Excalibur Documentation

Release 0.4.3

Camelot Developers

Jul 17, 2020

Contents

1 Using Excalibur **3**

2 Why Excalibur? **5**

3 Support us on OpenCollective **7**

4 The User Guide **9**

4.1 Introduction 9

4.2 Installation of Excalibur 10

4.3 How-to Guides 10

4.4 Usage with screenshots 12

4.5 FAQ 15

5 The Contributor Guide **17**

5.1 Contributor's Guide 17

Release v0.4.3. (*Installation*)

chat on gitter

Excalibur is a web interface to extract tabular data from PDFs, written in **Python 3**! It powered by Camelot.

Note: Excalibur only works with text-based PDFs and not scanned documents. (As Tabula explains, "If you can click and drag to select text in your table in a PDF viewer, then your PDF is text-based".)

CHAPTER 1

Using Excalibur

Note: You need to install ghostscript before moving forward.

After *installing Excalibur with pip* (), you can initialize the metadata database using:

```
$ excalibur initdb
```

And then start the webserver using:

```
$ excalibur webserver
```

That’s it! Now you can go to http://localhost:5000 and start extracting tabular data from your PDFs.

1. **Upload** a PDF and enter the page numbers you want to extract tables from.
2. Go to each page and select the table by drawing a box around it. (You can choose to skip this step since Excalibur can automatically detect tables on its own. Click on “**Autodetect tables**” to see what Excalibur sees.)
3. Choose a flavor (Lattice or Stream) from “**Advanced**”.
    a. **Lattice**: For tables formed with lines.
    b. **Stream**: For tables formed with whitespaces.
4. Click on “**View and download data**” to see the extracted tables.
5. Select your favorite format (CSV/Excel/JSON/HTML) and click on “**Download**”!

Note: You can also download executables for Windows and Linux from the releases page and run them directly!

CHAPTER 2

Why Excalibur?

- Extracting tables from PDFs is hard. A simple copy-and-paste from a PDF into an Excel doesn't preserve table structure. **Excalibur makes PDF table extraction very easy**, by automatically detecting tables in PDFs and letting you save them into CSVs and Excels.
- Excalibur uses Camelot under the hood, which gives you additional settings to tweak table extraction and get the best results. You can see how it performs better than other open-source tools and libraries in this comparison.
- You can save table extraction *settings* (like table areas) for a PDF once, and apply them on new PDFs to extract tables with similar structures.
- You get complete control over your data. All file storage and processing happens on your own local or remote machine.
- Excalibur can be configured with MySQL and Celery for parallel and distributed workloads. By default, sqlite and multiprocessing are used for sequential workloads.

CHAPTER 3

Support us on OpenCollective

If Excalibur helped you extract tables from PDFs, please consider supporting its development by becoming a backer or a sponsor on OpenCollective!

CHAPTER 4

The User Guide

This part of the documentation focuses on instructions to get you up and running with Excalibur.

4.1 Introduction

Excalibur is a web interface built on top of Camelot, which is a Python library to extract tabular data from PDFs.

4.1.1 What's in a name?

Camelot was named after The Camelot Project (also the name of a castle in the Arthurian legend). To follow the same theme, this project was named Excalibur, which is the legendary sword of King Arthur.

4.1.2 Why another tool?

There are both open (Tabula, pdfplumber) and closed-source (Smallpdf, Docparser) tools that are widely used to extract data tables from PDFs. They either give a nice output or fail miserably. There is no in between. This is not helpful since everything in the real world, including PDF table extraction, is fuzzy. This leads to the creation of ad-hoc table extraction scripts for each type of PDF table.

Excalibur uses Camelot under the hood, which was created to offer users complete control over table extraction. If you can't get your desired output with the default settings, you can tweak them and get the job done!

Here is a comparison of Camelot's output with outputs from other open-source PDF parsing libraries and tools.

4.1.3 Excalibur LICENSE

> MIT License
>
> Copyright (c) Camelot Developers 2018

4.2 Installation of Excalibur

This part of the documentation covers the steps to install Excalibur.

