PDF Entity Annotation Tool (PEAT)

January 2023
Version 1.0

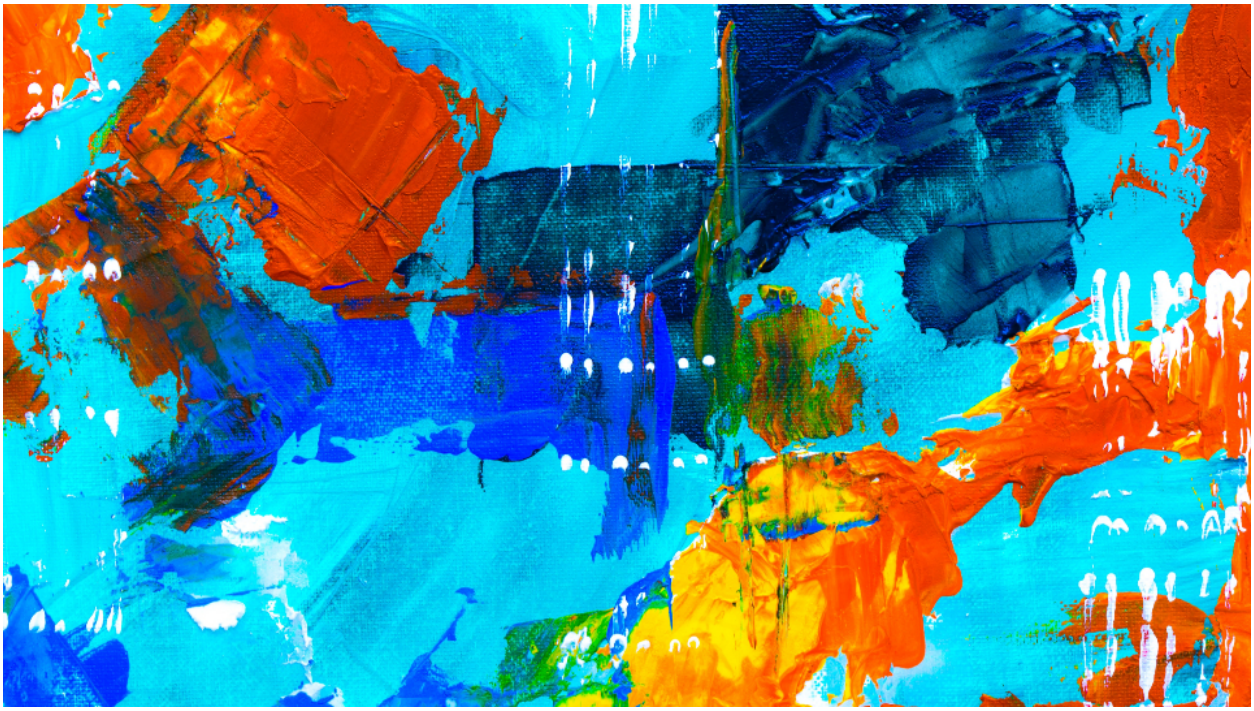


Table of Contents

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 Installation

1.1 Scope and Purpose

The purpose of this project is to further the research and development of tools that NCEA can use in their creation of machine-readable datasets and machine learning research. This effort consists of the following objectives:

- Research and develop software for NCEA that provides the ability to annotate scientific publications for use in machine learning algorithms. This software should be able to accept a list of tags provided by NCEA, allow the user to apply these tags to PDF documents in a web interface, and then extract out the information needed in machine-readable formats that can be used for machine learning.

1.2 Source Code Access

This describes the process of accessing the source code.

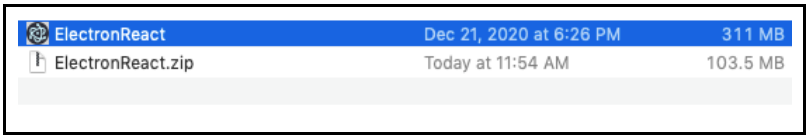
1. https://github.com/USEPA/peat

1.2 Installation Mac OSX

1. Download latest version from

2. Double click ‘mac.zip’ to unzip the file.

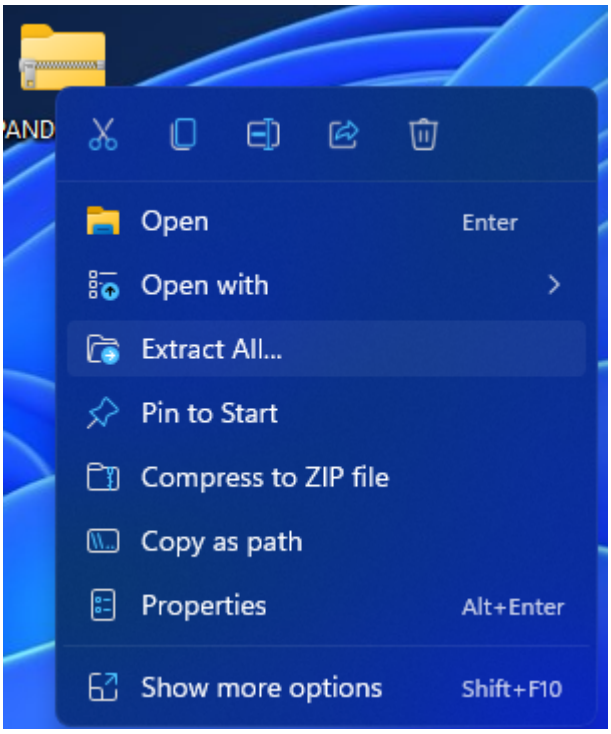
3. Double click ‘ElectronReact’ to launch the application.



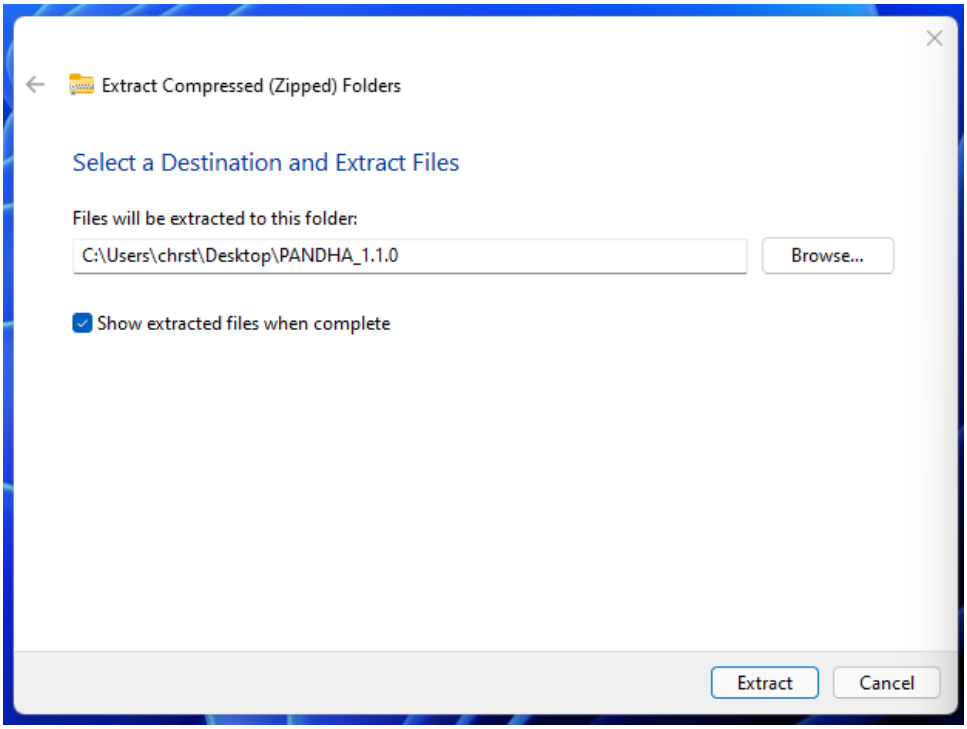
1.3 Installation Windows

1. Download latest version from (PEAT_1.1.0)
https://usepa.sharepoint.com/sites/DOEORNLCPAD/Shared%20Documents/General/Release%20Candidate/PANDHA_1.1.0.zip

