Improving large-scale DLA datasets through semantic validation and relation-aware multimodal LLMs

Lorenzo Massai[1,*], Simone Marinai[1]

[1]DINFO - University of Florence, via S. Marta, 3, Firenze, Italy

Abstract

Machine learning models for Document Layout Analysis (DLA) rely on large-scale datasets such as DocBank and many others. However, automatic generation of these datasets from PDF or LaTeX sources often introduces errors in region annotations, including missing captions, misplaced regions and overlapping elements. Such inconsistencies negatively impact both training quality and evaluation reliability, propagating noise or bias through downstream models.

In this work we present a preliminary version of a framework that improves the most widespread DLA datasets quality by assessing and correcting layout coherence in scholarly documents. The proposed system builds a relational model of article structure, integrating geometric (e.g. *near*, *below*, *lower page*), sequential (e.g. *follows*, *precedes*), and semantic (e.g. *disjoint*, *subclass*) relations among layout elements. These relations can be learned directly from document data, enabling automatic detection and correction of violations, including semantic and hierarchical inconsistencies.

Error typologies that can be detected include those derived from class reading order (e.g. *title* → *author* → *abstract*) and proximity, which are found to be the most frequent. The detection of such violations highlights missing or misplaced components and guide automated correction that we perform through LLMs.

The resulting datasets provide a more reliable ground truth with improved structural and semantic coherence for training and evaluating document segmentation and understanding models, with reduced noise and bias inherent in automatically generated annotations.

Keywords

document layout analysis, dataset cleaning, large language models, multimodal feature extraction, ontology learning, scholarly document processing, visual document understanding

1. Introduction

Recent advances in Document Layout Analysis (DLA) are driven by the availability of large annotated datasets such as PubLayNet and DocBank, which provide millions of labeled pages extracted from scientific literature. These resources have enabled deep learning architectures for detecting layout elements such as titles, paragraphs, tables, and figures. Nevertheless, the automatic procedures used to build such datasets, which are mainly based on XML or LaTeX extraction, introduce a substantial amount of noise. Errors such as captions without figures or tables, overlapping text blocks, or missing elements are frequent and propagate through model training, ultimately reducing accuracy and interpretability. Despite the scale of these datasets, their internal logical coherence has received little systematic analysis.

This work addresses this gap by introducing a relation based post processing system that checks and corrects layout inconsistencies in large-scale scholarly datasets. The main idea is that layout elements within a scholarly article are not independent objects but participate in structured relations: *geometric*, *directional*, and *semantic*, which can be modeled and validated. By verifying these relations, it becomes possible to automatically identify annotation errors and correct them, improving dataset quality without reannotation. We define four main categories of relations:

- **Proximity relations** (*near*) express geometric adjacency between layout elements. They are primarily used to detect misplaced or missing components such as captions, figures, and tables.

IRCDL 2026: 22nd Conference on Information and Research Science Connecting to Digital and Library Science, February 19-20, 2026, Modena, Italy

*Corresponding author.

lorenzo.massai@unifi.it (L. Massai); simone.marinai@unifi.it (S. Marinai)

0000-0002-8252-0549 (L. Massai); 0000-0002-6702-2277 (S. Marinai)

- **Directional relations** (*below*, *upon*) describe relative geometric positioning along a given axis, e.g. that figures are typically below captions and footers are typically below all other layout classes.
- **Semantic relations** (*disjoint*, *subclass*, *equivalent*) capture ontological constraints between layout classes, preventing incompatible overlaps and enforcing hierarchical consistency.
- **Reading order relations** (*follows*, *precedes*) describe the logical progression of textual components across the first page of a scholarly article. Typical ordered chains such as *title* $\rightarrow$*author* $\rightarrow$ *abstract* define the expected hierarchy of presentation and can reveal missing or misplaced metadata blocks.

Proximity and directional are **visual** relations and capture the expected relative arrangement of elements. Our violation detection strategy aims to check the consistency of the layout classes for which such patterns are statistically frequent, which ensure robustness by focusing on recurrent and well supported configurations (e.g. footers are expected at the bottom of the page and detecting other layout elements positioned below a bounding box labeled as *footer* is a strong indicator of an annotation error). By modeling and validating these dependencies, our strategy can automatically detect anomalies such as failures to recognize the title bounding box, the author section and the abstract. This allows us to select page segments where most likely errors occur, subsequently correcting them through a multimodal LLM to ensure that the learned geometric, semantic, and relative order rules are preserved.

Unlike previous hard coded validators, our strategy involves learning relation patterns directly from document images and document ontologies, and explores the use of LLMs for corrections. Geometric and semantic embeddings are derived from co-occurrence statistics across annotated corpora, while frequent relational patterns (e.g. caption-figure proximity, title $\rightarrow$ author $\rightarrow$ abstract order) are made explicit. Once learned, these relations are used to check annotations for inconsistencies, categorize them, and apply LLM based corrective actions such as removing incorrect bounding boxes.

We evaluate the coherence of annotations on a small subset of a major dataset for layout analysis (DocBank), applying our strategy as post processing. Our preliminary analysis on DocBank reveals the presence of various structural annotation violations, suggesting that targeted correction strategies could improve the visual and semantic consistency of resources for downstream applications.

Beyond our experiments, the proposed framework provides a general methodology for assessing datasets reliability in scholarly document understanding. By exploiting layout relations as measurable constraints, our strategy establishes a bridge between structural analysis and dataset quality assurance. Researchers leveraging large document datasets can thus benefit from more coherent training material, supporting reproducible, explainable, and semantically grounded document analysis systems.

2. Related work

Document segmentation aims to identify and classify the structural regions of a page (titles, headings, paragraphs, figures, tables, captions, equations, references); this is a foundational step in scholarly document understanding. Over the years, methods and datasets have coevolved: early heuristics and connected component labeling were gradually replaced by learning based detectors and large, automatically annotated corpora that enable deep models to generalize across styles and publishers [1]. In parallel, the NLP community has developed pipelines that recover logical and bibliographic structure directly from PDFs or LaTeX sources, often combining OCR, layout analysis and TEI-XML conversion [2]. These pipelines rely on semantic processing techniques, including named entity recognition [3], dependency parsing [4], semantic labeling [5], and semantic relation extraction [6, 7], to transform raw document content into structured knowledge.

Classical approaches for document segmentation adopt bottom up grouping of connected components (RLSA [8], Docstrum [9]) or top down partitioning (X-Y cut [10]), often augmented with hand crafted features to classify regions. Statistical classifiers (SVMs, early neural networks) improve robustness but are limited by dataset scale and feature design. Logical structure recovery is usually performed exploiting visual features, with a few exceptions that rely more directly on text and font information [11, 12, 13], making use of transformers and transfer learning for mapping sequences of text blocks into hierarchical structures.

Deep learning approaches cast document segmentation as object detection or instance segmentation, using Faster/Mask R-CNN variants and one-stage detectors for region recognition. Token-aware multimodal transformers combining OCR, token boxes, and visual features improve semantic labeling, while graph models and GNNs capture spatial and logical relations. Systems such as DocParser [14] and DocStruct [15], used in hierarchical frameworks like HRDOC [16], extract regions and relations but are typically limited to single-page context.

Whole-document analysis further increases the problem complexity, requiring the modeling of relations between elements in different pages or semantic regions that span multiple pages. Demirtas et al. [17] tackle this issue by extracting inter-page dependencies in the form of triples using a neural dependency parser. More broadly, the complexity of document structure induction from PDFs has led the NLP community toward manual annotation of documents [18, 19], sometimes complemented with data augmentation strategies [20] to mitigate the limited size of carefully labeled design corpora.

