Doctron: A web-based collaborative annotation tool for ground truth creation in IR

Ornella Irrera
ornella.irrera@unipd.it
University of Padua
Padua, Italy

Stefano Marchesin
stefano.marchesin@unipd.it
University of Padua
Padua, Italy

Farzad Shami
farzad.shami@studenti.unipd.it
University of Padua
Padua, Italy

Gianmaria Silvello
gianmaria.silvello@unipd.it
University of Padua
Padua, Italy

Abstract

In Information Retrieval (IR), ground truth creation is a crucial yet resource-intensive task that relies on human experts to build test collections – essential for training and evaluating retrieval models. Large-scale evaluation campaigns, such as TREC and CLEF, demand significant human effort to produce reliable, high-quality annotations. To ease this process, tailored annotation tools are pivotal to supporting assessors and streamlining their workload.

To this end, we introduce Doctron, a web-based, dockerized annotation tool designed to streamline ground truth creation for IR tasks. Doctron enables the annotation of both textual documents and images. It supports annotating textual passages, identifying relationships, tagging and linking entities, evaluating document relevance to a topic with graded labels, and performing object detection. It offers a collaborative environment where teams can work with defined user roles and permissions. The integration of Inter Annotator Agreement (IAA) measures helps to identify inconsistencies between annotators, thereby ensuring the reliability and high quality of the annotated ground truth data.

CCS Concepts

• **Information systems** → **Information systems applications**; **Information retrieval**.

ACM Reference Format:
Ornella Irrera, Stefano Marchesin, Farzad Shami, and Gianmaria Silvello. 2024. Doctron: A web-based collaborative annotation tool for ground truth creation in IR. In *Proceedings of The 48th International ACM SIGIR Conference on Research and Development in Information Retrieval July 13-18, 2025 (SIGIR '25).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

SIGIR '25, Padova, Italy,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Ground truth creation of large corpora is crucial for Information Retrieval (IR), providing an essential basis for training, evaluating, and improving search systems. To this end, manual annotation, involving human assessors labeling documents, is still the de facto standard, contributing to the robustness and reliability of test collections central to the IR progress. The creation of a ground truth is a complex workflow, typically involving multiple stakeholders, such as domain experts, data annotators, and project coordinators, each contributing to different process phases. These phases include establishing annotation guidelines, selecting the appropriate annotation tool, preprocessing the dataset to make it compliant with the tool, performing the actual labeling, ensuring consistency by resolving discrepancies, evaluating annotation quality, and revising the annotations if needed. Undeniably, creating a ground truth is a time-consuming and resource-intensive task, often representing a bottleneck in the overall project workflow [21, 25]. In this context, choosing the appropriate annotation tool can lead to significant time savings, streamlining numerous tasks and reducing the effort required from assessors. Several factors influence this choice, including the objectives and complexity of the task, the supported input and output formats, the tool ease of use and installation, and the availability of collaborative features that allow multiple users to work together in real-time, enhancing the annotation process.

In recent years, various reviews evaluated the effectiveness of annotation tools, compared their features, and helped researchers choosing the most suitable option for their specific tasks [4, 21, 22]. Some annotation tools are tailored to specific domains, addressing the unique requirements of particular research areas. Notably, many of these tools are specifically designed for the biomedical domain [2, 8, 10, 11, 26], usually supporting tasks like document classification, Named Entity Recognition (NER), Named Entity Recognition and Linking (NER+L), and relation annotation, while integrating domain-specific ontologies. General-purpose tools [6, 9, 12, 20, 23, 24, 28, 31, 32], on the other hand, are highly customizable to suit different research domains, offering a wide range of features suitable for various annotation tasks. In addition to textual data, they often support images, videos, and audio, making them adaptable to multimodal tasks.

However, most of the available tools are designed for general annotation tasks rather than specifically for IR. As a result, they

often lack features to assess the relevance of documents to specific topics or to annotate passages relevant to those topics.

Contributions. To overcome this limitation, in this paper we introduce Doctron, a web-based dockerized annotation tool that allows annotators to collaboratively annotate collections of documents. Doctron represents an advancement in annotation tools, particularly in its *focus on IR*. It is explicitly designed to streamline ground truth creation for IR, allowing users to annotate documents based on their relevance to topics and utilize graded labels. Additionally, Doctron introduces a *robust collaborative environment* through role-based permissions, facilitating efficient teamwork among annotators, reviewers, and administrators. This framework enhances collaboration compared to other existing tools, which often lack such advanced functionalities.

In terms of features, Doctron goes beyond standard capabilities offered by other tools, which may only support multilabel annotation or basic NER. It extends its offerings to include *passage annotation, object detection, and the ability to handle both textual and image data*, effectively addressing diverse needs in IR annotations.

Doctron also incorporates *Inter Annotator Agreement (IAA) metrics*, allowing teams to actively evaluate the consistency of annotations across different users. This integration offers a measure of reliability that many existing tools fail to effectively track. Moreover, Doctron integrates `ir_dataset` [18], allowing users to upload and (re-)annotate test collections typically adopted by the IR community. The user-friendly customization of Doctron sets it apart; with its *intuitive interface*, it allows users – even those lacking expertise in annotation – to easily customize templates and workflows. This is in contrast to other tools that often require extensive setup and configuration. Finally, Doctron is *completely free and open-source*, providing extensive functionalities without hidden costs. This approach differs from many other platforms that operate under freemium models.

Doctron can be accessed online at https://doctron.dei.unipd.it/ as a cloud-based service.[1] Users also have the option to download a Dockerized version of the tool from GitHub[2], which can be installed and operated on a local server for various project purposes.

Outline. The rest of this paper is organized as follows. Section 2 provides an overview of the available annotation tools; Section 3 describes Doctron, focusing on its user interface, architecture, and functionalities; Section 4 presents the qualitative and quantitative evaluation, comparing Doctron with other annotation tools. Finally, Section 5 draws some final remarks and outlines future work.

2 Related work

The annotation tools developed over the years can be roughly classified into domain-specific and general-purpose.

