DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis

Birgit Pfitzmann
IBM Research
Rueschlikon, Switzerland
bpf@zurich.ibm.com

Christoph Auer
IBM Research
Rueschlikon, Switzerland
cau@zurich.ibm.com

Michele Dolfi
IBM Research
Rueschlikon, Switzerland
dol@zurich.ibm.com

Ahmed S. Nassar
IBM Research
Rueschlikon, Switzerland
ahn@zurich.ibm.com

Peter Staar
IBM Research
Rueschlikon, Switzerland
taa@zurich.ibm.com

ABSTRACT

Accurate document layout analysis is a key requirement for high-quality PDF document conversion. With the recent availability of public, large ground-truth datasets such as PubLayNet and DocBank, deep-learning models have proven to be very effective at layout detection and segmentation. While these datasets are of adequate size to train such models, they severely lack in layout variability since they are sourced from scientific article repositories such as PubMed and arXiv only. Consequently, the accuracy of the layout segmentation drops significantly when these models are applied on more challenging and diverse layouts. In this paper, we present *DocLayNet*, a new, publicly available, document-layout annotation dataset in COCO format. It contains 80863 manually annotated pages from diverse data sources to represent a wide variability in layouts. For each PDF page, the layout annotations provide labelled bounding-boxes with a choice of 11 distinct classes. DocLayNet also provides a subset of double- and triple-annotated pages to determine the inter-annotator agreement. In multiple experiments, we provide baseline accuracy scores (in mAP) for a set of popular object detection models. We also demonstrate that these models fall approximately 10% behind the inter-annotator agreement. Furthermore, we provide evidence that DocLayNet is of sufficient size. Lastly, we compare models trained on PubLayNet, DocBank and DocLayNet, showing that layout predictions of the DocLayNet-trained models are more robust and thus the preferred choice for general-purpose document-layout analysis.

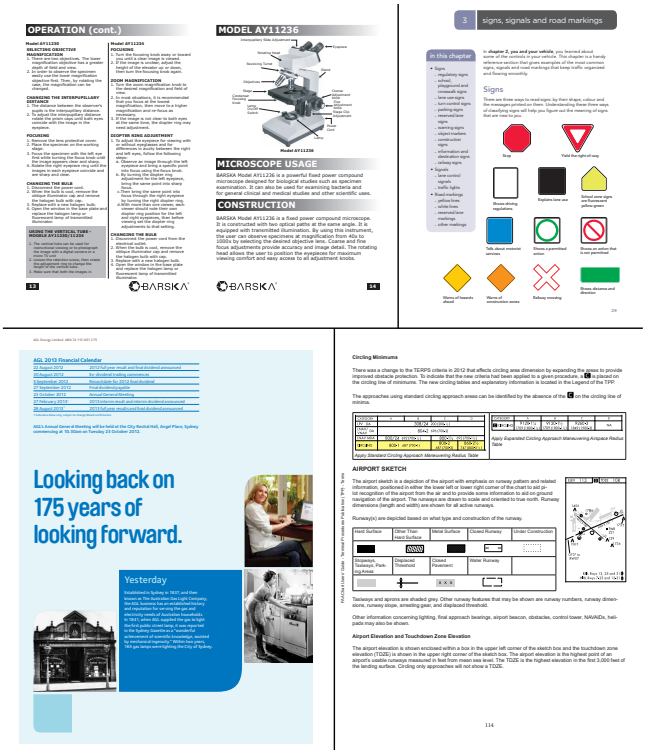

Figure 1: Four examples of complex page layouts across different document categories

CCS CONCEPTS

• **Information systems** → **Document structure**; • **Applied computing** → **Document analysis**; • **Computing methodologies** → **Machine learning**; **Computer vision**; *Object detection*;

KEYWORDS

PDF document conversion, layout segmentation, object-detection, data set, Machine Learning

ACM Reference Format:
Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter Staar. 2022. DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539043

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3539043

1 INTRODUCTION

Despite the substantial improvements achieved with machine-learning (ML) approaches and deep neural networks in recent years, document conversion remains a challenging problem, as demonstrated by the numerous public competitions held on this topic [1–4]. The challenge originates from the huge variability in PDF documents regarding layout, language and formats (scanned, programmatic or a combination of both). Engineering a single ML model that can be applied on all types of documents and provides high-quality layout segmentation remains to this day extremely challenging [5]. To highlight the variability in document layouts, we show a few example documents from the DocLayNet dataset in Figure 1.

