Autodive: An Integrated Onsite Scientific Literature Annotation Tool

Yi Du[1,2,3]*, Ludi Wang[1], Mengyi Huang[1,2], Dongze Song[1], Wenjuan Cui[1,2], Yuanchun Zhou[1,2,3]*

[1] Computer Network Information Center, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] University of Science and Technology of China
$\{duyi, wld, myhuang, songdongze, wenjuancui, zyc\}@cnic.cn$

Abstract

Scientific literature is always available in Adobe's Portable Document Format (PDF), which is friendly for scientists to read. Compared with raw text, annotating directly on PDF documents can greatly improve the labeling efficiency of scientists whose annotation costs are very high. In this paper, we present Autodive, an integrated onsite scientific literature annotation tool for natural scientists and Natural Language Processing (NLP) researchers. This tool provides six core functions of annotation that support the whole lifecycle of corpus generation including i)annotation project management, ii)resource management, iii)ontology management, iv)manual annotation, v)onsite auto annotation, and vi)annotation task statistic. Two experiments are carried out to verify efficiency of the presented tool. A live demo of Autodive is available at http://autodive.sciwiki.cn, and a video demo http://autodive.sciwiki.cn/introVideo/introduce-v1.0.mp4. The source code is available at https://github.com/Autodive.

1 Introduction

Influential applications such as AlphaFold2 (Jumper et al., 2021), and mobile robotic chemist(Burger et al., 2020) rely on high-quality databases and domain knowledge, some of which are constructed from scientific literature. The traditional method of building such databases still needs the manual annotation of numerous professionals. With the development of scientific research and the enhancement of interdisciplinary integration, the number of scientific articles has increased explosively, which also requires the annotators who are assigned to build the domain-specific database to have a suitable background and are familiar with annotation of literature data. Some research attempts to assist the manual annotation process with the intelligent natural language model, such as Named Entity Recognition (NER) and Relation Identification (RI). The intelligent method can automatically extract knowledge from articles, and form high-quality databases after expert proofreading (Cruse et al., 2022; Yan et al., 2022). However, the performance of the general model in specific fields is not satisfactory, and the construction of a specific intelligent model needs high-quality databases. Thus, An easy-to-use annotation tool with a graphical user interface that allows the labeling of text efficiently and consistently is crucial and necessary.

Annotation tools play a crucial role during the database-making process in the field of biology(López-Fernández et al., 2013), material(Corvi et al., 2021), and chemistry(Swain and Cole, 2016). Although the majority of released annotation tools mainly focus on the annotation of multimedia such as image and video, there are still text annotation tools such as AnnIE(Friedrich et al., 2021), TS-ANNO(Stodden and Kallmeyer, 2022) and Doccano(Nakayama et al., 2018). However, most of the tools need to convert the input file from PDF format to plain text, which may cause additional labor costs for resource preparation. Moreover, the annotation of specific data required professionals with knowledge of different fields. Compared with raw text, annotating directly on PDF documents conforms to the reading habits of the professionals with the original images, tables and layout information and can greatly improve the labeling efficiency of them whose annotation costs are very high. We refer to this need as **Onsite Annotation**, which includes the abilities to display and direct annotate on the original PDF documents.

To meet the needs of professionals on direct annotation, the new tool should support direct scientific literature annotation in PDF format. That means this tool should display the original literature in PDF format directly, so users can read the complete PDF content in the annotation inter-

[*] Correspond Author

face, where they annotate the entity and relation of scientific literature through scheduled annotation operation. In this way, this tool retains the layout information of the original PDF literature to fit the reading habits of annotators. Different from traditional text annotation projects, scientific annotation projects require annotation tools to process long literature, annotate complex entities and export appropriate annotation results to a formed database.

Besides onsite annotation, integration is also a valuable factor during the design of the tool. First, the existing domain-specific ontology should be easily integrated which can reduce the time costs of ontology design. Second, the tool should be well integrated with existing and new Named-entity Recognition (NER) models flexibly. Third, integration with existing file systems or literature collections, such as Pubmend Central and self-organized document folders, should also be valued. By integrating with an existing data source, a researcher can deploy this tool locally without leakage of unpublished or copyrighted documents.

Hence, we propose an open-sourced annotation tool Autodive, with its contributions summarized below:

(1) Onsite Annotation. Autodive supports onsite annotation of PDF documents for professionals. They can annotate directly on PDF documents, and get instant visual feedback.

(2) Integration. Autodive can integrated external modules that may assist the whole annotation process, including corpus management, ontology construction, manual and intelligent annotation.

(3) Domain Verified. The effectiveness has been verified by two tasks, including catalytic material annotation and scientific dataset annotation.

2 Architecture

The overall architecture of Autodive is shown in Fig.1 with three layers which are Data Source, Server Layer, and Frontend. The core component of Autodive is the **Server Layer**. *Data Source Adapter* can integrate specific **Data Sources** such as Pubmed Central or File System with Autodive. Three core server-side engines play important roles. The first is the *Regular Expression Parsing Engine*. It can help Autodive extract entities by predefined regular expressions. The second is the *Ontology Management Engine*. It can load and parse ontology files with OWL format and save the designed ontology to OWL or self-defined JSON files. The third engine is the *NER Model Management Engine*. It is a model base that is extendable and friendly to newly trained NER models. The annotation engine is the bridge between the server-side and the frontend. This engine provides locating and feedback of entities. **Frontend** is implemented mainly using a progressive JavaScript framework Vue.js.

