A Dataset for Analyzing Complex Document Layouts in the Digital Humanities and its Evaluation with Krippendorff's Alpha[⋆]

David Tschirschwitz[0000−0001−5344−4172],
Franziska Klemstein[0000−0003−3137−6732],
Benno Stein[0000−0001−9033−2217], and
Volker Rodehorst[0000−0002−4815−0118]

Bauhaus-Universität Weimar, Germany
david.tschirschwitz@uni-weimar.de

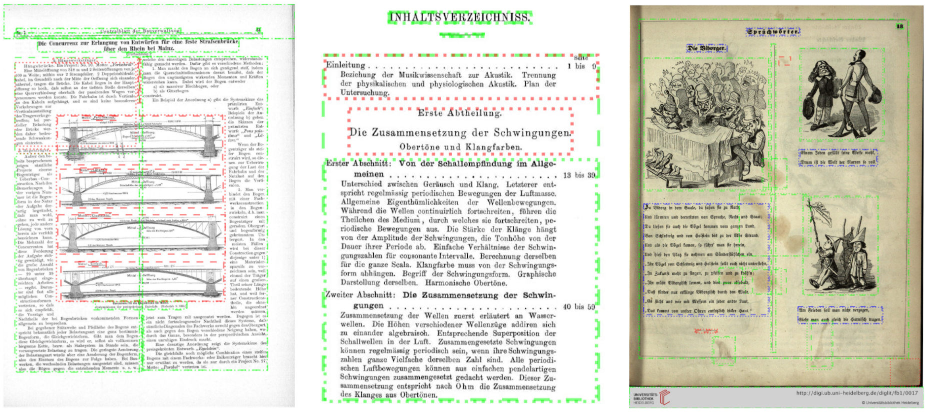
Abstract. We introduce a new research resource in the form of a high-quality, domain-specific dataset for analyzing the document layout of historical documents. The dataset provides an instance segmentation ground truth with 19 classes based on historical layout structures that stem (a) from the publication production process and the respective genres (life sciences, architecture, art, decorative arts, etc.) and, (b) from selected text registers (such as monograph, trade journal, illustrated magazine). Altogether, the dataset contains more than 52,000 instances annotated by experts. A baseline has been tested with the well-known Mask R-CNN and compared to the state-of-the-art model VSR [55]. Inspired by evaluation practices from the field of Natural Language Processing (NLP), we have developed a new method for evaluating annotation consistency. Our method is based on Krippendorff's alpha (K-α), a statistic for quantifying the so-called "inter-annotator-agreement". In particular, we propose an adaptation of K-α that treats annotations as a multipartite graph for assessing the agreement of a variable number of annotators. The method is adjustable with regard to evaluation strictness, and it can be used in 2D or 3D as well as for a variety of tasks such as semantic segmentation, instance segmentation, and 3D point cloud segmentation.

Keywords: Document layout analysis · Digital humanities · Instance segmentation · Inter-annotator-agreement.

1 Introduction

Research in the digital humanities requires experts to interpret large corpora and to be able to search these corpora for specific information. Existing tools

[⋆] This version of the contribution has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1007/978-3-031-16788-1_22. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms.



(a) K-α = 0.33. (b) K-α = 0.51. (c) K-α = 0.92.

Fig. 1: Comparing images with different K-α values. Dashed lines marks annotator A and dotted lines signifies annotator B. The colors indicate matching (green), missed (blue) or class wise disagreed (red) annotations.

and analysis methods in digital humanities focus on digitized text, while visual content in the form of images is often neglected. In "The visual digital turn" by Wevers et al. [51] this bias is explained by the availability of Optical Character Recognition (OCR) and other powerful tools for the analysis of text. However, to get a "holistic understanding" of a document, images have to be included into the analysis as well. To extract structured data, an understanding of the document layout is necessary: Knowing the document layout allows precise extraction of layout elements (e.g., images, equations, editorial notes) and to handle them in a suitable manner depending on the type of information they contain [38]. Moreover, layout information is a useful preprocessing step for image processing and related recognition tasks in a document understanding system [16,13,47].

Several commercial tools do exist to execute document layout analysis, such as Kofax OmniPage [30], Abbyy FineReader [6] or PRImA Aletheia [43,14]. Additionally, large public datasets like PubLayNet [56] or DocBank [35] can be used to train highly-accurate data-driven methods like deep neural networks for document layout analysis. For historical data, datasets like HJDataset [49] or IMPACT [42] are available. While all of these approaches work to some degree for the digital humanities domain, the selection of layout elements in these datasets is often not designed to identify specific artistic elements in the layout (e.g., decorations or frames). This makes it impossible to capture specific historical production processes as well as the importance of the publication as an artistic medium.

In order to address these issues, we created a high-quality dataset for document layout analysis with fine-grained annotations, including instance segmentation level ground truth data (see Figure 1). The dataset intends to serve the community as a resource to further test and enhance methods running on historical document layout analysis. As our benchmarking shows, state-of-the-art models [55] designed for contemporary literature document layout analysis fall short compared to well established generic approaches [23] for instance segmentation. Since the creation process of the dataset had very specific requirements by the domain experts, we also introduce a new method to evaluate annotation consistency. Our method is an adaptation of Krippendorff's alpha [33,22] (K-α) for computer vision, a statistical measurement, which is commonly used in Natural Language Processing (NLP) and other Machine Learning (ML) subfields [39]. Our comprehensive, stage-less and customizable method can be used in multiple scenarios. Such as, the annotation consistency evaluation on a subset of a finalized dataset to describe the quality, on the entire dataset to provide feedback if disagreements between annotators occur which allows filtering or correction of low-quality annotations or the identification of specific hard examples. Our work makes two key contributions:

- A new high-quality historical document layout analysis dataset with over $52,000$ instances and a benchmark on our dataset.
- We also propose a new method for evaluating annotation consistency in computer vision, extending K-α by treating annotations as a multipartite graph.

2 Related Work

State-of-the-art techniques rely on data-driven deep learning techniques. Two general approaches to analyse the layout seem generally feasible. For born-digital documents which contain embedded text that can be used as input data, models like BERT [17] or LayoutLM [54] have been used with great success. On the other hand, models like Faster R-CNN [44] or Mask R-CNN [23] can be utilized using the visual cues. As a hybrid solution models can also use both aspects like the current best performing layout model VSR [55]. Various public datasets exist to train and benchmark these models like PubLayNet [56], DocBank [35] or IMPACT [42]. Using models that rely on text-based ground truth is problematic for historical none-digital-born documents, since they are often missing ground truth texts. OCR methods can be used to obtain the text, however, while these OCR results are often highly accurate, some errors are still occurring and these are passed to models that use text as input data. Manually creating ground truth text data for a historical dataset to use them in addition to the image for training, does not seem viable, since during inference no textual ground truth would be available as well. This could be considered a limiting factor for using models like BERT, LayoutLM or VSR on historical documents.

