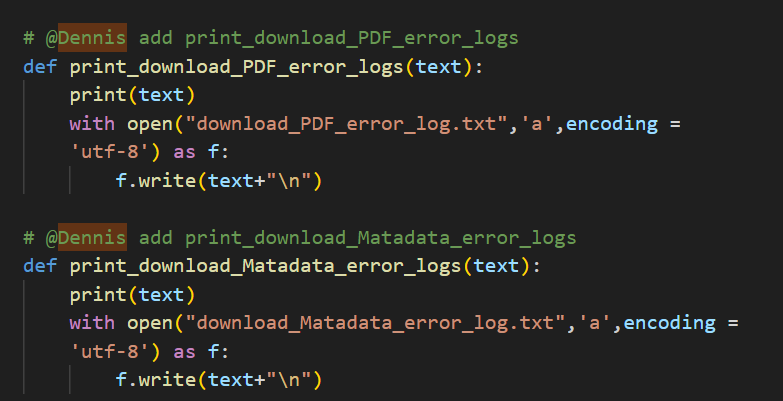
This document is a record of specific modifications, including:

1. Any annotations in the document that contain **'@Dennis'** are the code changes Dennis made. Looked like this:



1. All modifications for the school shifters are detailed in this document using Auburn as an example. Please refer to Auburn.

A diagram of a process

Description automatically generated

**Auburn: finished**

Use ETD OAI-PMH

https://etd.auburn.edu/oai/request?verb=ListSets

sitemap: https://etd.auburn.edu//sitemap?map=0

copyright: <https://graduate.auburn.edu/current-students/auetd-info/>

Modify ideas:

1. The AirForce etdrepo is: ETDMiner/webcrawler/etdrepo,

But the Auburn etdrepo is: ETDMiner/etdrepo

**Standardizing here, all future etdrepo paths will be “ETDMiner/etdrepo”.**

1. Copyright: embargo, is different with other universities.
2. Add new fields to the database table **etds**: haspdf, timestamp\_metadata, timestamp\_pdf.

**Add new code to extract the haspdf, timestamp\_metadata, and timestamp\_pdf data.**

1. Add new function: exists\_in\_etds(soup), used to check if this record is in the database etds, if yes, return etdid, or return None.
2. Add function: empty\_fields(soup,etdid), used to return a empty\_fields list
3. Add function: update\_empty\_field(soup,empty\_list,etdid), Used to update empty fields in the database etds based on newly downloaded
4. Add function: extract\_all\_field(soup), used to extract all field from matadata, and return a JSON.
5. Change author and advisor to list, used to deal with multiple people.

A screen shot of a computer program

Description automatically generated

1. Add copyright and identifier:

A screen shot of a computer code

Description automatically generated

Degree : None

Department: None

Discipline: None

Second\_identifier: None

 <dim:field mdschema="dc" element="rights" lang="en\_US">EMBARGO\_GLOBAL</dim:field>

  <dim:field mdschema="dc" element="subject" lang="en\_US">Interdepartmental Pharmacy</dim:field>

  <dim:field mdschema="dc" element="title" lang="en\_US">Potentially Inappropriate Medication Use and Access to Medication Management Services in Older Adults Living with Alzheimer's Disease and Related Dementia</dim:field>

  <dim:field mdschema="dc" element="type" lang="en\_US">PhD Dissertation</dim:field>

  <dim:field mdschema="dc" element="embargo" qualifier="length" lang="en\_US">MONTHS\_WITHHELD:24</dim:field>

  <dim:field mdschema="dc" element="embargo" qualifier="status" lang="en\_US">EMBARGOED</dim:field>

  <dim:field mdschema="dc" element="embargo" qualifier="enddate" lang="en\_US">2025-05-03</dim:field>

  <dim:field mdschema="dc" element="creator" qualifier="orcid" lang="en\_US">0000-0002-1348-7968</dim:field>

</dim:dim>

1. Edit function insertETDs(soup)
2. Add function: final\_pdf\_dir(etdid), "Used to return the PDF address within the final etdrepo folder.
3. Add function: final\_xml\_dir(etdid), "Used to return the xml address within the final etdrepo folder.
4. Add function: insert\_haspdf\_timestamp (final\_dir,etdid), used to insert haspdf and timestamp\_pdf
5. Add function: insert\_metadata\_timestamp(final\_dir,etdid)
6. Add function: etdrepo\_check(final\_dir,etdid), used to check if the pdf/xml exist in the etdrepo
7. Add movehtml() function:

A screen shot of a computer code

Description automatically generated

Finished crawling, test shifter, this is the result:

A screen shot of a computer

Description automatically generated

**AirForceInstitute: finished**

Crawling: Using sitemap to download.