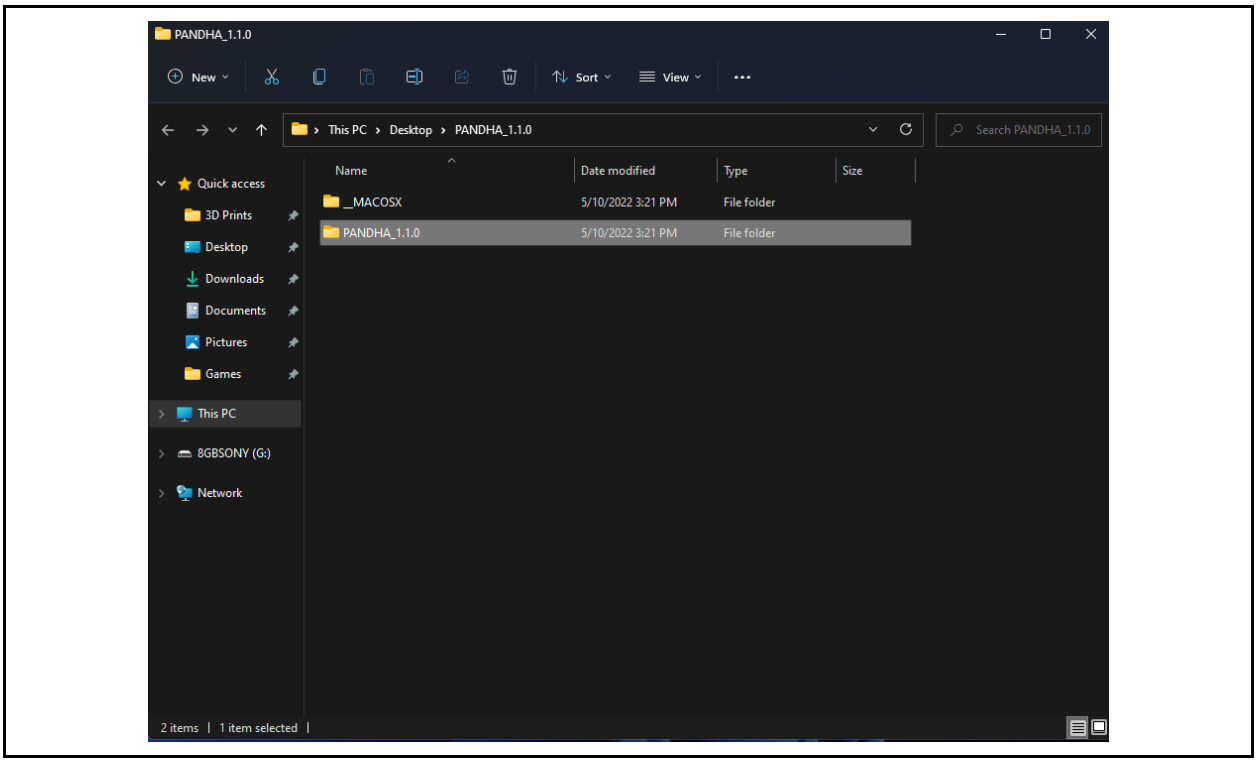
2. Right click PANDHA_1.0.0.zip’ and select ‘Extract All’



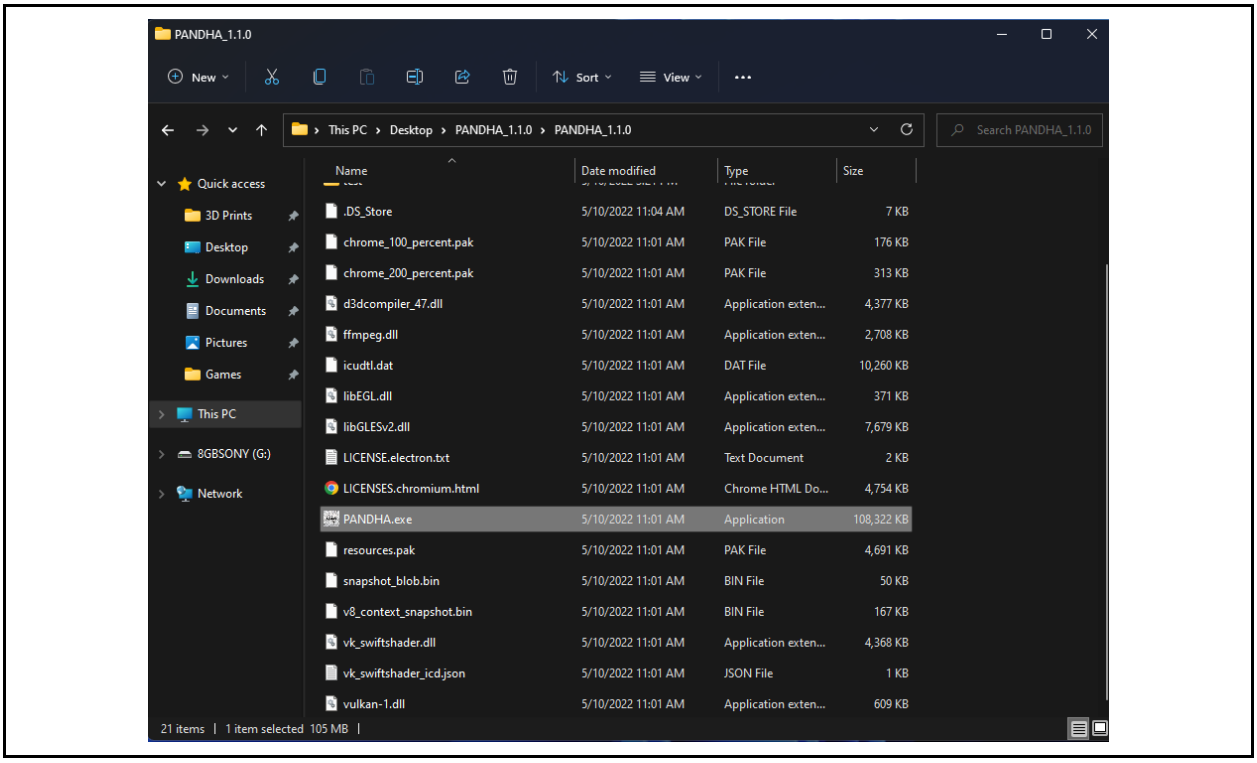
3. Select location and hit *Extract*



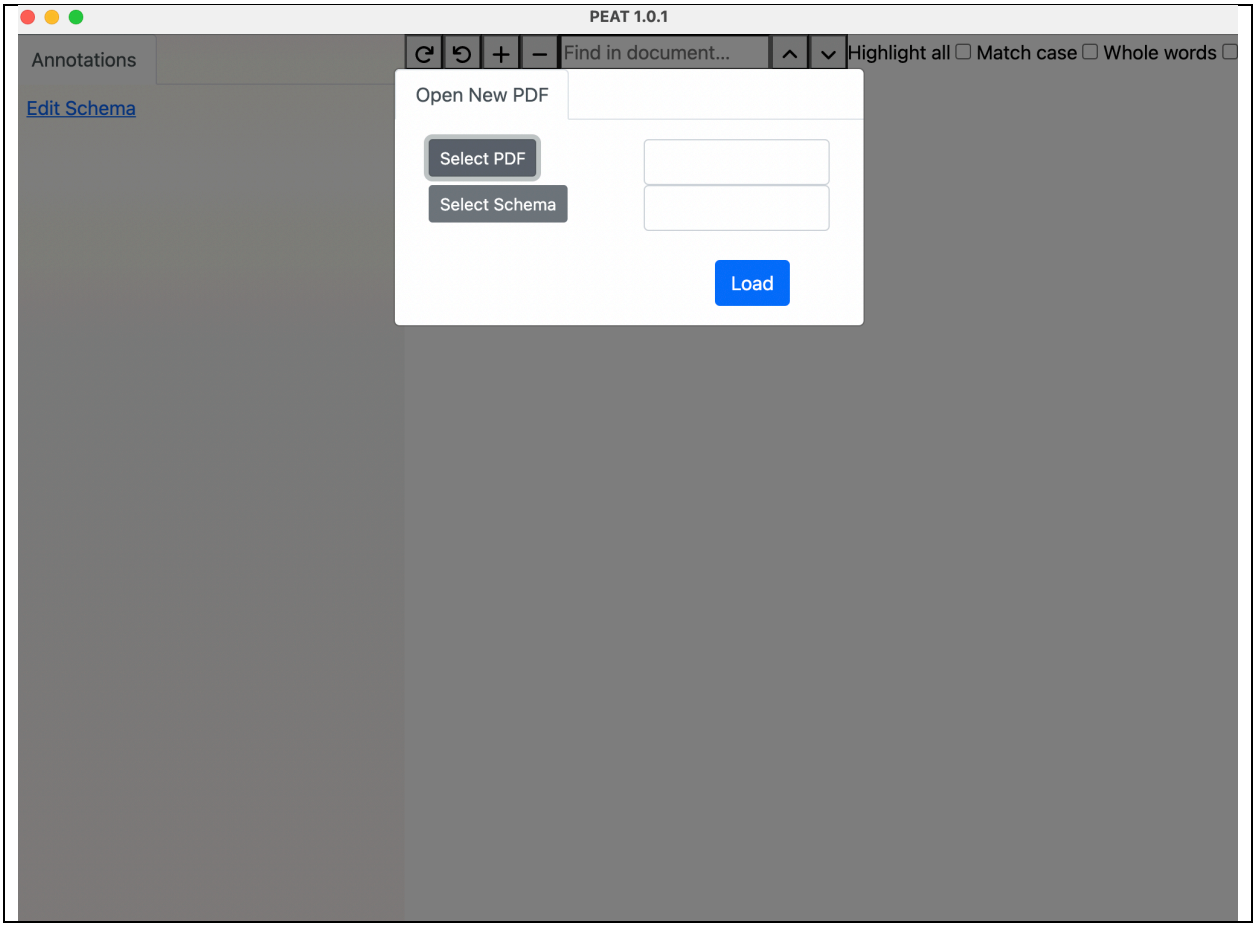
4. From the extracted location double click the *PANDHA_1.1.0*Folder