Domain-specific tools. Domain-specific tools are tailored to fulfill the domain's requirements, implementing ad hoc functionalities. Most of the currently available tools are designed for the biomedical domain. They include MyMiner [26], which is an offline annotation tool that supports multilabel annotation, NER and NER+L. TeamTat [11], the last version of EzTag [16] and BioQRator [15], which is a text annotation tool developed for performing NER+L and relation extraction. A powerful tool is TagTog [2] that supports various document annotation templates and offers more robust automatic annotation, IAA metrics, and collaboration. MedTAG [8] is a dockerized collaborative annotation tool that supports multilabel annotation, and NER+L. The most recent tool is MetaTron [10] supporting multilabel annotation, NER, NER+L, and the annotation of relationships. It provides various statistics through various IAA measures and includes user roles and annotation review functionalities. All these tools do not provide the capability to annotate documents about a specific topic or assess a document's relevance to that topic. Furthermore, these tools typically lack the flexibility to include annotation dimensions that extend beyond mentions annotation, NER(+L), and relation extraction.

General-purpose. General-purpose annotation tools address various tasks across various data types, including text, images, audio, and video. For instance, Doccano [20] is a free, dockerized collaborative tool that facilitates multilabel annotation, NER, and object detection. However, it lacks native support for multigraded relevance annotation. Although workarounds can be implemented to include this feature, they may compromise usability and increase the number of clicks required to annotate a topic-document pair. Tools like brat [28] and PrettyTags [6] enable semantic annotations through NER and relationship annotation but do not provide customizable options for dimensions beyond these functions. Similarly, Yedda [32] is focused solely on annotating textual spans and lacks broader annotation dimensions. LightTag [24], while tailored for collaborative text annotation, specializes in NER and label management without an online version as of February 2025. INCEpTION [12], the advanced version of WebAnno [33], tackles complex workflows but is more complicated and does not directly support relevance annotation. GATE Teamware [1] and its updated version, GATE Teamware 2 [31], focus exclusively on document classification and lack support for annotating textual spans. POTATO [23] also covers various tasks like sentiment analysis, but does not enhance annotation dimensions relevant to IR. Commercial platforms such as SuperAnnotate[3], Ubiai[4], and Prodigy[5] integrate active learning and collaboration features but mostly come with associated costs. BasicAI[6] and PDFAnno [27] primarily focus on image/video annotation and PDF comments and do not offer free usage. In contrast, Doctron provides a comprehensive framework for annotating documents based on their relevance to specific topics. It incorporates customizable annotation dimensions that extend beyond standard functions like NER and relation extraction, positioning it as a vital solution for IR tasks. However, general-purpose tools are not readily adaptable to handle tasks related to IR involving the typical topic-document relevance assessment. Although they support document annotation, they require significant configuration to incorporate topic-document relevance or passage identification, making them not straightforward choices for IR-specific tasks. Only two tools in the literature implement features explicitly designed

[1] Access is currently granted only to reviewers with login: demo and password: demo.
[2] https://github.com/meta-doc-dev/DocTron

[3] https://www.superannotate.com/
[4] https://ubiai.tools/
[5] https://prodi.gy/
[6] https://www.basic.ai/

for use in IR annotation tasks. DocTAG [9] is a dockerized annotation tool explicitly designed for IR and is one of the few tools that enables the relevance assessment of a set of documents about a set of topics selected and uploaded by the annotator. It allows users to annotate documents collaboratively and supports multilabel annotation, passage annotation, and entity linking. However, it does not implement annotation curation and is not role-based. Label-Studio [29] is a collaborative manual annotation tool not designed explicitly for IR. However, like DocTAG, it allows configuring a set of topics used in document retrieval and relevance assessment for a specific query. For this reason, we categorize it as IR-specific. LabelStudio supports annotating textual documents, images, videos, and audio and promotes automatic data labeling. However, features such as collaboration, user roles, automatic labeling, and annotation reviews require a payment subscription.

3 Doctron

Doctron is based on three main concepts: document, topic, and annotation template. A *document* is the default annotable unit, which can be a piece of text or an image. A *topic* is a structured description of an information need. Topics can be textual information characterized by topic number, title, description, narrative, or images. The *annotation template* identifies how a topic-document pair is annotated and describes the specific annotation type and the criteria for assessing relevancy. In the annotation for passage retrieval, for example, the passage is typically labeled according to its relevance to the topic.

Doctron is cloud-based and distributed as a Docker container for easy deployment. It is platform-independent and can be installed in any hosting environment. It ensures privacy and security, giving research groups or organizations full control over data access and sharing.

Doctron was built by revising, updating, and extending MetaTron [10], which targets solely the biomedical domain. We built on MetaTron's layout, functionalities, and data schema. On top of this code base, we introduced topic-document pair annotation, new templates like graded relevance, passages annotation, image support, and object detection. Additionally, we completely redesigned the collection creation process, dashboard, and IAA modules.

Annotation workflow. In Doctron, there are three types of users: annotator, reviewer, and administrator. The *annotator* adds annotations to documents, providing their insights and information based on the predefined annotation templates. The *reviewer(s)* have the highest level of expertise; they have full access to the annotators' work and can update the entire set of annotations to ensure quality and consistency. The *administrator(s)* manage the collection, oversees user roles, defines the set of annotation guidelines, and configures annotation templates and settings. They have access to all the annotations and the work of the reviewers. They can modify and update the annotations of all the annotators – reviewers included – and keep track of the annotation status and progress.

In Figure 1, we illustrate the statistical and annotation workflows of Doctron, describing the process through which the annotations are refined and finalized. The annotation workflow (blue lines in Figure 1) begins with the annotators analyzing and annotating a designated set of topic-document pairs. Their initial annotations

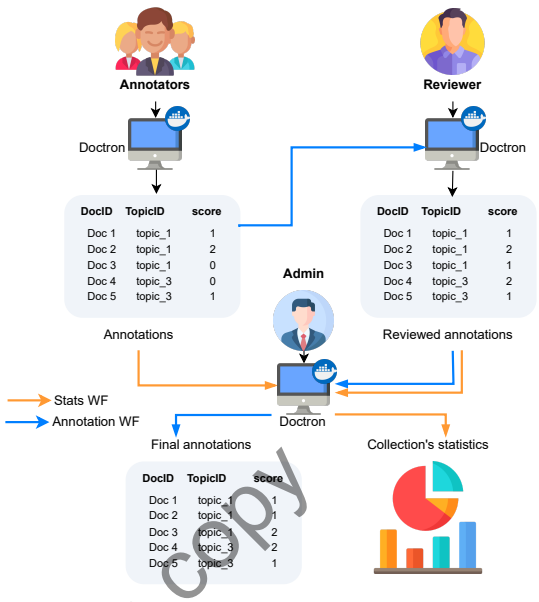

Figure 1: Doctron workflows are as follows: the statistics workflow, shown in orange, is dedicated to computing relevant statistics for the collection, while the workflow for generating the final set of annotations is represented in blue.

might undergo a review process, during which one or more reviewers examine the work, making necessary updates to produce a refined version. In the final stage, administrators can take the entire set of annotations, fix and review them, and generate the final annotation set. The statistics workflow (orange lines in Figure 1) considers the annotations made by the annotators and those of the reviewers. These are used to generate the collections' statistics, enhancing the overall analysis and evaluation of the annotation process. Using both the initial annotations provided by the annotators and those of the reviewers, the statistical workflow ensures the identification of patterns, inconsistencies, and potential areas for improvement. It allows the identification of which documents need to be reviewed.