4.2.1 Using pip

After installing ghostscript, which is one of the requirements for Camelot (See install instructions), you can simply use pip to install Excalibur:

```
$ pip install excalibur-py
```

4.2.2 From the source code

After installing ghostscript, clone the repo using:

```
$ git clone https://www.github.com/camelot-dev/excalibur
```

and install Excalibur using pip:

```
$ cd excalibur
$ pip install .
```

4.3 How-to Guides

Excalibur’s architecture is heavily inspired from Airflow, so you may experience déjà vu while reading this page of the documentation. Airflow LICENSE.

4.3.1 Setting Configuration Options

The first time you run Excalibur, it will create a file called `excalibur.cfg` in your `$EXCALIBUR_HOME` directory (`~/excalibur` by default). This file contains Excalibur’s configuration and you can edit it to change any of the settings.

For example, the metadata database connection string can be set in `excalibur.cfg` like this:

```
[core]
sql_alchemy_conn = my_conn_string
```

4.3.2 Resetting the Metadata Database

Warning: The following command will wipe your Excalibur metadata database, removing all information about uploaded files, saved settings and finished/in-progress jobs.

You can reset the metadata database using:

```
$ excalibur resetdb
```

4.3.3 Using the MySQL Database Backend

Excalibur uses SqlAlchemy to connect to a database backend. By default, a sqlite database is used. To use MySQL, you need to first install MySQL and then create a database and a user.

Installing MySQL

To use the MySQL database backend, you need to install Excalibur using:

```
$ pip install excalibur-py[mysql]
```

You can install MySQL using your system's package manager. For Ubuntu:

```
$ sudo apt update
$ sudo apt install mysql-server libmysqlclient-dev
```

And then set it up using:

```
$ sudo mysql_secure_installation
```

Setup

Now you can create the a database and a user for Excalibur:

```
> CREATE DATABASE excalibur CHARACTER SET utf8 COLLATE utf8_unicode_ci;
> GRANT ALL ON excalibur.* TO 'excalibur'@'%' IDENTIFIED BY '1234';
```

Finally, you need to change the `sql_alchemy_conn` in `excalibur.cfg` to:

```
[core]
sql_alchemy_conn = mysql://excalibur:1234@localhost:3306/excalibur
```

And initialize the metadata database using:

```
$ excalibur initdb
```

4.3.4 Scaling Out with Celery

`CeleryExecutor` is one of the ways you can scale out the number of workers. For this to work, you need to setup a Celery backend (RabbitMQ, Redis, . . .) and change your excalibur.cfg to point the executor parameter to `CeleryExecutor` and provide the related Celery settings.

For more information about setting up a Celery broker, refer to the exhaustive Celery documentation on the topic.

To kick off a worker, you need to setup Excalibur and kick off the worker subcommand:

```
$ excalibur worker
```

Your worker should start picking up tasks as soon as they get fired in its direction.