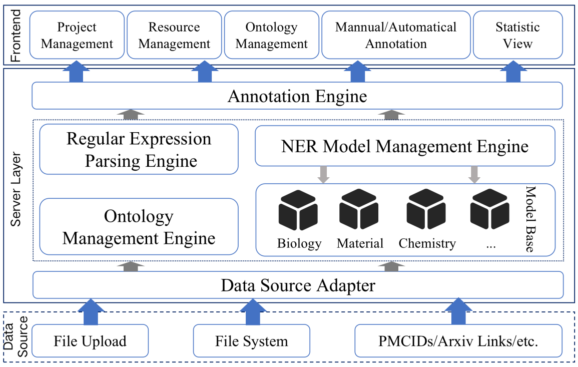


Figure 1: Architecture of Autodive.

As shown in Fig.2, the complete literature in PDF format is shown on the right part of the annotation page of Autodive, where annotators can read the literature and annotate the entity. Users click the mouse to select words on the displayed PDF document, and click the right mouse to select the type of entities or relationships. When connecting two annotated entities, users can annotate relation of them. Shortcut-key annotation is allowed for the increase of efficiency of the annotation project. All identified or annotated entities and relations are listed in the entity-label-list and relation-label-list with predefined ontology.

3 Modules

Autodive integrates management, annotation, and optimization through six modules and is specially built for scientific literature annotation projects. Users use the **Project Management** module to develop and manage personalized annotation projects. The **Resource Management** module allows users to manage their own literature resources pool. The **Ontology Management** module pre-defines the knowledge ontology, which is defined and standardized by the annotation project administrator. The basic components for annotating are the **Manual Annotation** module and the **Auto Annotation** module. Trained automatic annotation models are saved in the background in order to improve the productivity of manual annotation work. Finally,

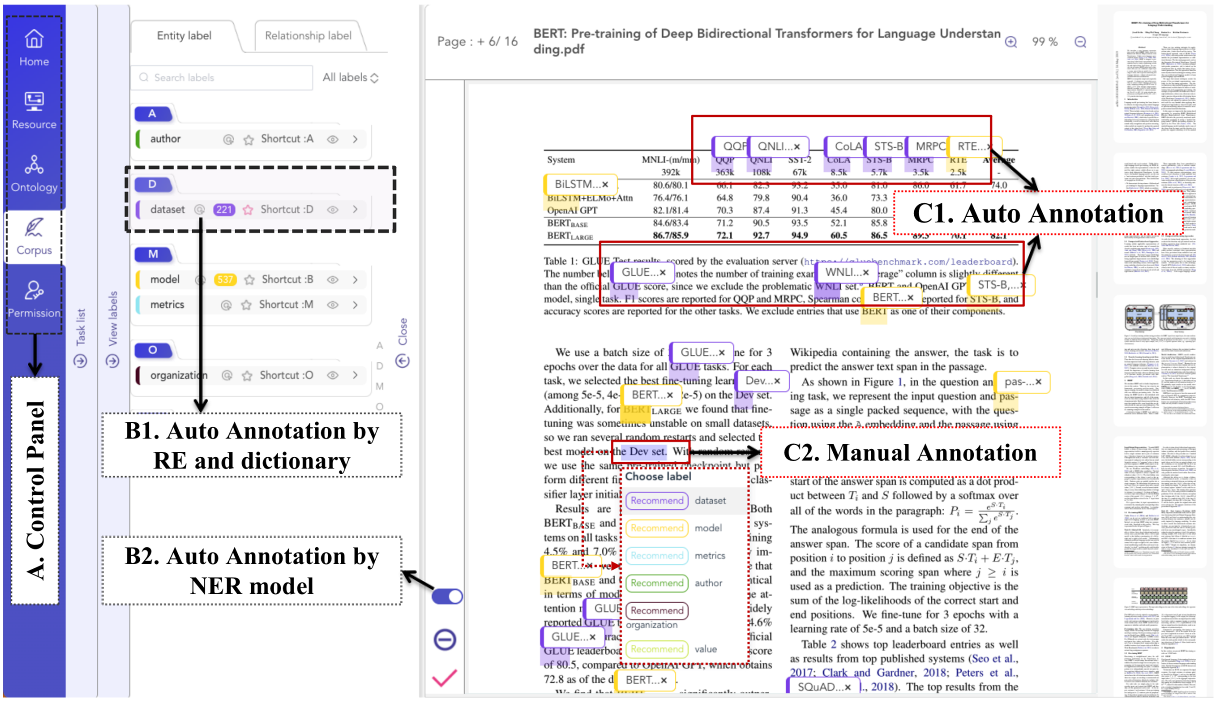


Figure 2: Overview of Autodive annotation interface. The left part of the interface is A. Control Panel that integrates all functions of Autodive. B1 and B2 demonstrate usage of three auto annotation functions, while C1 and C2 show the results of both auto annotation(C1) and manual annotation(C2).

we created a **Statistic View** module for Autodive to represent the overall progress of the annotation project.

3.1 Project Management

The annotation lifecycle starts with the creation of one annotation project. Autodive designs project management module that includes project creation and annotator administration. As shown in Fig.3(a), users submit necessary information such as project name while creating an annotation project. Domain information is also encouraged to fill out so that the auto-annotation model can accelerate the annotation.

There are three different kinds of roles. The creator of the project, also plays as the administrator, has the ability to invite annotators. If the invitation email is approved, invitees will be added straight to the relevant annotation project as administrators or annotators. The administrator also has the rights to assign annotation tasks to project members.