Architecture. The architecture of Doctron follows a three-tier design. It includes: (i) the *data layer* implemented with a PostgreSQL database storing documents, topics and annotations, guaranteeing the persistence and integrity of annotated data; (ii) the *business logic* implemented with Django, [7] a Python web framework that is responsible for handling core application functionalities such as processing requests from the front-end and interacting with the PostgreSQL database to store and retrieve documents, topics, and annotated data; (iii) the *presentation layer* developed using React.js, [8] which offers an intuitive and interactive platform for annotators to perform annotation tasks.

User interface. Doctron's user interface has been designed to be intuitive and to facilitate and speed up annotators' work. On login, the user is asked to provide an annotation template. The system

[7] https://www.djangoproject.com/
[8] https://react.dev/

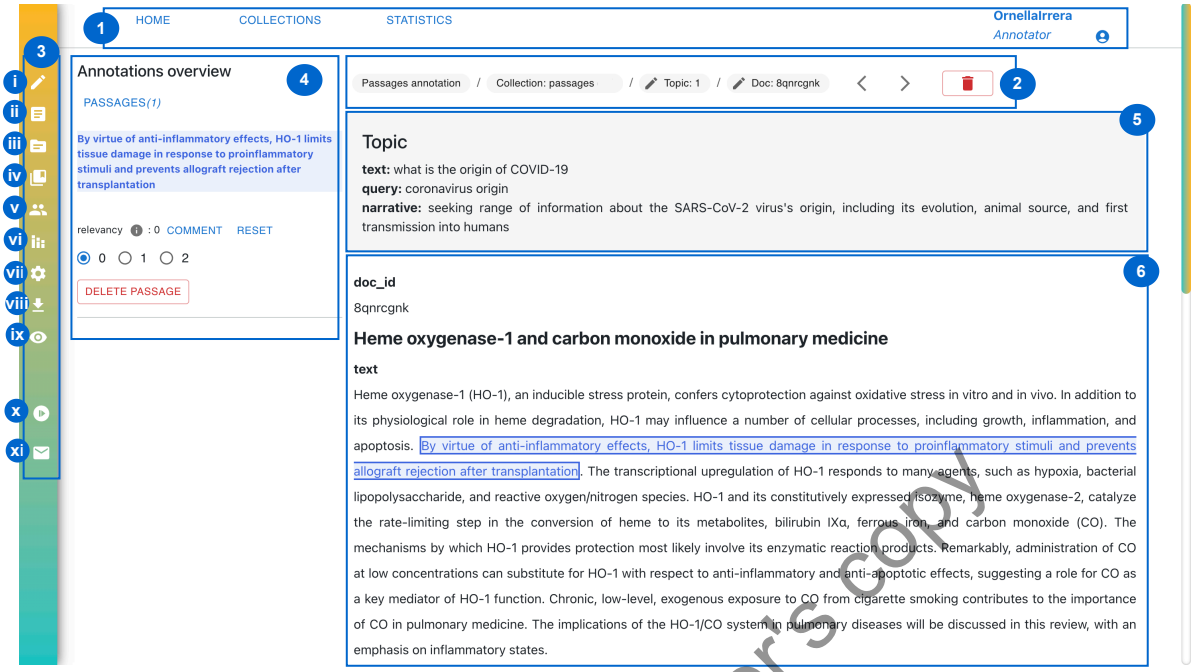

Figure 2: Doctron user interface with passages annotation template.

loads the available collections created for the required template and loads the latest opened document and topic. The main interface is illustrated in Figure 2; we report the passage retrieval annotation template in this case.

The main header ❶ allows the user to navigate to the *Home*, *Collections*, and *Statistics* pages. The *Home* page redirects the user to the annotation interface. On the *Collections* page, annotators can access all the collections shared with them. The *Statistics* page provides the annotators access to their statistics and IAA metrics.

The document header ❷ displays (from left to right) the collection's annotation template, the collection's name, the topic's identifiers, and the document's identifier. By clicking on the topic and document identifiers, it is possible to add some notes about the topic and document that, in turn, can be considered by the collection administrators. Two arrows allow the user to navigate to the next and previous documents in the collection. Finally, the annotator can delete all the document's annotations.

The left sidebar ❸ is designed to offer a range of functionalities that can be easily accessed directly from the main annotation interface, reducing the need for multiple actions. These functionalities include (from top to bottom): (i) role switching (e.g., changing from *annotator* to *reviewer* or *admin*), (ii) access to documents lists with the option to switch between them, (iii) access to topics list, (iv) access to the collections list for the selected annotation template, (v) accessing the list of members who have annotated the current topic-document pair (if any), with the option to view their set of annotations, (vi) access to the user's document statistics, (vii) system settings (e.g., modifying entity and tag colors, line height, and font size), (viii) annotations download, allowing the user to choose the format and decide whether to download annotations for the current document or the entire collection, (ix) hide or show textual parts, (x) a demo with tutorial videos, and (xi) access to instructions.

The left panel ❹ provides an overview of the user's annotations for the current topic-document pair. It enables the user to interact with the annotations by adding, updating, or removing them and including comments that explain or justify each annotation, which the administrators and reviewers can view.

The main area of the user interface is focused on the topic-document pair being annotated. Information about the topic (such as ID, title, narrative, and description for text-based topics, or ID and image for image-based topics) is displayed in the gray panel ❺, while the document to be annotated (whether text or image) is shown below the topic in ❻.

Annotation templates. In Doctron, an *annotation template* provides structured guidelines for annotators to assess and label documents based on specific topics. Each template outlines the types of annotation that can be applied to a document. Doctron offers seven templates described below.

