Demonstrator (LSD) the data of the Passau Diocesan Archive is the basis. The dataset is named $ABP_S_1847-1878$ and is described in detail in the following section.

3.1 ABP_S_1847-1878

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. The scans originate from 212 pastoral districts (mainly parishes) with their own record keeping in the time between the uproar of 1848 and the beginning of the German Empire in 1871. This period is marked by massive social, economic and technical transformations.

| $ \begin{array}{ c } \hline logal block Table Lake Lake Lake Lake Lake Lake Lake Lak$ | Ingel bler sakesmant Ingel bler sakes | I logat Make Quant members I logat I logat <thi logat<="" th=""> I logat <thi log<="" th=""><th>ed. Angla. Degrant Stands Anglant Stal Allan Alexington. Sugaration the start Sungipulant Standsgring Start Stard Top Alemant Alemant</th></thi></thi> | ed. Angla. Degrant Stands Anglant Stal Allan Alexington. Sugaration the start Sungipulant Standsgring Start Stard Top Alemant Alemant |
|---|--|--|--|
| $ \begin{array}{ $ | Image: | Dourner. I dot Wroics Bits tarking Bits tarkin | el Agel. "Agene Stander Agene Steller Allen Sterreigen Stegenstern ner: """""""""""""""""""""""""""""""""""" |
| Writes Burget Burget< | Venon Plant Plant <th< th=""><th>Wenson Due extindy Recrit decounset <thdecounset< th=""> <thdecounset< th=""></thdecounset<></thdecounset<></th><th>une Frenificient Some Siging life tout 24 Mont Mont</th></th<> | Wenson Due extindy Recrit decounset Decounset <thdecounset< th=""> <thdecounset< th=""></thdecounset<></thdecounset<> | une Frenificient Some Siging life tout 24 Mont Mont |
| $\frac{1}{100} + \frac{1}{100} + \frac{1}{100} + \frac{1}{100} + \frac{1}{100} + \frac{1}{1000} + \frac{1}{10000} + \frac{1}{10000000000000000000000000000000000$ | intervent inter | Rection: Control Contro Control <thcontrol< th=""> <th< th=""><th>And a second sec</th></th<></thcontrol<> | And a second sec |
| $ \begin{array}{ c c c c c c c c c c c c c c c c c c c$ | $\frac{1}{1} \frac{1}{1} \frac{1}$ | Berker Gala ABP (572, Edso) Gala ABP (572, Edso) </td <td>The second and the second second second</td> | The second and the second second second |
| $ \begin{array}{c c c c c c c c c c c c c c c c c c c $ | UNITY IN INTERCE INTER | Offer ABP 6722; Ede0 Other Adv Adv 119/19 # 119 # 100 G G G <td< td=""><td>the for I I</td></td<> | the for I I |
| 19/10 1 <td>Interformer de la construit de la cons</td> <td>119/19 1 119/19 1 110/19 1 110/19 1 110/19 1 110/19 1 110/19 1 110/19 1 110/19 1 110/19</td> <td>And Man Vat Mrs R. A. Mosmiller</td> | Interformer de la construit de la cons | 119/19 1 119/19 1 110/19 1 110/19 1 110/19 1 110/19 1 110/19 1 110/19 1 110/19 1 110/19 | And Man Vat Mrs R. A. Mosmiller |
| $ \begin{array}{c c c c c c c c c c c c c c c c c c c $ | 0 Trie Person Update | D Title Page Upbader Up 2006 Ref Mannich Simon 5, deplotted 30 exclangeBill. 10 | 2 5/2/all Ha Add Marmiller H.S. No. A. |
