



Datasets and annotations for layout analysis of scientific articles

Andrea Gemelli¹ · Simone Marinai¹ · Lorenzo Pisaneschi¹ · Francesco Santoni¹

Received: 15 May 2023 / Revised: 16 October 2023 / Accepted: 23 January 2024
© The Author(s) 2024

Abstract

For a long time now, datasets containing scientific articles have been crucial to the analysis and recognition of document images. These document collections have frequently served as a testing ground for cutting-edge methods for optical character recognition, layout analysis, and document understanding in general. We thoroughly analyze and compare many datasets proposed for layout analysis of scientific documents, ranging from small collections of scanned papers to modern large-scale datasets containing digital-born papers, which have been proposed to train deep learning-based methods. Furthermore, we outline a detailed taxonomy of the annotation procedures used considering manual, automatic, and generative approaches, and we analyze their benefits and drawbacks. This survey is meant to provide the reader with a review of the most used benchmarks together with detailed information on data, annotations, and complexity, helping scholars to identify the most suitable dataset for their tasks of interest. We also discuss possible open problems to further enhance datasets to support research in the layout analysis of scientific articles.

Keywords Document understanding · Document layout analysis · Data collection · Data annotation

1 Introduction

Several research and application fields require annotated datasets to advance the development of intelligent systems. Among many, ImageNet [1] permitted the growth of novel approaches that have guided the creation of some of the most modern learning systems. Document Image Analysis and Recognition (DIAR) is not an exception, and several well-known benchmark datasets allowed researchers to advance the state of the art in DIAR and in general in pattern recognition. For instance, in the 1990s the NIST [2] and MNIST [3] datasets of handwritten digits have been instrumental for significant advances of techniques for pattern recognition [4]. DIAR is not limited to isolated character recognition, but

encompasses several tasks ranging from pre-processing, to layout analysis with the overall aim of achieving document understanding in many application domains [5]. Many application areas deal with proprietary data that cannot be made publicly available, due to copyright and privacy issues such as financial documents or health records. These latter difficulties, along with the annotation effort required for large quantities of documents, are usually the main challenges faced when creating a new benchmark dataset for DIAR.

Among other tasks, Document Layout Analysis (DLA) research advanced significantly in the 1990s thanks to a new collection of scanned pages of scientific articles. Similar to NIST, the UW datasets [6, 7] set a milestone for evaluating research progresses. Thereafter, scientific articles have been widely used as benchmark sources of data due to their availability, in terms of quantity and accessibility, and their rich semantic structure that allows researchers to focus on different tasks in the document understanding pipeline, e.g., Table Detection (TD) and Table Recognition (TR). Although some tasks are nowadays basically solved (e.g., physical layout analysis, that is, the identification of homogeneous regions of text in the page), there is still space for research in the analysis of challenging regions of documents (e.g. tables and graphical illustrations) as well as for the overall understand-

✉ Simone Marinai
simone.marinai@unifi.it

Andrea Gemelli
andrea.gemelli@unifi.it

Lorenzo Pisaneschi
lorenzo.pisaneschi@unifi.it

Francesco Santoni
francesco.santoni@unifi.it

¹ Dipartimento di Ingegneria dell'Informazione (DINFO),
Università degli Studi di Firenze, Via di Santa Marta, 3,
Florence, Italy

ing of scientific articles published with non-common styles and layouts.

In this paper, we aim to provide a guide to the *different datasets* that have been proposed over the past 30 years for supporting research on DLA over scientific articles: other recent surveys focused either on historical document collections [8] or state-of-the-art methods for page object detection [9], and we suggest the reader to go through them to a broader comprehensive overview of the DIAR field. In addition to a comprehensive inventory of datasets, highlighting their strengths and limitations, we focus our attention on the *annotation procedures* that have been proposed for such collections, with an analysis of the advantages and disadvantages of different approaches. We reviewed some of the most important state-of-the-art methods tested on the collections in this survey, but for a complete overview of DLA, we would like to refer to one survey over the most important methods proposed to tackle this task [10].

We focus on DLA of scientific articles, for three main reasons:

- to investigate the most used annotation procedures, along with the challenges of creating large and qualitative datasets with good annotations;
- to provide researchers with an overview of available datasets along with their details, to help the identification of the best suitable benchmarks to develop and test novel algorithms;
- to outline how document collections have changed until today, posing questions and open problems that could further enhance the DIAR research.

The paper is organized as follows. In Sect. 2, we describe DLA and the principal techniques used to tackle it, with a particular focus on the analysis of scientific articles. Then, a detailed description of annotation procedures is depicted in Sect. 3. After an overview of the collections reported in this survey (Sect. 4), we divide them in three main categories, starting from small scale fully annotated in Sect. 5, mostly containing scanned documents. Then, in Sect. 6, partially annotated collections are described, focused only on challenging parts of scientific documents such as tables and figures. Finally, large-scale fully annotated datasets are listed in Sect. 7. For completeness, in Sect. 8 we provide a broader overview of significant datasets that are related to DLA for different types of documents, and in Sect. 9 we discuss and summarize the impact and complexity of each collection presented in this survey, along with the latest state-of-the-art methods tested on the datasets. We then discuss identified open problems and challenges to the field of DLA for scientific articles in Sect. 10. Finally, we outline the conclusions in Sect. 11.

2 Document layout analysis of scientific articles

In addition to Optical Character Recognition (OCR) of printed or handwritten characters, one of the most investigated tasks in Document Image Analysis and Recognition has been Document Layout Analysis which aims at finding regions in a page, such as text or figures (physical layout analysis) and recognizing and classifying them, e.g., discriminating text blocks as title or paragraph (logical layout analysis). In physical layout analysis, the aim is to identify homogeneous regions (usually by means of bounding boxes) [10, 11]. Since many tasks can be addressed in DLA of scientific articles, we summarize the main ones in Table 1 together with the acronyms used in this paper.

Over the years, several methods have been proposed attempting to solve DLA, following the application of novel techniques and the gathering of larger collections of annotated data. Ranging from the early 1990s up to nowadays, it is possible to broadly divide the different techniques into three main groups: heuristics, statistical machine learning, and deep learning methods. The first two groups are described in [12], dividing different approaches depending on two criteria.

The first criterion refers to *how* the document is analyzed, either using bottom-up, top-down, or hybrid techniques. Bottom-up techniques start gathering information at the pixel level and then iteratively group them into larger areas, from connected components (CCs) up to larger meaningful areas of text or non-text (e.g., figures). Representative algorithms from this group are RLSA [13], Docstrum [14], and Voronoi diagrams [15]. On the contrary, top-down techniques start from the whole document until basic components are found in subsequent steps, like in the X-Y cut algorithm [16]. Finally, hybrid methods are compositions of the aforementioned ones.

The second criterion discriminates the techniques considering *what* is analyzed either the physical or the logical document layout. The first one aims at the identification of homogeneous regions in the page while the latter at assigning functional information, a label, to these regions. Methods are categorized on these terms depending on the downstream task they are used for. To cite some, Strouthopoulos and Papanarkos [17] propose an Artificial Neural Network ANN to classify 8×8 document patches as graphics or halftones. Wu et al. [18] segment text regions using a series of split-or-merge operations guided by a binary SVM classifier. Once the page objects are segmented and/or classified, some post-processing techniques could be considered to generalize the results over different layouts [10]. It is worth to notice that most methods for layout analysis have been demonstrated and tested on collections of digitized scientific articles.

Table 1 Acronyms of tasks addressed in layout analysis of scientific articles

Acronym	Full name	Description
DLA	Document layout analysis	Detect and recognize objects inside documents, such as figures, tables, text
TD	Table detection	Find bounding box coordinates of tables (also called table extraction: TE)
TSR	Table structure recognition	Identify columns, rows, spanning and empty cells of tables
TR	Table recognition	Includes TSR and the analysis of the content (e.g., reconstructing the table in a machine-readable format)
TFA	Table functional analysis	Discrimination between header and cells in tables
TTC	Table type classification	Classifying tables given their content
FE	Figure extraction	Aims at the extraction of images and tables. It may consider also captions

More recently, deep learning techniques have been used also for DLA, taking advantage of larger document collections. In a recent paper summarizing models, tasks, and datasets for document AI [19], the most important techniques are listed and divided into three broad categories: Convolutional Neural Networks (CNN), Graph Neural Networks (GNN), and Transformers. For instance, Faster R-CNN [20]- and Mask R-CNN [21]-based architectures have been widely used across several benchmarks to detect page objects, while LayoutLM [22] has been the first transformer-based architecture applied to DIAR. Graphs have been mainly used for information extraction over administrative documents [23, 24] and TD/TR [25]. In the ICDAR2021 competition [26], Zhang et al. [27] achieve the SOTA on the Document Layout Recognition track, proposing a multi-modal Mask-RCNN-based object detection framework that makes use of vision, language, and geometry. A more detailed description of state-of-the-art methods tested on the datasets presented in this survey is found in Sect. 9.3.

It is useful to observe that the majority of methods used so far for DLA strongly rely on supervised learning. This is why the amount of labeled data has been always at the same time an important and complex issue to address in DIAR. There are two main problems related to annotated collections of documents: (i) not all types of documents are publicly available due to, for example, policy issues; therefore, many benchmarks are composed by scientific articles; (ii) not all available data come with structured information for automatic annotation, forcing to choose either to manually inspect a small amount of data or exclude a vast amount of unlabeled documents. These problems highly affect the proposed DLA frameworks both in terms of robustness and generalization, due to the lack of variability in available benchmarks. A solution to fill the gap between expensive annotation procedures and large automatically labeled collections is the generation of synthetic data that, by construction, come with annotations (e.g. [28]). Even if this solution opens new possibilities, it is not easy to generate data as faithfully as possible to real data

and to avoid the trained algorithms dropping drastically in the wild.

In the next chapters, we describe the peculiarities of scientific document datasets and then analyze traditional and novel annotation procedures including some generative methods.

3 Annotation procedures

As depicted before, the state-of-the-art techniques proposed to tackle DLA have been often supervised oriented, thus requiring carefully annotated data. In particular in the last years, an increasing demand of larger datasets for deep learning methods has started to open new challenges on how to annotate such collections. Throughout the years, procedures for annotating documents have been proposed trying to maximize two different measures, usually inversely proportional: quality and amount of annotations, taking into consideration also time, costs, and data variability. As summarized in Fig. 1, we broadly categorize the annotation procedures in three different classes:

- *Manual*. A designated team is given a set of rules to be followed, along with an annotation tool to both help them and to output the labels in a standard format.
- *Automatic*. A set of techniques, algorithmic or based on machine learning, that scale with the amount of data to be annotated. The only constraint is that scientific articles in PDF need additional structured files, such as L^AT_EX, XML, or HTML. Human intervention is usually required mostly to check the quality of a subset of randomly selected samples.
- *Generative*. Generative models can synthesize data from documents of a given domain. Along with generated samples, annotations are also generated exploiting deep learning and transformers architectures.