Graded labeling. This annotation template consists of labels (e.g., relevance) associated with a range of values (e.g., integers from 0 to 3). Annotators must assign a value to each label, indicating that the label's value is for the document concerning the specific topic. This template can be used for text-based and image-based topics and documents, as the graded labeling is applied at the document level and is not limited to specific sections of a text or portions of an image. The label "relevancy" and its corresponding ranges (0,1,2) are displayed in panel ❹ in Figure 2, allowing the annotators to update or remove the assigned grade directly from the interface while viewing the document. An example of test collection that

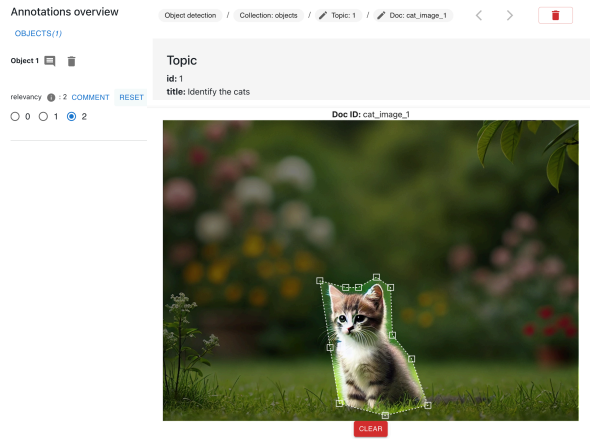

Figure 3: Doctron object annotation interface. The cat is the identified object, while the graded label is assigned on the panel on the left.

can be annotated with this template is the TREC Robust Track [30], where each document is assessed with graded relevance for a given topic. We used TREC Robust 2004 as a reference collection for designing this template.

Passage annotation. Passages are brief sections—comprising one or more sentences—within a textual document. Doctron allows users to identify these passages and annotate them accordingly. Users can select a passage by dragging from the first to the last character and then associate a graded label that indicates the significance of that passage concerning a topic. Additionally, Doctron supports loading pre-annotated passages, simplifying graded label assignments. An example of passage annotation is illustrated in Figure 2. The annotation template for passages was designed with MS MARCO[9] in mind, which is used in the TREC Deep Learning (DL) Track [5], where the goal is to classify passages based on their relevance to a query.

Object annotation. Object annotation involves identifying a specific area within an image (object detection). This process requires selecting the object's perimeter by clicking on and assigning one or more graded labels to the selected region. Similarly to the passage annotation template, Doctron also allows for the loading of object coordinates within documents, enabling annotators to focus on assigning graded labels to the objects. Figure 3 illustrates the object detection and annotation process. The detected object is the cat, annotated as highly relevant (relevance: 2).

Entity tagging (NER). Entity tagging is the process of identifying *mentions* – specific words or phrases in a text that refer to particular entities – and labeling them with predefined *tags*, which are the categories to associate with the mentions. This process is also known as NER. In Doctron, a mention is first identified by selecting one or more words, and then the tags are assigned. A mention can be associated with one or more tags. Additionally, Doctron allows users to define new tags on the fly without the need to load the entire set of tags in advance.

[9]https://microsoft.github.io/msmarco/TREC-Deep-Learning

Entity linking (NER+L). Entity linking associates the mentions identified in the textual document with their corresponding entries in a knowledge base, ontology, or concept schema, such as Wikipedia, Wikidata, or a custom database. Once an entity mention is extracted from the text, entity linking resolves ambiguity by associating the mention with a unique and specific entry in the knowledge base. Doctron also enables the specification of new entities on the fly, without loading the entire ontology in advance, in the same way as the previous template.

Relationship annotation. A relationship comprises three main components: a subject, a predicate, and an object, with the ties always starting from the subject and ending at the object. Each relationship component can be represented by an ontological concept, a tag, or a textual mention (either with or without associated concepts and tags).

Fact annotation. Facts are triples consisting of a subject, a predicate, and an object. All components are ontological concepts or tags, and none are textual mentions in the document. Like graded labeling, fact annotation is at the document level and independent of specific portions of the document.

Collection management and customization. In Doctron, a collection comprises a set of annotators (at least one), a set of documents (either textual or images), a set of topics (either textual or images), and is associated with an annotation template. Each annotator/reveiwer in Doctron can annotate multiple collections of documents.

Input format. In Doctron, topics and documents are schema-free, meaning they do not require a predefined structure or a fixed set of information. A wide range of input formats are allowed, with textual topics uploaded in JSON and image-based topics in PNG or JPG. Doctron also allows for subtopics, enabling hierarchical organization for detailed annotations. Documents can be uploaded in JSON, CSV, TXT, or PDF formats, thanks to GROBID [17] integration, which extracts structured data from PDFs. Additionally, Doctron supports importing pre-annotated data, allowing users to refine or build on existing annotations instead of starting from scratch.

Output format. Document annotations can be downloaded in TREC-like (i.e., `qrels` format) and custom formats. The TREC-like format structures annotations according to TREC guidelines, making it suitable for benchmarking and comparing retrieval system performance. This format typically includes document identifiers, annotation offsets, and relevance scores. In contrast, custom formats are tailored to each annotation template and are available in CSV and JSON.

API integration. The integration with `ir_datasets`[10] [18] allows users to upload collections by simply specifying their URL. Users can load documents directly from `ir_datasets`, which will be uploaded offline, independent of user interaction. Additionally, users can choose to upload their topics or use those provided by `ir_datasets` for the selected collection, offering full customization and allowing users to adapt the platform to their specific needs. Doctron also integrates with the PubMed REST API, enabling users to import abstracts from PubMed by specifying their PubMed IDs. The system automatically retrieves these abstracts and adds them

[10]https://ir-datasets.com/

to the collection, including metadata such as title, authors, year, and venue. This facilitates the creation of medical IR collections, like TREC CDS 14-16[11].

Annotation template customization. Doctron offers complete customization of annotation templates. Users can define the set of tags for entity tagging, the concepts to perform entity linking, providing the URLs pointing to concepts in the desired knowledge base, the predicates for relationship annotation, and the labels (along with their value ranges) for graded labeling, passages, and object annotation. New tags, concepts, predicates, and labels can be added anytime.

Annotation rounds. Collections can have multiple rounds of annotations, meaning users can revise or refine their annotations over time. Each round allows for updates or corrections based on reviewers', admins', or other annotators' feedback.

Collection modality. Collections can be configured in two modalities – i.e., collaborative and competitive – each determining what annotators can see and annotate. In *collaborative* mode, annotators can annotate all documents and view each other's annotations. Conversely, in *competitive* mode, annotators have no access to each other's annotations. This guarantees unbiased results by preventing external influence during the annotation process and ensures that evaluations are solely based on individual contributions.

