Enhancing Ground Truth Creation for IR Through Web-Based Collaborative Annotation

Ornella Irrera*, Stefano Marchesin and Gianmaria Silvello

Department of Information Engineering, University of Padova, via Gradenigo 6/B, 35131, Padua, Italy

Abstract

Ground truth creation plays a key role in Information Retrieval (IR), serving as the foundation for building test collections used to train and evaluate retrieval systems. However, this process is often time-consuming and demands considerable effort from human experts. Evaluation campaigns like TREC and CLEF highlight the scale of this challenge, requiring extensive manual annotation to ensure the accuracy and reliability of the data.

To reduce this workload and support assessors more effectively, we present Doctron. Doctron is a collaborative, web-based, containerized platform designed to simplify ground truth creation in IR. It supports both text and image annotation, including features like entity tagging and linking, passages annotation, graded labeling, and object detection. The platform enables team collaboration through role-based permissions and integrates Inter Annotator Agreement (IAA) metrics to ensure annotation consistency and quality.

Keywords

Manual annotation, Ground truth creation, Evaluation

1. Introduction

Creating ground truth datasets for large corpora is fundamental to Information Retrieval (IR), as they are the foundation for training, evaluating, and enhancing search systems. Manual annotation – where human assessors label documents – remains the standard approach, playing a key role in ensuring the reliability and robustness of test collections that contribute to the progress in IR. In the context of large-scale evaluation initiatives like Text REtrieval Conference (TREC) and Conference and Labs of the Evaluation Forum (CLEF), the creation of high-quality annotated corpora is crucial for enabling researchers to benchmark models and foster progress in the field [5].

However, generating such datasets is a demanding and resource-intensive process involving multiple stakeholders – such as domain experts, annotators, and project coordinators – in a complex workflow. This process encompasses defining guidelines, selecting and configuring tools, preprocessing data, performing annotations, resolving inconsistencies, assessing quality, and making revisions. Due to its complexity, this workflow often becomes a bottleneck in IR projects [13, 17].

Choosing an appropriate annotation tool can streamline the process and ease assessors' work. In recent years, several reviews have assessed the effectiveness of annotation tools, compared their functionalities, and provided guidance to researchers in selecting the most appropriate tool for their specific needs [2, 13, 14]. Some annotation tools are specifically designed for certain domains, addressing the unique demands of particular research contexts. A significant number of these are tailored for the biomedical field [1, 4, 7, 8, 18], typically supporting tasks such as document classification, Named Entity Recognition (NER) and Named Entity Recognition and Linking (NER+L), and relation annotation, often with integration of domain-specific ontologies. Conversely, general-purpose tools [3, 5, 9, 12, 15, 16, 19–21] offer high customizability to accommodate a wide range of annotation tasks. Nevertheless, most annotation tools do not specifically address the needs of IR. They often lack essential features like relevance assessments for topic-document pairs or passage-level annotations, and are difficult to be properly configured. As a result, choosing and adapting a tool for IR is often complex, making it

IIR'25: The 15th Italian Information Retrieval Workshop. September 03–05, 2025, Cagliari, Italy

*Corresponding author.

ornella.irrera@unipd.it (O. Irrera); stefano.marchesin@unipd.it (S. Marchesin); gianmaria.silvello@unipd.it (G. Silvello)

0000-0003-2284-5699 (O. Irrera); 0009-0008-9269-3639 (S. Marchesin); 0000-0003-4970-4554 (G. Silvello)