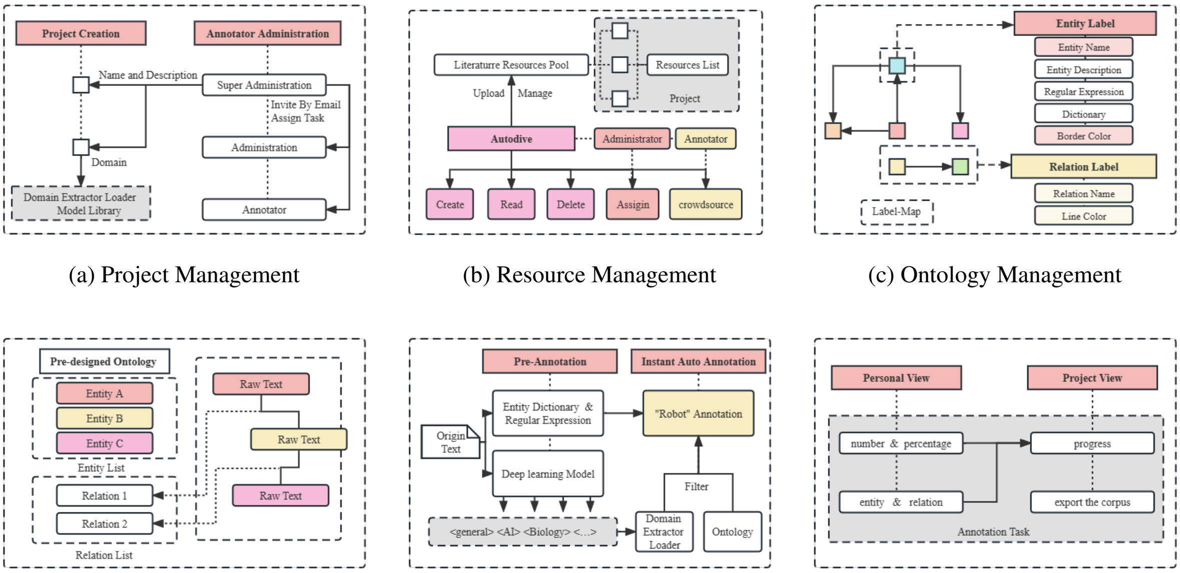
3.2 Resource Management

For a literature annotation project, the initial and critical step is to control which documents to be annotated and where to find them. The quantity and diversity of literature affect the quality of annotation data, and therefore the accuracy of intelligent models. Autodive allows users to upload and manage their own literature resources pool, and form a list of documents relating different annotation projects by associating the literature resources with them. Besides uploading PDF documents directly, this tool also provides standard API (Application Programming Interface) that can import scientists' own literature resources.

After initialing the list of annotation resources, the administrator can assign the annotation resources to other annotators, and the annotators can complete the follow-up task in the form of crowdsourcing, which is as shown in Fig.3(b).

3.3 Ontology management

Ontology means what kinds of entities and relations to be annotated in the annotation project, which is defined and controlled by the project administer to fit task requirements. Well defined ontology can enhance the efficiency of annotation. The process of ontology design is shown in Fig.3(c). The first basis component of ontology management is load and design of one ontology. Considering the complexity of the link between entity-labels, the same relation-label category may distribute to numerous entity-labels pairs, and multiple relation-labels may distribute to the same entity-labels pair. The



(a) Project Management

(b) Resource Management

(c) Ontology Management

(d) Manual Annotation

(e) Auto Annotation

(f) Statistic View

Figure 3: Operation process of six core modules in Autodive.

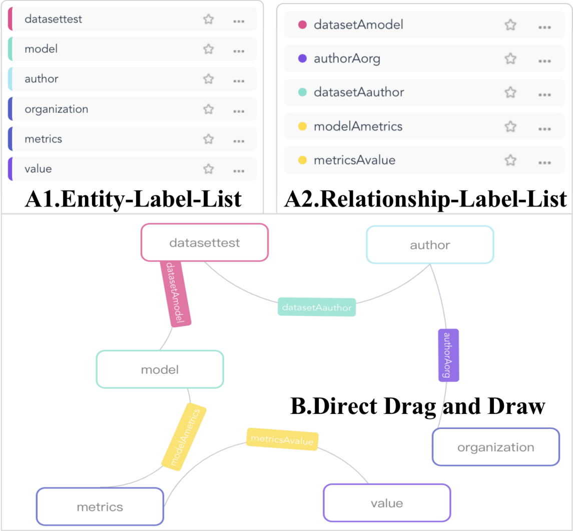


Figure 4: Two different display modes of ontology management.

label-map will display everything mentioned above. Autodive also allow users download the designed ontology for subsequent research or application.

During the design process of ontology, Autodive allow users upload the corresponding dictionary and fill up the required regular expression. It is advantageous to simplify some annotation projects when the entity-label has a clear thesaurus to-be-annotated terms or words that have a unified structure. Two display modes of ontology is shown in Fig.4.

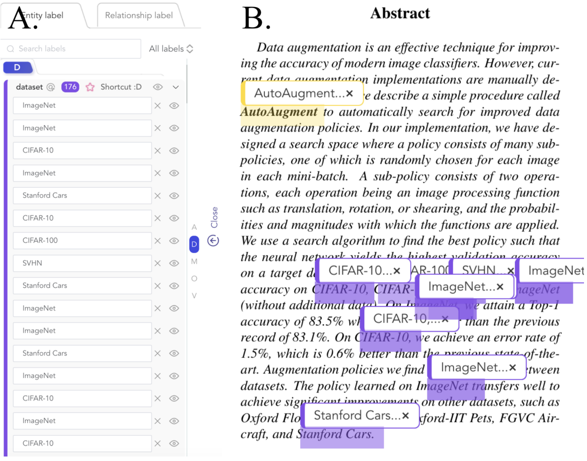
3.4 Manual Annotation

The manual annotation of an entity in Autodive can select raw text instead of drawing a bounding box in PDF document, this mode is not like PAWLS(Neumann et al., 2021). After selecting, the user decides what kind of entity the text is. This way of entity annotation is more precise, especially in annotating text that has line wrap in document. The manual annotation also provide a controlled annotation of relationship. Annotator select two annotated entity, then Autodive will recommend possible relationship that pre-designed in ontology. This recommendation step can also increase the efficiency of annotation.

