Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 139–149 DOI: 10.15439/2024F8851 ISSN 2300-5963 ACSIS, Vol. 41

Converting German Historical Legal Documents to TEI XML including challenges with Table Extraction

Thomas Reiser* , Petra Steiner†
* University of Koblenz, Germany,
Email: treiser@uni-koblenz.de
† Federal Institute for Vocational Education and Training (BIBB), Bonn, Germany,
Email: steiner@bibb.de

***Abstract*—The occupations archive at the Federal Institute for Vocational Education and Training contains thousands of historical German Vocational Education and Training (VET) and Continuing Vocational Education and Training (CVET) regulations from the last 100 years. However, these are hardly accessible because they are currently only available in their original paper form. We present a workflow that transcribes images of these regulations into the TEI XML format which preserves the logical document structure and stores metadata. This paper addresses issues caused by poor page segmentation of the applied optical character recognition (OCR) methods and presents rules that can reconstruct a large part of the documents' hierarchy. A straightforward table recognition method for tables with borders is presented, as well as a metadata extraction procedure for the selected data set. While our approach is generic and functional, further research is necessary to develop a fully automated and more robust workflow.**

I. Introduction

IT IS evident that legal texts constitute an indispensable component of labor market research. The occupations archive at the Federal Institute for Vocational Education and Training (Bundesinstitut für Berufsbildung, BIBB) encompasses a multitude of historical regulations pertaining to vocational education and training (VET) and continuing vocational education and training (CVET). The regulations in the occupations archive are from the 1920s, the Third Reich, German Democratic Republic, and also the Federal Republic Germany. However, these documents are currently accessible to only a select few individuals with access to this archive. The objective is to digitize the archive by generating transcripts of the documents in the Text Encoding Initiative (TEI) XML format. This format is capable of capturing all logical text elements and layout elements such as page beginnings, footnotes, page headers, and line breaks.

A feasibility analysis and a preliminary draft of a pipeline for this process have been developed using a data set that is already available as digital images. The selected data set comprises 600 VET and 383 CVET regulations from 1969 to 2022. As the Vocational Training Act (Berufsbildungsgesetz, BBiG) of 1969 established a framework for the majority of these regulations and they are all published in the Federal Gazette (Bundesanzeiger), these documents are open to the public and allow for a well-structured rule-based approach. Once the documents in the archive are available as digital images, we aim to optimize the transcription process for them by applying state of the art layout analysis and structure recognition methods.

In this paper, our research questions are:

1) How can VET and CVET regulations from 1969 to 2022 that have been published in the Federal Gazette be digitized into fully structured TEI XML documents?
2) How are the selected documents structured?
3) How can errors in these documents be detected when there is hardly any ground truth available?

This paper presents the results of the feasibility analysis and is divided into five sections. The first section provides an introduction into related works about document digitization efforts and the selected document collection, consisting of over 900 German training regulations. The second section presents a brief overview of the state of the art in the generation of text hierarchy from unstructured texts, including layout analysis, OCR, and classification models for text hierarchy recognition. The third section outlines the methodological background and the utilized pipeline, including scan preprocessing methods, OCR, postprocessing steps from the layout analysis, recognition of different text elements such as lists, headings, etc., the transcription into the target data format, TEI XML, and table recognition. The fourth section presents the experimental results and evaluation of this novel approach by comparing extracted metadata to the small existing ground truth and analyzing properties of the generated transcripts. The last section presents the conclusions and an outlook for further research on documents of the occupations archive, once its regulations are available as digital images.

A. Motivation

The digitization of historical documents has gained particular interest in recent years, with numerous approaches emerging to address this task. Optical character recognition (OCR) represents a foundational technology for digitization, with a significant research focus and a range of established tools. This paper describes a basic transcription pipeline for German

training regulations that have been published in the Federal Gazette, to generate fully structured TEI XML documents, maintaining the logical document structure and text hierarchy of the digitized documents.

There are numerous methodologies to address the task of document digitization. One basic approach is the detection of the entire text within the document images, as exemplified by the methodology employed in the case of the Finnish newspaper digitization project[1], where the objective is to generate an ALTO XML document that contains all recognized text, utilizing the Tesseract OCR engine. An alternative approach is shown in [2], where the authors model the text structure of legal texts in Austria and align the recognized text to this predefined structure, thereby improving the structured recognition of text.

More advanced methods employ the OCR results to construct structured data from the text images. The authors of [3] use OCR to digitize invoice papers and to structure the recognized information, such as product description, quantity, and price. Similarly, in the study by [4], the names of judges at German federal courts from 1950 to 2019 were extracted from the Federal Gazette by applying OCR to these publications.

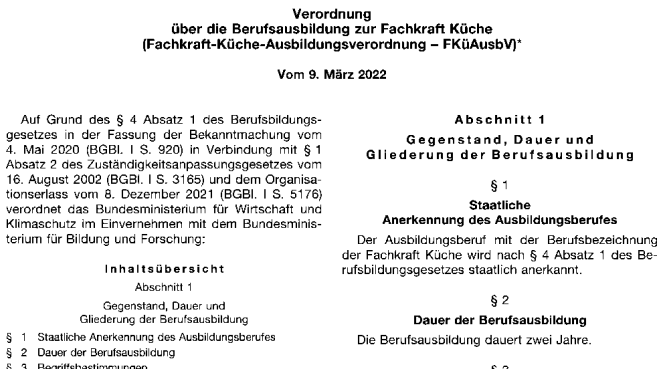
Not only table-structured data, but also graph data can be extracted from text images. T2KG, an NLP tool that can construct knowledge graphs from text, is presented in [5]. Another way to generate knowledge graphs from text using Open Information Extraction is introduced in [6]. Other approaches include contextual extraction and representation methods based on knowledge graphs, ontologies and taxonomies, see [7], [8], [9], [10], [11], [12].

For better accessibility, some OCR workflows are embedded into web applications. One of the largest efforts for this is OCR4all [13] which allows the usage of different preprocessing steps, segmentation methods, and OCR models. It also enables the interaction with each of the process steps so users can do corrections at intermediate results to improve the overall outcome. However, there are also less extensive tools which allow the management of digitized document collections, for example in [14], [15]

B. Data Set

The data set comprises 600 VET and 383 CVET regulations from 1969 to 2022. All of these regulations are published in the Federal Gazette and follow a similar layout structure. Each regulation begins with a short preamble, which is sometimes followed by a table of contents. A regulation page that illustrates most of the layout elements described here is shown in Figure 1.

Every regulation is comprised of multiple paragraphs. These paragraphs commence with a section sign (§) and are followed by the section number. One line below, the section headline is displayed. The paragraphs are further structured into sections. These sections can be either a single block of text or a set of enumerated sections, with the section number between parentheses ((1), (2),...).