Recent research emphasizes a data-centric AI perspective, where dataset quality and annotation consistency drive performance [21]. A fundamental aspect of this line of work is the availability and nature of datasets: small, manually annotated scanned corpora are more accurate, while being more subject to overfit; midsize task specific collections like table and figure datasets are not versatile and are hard to integrate with more general purpose collections; very large, automatically labeled digital born datasets where LaTeX/XML sources enable weak supervision are the most widespread, but are more subject to annotation errors and inconsistencies that a human annotator would typically avoid. Each category drives different design choices: large detection models require vast amounts of data and therefore rely on automatically constructed datasets, whereas detailed taxonomies and fine grained semantic classes are better supported by carefully curated manual annotations.

Effective solutions for layout extraction can also be found in commercial and open source systems. Grobid [22] is widely regarded as one of the best tools for extracting bibliographic data from scholarly articles [23], and supports multipage analysis with structured outputs for *title*, *authors*, *abstract*, *paragraphs*, *section headings* and more. Recent pipelines such as MinerU [24] and [25, 26, 27] are able to detect a rich set of classes (*titles*, *body paragraphs*, *images*, *image captions*, *tables*, *table captions*, *inline formulas*, *formula labels*), and discard *headers*, *footers*, *page numbers*, and *notes*. Among commercial solutions, Amazon Textract [28] offers high quality extraction of layout classes such as *title*, *header*, *footer*, *section title*, *page number*, *list*, *figure*, *table*, *formula* and *text*.

2.1. Datasets

Building on [1], we review the most used datasets and annotation criteria that are relevant for the segmentation of scientific articles.

Small-scale, manually annotated University of Washington collections provided early scanned pages and region ground truth that established evaluation protocols [29]. MARG (medical title pages) [30] and Prima [31] offer high quality XML annotations for scanned documents and are widely used in earlier segmentation evaluations. More recently, manually annotated corpora such as those in [18, 19] have been used to support fine grained layout analysis and to benchmark PDF extraction pipelines.

Task focused (tables/figures) Resources like Marmot [32], CS-150 [33] and CS-Large [34], FigureSeer [35], DeepFigures [36], and ScanBank [37] are datasets that concentrate on table and figure detection/extraction. They help develop specialized pipelines (table detection, table structure recognition, figure/caption pairing) and highlight domain shifts between digital born and scanned pages. Additional table focused corpora, such as those used in [38, 39], support evaluation of table localization and structure recognition under varied layout conditions.

Large-scale, automatically annotated PubLayNet [40] and DocBank [41] are the most used large-scale resources for page segmentation. PubLayNet provides hundreds of thousands of pages with region labels (text, title, list, figure, table) enabling robust detector training. DocBank (arXiv + LaTeX) supplies token level annotations bridging vision and NLP. TableBank [42], Table2LaTeX [43], PubTabNet [44], PubTables-1M [45], TabLeX [46], SciTSR [47] and related collections focus on table detection and table structure recognition at scale, leveraging LaTeX or XML to derive bounding boxes and structural

labels. DeepFigures [36] and PubTabNet [44] show that automatic alignment of XML/LaTeX with PDF can produce large corpora when carefully filtered, although residual alignment errors and labeling inconsistencies remain.

Cross domain and multiclass DocLayNet [48] and Dense Article Dataset (DAD) [49] provide manually verified, cross domain annotations with rich taxonomies (multiple layout classes) to improve generalization across document types. SciBank [50] and ScanBank [37] extend coverage to scientific subdomains and scanned theses, emphasizing variation in sources and acquisition quality. HRDOC [16] shifts the focus from flat layouts to hierarchical document structure, converting multipage documents into tree representations that support evaluation of multipage reconstruction methods.

Textual/large metadata (limited geometry) S2ORC [51] and UnarXive [52] are large parsed corpora (JSON/LaTeX) providing rich textual and citation metadata for millions of papers, widely used for retrieval, pretraining and weak supervision, but generally lacking geometric annotations unless paired with PDFs. S2ORC, built on Semantic Scholar [53], integrates citation graphs via Microsoft Academic Graph/S2AG [54], while UnarXive [55, 52] extracts text and citations from LaTeX using LaTeXML, Tralics, and Grobid, and incorporates metadata from OpenAlex [56]. These resources support high quality text and citation extraction and have been used to build ground truths for evaluating PDF conversion tools [57]. Article Regions Dataset (ARD) [58, 59] provides geometric annotations for single page documents, while other datasets [60, 61] adopt JSON based structural representations, highlighting the importance of flexible formats in large-scale scholarly document processing.

2.2. Annotation strategies and procedures

Annotation strategies fall into three classes: *manual* (high fidelity, low scale), *automatic* (LaTeX/XML alignment, high scale, possible noise), and *generative/synthetic* (layout generators, domain randomization, useful for augmentation). Automatic pipelines exploit LaTeX or XML tags to mark regions and then extract colored masks; alignment methods use string matching (Levenshtein, TF-IDF) to pair logical XML segments with PDF text. These methods are used to build datasets such as PubLayNet [40], DocBank [41], TableBank [42], PubTabNet [44] and PubTables-1M [45]. Manual tools (Aletheia, VIA, PAGE) are still used for quality control and domain specific corpora like DocLayNet and DAD, as well as in more recent efforts for hierarchical structure annotation [18, 19]. Synthetic generation and augmentation strategies [20] further enrich the diversity of training data when manual annotation is too costly.

The most widely used resources for training modern layout models are high scale, automatically constructed datasets. While these corpora are indispensable for scaling deep learning approaches, their annotation pipelines inevitably introduce geometric and semantic noise that is rarely revisited after the dataset release. A systematic post processing or consistency checking of these resources before they are used to train downstream models is still lacking. This work directly addresses this gap by proposing an automatic relation-driven refinement framework that exploits geometric, reading order and semantic relations among layout elements to detect and correct incoherencies in existing large-scale datasets.

2.3. Open challenges in document segmentation

In this scenario, our work targets several issues in scholarly document segmentation, including multipage versus single page analysis, taxonomy interoperability, annotation noise, and reading order extraction. By analyzing and normalizing these aspects, we are able to produce datasets that are more consistent, complete, and reliable, enabling improved training of models and better downstream performance in tasks such as information extraction and document parsing. Our research focuses on:

- **Single page/whole document segmentation:** most large corpora and models are page centric; handling multipage structures (multipage tables, distributed figure references, cross page sections) remains challenging [16, 17]. Multipage representations in datasets such as HRDOC [16] are still the exception rather than the rule.
- **Scanned documents:** digital born vs. scanned pages show strong differences in effectiveness; models trained on PubLayNet and DocBank may fail on scanned or older formats [1, 37, 50].