During the selection of documents that are part of the annotated dataset, a categorization to cover the presumably different types of layouts is necessary. Kise [31] provides a breakdown of the different layout types into several subcategories, namely, rectangular, Manhattan, non-Manhatten and two types of overlapping layouts (further called arbitrary complex [10]). By considering these layout types during data selection of documents to be annotated, a possible wider range of practice relevant cases were covered during training, that then can be beneficial for inference. While these layout types can give an orientation on how challenging different layouts might be, it is not given that learning one of these layout-types allows transfer to other documents of the same layout type. Our dataset contains three of these layout types (rectangular, Manhattan, and non-Manhattan), but while some of the historical sources tend to be hybrids, this creates another complexity. For this reason, we have added further layout components in addition to the typical ones such as text blocks, illustrations and tables.

Evaluating annotation consistency in computer vision is often done with the same metrics that are used for evaluating model performance like mean Average Precision (mAP) or the F_1 score. Many datasets rely entirely on the manual [37] or automated annotation pipeline [35,56] and no extra quality assurance step is taken, while in some cases a small sample of the dataset is double annotated and evaluated with such above mentioned metrics [21]. In the creation process of datasets in other domains (ML and NLP), agreements between annotators are calculated with statistical measurements like K-α [33,22] or Cohen's kappa [15]. Reasons to make the extra effort for these additional annotations are [7]:

- Validating and improving the annotation guideline and existing annotations.
- Identifying difficult cases or ambiguities.
- Assessment of the interpretation of the annotating data within the guideline range.
- Comparing annotator speed and accuracy.

Current approaches [41,46] for evaluating the inter-annotator agreement in computer vision rely on pixel-level comparison, that views the image as a long list of entries and then compares them to use them with the regular procedure of inter-annotator-agreement calculation. While this method can work for balanced datasets in some cases, it neglects criteria like the number of instances or at what point annotators really meant the same entity and did not just by chance overlapped in the same region. To remedy the above-mentioned limitations, we propose an adaptation of K-α, by interpreting annotations as a graph and rely on the Intersection over Union (IoU) in the first step of our calculation step. Thanks to the variety of different IoU versions, our method is flexible and universal applicable in the entire computer vision domain.

3 Evaluation Annotation Consistency

3.1 Data Quality Metric

To calculate K-α a multi-step procedure is applied, which is outlined in Figure 2. As a result of this calculation a single α value is reported which measures the inter-annotator-agreement, where $\alpha = 1$ is a perfect agreement, $\alpha = 0$ means there is no agreement and values < 0 indicate a disagreement. The general form of K-α is $\alpha = 1 - \frac{D_o}{D_e}$, where D_o is the observed disagreement and D_e is the expected disagreement, going further we will explain the calculation of α for nominal data.

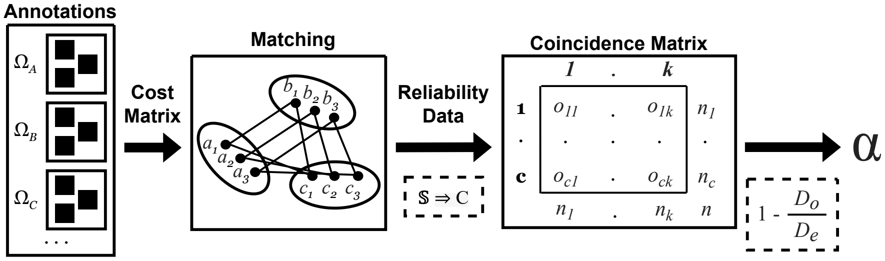


Fig. 2: Outline of the calculation process.

Assuming there are Ω annotators, where the first annotator is Ω_A, with a set of annotations $a = \{a_i\}_{i=1}^{N}$ and the second annotator is Ω_B with its own set of annotations $b = \{b_i\}_{j=1}^{M}$. Each such annotation a_i and b_i contains an arbitrary shape $\mathbb{S} \subseteq \mathbb{R}^n$ that is defined as A or B respectively. These shapes can be either in 2D or 3D space and the IoU is calculated with the following equation:

$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

In order to match all the entries of Ω_A and Ω_B, a cost matrix is created with the following cost function:

$$C(i, j) = 1 - IoU(a_i, b_j) \tag{2}$$

The two sets are viewed as a bipartite graph and can therefore be matched via the Hungarian algorithm. This requires $N = M$, for which the smaller set is padded with $\varnothing$ so that $N = M$. Finding the permutation matrix X is done by optimizing $argmin \sum_{i=1}^{N} \sum_{j=1}^{M} C_{i,j} X_{i,j}$. While an optimal solution exists for a bipartite graph ($|\Omega| = 2$), a multipartite graph, which exists if $|\Omega| > 2$, an optimal solution cannot necessarily be found since the problem is considered APX-complete [27,8]. However, for most cases $N = M$ is sufficiently small, so it can be assumed that an optimal solution can be found with a simple greedy matching between $|\Omega| > 2$. After matching, the obtained reliability data can be organized in a matrix. Instead of the shapes $\mathbb{S}$ now the classes contained in each annotation are used.

From the reliability data a coincidence matrix is calculated which contains the values $o_{ck} = \sum_u \frac{\text{Number of c-k pairs in unit u}}{m_u - 1}$, where m_u is the number of observers m for unit u. Further, we can calculate $n_c = \sum_k o_{ck}$ and $n = \sum_c n_c$. This allows calculation of α for nominal data using the following equation:

$$\alpha = 1 - \frac{D_o}{D_e} = \frac{(n-1)\sum_c o_{cc} - \sum_c n_c(n_c - 1)}{n(n-1) - \sum_c n_c(n_c - 1)} \tag{3}$$

To discourage annotators from missing entries, the $\varnothing$ is replaced by a filler class instead of a ”cannot code” or ”no data available” entry like in the canonical Krippendorff alpha version. This leads to worse agreement scores if one annotator missed an entry that others found. A calculation example is attached in Appendix 1. Our code for calculating K-α on COCO[1] formatted data can be found here, https://github.com/Madave94/KrippendorffAlphaComputerVision.

