

PDF Entity Annotation Tool (PEAT)

January 2023
Version 1.0

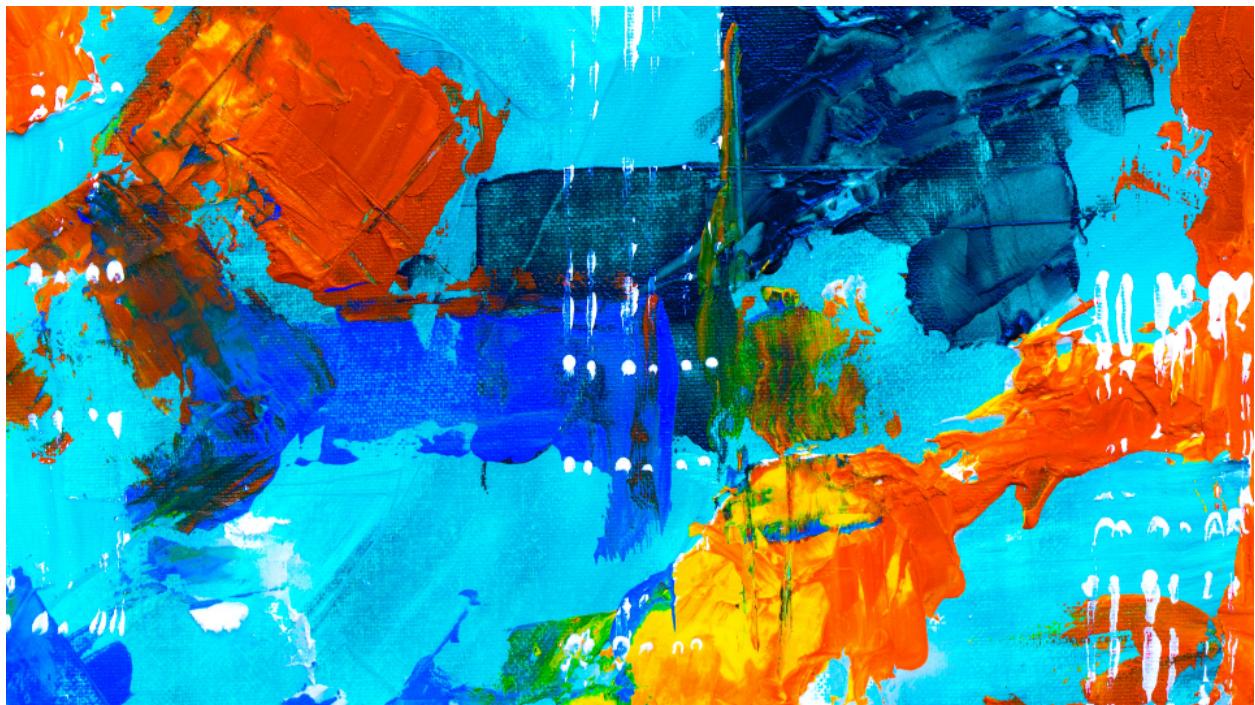


Table of Contents

Table of Contents	1
Document Revision.....	2
1 Application	3
1.0 Load PDF	3
1.1 Annotate PDF.....	5
1.2 Save Annotations	7
1.3 Load Annotations	9
1.4 Delete Annotations.....	11
1.5 Edit Schema.....	12
1.6 Auto Annotation.....	13

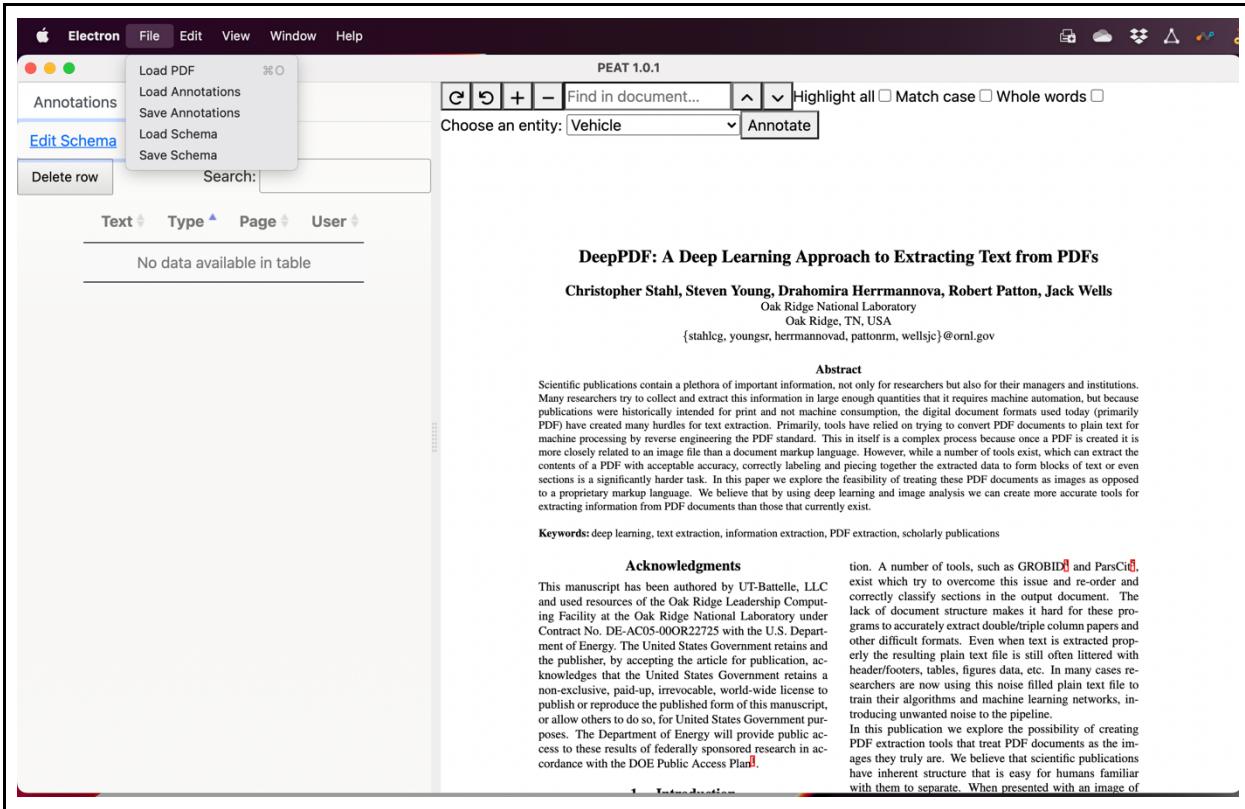
Document Revision

Date	Version Number	Document Changes
01/05/2021	0.1.0	Initial draft
07/20/2021	0.1.1	Updated to beta PEAT 0.1.1
09/29/2021	0.1.2	Update to beta PEAT 0.1.2
05/10/2022	0.1.3	Update to beta PEAT 0.1.3
01/24/2023	1.0.1	Updated to release PEAT 1.0.1

