PDF Entity Annotation Tool (PEAT)

Christopher G. Stahl [1¶], **Kristan J. Markey** [2], **Brian C. Jewell** [1], **Dahnish Shams** [2], **Michele M. Taylor** [2], **A. Amina Wilkins** [2], **Sean Watford** [2], **Andy Shapiro** [2], **and Michelle Angrish** [2]

1 Oak Ridge National Laboratory, USA **2** Office of Research and Development. United States Environmental Protection Agency ¶ Corresponding author

DOI: 10.21105/joss.05336

Software

- Review
- Repository
- Archive

Editor: Ana Trisovic

Reviewers:

- @flor14
- @RuneBlaze

Submitted: 24 January 2023
Published: 08 April 2025

Disclaimer: The views expressed in this manuscript are those of the authors and do not necessarily reflect the views or policies of the U.S. EPA, UT-Battelle, LLC, or U.S. Department of Energy.

Summary

While different text mining approaches – including the use of Artificial Intelligence (AI) and other machine based methods - continue to expand at a rapid pace, the tools used by researchers to create the labeled datasets required for training, modeling, and evaluation remain rudimentary. Labeled datasets contain the target attributes the machine is going to learn; for example, training an algorithm to delineate between images of a car or truck would generally require a set of images with a quantitative description of the underlying features of each vehicle type. Development of labeled textual data that can be used to build natural language machine learning models for scientific literature is not currently integrated into existing manual workflows used by domain experts. Published literature is rich with important information, such as different types of embedded text, plots, and tables that can all be used as inputs to train ML/natural language processing (NLP) models, when extracted and prepared in machine readable formats. Currently, both normalized data extraction of use to domain experts and extraction to support development of ML/NLP models are labor intensive and cumbersome manual processes. Automatic extraction of data and information from formats such as PDFs that are optimized for layout and human readability, not machine readability. The PDF (Portable Document Format) Entity Annotation Tool (PEAT) was developed with the goal of allowing users to annotate publications within their current print format, while also allowing those annotations to be captured in a machine-readable format. One of the main issues with traditional annotation tools is that they require transforming the PDF into plain text to facilitate the annotation process. While doing so lessens the technical challenges of annotating data, the user loses all structure and provenance that was inherent in the underlying PDF. Also, textual data extraction from PDFs can be an error prone process. Challenges include identifying sequential blocks of text and a multitude of document formats (multiple columns, font encodings, etc.). As a result of these challenges, using existing tools for development of NLP/ML models directly from PDFs is difficult because the generated outputs are not interoperable.

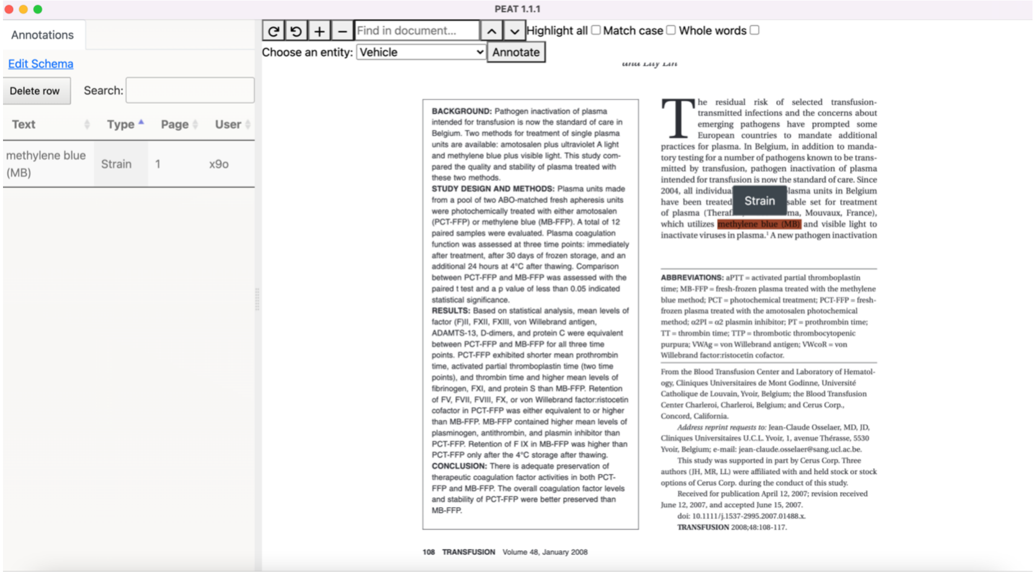
We created a system that allows annotations to be completed on the original PDF document structure, with no plain text extraction. The result is an application that allows for easier and more accurate annotations. In addition, by including a feature that grants the user the ability to easily create a schema, we have developed a system that can be used to annotate text for different domain-centric schemas of relevance to subject matter experts. Different knowledge domains require distinct schemas and annotation tags to support machine learning.