3.2 Method Properties

No Pipeline. Human annotated datasets often rely on multi-stage annotation pipelines [21,37]. For example, the COCO dataset uses a pipeline where the first three steps are category labeling, instance spotting and instance segmentation. This requires to a certain degree that the first stage (category labeling) is finished before the second stage (instance spotting) can start. This however, is not required when using the here presented adaptation of K-α. Since there are no stages in the annotation process, annotators can annotate independent from each other and annotations can be extended even after the initial dataset was finished (e.g., active learning). The process can only be considered stage-less within each annotation iteration and after creation of a final guideline.

Cherry-Picking Annotations. Besides determining the quality of annotations K-α allows the evaluation of rater vitality [39]. This evaluation is a way of measuring how well a single annotator compares to the entire annotator group. It is defined as:

$$v_i = \alpha_{\Omega} - \alpha_{\Omega \setminus \Omega_i} \tag{4}$$

In case the annotator performs better than the rest of the group, the value will be positive, while the value is negative if the annotator performs worse than the group. With the help of this measurement, annotators with inconsistent quality can be excluded from the annotation process or the final dataset.

Shape and Dimension Independence. Since the IoU is used as an evaluation metric, the entire annotation consistency evaluation method is dimension independent and can be used in 2D or 3D space. Furthermore, it can be used to evaluate annotations created to solve different tasks like semantic segmentation, object detection or instance segmentation. It is also easy to adapt for specific use cases such as action detection where the Generalized Intersection over Union (GIoU) [45] could be used.

Customizability. Different task, might have more-or-less strict annotation requirements or allow more ambiguity. In order to adapt which annotations are accepted, the α threshold can be increased, which would require a higher agreement between annotators for an image to be accepted. Further, if a larger overlap

[1] Format definition: https://cocodataset.org/#format-data

between annotations is necessary the IoU threshold could be increased, so that only more overlapping areas are considered to be matching. To allow more ambiguity, K-α can be used in the canonical form that does not penalize missing data as much by including a default value that provides no reliability data instead of a filler class as described above. Lastly, adoptions could be made by including the classes during calculation of the cost function [12] and allow a more ”trustful” matching. Some adaptation examples can be found in Appendix 1.

Precise Guideline Requirement. A challenging part of any annotation process is the exact formulation of the annotation guideline. This process gets more difficult for more complex data. There are domains like medicine or civil engineering that require expert knowledge to formulate and annotate the data. While for simpler classes a pipeline approach might be more accessible, since for each stage only a single step needs to be finished, for more complex tasks, the context often matters and only a more holistic approach will yield success. Our method is more suited for complex tasks, where during the guideline definition, domain experts and data scientists are in an iterative process of further and further refining the exact annotation guideline.

Additional Annotations. Manual annotated data are expensive to obtain. On the other hand, it is also necessary to ensure sufficient quality of the ground truth data. Other datasets would often rely entirely on their annotation pipeline, either manual [37] or automated [56,35] without further quality evaluation. A better approach is to take a sample of the dataset by double annotating a subset [21] to evaluate the quality of the dataset. By using the here presented method, it is possible to evaluate such an annotation subset, however we see the application on more complex data that might contain some ambiguity or requires very strict quality control. An open question regarding these multi-per-image available annotations would be, if anything useful can be yielded by using them for training. A possible approach would be to use the annotations with some kind of a probability map similar to distillation learning [26].

4 TexBiG Dataset

4.1 Dataset Design

Due to the numerous digitization projects in the past decades, fundamental archival holdings and (historical) publications have been transferred into digital collections, which are thus highly relevant for research in the humanities. Although numerous digital collections from different disciplines are available, these have so far often only been the starting point for the analysis of one specific domain or just a few layout classes, e.g. images of newspapers [51] or headlines and visual content [34]. The analysis of the complexity of the entire layout of an investigation domain as well as the comparison between different domains has been missing so far.

The aim of creating the dataset is to be able to analyze the intersections and differences between various domains of investigation in the period from 1880 to 1930 concerning their layout and their respective text-image relation. The starting point for this analysis is the Virtual Laboratory [1], which is a collection of sources on the history of experimental life sciences for the period 1830 to about 1930. As part of the dataset creation process, this corpus of texts on the history of nature and science was expanded to include artistic, applied arts, and humorous-satirical text sources.

We made this selection because a search for new values can be observed in various areas of life and society in the late 19th century. This search is reflected, among other things, in an increased interest of artists in technology, industry and life sciences [18]. This interest culminated at the beginning of the 20th century in various social reform movements around the world, with a great impact on developments in technology, science, also politics [9,40]. The sources we have selected are exemplary for their respective research domains. They were selected concerning their significance and relevance for the domain in the time frame investigated.

For the analysis of genre- and media-specific layout structures of historical documents, instance segmentation is necessary because it recognizes objects in images along with their associated shape, as well as very fine structures. This is of particular importance to us because layouts change significantly over time. For example, there are frames and decorations on the pages or even drawings that need to be recognized in a shape-specific way, partly because they vary greatly between the domains of investigation, but also because they converge over time. The need for this is evident in the artistic-experimental works, which are characterized by a high degree of innovative layout design. Particularly in the course of the 1920s, there is an effort to break up the two-dimensionality of the book by disrupting clear layout structures [32]. In addition, new pictorial elements, such as the symbol of the arrow, are invented, which then develop out of the artistic-experimental domain and only a little later find application in other domains, especially in sciences, as well [20].

The documents we selected come from different genres and domains with a range of production processes and vary in page count (see: Table 1, for further information on the sources, see: Appendix 2).

Table 1: Document selection and layout type

Name	Layout	Pages	Domain	Year
”Pädagogisches Skizzenbuch”[29]	non-Manhatten	61	Art	1925
”Zeitschrift für Physiologie und Psychologie der Sinnesorgane”[4]	Manhatten	493	Life Sciences	1907
”Lehre von den Tonempfindungen”[25]	Manhatten	658	Life Sciences	1863
”Das Kunstgewerbe”[5]	mixed type	454	Applied Arts	1890–1892
”Fliegende Blätter”[2]	mixed type	196	Satiric	1844–1845
”Centralblatt der Bauverwaltung”[3]	mixed type	395	Architecture	1881

4.2 Dataset Construction

Creation of the dataset was done by a selected group of annotators, all with extended knowledge of the application domain. An initial guideline was developed by the organizers and presented to these annotators. After creation of the guideline each iteration of the annotation process can be considered stage-less, since each iteration does only require a single processing step by each annotator. According to the guideline, annotators started ground truthing the dataset. On the way to the final dataset, at multiple points, the preliminary dataset was evaluated on K-α and trained on intermediate models. These evaluations and results of the model served as an orientation to further modify and refine the annotation guideline. Annotators that got assigned a set of documents, did work the entire annotation process on these annotations, no reassignments were made in between iterations.