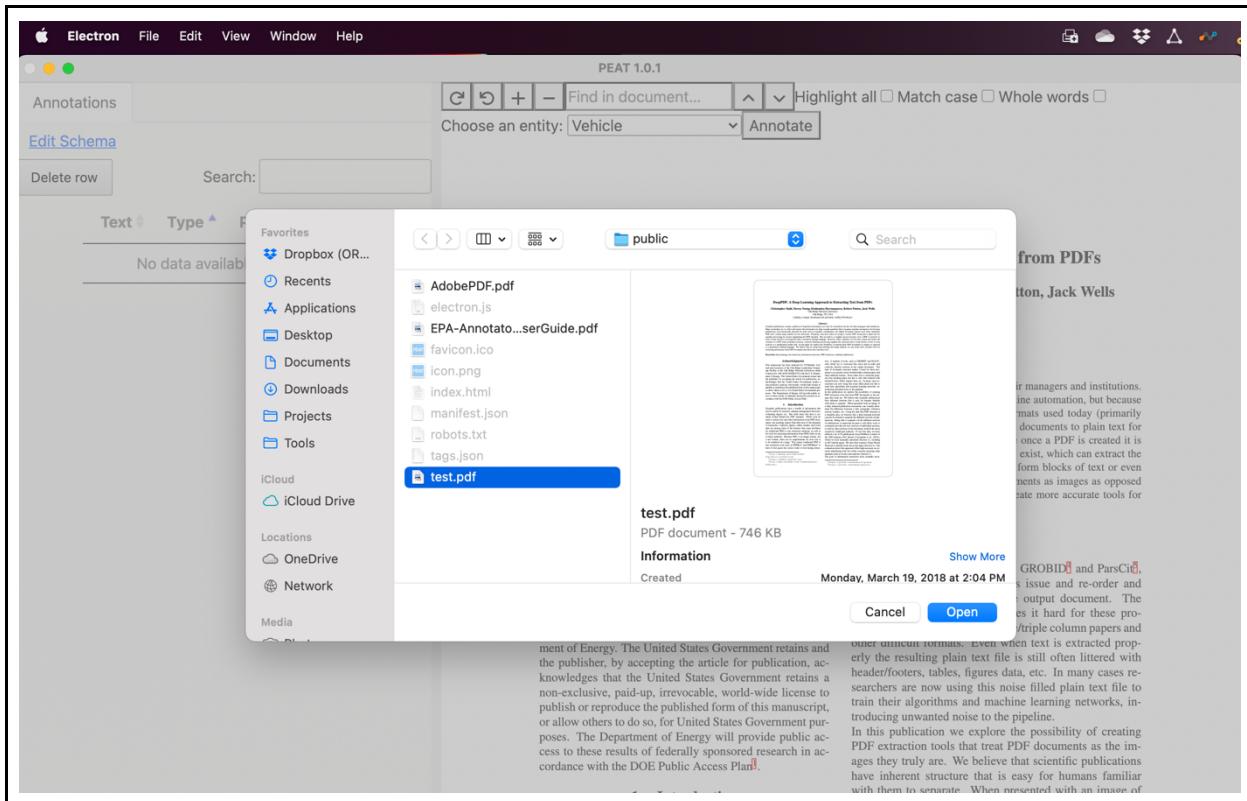
1 Application

1.0 Load PDF

1. Click *File* in the menu bar and select *Load PDF*.



2. Select the PDF file from your computer and click *Open*.



PEAT 1.0.1

Annotations
Edit Schema
Delete row
Search:

Text Type Page User
No data available in table

Find in document... Highlight all Match case Whole words
Choose an entity: Vehicle Annotate

DeepPDF: A Deep Learning Approach to Extracting Text from PDFs

Christopher Stahl, Steven Young, Drahomira Herrmannova, Robert Patton, Jack Wells
Oak Ridge National Laboratory
Oak Ridge, TN, USA
{stahlcg, youngsr, herrmannovad, pattonrm, wellsjc}@ornl.gov

Abstract
Scientific publications contain a plethora of important information, not only for researchers but also for their managers and institutions. Many researchers try to collect and extract this information in large enough quantities that it requires machine automation, but because publications were historically intended for print and not machine consumption, the digital document formats used today (primarily PDF) have created many hurdles for text extraction. Primarily, tools have relied on trying to convert PDF documents to plain text for machine processing by reverse engineering the PDF standard. This in itself is a complex process because once a PDF is created it is more closely related to an image file than a document markup language. However, while a number of tools exist, which can extract the contents of a PDF with acceptable accuracy, correctly labeling and piecing together the extracted data to form blocks of text or even sections is a significantly harder task. In this paper we explore the feasibility of treating these PDF documents as images as opposed to a proprietary markup language. We believe that by using deep learning and image analysis we can create more accurate tools for extracting information from PDF documents than those that currently exist.