5. Double click *PEAT.exe* to start the application



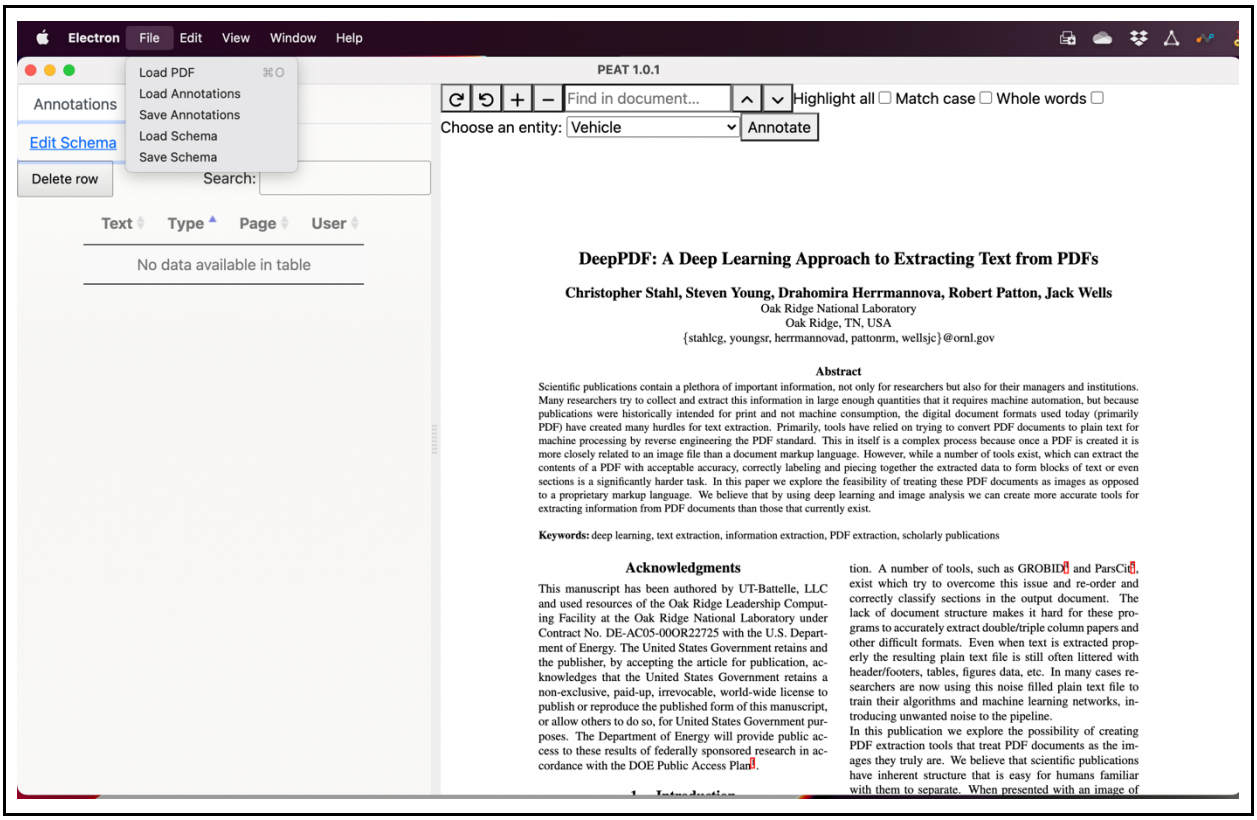
7. Select the PDF and Schema (tags.json is including in the PEAT/test folder) and click *Load*