After all pages have been annotated, multiple correction iterations were done. In these correction iterations annotators would only receive information about which pages had a low agreement or disagreement, we set K-$\alpha < 0.8$ as the threshold for pages to be reviewed. During the last iterations of the annotation process we decided to apply a stricter criterion to find overlooked annotations, which is easily adjustable with our K-α method.

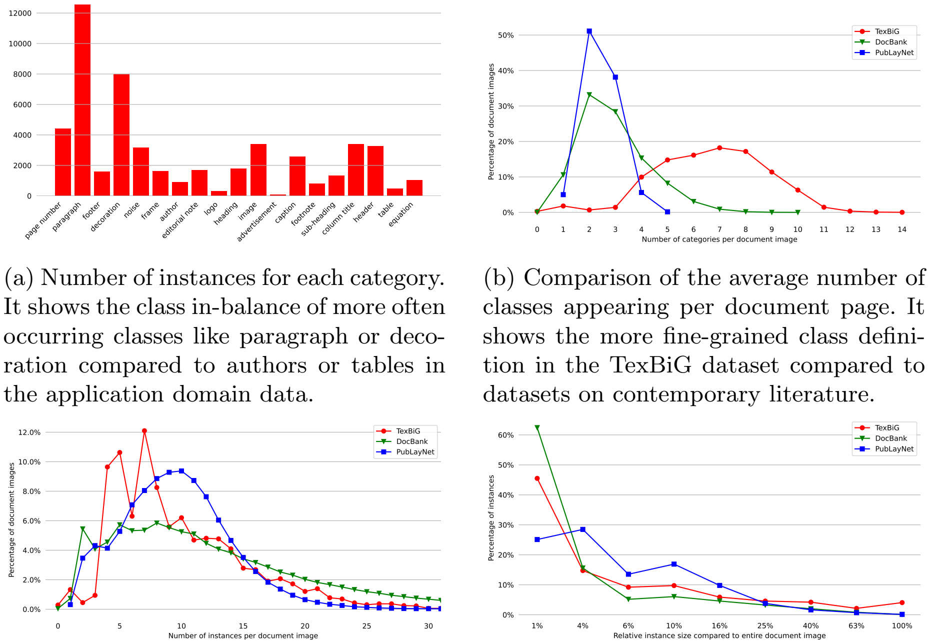
A principal that we applied to our annotation process, is that annotators would generally not be allowed to view the annotations of their cross-annotators (annotators assigned to the same document image). However, for guideline re-evaluation purpose and solving of possible annotation ambiguities, selected annotations have been reviewed by the entire annotator group. We assume that it is more aspirational that annotators only rely on the guideline as their reference and do not discuss unclear cases with each other. In the best case this leads to a more and more refined guideline. Hence, the designed annotation guideline can later be reused for dataset extension.

For the annotation process, we used the computer vision annotation tool (CVAT) [48]. Besides our annotations we will also publish the annotation guideline as a supplement. The dataset and guideline can be found at https://doi.org/10.5281/zenodo.6885144.

4.3 Dataset Analysis

Figure 3 shows several comparisons with PubLayNet [56] and DocBank [35] to get an idea about characteristics and properties of our dataset.

Unsurprisingly, there is an unbalance in the dataset as shown in Figure 3a as it is often the case in document analysis. This is further increased by research relevant classes like advertisements, logos, or frames. Due to the larger number of classes, the number of categories found per page as shown in Figure 3b is also significantly larger than in other datasets. Reasons for this are due to; many classes that are "auxiliary" to the main text body like headers, footers and decorations. On the other hand, the total number of instances per document page appears slightly lower than in PubLayNet and a bit higher than in DocBank



(a) Number of instances for each category. It shows the class in-balance of more often occurring classes like paragraph or decoration compared to authors or tables in the application domain data.

(b) Comparison of the average number of classes appearing per document page. It shows the more fine-grained class definition in the TexBiG dataset compared to datasets on contemporary literature.

(c) The number of instances per document page shows that some classes that are split into multiple instances are combined in the TexBiG dataset, according to the requirements of the domain experts.

(d) Relative size of instances compared to the document page size. It shows that the TexBiG dataset contains a higher number of larger classes compared to other datasets.

Fig. 3: Dataset statistics.

as depicted in Figure 3c. Lastly, TexBiG contains more large regions $> 40\%$ of the document page compared to the benchmarked datasets, these details can be found in Figure 3d. In Table 2 the split of the dataset is shown. A split 70-15-15 for train, validation and test respectively subset was chosen. Some prior information is included in the split since data are divided according to the different layout types so that each layout type is roughly represented the same in each subset.

4.4 Dataset Quality Evaluation

In this section the method previously explained is applied to evaluate the quality of the dataset. While technically values of K-α could be between -1 and +1, our dataset contains exactly one data point with a value of 0 and none below 0. While we think that someone might only want to use a subset of high-quality data for training and testing purposes, we release the entire dataset and leave the decision to researchers using our dataset to opt by themselves on whichever

Table 2: Semantic structure statistic of training, validation and test sets in TexBiG.

Split	Train		Validation		Test		All	
Advertisement	61	0.17%	10	0.12%	18	0.23%	89	0.17%
Author	623	1.72%	142	1.72%	142	1.81%	907	1.73%
Caption	1816	5.01%	408	4.94%	360	4.58%	2584	4.93%
Column title	2341	6.45%	504	6.11%	547	6.97%	3392	6.48%
Decoration	5413	14.93%	1335	16.18%	1247	15.88%	7995	15.27%
Editorial note	1206	3.33%	265	3.21%	221	2.81%	1692	3.23%
Equation	764	2.11%	163	1.98%	119	1.52%	1046	2.0%
Footer	1093	3.01%	244	2.96%	244	3.11%	1581	3.02%
Footnote	544	1.5%	125	1.51%	120	1.53%	789	1.51%
Frame	1132	3.12%	237	2.87%	265	3.37%	1634	3.12%
Header	2267	6.25%	479	5.81%	507	6.46%	3253	6.21%
Heading	1224	3.37%	334	4.05%	234	2.98%	1792	3.42%
Image	2339	6.45%	536	6.5%	508	6.47%	3383	6.46%
Logo	218	0.6%	46	0.56%	34	0.43%	298	0.57%
Noise	2225	6.13%	466	5.65%	471	6.0%	3162	6.04%
Page number	3072	8.47%	655	7.94%	689	8.77%	4416	8.43%
Paragraph	8633	23.8%	2030	24.6%	1887	24.03%	12550	23.96%
Sub-heading	955	2.63%	207	2.51%	168	2.14%	1330	2.54%
Table	342	0.94%	65	0.79%	72	0.92%	479	0.91%
Total	36268	69.24%	8251	15.75%	7853	14.99%	52372	100.0%