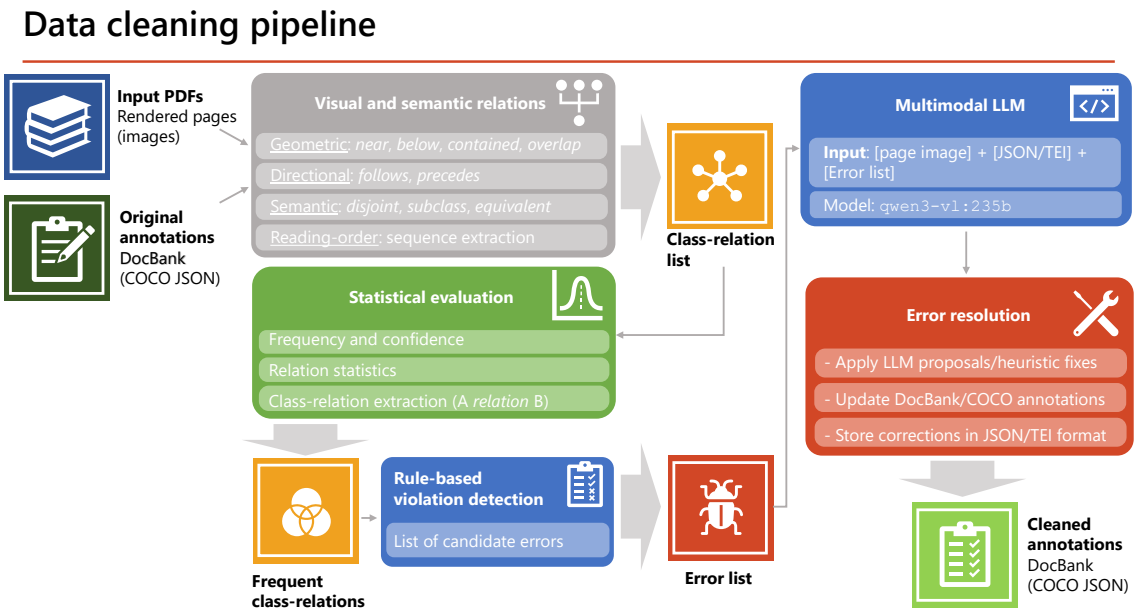


Figure 1: Post processing pipeline: from error detection to error resolution

- **Taxonomy interoperability:** inconsistent class sets hinder transfer and dataset merging [1, 48, 24, 28]. Aligning dataset label spaces with document ontologies is an open research direction.
- **Annotation noise and calibration:** automatic annotation pipelines need robust filtering and confidence estimation to bound label noise [36, 44, 57]. Large text corpora such as S2ORC and UnarXive [51, 52] lack consistent geometric grounding, making hard to reason about layout and content.
- **Reading order extraction:** most datasets lack explicit reading order annotations, forcing models to infer sequences from geometry or OCR flow [62], [63]. Common patterns (e.g. *title* → *author* → *abstract* or *figure* → *caption*) are therefore unstable, making it difficult to detect misplaced or missing front matter elements without dedicated post processing.

While several works address document layout analysis through rule based post processing or consistency checks embedded in downstream pipelines, these methods are commonly not designed as general purpose tools for dataset inspection and correction. To the best of our knowledge, there is currently no prior work that explicitly addresses automatic inspection and correction of document layout analysis datasets at the annotation level. In this sense, the present work constitutes a first step toward a dedicated framework for dataset-level validation and correction in DLA.

3. Framework definition

The proposed framework (Figure 1) operates as a pipeline for post processing and cleaning large document layout datasets such as DocBank. The architecture is modular and integrates components for XML extraction, visual relation detection, statistical evaluation, and dataset correction. Each module of the pipeline serves one of the project aims: learning and verifying the relations between elements, checking coherence, and generating a corpus that is as error-free as possible.

3.1. Relation extraction from page images

The spatial and relational analysis of document elements is analyzed parsing the JSON COCO annotations. Each annotated page is loaded together with its bounding boxes and categories. The system computes pairwise relations between elements through geometric properties such as distance, containment, and relative positioning. The result of this computation is stored as axioms: [classA, *relation*, classB] and aggregated across the dataset.

Proximity relation definition (near)

Proximity relations (Figure 2) such as $near$ express geometric adjacency and are used to detect missing or misplaced captions, figures, and tables.

Geometric and logical rules:

1. **Definition of Proximity ($near$):** Two elements (a, b) are considered $near$ if the distance between at least one pair of their bounding box edges is less than an estimated threshold. This value is approximated as the median distance between table/caption and figure/caption annotations learned from the DocBank dataset, and is found to be ≈ 30 pixels. Since the proposed framework is designed to operate at scale on large corpora containing exclusively scholarly documents (like DocBank), these confidence measures are able to capture highly recurrent layout patterns in this domain (e.g. Figure 3). Consequently, rare or borderline configurations are excluded by the filtering criteria derived from these statistics.

$$(a, b) \in near \Leftrightarrow distance(a, b) < 30px$$

$$(a, b) \in distant \Leftrightarrow (a, b) \notin near$$

2. **Hard proximity constraint ($hard\_near$):** A pair of classes (A, B) satisfies the hard proximity constraint if: $(\forall a \in \text{ClassA} : \exists b \in \text{ClassB} : (a, b) \in near) \wedge (\forall b \in \text{ClassB} : \exists a \in \text{ClassA} : (a, b) \in near)$.

3. **Error rule:** An error occurs if a hard proximity constraint is not satisfied (i.e., elements that should be near are distant).

$$(\text{ClassA}, \text{ClassB}) \in hard\_near \wedge a \in \text{ClassA} \wedge b \in \text{ClassB} \wedge (a, b) \in distant \Rightarrow \text{ERROR}(a, b)$$

Directional relation definition (below)

Directional relations (Figure 2) such as *above* and *below* compare the vertical positions of layout regions on the page, supporting error detection including misplaced titles or footers (e.g. Figure 4). This relation is more informative on the first page of the article, hence the analysis is restricted to that page.

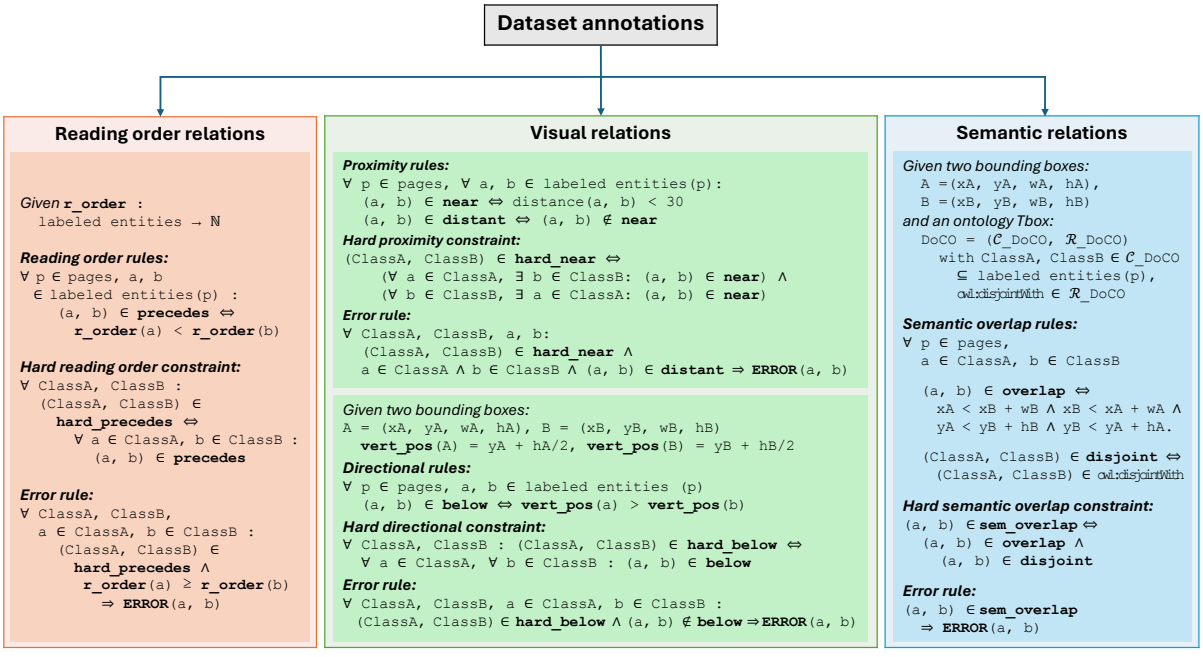
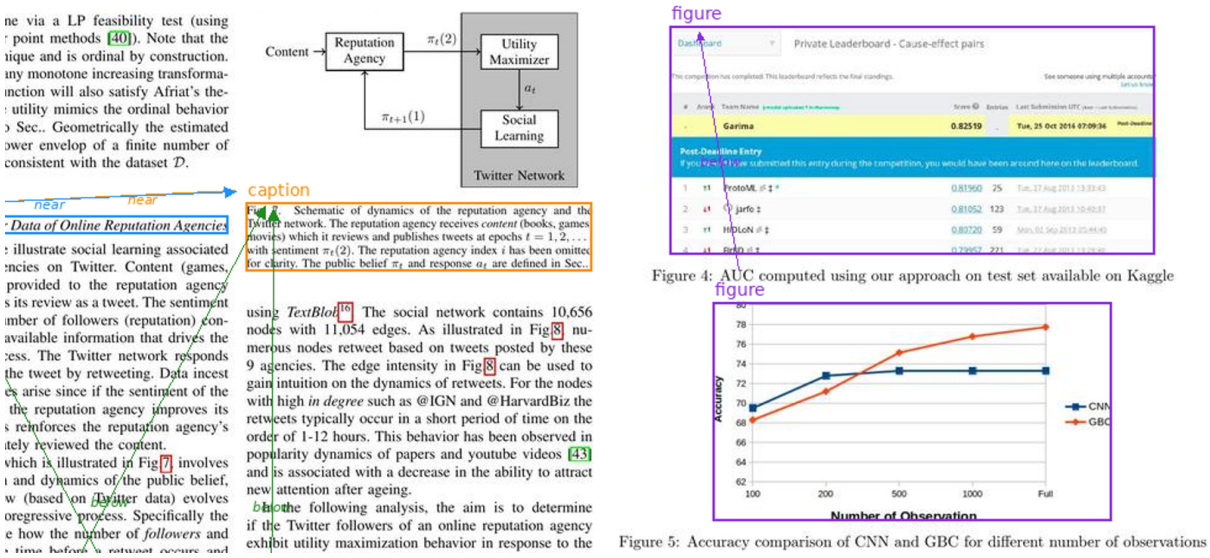