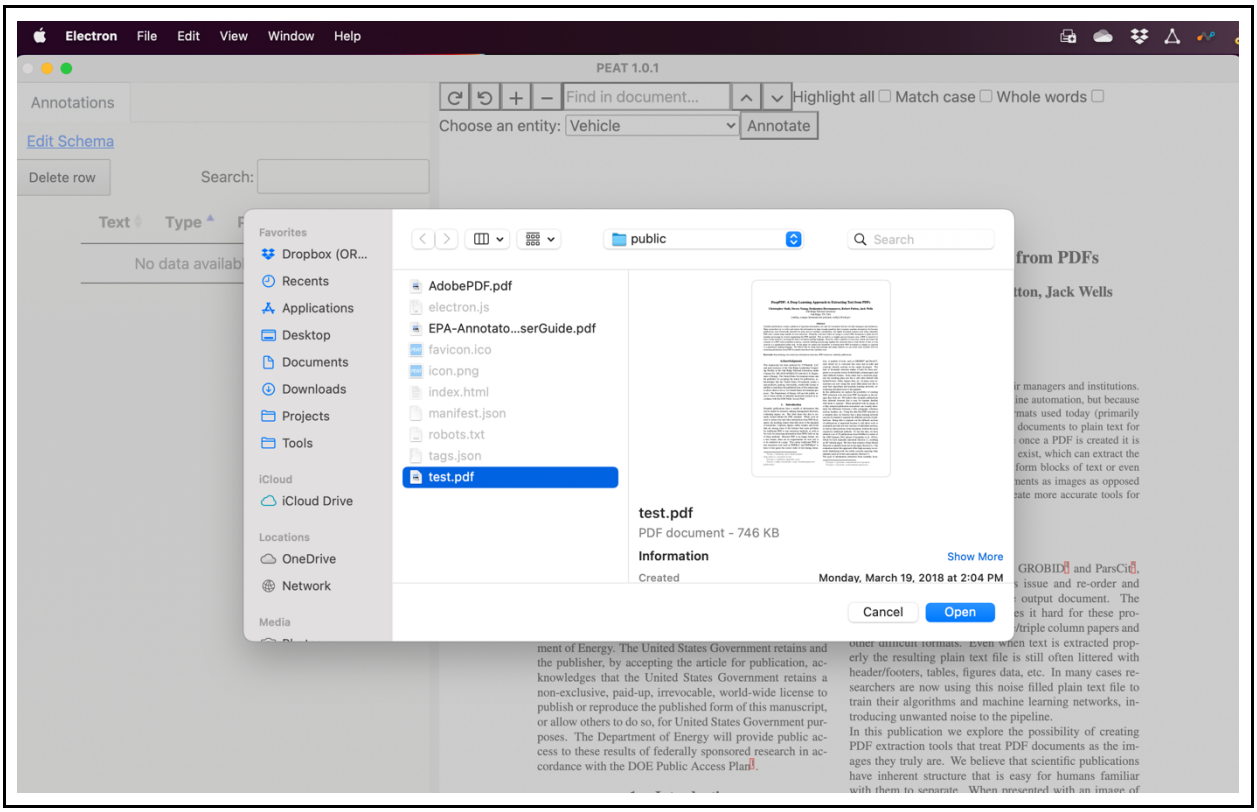
2 Application

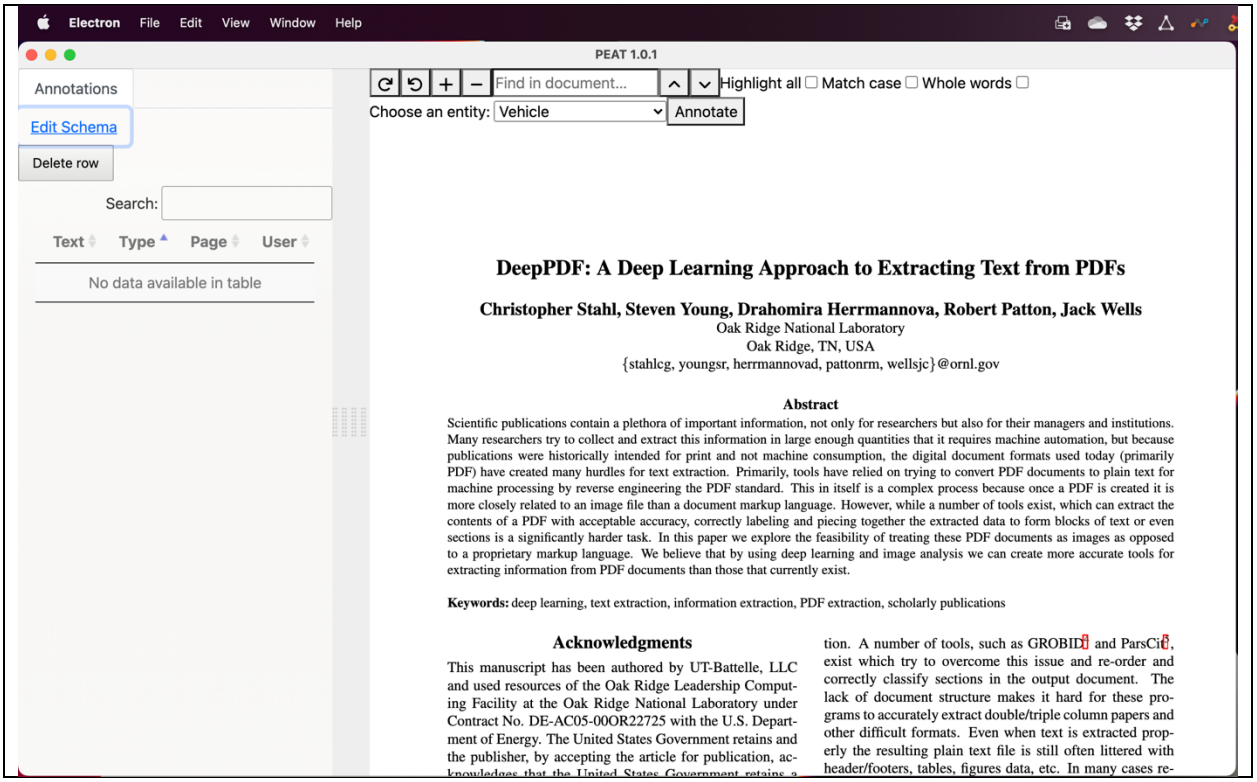
2.0 Load PDF

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



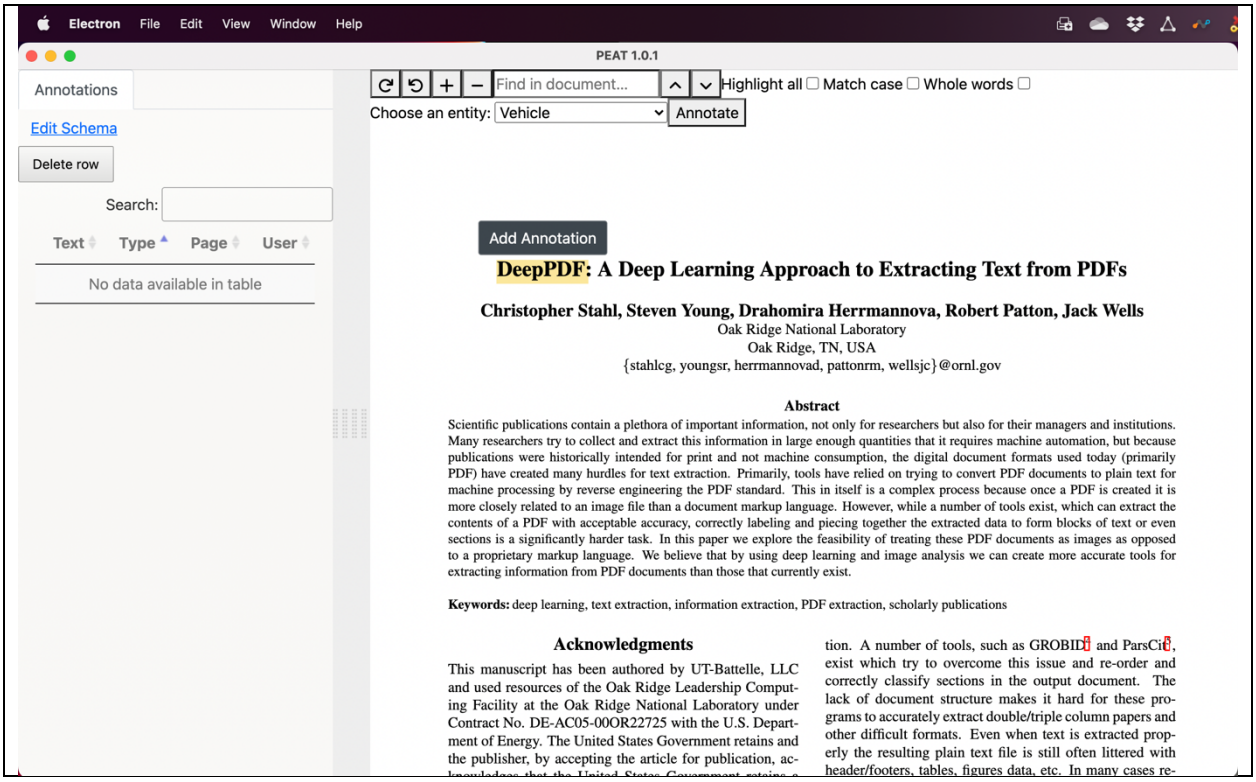
2. Select the PDF file from your computer and click *Open*.


Electron File Edit View Window Help
PEAT 1.0.1
Annotations
Edit Schema
Delete row
Search:
Text Type
No data availab
Find in document...
Highlight all Match case Whole words
Choose an entity: Vehicle
Annotate
Favorites
Dropbox (OR...
Recents
Applications
Desktop
Documents
Downloads
Projects
Tools
iCloud
iCloud Drive
Locations
OneDrive
Network
Media
public
Search
AdobePDF.pdf
electron.js
EPA-Annotato...serGuide.pdf
favicon.ico
icon.png
index.html
manifest.json
robots.txt
tags.json
test.pdf
test.pdf
PDF document - 746 KB
Information
Show More
Created
Monday, March 19, 2018 at 2:04 PM
Cancel
Open
from PDFs
tton, Jack Wells

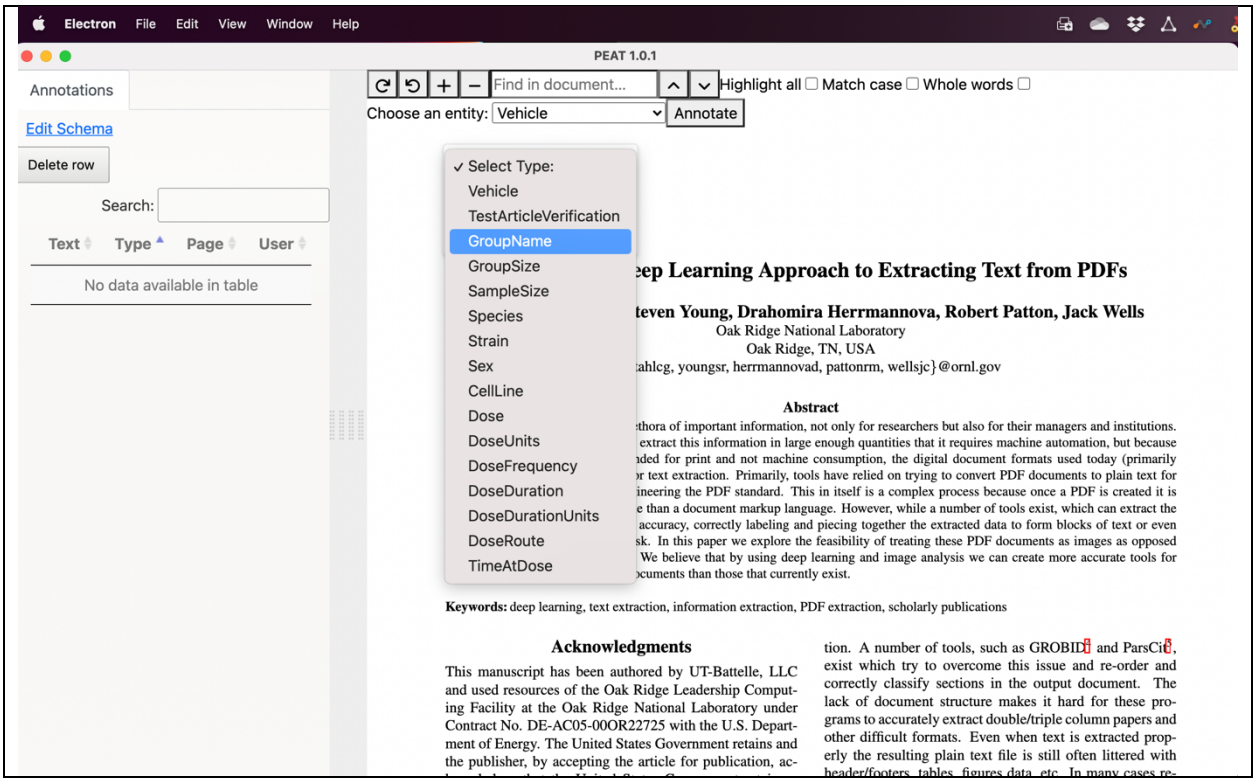

Electron File Edit View Window Help
PEAT 1.0.1
Annotations
Edit Schema
Delete row
Search:
Text Type Page User
No data available in table
Find in document...
Highlight all Match case Whole words
Choose an entity: Vehicle
Annotate
DeepPDF: A Deep Learning Approach to Extracting Text from PDFs
Christopher 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, 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 header/footers, tables, figures data, etc. In many cases re-

2.1 Annotate PDF

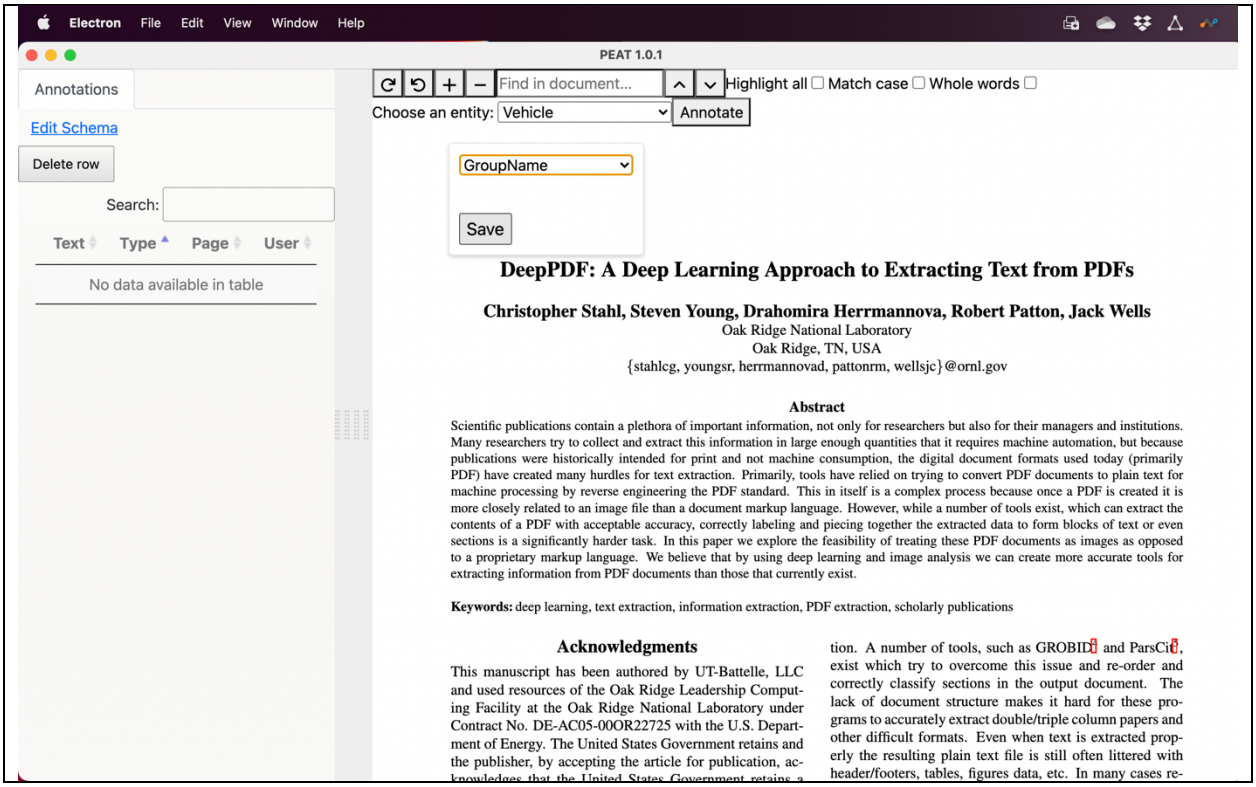
1. Highlight text you wish to annotate and select *Add Annotation*.

