



Project Figures

Jonathan Reynosa & Sa Hyun Min

CS 4624 Multimedia, Hypertext, and Information Access
Dr. Edward A. Fox
Virginia Tech, Blacksburg VA, 24061
5/11/2022

Table of Contents

Problem and Motivation

Deliverables and Requirements

Project Timeline

Methodology

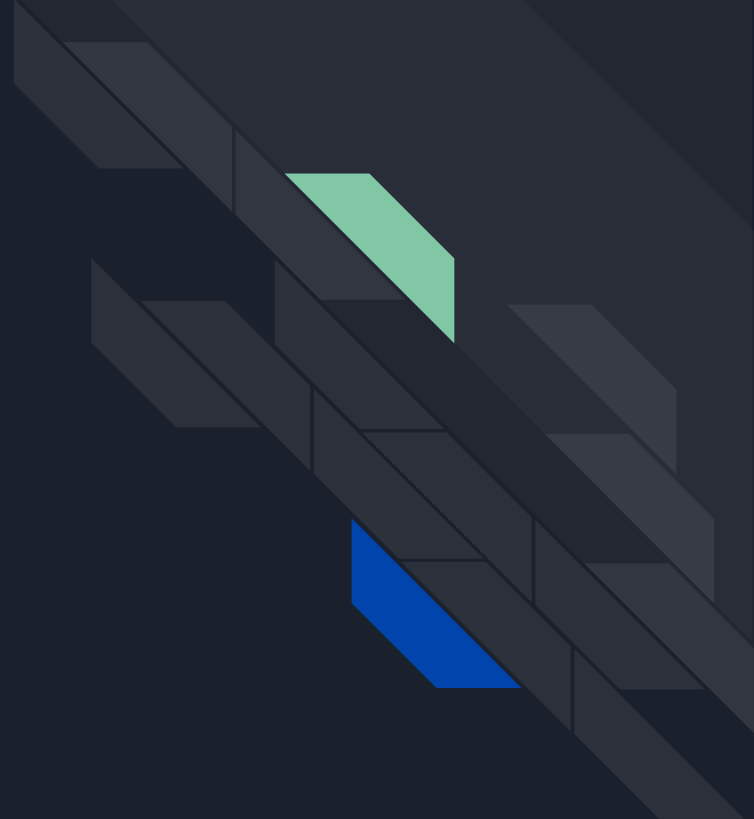
Final Design

Lessons Learned

Future Work

Acknowledgements

References





Problem and Motivation

- Problem
 - Need a way to search through various pdfs and return matching images/graphs within the ETDs.
- Motivation
 - Electronic theses and dissertations (ETDs) contain a rich amount of figures used for demonstrating scholarly results and observations
 - Challenge to effectively search for figures existing in a large number of research documents stored in PDF files



Deliverables and Requirements

- Deliverables
 - 1000 processed ETDs and their respective figures, captions, and document metadata.
 - Web-Based Interface built upon ODU's current user interface
- Requirements
 - Index figure data making use of Elasticsearch
 - Allow users to upload PDF files that would be searched
 - Build a web-based interface
 - Accepts text queries
 - Returns the appropriate figures and metadata