| Open Description Open Description <t< td=""><td>Second Second Second<</td><td>1998 All Manich Stems S. Applicated 30 exchangebit. 50 2019 All Schnick Joseph S. Applicated 30 exchangebit. 50 2019 All Schnick Joseph S. Applicated 30 exchangebit. 50 2014 All Schnick Joseph S. Applicated 30 exchangebit. 70 2014 All Schnick Joseph S. Applicated 30 exchangebit. 70 2014 All Schnick Joseph S. Applicated 30 exchangebit. 71 2014 All J. Chenk Joseph S. Applicated 30 exchangebit. 71 2014 All J. Chenk Joseph S. Applicated 30 exchangebit. 71 2013 All J. Chenk Joseph S. Applicated 30 exchangebit. 71 2014 All J. Chenk J. Schnickted 30 exchangebit. 71 2013 All J. Chenk J. Schnickted 30 exchangebit. 71 2014 All J. Chenk J. Schnickted 30 exchangebit. 71 2013 All J. Chenk J. Schnickted 30 exchangebit.</td><td>of device they device thing as is handlass constant</td></t<> | Second Second< | 1998 All Manich Stems S. Applicated 30 exchangebit. 50 2019 All Schnick Joseph S. Applicated 30 exchangebit. 50 2019 All Schnick Joseph S. Applicated 30 exchangebit. 50 2014 All Schnick Joseph S. Applicated 30 exchangebit. 70 2014 All Schnick Joseph S. Applicated 30 exchangebit. 70 2014 All Schnick Joseph S. Applicated 30 exchangebit. 71 2014 All J. Chenk Joseph S. Applicated 30 exchangebit. 71 2014 All J. Chenk Joseph S. Applicated 30 exchangebit. 71 2013 All J. Chenk Joseph S. Applicated 30 exchangebit. 71 2014 All J. Chenk J. Schnickted 30 exchangebit. 71 2013 All J. Chenk J. Schnickted 30 exchangebit. 71 2014 All J. Chenk J. Schnickted 30 exchangebit. 71 2013 All J. Chenk J. Schnickted 30 exchangebit. | of device they device thing as is handlass constant |
| 2009 AP Schnickland 0 exklogBith To 2019 AP Schnickland 0 exklogBith To 2020 AP Schnickland 0 exklogBith To 2021 AP Schnickland 0 exklogBith To 2020 AP Schnickland 0 exklogBith To 2021 AP Schnickland 0 exklogBith To AP Schnickland To 2021 AP Schnickland 0 exklogBith To exklogBith To AP Schnickland To <t< td=""><td>2009 APB Antony Looph, Septextel 20 examples. The analysis application of the analy</td><td> 2009 All Pathony, heaps, Sapericad 2010 All Pathony, heaps, Sapericad 2011 All Pathony, heaps, Sapericad 2011 All Pathony, Sapericad 2012 All Pathony, heaps, Sapericad 2013 All Pathony, heaps, Sapericad 2013 All Pathony, Sapericad 2014 All Pathony, heaps, Sapericad 2015 All Pathony, Sapericad 2016 All Pathony, Sapericad 2016 All Pathony, Sapericad 2016 All Pathony, Sapericad 2016 All Pathony, Sapericad 2017 All Pathony, Sapericad 2018 All Pathony, Sapericad 2018 Ofter, All Pathony, Sapericad 2019 Ofter, All Pathony, Saperic</td><td>a modelfe light and a granding</td></t<> | 2009 APB Antony Looph, Septextel 20 examples. The analysis application of the analy | 2009 All Pathony, heaps, Sapericad 2010 All Pathony, heaps, Sapericad 2011 All Pathony, heaps, Sapericad 2011 All Pathony, Sapericad 2012 All Pathony, heaps, Sapericad 2013 All Pathony, heaps, Sapericad 2013 All Pathony, Sapericad 2014 All Pathony, heaps, Sapericad 2015 All Pathony, Sapericad 2016 All Pathony, Sapericad 2016 All Pathony, Sapericad 2016 All Pathony, Sapericad 2016 All Pathony, Sapericad 2017 All Pathony, Sapericad 2018 All Pathony, Sapericad 2018 Ofter, All Pathony, Sapericad 2019 Ofter, All Pathony, Saperic | a modelfe light and a granding |