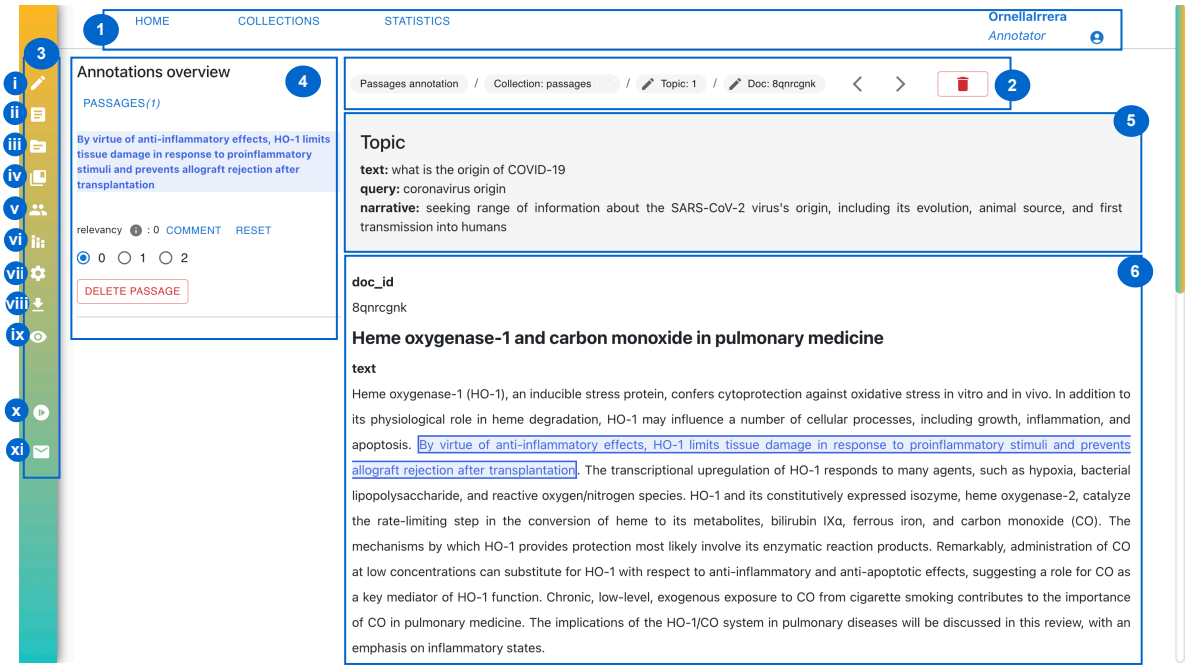

Figure 1: Doctron user interface with passages annotation template.

impractical for non-expert users.

To address the limitations of existing annotation tools, we introduce Doctron [6], a web-based, open-source, Dockerized platform designed for collaborative document annotation, with a strong focus on IR. It supports topic-based and graded relevance annotation, offering advanced features such as passage-level annotation, object detection, and support for both textual and image data. The platform enables efficient teamwork through role-based Inter Annotator Agreement (IAA) metrics to assess annotation consistency. Integration with the `ir_dataset` library [11] allows easy use and re-annotation of standard IR test collections. Doctron implements an intuitive, customizable interface accessible even to non-technical users, setting it apart from more complex tools. The platform is available as a cloud service at https://doctron.dei.unipd.it/ (currently accessible providing username: *demo* and password: *demo*), and can also be installed as a Docker container[1] for local deployment.

Doctron [6] was presented as a resource paper at SIGIR 2025.

2. Doctron

Doctron is a web-based, platform-independent annotation tool designed for document-topic pair annotation. Distributed as a Docker container, it enables easy deployment and ensures privacy by allowing local installation. Its architecture follows a three-tier design: a data layer with a PostgreSQL database storing documents, topics, and annotations; a business logic layer implemented with Django [2], handling core functionalities such as processing requests and interacting with the database; and a presentation layer developed with React.js [3], providing an interactive platform for annotators.

In Doctron, there are three user types: annotators, reviewers, and administrators. *Annotators* add annotations to documents. *Reviewers*, with the highest expertise, have full access to annotators' work and can update annotations to ensure quality and consistency. *Administrators* manage the collection, oversee annotators and reviewers, set annotation guidelines, and configure templates and settings. They can modify and update all annotations, including those of reviewers, and track progress.

[1] https://github.com/meta-doc-dev/DocTron
[2] https://www.djangoproject.com/
[3] https://react.dev/

The user interface of Doctron is presented in Figure 1. Upon login, users select an annotation template, and the system loads the relevant collections and the last opened topic-document pair. The main header ❶ provides access to the Home, Collections, and Statistics pages. The document header ❷ displays key identifiers and offers navigation and annotation reset options. A multifunctional left sidebar ❸ enables quick access to features like role switching, document/topic lists, settings, annotation downloads, and tutorials. The left panel ❹ summarizes the user's annotations for the current pair, allowing edits and comments. The main area presents the topic details ❺ and the document to annotate ❻, with the layout adapting to both text and image-based content.