Verordnung
über die Berufsausbildung zur Fachkraft Küche
(Fachkraft-Küche-Ausbildungsverordnung – FKüAusbV)*

Vom 9. März 2022

Auf Grund des § 4 Absatz 1 des Berufsbildungsgesetzes in der Fassung der Bekanntmachung vom 4. Mai 2020 (BGBl. I S. 920) in Verbindung mit § 1 Absatz 2 des Zuständigkeitsanpassungsgesetzes vom 16. August 2002 (BGBl. I S. 3165) und dem Organisationserlass vom 8. Dezember 2021 (BGBl. I S. 5176) verordnet das Bundesministerium für Wirtschaft und Klimaschutz im Einvernehmen mit dem Bundesministerium für Bildung und Forschung:

Inhaltsübersicht

Abschnitt 1

Gegenstand, Dauer und Gliederung der Berufsausbildung

§ 1 Staatliche Anerkennung des Ausbildungsberufes
§ 2 Dauer der Berufsausbildung
§ 3 Begriffsbestimmungen

Abschnitt 1

Gegenstand, Dauer und Gliederung der Berufsausbildung

§ 1

Staatliche Anerkennung des Ausbildungsberufes

Der Ausbildungsberuf mit der Berufsbezeichnung der Fachkraft Küche wird nach § 4 Absatz 1 des Berufsbildungsgesetzes staatlich anerkannt.

§ 2

Dauer der Berufsausbildung

Die Berufsausbildung dauert zwei Jahre.

Fig. 1: Excerpt of the first page of the VET regulation for kitchen qualified professionals from 2022.

These enumerations are hierarchically structured, with each element capable of containing a further enumeration (a), b), ...), which itself can be further nested (aa), bb), ...). In some regulations, the paragraphs are grouped in sections (1. Abschnitt or Abschnitt 1) or parts (Erster Teil or Teil 1). There are also a few exceptions where the sections are grouped in parts. The following combinations of the aforementioned structure elements are possible:

1) paragraphs
2) paragraphs with table of contents
3) paragraphs in sections
4) paragraphs in sections with table of contents
5) paragraphs in parts
6) paragraphs in parts with table of contents
7) paragraphs in sections in parts with table of contents

II. Related Literature

In order to get a large digitized text corpus, many approaches can be considered. Structuring the documents takes text digitization a step further because not only is text recognized, but also logical units, such as paragraphs, section, listings, tables, and figures need to be detected to make the output more meaningful.

While we plan to develop a more advanced approach for documents of the occupations archive, we conducted a feasibility and usability analysis of such a pipeline a first subset of VET and CVET regulations that have been openly available in the Federal Gazette. The results of these are discussed in this paper.

In first approaches for document structuring from the 2000s, human knowledge about the document has been used to define text- and layout-based rules to extract the text structure [16], [17], [18]. The first version of our digitization workflow also uses predefined rules, but we aim to recognize patterns using layout and text features automatically in future research to structure texts with different layouts. Similar to [2], a certain structure is defined to recreate the text hierarchy. There are a few different possible structures consisting of the same structure elements which are also described in subsection I-B.

Although some standard conversion tools, like Vertopal, to convert files that contain text into markup languages exist [19], they assume that the text in the documents that need to be converted contain correct structural information, which cannot be guaranteed by default OCR models. While they can be used to create files in HTML or TEI XML, the output files often do not represent the text hierarchy or logical document structure but are rather another representation of recognized text areas, lines, and text.

For metadata extraction, machine learning models like they are used in GROBID [20] can be used to extract metadata like title and authors. It also allows the recognition of references and citations, and the detection of the abstract. Despite it having a lot of useful features, it was trained on scientific articles and performs especially well on them. To use GROBID on legal texts like the training regulations in this article, it would be necessary to fine tune the model, for which training data would be required. Besides this, it again depends on good page segmentation and reliable text recognition.

While some other tools like PdfPig[1] or PDFMiner[2] also allow layout analysis from words in a PDF file, they are based on heuristics that do not match all document layouts, and again, depend on previously recognized words.

In a more recent publication [21], a tree structured document hierarchy is generated using the HELD model where text areas are inserted at a specific level in the tree hierarchy depending on the output of a binary classification function. This function compares features of the element that needs to be inserted and elements that have already been classified, e.g., format features or consecutive numbers in this text level. While this is a promising approach, it again requires training data and is not straightforward to implement. Besides this, it again depends on correct page segmentation which is at the moment not guaranteed. Because, in this article, we focus on a smaller feasibility analysis and the classification of elements is done by iterating over each element, it is easier and more efficient to use predefined rules, as mentioned above. This approach may become more relevant on the target data set of the occupations archive.

Our method aims at using out of the box OCR models and preprocessing techniques with rule based postprocessing. While OCR-D [22] is a tool that combines preprocessing with OCR, it was not fast enough on the available hardware to be efficiently used within the given time constraints for our use case. Because it has shown comparable results to other commerical alternatives in [23], [24], the Tesseract OCR engine has been selected for this work.

Although Tesseract delivers state of the art OCR results, it is still no one fits all solution. As mentioned before, many tools depend on correct layout analysis. To improve the results of layout analysis, many researchers currently employ object detection methods to pretrain models such as Faster R-CNN, YOLO, etc. to recognize text regions and classifying them into heading, list item, title, and more [25], [26], [27], [28].

[1]https://github.com/UglyToad/PdfPig
[2]https://github.com/pdfminer/pdfminer.six

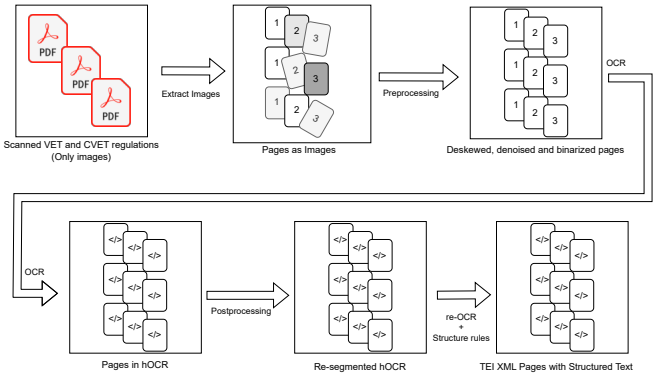


Fig. 2: Overview of the presented pipeline.

Similar to the layout analysis, also tables are recognized in a similar manner, using deep learning object detection methods [29], [30]. However, again, due to the lack of labeled training data for this data set, these are not applied in the feasibility analysis in this paper, but will likely be used on the regulations in the occupation archive.