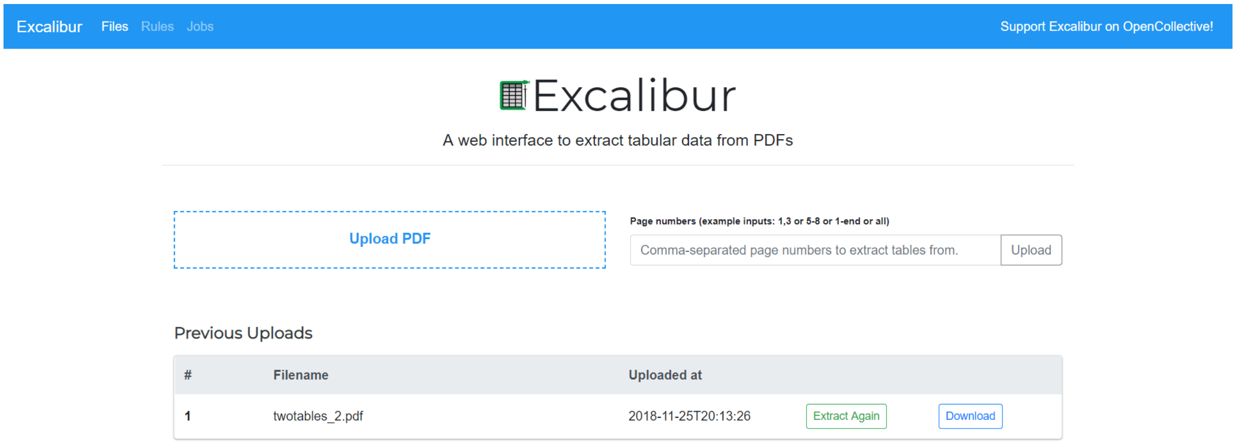
4.4 Usage with screenshots

This part of the documentation demonstrates usage of the web interface.

A table extraction workflow on Excalibur can be broken down into three simple steps.

4.4.1 Upload a PDF

When you open http://localhost:5000, you will see the following screen.


Excalibur
Files
Rules
Jobs
Support Excalibur on OpenCollective!
Excalibur
Upload PDF
Page numbers (example inputs: 1,3 or 5-8 or 1-end or all)
Comma-separated page numbers to extract tables from.
Upload
Extract Again
Download

A web interface to extract tabular data from PDFs

Previous Uploads

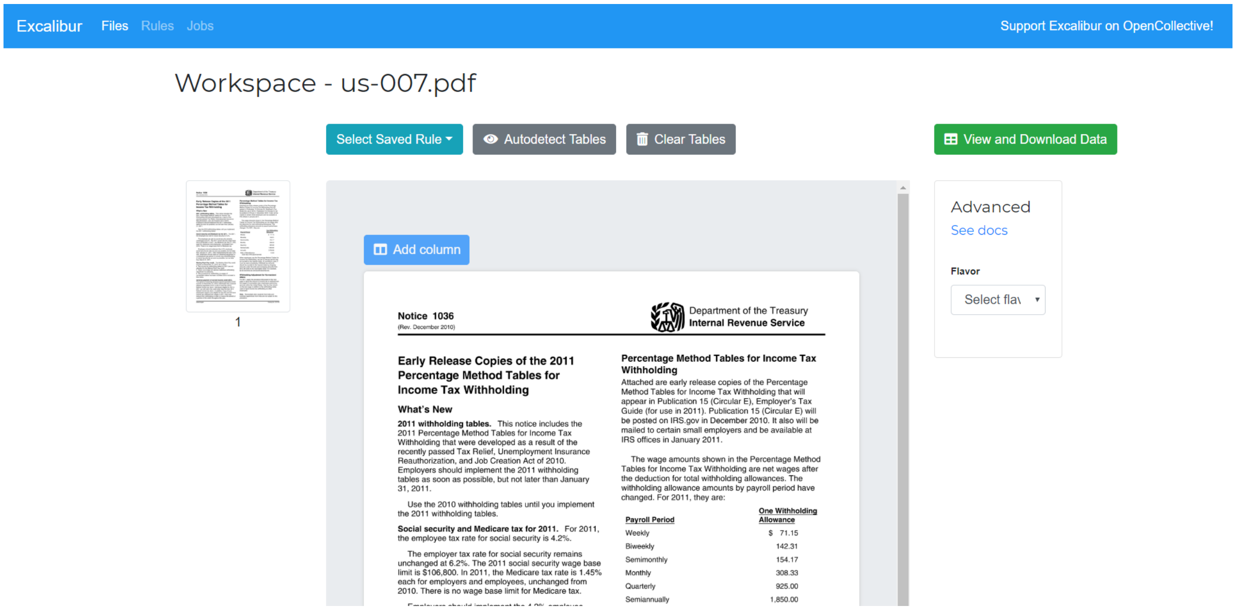
#	Filename	Uploaded at		
1	twotables_2.pdf	2018-11-25T20:13:26	Extract Again	Download

This is where you upload a PDF, select page numbers you want to extract tables from and click on “Upload”.