Figure 2: Rules for error detection



(a) (b)

Figure 3: Examples of proximity violation. Figures are statistically found to be near a caption. In this example **(a)**, the absence of the figure bounding box *near* the caption is indicative of a recognition error, as well as in **(b)** the absence of both caption bounding boxes *near* the figures (best viewed in color).

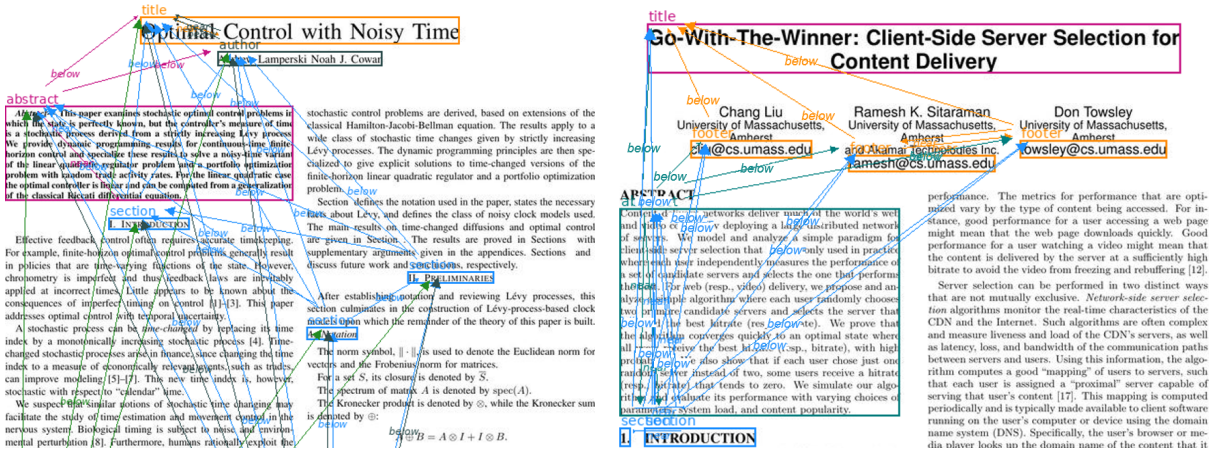
Geometric and logical rules:

1. **Definition of Below (*below*):** An element a is considered *below* an element b if its vertical position is greater than b's vertical position.

$$(a, b) \in below \Leftrightarrow vertical\_pos(a) > vertical\_pos(b)$$

   Example: The *footer* class is statistically found to be below every other region (Table 1).
2. **Hard directional constraint (*hard_below*):** $(\text{ClassA}, \text{ClassB}) \in hard\_below \Rightarrow \forall a \in \text{ClassA}, \forall b \in \text{ClassB} : (a, b) \in below$.
3. **Error rule:** An error occurs if an element that should always be *below* another, according to the hard constraint, is not.

$$(\text{ClassA}, \text{ClassB}) \in hard\_below \wedge a \in \text{ClassA} \wedge b \in \text{ClassB} \wedge (a, b) \notin below \Rightarrow \text{ERROR}(a, b)$$



(a) (b)

Figure 4: Example of directional violation. In **(a)** the order of bounding boxes is found as expected. In **(b)**, according to the relation analysis, the footer class should be below each other class and it is detected as a candidate error (best viewed in color).

Table 1
Preliminary confidence analysis of directionality for the *below* relation on a subset of 1000 documents from the DocBank dataset.

Class A	Class B	Directionality (%)	Support	Class A	Class B	Directionality (%)	Support
section	*title*	0,998	252	*footer*	*list*	0,988	42
author	*title*	0,997	165	*date*	*title*	0,986	36
abstract	*title*	0,996	138	*footer*	*section*	0,984	528
footer	*title*	0,995	108	*figure*	*title*	0,977	21
abstract	*author*	0,995	279	*caption*	*title*	0,974	18
equation	*title*	0,991	54	*caption*	*figure*	0,940	1257

Semantic relation definition (overlap + disjoint)

Semantic relations (Figure 2) such as *disjoint* and *subclass* encode constraints derived from the document ontology DoCO[1] [64]. These relations ensure that incompatible classes (e.g. caption and section) do not overlap in the visual representation of a document.

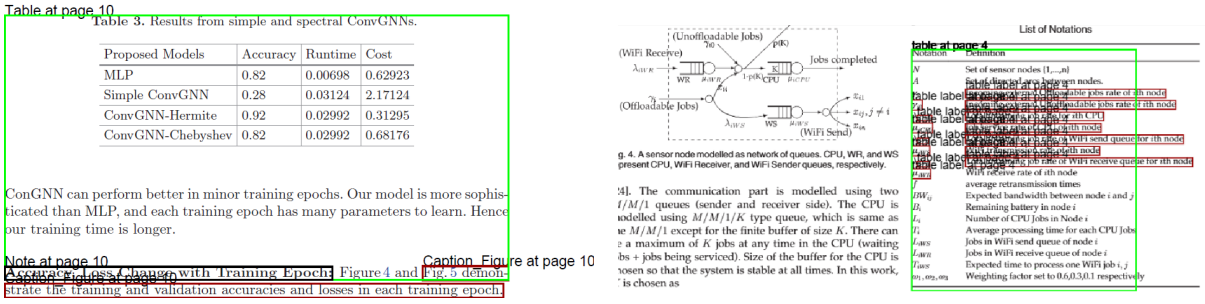


Figure 5: Examples of overlapping bounding boxes corresponding to disjoint classes (best viewed in color).

To enable the application of the constraints defined in the ontology for the considered layout elements, we exploit the DoCO relations and verify the admissibility of overlaps. Layout elements bounding boxes are represented as rectangles $(x, y, width, height)$ on a page, allowing the identification of *geometric* overlaps; the relations defined in the ontology allow us to further refine this set excluding the layout classes that are involved in a *disjoint* relation, allowing to formally check for geometric overlaps and detect *disjoint* constraint violations. We call these specific kind of errors *semantic* overlaps, corresponding to recognition errors when disjoint classes geometrically overlap (e.g. Figure 5). Overlaps are classified as admissible or not (errors) based on the `owl:disjointWith` constraints in the ontology.