TEI XML has been selected as the target data format because it is recommended by the German Research Foundation (DFG) as a good standard for long time archives of documents [31]. While PDF is a proprietary standard that is stored in binary files, markup languages like XML can be read by almost any computer without installing additional software that can read PDF files. Besides that, these files can be efficiently stored in XML databases like eXist-db [32] to manage the document collection. eXist-db also allows the addition of plugins, such as a versioning plugin which allows memory-efficient saving of different versions of the same documents and a fast restore, if an older version is required. Especially in this automated setting, where errors need to be expected, this can be very helpful. Another useful plugin is TEI Publisher [33] which allows a user friendly management of the XML database as well as viewing the documents in a well human-readable ways and editing uploaded TEI XML files. Besides this, it allows viewing the text along with the original images which can facilitate the error correction process. Therefore, TEI XML has been found to be a suitable choice for a target data format when digitizing larger text corpora.

III. Methodology

In this section, the developed pipeline is presented step by step. An illustration is depicted in Figure 2. The pipeline starts with a list of PDF or image files. If PDF files are given, the pages are extracted as images at 300 dots per inch (DPI) which is considered to be a good trade-off between image quality and storage efficiency [31].

A. Preprocessing

A common step when it comes to document digitization with OCR is image preprocessing [34]. The scanned pages

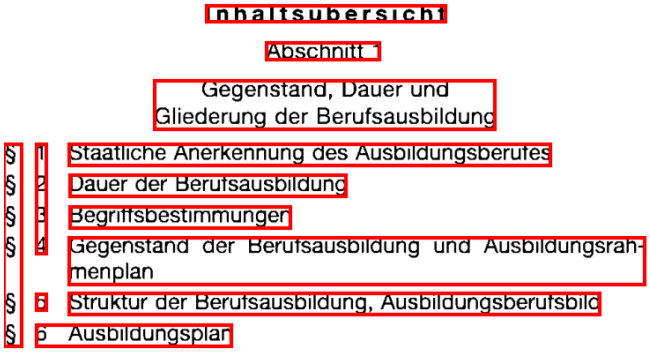


Fig. 3: Bounding boxes of lines recognized by Tesseract in the table of contents of the VET regulation for kitchen qualified professionals in 2022.

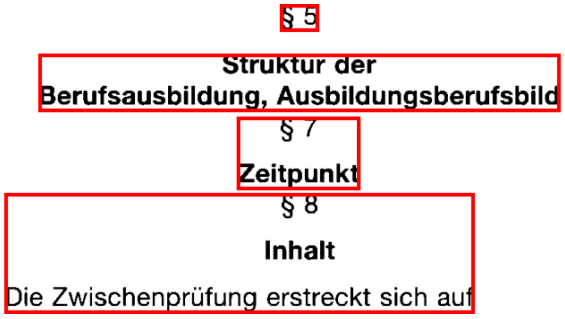


Fig. 4: For each of the three paragraph headlines, Tesseract recognized the text area in a different way.

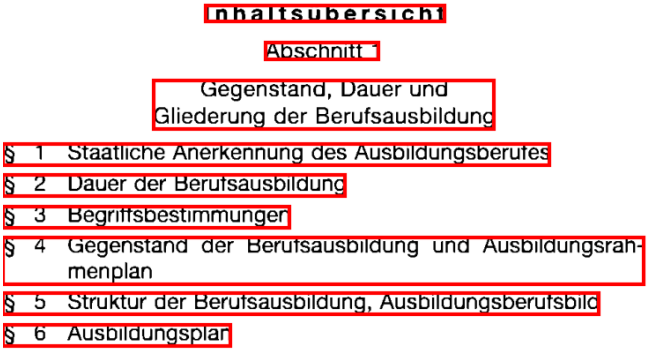


Fig. 5: Bounding boxes of lines after correcting the errors of the Tesseract output shown in Figure 3.

can be skewed, contain noise such as water damages or ink stains, white margins from the paper, and are usually in color. For most OCR engines, a clean black and white image with no noise and straight pages is required. To achieve this from the original images, scantailor[3] is used, an open-source tool that supports many different preprocessing steps, including image binarization, page deskewing, content detection, and noise removal. These steps are applied and afterwards, OCR is executed.

B. OCR

The character recognition is done by an OCR engine. One of the most frequent used OCR engines in literature is Tesseract which is open-source and can in some cases give comparable results to commercial alternatives like ABBYY FineReader [23], [24]. It is able to recognize most text areas and characters correctly, but sometimes has trouble with text that is aligned on the same width across multiple lines as shown in Figure 3.

Another problem is the uneven indentation of text in the document. As shown in Figure 4, Tesseract paritally recognizes the paragraph headlines as paragraph number and headline in two text areas, one text area, or even including a part of the paragraph's text.

To fix these errors, a postprocessing step is introduced that fixes the most frequent OCR errors that have been found by investigating the results by hand.

C. Post-Correction

To solve the aforementioned issues, conditions on when and how to restructure bounding boxes are introduced. Recognized text areas are defined by bounding boxes (bboxes), consisting of left, upper, right, and lower border. For the issue shown in Figure 3, a threshold was defined and all text areas on the same height are found and sorted ascending by their right border. Afterwards, if the distance between the left border of a text element in one of the sorted clusters to the right border of its predecessor is smaller than the predefined threshold or the elements overlap, they need to be merged. Once all areas that need to be merged have been identified, the area with the furthest right border in each of these clusters is expanded to the border that is the furthest left in the cluster. After all merges have been completed, each element that has been used to expand another text area, is removed. The result is shown in Figure 5.

For the inconsistent assignment of lines to text areas as shown in Figure 4, a measure is defined for when to split text areas. First, the line distances for subsequent lines in each text area are computed. These distances for an example page are plotted in Figure 6.

As shown in this example, there seem to be two mass centers of line distances: One for lines that belong into the same text area, another for lines that should be within two different text areas. A common way to detect a fixed number of clusters is k-Means Clustering [35]. Setting $k = 2$ will detect both clusters in the line distances. If the line distance is larger than the largest value of the cluster with the lower centroid, marked by the red line in the plot, the text area is split between the two respective text lines. The result will cause all paragraph headlines to be split in two boxes, *§ paragraph number* and *paragraph headline*. The resulting bboxes of the text areas are now as shown in Figure 7. The text areas now only contain coherent text segments. This

[3]https://github.com/trufanov-nok/scantailor-universal

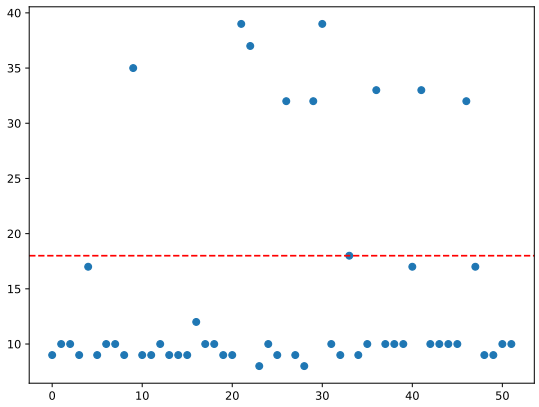


Fig. 6: Line distances within text areas of the second page of the VET regulation for kitchen qualified professionals in 2022.