You can also see previously uploaded PDFs, extract tables from them again by clicking on “Extract Again” or download tables that were extracted last time by clicking on “Download”.

4.4.2 Select table areas and other settings

Clicking on “Upload” or “Extract Again” will take you to a **Workspace**.



Workspace - us-007.pdf

Flavor: Select flav

Notice 1036
(Rev. December 2010)

Department of the Treasury
Internal Revenue Service

Early Release Copies of the 2011 Percentage Method Tables for Income Tax Withholding

What's New

2011 withholding tables. This notice includes the 2011 Percentage Method Tables for Income Tax Withholding that were developed as a result of the recently passed Tax Relief, Unemployment Insurance Reauthorization, and Job Creation Act of 2010. Employers should implement the 2011 withholding tables as soon as possible, but not later than January 31, 2011.

Use the 2010 withholding tables until you implement the 2011 withholding tables.

Social security and Medicare tax for 2011. For 2011, the employee tax rate for social security is 4.2%.

The employer tax rate for social security remains unchanged at 6.2%. The 2011 social security wage base limit is $106,800. In 2011, the Medicare tax rate is 1.45% each for employers and employees, unchanged from 2010. There is no wage base limit for Medicare tax.

Percentage Method Tables for Income Tax Withholding

Attached are early release copies of the Percentage Method Tables for Income Tax Withholding that will appear in Publication 15 (Circular E), Employer's Tax Guide (for use in 2011). Publication 15 (Circular E) will be posted on IRS.gov in December 2010. It also will be mailed to certain small employers and be available at IRS offices in January 2011.

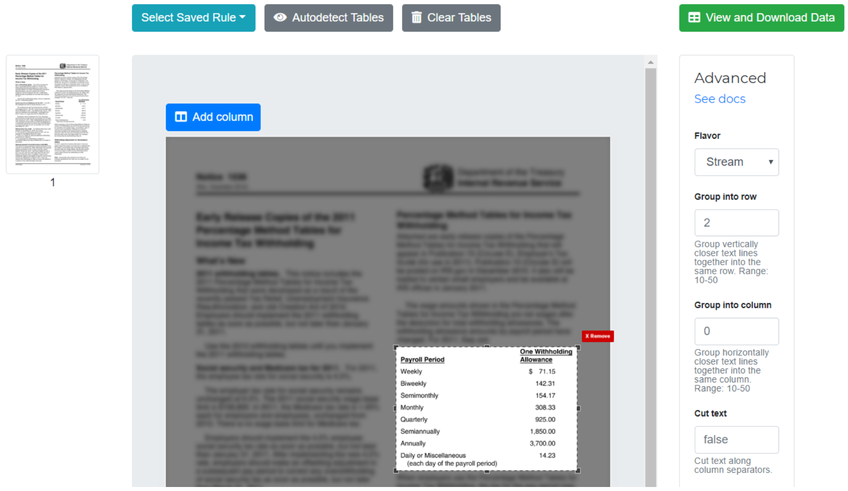
The wage amounts shown in the Percentage Method Tables for Income Tax Withholding are net wages after the deduction for total withholding allowances. The withholding allowance amounts by payroll period have changed. For 2011, they are:

Payroll Period	One Withholding Allowance
Weekly	$ 71.15
Biweekly	142.31
Semimonthly	154.17
Monthly	308.33
Quarterly	925.00
Semiannually	1,850.00

Here you can select table areas by clicking and dragging across a PDF page's image. You can also tweak Camelot's advanced settings. Let's extract the table from this PDF file as an example:

Note: This is a worst-case example where the table is buried deep inside the text. In most cases, you don't need to select table areas, columns or change the advanced settings since Excalibur can do that automatically. You can click on "Autodetect tables" to see what table areas Excalibur detected.