Collaborative and competitive features. Depending on the modality assigned to the collection, Doctron provides some modality-based features designed to to enhance teamwork and streamline annotation. In both modalities, documents and topics can be partitioned among annotators to efficiently distribute workload, allowing simultaneous work on different parts of the collection and minimizing redundancy. In competitive modality, Doctron enables the creation of a *honeypot* – a set of documents selected by the collection administrator and assigned to all annotators. This common basis enables comparison of annotations, helping assess quality and ensure consistency. In crowdsourcing, the honeypot provides a standardized reference point, improving the reliability of the dataset and helping identify discrepancies in annotations.

When the collaborative mode is enabled, annotators can view each other's annotations and copy them into their workspaces. Additionally, Doctron includes an automatically generated ground truth based on majority voting. This feature creates a set of annotations for each document by selecting the ones made by more than half of the annotators. Annotators can copy this ground truth into their own workspaces as needed. This ground truth can serve as a starting point for annotators working on the collection or be used to identify ambiguous annotations.

Collection statistics and IAA features. From the *Statistics* tab in the main header ❶ in Figure 2, annotators can access the statistics related to the collections shared with them.

Individual statistics. Individual statistics offer an overview of the annotator's work, including details such as the number of annotated and unannotated documents for each topic and other information related to the document template. For each document annotated concerning a topic, the statistics provide some information that depends on the annotation template of the collection, such as the assigned labels and their corresponding grade (for graded labeling), the number of identified passages and their assigned graded labels (for passage annotation), the number of objects and their graded assigned labels (for object annotation), the number of tagged or linked mentions (for NER and NER+L), and the number of relations and facts (for relationship and fact annotation). Administrators can access the individual statistics of each annotator of the collection –reviewers included, to monitor their progress and reassign tasks if necessary.

Global statistics. Global statistics apply to all annotators within the collection, providing the same type of information as individual statistics but aggregated across all contributors. Additionally, for each annotated document, they give the number of annotators involved and their usernames.

Inter Annotator Agreement (IAA). IAA measures the level of agreement between annotators of a collection, offering insights into the annotations reliability. It helps detect inconsistencies among annotators and identifies documents or topics that may need further review by administrators or reviewers. In Doctron, IAA metrics are available only to the administrators and the collection reviewers. Doctron implements three IAA metrics to assess agreement: Cohen's Kappa, Fleiss's Kappa, and Krippendorff's Alpha.

Coehn's Kappa [3] is a statistical measure of inter-annotator agreement between two annotators. It is defined as: $\kappa = \frac{P_o - P_e}{1 - P_e}$, where P_o represents the observed agreement among the annotators, calculated as the proportion of instances where the annotators assign the same annotation, and P_e the probability that two annotators independently make the same annotation.

Fleiss's Kappa [7] is a statistical measure used to assess the agreement between multiple annotators annotating a set of documents. It generalizes Cohen's Kappa to more than two annotators. It is defined as $\kappa_F = \frac{\bar{P} - \bar{P_e}}{1 - \bar{P_e}}$, where $\bar{P}$ is the average observed agreement, and $\bar{P_e}$ represents the expected agreement by chance across multiple annotators. It reflects the likelihood that the annotators would agree by chance.

Krippendorff's Alpha [13] measure offers greater flexibility than Fleiss's Kappa, as it can handle missing data. This makes it particularly useful when one or more annotators have not annotated all the documents assigned to them. It is computed as $\alpha = 1 - \frac{D_0}{D_e}$, where D_0 and D_e represent the observed and the expected disagreements, respectively. The values of Coehn's Kappa, and Fleiss' Kappa range from -1 (complete disagreement) to 1 (perfect agreement) [19], while Krippendorff's Alpha range from 0 to 1 [14].

In Figure 4, we present the statistics dashboard of Doctron. It is possible to select a collection from the panel on the left (❶). The annotator can choose which statistics to visualize in the header ❷: Individual, Global statistics, or IAA metrics –In Figure 4, we report IAA metrics. In ❸ three cards provide an overview of the annotated documents. The table in ❹ reports the values for Fleiss' Kappa and Krippendorff's alpha measures for each document annotated for the selected topic (topic 7). In ❺, the Cohen's Kappa agreement for a topic-document pair is illustrated. Cohen's Kappa is represented as a symmetric matrix, where annotators are listed along both the rows and columns, and each cell indicates the level of agreement between the annotator in the corresponding row and the one in the corresponding column. The cells are shaded in a blue gradient with

[11] http://www.trec-cds.org

Figure 4: Overview of the statistics and IAA metrics dashboard.

six levels (also defined in [19]), ranging from white to dark blue, to reflect the increasing annotators agreement.

4 Evaluation

This section provides a qualitative and quantitative evaluation to compare Doctron with various annotation tools. The qualitative evaluation highlights the key functionalities of each tool, while the quantitative evaluation measures their performance in collection creation, multilabel annotation, passage annotation, NER.

Qualitative evaluation. In the qualitative analysis, we compare the functionalities of nine annotation tools with those of Doctron. The selection of tools was based on two criteria: (i) availability as either an online service or a locally installable version, and (ii) the provision of essential features for free or via a free demo. In addition, we focus on the features available in the free versions of the tools. Nine comply with our selection criteria: MetaTron, Doctag, Doccano, LabelStudio, TeamTat, INCEpTION, brat, TagTog, POTATO. We included MetaTron because it served as the foundation for the development of Doctron, and we show that Doctron offers more advanced functionalities. While TeamTat and TagTog are designed for the biomedical domain, they can be adapted for general purposes with minimal modifications. Although the online version of TagTog is no longer available, we still included it in our evaluation because, as of 2022, the online version was up and running and we used it in previous experiments. We excluded WebAnno from our evaluation since it is essentially the predecessor of INCEpTION. Tools like BasicAI, which are focused on labeling image-based data, were also excluded, as were LightTag, which is no longer available, and SuperAnnotate, UBIAI, and Prodigy, as they do not provide free versions or online demos. PrettyTags was not considered either, as it is currently unavailable.

We compared the annotation tools according to three criteria. **Technical criteria** concern the accessibility, usability, and availability of the tools. The criteria include (T1) the availability of open source code, (T2) free of charge, (T3) the ease of installation and use, and (T4) the availability online – entirely or as a demo. **Data criteria** concern the ability of the tools to manage various data sources, as well as different input, output, and annotation formats. They include: (D1) configurable annotation schema, (D2) support for the upload of pre-annotated data, (D3) configurable output data, (D4) configurable input data, (D5) PubMed integration, and (D6) `ir_dataset` integration or other custom IR data sources. **Functionalities criteria** concern the tools' features and capabilities for the annotation. They include (F1) topic-document annotation, (F2) multilabel annotation, (F3) graded labeling, (F4) passages annotation, (F5) images annotation and object detection, (F6) relationships or fact annotation, (F7) NER or NER+L, (F8) users and roles, (F9) IAA integration, (F10) ontologies, (F11) data privacy, (F12) multilingual support, (F13) guidelines definition, (F14) built-in predictions, (F15) keyboard shortcuts.