3.5 Auto Annotation

Autodive provides pre-annotation by three ways, they are regular expression parsing, dictionary mapping, and external NER models. It is an evident advantage for dictionary and regular expression annotation since they basically annotate "the proper terms". However, when there are unavoidable out-of-vocabulary (OOV) words in the to-be-annotated literature, intelligent model annotation using external NER models is a more effective choice.

It is evident that in various scientific research domains, the model of private domain learned the typical information during training, allowing it to accurately recognize the label. Autodive uses the **Domain Extractor Loader** and the **Model Li-**



Abstract

Data augmentation is an effective technique for improving the accuracy of modern image classifiers. However, current data augmentation implementations are manually de- e describe a simple procedure called ***AutoAugment*** *to automatically search for improved data augmentation policies. In our implementation, we have designed a search space where a policy consists of many subpolicies, one of which is randomly chosen for each image in each mini-batch. A sub-policy consists of two operations, each operation being an image processing function such as translation, rotation, or shearing, and the probabilities and magnitudes with which the functions are applied. We use a search algorithm to find the best policy such that the neural network yields the highest validation accuracy on a target d accuracy on CIFAR-10, CIFAR- ageNet (without additional da we attain a Top-1 accuracy of 83.5% wh than the previous record of 83.1%. On CIFAR-10, we achieve an error rate of 1.5%, which is 0.6% better than the previous state-of-the-art. Augmentation policies we find tween datasets. The policy learned on ImageNet transfers well to achieve significant improvements on other datasets, such as Oxford Flo xford-IIT Pets, FGVC Aircraft, and Stanford Cars.*

Figure 5: Auto annotation. A shows list of auto extracted labels.B displays instant auto annotation and onsite feedback.

brary to choose specific domain prediction models. When starting an annotation project, users should select the corresponding field of the project. In advance, we saved annotation models in several scientific domains in model library. Furthermore, when the project reaches unexcited fields, users are prompted to select a model from the "generic domain" and fill up the domain details.

One highlight function of the auto annotation is **Instant Auto Annotation**. As shown in Fig.5, the annotation results can be presented right on the complete text with a single click. To support this function, Autodive constructed a robot annotation layer, and the annotation on the source document is displayed in the same way as the manual annotation results. The text parsing tool will first import the text into the auto annotation model, and the model will return the annotation results. The label will then be filtered according to the ontology established by the project, and it will be matched to the precise spot on the document. The auto annotation model adds a robot annotation layer to the literature display layer. The deep learning model will undoubtedly take some time. Autodive chooses to extract literature in advance during background free time, saves the automatic annotation results in the database, and then performs specific matching annotation based on the model domain and ontology chosen by each announcer, reducing the waiting time of specific users.

3.6 Statistic View

To help the administrator and annotator know the status and progress of one annotation task, Autodive provides **Personal View** and **Project View** in this module. In Personal View, the number and percentage of current annotation task are provided so that annotator can evaluate his/her task. Besides, the number of annotated and auto-recognized entity and relationship is also shown in the view. In Project View, functions are provided to help administrator understand current progress of all annotators and their assigned tasks, such as the distribution of annotated entity and relationship, number and percentage of each annotator and his/her task, and so on. In Project View, user can export the corpus for further use.

4 Evaluation & Case Study

4.1 Annotation Tools Evaluation

We compared Autodive with other open sourced annotation tools, including AnnIE(Friedrich et al., 2021), Doccano(Nakayama et al., 2018), WebAnno(Yimam et al., 2013), INCEpTION(Klie et al., 2018), PDFAnno(Shindo et al., 2018) and PAWLS(Neumann et al., 2021), for annotation function comparison.

In order to match the need of scientific literature annotation, we design the evaluation metrics as bellows: The first is **[A].Availability**, which includes *[A1].Activity* and *[A2].Online Service*. The second is **[B].Onsite Support**, which includes *[B1].Onsite PDF Display* and *[B2].Onsite PDF Annotation*. The third is **[C].Function Integration**, which includes *[C1].Integration with File System*, *[C2].Integration with Ontology*, and *[C3].Integration with Pre-annotation Model*.The last is other functions such as **[D].Team Annotation** and **[E].Statistics**. A deeper description of these metrics is given in Definition of Evaluation Metrics.

The comparison results are shown in Tab.1. As shown in Tab.1, Autodive is superior to most active tools in the function of onsite PDF annotation. PDFAnno, PAWLS, and INCEpTION have functions for PDF annotation. However, PDFAnno has not been maintained for over 3 years. Compared with PAWLS, Autodive provides more integration functions with file systems and NER models which also depends on onsite annotation mode. Autodive provides a different integration mode with pre-annotation models and a more intuitive statistics view when compared with the latest version of INCEpTION.

	Availability		Onsite		Integration			Team	Statistics
Tools	**[A1]**	**[A2]**	**[B1]**	**[B2]**	**[C1]**	**[C2]**	**[C3]**	**[D]**	**[E]**
AnnIE	-	-	-	-	-	✓	-	-	-
Doccano	✓	✓	-	-	✓	✓	url	✓	✓
WebAnno	-	-	-	-	✓	✓	-	-	-
PDFAnno	-	-	✓	block	-	-	-	-	-
PAWLS	-	✓	✓	block	-	✓	✓	-	-
INCEpTION(*)	✓	✓	✓	text	✓	✓	url	✓	-
Autodive(**ours**)	✓(*)	✓	✓	text	✓	✓	✓	✓	✓

Table 1: Function Comparison of Text Annotation tools. Autodive will remain active (A1) and expand to support more formats, such as iamges and tables in documents. Onsite PDF Annotation (B2) have two different modes: annotating by drawing a block(**block**) and annotating by selecting a raw text(**text**). Integration with Pre-annotation Model (C3) have two different modes: directly using an external API(**url**) or integrated with a pre-trained model(✓). (*) the version of INCEpTION we compared is the latest stable version V26.8.