Since the table is formed using whitespaces, you need to select the **Stream** flavor from "**Advanced**". You also need to select a table area in this case for the reason mentioned in the note above.



Flavor: Stream

Group into row: 2

Group vertically closer text lines together into the same row. Range: 10-50

Group into column: 0

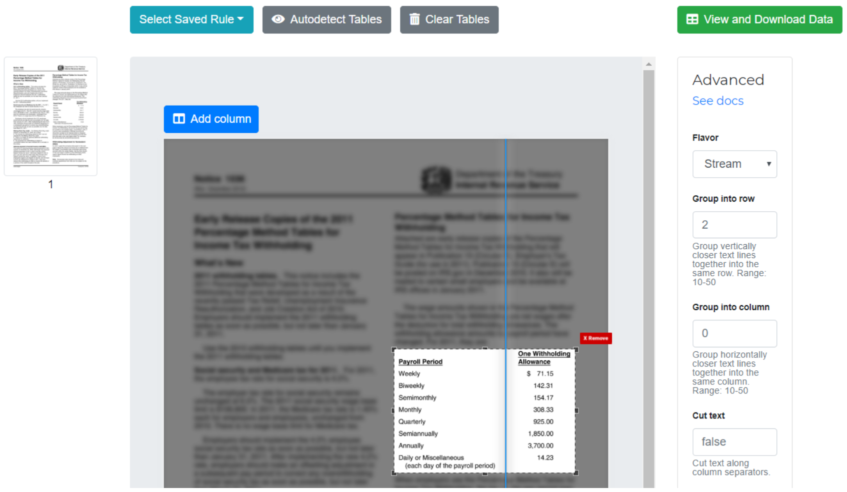
Group horizontally closer text lines together into the same column. Range: 10-50

Cut text: false

Cut text along column separators.

Payroll Period	One Withholding Allowance
Weekly	$ 71.15
Biweekly	142.31
Semimonthly	154.17
Monthly	308.33
Quarterly	925.00
Semiannually	1,850.00
Annually	3,700.00
Daily or Miscellaneous (each day of the payroll period)	14.23

Optionally, you can also add a column on a page by clicking on "Add column" button.



Advanced

See docs

Flavor: Stream

Group into row: 2

Group vertically closer text lines together into the same row. Range: 10-50

Group into column: 0

Group horizontally closer text lines together into the same column. Range: 10-50

Cut text: false

Cut text along column separators.

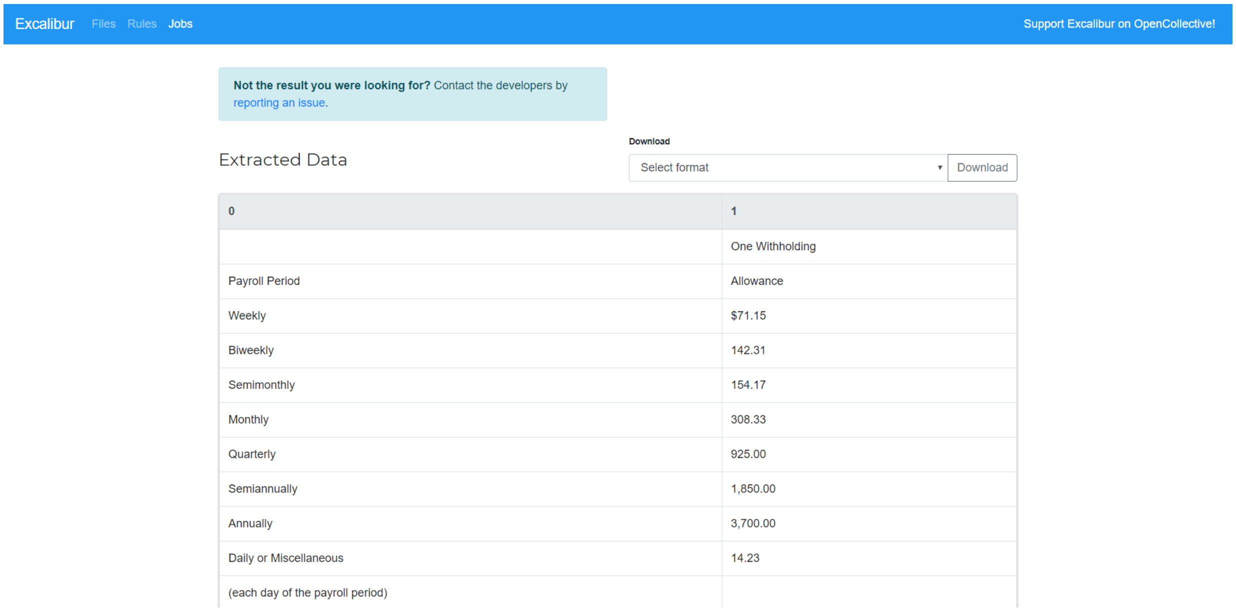
Note: To delete a column, double-click on it.