In Table 2, we list the tools for document annotations used in the literature. We report if it is possible to use them through

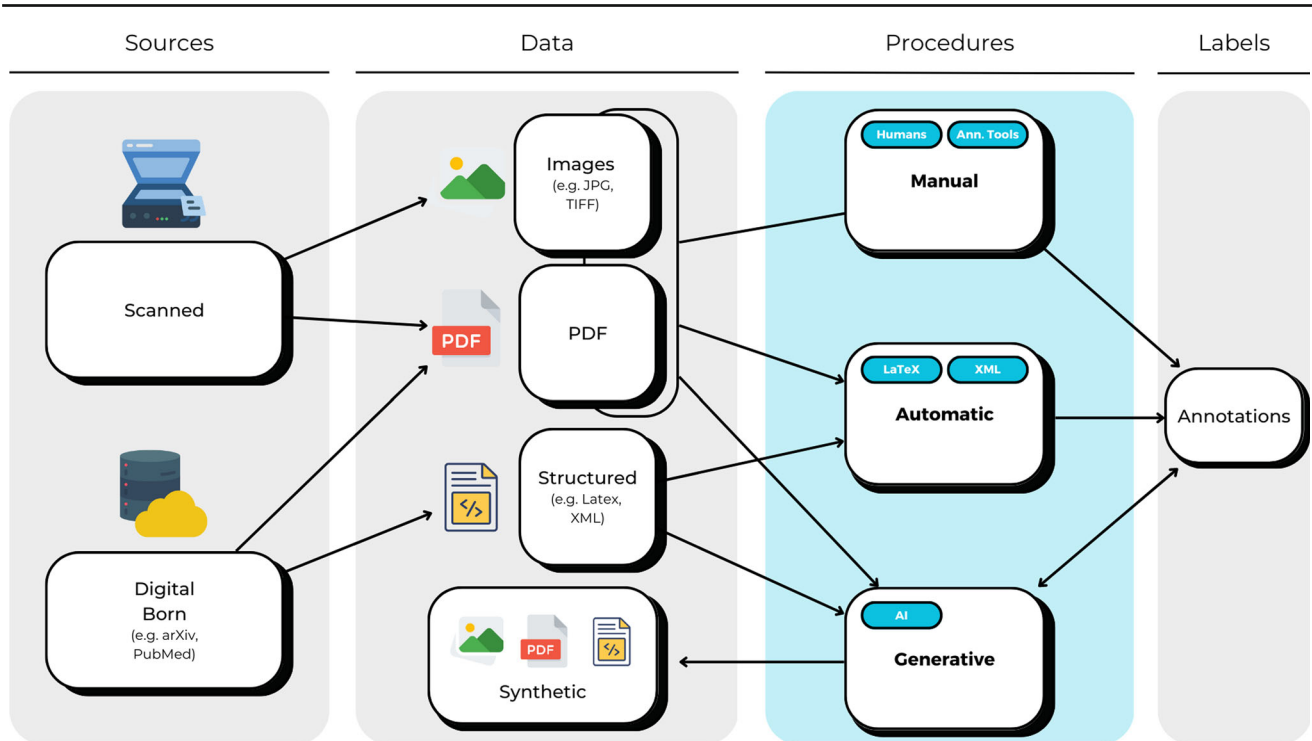


Fig. 1 Different pipelines to create labeled data depend on the combination of three main factors: the sources from where to crawl the data (left), which data are available (center), and which annotation proce-

dures to involve (right). Synthetic documents do not require to follow any specific annotation procedure since labels are given within the generated data

Table 2 Main tools used to support annotation

Tool name	Availability				Input	Procedure	Output
	Online	Offline	API	Desktop			
Aletheia [29]	✗	✓	✗	✓	Img	M	PAGE [30]
LabelImg [31]	✗	✓	✗	✓	Img	M	PASCAL-VOC [32]
LabelMe [33]	✓(limited)	✓	✗	✓	Img	M	JSON
CCS [34]	✓	✗	✓	✗	PDF	M, ML	JSON, XML
VOTT [35]	✓	✓	✓	✓	Img	M	JSON, CSV, PASCAL-VOC
VIA [36]	✓	✓	✓	✓	Img	M	CSV, JSON, COCO
FigureSeer[37]	✗	✗	✗	✗	PDF	ML	Structured data
Tabula [38]	✗	✓	✓	✓	PDF	A	CSV, XSLX
Pdfminer [39]	✗	✓(lib.)	✗	✗	PDF	A	Python obj., text
Pymupdf [40]	✗	✓(lib.)	✗	✗	PDF	A	Python obj
Lxml [41]	✗	✓(lib.)	✗	✗	XML	A	Python obj
Grobid [42]	✓	✓	✓	✗	PDF	A, ML	XML/TEI
Pdfplumber[43]	✗	✓	✗	✓	PDF	A	CSV

The Page format and Pascal-Voc formats are specific XML structures. Python obj. indicates that the mentioned tool is a library which outputs an annotation as a Python object. Procedures can be Automatic (A), Manual (M), and/or supported by Machine Learning models (ML)

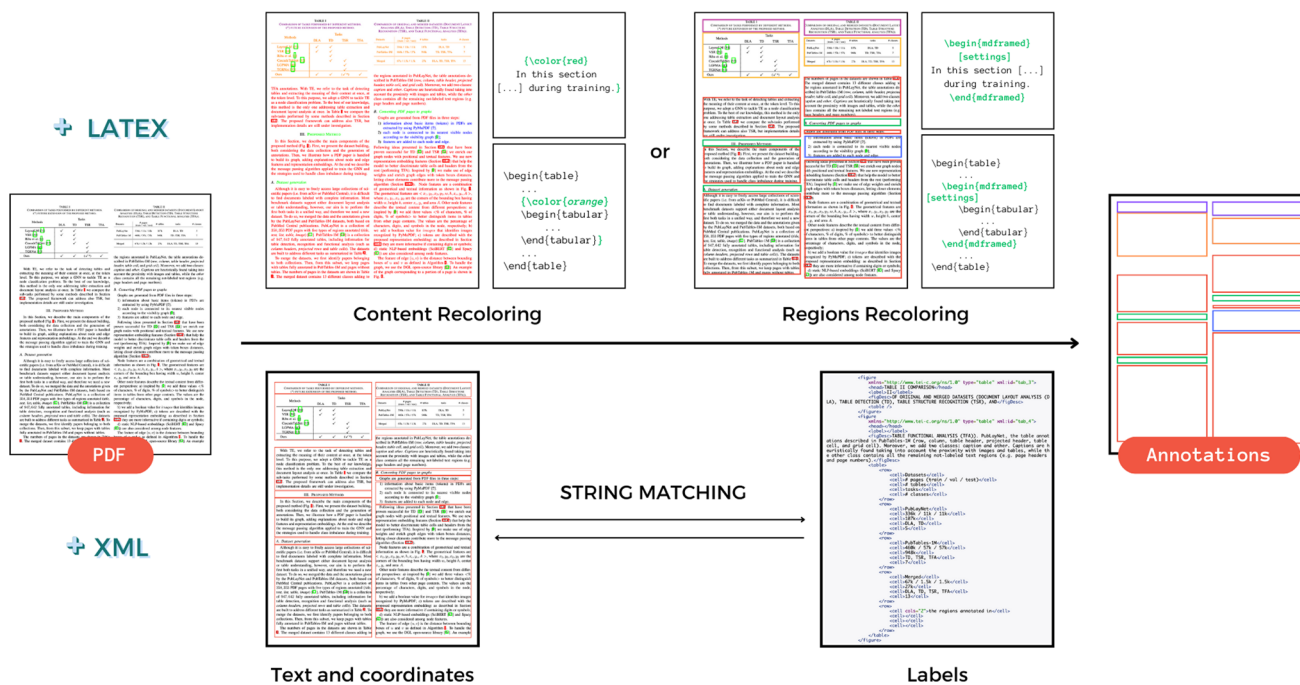


Fig. 2 General scheme adopted by automatic annotation. The upper part summarizes the two approaches based on recoloring L^AT_EX source code and compiling: both “Content Recoloring” and “Regions Recoloring” make use of RegEx matching to find target commands, injecting custom code, and compiling the new PDF colored version. Then, using

RGB values regions and labels are extracted. On the lower part, XML is used to give logical meaning to the content and locations extracted from PDF through PDFMiner-based [39] tools. To match XML and PDF content, algorithms such as Levenshtein distance and Bag of Word are usually involved

an online service, a local application or calling an API. We also summarize the annotation process for each tool, outlining the input data, the procedure, and the output data format.

3.1 Manual annotations

Manual techniques do not scale well with the dataset size. In addition, guidelines need to be defined so that different annotators follow the same rules, to produce coherent and homogeneous annotations among data. This approach has been the foundation of most of datasets across different machine learning fields; today it is often not preferred due to its high cost both in terms of money and time. However, researchers continue to propose manually annotated data since the human supervision is capable of yielding qualitative annotations and gather important information that are not easily available through automatic annotations.

3.2 Automatic annotations

PDF files of scientific articles are often not semantically structured. It is not trivial to access information such as text, tables, and titles and to detect them easily in the page layout starting from a PDF. However, scientific documents are often stored in other structured formats along with PDF. The

information in these files, such as L^AT_EX and XML, makes it possible to automatically identify objects in the layout. The general annotation scheme in this case is depicted in Fig. 2. We describe in detail the two main annotation approaches in the next subsection, while the differences between datasets can be found in their relative sections. We also include in this category *semi-automatic* annotations procedures, meaning the usage of automatic ones along with human supervision during the process.

3.2.1 From L^AT_EX format

ArXiv is a huge collection of scientific documents made available by the authors. From arXiv, many researchers collect data to build document datasets. Some datasets presented in this survey [44–46] have been built starting from arXiv, because most papers in the repository are stored including both PDF and L^AT_EX files, which allows researchers to implement automatic pipelines to annotate documents much faster than humans. This approach is visually described in the top part of Fig. 2: usually, the source L^AT_EX code is injected with custom commands, e.g., `\color{}` or `\begin{mdframed}... \end{mdframed}`, around detected target regions such as tables, lists, images. Then, a new version of the PDF is compiled and leveraging infor-

mation extraction tools that can identify colored text, regions are extracted in terms of bounding boxes and labels. The two main approaches are based on *content* or *regions recoloring* according to the granularity of the colored items: in the first case individual words and in the second one bounding boxes of regions. The drawbacks of automatic annotation based on L^AT_EX are that sometimes changing the source code results in either layout alteration or compilation errors.

3.2.2 From XML format

It is possible to approach automatic annotation by also considering XML files together with the corresponding PDF files. This is the case for datasets created from PubMed (e.g., [47] [48] [49]). PubMed Central (PMC) is a free full-text archive of biomedical and life sciences journal literature hosted at the US National Institutes of Health's National Library of Medicine (NIH/NLM). It is freely accessible since 1997. In PubMed, documents are presented in PDF and XML formats, making possible to integrate information in PDF and XML files. As depicted in the bottom part of Fig. 2, region coordinates and text are first extracted from the PDF using PDFMiner-based tools. Then, involving string matching techniques such as Levenshtein distance or Bag of Word (BoW), different regions are associated with different tags. Finally, labels are attached to the final layout annotations. The drawbacks of automatic annotation based on XML are that this structured information is not as widely available as L^AT_EX for scientific documents and cannot be easily edited to compile a modified PDF.

3.3 Generative models

An alternative solution is the generation of synthetic documents together with their annotations. Using generative methods, it is possible to automatically create annotations for an arbitrary amount of data, with layout and content variations. These approaches can use as starting point automatic or manually labeled documents, and they learn how to create new layouts and contents which are reasonably similar to real ones. However, despite the recent improvement, creating synthetic documents is still a difficult task.

Differently from manual and automatic procedures, in this case it is not possible to outline a general approach. Even if none of the datasets presented in this survey is generated, we found interesting to revise, to the best of our knowledge, the most important contributions developed so far in this direction. Document Domain Randomization [50] is a method which attempts to create document pages by diversifying real documents examples. DDR models textual and non-textual contents of a starting set of data following rules which are defined by the users, such as document layouts, font and style. DocSynth [51] is a model which, given a set of document

images, generates new variations leveraging a Generative Adversarial Network (GAN). This method generates images to augment existing datasets. Similar to DocSynth, READ (Recursive Autoencoders for Document Layout Generation) [52] is an architecture which relies on a Recurrent Neural Network (RNN) and Variational Auto-Encoder (VAE). Starting from a training set of document examples, with semantic-based labels, READ is able to generate large quantities of variable and plausible synthetic layouts similar to the original ones.

More recently, transformer-based architectures, on top of their success on NLP tasks, started to be used to learn and generate document layouts as well. LayoutTransformer [53] leverages self-attention to learn and understand the existing relationships between regions in a given layout. The authors based the generation of new scientific papers on data in the PubLayNet [49] dataset. Inspired by this work, we proposed an approach to generate layout and content of scientific papers in high-resolution PDF format [54], filling the regions with generated text and tables and images crawled from [55].

3.4 Discussion

One of the most challenging aspects when creating a new dataset is how to annotate the data, to scale with the dimension of the collection and include in the data a good amount of variability. Manual annotations are time consuming and expensive, but ensure good quality; on the contrary, automatic procedures can generate more annotated data, but require often human intervention due to errors. Document generation could be a good alternative to meet the pitfalls of the aforementioned approaches, but it is still a complex task: it is therefore an excellent opportunity to advance research in this direction.

4 Categorization of datasets

In this survey, we focus on scientific document collections, since they have been widely used by the research community to test and compare different approaches for DLA. Other related datasets, either considering scientific papers for different tasks or addressing DLA for different documents, are summarized in Sect. 10.

Following the description of the annotation procedures given before, we generally divide the datasets as shown in Fig. 3. This graphic is useful to visualize different procedures used for the creation of datasets. Generative-based techniques are used during model training for several purposes and, to the best of our knowledge, there are no collections published to be reported here.