4.2 NER of Cu-based Electrocatalysts

A high-quality corpus of catalysts may assist domain scientists in the discovery of catalysts based on a descriptor-optimization (Tran and Ulissi, 2018; Zhong et al., 2020). In this case, a corpus of Cu-based electrocatalysts for CO_2 reduction has been generated using the presented tool. At the beginning of the process, one senior scientist creates an annotation project and served as the administrator of the project by **Project Management Module**. After creating the project, the scientist finds the literature that needs annotation and assigns the literature to potential annotators using **Resource Management Module**. At the same time, he designs the ontology of Cu-based electrocatalysts with this assist of **Ontology Management Module** . There are 5 postgraduates with experience in experimental catalysis who used the tool to construct the corpus using **Manual Annotation Module** and **Auto Annotation Module**. During the annotation process, the administrator and annotators can view the rate of progress at any time by **Statistic View**. After annotation, the senior scientist exports the corpus and review all the annotated entity and relationship. In this real case, the corpus contains a collection of 6,086 records extracted from 835 publications with nine types of knowledge, including material, regulation method, product, faradaic efficiency, cell setup, electrolyte, synthesis method, current density, and voltage. This annotated corpus can be accessed publicly(Wang, 2023)(Wang et al., 2023).

4.3 Auto Annotation of AI Dataset and Model

In this case, we try to demonstrate the ability of auto annotation. A basic auto annotation project requires related model prepared and continuous annotation data. In order to analyse the effect of **Auto Annotation Module** quantitatively, we designed an experiment with a poorly correlated public dataset and increased proportion of annotated data to evaluate the correctness of auto annotation, as in a real annotation project. Firstly we trained an annotation model using the SciERC(Luan et al., 2018) dataset that mainly focused on the field of artificial intelligence. After deploying the model to the Autodive backend, we chose a number of abstracts from publications in the field of artificial intelligence on paperswithcode.com to simulate the automatic annotation effect. We designed ontology with "Model" and "Dataset" entities. All data contains 7,420 "Datasets" entities and 42,696 "Model" entities.

Training and updating the NER model during the annotation process helps fit the annotation project and improve the correctness. Tab.2 displays the results of this simulation. The zero-shot shows the correctness without any "annotation" data. With the updates of the auto annotation model through increasing sample size, the increased correctness of auto annotation module shows effectiveness of Auto Annotation Module, which helps annotators train their auto annotation models. As we can see, well integrated auto annotation model might meet the needs of scientific literature annotation, and it may perform better in specific projects.

5 Related Work

Mariana Neves(Neves and Ševa, 2021) gave a comprehensive review of existing document annotation tools. It splits the criteria of document annotation

	Sample Size			
Tools	**zero-shot**	**0.2**	**0.5**	**0.8**
Autodive	0.32	0.55	0.82	0.90

Table 2: Experiment Result

tools into four categories, which are publication, technical, data, and functional. In this review, it rates WebAnno(Yimam et al., 2013) as the best tool, which also extends to a new human-in-the-loop tool INCEpTION(Klie et al., 2018). It also mentioned that the top two missing functions of current tools are the support of document-level annotation, integration with existing corpus and pre-annotation, especially model-based pre-annotation.

Kinds of new annotation tools are reported in recent years. Many tools focus on specific tasks or functions, such as TeamTat(Islamaj et al., 2020), QuickGraph(Bikaun et al., 2022), DoTAT(Lin et al., 2022) and FAST(Kawamoto et al., 2021). Both task specified and common document annotation tools such as WebAnno(Yimam et al., 2013), Doccano(Nakayama et al., 2018), AnnIE(Friedrich et al., 2021) and TS-ANNO(Stodden and Kallmeyer, 2022) need a pre-process that convert document to pure text, which is a time-consuming work. By consulting with domain experts, the converter process also causes confuses reading, especially in the typeset document such as scientific literature.

As for annotation tools for PDF documents, PDFAnno(Shindo et al., 2018) and PAWLS(Neumann et al., 2021) are the two most relevant tools with our present tool. PDFAnno converts PDF documents into pure text without retaining PDF structure information, whose annotation mode is similar to our tool. However, it has not been maintained for over 3 years. PAWLS is a recent tool that supports PDF annotation with labels and structure. It has the advantage of annotation the meta or structural information by drawing a bounding box rather than selecting raw text. Autodive is inspired by PAWLS in the requirement of PDF annotation and surpasses it in integration function and annotation mode.

6 Discussion

We created Autodive, a collaborative scientific literature annotation tool that offers a comprehensive solution for the whole lifecycle of annotation, especially for scientific literature annotation. In addition, we provide automated annotation for annotators and can integrated with a variety of NER models. We found that Autodive can not only be utilized for scientific literature, but also for any editable PDF file. Also, Autodive is released as an open source project under Apache 2.0 license.

Autodive also has some limitations. First, accuracy of NER models. For text annotation projects, a more accurate NER model can ensure the accuracy of auto annotation. Second, collision in teamwork. Reduce annotation disputes in annotation projects caused by diverse knowledge perspectives between annotators and obtain more accurate annotation data. Third, there are PDF files that cannot be changed. Some obsolete or illegible PDF documents are difficult to process and cannot be used in the current Autodive version.