Note: The Lattice flavor doesn't need columns so the "Add column" button will be disabled when you select that flavor.

Finally, you can click on "View and Download Data". This will save the table areas, columns and advanced settings as a preset which can be used in the future on PDFs with similar table structures.

4.4.3 View and download data

Now you can view all the extracted tables and download them in CSV, Excel, JSON or HTML formats.



Not the result you were looking for? Contact the developers by reporting an issue.

Extracted Data

0	1
	One Withholding
Payroll Period	Allowance
Weekly	$71.15
Biweekly	142.31
Semimonthly	154.17
Monthly	308.33
Quarterly	925.00
Semiannually	1,850.00
Annually	3,700.00
Daily or Miscellaneous	14.23
(each day of the payroll period)	

To know more about how **Lattice** and **Stream** work, check out Camelot's how it works documentation.

4.5 FAQ

This part of the documentation answers some common questions. If you want to add a question (with or without an answer), you can open an issue on GitHub or submit a pull request which updates this file.

4.5.1 What is a workspace?

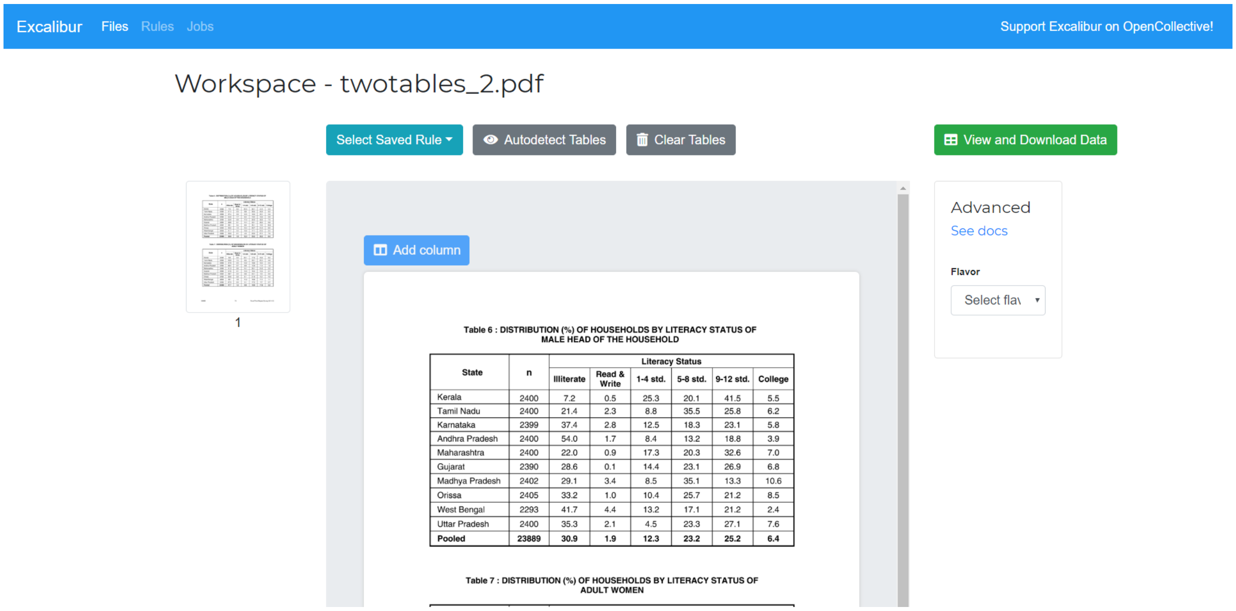


Table 6 : DISTRIBUTION (%) OF HOUSEHOLDS BY LITERACY STATUS OF MALE HEAD OF THE HOUSEHOLD

State	n	Literacy Status					
		Illiterate	Read & Write	1-4 std.	5-8 std.	9-12 std.	College
Kerala	2400	7.2	0.5	25.3	20.1	41.5	5.5
Tamil Nadu	2400	21.4	2.3	8.8	35.5	25.8	6.2
Karnataka	2399	37.4	2.8	12.5	18.3	23.1	5.8
Andhra Pradesh	2400	54.0	1.7	8.4	13.2	18.8	3.9
Maharashtra	2400	22.0	0.9	17.3	20.3	32.6	7.0
Gujarat	2390	28.6	0.1	14.4	23.1	26.9	6.8
Madhya Pradesh	2402	29.1	3.4	8.5	35.1	13.3	10.6
Orissa	2405	33.2	1.0	10.4	25.7	21.2	8.5
West Bengal	2293	41.7	4.4	13.2	17.1	21.2	2.4
Uttar Pradesh	2400	35.3	2.1	4.5	23.3	27.1	7.6
Pooled	**23889**	**30.9**	**1.9**	**12.3**	**23.2**	**25.2**	**6.4**

Table 7 : DISTRIBUTION (%) OF HOUSEHOLDS BY LITERACY STATUS OF ADULT WOMEN

A workspace is the part of the web interface where you can select table areas, add columns and tweak Camelot's advanced settings.

4.5.2 What are rules?