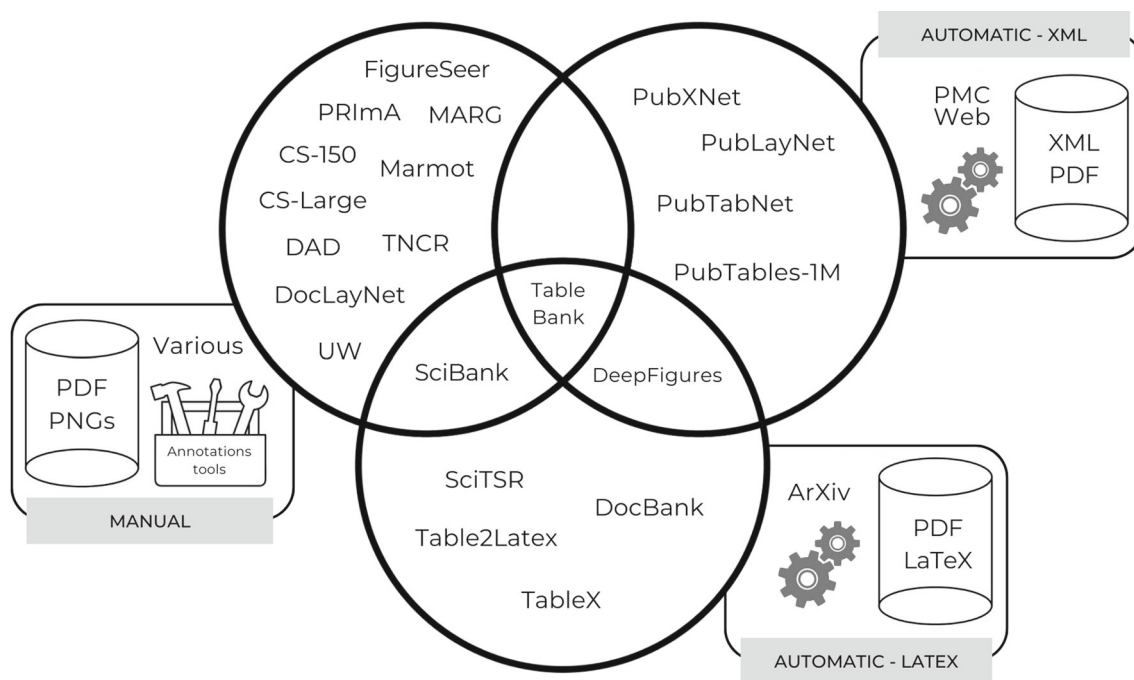


Fig. 3 Datasets overview based on different annotations procedures, enriched with information on where the documents have been crawled from. In particular, SciBank and TableBank specifically state that some

human intervention occurred for quality checking. TableBank and DeepFigures contain used automatic procedures based on $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ and XML

Moreover, considering *what* has been annotated in addition to *how*, we divide the datasets into three categories:

- *small-scale fully annotated*, where the manual annotation is commonly used and document images come from scanners or digital cameras.
- *partially annotated* documents, where the focus is mainly on finding and recognizing specific regions, such as tables or figures, two important objects for DLA.
- *large-scale fully annotated*, collections—where automatic labeling is preferred because of the huge amount of data.

These three main approaches for building datasets roughly follow also the timeline of research of the last 30 years. In the timeline shown in Fig. 4, the datasets are arranged in a chronological order and labeled according to the previous categories. The first proposed datasets contained scanned documents manually annotated, while most recent collections are built automatically and in general contain digital-born documents. Small-scale fully annotated datasets (Sect. 5) are represented as red triangles, partially annotated datasets (Sect. 6) as green circles, and large-scale fully annotated datasets (Sect. 7) as blue squares.

5 Small-scale fully annotated datasets

In this section, we present small datasets, that is, with a maximum order of about 10,000 pages. It is no coincidence that these collections are mostly composed of scanned documents, considering the resources required to obtain and annotate the data. Moreover, these datasets are have been released before 2009 except for ScanBank [57] (2021), which is also the larger among the others. These datasets are summarized and compared in Table 3.

5.1 University of Washington datasets

In the 1990s, the research group lead by Prof. Haralick released three datasets aimed at fostering research on the development of OCR tools and on document analysis in general. The aim has been surely achieved since the three datasets that came out from this work have been widely used by researchers and laid out the use of benchmark datasets and open evaluation metrics to support researchers to design and evaluate novel algorithms [6, 7]. In short, the UW datasets have been the model for the subsequent collections of data arranged by researchers in the community.

The *English Document Database I* was released in 1993 and consists of scanned pages from English technical journals randomly selected from the University of Washington library. The dataset was distributed as CD-ROM delivered by express

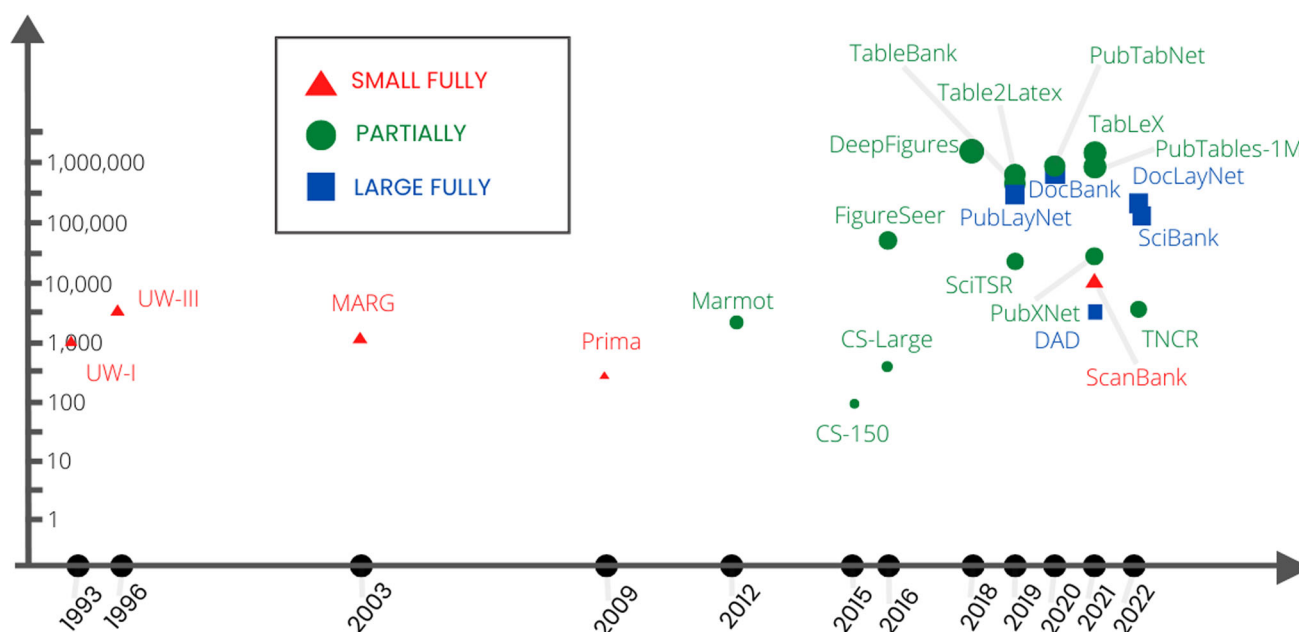


Fig. 4 Timeline of small-scale fully annotated (red triangles), partially annotated (green circles), and large-scale fully annotated (blue squares) datasets. On the y-axis, the size of each dataset is in logarithmic scale. Symbols are scaled according to the size of the corresponding dataset

Table 3 Small-scale fully annotated document collections comparison

Name	Av.	Data	Source	Size	Tasks	Format	Annotations		
							Input	Procedure	Output
UW-1 [6]	✗	P	Library	1147	OCR, DLA	TIFF	–	Manual	ASCII
UW-III [7]	✗	P	Library	3796	OCR, DLA	TIFF	–	Manual	ASCII
MARG [56]	✗	P	DL	1553	DLA	TIFF	–	n/a	XML
Prima [29]	✓	P	Various	305	DLA	Image	–	Manual	PAGE-XML [30]
ScanBank [57]	✓	D	DL	10,182	FE	Image	–	n/a	JSON

Full description of tasks is given in Table 1. DL refers to Digital Libraries. Data are either Pages (P) or Documents (D). Input blank lines mean that no additional files have been used for labeling

mail. It contains image files (bi-level and gray scale) and corresponding ground-truth information. The images came from scanned documents (from the journals themselves or photocopies) and synthetic ones generated with L^AT_EX. Additional information is related to the noise level in the document and metadata characterizing the provenance of the document.

Concerning the ground-truth information, pages are first decomposed into “header,” “footer,” and “live-matter” areas. The header is text appearing on top of the page and for scientific journals; it usually includes information like the name of the article, the journal, the authors, and the page number. Apart from the “footer,” the remaining part of the page is referred to as the “live matter.”

Each of these zones is described by its bounding box and is often decomposed at a finer level into sub-zones of different types (e.g. text, figure, tables, half-tones, and mathematical equations) that are again described by their bounding boxes. For each zone, additional information is provided, including

zone semantic meaning (e.g., for text zones we can have a section heading or a reference list item), the dominant font, the font style, etc. At the lower level, for each text zone, its contents in terms of ASCII text are provided, using escape sequences for non-ASCII symbols.

5.2 Medical Article Records Groundtruth

The Medical Article Records Groundtruth Dataset (MARG) contains bi-level images mostly coming from journal articles [56]. Ground-truth information is gathered from the output of the MARS (Medical Article Records System) system that combines scanning, OCR, document image analysis, and lexical analysis algorithms. MARS has been developed by an R&D team at the National Library of Medicine (NLM) and can automatically extract bibliographic metadata from paper-based biomedical journals to populate the Library’s database

(Medline). The ground-truth data, provided in XML, include page, zone, line, word, and character-level information.

The MARG dataset contains only title pages where the title, authors, affiliation, and abstract are organized in nine different layouts. For instance, in the layout of type “A” the title, authors, affiliation, and abstract are centered at the top of the title page and the rest of the page has a two-column layout.

5.3 PRImA

The dataset produced by the PRImA research group [29] is a realistic and highly accurate collection of several documents with a wide variety of layouts. It is easily accessible through a web application,¹ allowing researchers to identify subsets of interest. One subset was used for the ICDAR2009 Page Segmentation Competition [58]. The dataset is composed of scientific publications, magazines, and technical journals from mainstream publications related to news, business, and technology.

It contains 1240 ground-truth images in a 7 to 1 ratio in favor of magazine pages [29]; as of September 2022, it is composed of 305 images with 265 magazine pages and 40 technical articles. Metadata associated with each document contain administrative and bibliographic information (title, publication, author, copyright holder, etc.), scanning features, and content information such as the presence of images and/or graphical items, number of columns, and variety of font sizes. The dataset was built from scans of documents at 300 dpi in 24-bit color. Steps to eliminate the skew both in digitization and post-scanning are introduced. The scans are then cropped and binarized

The ground truth is stored in an XML-based format whose schema is part of the PAGE [30] image representation framework. Different region types are represented: text, image, line drawing, graphic, table, chart, separator, maths, noise, and frame. The textual regions are hierarchically divided into paragraphs, text lines, words, and glyphs. Each region is outlined with a polygon having only horizontal and vertical edges. For each region type, different information is collected; text regions are described by language, font, reading direction, text color, background color, and logical label. Trained operators added the region boundaries and region-level metadata through the semi-automated ground-truthing tool *Aletheia*. After a set of non-overlapping imprecise outlines is marked by the operator, *Aletheia* fits the set to the region contents like a shrink wrap operation. The tool can outline pages, regions, text lines, words, and glyphs.

The MARG and Prima are similar to UW datasets in size, but they concentrate on Layout Analysis using more modern annotation formats (XML).

¹ PRImA Layout Analysis Dataset.

5.4 ScanBank

After years of automatic labeling of digital-born documents labeling (Sects. 6 and 7), the ScanBank Dataset [57] collects and annotates scanned electronic theses and dissertations (ETDs) to be used for figure extraction. To motivate the need for this type of dataset, the authors claim that methods such as DeepFigure [47], trained over only digital-born documents, are unable to generalize over scanned ones that present a more challenging task. A total of 10,182 images of pages and 3300 figures were obtained across the 70 sampled ETDs, crawled from MIT’s DSpace repository². Each PDF page is converted into an image and then labeled manually with VGG Image Annotator (VIA) [36]. The resolution is scaled to 100 DPI to resemble a “screenshot” of a page from the PDF. Captions, table of contents, and list of tables and figures are labeled too, using bounding boxes. The authors also propose some data augmentation techniques trying to simulate actual scanned documents in a large collection of annotated scientific papers.