In our future project, we intend to expend the scope of literature annotation to include graphics and tables in addition to text. Additionally, autodive will support more file types, including the ability to download in JSON/CoNLL format and to upload plain text and pictures.

Acknowledgements

We thank all anonymous reviewers. This work is supported by the National Key Research and Development Plan of China under Grant No. 2022YFF0712200 and 2022YFF0711900, the Natural Science Foundation of China under Grant No. 61836013, Beijing Natural Science Foundation under Grant No. 4212030, Youth Innovation Promotion Association CAS.

References

Tyler Bikaun, Michael Stewart, and Wei Liu. 2022. Quickgraph: a rapid annotation tool for knowledge graph extraction from technical text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Benjamin Burger, Phillip M Maffettone, Vladimir V Gusev, Catherine M Aitchison, Yang Bai, Xiaoyan Wang, Xiaobo Li, Ben M Alston, Buyi Li, Rob Clowes, et al. 2020. A mobile robotic chemist. *Nature*, 583(7815):237–241.

Javier Corvi, Carla Fuenteslópez, José Fernández, Josep Gelpi, Maria-Pau Ginebra, Salvador Capella-Guitierrez, and Osnat Hakimi. 2021. The biomaterials annotator: a system for ontology-based concept annotation of biomaterials text. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 36–48.

Kevin Cruse, Amalie Trewartha, Sanghoon Lee, Zheren Wang, Haoyan Huo, Tanjin He, Olga Kononova, Anubhav Jain, and Gerbrand Ceder. 2022. Text-mined dataset of gold nanoparticle synthesis procedures, morphologies, and size entities. *Scientific Data*, 9(1):1–12.

Niklas Friedrich, Kiril Gashteovski, Mingying Yu, Bhushan Kotnis, Carolin Lawrence, Mathias Niepert, and Goran Glavaš. 2021. Annie: an annotation platform for constructing complete open information extraction benchmark. *arXiv preprint arXiv:2109.07464*.

Rezarta Islamaj, Dongseop Kwon, Sun Kim, and Zhiyong Lu. 2020. Teamtat: a collaborative text annotation tool. *Nucleic acids research*, 48(W1):W5–W11.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Shunyo Kawamoto, Yu Sawai, Kohei Wakimoto, and Peinan Zhang. 2021. Fast: fast annotation tool for smart devices. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 372–381.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *proceedings of the 27th international conference on computational linguistics: system demonstrations*, pages 5–9.

Yupian Lin, Tong Ruan, Ming Liang, Tingting Cai, Wen Du, and Yi Wang. 2022. Dotat: a domain-oriented text annotation tool. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–8.

Hugo López-Fernández, Miguel Reboiro-Jato, Daniel Glez-Peña, Fernando Aparicio, Diego Gachet, Manuel Buenaga, and Florentino Fdez-Riverola. 2013. Bioannote: a software platform for annotating biomedical documents with application in medical learning environments. *Computer methods and programs in biomedicine*, 111(1):139–147.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: text annotation tool for human. Software available from https://github.com/doccano/doccano.

Mark Neumann, Zejiang Shen, and Sam Skjonsberg. 2021. Pawls: Pdf annotation with labels and structure. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, page 258–264.

Mariana Neves and Jurica Ševa. 2021. An extensive review of tools for manual annotation of documents. *Briefings in bioinformatics*, 22(1):146–163.

Hiroyuki Shindo, Yohei Munesada, and Yuji Matsumoto. 2018. Pdfanno: a web-based linguistic annotation tool for pdf documents. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Regina Stodden and Laura Kallmeyer. 2022. TS-ANNO: an annotation tool to build, annotate and evaluate text simplification corpora. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 145–155, Dublin, Ireland. Association for Computational Linguistics.

Matthew C Swain and Jacqueline M Cole. 2016. Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of chemical information and modeling*, 56(10):1894–1904.

Kevin Tran and Zachary W Ulissi. 2018. Active learning across intermetallics to guide discovery of electrocatalysts for co2 reduction and h2 evolution. *Nature Catalysis*, 1(9):696–703.

Ludi Wang. 2023. A corpus of CO2 Electrocatalytic Reduction Process extracted from the scientific literature.

Ludi Wang, Yang Gao, Xueqing Chen, Wenjuan Cui, Yuanchun Zhou, Xinying Luo, Shuaishuai Xu, Yi Du, and Bin Wang. 2023. A corpus of co2 electrocatalytic reduction process extracted from the scientific literature. *Scientific Data*, 10(1):175.

Rongen Yan, Xue Jiang, Weiren Wang, Depeng Dang, and Yanjing Su. 2022. Materials information extraction via automatically generated corpus. *Scientific Data*, 9(1):1–12.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. Webanno: a flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6.

Miao Zhong, Kevin Tran, Yimeng Min, Chuanhao Wang, Ziyun Wang, Cao-Thang Dinh, Phil De Luna, Zongqian Yu, Armin Sedighian Rasouli, Peter Brodersen, et al. 2020. Accelerated discovery of co2 electrocatalysts using active machine learning. *Nature*, 581(7807):178–183.

A Definition of Evaluation Metrics

The definition of evaluation metrics among kinds of annotation tools are shown belows.

[A] Availability. **[A1] Activeness** : The annotation tool is still active and updated steadily. **[A2] Code Availability**: The annotation tool is open sourced.

[B] Onsite Support. **[B1] Onsite PDF Display**: The annotation tool provides a complete display with the structural information of the literature to suit the reading habits of scientific annotators. **[B2] Onsite PDF Annotation**: The annotation tool provides direct annotation on PDF documents, including **Text PDF Annotation** (annotate text directly) and **Block PDF Annotation** (annotate by drawing frames).