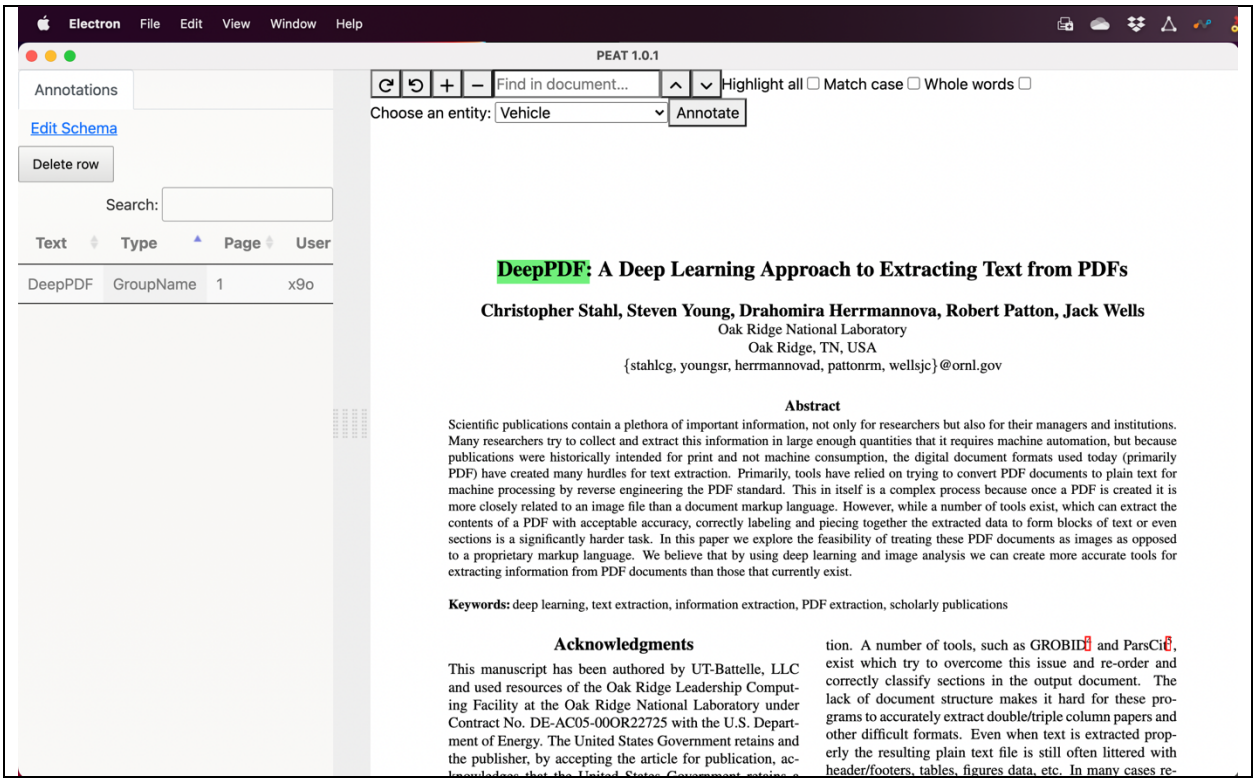


2. Select the annotation type.


Electron File Edit View Window Help
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
Select Type:
Vehicle
TestArticleVerification
GroupName
GroupSize
SampleSize
Species
Strain
Sex
CellLine
Dose
DoseUnits
DoseFrequency
DoseDuration
DoseDurationUnits
DoseRoute
TimeAtDose
eep Learning Approach to Extracting Text from PDFs
teven Young, Drahomira Herrmannova, Robert Patton, Jack Wells
Oak Ridge National Laboratory
Oak Ridge, TN, USA
tahlcg, youngsr, herrmannovad, pattonrm, wellsjc}@ornl.gov
Abstract
thora of important information, not only for researchers but also for their managers and institutions.
extract this information in large enough quantities that it requires machine automation, but because
ded for print and not machine consumption, the digital document formats used today (primarily
r text extraction. Primarily, tools have relied on trying to convert PDF documents to plain text for
ineering the PDF standard. This in itself is a complex process because once a PDF is created it is
e than a document markup language. However, while a number of tools exist, which can extract the
accuracy, correctly labeling and piecing together the extracted data to form blocks of text or even
sk. In this paper we explore the feasibility of treating these PDF documents as images as opposed
We believe that by using deep learning and image analysis we can create more accurate tools for
cuments 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 header/footers, tables, figures data, etc. In many cases re-

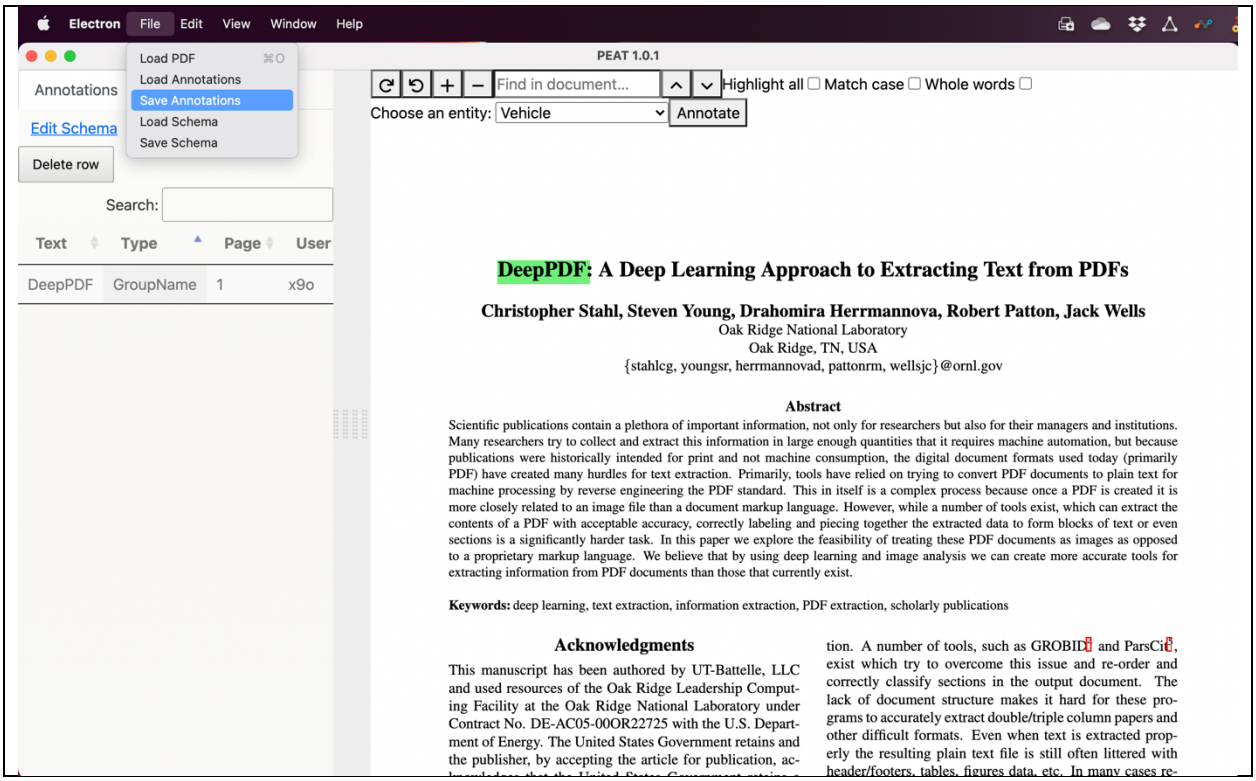
3. Hit save


Electron File Edit View Window Help
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
GroupName
Save
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, 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 header/footers, tables, figures data, etc. In many cases re-