All Disjoint Classes
back matter, body matter, captioned box, chapter, complex run-in quotation, footnote, formula, formula box, front matter, list, part, section, table
abstract, afterword, appendix, colophon, foreword, glossary, index, list of figures, list of tables, preface, TOC
label, paragraph, subtitle, title
list of authors, list of contributors, list of organizations
sentence, simple run-in quotation, text chunk

Table 2
General axioms of the DoCO ontology; underlinings are involved in the *not admissible* overlaps shown in Figure 5a.

Geometric and ontological rules

To exploit these relations for layout recognition improvements, we analyze the interactions among ontology classes and their corresponding spatial instances. We distinguish between *geometric overlaps*

[1] https://sparontologies.github.io/doco/current/doco.html#d4e145

and *semantic overlaps*: while some geometric overlaps between bounding boxes are admissible (e.g. a *Formula* overlapping *Abstract*), semantic overlaps indicate recognition errors when the associated overlapping classes are *disjoint* in the ontology [65].

1. **Definition of overlap (*overlap*):** Two bounding boxes A and B overlap if:

$$x_A < x_B + \text{width}_B \wedge x_B < x_A + \text{width}_A \wedge y_A < y_B + \text{height}_B \wedge y_B < y_A + \text{height}_A$$

with intersection over union (IoU) ≥ 0.1.
On the contrary, they do not overlap if one is entirely to the right or below the other:

$$x_A + \text{width}_A \leq x_B \quad \vee \quad x_B + \text{width}_B \leq x_A$$

$$y_A + \text{height}_A \leq y_B \quad \vee \quad y_B + \text{height}_B \leq y_A$$

2. **Ontological disjoint constraint (*disjoint*):** ClassA and ClassB are disjoint if the ontology specifies a *disjoint* relation:

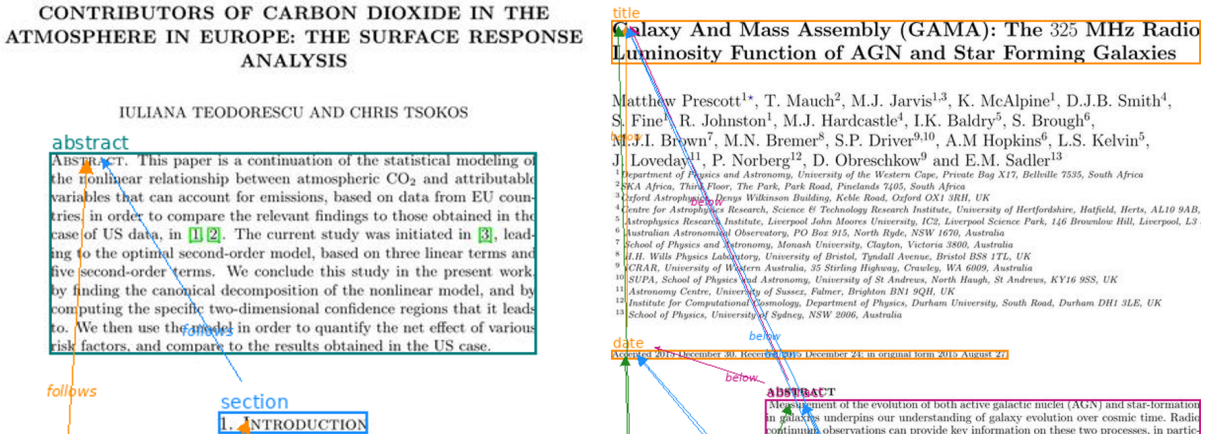
$$(\text{ClassA}, \text{ClassB}) \in \texttt{owl:disjointWith}$$

3. **Semantic overlap error rule:** A semantic error occurs when elements of disjoint classes geometrically overlap:

$$a \in \text{ClassA} \wedge b \in \text{ClassB} \wedge (a, b) \in \textit{overlap} \wedge (\text{ClassA}, \text{ClassB}) \in \texttt{owl:disjointWith} \Rightarrow \text{ERROR}(a, b)$$

Reading order relation definition (follows, precedes)

Reading order relations (Figure 2) such as *follows* describe the logical progression of textual components, applied primarily to extract the relative order of front matter elements (e.g. Figure 6). These include typical ordered chains such as *title* → *author* → *abstract* on the first page. Violations can reveal missing or misplaced metadata blocks.



(a) (b)

Figure 6: Examples of class reading order violation. The typical class reading order in the first page is found to be title, author, abstract, section. In case **(a)** title and author are missing, indicating likely a recognition error. The same holds for the lack of the author section between title and abstract in **(b)** (best viewed in color).

Once the expected sequence is statistically derived, any deviation from this pattern, either an incorrect ordering or the absence of one of the required classes, can be flagged as an error.

Logical rules:

1. These relations describe the expected position of elements; for instance, on the first page, a title is statistically found to be above the author list, which in turn precedes the abstract.
2. A violation occurs when the extracted order does not respect the learned sequence (e.g. *title* → *author* → *abstract*), or when one of the expected classes is missing in known sequences.

3.2. Visualization and statistical confidence

To support qualitative inspection and facilitate error analysis, we provide a dedicated visualization module based on `ImageDraw`. This component renders both bounding boxes and the detected relations directly on the page image, encoding relations as colored arrows annotated with textual labels. The resulting preview allows for rapid visual validation of both the spatial configuration of the layout elements and the relations inferred by the constraint analysis.

Beyond visualization, the framework computes a set of frequency based statistics that quantify the reliability of each detected relation. These statistics are derived from triples in the form (classA, *relation*, classB) extracted from the annotated corpus and are used to estimate confidence at three complementary levels: local, global, and document level.

- At the **local (page-level)** scale, confidence is estimated by measuring how frequently a given relation instance occurs within a page relative to other relations of the same type. This captures whether a relation is locally dominant or anomalous in the context of a single layout, and helps identify isolated violations that deviate from the page's prevailing structural pattern.
- At the **global (corpus-level)** scale, we compute normalized frequencies over the entire dataset. In particular, we measure how often a specific (classA, *relation*, classB) triple occurs relative to all instances of the same relation, and its frequency relative to all extracted relations. These global statistics capture how characteristic a relation is for a given pair of classes and serve as a prior reflecting dataset-wide regularities.
- At the **document-level**, confidence is estimated by measuring the consistency of a relation across multiple pages belonging to the same document. Specifically, we compute the proportion of documents in which a class A appears that also contain at least one instance of the relation (classA, *relation*, classB). This metric rewards relations that recur systematically within documents and penalizes sporadic or document-specific artifacts, thereby filtering out noise introduced by individual page layouts.

To support these three dimensions, we compute several associative and directional measures to further characterize the strength and asymmetry of relations. These include a smoothed directionality score that compares the frequency of classA $\rightarrow$ classB against its inverse, a log-odds ratio quantifying how strongly a relation favors one direction over the other, and a normalized pointwise mutual information (NPMI) score capturing the degree of statistical association between the two layout classes beyond chance. This multilevel confidence estimation makes corrections interpretable and reproducible, supporting both automated analysis and human validation by leveraging relation frequencies at the page, document, and corpus levels. While the framework is designed to support the definition of decision thresholds over individual confidence measures, in the current prototype such thresholds are not yet fixed. Instead, the statistics are computed and analyzed comparatively to assess which measures are most suitable for characterizing stable layout regularities. Each measure captures a different notion of regularity, spanning local layout frequency, corpus-level recurrence, document consistency, directional asymmetry, and semantic association. In this setting, we investigate how these measures behave in practice to identify which ones are most informative to highlight stable layout patterns. Among the semantic association metrics, Normalized Pointwise Mutual Information (NPMI) emerges as particularly promising, as it emphasizes semantic associations between the classes while mitigating the effect of globally frequent classes. For this reason, NPMI is currently being explored as a soft criterion to prioritize relations that are likely to reflect recurrent structural regularities, rather than as a single decision function.