| 2333 AB_{2} Schnicker (spec), Specification B_{2} AB_{2} | 2333 AB 2-Annotation (seep), 3-plotted 30 analogben. Th Th 2334 AB 2-Annotation (seep), 3-plotted 30 analogben. Th Th Th 2334 AB 2-Annotation (seep), 3-plotted 30 analogben. Th | 2031 All 25-bindeset joorph 5, depicted 2043 All 25-bindeset joorph 5, depicted 2054 All 25-bindese joorph 5, depicted 2056 All 25-bindese joorph 5, depicted 2066 All 25-bindese joorph 5, depicted 2076 All 26-bindese joorph 5, depicted 2086 All 26-bindese joorph 5, depicted 2087 All 26-bindese joorph 5, depicted 2088 All 26-bindese joorph 5, depicted 2088 All 26-bindese joorph 5, depicted 2089 All 26-bindese joorph 5, depicted 2080 All 26-bindese | Sould They South May 20 |
| $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 4 | 2.32 Abs 2.33 Abs 2.34 Abs Abs< | a File fort II II - |
| $ \begin{array}{cccccccccccccccccccccccccccccccccccc$ | 2009 By James Wolfrag S, Japlanter 90 consequence 10 consequenc | 2010 AP Johns Walang Schwards 0 2010 AP Johns Walang Schwards 1 2010 AP Johns Walang Schwards 1 2011 AP Johns Walang Schwards 1 2011 AP Johns Walang Schwards 1 2011 AP Johns Walang Schwards 1 2012 AP Johns Walang Schwards 1 2013 AP Johns Walang Schwards 1 2014 AP Johns Walang Schwards 1 2015 AP Johns Walang Schwards 1 2015 AP Johns Walang Schwards 1 2016 AP Johns Walang Schwards 1 2016 AP Johns Walang Schwards 1 2017 AP Johns Walang Schwards 1 2018 Gerta AP Johns Walang Schwards 1 2019 Gerta AP Johns Walang Schwards 1 2019 Gerta AP Johns Walang Schwards 1 2010 Gerta AP Johns Walang Schwards 1 2010 Gerta AP Johns Walang Schwards 1 2010 Gerta AP Johns Walang Schwards 1 2014 Gerta AP Johns Walang Schwards 1 2014 Gerta AP Johns Walang Schward 1 2015 Gerta AP Johns Walang Schward 1 2014 Gerta AP Johns Walang Schward 1 2015 Gerta AP Johns Walang Schward 1 2014 Gerta AP Johns Walang Sch | Heldisch 19 |
| $ \begin{array}{c c c c c c c c c c c c c c c c c c c $ | 2 2 3 3 3 4 3 4 3 4 | 2029 AP Frankenger Gronz, Sapital. 2029 AP Frankenger Gronz, Sapital. 2020 AP Frankenger Gronz, Sapital. 2020 AP Frankenger Gronz, Sapital. 2020 AP Frankenger Gronz, Sapital. 2021 AP Dol. Joseph S. Ageitated 2020 aP Frankenger Gronz, Sapital. 2021 AP Magnetic Version (Sapital. 2022 AP Magnetic Version (Sapital. 2023 AP Frankenger Gronz, Sapital. 2024 AP Magnetic Version (Sapital. 2024 AP Magnetic Version (Sapital. 2025 AP Stranger Joseph S. Ageitated 2026 The AP Stranger Joseph S. Ageitated 2026 AP Stranger Johan April. 2026 AP Stranger Johan April. 2026 AP Stranger Johan April. 2027 AP Stranger Johan April. 2028 AP Stranger Johan April. 2029 AP Stranger Johan April. 2029 AP Stranger Johan April. 2020 AP Stranger Johan | Sull May der & May |
| $ \begin{array}{c c c c c c c c c c c c c c c c c c c $ | 2.232 AB junct week and you have a standard week and you have a standar | Jang Picker, Michael S, depicted 20 enalog@eta. In en | mill Spal |
| $ \begin{array}{c} 2277 AB} \ black \ b$ | Stor As productions 5, definitions of the storage of the storag | All Dick, Josep, S., dopierted 00 contang@bit. Th All Dick, Josep, S., dopierted 10 contang@bit. W Contang. Dick, State, J. State, | Jan 7 Mly & g. May |