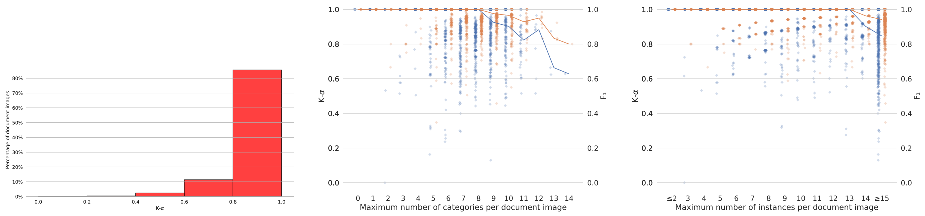
subset they use. In Figure 4, a quality comparison of the data is shown, starting with a sorting into different quality regiments in subfigure 4a. The two Figures 4b and 4c show a comparison between K-α and the F_1 score, both using an IoU of 0.5. It illustrates two things. First, with more document images containing more instances and classes the agreement decreases, which implies that the data gets more complex. Comparing α with F_1 shows that α is a more critical metric since the values are lower and more spread along the y axis, hence allowing a more differentiated evaluation. Three example images in Figure 1 show how document images look if they contain a very high, high or low agreement. For visualization purpose we only used document images with a maximum of two annotators.

4.5 Benchmarking

As our baseline approach we use Mask R-CNN [23] with the Detectron2 [52] framework. After some hyper-parameter tuning, we found that the model worked best on our dataset with a ResNet50 [24] backbone, a feature pyramid network with 3 × 3 convolutions (FPN) [36] and weights from a model trained on PubLayNet[2]. The results can be found in Table 3 and the model is compared with the current state-of-the-art model VSR [55] for document layout analysis.

While VSR showed significant improvements on DocBank or PubLayNet, the results on our dataset are not as impressive. We tried a multitude of configurations to improve our results including changes to the ResNeXt-101 [53]

[2] Weights: https://github.com/hpanwar08/detectron2



(a) Overall quality distribution.

(b) Quality distribution by number of different classes per page.

(c) Quality distribution by number of different instances per page.

Fig. 4: Dataset quality evaluation. In Figure 4b and 4c the F_1 score (orange) is compared with K-α (blue), with the line showing the median.

Table 3: Benchmark results on TexBiG comparing Mask R-CNN and VSR for object detection and instance segmentation. Extended version in Appendix 3.

Task	Bounding Box		Mask	
Model	Mask R-CNN	VSR	Mask R-CNN	VSR
mAP	73.18	75.90	65.43	65.80
mAP_{50}	85.12	87.80	80.31	82.60

backbone to a ResNeXt-50 backbone, switching BERT's [17] weights from the English version[3] to a German version[4] and using both pre-trained models the authors provide. Even after an extended hyper-parameter search, we could only exceed the results that Mask R-CNN achieves slightly. We assume that there are multiple underlying issues, related to different properties of historical document layout analysis data.

Firstly, there is no existing ground truth text, we used Tesseract [28] to generate the text that was then used by the model to analyze the layout. We have to assume that the OCR is not perfect, hence the errors then propagate into the network and cause inconsistencies in the training data that do not exist for the plain layout data. This is an issue not specific to our dataset since none-born-digital documents do not have a perfect textual representation, which means that models using this kind of information need to deal with inherently noisy or slightly faulty data. Furthermore, the pre-trained models are trained on English data, while our data is mainly in German, we, therefore, could assume that the semantic backbone branch of VSR is not as applicable with the German word embeddings as with the English ones. This would mean that Mask R-CNN has a significant advantage since it has a reasonable weight initialization from transfer learning. The code to our VSR implementation can be found at https://github.com/Madave94/VSR-TexBiG-Dataset.

[3] Weights: https://huggingface.co/bert-base-uncased
[4] Weights: https://huggingface.co/dbmdz/bert-base-german-uncased

5 Conclusion

We have introduced a stage-less evaluation method to create high-quality annotations for data. The method can be generally applied in computer vision including 2D and 3D data, covering a variety of tasks. The method was developed during the creation of a domain-specific dataset (which will be made publicly available) that provides the community with a research resource that can be used for the development and design of new architectures and methods. When the dataset was benchmarked on a state-of-the-art model, it has shown that there is a gap between models, which were developed on contemporary layout analysis datasets and their application to historical documents. Lastly, with the current state of the dataset, models trained on it can already become a research tool that can be useful for researchers in digital humanities.

Acknowledgments. This work was supported by the Thuringian Ministry for Economy, Science and Digital Society / Thüringer Aufbaubank (TMWWDG / TAB). In addition, the following persons should be mentioned, without whom the project would not have been feasible: Sascha Breithaupt, Johannes Hess, Henrik Leisdon, Josephine Tiede and Ina Tscherner. Lastly, we would like to thank Christian Benz and Jan Frederick Eick for their in-depth discussion and feedback and in particular Henning Schmidgen.

References

1. The Virtual Laboratory, https://vlp-new.ur.de/
2. Fliegende Blätter (1845–1944), https://nbn-resolving.org/urn:nbn:de:bsz:16-diglit-35697
3. Centralblatt der Bauverwaltung (1881–1931), https://digital.zlb.de/viewer/image/14688302_1881/1/
4. Zeitschrift für Psychologie und Physiologie der Sinnesorgane (1890-1909), https://ia804503.us.archive.org/25/items/bub_gb_2dIbAAAAMAAJ/bub_gb_2dIbAAAAMAAJ.pdf
5. Das Kunstgewerbe (1890–1895). https://doi.org/10.11588/diglit.18553, http://kunstgewerbe.uni-hd.de
6. ABBYY Development Inc.: ABBYY FineReader PDF 15, https://pdf.abbyy.com/de/finereader-pdf/
7. Artstein, R.: Inter-annotator agreement. In: Handbook of linguistic annotation, pp. 297–313. Springer (2017)
8. Ausiello, G., Crescenzi, P., Gambosi, G., Kann, V., Marchetti-Spaccamela, A., Protasi, M.: Complexity and approximation: Combinatorial optimization problems and their approximability properties. Springer Science & Business Media (2012)
9. Baumgartner, J. (ed.): Aufbrüche - Seitenpfade - Abwege : Suchbewegungen und Subkulturen im 20. Jahrhundert; Festschrift für Ulrich Linse. Königshausen & Neumann, Würzburg (2004)
10. Binmakhashen, G.M., Mahmoud, S.A.: Document Layout Analysis: A Comprehensive Survey. ACM Computing Surveys **52**(6), 1–36 (Jan 2020). https://doi.org/10.1145/3355610, https://dl.acm.org/doi/10.1145/3355610