3.3. Violation detection and LLM correction

The axioms collected from the annotated pages form the structural constraints of the dataset that are exploited to detect recognition errors. From these relations, high confidence associations are selected as requirements. Based on the existence of a relation among the present layout classes, the system may detect a *violation*. The violations that we consider are the following:

- **Proximity violations:** figure without nearby caption, or caption not linked to any table/figure;
- **Reading order violations:** incorrect ordering in first page sequence (*title* $\rightarrow$ *author* $\rightarrow$ *abstract*);

- **Directional violations:** elements consistently appearing below the expected position (e.g. misplaced *header* or *footer*);
- **Semantic violations:** overlapping and mutually exclusive regions such as *title* intersecting *caption*.

At current stage of development, the framework is limited to correction of semantic violations. After the heuristic and relation based error detection phase, error correction is enforced using a multimodal vision/language LLM. The model takes as input the page image containing the detected violation, the JSON COCO representation, and the list of inconsistencies. By jointly reasoning over visual layout and textual semantics, the LLM removes incorrect bounding boxes, enhancing the overall correctness of the input dataset. To implement this, the coordinates of layout classes involved in each error relation are integrated into a dedicated module that interacts without supervision with the open source multimodal `qwen3-vl:235b-cloud` model via the Ollama API, with temperature set to 0 to encourage low variance behavior. The labeled regions are given in the JSON COCO format: {`"id"`, `"image_id"`, `"category_id"`, `"segmentation"`, `"area"`, `"bbox"`, `"iscrowd"`}. Given the image, its annotations, and the error list in the same format, the LLM is asked to classify the bounding boxes in each error relation as *correct* or *incorrect*, guided by prompts that indicate the meaning and expected content of each layout class based on the retrieved context.

Currently, the system supports only semantic overlaps correction; in these cases the LLM evaluates the consistency of the content within the bounding boxes and any box classified as inconsistent is removed by the correction logic. The prompt includes the page image together with the labels and coordinates of the overlapping bounding boxes, expressed in the DocBank COCO format, and explicitly highlights the pair of blocks involved in the semantic violation.

The LLM is instructed to determine which of the overlapping bounding boxes is incorrect by reasoning over the expected semantic content (class label) associated with each region. To ensure correct integration within the workflow and reduce hallucinations, the prompt constrains the LLM to select one of the provided bounding boxes without altering labels or coordinates, and to return only the incorrect block in the original format, so that the rest of the logic can remove it from the annotations and update the remaining indexes accordingly. For relational constraints, such as *below* and *upon* relations, the LLM can be leveraged as well to assign the appropriate label to boxes whose positions deviate from expected spatial patterns. Proximity violation corrections (*near*) can effectively be performed adding bounding boxes based on detected omissions. At current stage of development this functionality, along with the handling of reading order violations, has not yet been implemented. Compared to identifying segmentation errors without any prior indication of their location, this strategy is expected to produce better results because it directs the model to specific text regions that are presumed to contain the errors, thereby making the correction process more effective.

The corrections performed by the LLM can be integrated into a new version of the dataset (e.g. `DocBank-Clean`) that builds new annotation files preserving the COCO format, proposing an improved version in which geometric inconsistencies are fixed by adjusting coordinates or merging unpaired bounding boxes; sequential violations are corrected according to the learned *follows*/*precedes* relations, and semantic conflicts are resolved through the ontology based checks.

3.4. Discussion

We conducted exploratory analysis on a limited subset of the DocBank dataset (1000 pages). Within this sample, our study reveals several annotation inconsistencies among geometric and semantic layout constraints. Beyond early feasibility analysis, quantitative evaluation (including large-scale precision estimates and downstream DLA experiments on original versus cleaned data) is left to future work, as it requires curated ground truth and dedicated experimental protocols. Although the scale of the experiments does not allow for statistically significant conclusions, our early results show that structural inconsistencies can be systematically detected through the application of explicit layout constraints. The current results indicate that the system can automatically identify and resolve identified violations, suggesting that extending the analysis to the entire data set could improve structural coherence. To the best of our knowledge, this is the first approach that explicitly addresses structural layout inconsistencies in large-scale DLA datasets by formalizing layout relations and combining them with multimodal LLMs;

prior pipelines that aim to improve model outputs rely on implicit heuristics and do not reason explicitly about layout relational coherence. Although the current prototype employs a very large LLM, many violations admit simpler resolution strategies, performing lightweight classification tasks (e.g. wrong labels or proximity errors). These cases can also be handled using less demanding models, or even purely symbolic methods, which may suffice for most corrections. Our choice is motivated by the cloud-based availability of the model, which avoids the local computational overhead.

Quantitative downstream evaluation is left for future work; the present contribution focuses on dataset inspection and correction. The results illustrate the potential of explicit layout constraints to support the analysis and improvement of internal consistency in large-scale document layout datasets. The framework is designed to be extensible, supporting future large-scale evaluations and the development of cleaner, semantically consistent resources for training vision-language document models.

4. Conclusions and future work

This work introduces a relational and multimodal strategy for assessing and ensuring the structural coherence of large-scale document layout datasets. Instead of treating layout elements as independent annotations, we model their mutual dependencies through geometric, sequential and semantic relations, and use these relations to reveal incoherent or implausible configurations in automatically generated ground truth, exploiting relations learned directly from data. The proposed system combines heuristic relation extraction, statistical confidence modeling, and a multimodal vision/language LLM that operates as a corrective layer over existing annotations. In a preliminary proof of concept on a small subset of DocBank, this relational perspective showed effective in detecting typical sources of incoherence, such as missing or misplaced captions and front matter elements, and in suggesting corrections that lead to more consistent page structures. These early experiments performed on the DocBank dataset reveal that even widely used resources are affected by non negligible structural noise stemming from PDF-to-XML alignment, incomplete region labeling, and violations of expected reading order. Although the current study does not aim at a definitive quantitative assessment, it describes the potential of relation-driven post processing for dataset curation, introducing automatic corrections and explicit coherence indicators. The definition of the framework and the DocBank based prototype lay the foundation for future cleaned releases and for more extensive benchmarking on larger portions of DocBank and on additional corpora such as PubLayNet and DocLayNet, with the broader goal of supplying more reliable training material for document understanding models.

Future work

Future developments will extend this strategy in several directions. We plan to scale the coherence analysis to full datasets such as DocBank, PubLayNet and DocLayNet, running the pipeline on all pages, extracting more relations, and computing relevance metrics. This will allow a systematic comparison of structural noise across corpora and the release of cleaned variants (e.g. `DocBank-Clean`, `PubLayNet-Clean`, `DocLayNet-Clean`) together with relation analysis and evaluations.

We also aim to generalize the model to multilingual articles, specifically where reading order and typographic conventions differ from Western layouts. By combining language aware signals with geometric reasoning, we plan to study how relational distributions change across languages and how coherence constraints must be adapted to support archives such as arXiv, HAL, and CNKI.