A key problem in the process of document conversion is to understand the structure of a single document page, i.e. which segments of text should be grouped together in a unit. To train models for this task, there are currently two large datasets available to the community, PubLayNet [6] and DocBank [7]. They were introduced in 2019 and 2020 respectively and significantly accelerated the implementation of layout detection and segmentation models due to their sizes of 300K and 500K ground-truth pages. These sizes were achieved by leveraging an automation approach. The benefit of automated ground-truth generation is obvious: one can generate large ground-truth datasets at virtually no cost. However, the automation introduces a constraint on the variability in the dataset, because corresponding structured source data must be available. PubLayNet and DocBank were both generated from scientific document repositories (PubMed and arXiv), which provide XML or LaTeX sources. Those scientific documents present a limited variability in their layouts, because they are typeset in uniform templates provided by the publishers. Obviously, documents such as technical manuals, annual company reports, legal text, government tenders, etc. have very different and partially unique layouts. As a consequence, the layout predictions obtained from models trained on PubLayNet or DocBank is very reasonable when applied on scientific documents. However, for more *artistic* or *free-style* layouts, we see sub-par prediction quality from these models, which we demonstrate in Section 5.

In this paper, we present the DocLayNet dataset. It provides page-by-page layout annotation ground-truth using bounding-boxes for 11 distinct class labels on 80863 unique document pages, of which a fraction carry double- or triple-annotations. DocLayNet is similar in spirit to PubLayNet and DocBank and will likewise be made available to the public[1] in order to stimulate the document-layout analysis community. It distinguishes itself in the following aspects:

(1) *Human Annotation*: In contrast to PubLayNet and DocBank, we relied on human annotation instead of automation approaches to generate the data set.
(2) *Large Layout Variability*: We include diverse and complex layouts from a large variety of public sources.
(3) *Detailed Label Set*: We define 11 class labels to distinguish layout features in high detail. PubLayNet provides 5 labels; DocBank provides 13, although not a superset of ours.
(4) *Redundant Annotations*: A fraction of the pages in the DocLayNet data set carry more than one human annotation. This enables experimentation with annotation uncertainty and quality control analysis.
(5) *Pre-defined Train-, Test- & Validation-set*: Like DocBank, we provide fixed train-, test- & validation-sets to ensure proportional representation of the class-labels. Further, we prevent leakage of unique layouts across sets, which has a large effect on model accuracy scores.

All aspects outlined above are detailed in Section 3. In Section 4, we will elaborate on how we designed and executed this large-scale human annotation campaign. We will also share key insights and lessons learned that might prove helpful for other parties planning to set up annotation campaigns.

In Section 5, we will present baseline accuracy numbers for a variety of object detection methods (Faster R-CNN, Mask R-CNN and YOLOv5) trained on DocLayNet. We further show how the model performance is impacted by varying the DocLayNet dataset size, reducing the label set and modifying the train/test-split. Last but not least, we compare the performance of models trained on PubLayNet, DocBank and DocLayNet and demonstrate that a model trained on DocLayNet provides overall more robust layout recovery.

2 RELATED WORK

While early approaches in document-layout analysis used rule-based algorithms and heuristics [8], the problem is lately addressed with deep learning methods. The most common approach is to leverage object detection models [9–15]. In the last decade, the accuracy and speed of these models has increased dramatically. Furthermore, most state-of-the-art object detection methods can be trained and applied with very little work, thanks to a standardisation effort of the ground-truth data format [16] and common deep-learning frameworks [17]. Reference data sets such as PubLayNet [6] and DocBank provide their data in the commonly accepted COCO format [16].