11. Bruening, U.: Bauhausbücher. Grafische Synthese - synthetische Grafik. Neue Bauhausbücher pp. 281–296 (2009)
12. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision (ECCV). pp. 213–229. Springer (2020)
13. Clausner, C., Antonacopoulos, A., Pletschacher, S.: ICDAR2019 Competition on Recognition of Documents with Complex Layouts – RDCL2019 p. 6
14. Clausner, C., Pletschacher, S., Antonacopoulos, A.: Aletheia - an advanced document layout and text ground-truthing system for production environments. In: 2011 International Conference on Document Analysis and Recognition. pp. 48–52 (2011). https://doi.org/10.1109/ICDAR.2011.19
15. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
16. Dengel, A., Shafait, F.: Analysis of the logical layout of documents. In: Doermann, D., Tombre, K. (eds.) Handbook of Document Image Processing and Recognition, chap. 6. Springer London, London (2014). https://doi.org/10.1007/978-0-85729-859-1, http://link.springer.com/10.1007/978-0-85729-859-1
17. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
18. Flach, S., Weigel, S. (eds.): WissensKünste : das Wissen der Künste und die Kunst des Wissens = the knowledge of the arts and the art of knowledge. VDG, Weimar (2011), http://www.gbv.de/dms/weimar/toc/64247172X_toc.pdf
19. Froschauer, E.M.: "An die Leser!": Baukunst darstellen und vermitteln ; Berliner Architekturzeitschriften um 1900. Wasmuth, Tübingen (2009)
20. Giedion, S.: Mechanization takes command a contribution to anonymous history. University of Minnesota (1948)
21. Gupta, A., Dollar, P., Girshick, R.: LVIS: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
22. Hayes, A.F., Krippendorff, K.: Answering the call for a standard reliability measure for coding data. Communication methods and measures **1**(1), 77–89 (2007)
23. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2961–2969 (2017)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
25. Helmholtz, H.v.: Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik. Braunschweig: F. Vieweg (1863), https://vlp-new.ur.de/records/lit3483
26. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 **2**(7) (2015)
27. Kann, V.: Maximum bounded 3-dimensional matching is max snp-complete. Information Processing Letters **37**(1), 27–35 (1991)
28. Kay, A.: Tesseract: an open-source optical character recognition engine. Linux Journal **2007**(159), 2 (2007)
29. Klee, P.: Pädagogisches Skizzenbuch. Bauhausbücher ; 2, Langen, München, 2. aufl. edn. (1925). https://doi.org/10.11588/diglit.26771, http://digi.ub.uni-heidelberg.de/diglit/klee1925
30. Kofax, Inc.: OmniPage Ultimate, https://www.kofax.de/products/omnipage

31. Koichi, K.: Page segmentation techniques indocument analysis. In: Doermann, D., Tombre, K. (eds.) Handbook of Document Image Processing and Recognition, chap. 5. Springer London, London (2014). https://doi.org/10.1007/978-0-85729-859-1, http://link.springer.com/10.1007/978-0-85729-859-1
32. Krauthausen, K.: Paul Valéry and geometry : instrument, writing model, practice. Preprint / Max-Planck-Institut für Wissenschaftsgeschichte 406, Max-Planck-Inst. für Wissenschaftsgeschichte, Berlin (2010)
33. Krippendorff, K.: Computing krippendorff's alpha-reliability (2011), https://repository.upenn.edu/asc_papers/43
34. Lee, B.C.G., Mears, J., Jakeway, E., Ferriter, M., Adams, C., Yarasavage, N., Thomas, D., Zwaard, K., Weld, D.S.: The newspaper navigator dataset: extracting headlines and visual content from 16 million historic newspaper pages in chronicling america. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. pp. 3055–3062 (2020)
35. Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., Zhou, M.: Docbank: A benchmark dataset for document layout analysis. arXiv preprint arXiv:2006.01038 (2020)
36. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2117–2125 (2017)
37. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV). pp. 740–755. Springer (2014)
38. Marinai, S.: Introduction to document analysis and recognition. In: Machine learning in document analysis and recognition, pp. 1–20. Springer (2008)
39. McCulloh, I., Burck, J., Behling, J., Burks, M., Parker, J.: Leadership of data annotation teams. In: 2018 International Workshop on Social Sensing (SocialSens). pp. 26–31 (2018). https://doi.org/10.1109/SocialSens.2018.00018
40. McLoughlin, W.G.: Revivals, awakening and reform. University of Chicago Press (1978)
41. Nassar, J., Pavon-Harr, V., Bosch, M., McCulloh, I.: Assessing data quality of annotations with krippendorff alpha for applications in computer vision. arXiv preprint arXiv:1912.10107 (2019)
42. Papadopoulos, C., Pletschacher, S., Clausner, C., Antonacopoulos, A.: The IMPACT dataset of historical document images. In: Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing - HIP '13. p. 123. ACM Press, Washington, District of Columbia (2013). https://doi.org/10.1145/2501115.2501130, http://dl.acm.org/citation.cfm?doid=2501115.2501130
43. Pattern Recognition & Image Analysis Research Lab: Aletheia document analysis system, https://www.primaresearch.org/tools/Aletheia
44. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:1506.01497 [cs] (Jan 2016), http://arxiv.org/abs/1506.01497, arXiv: 1506.01497
45. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
46. Ribeiro, V., Avila, S., Valle, E.: Handling inter-annotator agreement for automated skin lesion segmentation. arXiv preprint arXiv:1906.02415 (2019)
47. Richarz, J., Fink, G.A., et al.: Towards semi-supervised transcription of handwritten historical weather reports. In: 2012 10th IAPR International Workshop on Document Analysis Systems. pp. 180–184. IEEE (2012)