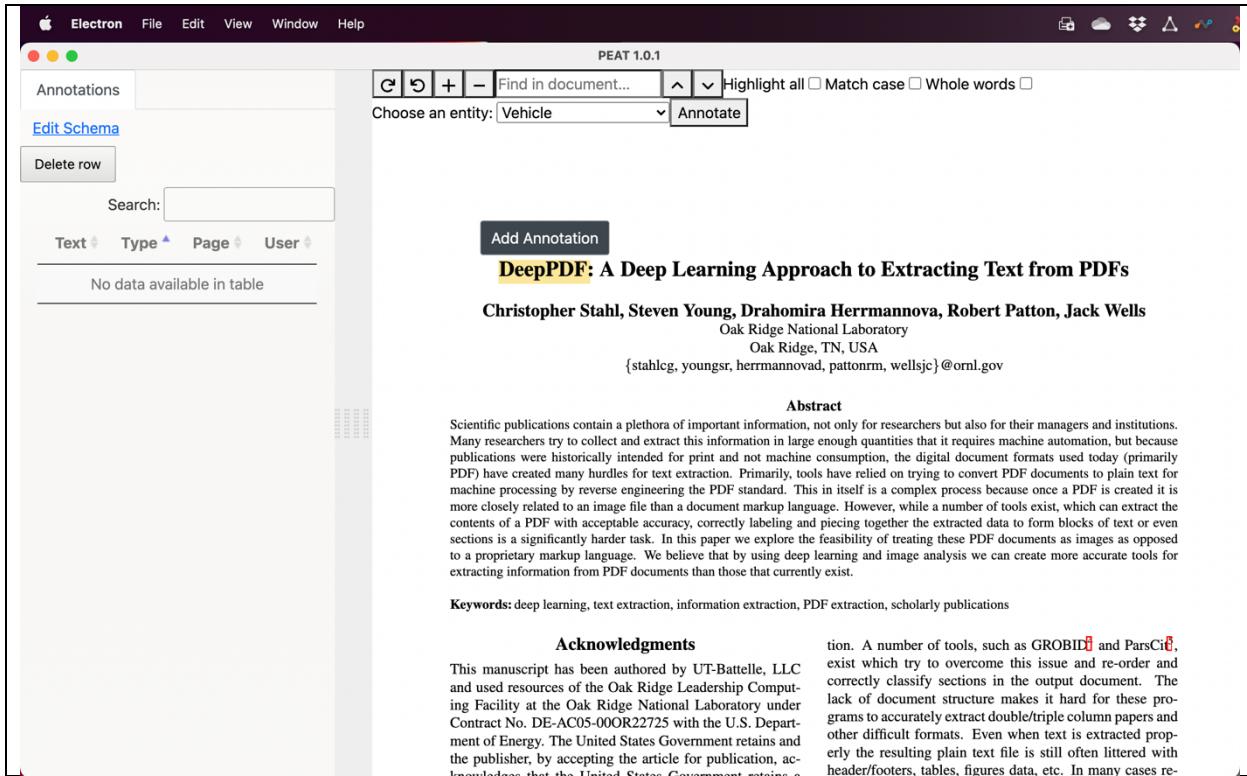
Keywords: deep learning, text extraction, information extraction, PDF extraction, scholarly publications

Acknowledgments
This manuscript has been authored by UT-Battelle, LLC and used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a

tion. A number of tools, such as GROBID and ParsCit, exist which try to overcome this issue and re-order and correctly classify sections in the output document. The lack of document structure makes it hard for these programs to accurately extract double/triple column papers and other difficult formats. Even when text is extracted properly the resulting plain text file is still often littered with header/footer, tables, figures data, etc. In many cases re-

1.1 Annotate PDF

1. Highlight text you wish to annotate and select *Add Annotation*.



2. Select the annotation type.

PEAT 1.0.1

Annotations Find in document... Highlight all Match case Whole words

Choose an entity: Annotate

✓ Select Type:
 Vehicle
 TestArticleVerification
GroupName
 GroupSize
 SampleSize
 Species
 Strain
 Sex
 CellLine
 Dose
 DoseUnits
 DoseFrequency
 DoseDuration
 DoseDurationUnits
 DoseRoute
 TimeAtDose

Deep Learning Approach to Extracting Text from PDFs

Authors: Steven Young, Drahomira Herrmannova, Robert Patton, Jack Wells
 Oak Ridge National Laboratory
 Oak Ridge, TN, USA
 {stahlcg, youngsr, herrmannovad, pattonrm, wellsjc}@ornl.gov

Abstract
 Scientific publications contain a plethora of important information, not only for researchers but also for their managers and institutions. Many researchers try to collect and extract this information in large enough quantities that it requires machine automation, but because publications were historically intended for print and not machine consumption, the digital document formats used today (primarily PDF) have created many hurdles for text extraction. Primarily, tools have relied on trying to convert PDF documents to plain text for machine processing by reverse engineering the PDF standard. This in itself is a complex process because once a PDF is created it is more closely related to an image file than a document markup language. However, while a number of tools exist, which can extract the contents of a PDF with acceptable accuracy, correctly labeling and piecing together the extracted data to form blocks of text or even sections is a significantly harder task. In this paper we explore the feasibility of treating these PDF documents as images as opposed to a proprietary markup language. We believe that by using deep learning and image analysis we can create more accurate tools for extracting information from PDF documents than those that currently exist.

Keywords: deep learning, text extraction, information extraction, PDF extraction, scholarly publications

Acknowledgments
 This manuscript has been authored by UT-Battelle, LLC and used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a

This manuscript has been authored by UT-Battelle, LLC and used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a

tion. A number of tools, such as GROBID and ParsCit, exist which try to overcome this issue and re-order and correctly classify sections in the output document. The lack of document structure makes it hard for these programs to accurately extract double/triple column papers and other difficult formats. Even when text is extracted properly the resulting plain text file is still often littered with header/footer tables, figures data, etc. In many cases re-

3. Hit save

PEAT 1.0.1

Annotations Find in document... Highlight all Match case Whole words

Choose an entity: Annotate

GroupName

DeepPDF: A Deep Learning Approach to Extracting Text from PDFs

Authors: Christopher Stahl, Steven Young, Drahomira Herrmannova, Robert Patton, Jack Wells
 Oak Ridge National Laboratory
 Oak Ridge, TN, USA
 {stahlcg, youngsr, herrmannovad, pattonrm, wellsjc}@ornl.gov

Abstract
 Scientific publications contain a plethora of important information, not only for researchers but also for their managers and institutions. Many researchers try to collect and extract this information in large enough quantities that it requires machine automation, but because publications were historically intended for print and not machine consumption, the digital document formats used today (primarily PDF) have created many hurdles for text extraction. Primarily, tools have relied on trying to convert PDF documents to plain text for machine processing by reverse engineering the PDF standard. This in itself is a complex process because once a PDF is created it is more closely related to an image file than a document markup language. However, while a number of tools exist, which can extract the contents of a PDF with acceptable accuracy, correctly labeling and piecing together the extracted data to form blocks of text or even sections is a significantly harder task. In this paper we explore the feasibility of treating these PDF documents as images as opposed to a proprietary markup language. We believe that by using deep learning and image analysis we can create more accurate tools for extracting information from PDF documents than those that currently exist.