Statement of need

Data users are confronted with new data being published faster than manual data extractions are practical. Millions of publications are produced every year (White, 2021), potentially containing useful information for ongoing and future research, experiments, etc. With the sheer amount of data needing to be processed, it is unsustainable to manually keep up with the information being produced. The development and application of ML tools to help process and manage useful information from these publications is critically important to ensure that the most recent evidence is identified, evaluated, and extracted. Development of ML models requires manually labelled datasets, but currently, generation of those datasets is a high level of effort endeavor and not aligned with existing workflows used to manually summarize content presented in full-text (e.g., literature review workflows). Scientific literature is almost exclusively made available in print publications that are locked behind non-machine-readable formats, primarily PDF. The PDF format was designed to create digital versions of paper documents while maintaining their visual look and structure. While the format handles that task very well, it does not lend to the extraction of data to other formats. The opening of the PDF standard in 2008 (International Organization for Standardization, 2020) has helped facilitate PDF text extraction, but document formatting and content (e.g., coordinates of text within a document) can still be lost during PDF to text conversion using currently available tools (Xu et al., 2013).

Existing annotation solutions such as BRAT (Stenetorp et al., 2012), TeamTat (Islamaj et al., 2020), GATE (*GATE*, 2022), and Dextr (Walker et al., 2022) rely on first converting the PDF to a plain text layer before annotation by the user. Not only does this require the researcher to annotate unstructured plain text as opposed to a fully formatted PDF document, but also in many instances the PDF's format (double/triple columns, tables, etc) would fail to extract and render a document unusable. This is due to the nature of the PDF document creation process. Text can be split into several chunks (splits within words, lines, columns, etc.) and extraction tools use heuristics to attempt to put the chunks back together like a puzzle. A separate tool, PDFAnno (Shindo et al., 2018), improved on this by allowing annotation layers to be created on top of PDF documents but is no longer maintained or functional. Reac-PDF Annot (Tyurin, 2022) is a library that allows for basic annotations to be created in on top of PDFJS (Mozilla, 2025) based applications and was used as a baseline for our application.

PEAT was designed to take advantage of the latest advancements in PDF text extraction methods while also allowing the user to annotate and label the data directly in PDF format. This approach allows a user to work in a document structure they are familiar with, improving the user experience and facilitating the creation of labeled data for machine consumption and training of future machine learning models. PEAT is a portable, standalone application built off the Electron software framework (Electron Authors, 2025) and can be used on all major operating systems (Windows, Linux, and Macintosh) and provides an interface for users to annotate PDFs that are displayed in their intended format (Figure 1).



Text	Type	Page	User
methylene blue (MB)	Strain	1	x9o

BACKGROUND: Pathogen inactivation of plasma intended for transfusion is now the standard of care in Belgium. Two methods for treatment of single plasma units are available: amotosalen plus ultraviolet A light and methylene blue plus visible light. This study compared the quality and stability of plasma treated with these two methods.

STUDY DESIGN AND METHODS: Plasma units made from a pool of two ABO-matched fresh apheresis units were photochemically treated with either amotosalen (PCT-FFP) or methylene blue (MB-FFP). A total of 12 paired samples were evaluated. Plasma coagulation function was assessed at three time points: immediately after treatment, after 30 days of frozen storage, and an additional 24 hours at 4°C after thawing. Comparison between PCT-FFP and MB-FFP was assessed with the paired t test and a p value of less than 0.05 indicated statistical significance.

RESULTS: Based on statistical analysis, mean levels of factor (F)II, FXII, FXIII, von Willebrand antigen, ADAMTS-13, D-dimers, and protein C were equivalent between PCT-FFP and MB-FFP for all three time points. PCT-FFP exhibited shorter mean prothrombin time, activated partial thromboplastin time (two time points), and thrombin time and higher mean levels of fibrinogen, FXI, and protein S than MB-FFP. Retention of FV, FVII, FVIII, FX, or von Willebrand factor:ristocetin cofactor in PCT-FFP was either equivalent to or higher than MB-FFP. MB-FFP contained higher mean levels of plasminogen, antithrombin, and plasmin inhibitor than PCT-FFP. Retention of F IX in MB-FFP was higher than PCT-FFP only after the 4°C storage after thawing.

CONCLUSION: There is adequate preservation of therapeutic coagulation factor activities in both PCT-FFP and MB-FFP. The overall coagulation factor levels and stability of PCT-FFP were better preserved than MB-FFP.

The residual risk of selected transfusion-transmitted infections and the concerns about emerging pathogens have prompted some European countries to mandate additional practices for plasma. In Belgium, in addition to mandatory testing for a number of pathogens known to be transmitted by transfusion, pathogen inactivation of plasma intended for transfusion is now the standard of care. Since 2004, all individua lasma units in Belgium have been treated sable set for treatment of plasma (Theraf ma, Mouvaux, France), which utilizes methylene blue (MB) and visible light to inactivate viruses in plasma.[1] A new pathogen inactivation

ABBREVIATIONS: aPTT = activated partial thromboplastin time; MB-FFP = fresh-frozen plasma treated with the methylene blue method; PCT = photochemical treatment; PCT-FFP = fresh-frozen plasma treated with the amotosalen photochemical method; α2PI = α2 plasmin inhibitor; PT = prothrombin time; TT = thrombin time; TTP = thrombotic thrombocytopenic purpura; VWAg = von Willebrand antigen; VWcoR = von Willebrand factor:ristocetin cofactor.