48. Sekachev, B., Manovich, N., Zhiltsov, M., Zhavoronkov, A., Kalinin, D., Hoff, B., TOsmanov, Kruchinin, D., Zankevich, A., DmitriySidnev, Markelov, M., Johannes222, Chenuet, M., a andre, telenachos, Melnikov, A., Kim, J., Ilouz, L., Glazov, N., Priya4607, Tehrani, R., Jeong, S., Skubriev, V., Yonekura, S., vugia truong, zliang7, lizhming, Truong, T.: opencv/cvat: v1.1.0 (Aug 2020). https://doi.org/10.5281/zenodo.4009388, https://doi.org/10.5281/zenodo.4009388
49. Shen, Z., Zhang, K., Dell, M.: A Large Dataset of Historical Japanese Documents with Complex Layouts. arXiv:2004.08686 [cs] (Apr 2020), http://arxiv.org/abs/2004.08686, arXiv: 2004.08686
50. Stielau, A.: Kunst und Künstler im Blickfeld der satirischen Zeitschriften ’Fliegende Blätter’ und ’Punch’. Aachen University (1976)
51. Wevers, M., Smits, T.: The visual digital turn: Using neural networks to study historical images. Digital Scholarship in the Humanities (Jan 2019). https://doi.org/10.1093/llc/fqy085, https://academic.oup.com/dsh/advance-article/doi/10.1093/llc/fqy085/5296356
52. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)
53. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1492–1500 (2017)
54. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: Pre-training of Text and Layout for Document Image Understanding. arXiv:1912.13318 [cs] (Jun 2020). https://doi.org/10.1145/3394486.3403172, http://arxiv.org/abs/1912.13318, arXiv: 1912.13318
55. Zhang, P., Li, C., Qiao, L., Cheng, Z., Pu, S., Niu, Y., Wu, F.: Vsr: A unified framework for document layout analysis combining vision, semantics and relations. In: International Conference on Document Analysis and Recognition (ICDAR). pp. 115–130. Springer (2021)
56. Zhong, X., Tang, J., Yepes, A.J.: Publaynet: largest dataset ever for document layout analysis. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1015–1022. IEEE (2019)

Appendix 1

In this part some example calculations for K-α are illustrated. It shows different forms of how the metric can be adapted. As a helpful resource for further understanding we recommend reading [33]. There are two aspects that we assume will affect most adaptation possibilities: the creation of the cost matrix, which would change the creation of the reliability data and the handling of missing data when calculating K-α.

Figure 5 is used as an example and the resulting reliability matrix after the graph matching would look as shown in the matrix below. For this matching a IoU threshold of 0.5 is used and the normal matching is used.

	unit 1	unit 2	unit 3	unit 4	unit 5
Ω_A	a_1	a_2	a_3	a_4	$\varnothing$
Ω_B	b_1	b_2	b_3	b_4	$\varnothing$
Ω_C	$\varnothing$	c_2	c_3	c_4	c_1

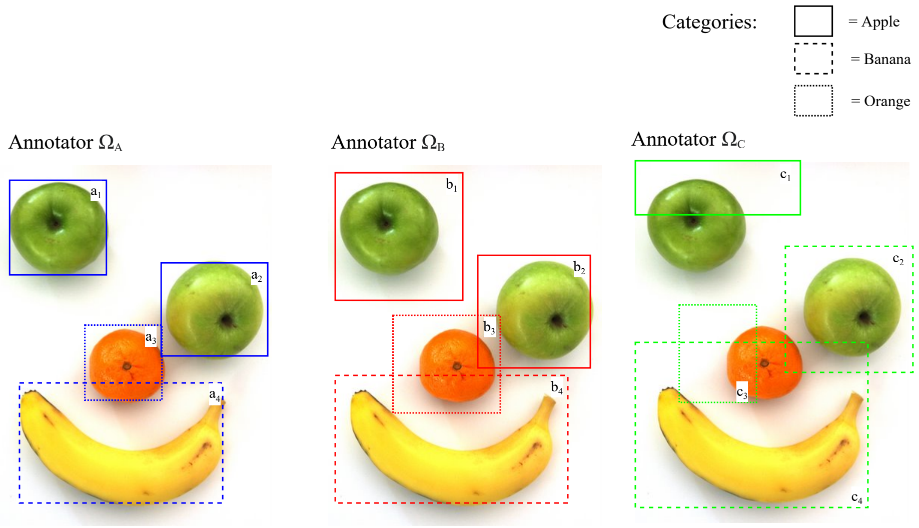


Fig. 5: Three annotator visual examples of matching with $IoU > 0.5$ threshold. Image source: https://commons.wikimedia.org/wiki/File:Smile_at_a_stranger.jpg

Strict matching, don't allow missing data

After the reliability data is calculated all values a_n, b_m and c_p are replaced with their classes and for this case which doesn't allow missing data $\varnothing$ is replaced with the filler class 0. Hence, the matrix would look as follow:

	unit 1	unit 2	unit 3	unit 4	unit 5
Ω_A	1	1	3	2	0
Ω_B	1	1	3	2	0
Ω_C	0	2	3	2	1

From here K-α is calculated the regular way, by first creating the coincidence matrix. Unit 1 contains $3(3-1)=6$ pairs, 2 matching **1-1** pairs, 2 mismatching **1-0** pairs and 2 mismatching **0-1** pairs, it contributes $\frac{2}{3-1}=1$ to the $o_{1,1}$ cell, $\frac{2}{3-1}=1$ to the $o_{1,0}$ cell and $\frac{2}{3-1}=1$ to $o_{0,1}$ cell. Unit 2 contains $\frac{3}{3-1}=6$ pairs, 2 matching **1-1** pairs, 2 mismatching **1-2** pairs and 2 mismatching **2-1** pairs, it contributes $\frac{2}{3-1}=1$ to the $o_{1,1}$ cell, $\frac{2}{3-1}=1$ to the $o_{1,2}$ cell and $\frac{2}{3-1}=1$ to the $o_{2,1}$ cell. Unit 3 contains $\frac{3}{3-1}=6$ pairs, 6 matching **3-3** pairs, $\frac{6}{3-1}=3$ to the $o_{3,3}$ cell. Unit 4 contains $\frac{3}{3-1}=6$ pairs, 6 matching **2-2** pairs, it contributes $\frac{6}{3-1}=3$ to the $o_{2,2}$ cell. Unit 5 contains $\frac{3}{3-1}=6$ pairs, 2 matching **0-0** pairs, 2 mismatching **0-1** pairs and 2 mismatching **1-0** pairs, it contributes $\frac{2}{3-1}=1$ to the $o_{0,0}$ cell, $\frac{2}{3-1}=1$ to the $o_{0,1}$ cell and $\frac{2}{3-1}=1$ to the $o_{1,0}$ cell. As an example the first value in the coincidence matrix $o_{0,0}$ is the sum of all value in the five units related to $o_{0,0}$, which is rather straight forward since only unit 5 contains **0-0** pairs, hence $o_{0,0}=1$. The coincidence matrix is as follows:

	0	1	2	3	
0	1	2	0	0	3
1	2	2	1	0	5
2	0	1	3	0	4
3	0	0	0	3	3
	3	5	4	3	15

Computing K-α is now done via equation 3, which means for our example:

$$\alpha = \frac{(15-1)(1+2+3+3)-[3(3-1)+5(5-1)+4(4-1)+3(3-1)]}{15(15-1)-[3(3-1)+5(5-1)+4(4-1)+3(3-1)]} = 0.49$$

Strict matching, but allow missing data

A second possible version build on the same example shown in Figure 5 that allows missing data, would transfer the reliability data slightly different. Instead

	unit 1	unit 2	unit 3	unit 4	unit 5
Ω_A	1	1	3	2	*
Ω_B	1	1	3	2	*
Ω_C	*	2	3	2	1

of 0 a * will be used indicating missing data, which won't be included in the calculation of α.

Calculating the coincidence matrix would be done in the same way as before for Unit 2, Unit 3 and Unit 4, but Unit 1 and Unit 5 are different. Unit 1 contains $2(2-1) = 2$ pairs, which are 2 matching **1-1** pairs it contributes $\frac{2}{2-1} = 2$ to the $o_{1,1}$ cell. Since Unit 4 only contains a single entry, no pairable unit can be found. The coincidence matrix would therefore be:

	1	2	3	
1	3	1	0	4
2	1	3	0	4
3	0	0	3	3
	4	4	3	11

This results in a calculation of alpha with the following values:

$$\alpha = \frac{(11-1)(3+3+3) - [4(4-1) + 4(4-1) + 3(3-1)]}{11(11-1) - [4(4-1) + 4(4-1) + 3(3-1)]} = 0.75$$

Appendix 2

Further information to the historical sources:

The selected documents were already available as digitized sources. They all come from publicly accessible digital collections. These are: the digital collections of University Library Heidelberg ("Pädagogisches Skizzenbuch" [29], "Das Kunstgewerbe" [5] and "Fliegende Blätter"[2]), the Internet Archive ("Zeitschrift für Physiologie und Psychologie der Sinnesorgane" [4]), the Virtual Laboratory ("Lehre von den Tonempfindungen" [25]) and the digital collections of the Berlin State Library ("Centralblatt der Bauverwaltung" [3]).

The Pedagogical Sketchbook by Paul Klee is part of the artistic-experimental domain. It is the second volume in the Bauhaus book series. The so-called Bauhaus books are a series of books published from 1925 to 1930 by Walter Gropius and Lazlo Moholy-Nagy. Although the books appeared as a series in the same publishing house (Albert Langen Verlag), the respective layout varied widely [11]. The publication sequence was also irregular. While in 1925 alone eight publications of the series could be published, in 1926 there were only two and in 1927, 1928, 1929 and 1930 one more volume each. The publication of Klee presented not only his artwork but also presented his art theoretical knowledge. At the same time, it presented aspects of his extensive lectures on visual form at the Bauhaus and conveyed his way of thinking and working on this topic.

Both the journal "Physiology und Psychologie der Sinnesorgane" and Hermann von Helmholtz's publication "Lehre von den Tonempfindungen" are part of the domain of life sciences. The different types of publications (journal and monograph) have different but typical layout components within their domain, which is why they were both integrated into the dataset.

The journal "Das Kunstgewerbe" appeared every fourteen days from 1890 to 1895 and belongs to the domain of applied arts. The individual issues had a length of 10 pages and a manageable number of illustrations, but the pages were often designed with decorative frames and ornaments.

The illustrated magazine "Fliegende Blätter" appeared from 1844 to 1944, at first irregularly several times a month, later regularly once a week. The humorous-satirical publication was richly illustrated and held in high esteem among the German bourgeoisie. At the same time, the "Fliegende Blätter" are significant both artistically and in terms of printing, due to the high quality of its layout [50].

The "Centralblatt der Bauverwaltung" was a professional journal intended to satisfy the need for information in the construction sector. The journal was first published in April 1881 by the publishing house Ernst & Sohn, in 1931 it was merged with the "Zeitschrift für Bauwesen", in 1944 the publication was discontinued. The Ministry of Public Works acted as publisher until 1919, and from 1920 to 1931 the Prussian Ministry of Finance. The journal was to serve as a supplement to the existing trade journals and, in contrast to these, was to have a faster publication schedule. Information about construction projects and competitions, projects currently being implemented, new technologies and amended legal framework conditions were to reach the readership more quickly than before and also address international developments. At the same time, however, the journal was to be less elaborately designed than the existing trade organs and art journals. Although the Ministry of Public Works was the editor and the structure of the journal was divided into ”official” and ”non-official” parts, it can nevertheless not be characterized as a purely official journal of authorities [19].

Appendix 3

Table 4: Benchmark results on TexBiG comparing Mask R-CNN and VSR for object detection and instance segmentation. This is the extended version of the table, including the different classes.

Task	Bounding Box		Mask	
Model	Mask R-CNN	VSR	Mask R-CNN	VSR
mAP	73.18	75.90	65.43	65.80
mAP_{50}	85.12	87.80	80.32	82.60
Advertisement	74.82	75.20	76.61	75.50
Author	52.49	55.40	47.30	47.80
Caption	44.43	52.10	45.48	49.30
Column title	86.89	89.20	85.73	82.00
Decoration	55.89	28.10	52.03	23.20
Editorial note	67.36	69.20	64.05	63.40
Equation	54.74	81.10	46.90	67.60
Footer	90.99	91.80	91.72	91.70
Footnote	83.26	84.30	83.58	82.30
Frame	91.99	91.50	0.00	0.00
Header	98.33	98.90	98.46	98.70
Heading	71.26	82.40	70.51	76.00
Image	70.45	73.10	60.99	61.80
Logo	87.78	89.80	78.33	81.40
Noise	82.57	83.30	65.59	65.60
Page number	78.29	69.30	75.33	60.70
Paragraph	86.90	90.50	86.60	89.20
Sub-heading	54.41	62.50	56.14	57.80
Table	57.58	74.40	57.88	76.30