Lately, new types of ML models for document-layout analysis have emerged in the community [18–21]. These models do not approach the problem of layout analysis purely based on an image representation of the page, as computer vision methods do. Instead, they combine the text tokens and image representation of a page in order to obtain a segmentation. While the reported accuracies appear to be promising, a broadly accepted data format which links geometric and textual features has yet to establish.

3 THE DOCLAYNET DATASET

DocLayNet contains 80863 PDF pages. Among these, 7059 carry two instances of human annotations, and 1591 carry three. This amounts to 91104 total annotation instances. The annotations provide layout information in the shape of labeled, rectangular bounding-boxes. We define 11 distinct labels for layout features, namely *Caption*, *Footnote*, *Formula*, *List-item*, *Page-footer*, *Page-header*, *Picture*, *Section-header*, *Table*, *Text*, and *Title*. Our reasoning for picking this particular label set is detailed in Section 4.

In addition to open intellectual property constraints for the source documents, we required that the documents in DocLayNet adhere to a few conditions. Firstly, we kept scanned documents

[1] https://developer.ibm.com/exchanges/data/all/doclaynet

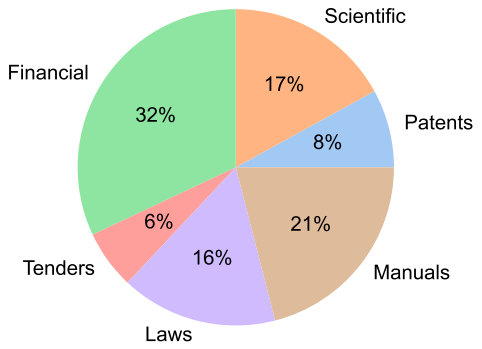

Figure 2: Distribution of DocLayNet pages across document categories.

to a minimum, since they introduce difficulties in annotation (see Section 4). As a second condition, we focussed on medium to large documents (> 10 pages) with technical content, dense in complex tables, figures, plots and captions. Such documents carry a lot of information value, but are often hard to analyse with high accuracy due to their challenging layouts. Counterexamples of documents not included in the dataset are receipts, invoices, hand-written documents or photographs showing "text in the wild".

The pages in DocLayNet can be grouped into six distinct categories, namely *Financial Reports*, *Manuals*, *Scientific Articles*, *Laws & Regulations*, *Patents* and *Government Tenders*. Each document category was sourced from various repositories. For example, Financial Reports contain both *free-style* format annual reports[2] which expose company-specific, artistic layouts as well as the more formal SEC filings. The two largest categories (*Financial Reports* and *Manuals*) contain a large amount of free-style layouts in order to obtain maximum variability. In the other four categories, we boosted the variability by mixing documents from independent providers, such as different government websites or publishers. In Figure 2, we show the document categories contained in DocLayNet with their respective sizes.

We did not control the document selection with regard to language. The vast majority of documents contained in DocLayNet (close to 95%) are published in English language. However, DocLayNet also contains a number of documents in other languages such as German (2.5%), French (1.0%) and Japanese (1.0%). While the document language has negligible impact on the performance of computer vision methods such as object detection and segmentation models, it might prove challenging for layout analysis methods which exploit textual features.

To ensure that future benchmarks in the document-layout analysis community can be easily compared, we have split up DocLayNet into pre-defined train-, test- and validation-sets. In this way, we can avoid spurious variations in the evaluation scores due to random splitting in train-, test- and validation-sets. We also ensured that less frequent labels are represented in train and test sets in equal proportions.

[2] e.g. AAPL from https://www.annualreports.com/

Table 1 shows the overall frequency and distribution of the labels among the different sets. Importantly, we ensure that subsets are only split on full-document boundaries. This avoids that pages of the same document are spread over train, test and validation set, which can give an undesired evaluation advantage to models and lead to overestimation of their prediction accuracy. We will show the impact of this decision in Section 5.