Keywords: deep learning, text extraction, information extraction, PDF extraction, scholarly publications

Acknowledgments
 This manuscript has been authored by UT-Battelle, LLC and used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a

This manuscript has been authored by UT-Battelle, LLC and used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a

tion. A number of tools, such as GROBID and ParsCit, exist which try to overcome this issue and re-order and correctly classify sections in the output document. The lack of document structure makes it hard for these programs to accurately extract double/triple column papers and other difficult formats. Even when text is extracted properly the resulting plain text file is still often littered with header/footer tables, figures data, etc. In many cases re-

The screenshot shows the PEAT 1.0.1 application window. At the top, there's a menu bar with Apple, Electron, File, Edit, View, Window, and Help. Below the menu is a toolbar with icons for search, find, highlight, and other functions. A search bar says "Find in document..." with options for "Highlight all", "Match case", and "Whole words". A dropdown menu says "Choose an entity: Vehicle" and has an "Annotate" button.

The main area displays a PDF document titled "DeepPDF: A Deep Learning Approach to Extracting Text from PDFs" by Christopher Stahl, Steven Young, Drahomira Herrmannova, Robert Patton, and Jack Wells. The authors are from Oak Ridge National Laboratory, Oak Ridge, TN, USA, and their email is {stahlcg, youngsr, herrmannovad, pattonrm, wellsjc}@ornl.gov.

Abstract

Scientific publications contain a plethora of important information, not only for researchers but also for their managers and institutions. Many researchers try to collect and extract this information in large enough quantities that it requires machine automation, but because publications were historically intended for print and not machine consumption, the digital document formats used today (primarily PDF) have created many hurdles for text extraction. Primarily, tools have relied on trying to convert PDF documents to plain text for machine processing by reverse engineering the PDF standard. This in itself is a complex process because once a PDF is created it is more closely related to an image file than a document markup language. However, while a number of tools exist, which can extract the contents of a PDF with acceptable accuracy, correctly labeling and piecing together the extracted data to form blocks of text or even sections is a significantly harder task. In this paper we explore the feasibility of treating these PDF documents as images as opposed to a proprietary markup language. We believe that by using deep learning and image analysis we can create more accurate tools for extracting information from PDF documents than those that currently exist.

Keywords: deep learning, text extraction, information extraction, PDF extraction, scholarly publications

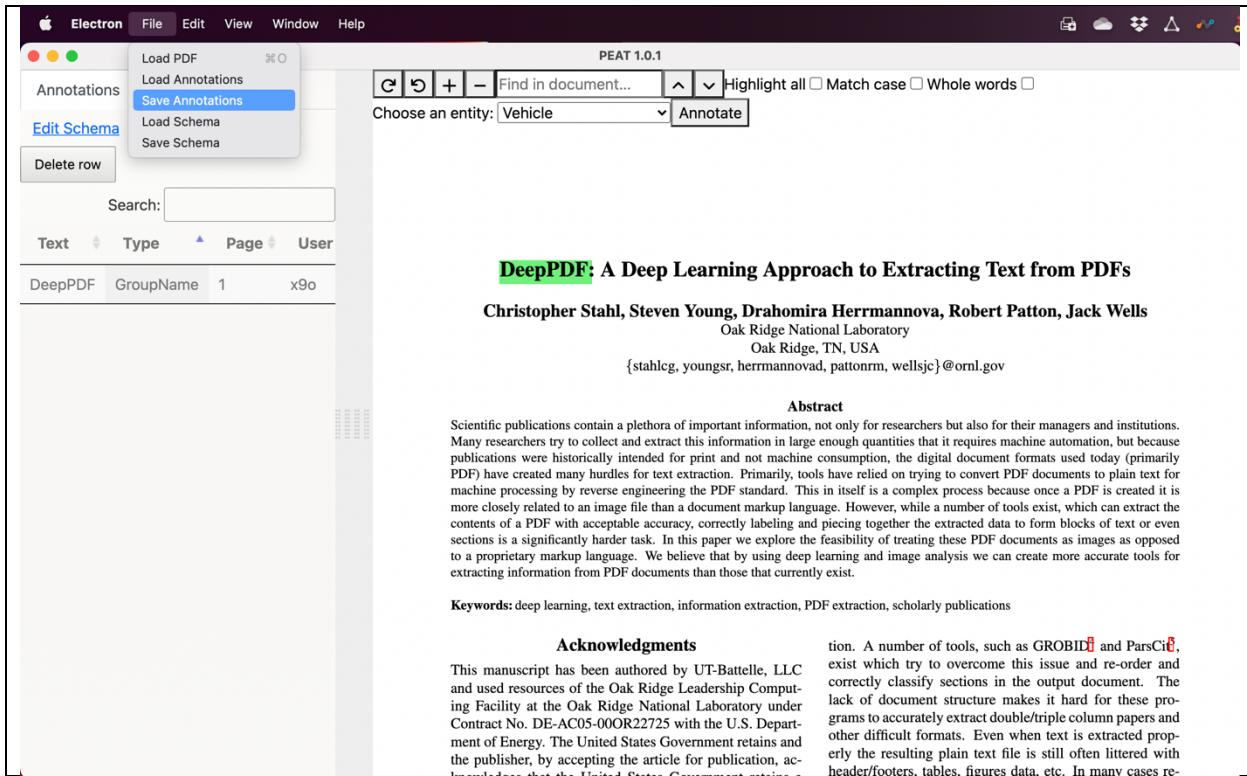
Acknowledgments

This manuscript has been authored by UT-Battelle, LLC and used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a