Annoation templates. Annotators can annotate document-topic pairs in Doctron using seven annotation templates. These templates include (i) *graded labeling*, where annotators assign labels (e.g., relevance) with a range of values (e.g., 0 to 3) to a document (textual or image) regarding a specific topic; (ii) *passage annotation*, where annotators select specific passages within a textual document and assign a graded label indicating its relevance to a topic; (iii) *object annotation*, which involves identifying and labeling objects within images by selecting the object's perimeter and assigning one or more graded labels to the selected area; (iv) *NER*, where annotators identify mentions of entities in a text and label them with predefined tags; (v) *NER+L*, which adds the step of linking identified entities to specific entries in external knowledge bases like Wikipedia or Wikidata; (vi) *relationship annotation*, where annotators identify relationships between a subject, predicate, and object, which may be represented by ontological concepts, tags, or mentions; and (vii) *fact annotation*, which involves annotating factual triples (subject, predicate, object), where all components are ontological concepts or tags, and none are textual mentions from the document.

Collections and Customization. A collection in Doctron includes a group of annotators (at least one), a set of documents (either text or images), a set of topics (either text or images), and is associated with an annotation template. Each user in Doctron has the ability to annotate multiple document collections, which exist independently from one another. Documents in Doctron are schema-free and can be uploaded in several formats including: JSON, CSV, TXT and PDF (thanks to the integration with GROBID [10]) for textual documents, and JPG and PNG for image documents. Doctron supports integration with `ir_datasets` [11] and the PubMed REST API, allowing users to import documents and topics from URLs or external sources with full customization. Annotation templates are fully configurable, enabling the definition of custom tags, ontological concepts, and labels. Collections can go through multiple annotation rounds and be set in either *collaborative* or *competitive* mode: the former allows annotators to view other annotators' work, the latter instead hides the identities of other users. This ensures flexibility, as well as control over the annotation process.

Collection Statistics. Doctron provides two levels of statistical reporting: individual and global. Individual statistics give each annotator an overview of their work, including the number of annotated and unannotated documents per topic, as well as annotation-specific metrics depending on the template used (e.g., graded labels, identified passages, tagged entities, or extracted relations and facts). Administrators and reviewers can also monitor these statistics to track annotator activity and spot inconsistencies or documents that need to be reannotated. Global statistics aggregate this information across all annotators in a collection, also listing which annotators worked on each document. To evaluate annotation quality and consistency, Doctron includes IAA metrics – Cohen's Kappa, Fleiss's Kappa, and Krippendorff's Alpha – which are accessible by administrators and reviewers. These metrics provide insight into annotation reliability by quantifying the degree of agreement among annotators. Together, these features offer a comprehensive view of annotation performance and ensure quality control throughout the annotation process.

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	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓

3. Evaluation

We performed a qualitative evaluation where we compared Doctron functionalities to those provided by other 9 tools: MetaTron [7], Doctag [5], Doccano, LabelStudio, TeamTat [8], INCEpTION [9], brat [19], TagTog [1], POTATO [15]. We used three sets of criteria: technical, data-related, and functionality-based. These criteria cover aspects such as ease of use, data integration, annotation capabilities, and support for information retrieval workflows. As shown in Table 1, Doctron emerges as the most complete solution, fulfilling 24 out of 25 criteria. It is the only tool that supports the combination of topic-document annotation, passage-level annotation, and graded labeling, while also offering TREC-like export and IR-specific dataset integration. Although it lacks built-in automatic predictions, Doctron supports easy integration of user-defined models, ensuring flexibility and customizability.

We performed a quantitative analysis to assess the efficiency of five annotation tools – Doctron, Doctag, INCEpTION, Doccano, and LabelStudio – by measuring the number of clicks and time required for two tasks: (i) creating and configuring a document collection, and (ii) annotating 15 documents using three typical IR-related templates: multilabel classification, passage annotation, and NER. While the annotation phase yielded comparable results across tools – due to similar implementations of task templates – significant differences emerged in the setup phase. Doctron and Doctag proved more efficient for creating collections, with fewer steps and a simpler setup—Doctron especially stood out for its ease of use. In contrast, general-purpose tools like INCEpTION and LabelStudio required more complex configurations to support IR-specific workflows. These findings show that general tools often demand extra effort for IR tasks, while Doctron offers a more integrated, ready-to-use solution. More details can be found in the original Doctron paper [6].

4. Final remarks

We presented Doctron, a portable, web-based annotation tool designed for the IR. It supports collaborative workflows with role-based access, offers a wide range of annotation templates (including graded labeling, passage annotation, NER and NER+L), and integrates with PubMed and `ir_dataset` ensuring easy and customizable collection setup. Doctron includes built-in IAA metrics and detailed statistics to ensure annotation quality. Through qualitative and quantitative evaluations, we demonstrated that Doctron stands out for its flexibility, rich feature set, and ease of use. As future work, we plan to integrate automatic annotation support using pre-trained models for NER/NER+L and Language Language Models (LLMs) for graded and passage-level annotation.