Another direction concerns generative data augmentation. Diffusion based layout generators and synthetic document simulators can be used to create controlled page structures with known relational ground truth. The multimodal LLM will be extended to bounding box correction and multipage reasoning, enabling detection and fixing of inconsistencies that span several pages (e.g. repeated titles, displaced references, figure and caption misalignments). Document level attention and memory mechanisms can be implemented to aggregate relational evidence and evaluate coherence beyond single page layouts. Finally, we plan to integrate coherence constraints directly into model training, using the coherence score as an auxiliary objective to encourage structurally consistent predictions. We also aim to evaluate whether training on corrected datasets improves downstream tasks such as information extraction or document parsing, and how such corrections impact overall performance.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] A. Gemelli, S. Marinai, L. Pisaneschi, F. Santoni, Datasets and annotations for layout analysis of scientific articles, Int. J. Document Anal. Recognit. 27 (2024) 683–705. URL: https://doi.org/10.1007/s10032-024-00461-2. doi:10.1007/S10032-024-00461-2.

[2] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, Ł. Bolikowski, Cermine: automatic extraction of structured metadata from scientific literature, International Journal on Document Analysis and Recognition (IJDAR) 18 (2015) 317–335.

[3] I. Keraghel, S. Morbieu, M. Nadif, Recent advances in named entity recognition: A comprehensive survey and comparative study, 2024. URL: https://arxiv.org/abs/2401.10825. arXiv:2401.10825.

[4] T. Dozat, C. D. Manning, Deep biaffine attention for neural dependency parsing, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017. URL: https://openreview.net/forum?id=Hk95PK9le.

[5] M. Palmer, D. Gildea, N. Xue, Semantic role labeling, Morgan & Claypool Publishers, 2011.

[6] L. Massai, Evaluation of semantic relations impact in query expansion-based retrieval systems, Knowledge Based Systems 283 (2024) 111183. URL: https://doi.org/10.1016/j.knosys.2023.111183. doi:10.1016/J.KNOSYS.2023.111183.

[7] J. A. Diaz-Garcia, J. A. D. Lopez, A survey on cutting-edge relation extraction techniques based on language models, Artificial Intelligence Review 58 (2025) 287.

[8] F. M. Wahl, K. Y. Wong, R. G. Casey, Block segmentation and text extraction in mixed text/image documents, Computer graphics and image processing 20 (1982) 375–390.

[9] L. O'Gorman, The document spectrum for page layout analysis, IEEE Transactions on pattern analysis and machine intelligence 15 (2002) 1162–1173.

[10] G. Nagy, S. C. Seth, Hierarchical representation of optically scanned documents, in: CSE Conference and Workshop Papers, 1984, pp. 347–349.

[11] P. Huang, A. R. Kashyap, Y. Qin, Y. Yang, M. Kan, Lightweight contextual logical structure recovery, in: A. Cohan, G. Feigenblat, D. Freitag, T. Ghosal, D. Herrmannova, P. Knoth, K. Lo, P. Mayr, M. Shmueli-Scheuer, A. de Waard, L. L. Wang (Eds.), Proceedings of the Third Workshop on Scholarly Document Processing, SDP@COLING 2022, Gyeongju, Republic of Korea, October 12 - 17, 2022, Association for Computational Linguistics, 2022, pp. 37–48. URL: https://aclanthology.org/2022.sdp-1.5.

[12] A. R. Kashyap, M. Kan, Sciwing- A software toolkit for scientific document processing, in: M. K. Chandrasekaran, A. de Waard, G. Feigenblat, D. Freitag, T. Ghosal, E. H. Hovy, P. Knoth, D. Konopnicki, P. Mayr, R. M. Patton, M. Shmueli-Scheuer (Eds.), Proceedings of the First Workshop on Scholarly Document Processing, SDP@EMNLP 2020, Online, November 19, 2020, Association for Computational Linguistics, 2020, pp. 113–120. URL: https://doi.org/10.18653/v1/2020.sdp-1.13. doi:10.18653/V1/2020.SDP-1.13.

[13] Y. Shinyama, Pdfminer: Python pdf parser and analyzer, 2015. URL: https://github.com/euske/pdfminer.

[14] J. Rausch, O. Martinez, F. Bissig, C. Zhang, S. Feuerriegel, Docparser: Hierarchical document structure parsing from renderings, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021,

Virtual Event, February 2-9, 2021, AAAI Press, 2021, pp. 4328–4338. URL: https://doi.org/10.1609/aaai.v35i5.16558. doi:10.1609/AAAI.V35I5.16558.

[15] Z. Wang, M. Zhan, X. Liu, D. Liang, Docstruct: A multimodal method to extract hierarchy structure in document for general form understanding, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020, volume EMNLP 2020 of *Findings of ACL*, Association for Computational Linguistics, 2020, pp. 898–908. URL: https://doi.org/10.18653/v1/2020.findings-emnlp.80. doi:10.18653/V1/2020.FINDINGS-EMNLP.80.

[16] J. Ma, J. Du, P. Hu, Z. Zhang, J. Zhang, H. Zhu, C. Liu, Hrdoc: Dataset and baseline method toward hierarchical reconstruction of document structures, in: B. Williams, Y. Chen, J. Neville (Eds.), Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, AAAI Press, 2023, pp. 1870–1877. URL: https://doi.org/10.1609/aaai.v37i2.25277. doi:10.1609/AAAI.V37I2.25277.

[17] M. Arif Demirtaş, B. Oral, M. Yasin Akpınar, O. Deniz, Semantic parsing of interpage relations, in: 2022 26th International Conference on Pattern Recognition (ICPR), 2022, pp. 1579–1585. doi:10.1109/ICPR56361.2022.9956546.

[18] M. Neumann, Z. Shen, S. Skjonsberg, Pawls: Pdf annotation with labels and structure, arXiv preprint arXiv:2101.10281 (2021).

[19] LabelImg, Resource available online, 2022. URL: https://www.github.com/tzutalin/labelImg.

[20] L. Pisaneschi, A. Gemelli, S. Marinai, Automatic generation of scientific papers for data augmentation in document layout analysis, Pattern Recognition Letters 167 (2023) 38–44. URL: https://www.sciencedirect.com/science/article/pii/S0167865523000247. doi:https://doi.org/10.1016/j.patrec.2023.01.018.

[21] D. Zha, Z. P. Bhat, K.-H. Lai, F. Yang, X. Hu, Data-centric ai: Perspectives and challenges, in: Proceedings of the 2023 SIAM international conference on data mining (SDM), SIAM, 2023, pp. 945–948.

[22] P. Lopez, Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications, in: European Conference on Research and Advanced Technology for Digital Libraries, 2009. URL: https://api.semanticscholar.org/CorpusID:27383212.

[23] D. Tkaczyk, A. Collins, P. Sheridan, J. Beel, Evaluation and comparison of open source bibliographic reference parsers: A business use case, arXiv preprint arXiv:1802.01168 (2018).

[24] B. Wang, C. Xu, X. Zhao, L. Ouyang, F. Wu, Z. Zhao, R. Xu, K. Liu, Y. Qu, F. Shang, et al., Mineru: An open-source solution for precise document content extraction, arXiv preprint arXiv:2409.18839 (2024).

[25] C. He, W. Li, Z. Jin, C. Xu, B. Wang, D. Lin, Opendatalab: Empowering general artificial intelligence with open datasets, arXiv preprint arXiv:2407.13773 (2024).

[26] Z. Zhao, H. Kang, B. Wang, C. He, Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception, 2024. URL: https://arxiv.org/abs/2410.12628. arXiv:2410.12628.

[27] B. Wang, Z. Gu, C. Xu, B. Zhang, B. Shi, C. He, Unimernet: A universal network for real-world mathematical expression recognition, 2024. arXiv:2404.15254.

[28] I. Amazon Web Services, Amazon textract, 2025. URL: https://aws.amazon.com/textract/.

[29] I. T. Phillips, S. S. Chen, R. M. Haralick, CD-ROM document database standard, Technical Report, 1993. URL: https://doi.org/10.1109/ICDAR.1993.395691. doi:10.1109/ICDAR.1993.395691.