For a PDF document, a set of table areas, columns and Camelot's advanced settings which you select form a table extraction rule.

Excalibur contains a rule manager where you can download saved rules and upload rules that you might have created on other Excalibur installations. To view all existing rules, click on "Rules" in the navigation bar on top.

While on the workspace, you can select a saved rule by clicking on the "Select saved rule" dropdown. After you click on a saved rule, the table areas, columns and Camelot settings contained in that rule will be loaded on the workspace.

4.5.3 What are jobs?

When you click on "View and Download Data" on the workspace, Excalibur starts a table extraction job using the settings you specified. You can view a list of all historical jobs by clicking on "Jobs" in the navigation bar on top.

To view and download the tables extracted in an older job, click on the "Download" button for that job.

CHAPTER 5

The Contributor Guide

If you want to contribute to the project, this part of the documentation is for you.

5.1 Contributor's Guide

If you're reading this, you're probably looking to contributing to Excalibur. *Time is the only real currency*, and the fact that you're considering spending some here is *very* generous of you. Thank you very much!

This document will help you get started with contributing documentation, code, testing and filing issues. If you have any questions, feel free to reach out to Vinayak Mehta, the author and maintainer.

5.1.1 Code Of Conduct

The following quote sums up the **Code Of Conduct**.

> **Be cordial or be on your way**. *–Kenneth Reitz*

Kenneth Reitz has also written an essay on this topic, which you should read.

As the Requests Code Of Conduct states, **all contributions are welcome**, as long as everyone involved is treated with respect.

5.1.2 Your first contribution

A great way to start contributing to Excalibur is to pick an issue tagged with the help wanted or the good first issue tags. If you're unable to find a good first issue, feel free to contact the maintainer.