In order to accommodate the different types of models currently in use by the community, we provide DocLayNet in an *augmented* COCO format [16]. This entails the standard COCO ground-truth file (in JSON format) with the associated page images (in PNG format, 1025×1025 pixels). Furthermore, custom fields have been added to each COCO record to specify document category, original document filename and page number. In addition, we also provide the original PDF pages, as well as sidecar files containing parsed PDF text and text-cell coordinates (in JSON). All additional files are linked to the primary page images by their matching filenames.

Despite being cost-intense and far less scalable than automation, human annotation has several benefits over automated ground-truth generation. The first and most obvious reason to leverage human annotations is the freedom to annotate any type of document without requiring a programmatic source. For most PDF documents, the original source document is not available. The latter is not a hard constraint with human annotation, but it is for automated methods. A second reason to use human annotations is that the latter usually provide a more natural interpretation of the page layout. The human-interpreted layout can significantly deviate from the programmatic layout used in typesetting. For example, "invisible" tables might be used solely for aligning text paragraphs on columns. Such typesetting tricks might be interpreted by automated methods incorrectly as an actual table, while the human annotation will interpret it correctly as *Text* or other styles. The same applies to multi-line text elements, when authors decided to space them as "invisible" list elements without bullet symbols. A third reason to gather ground-truth through human annotation is to estimate a "natural" upper bound on the segmentation accuracy. As we will show in Section 4, certain documents featuring complex layouts can have different but equally acceptable layout interpretations. This natural upper bound for segmentation accuracy can be found by annotating the same pages multiple times by different people and evaluating the inter-annotator agreement. Such a baseline consistency evaluation is very useful to define expectations for a good target accuracy in trained deep neural network models and avoid overfitting (see Table 1). On the flip side, achieving high annotation consistency proved to be a key challenge in human annotation, as we outline in Section 4.

4 ANNOTATION CAMPAIGN

The annotation campaign was carried out in four phases. In phase one, we identified and prepared the data sources for annotation. In phase two, we determined the class labels and how annotations should be done on the documents in order to obtain maximum consistency. The latter was guided by a detailed requirement analysis and exhaustive experiments. In phase three, we trained the annotation staff and performed exams for quality assurance. In phase four,

Table 1: DocLayNet dataset overview. Along with the frequency of each class label, we present the relative occurrence (as % of row "Total") in the train, test and validation sets. The inter-annotator agreement is computed as the mAP@0.5-0.95 metric between pairwise annotations from the triple-annotated pages, from which we obtain accuracy ranges.

		% of Total			triple inter-annotator mAP @ 0.5-0.95 (%)						
class label	Count	Train	Test	Val	All	Fin	Man	Sci	Law	Pat	Ten
Caption	22524	2.04	1.77	2.32	84-89	40-61	86-92	94-99	95-99	69-78	n/a
Footnote	6318	0.60	0.31	0.58	83-91	n/a	100	62-88	85-94	n/a	82-97
Formula	25027	2.25	1.90	2.96	83-85	n/a	n/a	84-87	86-96	n/a	n/a
List-item	185660	17.19	13.34	15.82	87-88	74-83	90-92	97-97	81-85	75-88	93-95
Page-footer	70878	6.51	5.58	6.00	93-94	88-90	95-96	100	92-97	100	96-98
Page-header	58022	5.10	6.70	5.06	85-89	66-76	90-94	98-100	91-92	97-99	81-86
Picture	45976	4.21	2.78	5.31	69-71	56-59	82-86	69-82	80-95	66-71	59-76
Section-header	142884	12.60	15.77	12.85	83-84	76-81	90-92	94-95	87-94	69-73	78-86
Table	34733	3.20	2.27	3.60	77-81	75-80	83-86	98-99	58-80	79-84	70-85
Text	510377	45.82	49.28	45.00	84-86	81-86	88-93	89-93	87-92	71-79	87-95
Title	5071	0.47	0.30	0.50	60-72	24-63	50-63	94-100	82-96	68-79	24-56
Total	1107470	941123	99816	66531	82-83	71-74	79-81	89-94	86-91	71-76	68-85