6 Partially annotated datasets

In this section, we present datasets that are intended for layout analysis considering only some specific parts, such as tables and figures. Table Detection and Figure extraction are important and challenging subtasks of layout analysis: they introduce a significant variability beyond text regions and require the use of more complex pipelines and techniques. The datasets summarized in Table 4 are relatively large and recent. As shown in Fig. 4, they are temporally located between the scanned datasets and digital-born datasets. Their magnitude varies between 100 (CS-150 [60]) and 1 M (TabLeX [66]) pages. In particular, it is possible to notice a particular focus of the community for figure extraction, in particular for tables, given the large amount of collections compared to other regions of the document layout. This is mainly due to two reasons: (i) tables come with nearly infinite variation in layout and finding methods generalizing over them requires a huge amount of labeled data and (ii) tables summarize important information and their detection and recognition are crucial to perform document understanding and information retrieval.

6.1 Marmot

Marmot [59] is a large standard Chinese and English dataset for table detection. It contains over 2000 PDF pages collected in a 1:1 proportion from Chinese e-books and English conference and journal papers. The ground-truth data were

² MITDspace.

Table 4 Partially annotated document collections comparison. Full description of tasks is given in Table 1

Name	Av.	Data	Source	Size	Tasks	Format	Annotations		
							Input	Procedure	Output
Marmot [59]	✓	P	Various	2000	TD	PDF	–	M	XML
CS-150 [60]	✓	P	Conf	150	FE	PDF	–	M	JSON
CS-Large [61]	✓	P	DL	346	FE	PDF	–	M	JSON
FigureSeer [37]	✓	P	DL	60,000	FE	PDF	–	M	–
DeepFigures [47]	✓	P	DL	1.3M+	FE	PDF	L ^A T _E X/XML	A	–
SciTSR [62]	✓	T	DL	15k	TSR	PDF(img)	L ^A T _E X	A	–
TableBank [63]	✓	P	DL	370k	TD, TSR	arXiv	L ^A T _E X/Word	A	JSON/HTML
Table2Latex [44]	✗	T	DL	450k	TSR	PNG	L ^A T _E X	A	–
PubTabNet [64]	✓	T	DL	516k+	TD, TSR	IMG(html)	PDF/XML	A	JSON
PubXNet [48]	✗	P	DL	24k+	TD, TSR	PDF	PDF/HTML	A	JSON
PubTables-1 M [65]	✓	P	DL	1 M	TD/TSR/TFA	PDF/XML	PDF/XML	A	JSON
TabLeX [66]	✓	T	DL	1 M+	TSR/TD	IMG	PDF/L ^A T _E X	A	–
TNCR [67]	✓	T	Various	6621	TD, TTC	IMG	PDF	A	IMG

Conf. (Conferences) are proceedings from NIPS, ICML, and AAAI. Annotations column describes how inputs are transformed to ground-truth (output) using a specific procedure. "DL" refers to Digital Libraries. Procedures are either Manual (M) or Automatic (A). Data are either Pages (P) or Tables (T). Input/output blank lines mean that no additional files have been used for labeling and/or produced as output

extracted with a semi-automatic tool, *Marmot*, and then double-checked by 15 people. The pages show a good variety in language types, page layouts, and table styles. The Chinese e-Book pages, provided by the Founder Apabi digital library, contain no more than 15 pages per book and are mostly in one-column layout. The English pages are both in one-column and two-column layouts. Half pages in the dataset contain at least one table and the other half contain complex layouts that can be mistaken as a table such as matrices and figures. Each page is described by a labeled ground truth, a page image at 600 dpi, and an XML description of objects attributes. The ground-truth data schema defines a set of tags that describe *leaves* (basic page units corresponding to text, image, or graph contents) and *composites* (logical components labeled by the tool and including more leaves). Each table is composed by three parts: *table caption*, *table footnote*, *table body*. Each part contains *textline* elements (an entire textline not crossing a page column) that in tables correspond to single lines in table cells. Textlines then contain text characters.

6.2 CS-150 and CS-Large

In 2015, the CS-150 dataset [60] has been released to address figure extraction. It is a collection of 50 papers from NIPS 2008–2013, 50 from ICML 2009–2014, and 50 from AAAI 2009–2014, selecting 10 published papers at random from each conference and year. Images, tables, and captions have been labeled by experts using the LabelMe tool [33]. In 2016, the authors released a larger version of the dataset called CS-

Large [61] by randomly sampling 346 papers from Semantic Scholar [68] published after 1999.

The authors also propose PDFFigures 2.0 [61], a tool built on top of the previous version for a wider range of different styles and layouts (tested on CS-Large). It is based on data-driven heuristics, which exploit formatting conventions used consistently in the computer science domain.

6.3 FigureSeer

The end-to-end framework FigureSeer [37] parses result figures to enable search and retrieval of results in research papers. To benchmark the framework, an annotated figure parsing dataset is offered. The dataset is produced using over 20,000 papers from five conferences (CVPR, ICML, ACL, CHI, AAAI) gathered from CiteSeerX indexed by Semantic Scholar. In the papers, there are more than 60,000 figures. Mechanical Turk [69] is used to annotate the figures.

6.4 DeepFigures

DeepFigures [47] is the first large automatically labeled dataset of scientific papers for figure extraction. It counts more than one million scientific papers, with 5.5 million induced labels of tables and images bounding boxes. This paper introduces one of the first approaches to automatically label large document collections.

Following the general automatic procedure schema described in Sect. 3.2, in DeepFigures both L^AT_EX and XML structured information has been used. In these scenarios, tables, figures, and captions have been surrounded by bounding box, using:

(i) difference of original and re-compiled PDFs according to RGB pixel values for $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ sources; (ii) dynamic programming to find the substrings in the PDF text with the smallest Levenshtein distance to the caption text in the XML file for images, bag of words similarity for tables (using PDF-Box [70]). The overall precision of the induced labels has been evaluated by hiring experts to check randomly selected papers from both sources, resulting in an average precision of 96.8%. The dataset has been used by the authors to train a deep learning method and then deployed in Semantic Scholar to extract figures from 13 million scientific documents.

6.5 SciTSR

SciTSR [62] is a large-scale table structure recognition dataset, which contains 15,000 tables in PDF and their corresponding structure labels obtained from $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ source files. The authors identify also a subset of 3600 tables, called SciTSR-COMP, focused on complicated tables containing multiple spanning cells, i.e., a cell that spans over two or more rows/columns.

Differently from the general annotation scheme for $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ structured information (as in Sect. 3.2.1), no recoloring has been used here. In this dataset, the authors extracted table regions looking for table sections, by means of RegEx matching `\begin{table} ... \end{table}` to generate the PDF table file without recoloring. Finally, they extracted cells, columns, and rows parsing `\\ , & , \multirow{}` and `\multicolumn{}` commands, from which structure labels have been generated for Table Structure Recognition.

6.6 TableBank

TableBank [63] is an image-based table detection and recognition dataset built with a novel weak supervision from Word and $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ documents gathered on the Internet. It contains 417,234 high-quality labeled tables from documents taken from a variety of domains. TableBank was built exploiting the Word and $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ files' mark-up syntax to detect table structures. The Word files are in DOCX format and have been crawled from the web. Table bounding boxes are contained inside internal Office XML code. $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ files have been bulk-downloaded from arXiv considering scientific papers published in the period 2014–2018. The TableBank dataset focuses on table detection (TD) and table structure recognition (TSR).

For the Table Detection task, TableBank authors have created 417,234 labeled tables from the crawled documents as follows: tables are detected and annotated at both $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ and XML level. In the latter, each table is bounded by `<w:tbl>` and `</w:tbl>` tags; the bounding box can be colored still using XML in this case, since the Office XML code can be compiled back to the Word file. A simi-

lar approach is performed for $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ tables, where bounding boxes are colored using the `fcolorbox` command. In this way, authors were able to annotate the tables, in a manner similar to what is done for DeepFigures [47].

Concerning Table Structure Recognition, TableBank contains 145,463 training instances from Word and $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ documents converting them into HTML tags sequences, passing through the XML information. For Word files, the XML file is already provided. $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ files are firstly converted to XML using the LaTeXXML³ toolkit. In both cases, `<cell_y>` and `<cell_n>` are extracted considering cells with and without text, respectively.

Table Detection experiments have been conducted using Faster RCNN Model (ResNeXt-152) [71].

6.7 Table2Latex-450k

Table2Latex-450k [44] is a dataset of 450K table images for scientific table recognition, digitally rendered from $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ files crawled from arXiv articles between July 1991 and November 2016.

As in SciTSR[62], the tables are extracted by means of RegEx matching, compiled into PDF and then converted into cropped images. Of the resulting 940,279 tables from 298,168 articles, only 465,957 were kept since the others resulted in compilation errors. The dataset is separated into a training/validation/test sets at article level due to the similarity shared by tables within the same article.

Associated with PNG images, the dataset also contains a tokenized and normalized version of the $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ source code to reduce the ambiguities and make it easier for neural models to learn a consistent mapping. To do so, PlasTex⁴, a tool to transform LaTeX into XML-DOM, has been used.

6.8 PubTabNet

The PubTabNet dataset [64] gathers scientific documents from PMC, with annotations automatically generated by matching PDF and XML files. Differently from the previous ones, this dataset has three main advantages. (i) the tables are typeset by the publishers of over 6,000 journals in PMC, which offers considerably more diversity; (ii) cells are categorized into headers and body cells, which is important when retrieving information from tables; (iii) the format of the output is HTML, which can be directly integrated into web applications. In addition, tables in HTML format are represented as a tree structure.

The PDF and XML files are matched through the algorithm of [49], following the general schema described in

³ LaTeXXML.

⁴ PlasTex.

Sect. 3.2.2 to detect table regions and subsequently to convert them to images. Table regions are identified within the two sources and kept if the cosine similarity of the TF-IDF features of the two texts is larger than 90% and their length differs by less than 10%; rare tables with formulas or no frequent characters are removed as well. Finally, each table is rendered as image along with its HTML ground-truth annotation, without hyperlinks and metadata unused for the downstream task.

6.8.1 FinTabNet and PubXNet

Two datasets are proposed in [48]: (i) PubXNet, is an enhanced version of PubTabNet [64] adding table cells labels for a joint subset of PubTabNet and PubLayNet [49], (ii) the FinTabNet dataset is proposed to perform table detection and structure recognition on financial documents. Since HTML was already available, authors directly matched that information with the source PDF for automatic labeling.

6.9 PubTables-1 M

PubTables-1 M [65] is a large dataset composed of nearly 1 million tables extracted from scientific articles. This dataset contains detailed information about table structure and attempts to address the problem of over-segmentation using a novel canonicalization technique to obtain more consistent ground truth. PubTables-1 M has been developed to solve table extraction problems and its three main subtasks: table detection, table structure recognition, and functional analysis. It aims at overcoming the large variety of formats, styles, and structures that define tables in scientific papers, contrasting over-segmentation problems, and the need for more data. PubTables-1 M contains rich annotation information for all rows, columns, and tables' headers. The annotation procedure follows the one described in Sect. 3.2.2: to align XML and PDF the Needleman–Wunsch algorithm has been used, along with custom noise removal heuristics to ensure a certain level of quality. PubTables-1 M contains 947,642 tables annotated for structure recognition. For table detection, there are 575,304 fully annotated tables.

6.10 TabLeX

TabLeX [66] is a dataset of table images generated from scientific articles downloaded from arXiv, created working directly on L^AT_EX as described in Sect. 3.2.1. The papers belong to eight categories: Physics, Computer Science, Mathematics, Statistics, Electrical Engineering, Quantitative Biology, Quantitative Finance, and Economics. TabLeX consists of two subsets: one for table structure extraction and the other for table content extraction. The dataset contains images in different aspect ratios and in a variety of fonts

to ease the development of tools that are able to robustly extract tables from scientific documents. The dataset is organized in sets for table structure detection (TSD) and table content detection (TCD), and each set is further split into short (TSD-250, TCD-250) and long (TSD-500, TCD-500) tables.

6.11 TNCR

TNCR (Table Net detection and classification dataset [67]) is an image dataset of scanned documents. Data are collected as PDF files from various public access libraries, for a total of 875,026 PDF pages; these pages are then parsed manually to select only table pages, obtaining a subset of images to be used for training a Faster-RCNN to annotate candidate tables. Finally, the chosen ones are manually annotated using one of the following labels, based on table type: full lined, no lines, merged cells, partial lined, and partial lined merged cells.