Acknowledgments

This work is supported by the HEREDITARY Project, as part of the European Union's Horizon Europe research and innovation programme under Grant Agreement No GA 101137074.

Declaration on Generative AI

During the preparation of this work, the author did not use any AI tool.

References

[1] J.M. Cejuela, P. McQuilton, L. Ponting, S. J. Marygold, R. Stefancsik, G. H. Millburn, and B. Rost. tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. *Database J. Biol. Databases Curation*, 2014, 2014.

[2] L. Colucci Cante, S. D'Angelo, B. Di Martino, and M. Graziano. Text annotation tools: A comprehensive review and comparative analysis. In *International Conference on Complex, Intelligent, and Software Intensive Systems*, pages 353–362. Springer, 2024.

[3] B. Di Martino, F. Marulli, M. Graziano, and P. Lupi. 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*, volume 278 of *Lecture Notes in Networks and Systems*, pages 636–645. Springer, 2021.

[4] F. Giachelle, O. Irrera, and Silvello G. Medtag: a portable and customizable annotation tool for biomedical documents. *BMC Medical Informatics Decis. Mak.*, 21(1):352, 2021.

[5] F. Giachelle, O. Irrera, and G. Silvello. 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*, volume 13186 of *Lecture Notes in Computer Science*, pages 288–293. Springer, 2022.

[6] O. Irrera, S. Marchesin, F. SHami, and G. Silvello. 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 (SIGIR '25), July 13–18, 2025, Padua, Italy*, 2025.

[7] O. Irrera, S. Marchesin, and 1a G. Silvello. Metatron: advancing biomedical annotation empowering relation annotation and collaboration. volume 25, page 112, 2024.

[8] R. Islamaj Dogan, D. Kwon, S. Kim, and Z. Lu. Teamtat: a collaborative text annotation tool. volume 48, pages W5–W11, 2020.

[9] J. C. Klie, M. Bugert, B. Boullosa, R. E. de Castilho, and I. Gurevych. 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*, pages 5–9. Association for Computational Linguistics, 2018.

[10] P. Lopez. GROBID: combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries, 13th European Conference, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009. Proceedings*, volume 5714 of *Lecture Notes in Computer Science*, pages 473–474. Springer, 2009.

[11] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, and N Goharian. 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*, pages 2429–2436. ACM, 2021.

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

[13] M. Neves and J. Seva. An extensive review of tools for manual annotation of documents. *Briefings in Bioinformatics*, 22(1):146–163, 2021.

[14] M. L. Neves and U. Leser. A survey on annotation tools for the biomedical literature. *Briefings in Bioinformatics*, 15(2):327–340, 2014.

[15] J. Pei, A. Ananthasubramaniam, X. Wang, N. Zhou, A. Dedeloudis, J. Sargent, and D. Jurgens. POTATO: the portable text annotation tool. In Wanxiang Che and Ekaterina Shutova, editors,

Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022 - System Demonstrations, Abu Dhabi, UAE, December 7-11, 2022, pages 327–337. Association for Computational Linguistics, 2022.

[16] T. Perry. Lighttag: Text annotation platform. In Heike Adel and Shuming Shi, editors, *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*, pages 20–27. Association for Computational Linguistics, 2021.

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

[18] D. Salgado, M. Krallinger, M. Depaule, E. Drula, A. V. Tendulkar, F. Leitner, A. Valencia, and C. Marcelle. Myminer: a web application for computer-assisted biocuration and text annotation. *Bioinformatics*, 28(17):2285–2287, 2012.

[19] P. Stenetorp, S. Pyysalo, G. Topic, T. Ohta, S. Ananiadou, and J. Tsujii. 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*, pages 102–107. The Association for Computer Linguistics, 2012.

[20] D. Wilby, T. Armakharm, I. Roberts, X. Song, and K. Bontcheva. 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*, pages 145–151. Association for Computational Linguistics, 2023.

[21] J. Yang, Y. Zhang, L. Li, and X. Li. YEDDA: A lightweight collaborative text span annotation tool. In Fei Liu and Thamar Solorio, editors, *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, System Demonstrations*, pages 31–36. Association for Computational Linguistics, 2018.