Difference with other universities:

1. Uri is under “Recommended Citation”, not under the “uri”

A screenshot of a computer

Description automatically generated

##### Every time you run the script, it will overwrite the existing PDF files.

##### Line142( in function extractPDF() ):

if downloadableUrl == None or os.path.exists(filepath):

        pass

        print(filepath," aleady exist, pass")

##### Line190( in function saveHTML() ):

if not os.path.exists(fileName):

    #print(response.decode('utf-8'))

        with open(filepath, 'wb') as file: # wb for override

            file.write(response)

(1) crawling:

Add print download error log function:

A computer screen with text and symbols

Description automatically generated with medium confidence

If the pdf already exists, do not re-download. (check the filepath not the database, leaving all the database checking and editing to the shifter)

A screen shot of a computer code

Description automatically generated

Handle the pdf download error, and print the error log:

A screen shot of a computer program

Description automatically generated

If metadata already exists, do not re-download. Handle the metadata download error, and print the error log:

A screen shot of a computer program

Description automatically generated

(2) shifter:

Extract **copyright**:

A screen shot of a computer program

Description automatically generated

Extract **AFIT** and **DTIC**:

A screen shot of a computer code

Description automatically generated

Extract **haspdf and timestamp**:

When the shifter run, the pdf and metadata will be renamed, and be moved to a new directory, so in the server, the original still exists or not? So when we check if haspdf and timestamp, we should check the final directory.

The etdId will create after insertETDs(), and we need etdId to find the final directory.

The original code just check if the downloadUrl exists:

Yes, insert the localrelpath and etdId, url to the table pdfs;

No, insert the localrelpath and etdId, url(is None) to the table pdfs;

Create a new function **insert\_haspdf\_timestamp**(pdf\_path,etdid):

A screen shot of a computer program

Description automatically generated

Create a new function **final\_dir(etdid)**:

A screen shot of a computer code

Description automatically generated

Finished crawling, test shifter, this is the result:

A screenshot of a computer program

Description automatically generated

**BallState: just xml, please check the crawler, the pdf should can be downloaded.**

sitemap: https://cardinalscholar.bsu.edu/sitemap0.xml

ListSets: https://cardinalscholar.bsu.edu/server/oai/request?verb=ListSets

Ball State Theses and Dissertations [**com\_20.500.14291\_7**] or [[col\_20.500.14291\_13](https://cardinalscholar.bsu.edu/server/oai/request?verb=ListRecords&metadataPrefix=oai_dc&set=col_20.500.14291_13)]

#@Dennis records = sickle.ListRecords(metadataPrefix='dim', set='com\_123456789\_7')

records = sickle.ListRecords(metadataPrefix='dim', set='col\_20.500.14291\_13')

this college just has metatdata, no pdf.

crawl\_delay = 5， so download speed is very slow, and almost 30 mins will show “HTTPError: 429 Client Error: Too Many Requests”, so crawl is very difficult.

A computer screen with text and numbers

Description automatically generated

The directory path does not exist.

When I tried to extract copyright(lisence.txt), the website git a failed status code:429 (Too Many Requests), and then I checked the robots.txt, Crawl-delay: 10, so I need to set time.sleep(15), and the shifter will be very slow, and sometime it still shows error 429.

Try: 30s still not work

Document:

A computer screen with white text

Description automatically generated

A screenshot of a computer program

Description automatically generated

And maybe my ip is banned by the website. I try a lot of times, the total count still 700.

**Boston University: disallow crawlers curl https://open.bu.edu/oai/request?verb=ListSets**

https://open.bu.edu/oai/request?verb=ListSets

Theses & Dissertations [com\_2144\_898]

requests.exceptions.HTTPError: 403 Client Error: Forbidden for url: https://open.bu.edu/oai/request?metadataPrefix=dim&set=com\_2144\_899&verb=ListRecords

from sitemap:

# Crawlers that are kind enough to obey, but which we'd rather not have

# unless they're feeding search engines.

User-agent: UbiCrawler

Disallow: /

**Calthch: finished**

use rdf to get all xml: https://thesis.library.caltech.edu/cgi/export/repository/RDFXML/caltechthesis.rdf

dump all the url to urls.txt

1. Calthch has no shifter.py, so I create a new shifter.py for this university
2. In the metadata, there have 2 types of meatdata: eprints and DC.