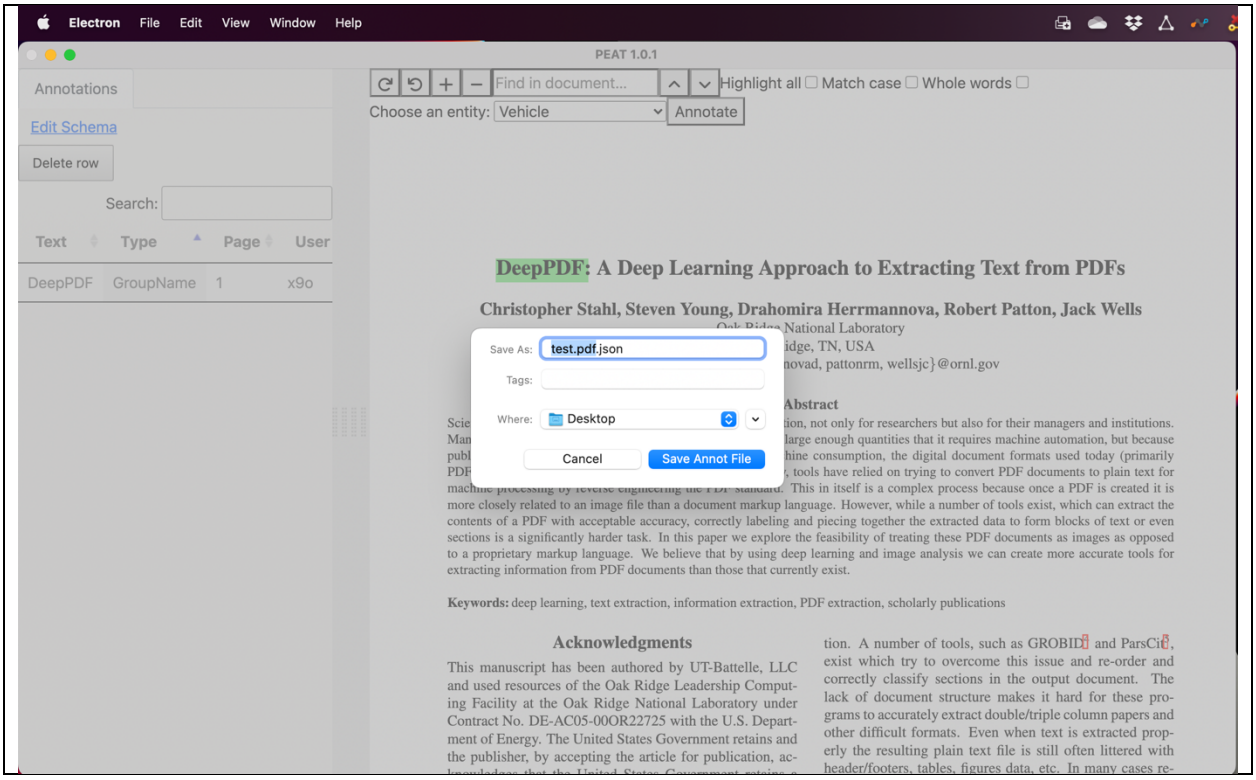


2.4 Save Annotations

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

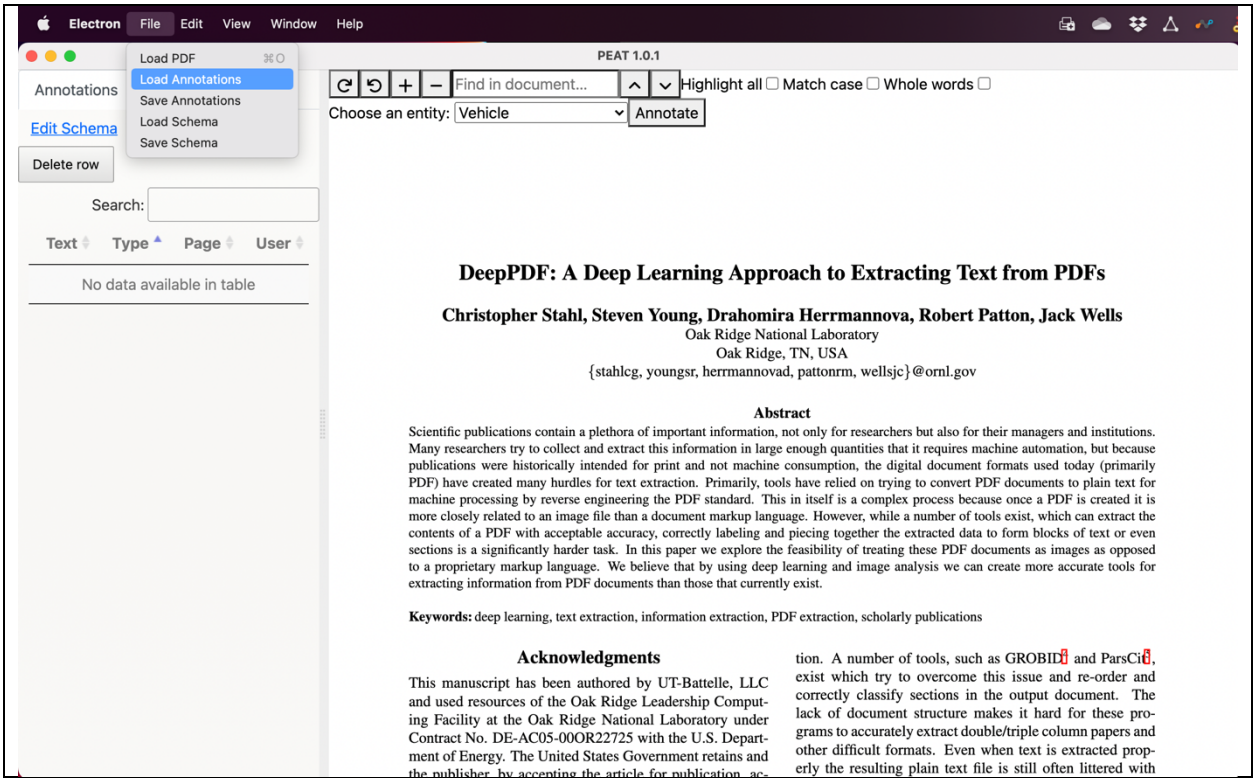


2. Select a save location on your computer and click *Save Annot File*.

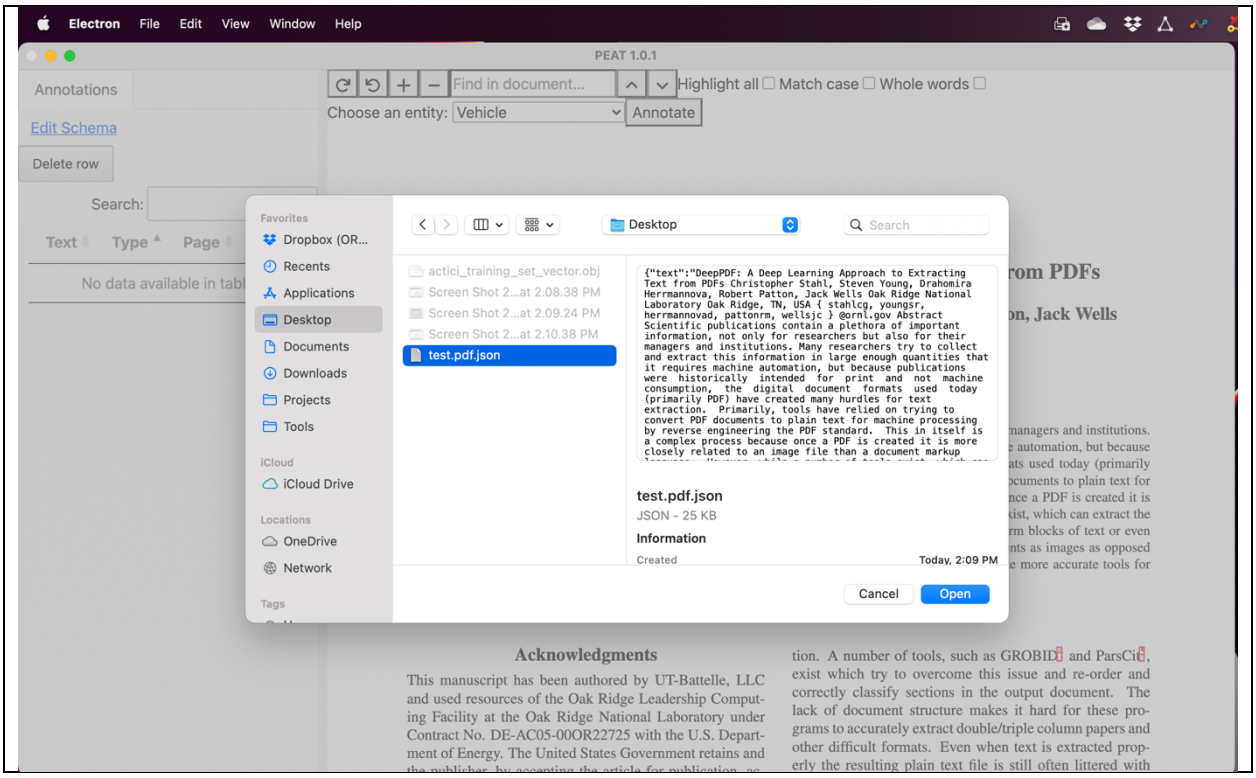


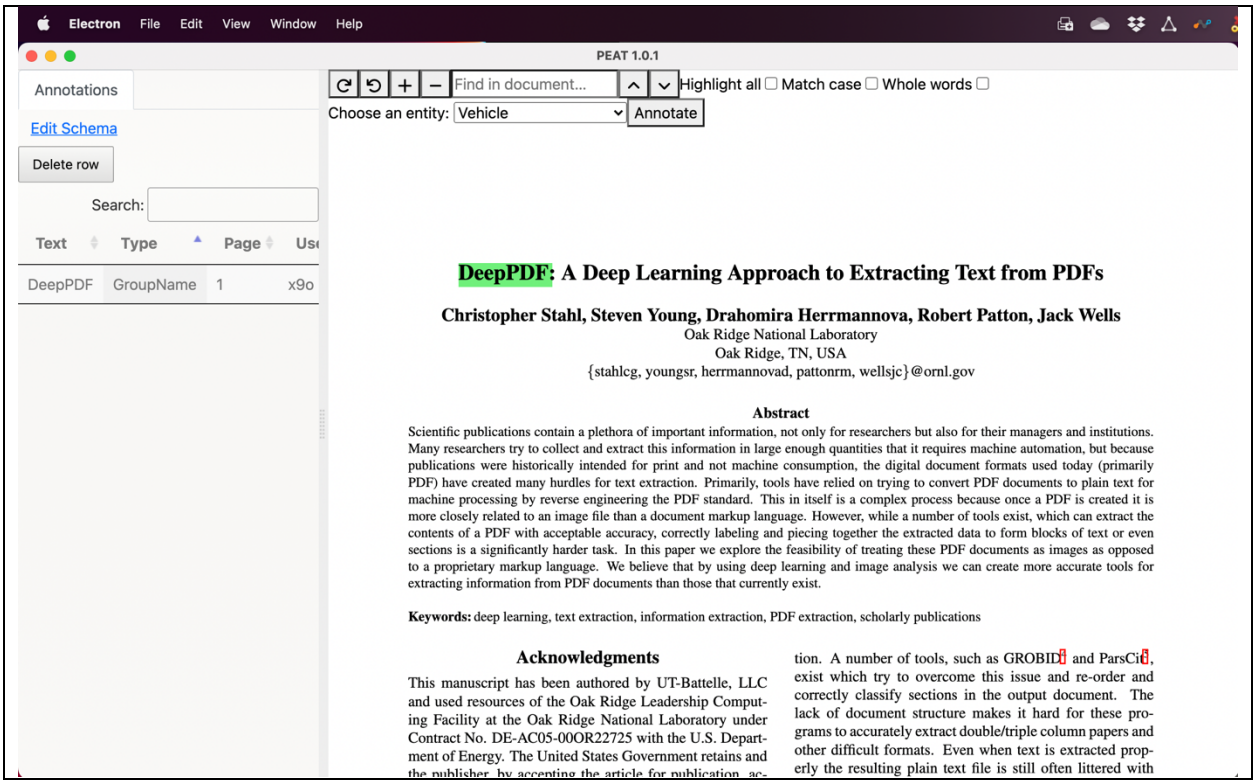
2.5 Load Annotations

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



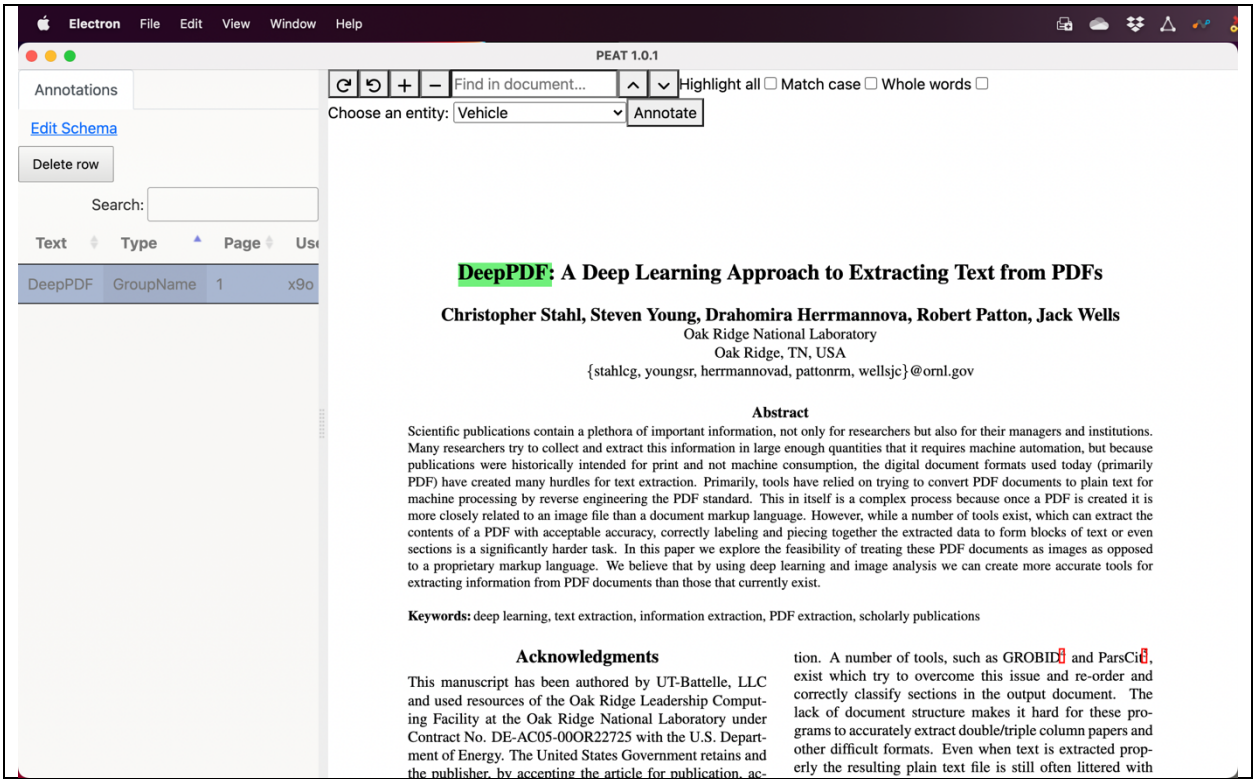
2. Select an annotation file and click *Open*


Electron File Edit View Window Help
PEAT 1.0.1
Annotations
Edit Schema
Delete row
Search:
Text Type Page
No data available in table
Find in document...
Highlight all Match case Whole words
Choose an entity: Vehicle
Annotate
Favorites
Dropbox (OR...
Recents
Applications
Desktop
Documents
Downloads
Projects
Tools
iCloud
iCloud Drive
Locations
OneDrive
Network