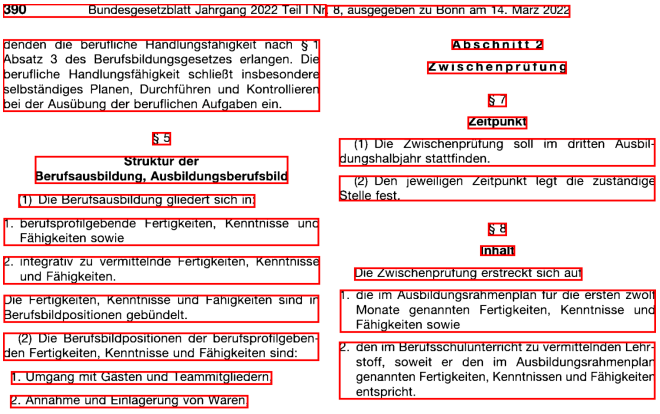


Fig. 7: The bounding boxes of Figure 4 after splitting the text areas according to the described procedure. Each text area now contains a single text element.

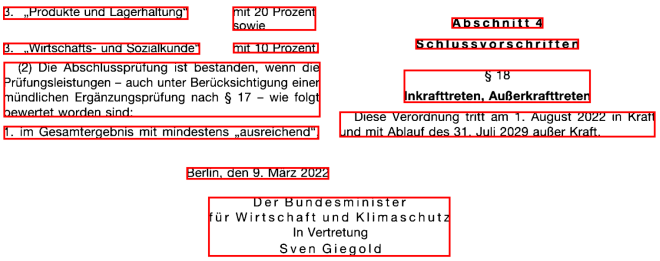


Fig. 8: All regulation bodies end with a centered text area containing the responsible federal minister and his ministry.

procedure is applied for each document and formalized as follows:

Let

- c be the set of text areas that have been recognized by Tesseract. Here, a text area is defined by its bounding box and the list of lines in this area.
- l_{c_i} be the list of detected lines in the text area $c_i \in c$. A line is defined by its bounding box and the list of words in it. Words are defined by their bounding box and text.
- $l^j_{c_i}$ be the j-th element of the list of lines l_{c_i} that is ordered by the y_1-coordinate of the bounding box of each line, i.e., its upper border.
- $bbox(l)$ be the bounding box of a line l, defined by x_1, y_1, x_2, y_2 where x_1 defines the left border, y_1 the upper border, x_2 the right border, and y_2 the lower border of the line.
- $d = \bigcup_{c_i \in c} \left\{ bbox(l^j_{c_i})_{y_1} - bbox(l^{j-1}_{c_i})_{y_2} \,\middle|\, j \in \{1 \dots |l_{c_i}|\} \right\}$ be the set of line distances between lines that are in the same text area.
- $kmeans_k(D) = \{D_1, ..., D_k\}$ be the result of k-Means clustering where k determines the number of clusters to detect and $D_i \subseteq D$ are the detected clusters.

The threshold to split text areas for this document is then calculated by:

$$split\_thresh = max\left\{argmin_{d_i \in kmeans(d)}\left\{min\left(d_i\right)\right\}\right\}$$

After these postprocessing steps have been applied to improve text segmentation, OCR is applied a second time for each resegmented text area to resolve errors that have been caused by bad segmentation.

D. Metadata Extraction

The texts in the selected corpus contain some metadata about themselves: title, release date, place of publication, and responsible federal ministry and minister. Since the layout is the same for all of the selected texts, the corresponding elements can be detected by patterns and the gathered information can be stored. For TEI XML documents, all metadata is stored in a `fileDesc` within a `teiHeader` element. The first element giving information about the document itself is the title which is the first element after the page header. In the OCR result, the title is extracted and encoded in a `title` tag. For the author, the last element of the body text is considered. This element contains the responsible federal ministry and minister. An example is shown in Figure 8. To extract the name of the federal minister, his ministry and, in some cases, his substitute, again, a regular expression is used:
`Der\s*Bundesminis[ftl]er\s*(.*?)(?: In\s* Vertretung|$)` for male and
`Die\s*Bundesminis[ftl]erin\s*(.*?)(?: In\s *Vertretung|$)` for female ministers.
`title` and `author` are stored in the `titleStmt`, with the `author` containing `persName` for his name and `orgName` for his organization's name.

Finally, the element above the author text area contains the city and publication date, as also seen in Figure 8. The text is split at the comma to separate city from date and both information is stored in the target document as `pubPlace` and

`date` within the `publicationStmt` of the `fileDesc`. As these elements are not relevant for the document content, they are no longer considered after being encoded in the metainformation part.

The `teiHeader` for the VET regulation for kitchen qualified professionals from 2022 is shown here:

```
<TEI version="3.3.0"
  ↪ xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Verordnung
                  <lb/>über die
                   ↪ Berufsausbildung zur
                   ↪ Fachkraft Küche
                  <lb/>(Fachkraft- Küche-
                   ↪ Ausbildungsverordnung
                   ↪ - FKüAusbV)*
                </title>
        <author>
          <persName>Sven
           ↪ Giegold</persName>
          <orgName>Bundesministerium für
           ↪ Wirtschaft und
           ↪ Klimaschutz</orgName>
        </author>
      </titleStmt>
      <publicationStmt>
        <publPlace>Berlin</publPlace>
        <date>9. März 2022</date>
      </publicationStmt>
    </fileDesc>
  </teiHeader>
  ...
</TEI>
```

E. Transcription to TEI XML

The regulations always have a page header which should be represented in the encoded document but not part of the text. To recognize these headers, the uppermost element and all elements on the same height are selected and their text is concatenated. Because the text is in a two column layout, Tesseract is sometimes not able to recognize the page header as a single line and splits it to fit into the two column layout. Therefore, all elements on the same height as the uppermost element on a page are also considered to be part of the header.

As target format for the image transcripts, TEI XML version P5 [36] has been selected which is the latest version at the time this article was written. For the transcription into fully structured TEI XML files, text based rules, i.e., regular expressions, have been defined in order to classify text elements that start a new paragraph, section or part. These rules are:

- parts (*Teil N* or *N-ter Teil*):
  `^Teil\s*\d\s*$` or `^\b\w+er\b\s*Teil\s*$`
- sections (*Abschnitt N* or *N. Abschnitt*):
  `^Abschnitt\s*\d+$` or
  `^\d+\.\s*Abschnitt\s*$`
- paragraphs (*§ N Headline*):
  `^(§|5|8|S|s|&|\$)*\s*\d+\s*$`

Text within a paragraph is also structured in a given hierarchy, the patterns for these categories are:

1) (1) Enumerated section:
   `^[\(\}\]]\d+[\)\}\]]`
2) 1. First level enumeration:
   `^\d+\s?[\.,]\s?.*`
3) a) Second level enumeration:
   `^[a-z]\s*[\]\)\}]\s`
4) aa) Third level enumeration:
   `^([a-z])\1{1}\s*[\]\)\}]\s`
5) aaa) Fourth level enumeration:
   `^([a-z])\1{2}\s*[\]\)\}]\s`
6) aaaa) Fifth level enumeration:
   `^([a-z])\1{3}\s*[\]\)\}]\s`
7) aaaa) Sixth level enumeration:
   `^([a-z])\1{4}\s*[\]\)\}]\s`
8) Everything that does not match any of the specified patterns is considered to be a non-enumerated paragraph section.