From the Blood Transfusion Center and Laboratory of Hematology, Cliniques Universitaires de Mont Godinne, Université Catholique de Louvain, Yvoir, Belgium; the Blood Transfusion Center Charleroi, Charleroi, Belgium; and Cerus Corp., Concord, California.

Address reprint requests to: Jean-Claude Osselaer, MD, JD, Cliniques Universitaires U.C.L. Yvoir, 1, avenue Thérasse, 5530 Yvoir, Belgium; e-mail: jean-claude.osselaer@sang.ucl.ac.be.

This study was supported in part by Cerus Corp. Three authors (JH, MR, LL) were affiliated with and held stock or stock options of Cerus Corp. during the conduct of this study.

Received for publication April 12, 2007; revision received June 12, 2007, and accepted June 15, 2007.

doi: 10.1111/j.1537-2995.2007.01488.x.

TRANSFUSION 2008;48:108-117.

108 **TRANSFUSION** Volume 48, January 2008

Figure 1: PEAT User Interface.

The portability is accomplished by embedding all necessary dependencies into the framework, which does result in a larger than usual footprint, PEAT is 500 MB. The application allows users to load PDFs directly from their file system along with data annotation forms with standard or customizable annotation types, labels, entities, and other features such as custom color highlighting. The application also includes features for users to edit and import/export data extraction schemas, export annotations of X and Y PDF coordinate structure (based on the image layer of the PDF), search and manipulate annotations, and save/load progress. Once a user has completed document annotation, the labeled data, full text, and all associated metadata is exportable in JSON format that can be processed by a variety of NLP model building applications such as Spacy or PyTorch (Ansel et al., 2024; Honnibal et al., 2020).

Conclusion and Future Work

In this work we demonstrated PEAT's ability to assist users in the creation of annotated datasets that provide a machine-readable output suitable for machine learning applications. It allows users to annotate PDF documents in their native form using standard or custom annotations suitable for direct use with named entity recognition tools. Additionally, the schema editor tool grants users the flexibility to customize their annotations for domain specific applications. PEAT provides the foundation for many additional features such as collaborative annotation, a hierarchical annotation system (i.e., groups of annotations forming "relationships"), auto annotation based on imported ontologies, and more. Finally, creating a pluggable architecture for generating and aligning text layers or segments which would be used for annotations and NLP processing would be desirable. These feature enhancements would further increase the user's ability to create more advanced annotation sets, laying the groundwork for the continued growth and evolution of machine learning applications.

Acknowledgements

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy and sponsored by U.S. Environmental Protection Agency (U.S. EPA) under agreement DW08992524301.The United States Government retains and the publisher, by accepting the article for publication, acknowledges that

the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan. The views expressed are those of the authors and do not necessarily represent the views or policies of the U.S. EPA. Any mention of trade names, products, or services does not imply an endorsement by the U.S. government or the U.S. EPA. The U.S. EPA does not endorse any commercial products, services, or enterprises.

References

Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., … Chintala, S. (2024). PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. https://doi.org/10.1145/3620665.3640366

Electron Authors. (2025). *Electron*. https://github.com/electron/electron

GATE: A full-lifecycle open source solution for text processing. (2022). The University of Sheffield. https://gate.ac.uk/overview.html

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. https://doi.org/10.5281/zenodo.1212303

International Organization for Standardization. (2020). *ISO 32000-2:2020 document management – portable document format – part 2: PDF 2.0*. https://www.iso.org/standard/75839.html

Islamaj, R., Kwon, D., Kim, S., & Lu, Z. (2020). TeamTat: a collaborative text annotation tool. *Nucleic Acids Research*, *48*. https://doi.org/10.1093/nar/gkaa333

Mozilla. (2025). *PDF.js*. https://github.com/mozilla/pdf.js

Shindo, H., Munesada, Y., & Matsumoto, Y. (2018). PDFAnno: a Web-based Linguistic Annotation Tool for PDF Documents. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. http://www.lrec-conf.org/proceedings/lrec2018/pdf/680.pdf

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). brat: A web-based tool for NLP-assisted text annotation. *Proceedings of the Demonstrations Session at EACL 2012*.

Tyurin, A. (2022). React-pdf-highlighter. In *GitHub repository*. Github. https://github.com/agentcooper/react-pdf-highlighter

Walker, V., Schmitt, C., Wolfe, M., Nowak, A., Kulesza, K., Williams, A., Shin, R., Cohen, J., Burch, D., Stout, M., Shipkowski, K., & Rooney, A. (2022). Evaluation of a semi-automated data extraction tool for public health literature-based reviews: Dextr. *Environ Int*. https://doi.org/10.1016/j.envint.2021.107025

White, K. (2021). Publications Output: U.S. Trends and International Comparisons. *National Center for Science and Engineering Statistics*. https://ncses.nsf.gov/pubs/nsb20214

Xu, C., Tang, Z., Tao, X., & Shi, C. (2013). Graph-based layout analysis for PDF documents. *Imaging and Printing in a Web 2.0 World IV*. https://doi.org/10.1117/12.2005608