The evaluation results are summarized in Table 1. The qualitative analysis identifies Doctron as the most comprehensive solution, meeting 24 out of 25 criteria. INCEpTION and MetaTron fulfill 18 criteria, followed by Doccano with 17, TagTog and DocTag with 16, and LabelStudio and TeamTat with 15. brat and POTATO meet 13 and 12 criteria, respectively. LabelStudio is the only tool lacking open-source code (T1) and, notably, only four out of ten tools are fully available online or offer an online demo (T4). LabelStudio does not meet any technical criteria; it is not completely free and has a cumbersome installation process. Additionally, its setup for

Table 1: Qualitative evaluation of 10 annotation tools. A ✓is placed if the criterion is met. The rows in gray highlight tools providing support for IR tasks. Doctron is represented in the row light-blue.

	Technical				Data						Functionalities														
Tool	T1	T2	T3	T4	D1	D2	D3	D4	D5	D6	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15
Metatron	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓				✓	✓	✓	✓	✓	✓	✓		✓	
Doctag	✓	✓	✓		✓	✓	✓	✓	✓		✓	✓		✓			✓	✓		✓	✓	✓			
Doccano	✓	✓	✓		✓	✓	✓	✓				✓			✓	✓	✓	✓			✓	✓	✓	✓	✓
LabelStudio					✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓				✓	✓	✓		✓
TeamTat	✓	✓		✓	✓	✓	✓	✓	✓							✓	✓	✓	✓	✓		✓		✓	
INCEpTION	✓	✓		✓	✓	✓	✓	✓				✓			✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
brat	✓	✓			✓	✓	✓	✓								✓	✓			✓	✓	✓			✓
tagtog	✓	✓			✓	✓	✓	✓				✓		✓		✓	✓	✓		✓	✓	✓	✓	✓	
POTATO	✓	✓	✓		✓	✓	✓	✓				✓	✓				✓				✓	✓		✓	
Doctron	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓

IR compliance is challenging and not universally applicable to the provided annotation templates. From a data perspective, Doctron uniquely integrates IR-specific data sources (D6). While all tools allow schema customization, support various input and output formats, and enable uploading of pre-annotated data (D1-D4), several tools, including Doccano, LabelStudio, and INCEpTION, do not accommodate complex and structured input data without preprocessing. Only three tools support topic-document annotation (F1), making them suitable for IR workflows. While all evaluated tools support multi-label annotation (F2), graded labeling (F3) is only available in Doctron and POTATO. Most tools facilitate relationship annotations (F6), as well as NER or NER+L (F7), but only four out of ten support passage and image annotation (F4, F5). Despite offering a range of templates, LabelStudio can only identify objects using rectangles, without the option to create polygons. All tools are multilingual (F12), and only MetaTron, TeamTat, and Doctron provide full integration for Inter-Annotator Agreement (IAA) (F9). The free version of LabelStudio, along with brat and POTATO, lacks user and role management or collaborative annotation features (F8). Almost all tools support ontologies (F10) and ensure data privacy (F11). Over half of the tools enable the definition of annotation guidelines and built-in predictions (F13, F14), while only five support keyboard shortcuts (F15).

Note that no tool simultaneously meets the requirements for topic-document annotation, passage annotation, and graded labeling. Doctron is the only tool that fully satisfies all these criteria while also offering the possibility to download annotations in a TREC-like format. Although Doctron does not offer pre-built models for automatic annotation, it provides an interface that allows users to easily integrate their own custom annotation methods. This makes adding automatic annotations a straightforward plug-and-play process, where users can upload prediction models that output results in the required format, without needing additional development.

Quantitative evaluation. Quantitative analysis involves comparing the performance of five annotation tools based on the number of clicks required and the total time taken to complete two tasks: (i) annotating a predefined set of documents and (ii) creating and customizing a document collection. While the efficient implementation of annotation templates is crucial for creating large annotated corpora, the ease of creating and customizing a collection is key to ensuring accessibility and usability for users with no experience. The annotation templates considered in our analysis include multi-label annotation, passage annotation, NER. The tools selected for the quantitative evaluation are chosen from those included in the qualitative evaluation, specifically: Doctron, Doctag, INCEpTION, Doccano, and LabelStudio. TagTog was excluded due to the unavailability of its online version and the inability to run its offline version. MetaTron was not considered, as Doctron is built upon it. TeamTat and brat were excluded because they do not support document labeling and passage annotation. Similarly, POTATO lacks support for passage annotation and both NER and NER+L tasks.

Table 2: Performances of the selected tools in collection creation and customization. The performances are assessed in terms of number of clicks and time taken create a collection.

	Clicks	Time [s]
Doctag	8	3.3
LabelStudio	20	6.5
Doccano	13	4.2
INCEpTION	35	10.3
Doctron	10	3.8

To assess the performance of the selected tools, we used Selenium[12], an open-source testing framework designed to automate web browsers. We created five web agents, each corresponding to one of the annotation tools. To run these experiments, we considered a collection with a sample of 15 documents extracted from the `TREC Robust 2004` dataset. Given that Doccano and INCEpTION do not support the definition of the topics, we considered only one topic in the present evaluation. The experiments have been carried out running the offline installable instances of the tools. As a consequence, the performances strictly depend on the machine they have been run on; some delays have been introduced to support requests and responses and make the tools run properly.

[12]https://www.selenium.dev

Table 3: Performances of the selected tools in multilabel annotation (MA), passages annotation (PA), entity tagging (NER). The performances are assessed in terms of number of clicks, and average time and standard deviation to perform 50 annotation rounds (reported as the AVG ± STD.)