7 Large-scale fully annotated datasets

In this section, we present large-scale datasets, which are summarized in Table 5. Large datasets for DLA have been collected taking advantage of publicly available collections of papers motivated by the increasing interest on document analysis exploiting machine learning. Researchers have been therefore able to start adopting deep learning techniques leveraging such huge datasets. In Fig. 4, these datasets are the most recent, intended to be used exploiting state-of-the-art methods for DLA. PubLayNet [49] is the first dataset of this type, counting 500k annotated PDF pages, and leveraging automatic annotation techniques. Following this thread, even larger datasets have been developed, looking for even more classes. Today, these datasets allow researchers to obtain the best results in DLA.

7.1 PubLayNet

PubLayNet [49] has been developed aiming to construct a dataset to be the largest ever created for DLA. PubLayNet is made up of publicly available scientific journals, specifically medical ones, in PDF and XML format, taken from PMC and labeled as described in Sect. 3.2.2 using PDFMiner [39].

Layout categories have been chosen to cover important document regions, such as paragraphs, titles, lists, images, and tables. Once all the PDF pages have been annotated, authors proceeded with quality control, discarding the noisy documents. Finally, documents were divided into training, development, and testing sets at the journal level to maximize the differences between sets and for a better evaluation of how well an object detector can generalize to unseen layouts and contents.

Table 5 Large-scale fully annotated document collections comparison

Name	Av.	Data	Source	Size	Tasks	Format	Annotations		
							Input	Procedure	Output
PubLayNet [49]	✓	P	DL	358k	DLA	PDF	XML/PDF	A	JSON
DocBank [45]	✓	P	DL	500k	DLA	PDF	PDF/L ^A T _E X	A	JSON
DAD [35]	✓	P	Various	5980	DLA	PDF	–	A	JSON
DocLayNet [72]	✓	P	Various	80k	DLA	PDF	PDF	M	JSON
SciBank [46]	✓	P	DL	74k	DLA	PDF	PDF/L ^A T _E X	A	CSV

Full description of tasks is given in Table 1. Annotations column describes how inputs are transformed to ground truth (output) using a specific procedure. “DL” refers to Digital Libraries. Data are always individual pages (P). Input blank lines mean that no additional files have been used for labeling

7.2 DocBank

DocBank [45] is a large-scale dataset built using a weak supervision approach. It enables models to integrate both the textual and layout information for downstream tasks. DocBank includes 500K document pages. The strength of the dataset is that it can be used for both NLP and Computer Vision tasks since its annotations are at the token level. DocBank documents are taken from arXiv, along with their L^AT_EX source files. The downloaded papers belong to many domains, such as Physics, Mathematics, and Computer Science; in general, DocBank is built to focus on scientific documents.

DocBank is a natural extension of the TableBank [63] dataset, and their annotation pipelines are quite similar. Working on L^AT_EX semantic as described in 3.2.1, twelve classes are identified: Abstract, Author, Caption, Equation, Figure, Footer, List, Paragraph, Reference, Section, Table, and Title.

PDFPlumber [43], a package based on PDFMiner, is then used to extract text and non-text elements from documents. Text is tokenized using white spaces; then, bounding boxes are defined as the rectangle with the upper left point of the first character and the right lower point of the last character of a token. Labels are determined on the color-mapping basis defined in the previous point.

Finally, the token-level dataset is converted to the region-level dataset, that is, tokens belonging to the same class are connected using the Breadth First Search algorithm to find connected components and then defining regions of tokens.

7.3 Dense Article Dataset

The Dense Article Dataset (DAD) [35] has been constructed to overcome the limited number of labeled classes in previous datasets. It covers research articles from multiple disciplines and captures the detailed elements of a research article in its original publication format. The dataset is composed of 450 open-access research articles from 14 different journals from

5 publishers (Elsevier, Springer, SAGEpublisher, Wiley, and IEEE).

The collected research articles are annotated with 43 classes that specialize three general categories: *front matters*, *body matters*, and *back matters*. 5980 pages are manually annotated with the Microsoft open-source tool "VOTT" [35]. The citations, due to their frequency (37k instances), are annotated by exploiting the output of Grobid [42] and then manually corrected. Annotations are saved as a JSON file with coordinates of the rectangular boxes.

7.4 DocLayNet

DocLayNet [72] is a dataset that consists of documents belonging to six distinct domains (Financial, Scientific, Patents, Manuals, Laws, and Tenders) collected from different free repositories. The main goal was to build an heterogeneous dataset to help trained detectors to produce more robust predictions without being domain specific. The dataset is composed of manual annotations (bounding boxes) for 80,863 PDF pages. Among these, 7059 carry two instances of manual annotations, and 1591 carry three, for 91,104 total annotation instances. Each bounding box belongs to one of the following 11 classes: Caption, Footnote, Formula, List-item, Page-footer, Page-header, Picture, Section-header, Table, Text, and Title. The authors avoided the use of scanned pages and looked for documents consisting of more than 10 pages, with many tables, figures, plots. Most documents in this dataset are written in English (95%).

DocLayNet is annotated using an augmented COCO format and the dataset is composed of three main parts: (i) original PDFs with text-cell coordinates and text (in JSON format); (ii) PNG files for each page (with size: 1025 × 1025 pixels); (iii) COCO annotations (in JSON format) for training, development, and test sets.

The authors considered manual annotations over automatic ones, also because structured source files were not available for the collected documents. DocLayNet manual annotations have been obtained using IBM CSS (Corpus Conversion Service) [34], a cloud-native platform that pro-

vides a visual annotation interface and allows for dataset inspection and analysis. To ensure homogeneity among different annotators, a 100-page annotation guideline has been written and administered to 32 selected experts. The whole annotation process lasted 6 months.

7.5 SciBank

SciBank [46] is a dataset containing 74,435 pages from 9635 scientific papers selected from arXiv. The annotation classes are 12: Caption, Equation, Inline equation, Figure, Keywords, Reference, Section, Subsection, Table, and Title. Annotations follow the general schema described in Sect. 3.2.1. The dataset consists of images and CSV files containing information about the page and regions on the page. For each annotation, in addition to the class label and the region coordinates, there is information about the paper, page number, size of the page, and corresponding PNG file.

Differently from Publaynet [49], TableBank [63], and DocBank [45], SciBank contains annotations for inline equation.

8 Datasets of other types of documents

Scientific publications have been widely used for DLA for two main reasons: (i) there is a significant amount of publicly available data on digital libraries that are easy to gather and usually come with additional structural information, such as XML or L^AT_EX that helps in the automatic annotation; (ii) scientific articles have a semantically rich non-trivial structure that is useful to train and test DLA algorithms. However, when it comes to transfer the information learned to other domains, it is nothing but trivial to keep the same performances due to high variability in layout and contents. Other benchmarks have been released to cope with this lack of information. First attempts to collect such data have been made by MediaTeam Document Database [73] and UvA Document Dataset [74].

The MediaTeam dataset contains 512 pages belonging to 198 documents of different categories (e.g., advertisements, manuals, and line drawing); the 58 “articles” are from newspapers. Originally all documents were stored in uncompressed TIFF format (total 7.2 GB). Due to the limited capacity of CD-ROM at the time, documents were distributed in compressed JPEG.

The UvA Dataset contains complex regions (the ground-truth is not limited to upright rectangles) and is focused on magazine pages with advertisements. Its most important feature is the inclusion of colored pages with respect to previous collections that were limited to gray-level or black and white document images.

One relevant task in DIAR is document classification that can be useful in DLA, since the identification of the specific document at hand can guide the analysis of different regions. In 2006, the RVL-CDIP (Ryerson Vision Lab Complex Document Information Processing) dataset [75] was released, containing 400,000 scanned gray-scale images with a great variety of contents in 16 different classes. It is still nowadays one of the largest collections of documents for DIAR related tasks. Some subsets have been extracted from RVL-CDIP during the years, trying to enhance the possible usage of this collection. Among others, we would like to cite Tobacco800 [76] and the annotated invoices gathered by Riba P. et al. [23]. The first one, composed of 1290 document images, is a realistic collection for document image analysis research since the images were scanned using a large variety of devices and tools over time. In addition, a significant percentage of Tobacco800 are consecutively numbered multi-page business documents, making it a valuable test for various content-based document image retrieval approaches. The second collection, instead, annotates a subsection of the RVL-CDIP invoices class for table detection and layout analysis, gathering some important information such as sender, receiver, and total amount.

Over several years, important competitions have been organized at the ICDAR conferences releasing benchmarks for TD and TSR that, nowadays, are still used to compare different methods. The most important ones to be cited are ICDAR 2013 [77] and ICDAR 2019 [78]. The first collection counts 156 tables crawled from two government sources *site:europa.eu* and *site:*.gov*. The organizers searched for tables keeping two pages after and before them, giving space for false positive detection. The second collection is more variable, both in layouts and contents. For the 2019 competition, 1639 tables are collected for TD and TR. Approximately half of them are handwritten, collected from the contribution of 23 different institutions, and the remaining are printed ones coming from different sources, such as scientific journals, forms, financial statements. The content varies between English and Chinese. Annotations and results are provided in XML format.

In 2020, IIT-AR-13K [79] is released. Manually annotated, it is one of the largest collections of business documents, collected from publicly available annual reports. There are five categories of objects that can be detected: tables, figures, natural images, logos, and signatures. The dataset presents also variability in content and layout: the reports are in different languages and scripts, while collected through ten years of twenty-nine different companies’ reports.

The most recent collection of this category is FinTab [80], a Chinese dataset for table extraction from financial documents that responds to the lack of diversity in tables of the currently existing datasets. It consists of 19 PDF files

with more than 1600 tables sourced from annual and semi-annual reports, debt financing, bond financing, collections of medium-term notes, short-term financing, and a prospectus. In total, the dataset adds up to 3329 pages with 2522 of them containing tables. The dataset has been manually reviewed and contains textual and structural ground-truth for the tables. The textual ground truth describes both the characters and string information while the structure ground truth is made related to cell and table line information; the ground truth is stored in JSON.

9 Impact of datasets

The release and widespread use of a research dataset can have a significant impact in the research on a given topic. In order to evaluate the impact of the datasets analyzed in this paper, we investigated three main features: i) the scientific impact, indirectly measured by the number of citations; ii) the dataset complexity, that is related to the quantity and quality of data and annotations, strictly connected to task to be tackled; iii) the most relevant methods evaluated on the datasets. The above factors are summarized in Table 1 and further discussed in this section.

While the impact can be measured with quantitative, albeit sometimes inaccurate, numerical values obtained from the number of citations to papers describing the datasets, the complexity is harder to be estimated. We therefore assess the dataset complexity by looking at some features of the data that are task dependent. The most important methods used are on the other hand identified by looking at research papers that describe techniques using the data. Impact and complexity can be inversely proportional: one dataset released several years ago could have pushed forward significantly the research in the past, but nowadays, given its small scale, it can be of little utility to train data-hungry machine learning models. On the opposite, a novel dataset can count on few citations, but could pose a new important challenge to the research community and can be useful to train novel models. Impact and complexity of the datasets are summarized in Table 6 and discussed in Sections 9.1 and 9.2. The tasks to be performed on a given dataset strongly affect the impact and complexity as well: for a fair comparison and discussion, a further division is therefore needed in our analysis. In reference to Table 1, four *macro*tasks are considered: Document Layout Analysis (DLA), Table Understanding (TD, TSR, TR, and TFA) and Figure Extraction (FE), and Document Classification for [75]. For each task, the most impactful datasets have been analyzed to identify the latest and most effective methods proposed as summarized in Table 6 and briefly discussed in Sect. 9.3. Some sample pages selected from the most impactful datasets per task are shown in Fig. 5.

In the rest of this section, we inspect the datasets on the basis of the three directions previously mentioned.

9.1 Impact

The impact of a dataset is evaluated in this work by the number of citations divided by the number of years passed from the dataset release to the time of writing (months are not taken into consideration). Citations are collected from Google Scholar and wrong references are usually not filtered out. This information is listed in the Impact columns in Table 6. Given a dataset, the identification of the most suitable paper to look for is not always easy. Nowadays, when a new dataset is released, it is often described in a research article that is also mentioned in the data repository. Users of the dataset are invited to cite this article when referencing to the data. For older datasets, it is less common to have an unique paper, or technical report, to be cited. For instance, papers discussing methods developed and tested on the UW-I/UW-III datasets referenced to the data by means of several citations [6, 7, 81–83] sometimes including more than one citation in the same article. In this case, the number of citations reported in Table 6 is obtained by a rough filtering of wrong results obtained with a Google search for "UW-III document." In our view, this is an underestimation of the number of actual references to the dataset. Similar considerations also apply to other earlier datasets like MARG [56], MediaTeam [73], and UvA [74]. Table 6 does not include information for SciBank [46] that is hosted in the IEEE DataPort, and has no citations yet.