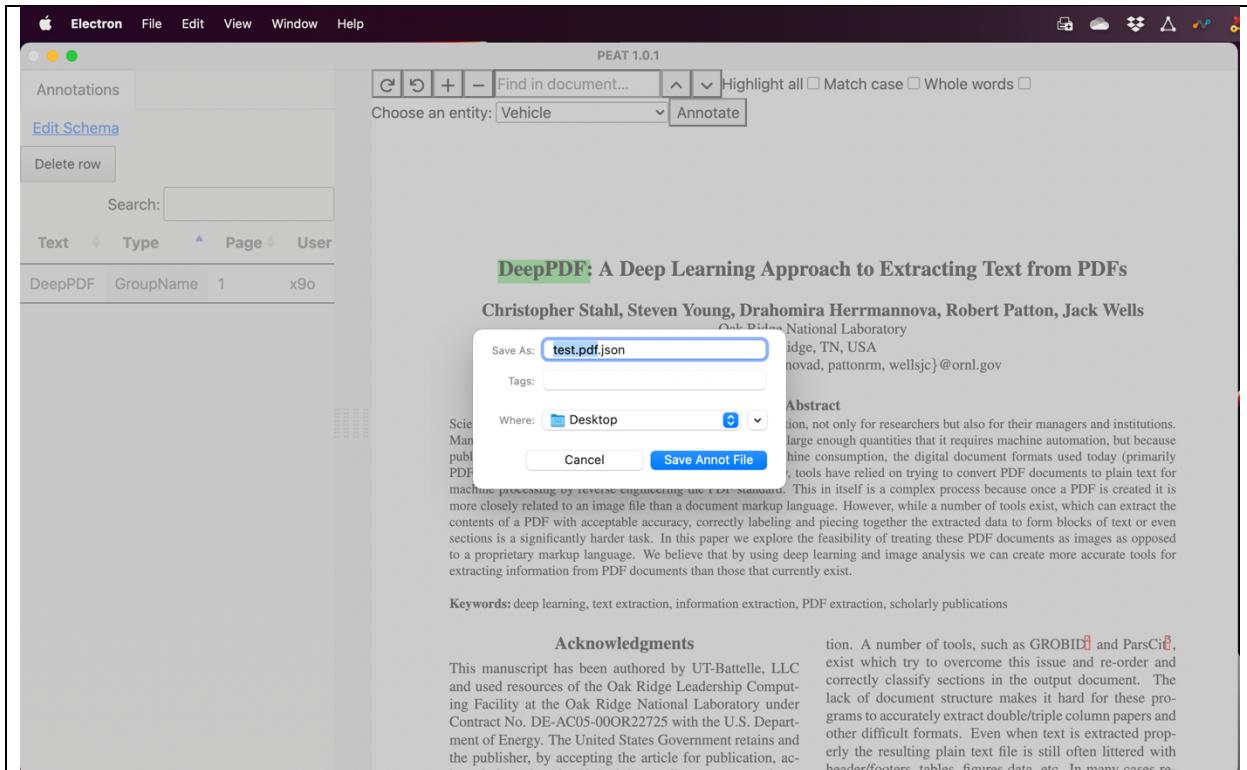
tion. A number of tools, such as GROBID and ParsCit, exist which try to overcome this issue and re-order and correctly classify sections in the output document. The lack of document structure makes it hard for these programs to accurately extract double/triple column papers and other difficult formats. Even when text is extracted properly the resulting plain text file is still often littered with header/footer, tables, figures data, etc. In many cases re-