| $ \begin{array}{cccccccccccccccccccccccccccccccccccc$ | Add Add <td> Alley Janz Yanz, Anne, Janzie Janz, Kanz, Janzie Janz, Kanz, Janzie Janz, Kanz, Janz, Jan</td> <td>and tillefor to a second</td> | Alley Janz Yanz, Anne, Janzie Janz, Kanz, Janzie Janz, Kanz, Janzie Janz, Kanz, Janz, Jan | and tillefor to a second |
| $ \begin{array}{cccccccccccccccccccccccccccccccccccc$ | 207 all phane, Mathine, deploted 0 excluded 1 exclu | 2207 ABP Product Mathing Anglement of a constraing@htt. W 2207 ABP Product Mathing Anglement of a constraing@htt. W 2208 Gitta ABP Societion, 201, deplicated is constraing@htt. W 2209 Gitta ABP Societion, 201, deplicated is constraing@htt. W 2200 Gitta ABP Societion, 201, deplicated is constraing@htt. W 2200 Gitta ABP Societion, 201, deplicated is constraing@htt. W 2200 Gitta ABP Societion, 201, deplicated is constraing@htt. W 2200 Gitta ABP Societion, 201, deplicated is constraing@htt. W 2200 Gitta ABP Societion, 201, deplicated is constraing@htt. W 2200 Gitta ABP Societion, 201, deplicated is constraing@htt. W 2200 Gitta ABP Societion, 201, deplicated is constraing@htt. W 2200 Gitta ABP Societion, 201, deplicated is constraing@htt. W 2200 Gitta ABP Societion, 201, deplicated is constraing@htt. W 2200 Gitta ABP Societion, 201, deplicated is constrained in the society of the s | · Very Max Soll Mars |
| $ \begin{array}{c c c c c c c c c c c c c c c c c c c $ | 242 247 AP Andrewing dagstered 10 excluding data Weight Weigh | Zerig All, Ander, Sond Augustand 13 exalting Better. W. 2014 [Englished] Zerig Ding, Aller, Sond Augustand 15 exalting Better. W. 2014 [Englished] Zerig Ding, Aller, Aller, Sond Augustand 15 exalting Better. W. 2014 [Englished] Zerig Ding, Aller, Aller, Sond Augustand, Sond 10 exalting Better. W. 2014 [Englished] Zerig Ding, Aller, Aller, Sond Augustand, Sond 10 exalting Better. W. 2014 [Englished] Zerig Ding, Aller, Aller, Sond Augustand, Sond 10 exalting Better. W. 2014 [Englished] Zerig Ding, Aller, Aller, Sond Augustand, Sond 10 exalting Better. W. 2014 [Englished] Zerig Ding, Aller, Aller, Sond Augustand, Sond 10 exalting Better. W. 2014 [Englished] Zerig Ding, Aller, Aller, Sond Augustand, Sond 10 exalting Better. W. 2014 [Englished] Zerig Ding, Aller, Aller, Sond Augustand, Sond 10 exalting Better. W. 2014 [Englished] Zerig Ding, Aller, Aller, Sond Augustand, Sond 10 exalting Better. W. 2014 [Englished] Zerig Ding, Aller, Aller, Sond Augustand, Sond 10 exalting Better. W. 2014 [Englished] Zerig Ding, Aller, Aller, Sond Augustand, Sond 10 exalt (Schult Schult Schu | 2 stylend 22 - Jusamilience 200 |
| $ \begin{array}{c} 2222 & Grave Apple decises, 202, 4 particular, 202 \\ Grave Apple decises, 202 \\ Gr$ | 2422 Cinc All Spectra (20) Application Signal A | 2242 Gitz ABP, election, 20.3, deplicited 55 exalting@stin. W 22467 Gitz ABP, election, 20.3, deplicited 155 exalting@stin. W 22469 Gitz ABP, Overview, D., garge, daylic. 50 exalting@stin. W 22469 ABP, Vinion, Nagenst, 5, depl. 30 exalting@stin. W 2247 ABP, Bringer, Johnen, Sayni, 5, depl. 30 exalting@stin. W 2247 ABP, Bringer, Johnen, Sayni, 5, depl. 30 exalting@stin. W 2247 ABP, Bringer, Johnen, Sayni, 5, depl. 30 exalting@stin. W 2247 ABP, Bringer, Johnen, Sayni, 5, depl. 30 exalting@stin. W 2247 ABP, Bringer, Johnen, Sayni, 5, depl. 30 exalting@stin. W 2248 ABV, Vinio, Sayni, 5, depl. 30 exalting@stin. W 2249 ABV, Charles, Sayni, 5, depl. 30 exalting@stin. W 2249 ABV, Charles, Sayni, 5, depl. 40 exalting 2240 ABV, Sayni, 5, depl. 40 exalting 2240 exalting 2240 ABV, Sayni, 5, depl. 40 exalting 2240 exalting | M. C. A. 2. A |