9.2 Complexity

We summarize the data complexity by looking at two information about the dataset content: the number of different types of labels and the type of data annotated.

- *Number of labels*: depending on the task, a larger number of classes to be distinguished are usually more informative and more challenging. For instance, for DLA it is quite easy to distinguish between text and not-text regions, but it is harder to identify the semantic role of text regions. In the case of Table Understanding, rather than looking at the number of classes it is more informative to take note of the type of information provided in the ground truth that can be: Table Position (TP), Table Type (TT), Cell Position (CP), and Cell Type (CT);
- *Type of data (ToD) labeled*: can be scanned (S), digital-born (D), or a mixture of the two (M). Usually digital-born document is easier to be parsed (e.g., there are no OCR errors), while images and low-quality scans make the document analysis harder. More recent datasets

Table 6 Impact numbers from Google Scholar citations; years do not take into account months

Document layout analysis	Impact				Complexity		Methods
	Release	Years (Y)	Citations (C)	C/Y	# Labels	ToD	
PubLayNet [49]	2019	4	309	77.2	5	D	[27, 92, 93, 95, 96, 100, 102–104]
DocBank [45]	2020	3	109	<u>36.3</u>	12	D	[27, 93]
DocLayNet [72]	2022	1	15	15.0	11	D	[96]
Prima [29]	2009	14	143	10.2	10	S	[103]
IIT-AR-13K [79]	2020	3	29	9.7	5	D	[92]
DAD [35, 104]	2022	1	9	9.0	43	D	[95]
UW-1 & UW-III [6, 7]	1993	30	200	6.7	35	M	[86]
MARG [56]	2003	20	40	2.0	5	S	[105–107]
Table understanding							
TableBank [63]	2020	3	137	45.7	TP, CP, CT	D	[92, 103]
PubTabNet [64]	2020	3	128	<u>42.7</u>	CP, CT	D	[94]
PubXNet [48]	2021	2	84	42.0	TP, CP, CT	D	[48]
ICDAR-2019 [78]	2019	4	139	34.7	TP, CP	M	[48, 92, 102]
PubTables-1M [65]	2022	1	27	27.0	TP, CP, CT	D	[108]
ICDAR-2013 [77]	2013	10	234	23.4	TP, CP	M	[48, 62, 92]
TNCR [67]	2022	1	19	19.0	TP, TT	D	[67]
SciTSR [62]	2019	4	69	17.2	CP	D	[62, 80]
FinTab [109]	2021	2	28	14.0	TP, CP	D	[80]
Table2Latex [44]	2019	4	39	9.7	CP, CT	D	[44]
Marmot [59]	2012	11	87	7.9	TP	D	[92]
TabLeX [66]	2021	2	8	4.0	CP, CT	D	n/a
Figure extraction							
FigureSeer [37]	2016	7	173	24.7	5	D	[37, 97]
DeepFigures [47]	2018	5	122	<u>24.4</u>	3	D	[47]
CS-Large [61]	2016	7	117	16.7	3	D	[47]
CS-150 [60]	2015	8	86	10.7	3	D	[61, 84]
ScanBank [57]	2021	2	5	2.5	3	S	n/a
Document classification							
RVL-CDIP [75]	2015	9	350	38.9	16	S	[100–102]

Table Understandings labels can be: table type (TT), table position (TP), cell type (CT), cell position (CP). Type of Document (ToD) can be: digital-born (D), scanned (S), or a mixture of the two (M)

nearly always deal with digital-born articles while earlier datasets focus on scanned pages. This change is due from one side to the increasing number of applications dealing with digital-born data and from the other hand to the development of techniques for automatically generating the ground truth that rely on information in digital-born documents.

This information summarizes the datasets descriptions reported in this paper and is shown in the Complexity columns in Table 6.

9.3 Methods

In this section, we briefly overview the most common methods tested on the main datasets analyzed in this paper. We already summarized the main approaches for DLA in Sect. 2. In particular, the methods analyzed in the following, and listed in the Methods column in Table 6, have been found across the best results for the two most impactful datasets for each task. For the remaining datasets, we report some specific methods tested on the data.

Among others and with the aim of a brief and general overview without aspiring to cover all the methods of interest, we report the following approaches.

- *Heuristic methods* for DLA of scientific articles are related to well known algorithms based, for instance, on the RLSA algorithm, projection profile, and area Voronoi diagram (Sect. 2). Other common approaches are based on the analysis of connected components in bi-level images. For instance, in [84] connected components are used to extract figures and captions from PDFs of biomedical articles. Most methods in this group have been tested on the UW datasets [85–89].
- When moving the attention to machine learning-based approaches, *object detectors* in images have been extensively explored in computer vision in general, and in DIAR in particular, by adopting seminal models like Faster R-CNN [90] and Mask R-CNN [91] demonstrating significant advancements in object detection performance. Using a dual backbone ResNext-101 with deformable convolution, DOLNet [92] achieve remarkable results over seven different benchmarks. Differently, HiM [93] and VSR[27] are state-of-the-art methods for DocBank [45] and [49], respectively, and use a RPN in conjunction with textual embedding and a graph structure for final document objects refinements in a multi-modal fashion. Considering a specific object of the document page, the GTE framework [94], composed of a series of vision-based neural networks, uses two levels of depth at table and cell level to detect table boundaries and its structure. Other relevant methods in this category are DeepLabV3+ [95], TNCR [67], and FigureSeer [37] for Figure Extraction and analysis.
- Since document layouts are, or either contain, highly structured items, several researchers focused their efforts on *graph-based* techniques to tackle several tasks. Even if more traditional approaches (like the XY-tree and the Area Voronoi diagram) can be regarded as graph based, the widespread development of geometric deep learning (and in particular Graph Neural Networks) gave rise to applications for DLA of scientific articles. These algorithms usually propose a graph representation of the document, considering clusters of similar semantics (e.g., words or paragraphs) as nodes and using distance metrics as edge connections. Compared to other methods, these techniques still highly rely on the quality of the document preprocessing, restricting their expressive power. One recent work belonging to this area is [25], where the authors show the potentiality of a GNN over a dense network for a custom dataset composed of 0.5 million synthetic tables. More recently, other proposals such as GLAM [96], GFTE [80], and GraphTSR [62] have shown

remarkable results on datasets such as DocLayNet [72], FinTab [48], and SciTSR [62]. Concerning the analysis of chart images, in [97] a semantic segmentation network is used to predict probability maps for line styles. From this output, a graph is built and the line tracing is performed using linear programming. We also recently proposed a graph-based architecture to detect and recognize tables and their surroundings on a subset of [49] and [65] relying on a graph structure of the page [98].

- Since the first proposal of the transformer architecture [99], the DIAR community began to explore the use of *transformer-based architectures* for several tasks, including DLA of scientific articles. LayoutLMv3 [100] is the first multi-modal architecture not relying on pre-trained visual extractors that mixing visual, textual, and linear embeddings in a transformer-based model is able to achieve state-of-the-art performances over PubLayNet [49] and other important benchmarks. Belonging to the same family, it is worth to mention also DocFormer [101] and DiT [102] that, similarly to LayoutLMv3, achieve remarkable results specially on ICDAR table competitions [77, 78] and RVL-CDIP [75] datasets.

10 Discussion and open problems

In this survey, we explored the datasets available to perform DLA for scientific articles, mainly focusing on the annotation procedures. The first collections included few documents manually annotated. More recently deep learning techniques gained an increasing interest and achieved important results. To train these models, large quantity of annotated data are needed along with automatic annotation techniques to address a growing amount of data.

It is worth mentioning that, despite technological and algorithmic developments, there are datasets such as DocLayNet [72] that are manually annotated. This opens a reflection on the trade-off needed among quality, heterogeneity, and quantity of annotated data and between the use of manual or automatic annotation techniques to build datasets.

The use of automatic annotation techniques allows researchers to label large datasets, which, however, are often scarcely heterogeneous. On the other hand, manual annotations make it possible to deal with datasets composed of widely varied data that are unfortunately not comparable in quantity to those annotated automatically.

The two approaches come with different limitations: manual procedures are expensive in terms of money and time and do not scale well. On the other side, automatic annotations are not applicable on any application domain, since additional structured information is needed, but is not always available. Moreover, they are not highly reliable for the quality of annotations compared to manual ones and often require

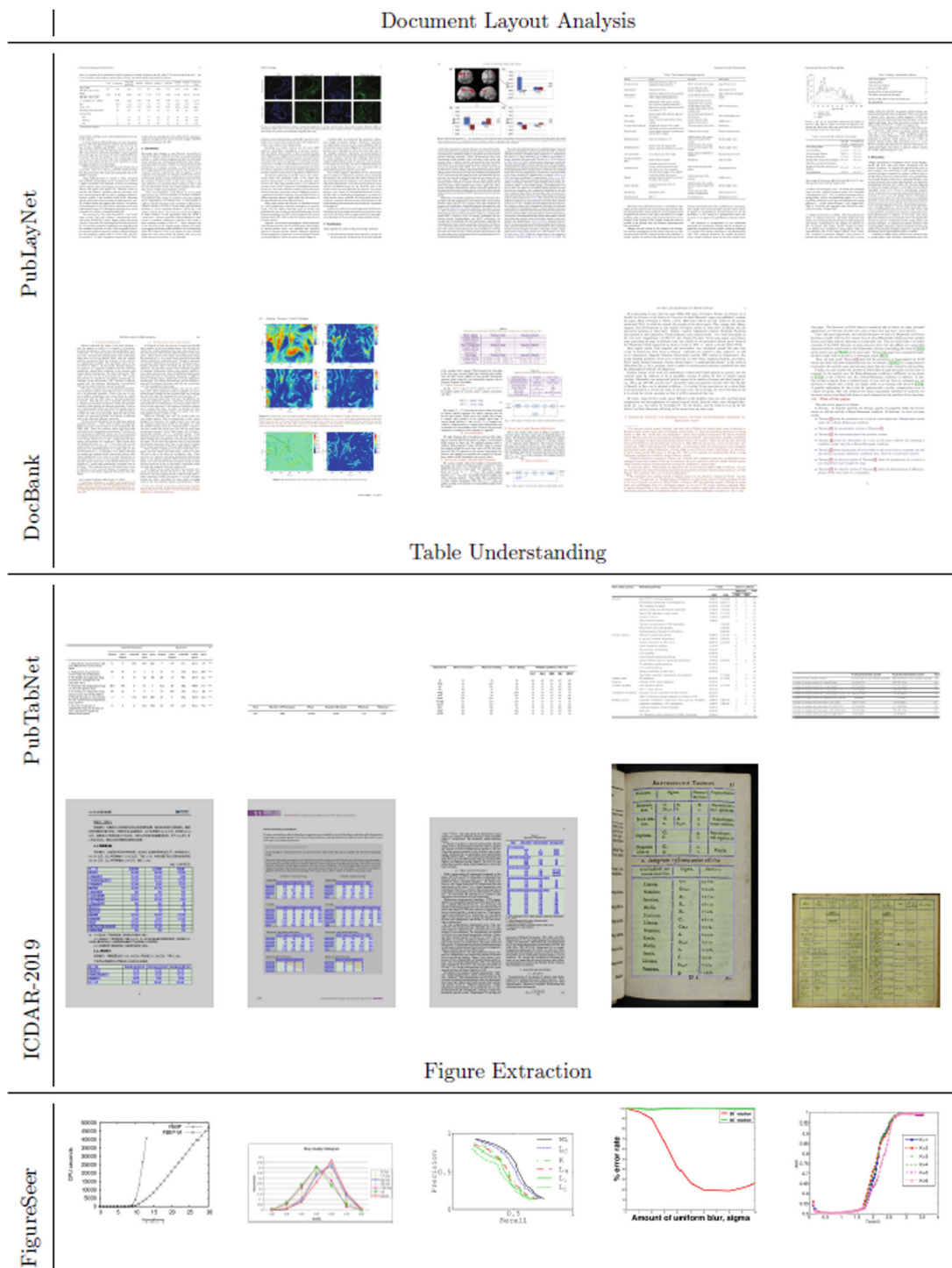


Fig. 5 Random samples selected from most impactful datasets per task (as described in Section 9). TableBank and PubMedX are skipped since they are an extension of DocBank and PubTabNet, containing similar

data. For figure extraction: FigureSeer contains directly annotated figures; DeepFigures examples are not shown since the authors provide only annotations

human supervision to check a representative sample of data. To cope with these limitations, generative procedures can be a viable alternative, as discussed in Sect. 3.3. An arbitrary large number of synthetically generated documents can help enhance the variability in contents and layout, providing also the corresponding annotations. On the other hand, the generation is restricted to the space of documents that the generator observed during training and there is still the risk of greatly diverting from reality. That is why generated datasets are usually not used in isolation, but combined with other collections.