1.2 Save Annotations

1. Click *File* in the menu bar and select *Save Annotations*.

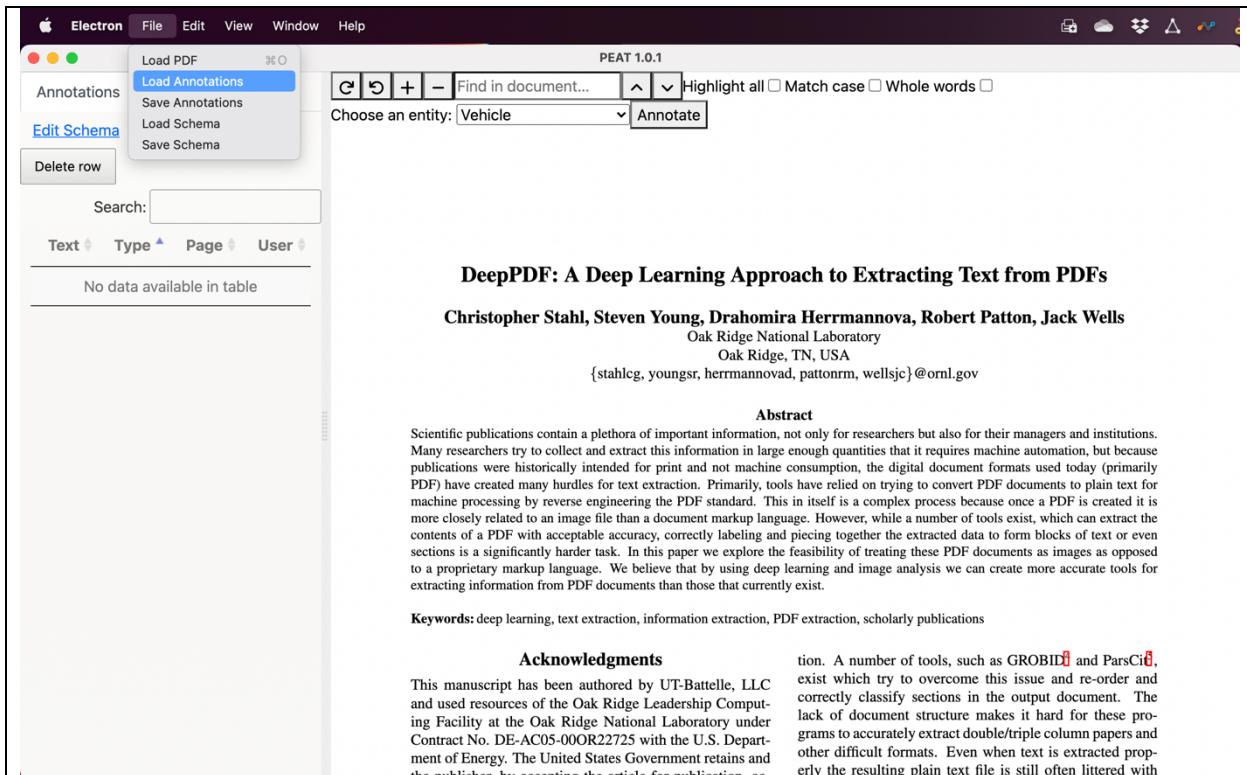


2. Select a save location on your computer and click *Save Annot File*.

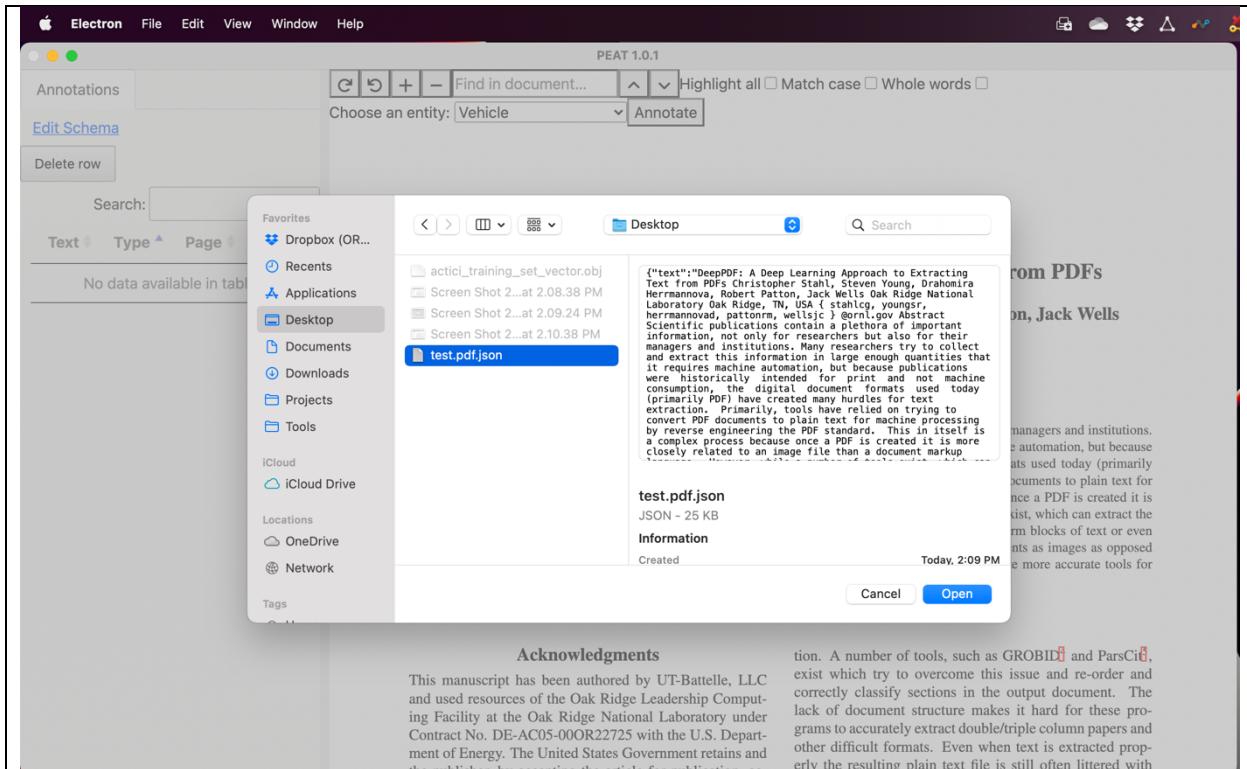


1.3 Load Annotations

1. Click *File* in the menu bar and select *Load Annotations*.



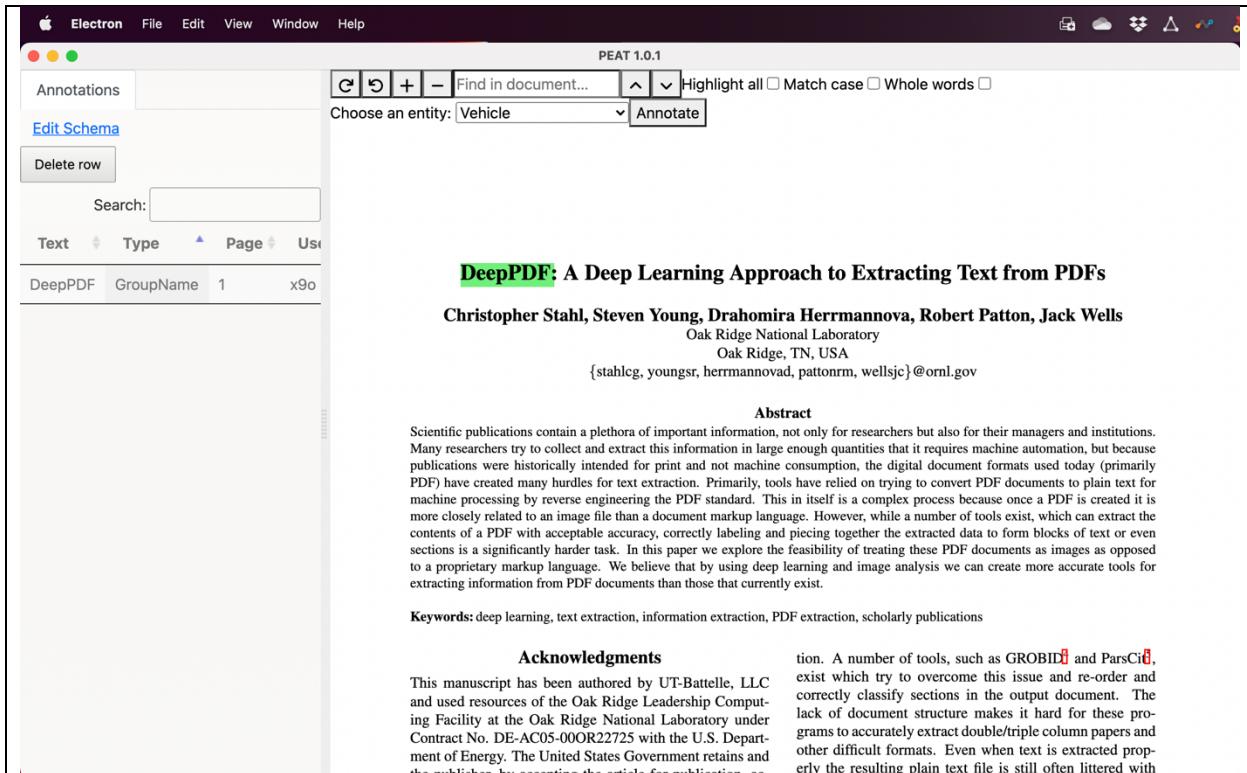
2. Select an annotation file and click *Open*



Acknowledgments

This manuscript has been authored by UT-Battelle, LLC and used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, ac-

tion. A number of tools, such as GROBID^[1] and ParsCit^[2], exist which try to overcome this issue and re-order and correctly classify sections in the output document. The lack of document structure makes it hard for these programs to accurately extract double/triple column papers and other difficult formats. Even when text is extracted properly the resulting plain text file is still often littered with



tion. A number of tools, such as GROBID^[1] and ParsCit^[2], exist which try to overcome this issue and re-order and correctly classify sections in the output document. The lack of document structure makes it hard for these programs to accurately extract double/triple column papers and other difficult formats. Even when text is extracted properly the resulting plain text file is still often littered with