Tags
Desktop
Search
acticl_training_set_vector.obj
Screen Shot 2...at 2.08.38 PM
Screen Shot 2...at 2.09.24 PM
Screen Shot 2...at 2.10.38 PM
test.pdf.json
{"text":"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
test.pdf.json
JSON - 25 KB
Information
Created
Today, 2:09 PM
Cancel
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
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

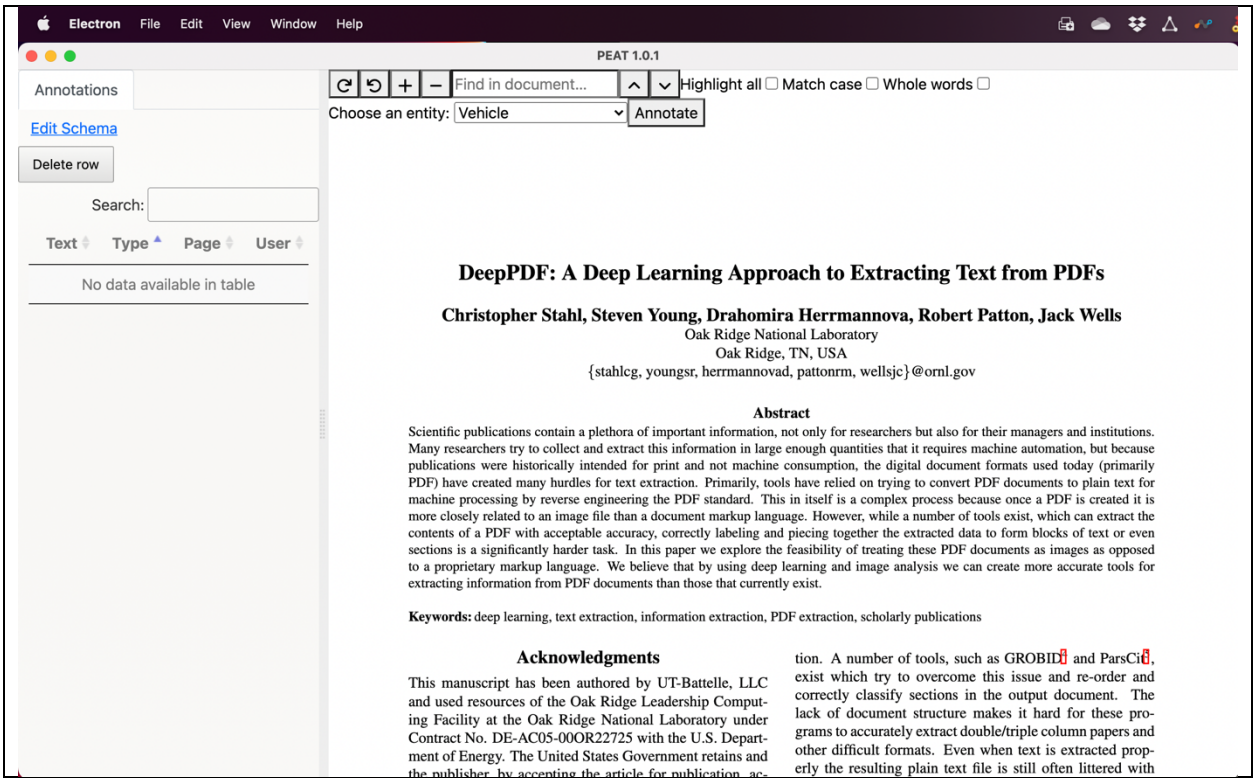

Electron File Edit View Window Help
PEAT 1.0.1
Annotations
Edit Schema
Delete row
Search:
Text Type Page Us
DeepPDF GroupName 1 x9o
Find in document...
Highlight all Match case Whole words
Choose an entity: Vehicle
Annotate
DeepPDF: A Deep Learning Approach to Extracting Text from PDFs
Christopher Stahl, 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
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