Project Timeline



Create
Foundation

Improve
Foundation

Make
Prototype

Improve
Prototype

Debug

Finish
Project

Front End
>>

Get server
running

Minor
changes to
front end

Process
one PDF
file

Process
1000 PDF
files

Troubleshoot
and Improve
Aesthetics

Finish
Project

Back End
>>

Explore
Framework

Process and
Extract
Figures

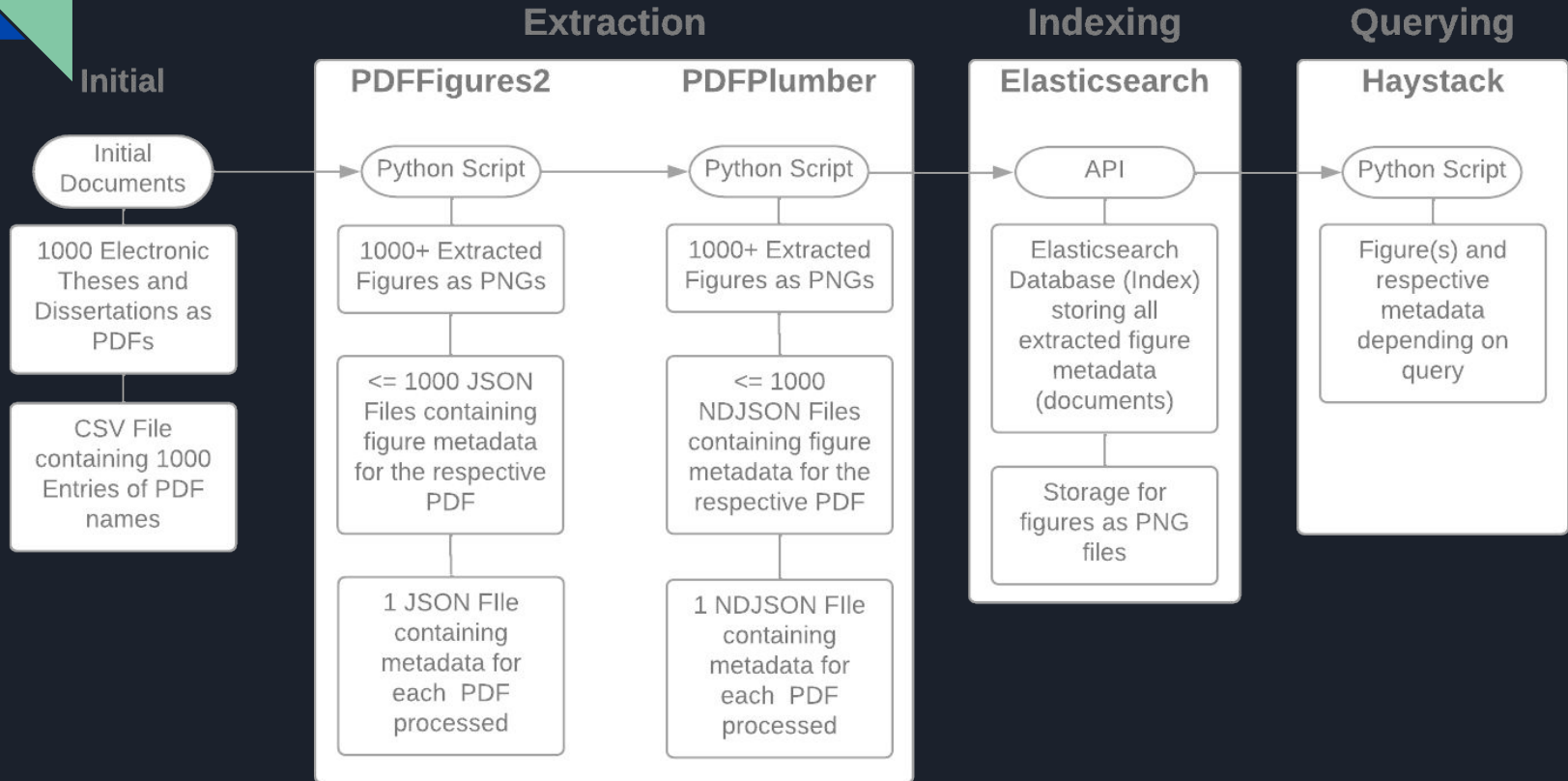
Index
figures and
data

Combine
Front End
and Back
End

Implement
Haystack and
improve
searching

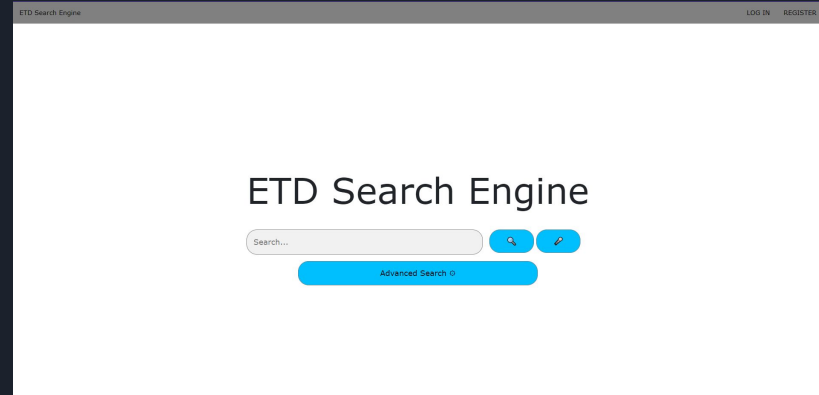
Finish
Project

Methodology - Back End



Methodology - Front End

- Front End
 - Running on localhost.
 - Written in PHP, HTML, and Javascript
 - Based on ETDUI
 - Features
 - Allow search for words in figure captions and image text
 - Allow users to upload PDFs



Final Design

Figure Search Engine

Search... 

Upload Files
Select files to upload:

Main page

>Allows users to upload PDFs

>uploaded PDFs would be searched

File Upload Page

>Allows users to see what PDFs they uploaded

>Also can see the name of the file and where the file is stored.

```
filename -> MATHHW8.pdf  
filetempName -> /tmp/phpFT4Fgt  
destFile -> /var/www/html/figures/src/figure_extraction/pdf_files/MATH HW8.pdf MATHHW8.pdf successfully uploaded
```

```
filename -> 20220419132300960.pdf  
filetempName -> /tmp/phpCxCvT  
destFile -> /var/www/html/figures/src/figure_extraction/pdf_files/20220419132300960.pdf 20220419132300960.pdf successfully uploaded
```

```
filename -> CkTest3.pdf  
filetempName -> /tmp/phpWCo85r  
destFile -> /var/www/html/figures/src/figure_extraction/pdf_files/Ck Test3.pdf CkTest3.pdf successfully uploaded
```

```
filename -> Unsafe.pdf  
filetempName -> /tmp/php1FKUBs  