All patterns are designed in a way that there is space for errors in the text recognition. If none of the mentioned patterns matches the recognized text, the text is considered as a non-enumerated paragraph section. The classification with these regular expressions is used to recognize headlines, paragraphs, and list items on different levels to recreate the original text hierarchy.

An excerpt of the generated TEI XML document of the VET regulation for kitchen qualified professionals from 2022 is shown here:

```
<TEI version="3.3.0"
  ↪ xmlns="http://www.tei-c.org/ns/1.0">
  ...
    <div n="1" type="abschnitt">
      <head>Abschnitt 1<lb/>Gegenstand,
       ↪ Dauer und<lb/>Gliederung der
       ↪ Berufsausbildung</head>
      <div n="1" type="paragraph">
        <head>8 1<lb/>Staatliche</head>
        <div type="section">
          <p>Anerkennung des
           ↪ Ausbildungsberufes</p>
        </div>
        <div type="section">
          <p>Der Ausbildungsberuf mit der
           ↪ Berufsbezeichnung<lb/>der
           ↪ Fachkraft Küche wird nach 8
           ↪ 4 Absatz 1 des
           ↪ Be-<lb/>rufsbildungsgesetzes
           ↪ staatlich anerkannt.</p>
        </div>
```

```
      </div>
      <div n="2" type="paragraph">
        <head>82<lb/>Dauer der
         ↪ Berufsausbildung</head>
        <div type="section">
          <p>Die Berufsausbildung dauert
           ↪ zwei Jahre.</p>
        </div>
      </div>
          ...
    </div>
  ...
</TEI>
```

F. Table Recognition

While the appendix of VET regulations from the late 1970s and later consists mostly of the tabular training schedules, in most of the CVET regulations' appendices, example certificates are included. This part focuses on encoding the tables in the appendices. All text areas in the appendix that have not been found as part of a table are encoded as a p-tag with the area's text in the appendix without any further structuring.

Table recognition is a common challenge in document digitization with a lot of ongoing research. As all tables in the selected data set have borders which are also recognized by Tesseract, these lines are used to detect table cells. An example on how the lines are recognized is shown by the red lines in Figure 10. If a line has higher width than height, it is considered to be a horizontal line. If its height is larger than its width, it is considered as a vertical line. Vertical lines are sorted ascending by their horizontal position and horizontal lines are sorted ascending by their vertical position. Then, each line in the sorted set of vertical lines is used with its predcessor to get the left and right border for a table box. The same method is used for horizontal lines to get the lower and upper border. In the detected cells, OCR is applied again and the recognized text is used to fill the table content. Starting from the second level enumeration pattern from subsection III-E, the patterns can be reused to recognize enumerations and plain text in the table cells.

Although this is a very simple approach, it mostly works on this data set. There are, however, some issues with this method. Like most of the pipeline, this procedure depends on the OCR results. In some cases, Tesseract was unable to recognize all lines, as seen in Figure 9, or only fractions of them.

To get the entirety of vertical lines that are only detected in fractions, all vertical lines are expanded to the lowermost and uppermost border of all lines. Horizontal lines are expanded to the left- and rightmost borders of all lines. Afterwards, all lines are compared on how close their borders are to each other to remove duplicate lines. Although this ensures that all lines are recognized, this method does not allow combined cells like they are found in the data set and also assumes that the upper, left, right, and lower border of the tables have been identified correctly. The detection of cells that need to be combined remains an open issue for now. Another issue with

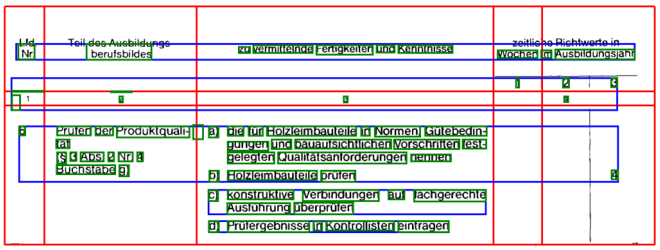


Fig. 9: In this case, the line between the last two table columns has not been detected at all.

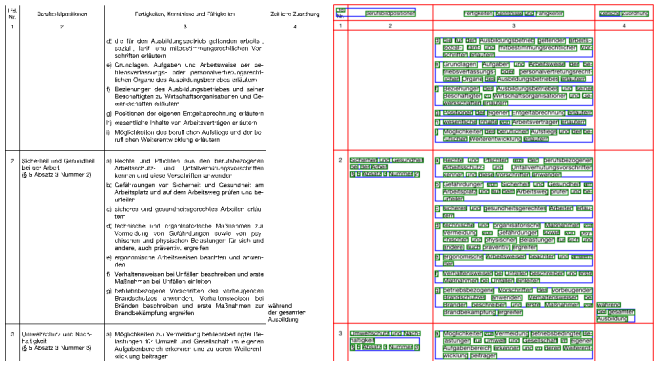


Fig. 10: The table lines and text on the left side are recognized as shown on the right side. Two of the horizontal table lines are detected longer than they actually are.

this procedure that needs to be addressed is caused by lines that are not recognized at all. In this case, two neighboring rows or columns will be considered as one.

Once we digitize the entire occupations archive, we will use more advanced table recognition methods because the older regulations, especially from the German Democratic Republic, contain borderless tables with complex structure and many empty cells.

IV. Evaluation

As there is hardly any ground truth available for a proper evaluation, a usability analysis as well as a small corpus analysis is conducted in order to find anomalies and weaknesses of the proposed procedure.

The web service used for document collection offers some metadata such as title, number of pages, and release year. To test the validity of the extracted metadata, extracted titles have been compared to the titles given in the original data. To allow some OCR errors when comparing the headlines, a maximum Levenshtein distance of eight was allowed between two headlines to be considered from the same regulation. Because there are many long titles, this distance allowed a reasonable amount of differences between the actual and detected titles.

For VET regulations, only three of the 600 regulation titles could not be mapped. This was due to the selected Levenshtein distance being too small or bad text segmentation by Tesseract.

With similar errors, six of the 383 CVET regulations could not be mapped to any of the titles in the ground truth. However, another error where the expected layout was not properly recognized by Tesseract was found such that the preamble text was considered to be the title.