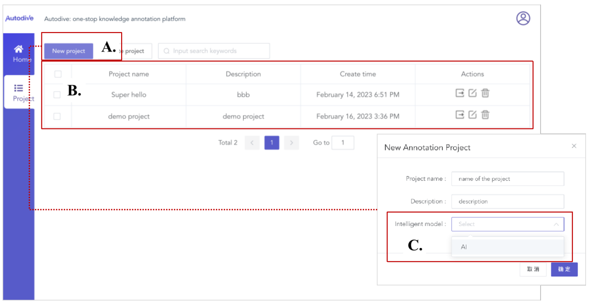
[C] Function Integration. **[C1] Integration With File System (Resource Management)**: The annotation tools has a file system which allows resource management such as uploading files. **[C2] Integration With Ontology**: The annotation tool enables the definition of ontology such as knowledge graph. **[C3] Integration With Pre-Annotation Model**: The annotation tool offers pre-annotation such as dictionary matching.

[D] Team Annotation: The annotation tool enables team annotation and the management of annotation results for all.

[E] Statistics: The annotation tool provides statistic view of project.

B Page Design

B.1 Project Management



Project name	Description	Create time	Actions
Super hello	bbb	February 14, 2023 6:51 PM	
demo project	demo project	February 16, 2023 3:36 PM	

Total 2 < 1 > Go to 1

New Annotation Project

Project name : name of the project

Description : description

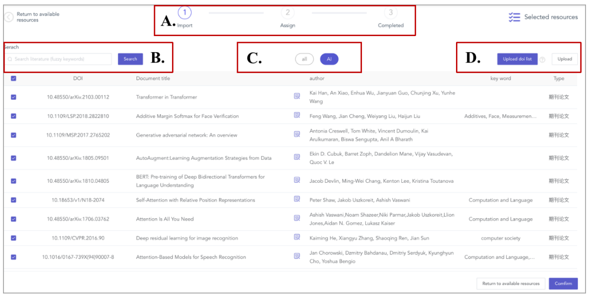
Intelligent model : Select

AI

取消 确定

Figure 6: Screenshot of project management. A is a button of creating new annotation project. Once click "New Project" button, user can input project and choose pre-annotation model,just like C. B shows the list of created annotation projects.

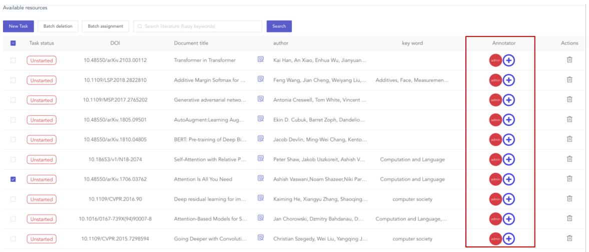
B.2 Resource Management



☑	DOI	Document title	author	key word	Type
☑	10.48550/arXiv.2103.00112	Transformer in Transformer	Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, Yunhe Wang		期刊论文
☑	10.1109/LSP.2018.2822810	Additive Margin Softmax for Face Verification	Feng Wang, Jian Cheng, Weiyang Liu, Haijun Liu	Additives, Face, Measuremen...	期刊论文
☑	10.1109/MSP.2017.2765202	Generative adversarial network: An overview	Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, Anil A Bharath		期刊论文
☑	10.48550/arXiv.1805.09501	AutoAugment:Learning Augmentation Strategies from Data	Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, Quoc V. Le		期刊论文
☑	10.48550/arXiv.1810.04805	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova		期刊论文
☑	10.18653/v1/N18-2074	Self-Attention with Relative Position Representations	Peter Shaw, Jakob Uszkoreit, Ashish Vaswani	Computation and Language	期刊论文
☑	10.48550/arXiv.1706.03762	Attention Is All You Need	Ashish Vaswani,Noam Shazeer,Niki Parmar,Jakob Uszkoreit,Llion Jones,Aidan N. Gomez, Lukasz Kaiser	Computation and Language	期刊论文
☑	10.1109/CVPR.2016.90	Deep residual learning for image recognition	Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun	computer society	期刊论文
☑	10.1016/0167-739X(94)90007-8	Attention-Based Models for Speech Recognition	Jan Chorowski, Dzmitriy Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, Yoshua Bengio	Computation and Language,...	期刊论文

Return to available resources Confirm

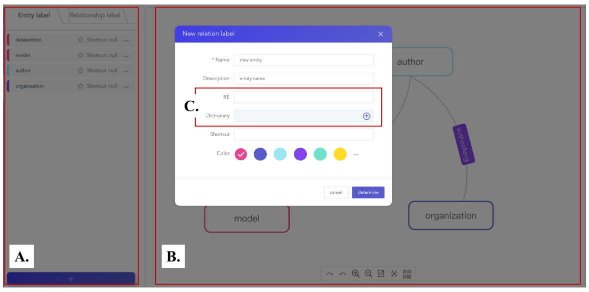
Figure 7: Screenshot of resource management. A shows the three steps of literature search and assign. B, C, D gave different ways to find or import literature files, such as direct search, using file tags or direct upload.