destFile -> /var/www/html/figures/src/figure_extraction/pdf_files/Unsafe.pdf Unsafe.pdf successfully uploaded
```

```
filename -> 3342102.pdf  
filetempName -> /tmp/phpzqYmxs  
destFile -> /var/www/html/figures/src/figure_extraction/pdf_files/3342102.pdf 3342102.pdf successfully uploaded
```

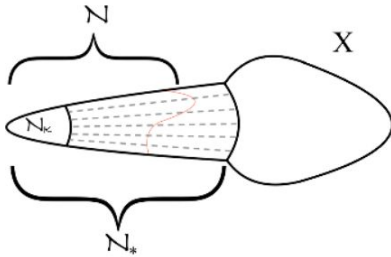
Final Design

Figure Search Engine

Figure



4 search results for Figure



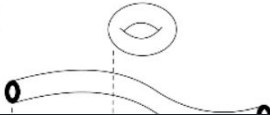
Caption(s): Figure 1.1: Hypersurface defined by FA in red

Figure Type: Figure

Image Text:

Download

$S^1 \times M$



Result page
>Figures are shown
>Search for keyword in caption /
words within figures.



Lessons Learned

- Front End
 - Running the base website.
 - Readme.txt directions vague.
 - Different Front End
 - Could have created our own original
- Back End
 - Project Scale
 - Learn various frameworks
 - Learn to work with various languages
 - Communication
 - Reaching out helped us in doing work faster



Future Work

- Front End
 - Run on actual website, not on localhost.
- Back End
 - Image Text
 - Extract text from images in a better manner
 - Add another framework to pipeline
 - Haystack
 - Increase precision and accuracy of searches



Acknowledgements



William Ingram

Assistant Dean and IT
Services Director
VT University Libraries



Jian Wu

Assistant Professor in
the CS Department
Old Dominion University



References

<https://canvas.vt.edu/courses/145290/pages/s2022project-figureextractionwebservice>

<https://canvas.vt.edu/courses/145290/pages/s2022project-figureextractionwebservice>

<https://opening-etds.github.io/>

<https://github.com/allenai/pdffigures2>

<https://www.elastic.co/what-is/elasticsearch>

<https://experts.vt.edu/5675-william-a-ingram>

<https://www.odu.edu/directory/people/j/j1wu>