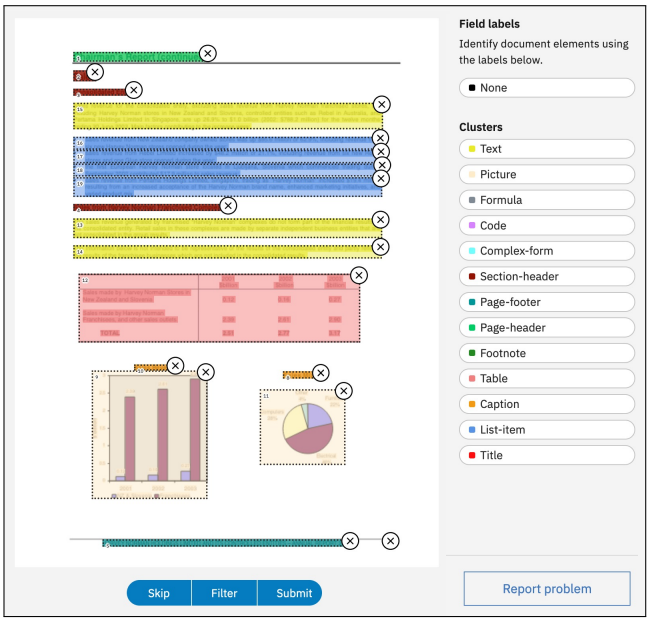

Figure 3: Corpus Conversion Service annotation user interface. The PDF page is shown in the background, with overlaid text-cells (in darker shades). The annotation boxes can be drawn by dragging a rectangle over each segment with the respective label from the palette on the right.

we distributed the annotation workload and performed continuous quality controls. Phase one and two required a small team of experts only. For phases three and four, a group of 40 dedicated annotators were assembled and supervised.

Phase 1: Data selection and preparation. Our inclusion criteria for documents were described in Section 3. A large effort went into ensuring that all documents are free to use. The data sources include publication repositories such as arXiv[3], government offices, company websites as well as data directory services for financial reports and patents. Scanned documents were excluded wherever possible because they can be rotated or skewed. This would not allow us to perform annotation with rectangular bounding-boxes and therefore complicate the annotation process.

Preparation work included uploading and parsing the sourced PDF documents in the Corpus Conversion Service (CCS) [22], a cloud-native platform which provides a visual annotation interface and allows for dataset inspection and analysis. The annotation interface of CCS is shown in Figure 3. The desired balance of pages between the different document categories was achieved by selective subsampling of pages with certain desired properties. For example, we made sure to include the title page of each document and bias the remaining page selection to those with figures or tables. The latter was achieved by leveraging pre-trained object detection models from PubLayNet, which helped us estimate how many figures and tables a given page contains.

Phase 2: Label selection and guideline. We reviewed the collected documents and identified the most common structural features they exhibit. This was achieved by identifying recurrent layout elements and lead us to the definition of 11 distinct class labels. These 11 class labels are *Caption*, *Footnote*, *Formula*, *List-item*, *Page-footer*, *Page-header*, *Picture*, *Section-header*, *Table*, *Text*, and *Title*. Critical factors that were considered for the choice of these class labels were (1) the overall occurrence of the label, (2) the specificity of the label, (3) recognisability on a single page (i.e. no need for context from previous or next page) and (4) overall coverage of the page. Specificity ensures that the choice of label is not ambiguous, while coverage ensures that all meaningful items on a page can be annotated. We refrained from class labels that are very specific to a document category, such as *Abstract* in the *Scientific Articles* category. We also avoided class labels that are tightly linked to the semantics of the text. Labels such as *Author* and *Affiliation*, as seen in DocBank, are often only distinguishable by discriminating on

[3] https://arxiv.org/