☐	Task status	DOI	Document title	author	key word	Annotator	Actions
☐	Unstarted	10.48550/arXiv.2103.00112	Transformer in Transformer	Kai Han, An Xiao, Enhua Wu, Jianyuan...		admin ⊕	
☐	Unstarted	10.1109/LSP.2018.2822810	Additive Margin Softmax for ...	Feng Wang, Jian Cheng, Weiyang Liu,...	Additives, Face, Measuremen...	admin ⊕	
☐	Unstarted	10.1109/MSP.2017.2765202	Generative adversarial netwo...	Antonia Creswell, Tom White, Vincent ...		admin ⊕	
☐	Unstarted	10.48550/arXiv.1805.09501	AutoAugment:Learning Aug...	Ekin D. Cubuk, Barret Zoph, Dandelio...		admin ⊕	
☐	Unstarted	10.48550/arXiv.1810.04805	BERT: Pre-training of Deep Bi...	Jacob Devlin, Ming-Wei Chang, Kento...		admin ⊕	
☐	Unstarted	10.18653/v1/N18-2074	Self-Attention with Relative P...	Peter Shaw, Jakob Uszkoreit, Ashish V...	Computation and Language	admin ⊕	
☑	Unstarted	10.48550/arXiv.1706.03762	Attention Is All You Need	Ashish Vaswani,Noam Shazeer,Niki Par...	Computation and Language	admin ⊕	
☐	Unstarted	10.1109/CVPR.2016.90	Deep residual learning for im...	Kaiming He, Xiangyu Zhang, Shaoqing...	computer society	admin ⊕	
☐	Unstarted	10.1016/0167-739X(94)90007-8	Attention-Based Models for S...	Jan Chorowski, Dzmitry Bahdanau, D...	Computation and Language,...	admin ⊕	
☐	Unstarted	10.1109/CVPR.2015.7298594	Going Deeper with Convoluti...	Christian Szegedy, Wei Liu, Yangqing J...	computer society	admin ⊕	

Figure 8: Screenshot of annotation task assignment. Highlighted part shows the function of assignment. Administrator of one project have permission to assign literature to different annotators.

B.3 Ontology Management


Entity label
Relationship label
datasettest ☆ Shortcut :null ...
model ☆ Shortcut :null ...
author ☆ Shortcut :null ...
organization ☆ Shortcut :null ...
A.
B.
C.
author
authorAorg
model
organization
自动排版

New relation label

* Name: new entity

Description: entity name

RE

Dictionary

Shortcut

Color

cancel determine

Figure 9: Screenshot of ontology management. A displays the list of all generated entities and relationships. B visualizes the constructed ontology. C is the interactive interface of entity design. In this interface, project owner can define a regular expression or upload a dictionary, so that Autodive can pre-annotate by the regular expression or dictionary.

B.4 Annotation

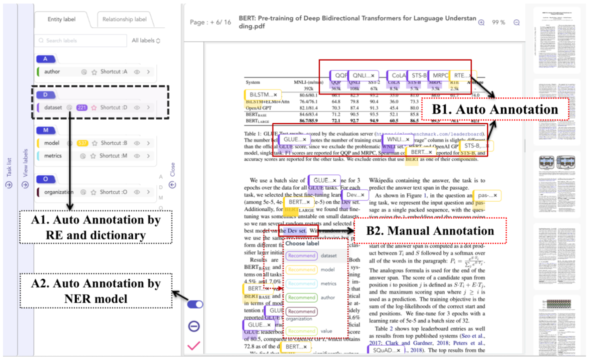


Figure 10: Screenshot of annotation. Similar to Fig. 4A1 and A2 demonstrate usage of three auto annotation functions, while B1 and B2 show the results of both auto annotation(B1) and manual annotation(B2).

B.5 Statistic View

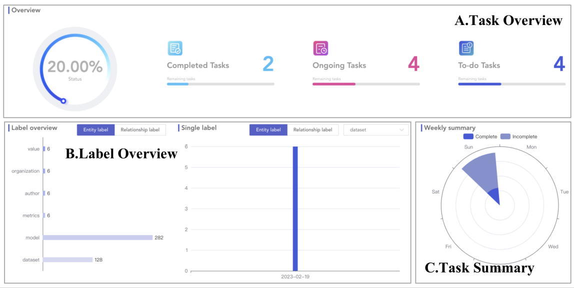
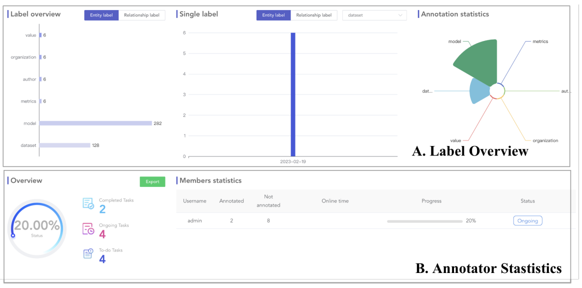


Figure 11: Screenshot of Personal View. Personal view allows annotator to see his/her own annotation status. A shows an overview of assigned task, including number of completed tasks, ongoing tasks and to-do tasks. B shows all the number of annotated or extracted entities and relationships. C shows a comparison between completed tasks and incomplete tasks on weekly view.



Username	Annotated	Not annotated	Online time	Progress	Status
admin	2	8		20%	Ongoing

Figure 12: Screenshot of Project View. Project view allows administrator to see overall annotation status and each annotator's progress. A shows summarized overview of all annotators. B shows each annotator's annotation progress.

C Demo Access

A live demo of Autodive is available at http://autodive.sciwiki.cn. The live demo provides two languages, English and simplified Chinese, which depends on the language setting of the web browser. It is allowed to convert language via the "head" button in the top right corner. We also provide a video demo at http://autodive.sciwiki.cn/introVideo/introduce-v1.0.mp4. The source code is available at https://github.com/Autodive. We provide a test account in the live demo using username *test* and password *autodive*. In this demo, we linked a resource library with dozens of open access (OA) scientific literature.

To use Autodive in production environment, users can also sign up with their own email address, upload their own literature that needs annotation, create personalized annotation project, assign an annotator, and complete their annotation task. Users can also deploy Autodive in their own server with personal literature collections.