1.4 Delete Annotations

1. Select annotation you wish to delete from the table in the side bar.

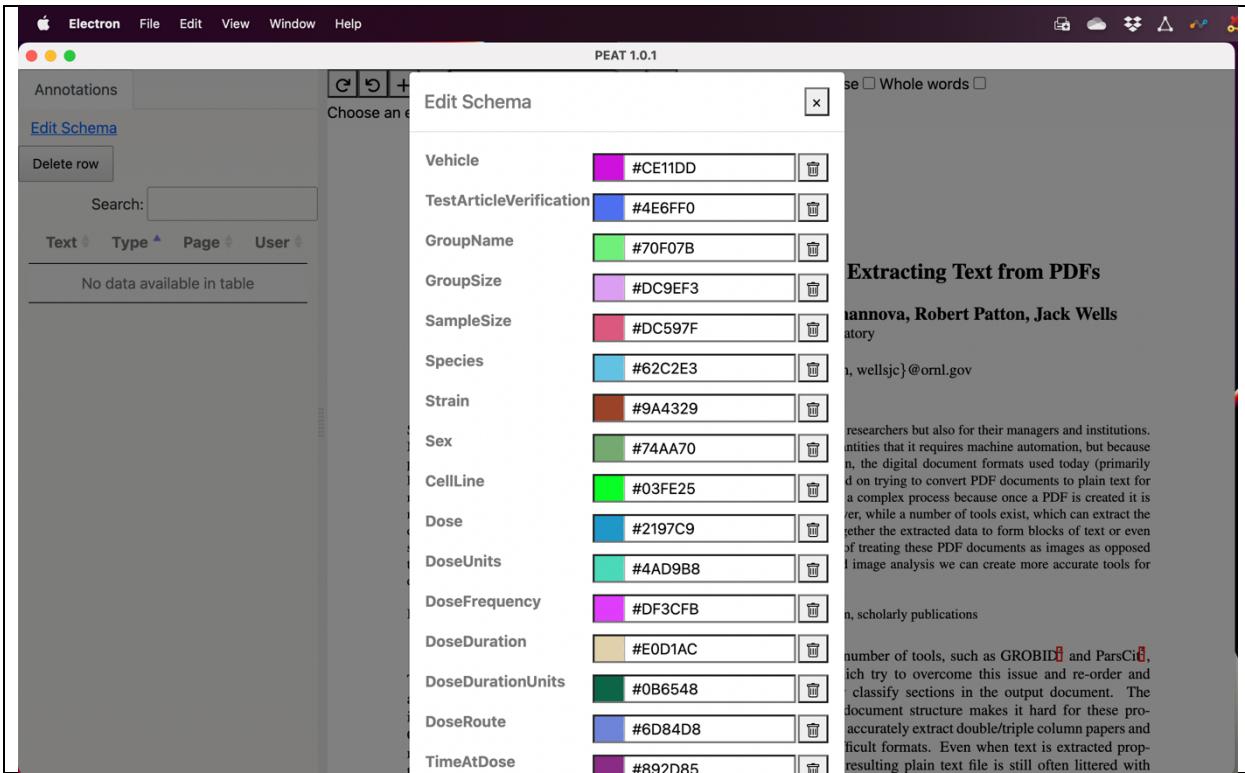
The screenshot shows the PEAT 1.0.1 software interface. On the left, there is a sidebar titled "Annotations" with a "Delete row" button and a search bar. The main area displays a PDF document titled "DeepPDF: A Deep Learning Approach to Extracting Text from PDFs". The document includes author information: Christopher Stahl, Steven Young, Drahomira Herrmannova, Robert Patton, Jack Wells, Oak Ridge National Laboratory, Oak Ridge, TN, USA, and their email addresses. Below the title is the abstract section, which discusses the challenges of extracting text from PDFs and how deep learning can help. The abstract concludes with a statement about the feasibility of treating PDF documents as images. The keywords listed are: deep learning, text extraction, information extraction, PDF extraction, scholarly publications. The acknowledgments section notes that the manuscript was authored by UT-Battelle, LLC, and used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher by accepting the article for publication ac-

2. Click *Delete selected row* button.

The screenshot shows the PEAT 1.0.1 application window. At the top, there's a toolbar with standard Mac OS X icons (File, Edit, View, Window, Help) and a title bar 'PEAT 1.0.1'. Below the toolbar is a search bar with buttons for 'Find in document...', 'Highlight all', 'Match case', and 'Whole words'. A dropdown menu 'Choose an entity' is set to 'Vehicle', and a button 'Annotate' is next to it. To the left, there's a sidebar with a 'Annotations' section, an 'Edit Schema' link, and a 'Delete row' button. Below the sidebar is a search field 'Search:' and a filter section with 'Text', 'Type', 'Page', and 'User' dropdowns. A message 'No data available in table' is displayed. The main content area shows a PDF document titled 'DeepPDF: A Deep Learning Approach to Extracting Text from PDFs' by Christopher Stahl, Steven Young, Drahomira Herrmannova, Robert Patton, and Jack Wells. The authors are listed as 'Christopher Stahl, Steven Young, Drahomira Herrmannova, Robert Patton, Jack Wells' from 'Oak Ridge National Laboratory, Oak Ridge, TN, USA' with email {stahlgc, youngsr, herrmannovad, pattonrm, wellsjw}@ornl.gov'. Below the title is an 'Abstract' section with a block of text about scientific publications and PDF extraction. Underneath the abstract is a 'Keywords' section with 'deep learning, text extraction, information extraction, PDF extraction, scholarly publications'. The 'Acknowledgments' section is present but contains no text. The right side of the document has some cut-off text.