Because the release years are also given in the ground truth, the regulations per year in the ground truth and the digitized collection are counted and compared. The results are shown in Figure 11. Dates that have been recognized by the pipeline but do not lie in the specified interval between 1969 and 2022 have been discarded. Additionally, in some documents, no release date was recognized at all. These two circumstances result in less digitized than original regulations that are considered in the plot.

Although there are some years with as many regulations in the digitized data set as in the ground truth, in most years, there are more documents in the ground truth. This issue has been caused by damages files that had metadata but could not be downloaded because the files were damaged. As shown in Figure 11b, in 2012, there have been more documents in the digitized data set than in the ground truth as a result of bad OCR. Due to less damaged files in the CVET regulations, there are more matches to the ground truth than for the VET regulations.

In subsection I-B, the different layout types have been presented. How these layout types are distributed among the VET regulations is shown in Figure 12.

As seen in Figure 12a, most VET regulations consist only of paragraphs that are not part of sections (Abschnitt) or parts (Teil). Also, only 12.6% of all regulations have a table of contents. All VET regulations with a table of contents consist of parts or sections which contain the paragraphs. Parts mostly appear in regulations for an industrial sector and regulations that cover multiple professions, e.g., the regulation on vocational training in the laboratory field of chemistry, biology and coatings (Verordnung über die Berufsausbildung im Laborbereich Chemie, Biologie und Lack), or in the regulation on vocational training in the weaving industry (Verordnung über die Berufsausbildung in der Weberei-Industrie). It becomes visible that some of the layout types can be found in only few regulations.

Like the VET regulations, most CVET regulations only consist of paragraphs. However, the second most relevant text structure type, consisting of sections and paragraphs without a table of contents, makes up almost a third of all CVET regulations. Less than 5% of all CVET regulations have a table of contents. Any other layout type is hardly present in the CVET regulations. The entire layout type distribution can be seen in Figure 12b.

For an insight how many parts, sections, and paragraphs most regulations consist of and to detect anomalies, bar charts with the number of them and how frequent this number occurs have been created. The number of parts for CVET regulations is shown in Figure 13.

There are two noticeable regulations with only one part. This shows one of the weaknesses with a rule-based approach that relies on text: In these regulations, a text area that contains only the line *written part* (schriftlicher Teil) exists. This matches the regular expression that matches patterns of n-th part (Erster Teil, ...) and causes the workflow to recognize it as a headline for a part although the regulations themselves contain no parts.

A similar anomaly can be seen when counting sections (Abschnitt) in the VET regulations, as seen in Figure 14.

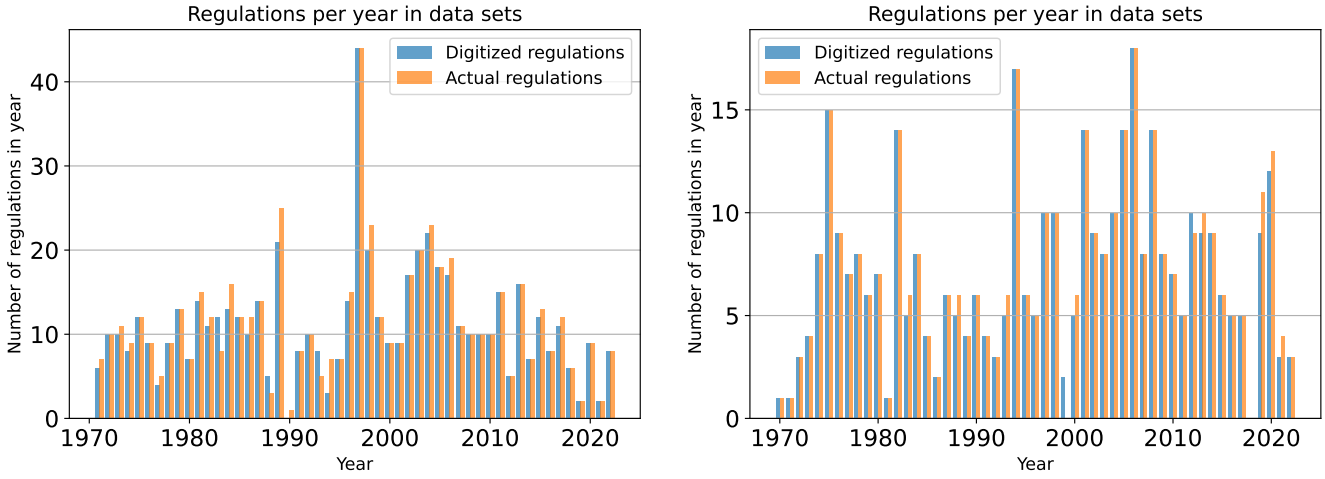
There are three regulations with only one section. In these three cases, however, only one section has been properly detected. This either hints at bad OCR results or a bug in the pipeline implementation that needs to be fixed.

V. Conclusions and Outlook

In this article we proposed a basic pipeline that takes images as input, applies image preprocessing, OCR, and postcorrection to process the OCR output into fully structured TEI XML. These transcripts have been used for a first insight into the digitized corpus of 600 VET and 383 CVET regulations that have been published in the Federal Gazette from 1969 to 2022 to detect anomalies which helped identifying weaknesses and bugs in the pipeline and its implementation. Although the basic approach looks promising on the selected documents, it strongly depends on the corpus-specific layout and good OCR results.

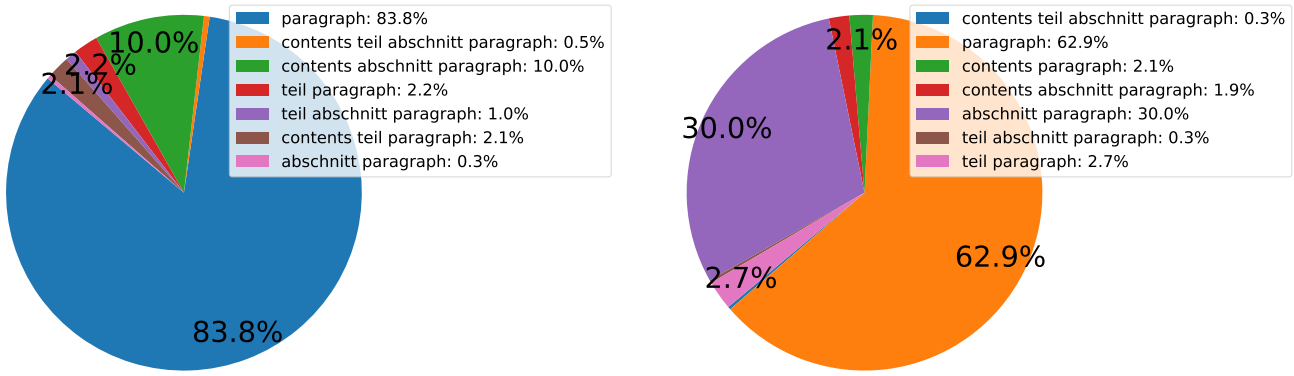
Although Tesseract performs well in many domains, it shows some limitations for the selected documents which make the described postprocessing necessary. It remains to try out different OCR engines, as some commercial tools like ABBYY FineReader not only recognize text, but also some text style information such as underlined, bold, and italic text. This work also shows how much information about a text can be gained using OCR and domain knowledge. Besides that, other tools such as layout-parser [25] or models trained on the annotated DocLayNet dataset [28] that also contain more information than just recognized text will be considered. There are also more advanced table recognition methods such as CascadeTabNet [30] and Table Transformer [37] that need to be evaluated on our data set in future research.