| 2401 Glauphander Ling verscher 2004 gehalten und gehalten wir der Steinen Beine B | 2401 Glughersteiner, Margeheiten und gesteiner, Margeheiten und gesteine | 2240 Ghey Berecken, Will Agalantiettel 115 er endangbette. Wi 2048 All Vinne Jehnen, Negrond, Sopel. 30 er endangbette. Wi 2048 All Vinne Jehnen, Negrond, Sopel. 30 er endangbette. Wi 2048 All Vinne Jehnen, Markin, Sopel. 30 er endangbette. Wi 2049 Tabelterepitter | The add the and the Main se _ R. A. Moremuiller |
| 240 Old B. Verview, S. garge, daffer, S. enanglight, W. exalting Bath, S. during, Statistical Mark, S. during, Statistical Mark, S. during, S. | Cell Control State and State and | 2240 Ohr, M.B., Verview, B., genge, daylie. 30 contang@bit. W 2447 ABP, Beinger, Johann, Magnith, S., daylie. 30 contang@bitt. W 2447 ABP, Beinger, Johann, Magnith, S., daylie. 30 contang@bitt. W 2447 ABP, Beinger, Johann, Magnith, S., daylie. 30 contang@bitt. W 2447 ABP, Beinger, Johann, Magnith, S., daylie. 30 contang@bitt. W 245 dialamenter: 4fr. Joinger, Markall, S., daylie, J., dayl | me state |
| Land All Application of the second | c constraints of sound sound constraints of sound sound constraints of sound sound sound sound constraints of sound sound sound constraints of sound sound | 2247 AD Discover day of the second se | In g. May S.H. May |
| $\begin{array}{c c c c c c c c c c c c c c c c c c c $ | 2204 TableTempiles II exalleg@bits_ II 4 </td <td>2004 TableTempiter 11 enalog@bit. Th st. deliterature fift-grigmentation fift-grigmentation fift-grigmentation fifthered for the first statement fifthered for the first statement fifthered for the first statement statement for the first statement st</td> <td>T Ista Thype It =</td> | 2004 TableTempiter 11 enalog@bit. Th st. deliterature fift-grigmentation fift-grigmentation fift-grigmentation fifthered for the first statement fifthered for the first statement fifthered for the first statement statement for the first statement st | T Ista Thype It = |
| 35. dang temi "Handay day day day day day day day day day | c | 33. Sing James' Hand you Bay Japan's Star Under Manage Start Stranger Stran | South Mar Low to Min |
| 34. And and the second | 1 Felixberger 2 Maximilian | soft Ander Single for Strange | a mile life to life a desperanted deligrantic and |
| 24. Under Hanne fange fanne i Steller i steller in steller i stell | c >> | 34 Ander Marine Hangebrief and State States and States | R.S. M. Andelitter |
| 1 Martin Andrea Martin Mar | 1 Felixberger 2 Maximilian | and standing in the second sec | the the sail to they is a south they |
| thereastic left it and the second state of the | Second de la construcción de | Webssonriel list R. and man and | The Man Heren |
| | < A skimilian | | · Sew the May South May |
| | < | | |
| | < | 2 Maximilian | |
| 2 Maximilian | <u>(</u> | | |
| 2 Maximilian | | | |
| Z Maximilian | < <u>></u> | | |
| 2 Maximilan | | < > | |
| Z Maximilian | | | |