Worked good right now!

A screenshot of a computer

Description automatically generated

1 pdf can not download, checked the original website page, shows need login to download.

Copyright : rights and copyright\_statement ‘||’

Check all the copyright, if combined with 2 or more, use ‘||’

Too many etds, the for loop just take 9999, so split the etds folder to 2 parts, and change the shifter, let it do 2 for loops to finish all.

**ColoradoWyoming**: the OAI not work

404 page not found <https://mountainscholar.org/oai/request?verb=ListSets>

**Cornell:**

404 page not found <https://ecommons.cornell.edu/oai/request?verb=ListSets>

New sickle site: https://ecommons.cornell.edu/server/oai/request?verb=ListSets

Harvest.py does not work. Re-write harvest.py

A screen shot of a computer

Description automatically generated

A screen shot of a computer screen

Description automatically generated

A screen shot of a computer

Description automatically generated

Right now work well

Build a download\_PDF\_error\_log.txt

Re-write shifter.py

A screen shot of a computer

Description automatically generated

**Duke: finished**

<https://dukespace.lib.duke.edu/oai/request?verb=ListSets> set='com\_10161\_1'

new sickle site:

https://dukespace.lib.duke.edu/server/oai/request?verb=ListSets

##### Dissertations [col\_10161\_4]

##### Masters Theses [col\_10161\_2493]

##### sickle = Sickle('https://dukespace.lib.duke.edu/server/oai/request')

##### records\_1 = sickle.ListRecords(metadataPrefix='dim', set='col\_10161\_4')

##### records\_2 = sickle.ListRecords(metadataPrefix='dim', set='col\_10161\_2493')

##### records = list(records\_1) + list(records\_2)

havrvest.py just download xml, no pdf. Edit harvest, make it download pdf.

Worked well right now

A screenshot of a computer program

Description automatically generated

**EastCarolina: finished**

<https://thescholarship.ecu.edu/oai/request?verb=ListSets>

##### Dissertations and Theses [com\_10342\_122]

##### Note：some pdfs are locked.

##### A screenshot of a computer program Description automatically generated

**EasternMichigan: finished**

##### Use sitemap

##### https://commons.emich.edu/theses/

total etds: 1186

Eastern Michigan University theses and dissertations are protected by copyright. **They may be viewed or downloaded from this site for the purposes of research and scholarship. Reproduction or distribution for commercial purposes is prohibited without written permission of the author.**

A screen shot of a computer

Description automatically generated

**GeorgeMason**: requests exceptions error

<http://mars.gmu.edu/oai/request?verb=ListSets>

##### etd @ Mason (Electronic Theses and Dissertations) [com\_1920\_2811]

##### new list of sets

https://mars.gmu.edu/server/oai/request?verb=ListSets

##### requests.exceptions.HTTPError: 404 Client Error: Not Found for url: <http://mars.gmu.edu/server/oai/request?metadataPrefix=dim&set=com_1920_2811&verb=ListRecords>

##### curl http://mars.gmu.edu/server/oai/request?metadataPrefix=dim&set=com\_1920\_2811&verb=ListRecords

##### [1] 963180

##### [2] 963181

##### [2]+ Done set=com\_1920\_2811

##### [dennis@DennisPC2 GeorgeMason (master)]$ <!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML 2.0//EN">

##### <html><head>

##### <title>403 Forbidden</title>

##### </head><body>

##### <h1>Forbidden</h1>

##### <p>You don't have permission to access this resource.</p>

##### </body></html>

##### try couple dates later.

**Gtech: not finished**

##### The old sitemap not work: https://smartech.gatech.edu/sitemap?map=0

##### Update the sitemap:

##### <https://repository.gatech.edu/sitemap0.xml>

##### <https://repository.gatech.edu/sitemap1.xml>

##### has 2 parsers, updated first parsers, create 2 txt

##### the second parser will use the 2 txts to parser, but the isItemThesis has some problem.

##### Re-write the isItemThesis() function:

##### A screen shot of a computer code Description automatically generated

Update the **savexml** function to prevent it from redownloading existing XML files.

check which URLs have been processed each time it runs and skip them. After processing each URL, it adds the URL to a set of processed URLs and saves this set to a file for the next run.