Problems and limitations bounded to the collection of documents and their annotations are nothing but solved. In other fields, such as computer vision and natural language processing, new paths have started to be traced (e.g., [110]) drawing inspiration from these fields some novel directions could be explored also for the DIAR community. Among others, we believe that addressing the following points could help to overcome the aforementioned limitations:

- *gathering wider variety of documents*: for instance, written in different languages and scripts, to make architectures more robust over different data distributions. We focused mainly on DLA for scientific literature (nearly always in written in English) since most benchmarks belong to this area. However, there are a several open access collections of documents on Internet that can be crawled and unsupervised techniques could be explored to work with them;
- *create multi-page collections*: usually the community is focused on single pages for training, while in real case scenarios there is interest also for understanding multi-page documents, such as in administrative applications. From the training point of view, using multiple pages would result in a more informative process for the system in use, while on the application perspective models usage could be extended to real-scenarios documents;
- *developing novel generative methods*: unfortunately not all types of documents are accessible due to privacy issues, e.g., medical records or legal documents. Recently, the community has seen the growth of new generative models based on diffusion [111] and unsupervised pre-training [112]. In this direction could be interesting to apply similar techniques to generate an arbitrary large stream of documents simply given a text prompt as input, on top of techniques similar to I2DFormer [113].

11 Conclusions

In this survey, we analyzed and compared the most used datasets for layout analysis, with a focus on those related to scientific publications. In particular, we drew a taxonomy of

the most common techniques to annotate data, either manual or automatic. We also discussed techniques (e.g., based on generative models) that might be used for the development of datasets to train intelligent systems able to generalize over larger distributions of documents.

We provided detailed information on the datasets proposed over the years, with reference to the most widely used tools for PDF analysis and labeling. Our aim was to summarize the annotation procedures proposed and to provide a general overview of available datasets, helping scholars to choose the most suitable data for experimenting their techniques.

We compared all the datasets described in the paper by looking at their estimated impact and complexity. We also presented a brief overview of the most used methods designed and tested on these datasets. A numerical comparing of experimental results achieved by different approaches on common data is challenging due to different articles addressing different sub-tasks, variations in data splits, and diverse performance measures. Adding this information would lead to a fragmented presentation. We remind that this survey primarily focuses on data generation/annotation algorithms rather than data utilization techniques, and we believe that a deeper analysis about these techniques should be addressed in a dedicated paper.

Finally, we discussed advantages and drawbacks of various approaches used for labeling datasets, proposing new perspectives, and interesting open problems to be investigated in the future.

Acknowledgements Francesco Santoni has been supported by the PAD project (ID FSE: 291424) partially funded by Regione Toscana (bando Assegni di Ricerca 2021 - POR FSE 2014-2020).

Funding Open access funding provided by Università degli Studi di Firenze within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). <https://doi.org/10.1109/CVPR.2009.5206848>

2. Grother, P.J.: NIST special database 19. Handprinted forms and characters database, National Institute of Standards and Technology **10** (1995)
3. Deng, L.: The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Process. Mag.* **29**(6), 141–142 (2012)
4. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324 (1998)
5. Marinai, S.: Introduction to document analysis and recognition. In: Marinai, S., Fujisawa, H. (eds.) *Machine Learning in Document Analysis and Recognition. Studies in Computational Intelligence*, vol. 90, pp. 1–20. Springer, Berlin (2008). https://doi.org/10.1007/978-3-540-76280-5_1
6. Phillips, I.T., Chen, S., Haralick, R.M.: Cd-rom document database standard. In: *Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR'93)*, pp. 478–483. IEEE (1993)
7. Liang, J., Rogers, R., Haralick, R.M., Phillips, I.T.: Uw-isl document image analysis toolbox: An experimental environment. In: *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, vol. 2, pp. 984–988. IEEE (1997)
8. Nikolaidou, K., Seuret, M., Mokayed, H., Liwicki, M.: A survey of historical document image datasets. *Int. J. Doc. Anal. Recognit.* (2022). <https://doi.org/10.1007/s10032-022-00405-8>
9. Nguyen, T.T., Le, H., Nguyen, T., Vo, N.D., Nguyen, K.: A brief review of state-of-the-art object detectors on benchmark document images datasets. *Int. J. Doc. Anal. Recognit. (IJ DAR)* **26**, 1–19 (2023)
10. Binmakhshen, G.M., Mahmoud, S.A.: Document layout analysis: a comprehensive survey. *ACM Comput. Surv.* (2019). <https://doi.org/10.1145/3355610>
11. Kise, K.: Page segmentation techniques in document analysis. In: Doermann, D.S., Tombre, K. (eds.) *Handbook of Document Image Processing and Recognition*, pp. 135–175. Springer, Berlin, Heidelberg (2014). https://doi.org/10.1007/978-0-85729-859-1_5
12. Marinai, S.: Learning algorithms for document layout analysis. In: Rao, C.R., Govindaraju, V. (eds.) *Handbook of Statistics. Handbook of Statistics*, vol. 31, pp. 400–419. Elsevier (2013). <https://doi.org/10.1016/B978-0-444-53859-8.00016-3>
13. Wahl, F.M., Wong, K.Y., Casey, R.G.: Block segmentation and text extraction in mixed text/image documents. *Comput. Graph. Image Process.* **20**(4), 375–390 (1982)
14. O’Gorman, L.: The document spectrum for page layout analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **15**(11), 1162–1173 (1993)
15. Kise, K., Sato, A., Iwata, M.: Segmentation of page images using the area Voronoi diagram. *Comput. Vis. Image Underst.* **70**(3), 370–382 (1998)
16. Nagy, G., Seth, S.C.: Hierarchical representation of optically scanned documents. In: *1984 7th International Conference on Pattern Recognition*, pp. 347–349 (1984)
17. Strouthopoulos, C., Papamarkos, N.: Text identification for document image analysis using a neural network. *Image Vis. Comput.* **16**(12–13), 879–896 (1998)
18. Wu, C.-C., Chou, C.-H., Chang, F.: A machine-learning approach for analyzing document layout structures with two reading orders. *Pattern Recognit.* **41**(10), 3200–3213 (2008)
19. Cui, L., Xu, Y., Lv, T., Wei, F.: Document AI: benchmarks, models and applications. *arXiv preprint arXiv:2111.08609* (2021)
20. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **28** (2015)
21. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969 (2017)
22. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: pre-training of text and layout for document image understanding. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1192–1200 (2020)
23. Riba, P., Dutta, A., Goldmann, L., Fornés, A., Ramos, O., Lladós, J.: Table detection in invoice documents by graph neural networks. In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 122–127 (2019). <https://doi.org/10.1109/ICDAR.2019.00028>
24. Gemelli, A., Biswas, S., Civitelli, E., Lladós, J., Marinai, S.: Doc2graph: a task agnostic document understanding framework based on graph neural networks. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) *Computer Vision - ECCV 2022 Workshops*, pp. 329–344. Springer, Cham (2023)
25. Qasim, S.R., Mahmood, H., Shafait, F.: Rethinking table recognition using graph neural networks. In: *2019 International Conference on Document Analysis and Recognition, ICDAR 2019*, pp. 142–147 (2019). <https://doi.org/10.1109/ICDAR.2019.00031>
26. Jimeno Yepes, A., Zhong, P., Burdick, D.: ICDAR 2021 competition on scientific literature parsing. In: *International Conference on Document Analysis and Recognition*, pp. 605–617 (2021). Springer
27. 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: *ICDAR*, vol. 12821, pp. 115–130 (2021)
28. Valveny, E.: In: Doermann, D., Tombre, K. (eds.) *Datasets and Annotations for Document Analysis and Recognition*, pp. 983–1009. Springer, London (2014). https://doi.org/10.1007/978-0-85729-859-1_32
29. Antonacopoulos, A., Bridson, D., Papadopoulos, C., Pletschacher, S.: A realistic dataset for performance evaluation of document layout analysis. In: *10th International Conference on Document Analysis and Recognition, ICDAR 2009*, pp. 296–300 (2009). <https://doi.org/10.1109/ICDAR.2009.271>
30. Pletschacher, S., Antonacopoulos, A.: The page (page analysis and ground-truth elements) format framework. In: *2010 20th International Conference on Pattern Recognition*, pp. 257–260 (2010). <https://doi.org/10.1109/ICPR.2010.72>
31. Tzutalin, D.: LabelImg. GitHub Repository <https://github.com/heartexlabs/labelImg> (2015)
32. Everingham, M., Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. *Int. J. Comput. Vis.* **88**(2), 303–338 (2010). <https://doi.org/10.1007/s11263-009-0275-4>
33. Wada, K.: Labelme: Image Polygonal Annotation with Python. GitHub <https://github.com/wkentaro/labelme> (2018)
34. Staar, P.W.J., Dolfi, M., Auer, C., Bekas, C.: Corpus conversion service: a machine learning platform to ingest documents at scale. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '18*, pp. 774–782. Association for Computing Machinery, New York, NY, USA (2018). <https://doi.org/10.1145/3219819.3219834>
35. Markewich, L., Zhang, H., Xing, Y., Lambert-Shirzad, N., Jiang, Z., Lee, R.K., Li, Z., Ko, S.: Segmentation for document layout analysis: not dead yet. *Int. J. Doc. Anal. Recognit.* **25**(2), 67–77 (2022). <https://doi.org/10.1007/s10032-021-00391-3>
36. Dutta, A., Zisserman, A.: The via annotation software for images, audio and video. In: *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 2276–2279 (2019)
37. Siegel, N., Horvitz, Z., Levin, R., Divvala, S.K., Farhadi, A.: FigureSeer: parsing result-figures in research papers. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *Proceedings of ECCV 2016, Part VII. LNCS*, vol. 9911, pp. 664–680. Springer, Berlin (2016). https://doi.org/10.1007/978-3-319-46478-7_41