Figure 1: Example image of the GTset_ABP with annotatedt GT (tables, baselines, transcription).

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. Stemming from than 590 individual hands, the data set, recoding an estimated number of 295,000 casualties, is also highly diverse with regards to writers. 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. Most of these printed layout categories comply with the given normative for content imposed by the Church. The vast majority of scans (15,147 images) even fall into one single template category. On 4,578 pages, the requested information was recorded in manually drawn tables or manually extended table prints. The images are openly available through the matricula online platform ⁵, records can be queried using a search engine supplied by the Diocese of Passau ⁶. The data are used by family historians as well as by historian scholars interested in age of those who died, the development or the spread of deadly diseases, etc.

The dataset is available in Transkribus (Collection ID 6285). Based on the ABP_S_1847 -1878 dataset a subset with annotated tables, baselines and tables is provided. The subset contains about 370 pages and is on Transkribus in Collection $GTset_ABP$ (Collection ID 6722).

Figure 1 shows an annotated image of the $GTset_ABP$ (Collection ID 6722) dataset. The table and the baselines are annotated, as well as the transcribed information.

3.2 Girona Dataset

The Girona Collection, has been provided by the Centre de Recerca d'Historia Rural (CRHR) from the Universitat de Girona (MoU partner of the READ project). The collection, called "Oficio de Hipotecas de Girona", is composed of a large number of notarial documents from the 17th century. It contains much more than morgage deeds, including such notarial acts as sales and purchase agreementes, leasing and Emphyteusis contracts, cretid operations, marriage contracts, manufacturing and commercial compaines, etc. The CRHR is interested in the perfect transcription of this collection.

So far, around 400 pages has been processed and the ground truth at two levels has been generated. First, a layout analysis of each page was manually done to indicate text blocks and lines, resulting in a dataset of around 16000 lines. Second, the pages were completely transcribed line by line by an expert paleographer using the CATTI technology (see Deliverable D7.5). Additionally to the diplomatic transcription of the document, some tags are added to the corresponding words in order to provide a more rich information about the content of the document: toponyms, antrhoponyms, trades, registry tupology, abbreviations and hyphenated words.

These pages are available in the READ platform in the RH_Girona collection. Image 2 shows an image of the Girona dataset.

3.3 Planned Datasets

For D5.10 the LSD dataset will be extended. Additionally, the $GTset_ABP$ will be the basis for further competitions (e.g. table recognition) at ICFHR 2018 (proposal planned). Also the cBad dataset will be extended with new document types (tables,

⁵http://data.matricula-online.eu/de/deutschland/passau/

⁶http://gendb.bistum-passau.de/



Figure 2: Example image of Girona dataset.

postcards) to create a new layout analysis dataset. Also a follow up competition of cBad is planned.

Currently a new dataset for writer identification is in progress. This dataset is created using the document images of the Passau Diocesan Archive, will be the first dataset which covers the evolving handwriting style over years. Currently, 28 writers have been selected with pages written within a time period of up to 20 years. Resulting in an uneven distributed dataset with about 1.800 document images. The pages will be analyzed further if preprocessing steps are needed and then the new dataset will be created. Furthermore, new tasks on this dataset for researchers have to be described which reveals the invariance of the methods to handwritings over decades.

References

- M. Diem, F. Kleber, S. Fiel, T. Grüning, and B. Gatos, "Icdar2017 competition on baseline detection (cbad)," in 14th IAPR International Conference on Document Analysis and Recognition (ICDAR17). IEEE Computer Society Press, 2017, pp. 1355–1360.
- [2] S. Fiel, F. Kleber, M. Diem, V. Christlein, G. Louloudis, N. Stamatopoulos, and B. Gatos, "Icdar 2017 competition on historical document writer identification (historical-wi)," in 14th IAPR International Conference on Document Analysis and Recognition (ICDAR17). IEEE Computer Society Press, 2017, pp. 1377–1382.
- [3] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, "Icdar 2017 competition on document image binarization (dibco 2017)," in 14th IAPR International Conference on Document Analysis and Recognition (ICDAR17). IEEE Computer Society Press, 2017, pp. 1395–1403.
- [4] J. A. Sanchez, V. Romero, A. H. Toselli, M. Villegasy, and E. Vidal, "Icdar2017 competition on handwritten text recognition on the read dataset," in 14th IAPR

International Conference on Document Analysis and Recognition (ICDAR17). IEEE Computer Society Press, 2017, pp. 1383–1388.