5.1.3 Setting up a development environment

After installing the dependencies, which include Tkinter and ghostscript, you can install the *dev* extra using pip:

```
$ pip install excalibur-py[dev]
```

Alternatively, you can clone the project repository, and install the *dev* extra using pip:

```
$ pip install -e ".[dev]"
```

5.1.4 Pull Requests

Submit a pull request

The preferred workflow for contributing to Excalibur is to fork the project repository on GitHub, clone, develop on a branch and then finally submit a pull request. Here are the steps:

1. Fork the project repository. Click on the ‘Fork’ button near the top of the page. This creates a copy of the code under your account on the GitHub.
2. Clone your fork of Excalibur from your GitHub account:

   ```
   $ git clone https://www.github.com/[username]/excalibur
   ```

3. Create a branch to hold your changes:

   ```
   $ git checkout -b my-feature
   ```

Always branch out from `master` to work on your contribution. It’s good practice to never work on the `master` branch!

Note: `git stash` is a great way to save the work that you haven’t committed yet, to move between branches.

4. Work on your contribution. Add changed files using `git add` and then `git commit` them:

   ```
   $ git add modified_files
   $ git commit
   ```

5. Finally, push them to your GitHub fork:

   ```
   $ git push -u origin my-feature
   ```

Now it’s time to go to the your fork of Excalibur and create a pull request! You can follow these instructions to do the same.

Work on your pull request

We recommend that your pull request complies with the following guidelines:

- Make sure your code follows pep8.
- In case your pull request contains function docstrings, make sure you follow the numpydoc format. All function docstrings in Excalibur follow this format (soon). Following the format will make sure that the API documentation is generated flawlessly.
- **Make sure your commit messages follow the seven rules of a great git commit message:**
  - Separate subject from body with a blank line

  - Limit the subject line to 50 characters
  - Capitalize the subject line
  - Do not end the subject line with a period
  - Use the imperative mood in the subject line
  - Wrap the body at 72 characters
  - Use the body to explain what and why vs. how

- Please prefix your title of your pull request with [MRG] (Ready for Merge), if the contribution is complete and ready for a detailed review. An incomplete pull request's title should be prefixed with [WIP] (to indicate a work in progress), and changed to [MRG] when it's complete. A good task list in the PR description will ensure that other people get a fair idea of what it proposes to do, which will also increase collaboration.
- If contributing new functionality, make sure that you add a unit test for it, while making sure that all previous tests pass. Excalibur uses pytest for testing (soon). Tests can be run using:

```
$ python setup.py test
```

5.1.5 Writing Documentation

Writing documentation, function docstrings, examples and tutorials is a great way to start contributing to open-source software! The documentation is present inside the `docs/` directory of the source code repository.

The documentation is written in reStructuredText, with Sphinx used to generate these lovely HTML files that you're currently reading (unless you're reading this on GitHub). You can edit the documentation using any text editor and then generate the HTML output by running *make html* in the `docs/` directory.

The function docstrings are written using the numpydoc extension for Sphinx. Make sure you check out how its format guidelines before you start writing one.

5.1.6 Filing Issues

We use GitHub issues to keep track of all issues and pull requests. Before opening an issue (which asks a question or reports a bug), please use GitHub search to look for existing issues (both open and closed) that may be similar.

Questions

Please don't use GitHub issues for support questions. A better place for them would be Stack Overflow. Make sure you tag them using the `python-excalibur` tag.

Bug Reports

In bug reports, make sure you include:

- Your operating system type and Python version number, along with the version numbers of NumPy, OpenCV and Excalibur. You can use the following code snippet to find this information:

```
import platform; print(platform.platform())
import sys; print('Python', sys.version)
import numpy; print('NumPy', numpy.__version__)
import cv2; print('OpenCV', cv2.__version__)
import excalibur; print('Excalibur', excalibur.__version__)
```

- The complete webserver traceback log. Just adding the exception message or a part of the traceback won't help us fix your issue sooner.
- Steps to reproduce the bug.
- A link to the PDF document that you were trying to extract tables from, telling us what you expected the interface to do and what actually happened.