38. Jeremy B. Merrill, M.T. Manuel Aristarán: TabulaPDF. <https://tabula.technology/>
39. Shinyama, Y.: PDFMiner - Python PDF Parser. GitHub <https://github.com/pdfminer/pdfminer.six>
40. Pymupdf: Pymupdf/pymupdf: python bindings for mupdf's rendering library
41. Behnel, S., Faassen, M., Bicking, I.: LXML: XML and HTML with Python. Lxml <https://github.com/lxml/lxml> (2005)
42. GROBID. GitHub <https://github.com/kermitt2/grobid/> (2008–2021)
43. Jsvine: Jsvine/pdfplumber. <https://github.com/jsvine/pdfplumber>
44. Deng, Y., Rosenberg, D.S., Mann, G.: Challenges in end-to-end neural scientific table recognition. In: Proceedings of ICDAR 2019, pp. 894–901. IEEE. <https://doi.org/10.1109/ICDAR.2019.00148>
45. Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., Zhou, M.: DocBank: a benchmark dataset for document layout analysis (2020)
46. Grijalva, F., Parra, C., Gallardo, M., Santos, E., Acuña, B., Rodríguez, J.C., Larco, J.: SciBank: a large dataset of annotated scientific paper regions for document layout analysis. IEEE Dataport (2022). <https://doi.org/10.21227/2yex-bt23>
47. Siegel, N., Lourie, N., Power, R., Ammar, W.: Extracting scientific figures with distantly supervised neural networks. In: Chen, J., et al. (eds.) Proceedings of ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018, pp. 223–232. ACM, New York, NY, USA (2018). <https://doi.org/10.1145/3197026.3197040>
48. Zheng, X., Burdick, D., Popa, L., Zhong, P., Wang, N.X.R.: Global table extractor (GTE): a framework for joint table identification and cell structure recognition using visual context. In: Winter Conference for Applications in Computer Vision (WACV) (2021)
49. 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 (2019)
50. Ling, M., Chen, J., Möller, T., Isenberg, P., Isenberg, T., Sedlmair, M., Laramée, R.S., Shen, H.-W., Wu, J., Giles, C.L.: Document domain randomization for deep learning document layout extraction. In: Document Analysis and Recognition—ICDAR 2021, pp. 497–513. Springer, Berlin (2021). https://doi.org/10.1007/978-3-030-86549-8_32
51. Biswas, S., Riba, P., Lladós, J., Pal, U.: DocSynth: A layout guided approach for controllable document image synthesis. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) 16th International Conference on Document Analysis and Recognition, ICDAR 2021. Lecture Notes in Computer Science, vol. 12823, pp. 555–568 (2021). https://doi.org/10.1007/978-3-030-86334-0_36
52. Patil, A.G., Ben-Eliezer, O., Perel, O., Averbuch-Elor, H.: READ: recursive autoencoders for document layout generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 544–545 (2020)
53. Gupta, K., Lazarow, J., Achille, A., Davis, L., Mahadevan, V., Shrivastava, A.: LayoutTransformer: layout generation and completion with self-attention. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, pp. 984–994 (2021). <https://doi.org/10.1109/ICCV48922.2021.00104>
54. Pisaneschi, L., Gemelli, A., Marinai, S.: Automatic generation of scientific papers for data augmentation in document layout analysis. Pattern Recognit. Lett. **167**, 38–44 (2023). <https://doi.org/10.1016/j.patrec.2023.01.018>
55. Deng, D., Wu, Y., Shu, X., Wu, J., Xu, M., Fu, S., Cui, W., Wu, Y.: Visimages: a corpus of visualizations in the images of visualization publications. arXiv preprint [arXiv:2007.04584](https://arxiv.org/abs/2007.04584) (2020)
56. Thoma, G.: Ground truth data for document image analysis. In: Symposium on Document Image Understanding and Technology (SDIUT), pp. 199–205 (2003)
57. Kahu, S.Y., Ingram, W.A., Fox, E.A., Wu, J.: Scanbank: A benchmark dataset for figure extraction from scanned electronic theses and dissertations. In: Downie, J.S., et al. (eds.) ACM/IEEE Joint Conference on Digital Libraries, JCDL 2021, pp. 180–191. IEEE. <https://doi.org/10.1109/JCDL52503.2021.00030>
58. Antonacopoulos, A., Pletschacher, S., Bridson, D., Papadopoulos, C.: ICDAR 2009 page segmentation competition. In: 2009 10th International Conference on Document Analysis and Recognition, pp. 1370–1374 (2009). <https://doi.org/10.1109/ICDAR.2009.275>
59. Fang, J., Tao, X., Tang, Z., Qiu, R., Liu, Y.: Dataset, ground-truth and performance metrics for table detection evaluation. In: 2012 10th IAPR International Workshop on Document Analysis Systems, pp. 445–449. IEEE (2012)
60. Clark, C.A., Divvala, S.K.: Looking beyond text: Extracting figures, tables and captions from computer science papers. In: Caragea, C., et al. (eds.) Scholarly Big Data: AI Perspectives, Challenges, and Ideas, Papers from the 2015 AAAI Workshop. AAAI Technical Report, vol. WS-15-13. AAAI Press (2015). <http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10092>
61. Clark, C.A., Divvala, S.K.: PDFFigures 2.0: mining figures from research papers. In: Adam, N.R., et al. (eds.) Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, JCDL 2016, pp. 143–152. ACM. <https://doi.org/10.1145/2910896.2910904>
62. Chi, Z., Huang, H., Xu, H.-D., Yu, H., Yin, W., Mao, X.-L.: Complicated table structure recognition (2019) [arXiv:1908.04729](https://arxiv.org/abs/1908.04729) [cs.IR]
63. Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., Li, Z.: Tablebank: table benchmark for image-based table detection and recognition. In: Proceedings of The 12th Language Resources and Evaluation Conference, pp. 1918–1925 (2020)
64. Zhong, X., ShafieiBavani, E., Jimeno Yepes, A.: Image-based table recognition: data, model, and evaluation. In: European Conference on Computer Vision, pp. 564–580. Springer (2020)
65. Smock, B., Pesala, R., Abraham, R.: PubTables-1M: towards comprehensive table extraction from unstructured documents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4634–4642 (2022)
66. Desai, H., Kayal, P., Singh, M.K.: TabLeX: A benchmark dataset for structure and content information extraction from scientific tables. In: ICDAR, pp. 554–569. Springer (2021)
67. Abdallah, A., Berendeyev, A., Nuradin, I., Nurseitov, D.: TNCR: table net detection and classification dataset. Neurocomputing **473**, 79–97 (2022). <https://doi.org/10.1016/j.neucom.2021.11.101>
68. Fricke, S.: Semantic scholar. J. Med. Libr. Assoc. **106**, 23 (2018). <https://doi.org/10.5195/JMLA.2018.280>
69. Crowston, K.: Amazon mechanical Turk: a research tool for organizations and information systems scholars. In: Bhattacharjee, A., Fitzgerald, B. (eds.) Shaping the Future of ICT Research. Methods and Approaches, pp. 210–221. Springer, Berlin, Heidelberg (2012)
70. The Apache Software Foundation: Apache PDFBox (2012). <http://pdfbox.apache.org/>
71. 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, pp. 1492–1500 (2017)
72. Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A.S., Staar, P.W.J.: DocLayNet: A large human-annotated dataset for document-layout segmentation. In: Zhang, A., Rangwala, H. (eds.) KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 3743–3751 (2022). <https://doi.org/10.1145/3534678.3539043>

73. Sauvola, J., Kauniskangas, H.: Mediateam document database ii, a cd-rom collection of document images. University of Oulu, Finland (1999)
74. Todoran, L., Worring, M., Smeulders, A.W.M.: The UvA color document dataset. *Int. J. Doc. Anal. Recognit.* **7**(4), 228–240 (2005). <https://doi.org/10.1007/s10032-004-0135-2>
75. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR)
76. Lewis, D., Agam, G., Argamon, S., Frieder, O., Grossman, D., Heard, J.: Building a test collection for complex document information processing. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 665–666 (2006)
77. Göbel, M., Hassan, T., Oro, E., Orsi, G.: ICDAR 2013 table competition. In: 2013 12th International Conference on Document Analysis and Recognition, pp. 1449–1453. IEEE (2013)
78. Gao, L., Huang, Y., Déjean, H., Meunier, J.-L., Yan, Q., Fang, Y., Kleber, F., Lang, E.: ICDAR 2019 competition on table detection and recognition (cTDaR). In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1510–1515. IEEE (2019)
79. Mondal, A., Lipps, P., Jawahar, C.: IIIT-AR-13K: a new dataset for graphical object detection in documents. In: International Workshop on Document Analysis Systems, pp. 216–230. Springer (2020)
80. Li, Y., Huang, Z., Yan, J., Zhou, Y., Ye, F., Liu, X.: GFTE: graph-based financial table extraction. In: International Conference on Pattern Recognition, pp. 644–658. Springer (2021)
81. Phillips, I.T., Chen, S., Ha, J., Haralick, R.M.: English document database design and implementation methodology. In: Proceedings of the 2nd Annual Symposium on Document Analysis and Retrieval, pp. 65–104 (1993)
82. Phillips, I.T.: User's reference manual for the UW English/technical document image database iii. UW-III English/technical document image database manual (1996)
83. Guyon, I., Haralick, R.M., Hull, J.J., Phillips, I.T.: Data sets for OCR and document image understanding research. In: Handbook of Character Recognition and Document Image Analysis, pp. 779–799. World Scientific (1997)
84. Li, P., Jiang, X., Shatkay, H.: Figure and caption extraction from biomedical documents. *Bioinformatics* **35**(21), 4381–4388 (2019)
85. Liang, J., Phillips, I.T., Haralick, R.M.: A statistically based, highly accurate text-line segmentation method. In: Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR'99 (Cat. No. PR00318), pp. 551–554. IEEE (1999)
86. Shafait, F., Keysers, D., Breuel, T.: Performance evaluation and benchmarking of six-page segmentation algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(6), 941–954 (2008)
87. Nayef, N., Ogier, J.-M.: Text zone classification using unsupervised feature learning. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 776–780. IEEE (2015)
88. Mao, S., Kanungo, T.: Automatic training of page segmentation algorithms: an optimization approach. In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, vol. 4, pp. 531–534. IEEE (2000)
89. Kanungo, T., Haralick, R.M., Baird, H.S., Stuezle, W., Madigan, D.: A statistical, nonparametric methodology for document degradation model validation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(11), 1209–1223 (2000)
90. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 28. Curran Associates, Inc. (2015). https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf
91. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
92. Mondal, A., Agarwal, M., Jawahar, C.: Dataset agnostic document object detection. *Pattern Recognit.* **142**, 109698 (2023)
93. Canhui, X., Yuteng, L., Cao, S., Honghong, Z., Hengyue, B., Yinong, C.: Him: hierarchical multimodal network for document layout analysis. *Appl. Intell.* **53**, 1–13 (2023)
94. Zheng, X., Burdick, D., Popa, L., Zhong, X., Wang, N.X.R.: Global table extractor (GTE): a framework for joint table identification and cell structure recognition using visual context. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 697–706 (2021)
95. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
96. Wang, J., Krumdick, M., Tong, B., Halim, H., Sokolov, M., Barda, V., Vendryes, D., Tanner, C.: A graphical approach to document layout analysis. In: International Conference on Document Analysis and Recognition, pp. 53–69. Springer (2023)
97. Kato, H., Nakazawa, M., Yang, H.-K., Chen, M., Stenger, B.: Parsing line chart images using linear programming. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2109–2118 (2022)
98. Gemelli, A., Vivoli, E., Marinai, S.: Graph neural networks and representation embedding for table extraction in pdf documents. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 1719–1726 (2022). <https://doi.org/10.1109/ICPR56361.2022.9956590>
99. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc. (2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
100. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: pre-training for document AI with unified text and image masking. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 4083–4091 (2022)
101. Appalaraju, S., Jasani, B., Kota, B.U., Xie, Y., Manmatha, R.: Docformer: end-to-end transformer for document understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 993–1003 (2021)
102. Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., Wei, F.: Dit: self-supervised pre-training for document image transformer. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 3530–3539 (2022)
103. Shen, Z., Zhang, R., Dell, M., Lee, B.C.G., Carlson, J., Li, W.: Layoutparser: a unified toolkit for deep learning based document image analysis. In: Document Analysis and Recognition—ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16, pp. 131–146. Springer (2021)
104. Ma, C., Lin, W., Sun, L., Huo, Q.: Robust table detection and structure recognition from heterogeneous document images. *Pattern Recognit.* **133**, 109006 (2023). <https://doi.org/10.1016/j.patcog.2022.109006>
105. Rusiñol, M., Frinken, V., Karatzas, D., Bagdanov, A.D., Lladós, J.: Multimodal page classification in administrative document

- image streams. *Int. J. Doc. Anal. Recognit.* **17**(4), 331–341 (2014). <https://doi.org/10.1007/s10032-014-0225-8>
106. Zulfiqar, A., Ul-Hasan, A., Shafait, F.: Logical layout analysis using deep learning. In: 2019 Digital Image Computing: Techniques and Applications (DICTA), pp. 1–5. IEEE (2019)
 107. Liu, L., Wang, Z., Qiu, T., Chen, Q., Lu, Y., Suen, C.Y.: Document image classification: progress over two decades. *Neurocomputing* **453**, 223–240 (2021)
 108. Smock, B., Pesala, R., Abraham, R.: Aligning benchmark datasets for table structure recognition. In: Fink, G.A., Jain, R., Kise, K., Zanibbi, R. (eds.) *Document Analysis and Recognition—ICDAR 2023*, pp. 371–386. Springer, Cham (2023)
 109. Zheng, X., Burdick, D., Popa, L., Zhong, X., Wang, N.X.R.: Global table extractor (GTE): a framework for joint table identification and cell structure recognition using visual context. In: *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3–8, 2021*, pp. 697–706. IEEE (2021). <https://doi.org/10.1109/WACV48630.2021.00074>
 110. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. *CoRR* [arXiv:2103.00020](https://arxiv.org/abs/2103.00020)
 111. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: *International Conference on Machine Learning*, pp. 8821–8831. PMLR (2021)
 112. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: *ICML (2021)*
 113. Naeem, M.F., Xian, Y., Gool, L.V., Tombari, F.: I2DFormer: learning image to document attention for zero-shot image classification. *Adv. Neural Inf. Process. Syst.* **35**, 12283–12294 (2022)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.