1.5 Edit Schema

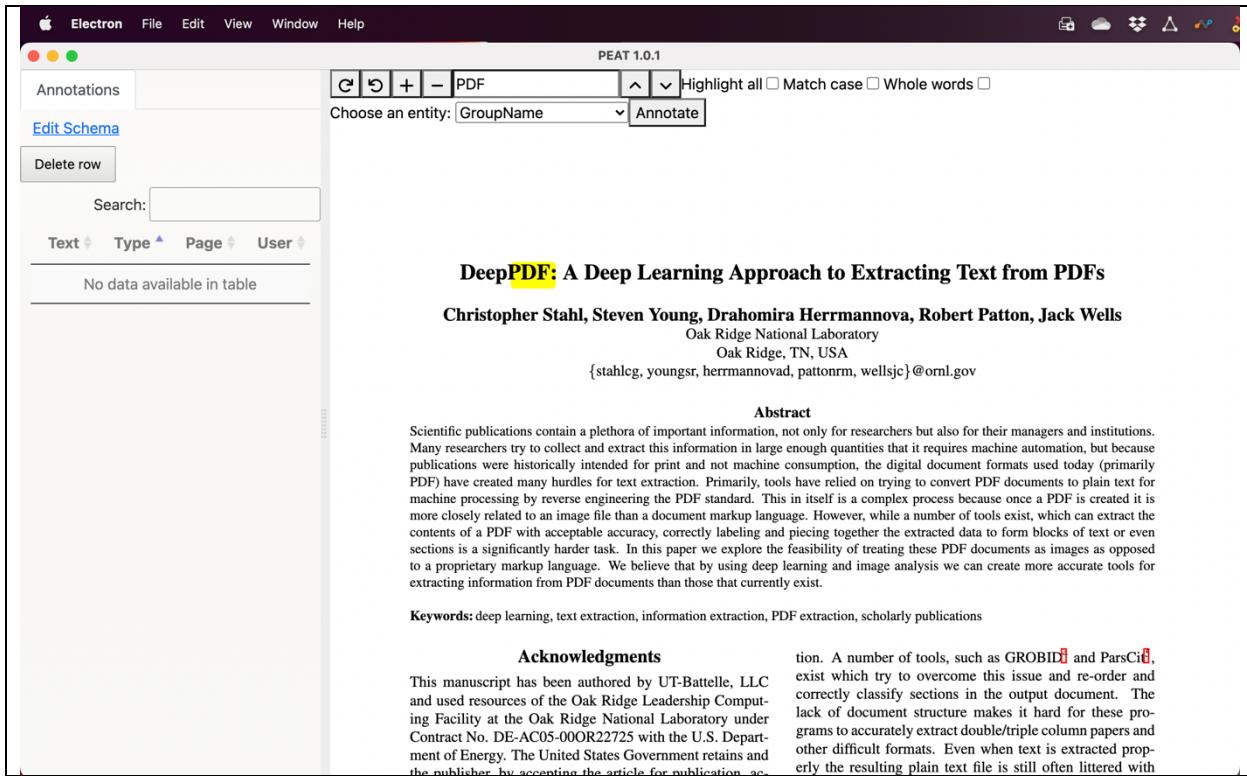
1. Click *Edit Schema* hyper-link



- Change existing entity
 - Click the text of any entity to edit that entities type.
 - Click the color selector to change the annotation color.
 - Click the trash can icon to delete that entity.
- Add new entity type
 - Click Add Entity Type to add a new entity.
- Save changes
 - Click the Save button.

1.6 Auto Annotation

1. Type word or phrase to be searched for in *Find in document* search bar



2. Using the arrows (Up or Down) a yellow highlight will cycle through matches found in the document.
3. Select entity type from the dropdown box.

Annotations

Choose an entity: **GroupName**

Annotations

Search:

Type ▲ Page ▲ User ▲

No data available in table

Vehicle

TestArticleVerification

GroupName

GroupSize

SampleSize

Species

Strain

Sex

CellLine

Dose

DoseUnits

DoseFrequency

DoseDuration

DoseDurationUnits

DoseRoute

TimeAtDose

Learning Approach to Extracting Text from PDFs

Christopher Stahl, Steven Young, Drahomira Herrmannova, Robert Patton, Jack Wells
Oak Ridge National Laboratory
Oak Ridge, TN, USA
`{stahlc, youngsr, herrmannovad, pattonrm, wellsjc}@ornl.gov`

Abstract

Scientific publications contain a plethora of important information, not only for researchers but also for their managers and institutions. Many researchers try to collect and extract this information in large enough quantities that it requires machine automation, but because publications were historically intended for print and not machine consumption, the digital document formats used today (primarily PDF) have created many hurdles for text extraction. Primarily, tools have relied on trying to convert PDF documents to plain text for machine processing by reverse engineering the PDF standard. This in itself is a complex process because once a PDF is created it is more closely related to an image file than a document markup language. However, while a number of tools exist, which can extract the contents of a PDF with acceptable accuracy, correctly labeling and piecing together the extracted data to form blocks of text or even sections is a significantly harder task. In this paper we explore the feasibility of treating these PDF documents as images as opposed to a proprietary markup language. We believe that by using deep learning and image analysis we can create more accurate tools for extracting information from PDF documents than those that currently exist.

Keywords: deep learning, text extraction, information extraction, PDF extraction, scholarly publications

Acknowledgments

This manuscript has been authored by UT-Battelle, LLC and used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, ac-

tion. A number of tools, such as GROBID and ParsCit, exist which try to overcome this issue and re-order and correctly classify sections in the output document. The lack of document structure makes it hard for these programs to accurately extract double/triple column papers and other difficult formats. Even when text is extracted properly the resulting plain text file is still often littered with

4. Click Annotate to add an annotation for the current selection.

Annotations

Choose an entity: **GroupName**

Annotations

Search:

Type ▲ Page ▲ User ▲

PDF GroupName 1 x90

PEAT 1.0.1

Highlight all Match case Whole words

DeepPDF: A Deep Learning Approach to Extracting Text from PDFs

Christopher Stahl, Steven Young, Drahomira Herrmannova, Robert Patton, Jack Wells
Oak Ridge National Laboratory
Oak Ridge, TN, USA
`{stahlc, youngsr, herrmannovad, pattonrm, wellsjc}@ornl.gov`

Abstract

Scientific publications contain a plethora of important information, not only for researchers but also for their managers and institutions. Many researchers try to collect and extract this information in large enough quantities that it requires machine automation, but because publications were historically intended for print and not machine consumption, the digital document formats used today (primarily PDF) have created many hurdles for text extraction. Primarily, tools have relied on trying to convert PDF documents to plain text for machine processing by reverse engineering the PDF standard. This in itself is a complex process because once a PDF is created it is more closely related to an image file than a document markup language. However, while a number of tools exist, which can extract the contents of a PDF with acceptable accuracy, correctly labeling and piecing together the extracted data to form blocks of text or even sections is a significantly harder task. In this paper we explore the feasibility of treating these PDF documents as images as opposed to a proprietary markup language. We believe that by using deep learning and image analysis we can create more accurate tools for extracting information from PDF documents than those that currently exist.

Keywords: deep learning, text extraction, information extraction, PDF extraction, scholarly publications

Acknowledgments

This manuscript has been authored by UT-Battelle, LLC and used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, ac-

tion. A number of tools, such as GROBID and ParsCit, exist which try to overcome this issue and re-order and correctly classify sections in the output document. The lack of document structure makes it hard for these programs to accurately extract double/triple column papers and other difficult formats. Even when text is extracted properly the resulting plain text file is still often littered with