	MA		PA		NER	
	Clicks	Time [s]	Clicks	Time [s]	Clicks	Time [s]
Doctag	60	49.4 ± 1.5	155	375.1 ± 1.1	602	645.5 ± 1.5
LabelStudio	45	47.4 ± 1.1	124	350.4 ± 2.5	390	515.4 ± 2.2
Doccano	45	46.8 ± 0.8	-	-	390	522.3 ± 1.7
INCEpTION	45	46.7 ± 0.9	-	-	424	511.3 ± 2.7
Doctron	45	45.6 ± 1.3	105	324.5 ± 2.3	390	512.1 ± 1.2

Collection creation and customization. Table 2 presents the number of clicks and the time required to create a collection and customize the tool, setting up the annotation environment for the multilabel annotation, with two labels defined: *relevancy* and *clarity*. Among the tools analyzed, Doctag and Doctron are the most efficient in both time –3.3 and 3.8 seconds– and number of clicks needed for configuration –8 and 10 clicks. The small difference is due to the way in which labels are defined: while in Doctag, a file with the labels has to be provided, in Doctron, labels are added directly via the user interface. The higher efficiency of Doctron and Doctag compared to the other tools evaluated is due to the fact that all necessary user inputs are collected during the creation of the collection, and do not require any other configuration. LabelStudio showed similar results with 6.5 seconds and 20 clicks: while the upload of the documents is easy, the configuration of the labels requires more actions. Doccano falls in between with 4.2 seconds and 13 clicks: it is efficient and user-friendly, but it requires more clicks than Doctag because the collection is created first and documents are loaded afterward. In contrast, Doctag and Doctron load all documents at the time of collection creation. INCEpTION is the least efficient (10.3 seconds and 35 clicks) due to its complexity, which requires users to have a strong understanding of the documentation –such as the concept of an *annotation layer*.

Annotation experiments. Table 3 presents the performance results in terms of the number of clicks and the average time taken to perform multilabel annotation (MA), passages annotation (PA), NER on 15 documents. The reported values are the averages across 50 annotation rounds, with each round representing a full annotation of all 15 documents. In the multilabel annotation (MA) process, each document received two labels: one from the set {*relevant*, *not relevant*} and another from the set {*clear*, *not clear*}. All documents were annotated in three clicks (one for each label and one to proceed to the next document), except for Doctag, which required an additional click to save the annotations. On average, all tools completed an annotation round in 45 to 49 seconds. For passage annotation (PA), we identified a total of 25 passages in the 15 documents –i.e., less than 2 passages for each document. Since Doccano and INCEpTION do not support passage annotation, they were excluded from these experiments. Among the evaluated tools, Doctron proved to be the most efficient, requiring 105 clicks and 324.5 seconds. In contrast, Doctag performed the worst, taking 375.1 seconds with 155 clicks. Doctag required additional clicks to identify the intended passage, which impacted the recorded metrics. LabelStudio falls in between with 124 clicks and 350 seconds. In entity tagging (NER), a total of 98 mentions were tagged in 15 documents –i.e., less than 7 mentions per document. Doctron, Doccano, and LabelStudio exhibited the same performance in terms of the number of clicks, as they followed identical steps for tagging an entity, resulting in an equal click count. INCEpTION required only 34 more clicks than these three tools, while Doctag, consistent with previous cases, recorded a higher number of actions. Regarding time taken, Doctron, INCEpTION, Doccano, and LabelStudio required between 512 and 523 seconds to annotate the 15 documents, with Doctron being the most efficient. Doctag took the longest time due to its annotation process, which involves more interactions. Doctag was originally designed to support NER+L and had to be adapted for NER in these experiments, which further contributed to its lower performance compared to the other tools. We observe that the primary factor influencing performance is the process of creating a collection and its adaptability to the IR domain. The implementation of annotation templates across the tested tools is similar and optimized to improve user experience, which explains the similarities in the number of actions. Variations in annotation time are mainly due to differences in backend and database implementations. These differences are minimal and not noticeable to users, suggesting that all tools are designed for fast and efficient annotation. However, it is important to highlight that we adjusted the configuration and settings of the tools compared against Doctron to ensure they operated under optimal conditions across all three annotation tasks. In contrast, Doctron streamlines the customization of both collection and annotation templates, providing users with an efficient and intuitive out-of-the-box experience.

5 Conclusions

This paper introduced Doctron, an annotation tool for the IR domain, distributed as a Docker container for privacy and portability. It supports collaboration, allowing users to share document collections, and offers role-based access for administrators, annotators, and reviewers. Doctron includes various annotation templates like graded labeling, passage annotation, object annotation, NER, and NER+L. To ensure annotation reliability, it provides collection-based statistics and implements IAA metrics such as Cohen's Kappa, Fleiss' Kappa, and Krippendorff's Alpha.

We performed two analyses—a qualitative and a quantitative comparison—of Doctron and other annotation tools. In the qualitative assessment, we evaluated the tools based on technical aspects, available functionalities, and flexibility in input and output formats. Our findings showed that Doctron stands out as the most flexible and comprehensive tool, offering a complete set of features vital for IR. In the quantitative analysis, we concluded that Doctron offers a more intuitive, customizable, and efficient workflow, making it highly suitable for both experienced and novice users.

In future work, we aim to integrate automatic models for NER and NER+L, along with LLMs for graded labeling and passage annotation. These automatic predictions will act as annotation assistants, providing initial annotations for annotators to refine, which will speed up the process and reduce manual effort.

References

[1] K. Bontcheva, H. Cunningham, I. Roberts, A. Roberts, V. Tablan, N. Aswani, and G. Gorrell. 2013. GATE Teamware: a web-based, collaborative text annotation framework. *Lang. Resour. Evaluation* 47, 4 (2013), 1007–1029. https://doi.org/10.1007/S10579-013-9215-6

[2] J.M. Cejuela, P. McQuilton, L. Ponting, S. J. Marygold, R. Stefancsik, G. H. Millburn, and B. Rost. 2014. tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. *Database J. Biol. Databases Curation* 2014 (2014). https://doi.org/10.1093/DATABASE/BAU033

[3] J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[4] L. Colucci Cante, S. D'Angelo, B. Di Martino, and M. Graziano. 2024. Text Annotation Tools: A Comprehensive Review and Comparative Analysis. In *International Conference on Complex, Intelligent, and Software Intensive Systems*. Springer, 353–362.

[5] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, and E. M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. *CoRR* abs/2003.07820 (2020). arXiv:2003.07820 https://arxiv.org/abs/2003.07820

[6] B. Di Martino, F. Marulli, M. Graziano, and P. Lupi. 2021. PrettyTags: An Open-Source Tool for Easy and Customizable Textual MultiLevel Semantic Annotations. In *Complex, Intelligent and Software Intensive Systems - Proceedings of the 15th International Conference on Complex, Intelligent and Software Intensive Systems (CISIS-2021), Asan, Korea, 1-3 July 2021 (Lecture Notes in Networks and Systems, Vol. 278)*, Leonard Barolli, Kangbin Yim, and Tomoya Enokido (Eds.). Springer, 636–645. https://doi.org/10.1007/978-3-030-79725-6_64

[7] J. L. Fleiss, J. C. Nee, and J. R. Landis. 1979. Large sample variance of kappa in the case of different sets of raters. *Psychological bulletin* 86, 5 (1979), 974.