2.6 Delete Annotations

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

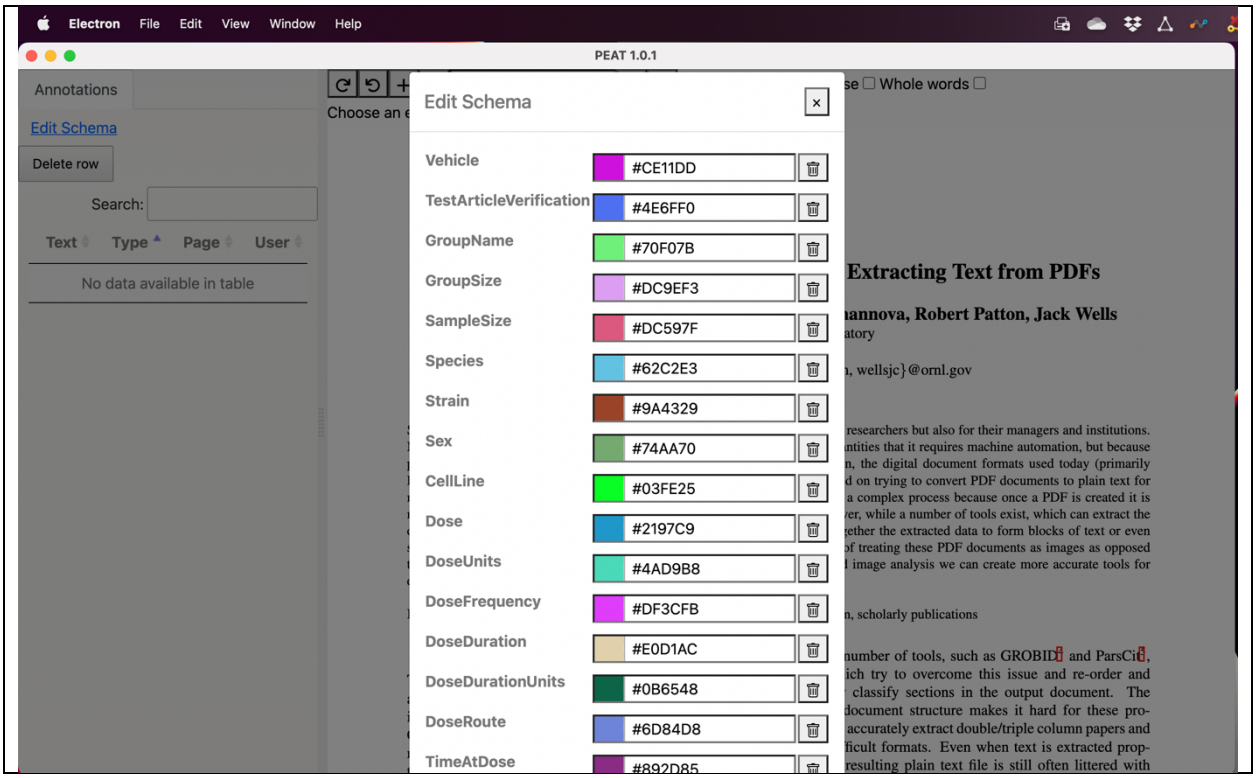


2. Click *Delete selected row* button.



2.6 Edit Schema

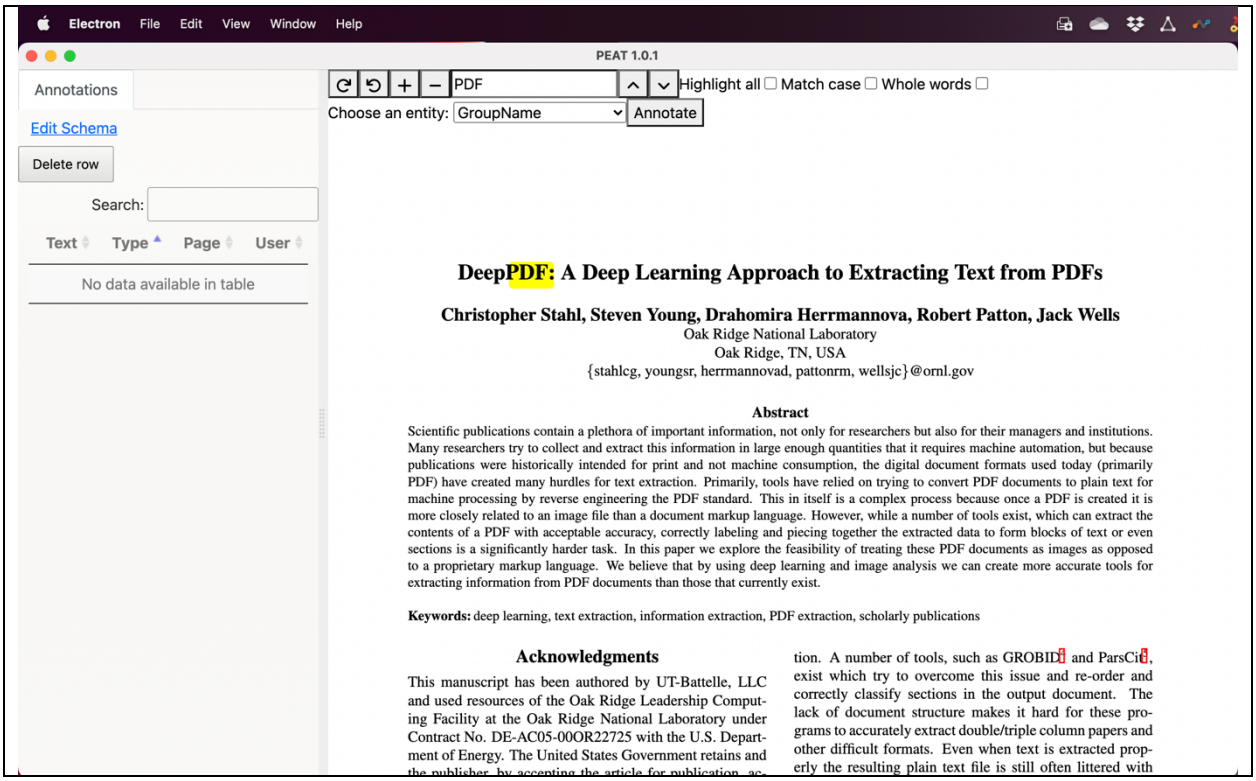
1. Click *Edit Schema* hyper-link



- 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.

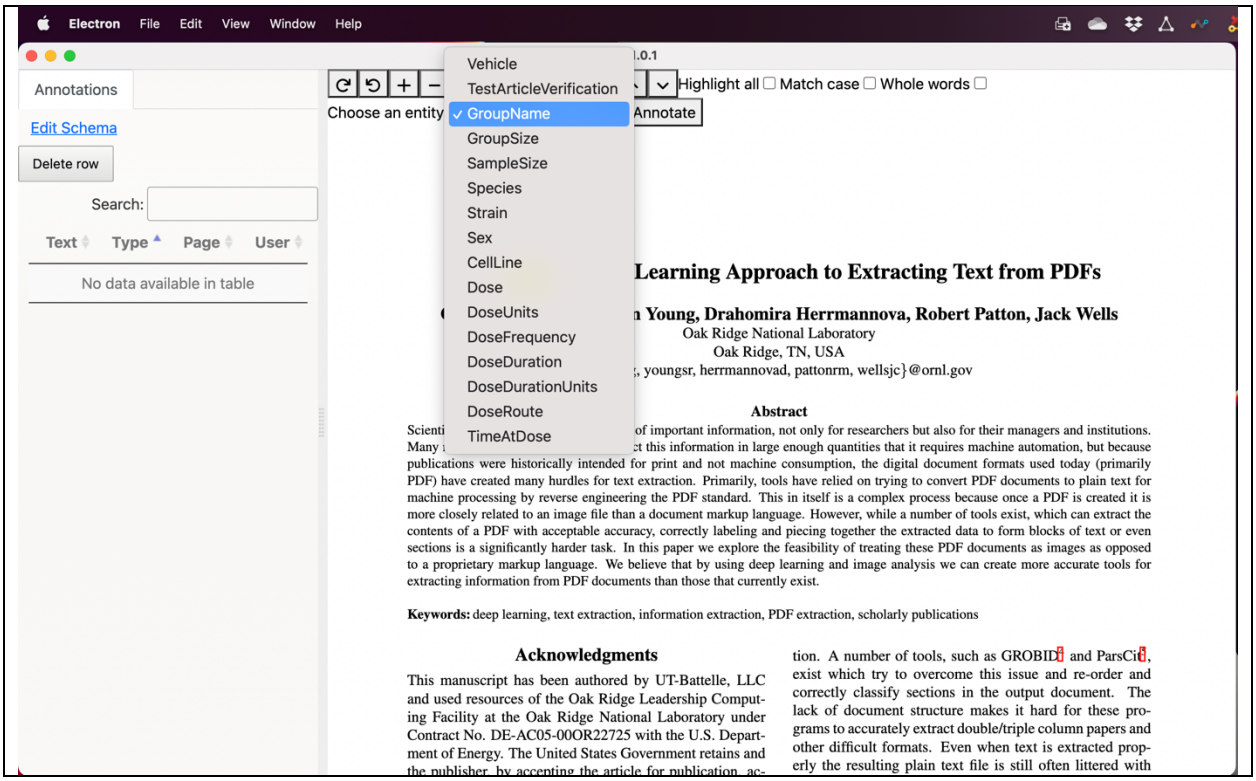
2.7 Auto Annotation

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.



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