[30] G. Ford, G. R. Thoma, Ground truth data for document image analysis, Technical Report, 2003.

[31] A. Antonacopoulos, D. Bridson, C. Papadopoulos, S. Pletschacher, A realistic dataset for performance evaluation of document layout analysis, in: 10th International Conference on Document Analysis and Recognition, ICDAR 2009, Barcelona, Spain, 26-29 July 2009, IEEE Computer Society, 2009, pp. 296–300. URL: https://doi.org/10.1109/ICDAR.2009.271. doi:10.1109/ICDAR.2009.271.

[32] J. Fang, X. Tao, Z. Tang, R. Qiu, Y. Liu, Dataset, ground-truth and performance metrics for table

detection evaluation, in: 2012 10th IAPR International Workshop on Document Analysis Systems, IEEE, 2012, pp. 445–449.
[33] C. A. Clark, S. K. Divvala, Looking beyond text: Extracting figures, tables and captions from computer science papers., in: AAAI Workshop: Scholarly Big Data, volume 6, 2015.
[34] C. Clark, S. Divvala, Pdffigures 2.0: Mining figures from research papers, in: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, 2016, pp. 143–152.
[35] N. Siegel, Z. Horvitz, R. Levin, S. Divvala, A. Farhadi, Figureseer: Parsing result-figures in research papers, in: European Conference on Computer Vision, Springer, 2016, pp. 664–680.
[36] N. Siegel, N. Lourie, R. Power, W. Ammar, Extracting scientific figures with distantly supervised neural networks, in: Proceedings of the 18th ACM/IEEE on joint conference on digital libraries, 2018, pp. 223–232.
[37] S. Y. Kahu, W. A. Ingram, E. A. Fox, J. Wu, Scanbank: A benchmark dataset for figure extraction from scanned electronic theses and dissertations, arXiv preprint arXiv:2106.15320 (2021).
[38] F. Shafait, R. Smith, Table detection in heterogeneous documents, in: D. S. Doermann, V. Govindaraju, D. P. Lopresti, P. Natarajan (Eds.), The Ninth IAPR International Workshop on Document Analysis Systems, DAS 2010, June 9-11, 2010, Boston, Massachusetts, USA, ACM International Conference Proceeding Series, ACM, 2010, pp. 65–72. URL: https://doi.org/10.1145/1815330.1815339. doi:10.1145/1815330.1815339.
[39] A. Gemelli, E. Vivoli, S. Marinai, Graph neural networks and representation embedding for table extraction in pdf documents, in: 2022 26th International Conference on Pattern Recognition (ICPR), IEEE, 2022. URL: http://dx.doi.org/10.1109/ICPR56361.2022.9956590. doi:10.1109/icpr56361.2022.9956590.
[40] X. Zhong, J. Tang, A. J. Yepes, Publaynet: largest dataset ever for document layout analysis, in: 2019 International conference on document analysis and recognition (ICDAR), IEEE, 2019, pp. 1015–1022.
[41] M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, M. Zhou, Docbank: A benchmark dataset for document layout analysis, arXiv preprint arXiv:2006.01038 (2020).
[42] M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, Z. Li, Tablebank: Table benchmark for image-based table detection and recognition, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 1918–1925.
[43] Y. Deng, D. Rosenberg, G. Mann, Challenges in end-to-end neural scientific table recognition, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2019, pp. 894–901.
[44] X. Zhong, E. ShafieiBavani, A. Jimeno Yepes, Image-based table recognition: data, model, and evaluation, in: European conference on computer vision, Springer, 2020, pp. 564–580.
[45] B. Smock, R. Pesala, R. Abraham, Pubtables-1m: Towards comprehensive table extraction from unstructured documents, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4634–4642.
[46] H. Desai, P. Kayal, M. Singh, Tablex: a benchmark dataset for structure and content information extraction from scientific tables, in: International conference on document analysis and recognition, Springer, 2021, pp. 554–569.
[47] Z. Chi, H. Huang, H.-D. Xu, H. Yu, W. Yin, X.-L. Mao, Complicated table structure recognition, arXiv preprint arXiv:1908.04729 (2019).
[48] B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar, P. Staar, Doclaynet: A large human-annotated dataset for document-layout segmentation, in: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining, 2022, pp. 3743–3751.
[49] L. Markewich, H. Zhang, Y. Xing, N. Lambert-Shirzad, Z. Jiang, R. K.-W. Lee, Z. Li, S.-B. Ko, Segmentation for document layout analysis: not dead yet, International Journal on Document Analysis and Recognition (IJDAR) (2022). URL: https://doi.org/10.1007/s10032-021-00391-3. doi:10.1007/s10032-021-00391-3.
[50] F. Grijalva, C. Parra, M. Gallardo, E. Santos, B. Acuña, J. C. Rodríguez, J. Larco, Scibank: a large dataset of annotated scientific paper regions for document layout analysis, IEEE Dataport (2022).

[51] K. Lo, L. L. Wang, M. Neumann, R. Kinney, D. S. Weld, S2orc: The semantic scholar open research corpus, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 4969–4983.
[52] T. Saier, J. Krause, M. Färber, unarxive 2022: All arxiv publications pre-processed for nlp, including structured full-text and citation network, in: 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL), IEEE, 2023, pp. 66–70.
[53] W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, J. Dunkelberger, A. Elgohary, S. Feldman, V. Ha, et al., Construction of the literature graph in semantic scholar, arXiv preprint arXiv:1805.02262 (2018).
[54] A. D. Wade, The semantic scholar academic graph (s2ag), in: Companion Proceedings of the Web Conference 2022, 2022, pp. 739–739.
[55] T. Saier, M. Färber, Bibliometric-enhanced arxiv: A data set for paper-based and citation-based tasks., in: BIR@ ECIR, 2019, pp. 14–26.
[56] J. Priem, H. Piwowar, R. Orr, Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, arXiv preprint arXiv:2205.01833 (2022).
[57] H. Bast, C. Korzen, A benchmark and evaluation for text extraction from pdf, in: 2017 ACM/IEEE joint conference on digital libraries (JCDL), IEEE, 2017, pp. 1–10.
[58] C. Soto, S. Yoo, Visual detection with context for document layout analysis, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3464–3470.
[59] C. Soto, S. Yoo, Visual detection with context for document layout analysis, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3464–3470.
[60] D. Kershaw, R. Koeling, Elsevier oa cc-by corpus, arXiv preprint arXiv:2008.00774 (2020).
[61] T. Bozsolik, J. Tricot, D. Rishi, B. Maltzan, S. Brinn, arxiv dataset, 2024. URL: https://www.kaggle.com/dsv/7548853. doi:10.34740/KAGGLE/DSV/7548853.
[62] Z. Wang, Y. Xu, L. Cui, J. Shang, F. Wei, LayoutReader: Pre-training of text and layout for reading order detection, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 4735–4744. URL: https://aclanthology.org/2021.emnlp-main.389/. doi:10.18653/v1/2021.emnlp-main.389.
[63] C. Zhang, Y. Guo, Y. Tu, H. Chen, J. Tang, H. Zhu, Q. Zhang, T. Gui, Reading order matters: Information extraction from visually-rich documents by token path prediction, 2023. URL: https://arxiv.org/abs/2310.11016. arXiv:2310.11016.
[64] A. Constantin, S. Peroni, S. Pettifer, D. Shotton, F. Vitali, The document components ontology (doco), Semantic web 7 (2016) 167–181.
[65] L. Massai, S. Marinai, et al., An integrated system for interacting with multi-page scholarly documents, in: Proceedings of IRCDL 2025, volume 3937, Marcella Cornia, Giorgio Maria Di Nunzio, Donatella Firmani, Stefano Mizzaro ..., 2025, pp. 14–26.