[8] F. Giachelle, O. Irrera, and Silvello G. 2021. MedTAG: a portable and customizable annotation tool for biomedical documents. *BMC Medical Informatics Decis. Mak.* 21, 1 (2021), 352. https://doi.org/10.1186/S12911-021-01706-4

[9] F. Giachelle, O. Irrera, and G. Silvello. 2022. DocTAG: A Customizable Annotation Tool for Ground Truth Creation. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 13186)*. Springer, 288–293. https://doi.org/10.1007/978-3-030-99739-7_35

[10] O. Irrera, M. Lissandrini, D. Dell'Aglio, and G. Silvello. 2024. Reproducibility and Analysis of Scientific Dataset Recommendation Methods. In *Proc. of the 18th ACM Conference on Recommender Systems (RecSys 2024)*. ACM Press.

[11] R. Islamaj, Kwon D., S. Kim, and Z Lu. 2021. TeamTat: A Collaborative Text Annotation Tool for Creating Gold-Standard Corpora. In *AMIA 2021, American Medical Informatics Association Annual Symposium, San Diego, CA, USA, October 30, 2021 - November 3, 2021*. AMIA.

[12] J. C. Klie, M. Bugert, B. Boullosa, R. E. de Castilho, and I. Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *COLING 2018, The 27th International Conference on Computational Linguistics: System Demonstrations, Santa Fe, New Mexico, August 20-26, 2018*, Dongyan Zhao (Ed.). Association for Computational Linguistics, 5–9. https://aclanthology.org/C18-2002/

[13] K. Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement* 30, 1 (1970), 61–70.

[14] K. Krippendorff. 2011. Computing Krippendorff's alpha-reliability.

[15] D. Kwon, S. Kim, S.Y. Shin, A. Chatr-aryamontri, and W. J. Wilbur. 2014. Assisting manual literature curation for protein-protein interactions using BioQRator. *Database J. Biol. Databases Curation* 2014 (2014). https://doi.org/10.1093/DATABASE/BAU067

[16] D. Kwon, S. Kim, C.H. Wei, R. Leaman, and Z. Lu. 2018. ezTag: tagging biomedical concepts via interactive learning. *Nucleic Acids Res.* 46, Webserver-Issue (2018), W523–W529. https://doi.org/10.1093/NAR/GKY428

[17] P. Lopez. 2009. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *Research and Advanced Technology for Digital Libraries*. Springer Berlin Heidelberg, Berlin, Heidelberg, 473–474.

[18] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, and N Goharian. 2021. Simplified Data Wrangling with ir_datasets. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. ACM, 2429–2436. https://doi.org/10.1145/3404835.3463254

[19] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[20] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, and X. Liang. 2018. Doccano: Text annotation tool for human. *Software available from https://github. com/doccano/doccano* 34 (2018).

[21] M. Neves and J. Seva. 2021. An extensive review of tools for manual annotation of documents. *Briefings Bioinform.* 22, 1 (2021), 146–163. https://doi.org/10.1093/BIB/BBZ130

[22] M. L. Neves and U. Leser. 2014. A survey on annotation tools for the biomedical literature. *Briefings Bioinform.* 15, 2 (2014), 327–340. https://doi.org/10.1093/BIB/BBS084

[23] J. Pei, A. Ananthasubramaniam, X. Wang, N. Zhou, A. Dedeloudis, J. Sargent, and D. Jurgens. 2022. POTATO: The Portable Text Annotation Tool. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022 - System Demonstrations, Abu Dhabi, UAE, December 7-11, 2022*, Wanxiang Che and Ekaterina Shutova (Eds.). Association for Computational Linguistics, 327–337. https://doi.org/10.18653/V1/2022.EMNLP-DEMOS.33

[24] T. Perry. 2021. LightTag: Text Annotation Platform. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021*, Heike Adel and Shuming Shi (Eds.). Association for Computational Linguistics, 20–27. https://doi.org/10.18653/V1/2021.EMNLP-DEMO.3

[25] J. Pustejovsky and A. Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

[26] D. Salgado, M. Krallinger, M. Depaule, E. Drula, A. V. Tendulkar, F. Leitner, A. Valencia, and C. Marcelle. 2012. MyMiner: a web application for computer-assisted biocuration and text annotation. *Bioinform.* 28, 17 (2012), 2285–2287. https://doi.org/10.1093/BIOINFORMATICS/BTS435

[27] H. Shindo, Y. Munesada, and Y. 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, Miyazaki, Japan, May 7-12, 2018*, Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (Eds.). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2018/summaries/680.html

[28] P. Stenetorp, S. Pyysalo, G. Topic, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012*, Walter Daelemans, Mirella Lapata, and Lluís Màrquez (Eds.). The Association for Computer Linguistics, 102–107. https://aclanthology.org/E12-2021/

[29] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov. 2020. Label studio: Data labeling software. *Open source software available from https://github. com/heartexlabs/label-studio* 2022 (2020).

[30] E. M Voorhees. 2005. The TREC robust retrieval track. *SIGIR Forum* 39, 1 (2005), 11–20. https://doi.org/10.1145/1067268.1067272

[31] D. Wilby, T. Armakharm, I. Roberts, X. Song, and K. Bontcheva. 2023. GATE Teamware 2: An open-source tool for collaborative document classification annotation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. EACL 2023 - System Demonstrations, Dubrovnik, Croatia, May 2-4, 2023*, Danilo Croce and Luca Soldaini (Eds.). Association for Computational Linguistics, 145–151. https://doi.org/10.18653/V1/2023.EACL-DEMO.17

[32] J. Yang, Y. Zhang, L. Li, and X. Li. 2018. YEDDA: A Lightweight Collaborative Text Span Annotation Tool. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, System Demonstrations*, Fei Liu and Thamar Solorio (Eds.). Association for Computational Linguistics, 31–36. https://doi.org/10.18653/V1/P18-4006

[33] S. M. Yimam, I. Gurevych, R. E. de Castilho, and C. Biemann. 2013. WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In *51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, Proceedings of the Conference System Demonstrations, 4-9 August 2013, Sofia, Bulgaria*. The Association for Computer Linguistics, 1–6. https://aclanthology.org/P13-4001/

Received 18 February 2025