Structured text contains more information and features than plain text that is mapped to regions on an image. Without having to label a large amount of training data, document structure information has been be extracted to get first insights into the digitized text corpus. In future research, we aim at the automated recognition of layout- and text-based rules in order to find these rules for any input document automatically and more flexibility for other documents. Once all regulations of the occupations archive have been digitized, the application of NLP and text mining methods will be much easier and more targeted because each paragraph, section and table are already identified, making a mapping to certain areas of a training program more feasible. Besides that, large, structured text corpora are also useful for the training of large language models and can support researchers in different domains to retrieve information from historical documents much faster.



(a) (b)

Fig. 11: VET (left) and CVET (right) regulations per year according to the ground truth and the extracted data.



(a) (b)

Fig. 12: Distribution of layout types across all VET (left) and CVET (right) regulations in the data set. The percentage describes how many VET (or CVET) regulations consist of these elements. E.g., in the left plot, "paragraph" describes the percentage of VET regulations that consist only of paragraphs, but are not grouped in sections and do not contain a table of contents.

References

[1] M. Koistinen, K. Kettunen, and J. Kervinen, "How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine," *Proc. of LTC*, pp. 279–283, 2017.

[2] A. Nabizai and H.-G. Fill, "Eine Modellierungsmethode zur Visualisierung und Analyse von Gesetzestexten," *Jusletter IT*, February 2017. [Online]. Available: http://eprints.cs.univie.ac.at/5131/

[3] V. N. Sai Rakesh Kamisetty, B. Sohan Chidvilas, S. Revathy, P. Jeyanthi, V. M. Anu, and L. Mary Gladence, "Digitization of Data from Invoice using OCR," in *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, 2022. doi: 10.1109/ICCMC53470.2022.9754117 pp. 1–10.

[4] H. Hamann, "The German Federal Courts Dataset 1950–2019: From Paper Archives to Linked Open Data," *Journal of empirical legal studies*, vol. 16, no. 3, pp. 671–688, 2019. doi: https://doi.org/10.1111/jels.12230

[5] N. Kertkeidkachorn and R. Ichise, "T2KG: An End-to-End System for Creating Knowledge Graph from Unstructured Text," in *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[6] J. L. Martinez-Rodriguez, I. Lopez-Arevalo, and A. B. Rios-Alvarado, "OpenIE-based approach for Knowledge Graph construction from text," *Expert Systems with Applications*, vol. 113, pp. 339–355, 2018. doi: https://doi.org/10.1016/j.eswa.2018.07.017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417418304329

[7] J. Dörpinghaus and A. Stefan, "Knowledge extraction and applications utilizing context data in knowledge graphs," in *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2019. doi: 10.15439/2019F3 pp. 265–272.

[8] J. Dörpinghaus, A. Stefan, B. Schultz, and M. Jacobs, "Context mining and graph queries on giant biomedical knowledge graphs," *Knowledge and Information Systems*, vol. 64, no. 5, pp. 1239–1262, 2022. doi: https://doi.org/10.1007/s10115-022-01668-7

[9] Y. Fettach, M. Ghogho, and B. Benatallah, "Knowledge graphs in education and employability: A survey on applications and techniques," *IEEE Access*, vol. 10, pp. 80 174–80 183, 2022. doi: 10.1109/ACCESS.2022.3194063

[10] J. Dörpinghaus, S. Klante, M. Christian, C. Meigen, and C. Düing, "From social networks to knowledge graphs: A plea for interdisciplinary approaches," *Social Sciences & Humanities Open*, vol. 6, no. 1, p. 100337, 2022. doi: https://doi.org/10.1016/j.ssaho.2022.100337

[11] J. Dörpinghaus, V. Weil, and J. Binnewitt, "Towards the analysis of longitudinal data in knowledge graphs on job ads," in *The Workshop on Computational Optimization*. Springer, 2022. doi:

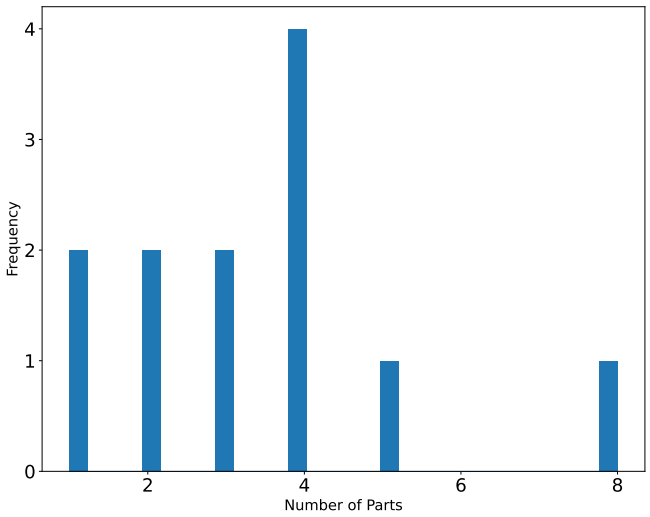


Fig. 13: The number of parts (Teil) in a CVET regulation on the x-axis and the frequency number of regulations with it on the y-axis.

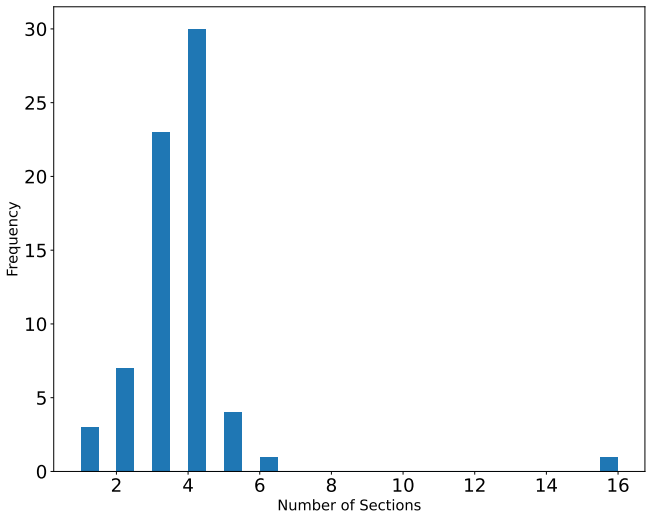


Fig. 14: The number of sections (Abschnitt) in a VET regulation on the x-axis and the frequency number of regulations with it on the y-axis.

https://doi.org/10.1007/978-3-031-57320-0_4 pp. 52–70.

[12] A. Fischer and J. Dörpinghaus, "Web mining of online resources for german labor market research and education: Finding the ground truth?" *Knowledge*, vol. 4, no. 1, pp. 51–67, 2024. doi: https://doi.org/10.3390/knowledge4010003

[13] C. Reul, D. Christ, A. Hartelt, N. Balbach, M. Wehner, U. Springmann, C. Wick, C. Grundig, A. Büttner, and F. Puppe, "OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings," *Applied Sciences*, vol. 9, no. 22, p. 4853, 2019. doi: https://doi.org/10.3390/app9224853

[14] J. M. Jayoma, E. S. Moyon, and E. M. O. Morales, "OCR Based Document Archiving and Indexing Using PyTesseract: A Record Management System for DSWD Caraga, Philippines," in *2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, 2020. doi: 10.1109/HNICEM51456.2020.9400000 pp. 1–6.

[15] S. Van Nguyen, D. A. Nguyen, and L. S. Q. Pham, "Digitalization of Administrative Documents A Digital Transformation Step in Practice," in *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, 2021. doi: 10.1109/NICS54270.2021.9701547 pp. 519–524.

[16] S. Tsujimoto and H. Asada, "Major components of a complete text reading system," *Proceedings of the IEEE*, vol. 80, no. 7, pp. 1133–1149, 1992. doi: 10.1109/5.156475

[17] J. v. Beusekom, D. Keysers, F. Shafait, and T. Breuel, "Example-based logical labeling of document title page images," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, 2007. doi: 10.1109/ICDAR.2007.4377049 pp. 919–923.

[18] S. Klink and T. Kieninger, "Rule-based document structure understanding with a fuzzy combination of layout and textual features," *International Journal on Document Analysis and Recognition*, vol. 4, no. 1, pp. 18–26, 2001. doi: https://doi.org/10.1007/PL00013570

[19] P. Pathirana, A. Silva, T. Lawrence, T. Weerasinghe, and R. Abeyweera, "A comparative evaluation of pdf-to-html conversion tools," in *2023 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, vol. 6, 2023. doi: 10.1109/SCSE59836.2023.10214989 pp. 1–7.

[20] P. Lopez, "Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications," in *Research and Advanced Technology for Digital Libraries*, M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, and G. Tsakonas, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. doi: https://doi.org/10.1007/978-3-642-04346-8_62. ISBN 978-3-642-04346-8 pp. 473–474.

[21] R.-Y. Cao, Y.-X. Cao, G.-B. Zhou, and P. Luo, "Extracting Variable-Depth Logical Document Hierarchy from Long Documents: Method, Evaluation, and Application," *Journal of Computer Science and Technology*, vol. 37, no. 3, pp. 699–718, 2022. doi: https://doi.org/10.1007/s11390-021-1076-7

[22] C. Neudecker, K. Baierer, M. Federbusch, M. Boenig, K.-M. Würzner, V. Hartmann, and E. Herrmann, "Ocr-d: An end-to-end open source ocr framework for historical printed documents," in *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, ser. DATeCH2019. New York, NY, USA: Association for Computing Machinery, 2019. doi: 10.1145/3322905.3322917. ISBN 9781450371940 p. 53–58. [Online]. Available: https://doi.org/10.1145/3322905.3322917

[23] A. P. Tafti, A. Baghaie, M. Assefi, H. R. Arabnia, Z. Yu, and P. Peissig, "OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym," in *Advances in Visual Computing*, G. Bebis, R. Boyle, B. Parvin, D. Koracin, F. Porikli, S. Skaff, A. Entezari, J. Min, D. Iwai, A. Sadagic, C. Scheidegger, and T. Isenberg, Eds. Cham: Springer International Publishing, 2016. doi: https://doi.org/10.1007/978-3-319-50835-1_66. ISBN 978-3-319-50835-1 pp. 735–746.

[24] M. Lundqvist and A. Forsberg, "A comparison of OCR methods on natural images in different image domains," 2020.

[25] Z. Shen, R. Zhang, M. Dell, B. C. G. Lee, J. Carlson, and W. Li, "Layoutparser: A unified toolkit for deep learning based document image analysis," pp. 131–146, 2021. doi: https://doi.org/10.1007/978-3-030-86549-8_9

[26] X. Zhong, J. Tang, and A. Jimeno Yepes, "Publaynet: Largest dataset ever for document layout analysis," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019. doi: 10.1109/ICDAR.2019.00166 pp. 1015–1022.

[27] Z. Shen, K. Zhang, and M. Dell, "A large dataset of historical japanese documents with complex layouts," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020. doi: 10.1109/CVPRW50498.2020.00282 pp. 2336–2343.

[28] B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar, and P. W. J. Staar, "Doclaynet: A large human-annotated dataset for document-layout segmentation," p. 3743–3751, 2022. doi: 10.1145/3534678.353904. [Online]. Available: https://doi.org/10.1145/3534678.3539043

[29] X. Zhong, E. ShafieiBavani, and A. J. Yepes, "Image-based table recognition: data, model, and evaluation," 2020. [Online]. Available: https://arxiv.org/abs/1911.10683

[30] D. Prasad, A. Gadpal, K. Kapadni, M. Visave, and K. Sultanpure, "Cascadetabnet: An approach for end to end table detection and structure recognition from image-based documents," 2020.

[31] R. Altenhöner, A. Berger, C. Bracht, P. Klimpel, S. Meyer, A. Neuburger, T. Stäcker, and R. Stein, "DFG-Praxisregeln "Digitalisierung". Aktualisierte Fassung 2022." Feb. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7435724

[32] W. Meier, "exist: An open source native xml database," in *Web, Web-Services, and Database Systems*, A. B. Chaudhri, M. Jeckle, E. Rahm, and R. Unland, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003. doi: https://doi.org/10.1007/3-540-36560-5_13. ISBN 978-3-540-36560-0 pp. 169–183.

[33] e editiones, "Tei publisher," accessed: 2024-07-15. [Online]. Available: https://teipublisher.com

[34] L. O'Gorman and R. Kasturi, *Document Image Analysis*. IEEE Computer Society Press Los Alamitos, 1995, vol. 39.

[35] J. Han, J. Pei, and H. Tong, *Data Mining: Concepts and Techniques*. Morgan kaufmann, 2022.

[36] TEI Consortium, eds., *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, TEI Consortium, 2024, last modified 2024-07-08. [Online]. Available: http://www.tei-c.org/Guidelines/P5/

[37] B. Smock, R. Pesala, and R. Abraham, "PubTables-1M: Towards comprehensive table extraction from unstructured documents," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. doi: https://doi.org/10.48550/arXiv.2110.00061 pp. 4634–4642.