A screen shot of a computer program

Description automatically generated

##### All the pdf are locked, can not be downloaded. Need to login the library.

##### Add parser3.py to deal with url--01.txt

**Illinois: finished**

##### The sickle = Sickle('https://www.ideals.illinois.edu/dspace-oai/request?') not work

##### Find the new OAI-PMH: <https://www.ideals.illinois.edu/oai-pmh>?verb=ListSets

##### Dissertations and Theses - Urban and Regional Planning [col\_2142\_5128] 161

[Records](https://www.ideals.illinois.edu/oai-pmh?verb=ListRecords&metadataPrefix=oai_dc&set=col_2142_5128)[Identifiers](https://www.ideals.illinois.edu/oai-pmh?verb=ListIdentifiers&metadataPrefix=oai_dc&set=col_2142_5128)

##### Graduate Dissertations and Theses at Illinois [com\_2142\_5130] 52064

[Records](https://www.ideals.illinois.edu/oai-pmh?verb=ListRecords&metadataPrefix=oai_dc&set=com_2142_5130)[Identifiers](https://www.ideals.illinois.edu/oai-pmh?verb=ListIdentifiers&metadataPrefix=oai_dc&set=com_2142_5130)

##### Graduate Dissertations and Theses at Illinois [col\_2142\_5131] 52064

[Records](https://www.ideals.illinois.edu/oai-pmh?verb=ListRecords&metadataPrefix=oai_dc&set=col_2142_5131)[Identifiers](https://www.ideals.illinois.edu/oai-pmh?verb=ListIdentifiers&metadataPrefix=oai_dc&set=col_2142_5131)

##### the metadataPrefix = ‘dim’ does not work, so use oai\_dc, and change all the dim to oai\_dc in the harvest.

files = tree.xpath('//oai\_dc:dc/oai\_dc:identifier', namespaces={'oai\_dc': 'http://www.openarchives.org/OAI/2.0/oai\_dc/'})

##### oai-phm can not get the pdf, so try to use sitemap to download the pdf.

for metadata in metadatas:

        filename = identifier.split('/')[-1].lower() + '.xml'

        (p / filename).open('wb').write(etree.tostring(metadata, pretty\_print=True))

        # @Dennis re-write the pdf download logic

        url = "http://hdl.handle.net/" + identifier

        print('download pdf Now on:',url)

        filename = download\_file\_stream(url, p, crawl\_delay=crawl\_delay)

        if filename is None:

            print(f'There was a problem downloading {url}')

##### getPDFdownloadUrl(soup) does not work, re\_write this function:

def getPDFdownloadUrl(soup):

    hrefValue = None

    # findClassess = soup.find\_all('tr', {'class':'ds-table-row odd'})

    # @Dennis

    findClassess = soup.find\_all('div', {'class':['dropdown-menu show','dropdown-menu']})

    if findClassess:

        for findClass in findClassess:

            # if findClass:

            #     imgTag = findClass.find('img')

            #     if imgTag['title'] == "application/pdf":

            #         anchortag = findClass.find('a')

            #         if anchortag:

            #             print(anchortag)

            #             hrefValue = anchortag['href']

            #             hrefValue = "https://www.ideals.illinois.edu/" + hrefValue

            #             # Check if there is download permission

            #             if 'isAllowed=n' in hrefValue:

            #                 hrefValue = None

            #             break

            # @Dennis

            if findClass:

                a\_element = findClass.find('a', class\_='dropdown-item')

                if a\_element and a\_element.find('i', {'clsss':'fa fa-download'}):

                    hrefValue = a\_element['href']

                    hrefValue = "https://www.ideals.illinois.edu/" + hrefValue

                    # Check if there is download permission

                    if 'isAllowed=n' in hrefValue:

                        hrefValue = None

                        break

    return hrefValue

##### when I use BeautifulSoup to find the download\_url, the content is loaded dynamically after the initial page load. So I may need to use a tool like Selenium to interact with the page dynamically, as BeautifulSoup can only parse the initial HTML content. For the time limited, I will not download pdf this time.

##### A screenshot of a computer program Description automatically generated

##### IndianaUniBloomington:

##### No harvest

##### No shifter

##### JohnHopkins: crawler does not work

##### <https://jscholarship.library.jhu.edu/oai/> does not work

##### new site: <https://jscholarship.library.jhu.edu/server/oai/request?verb=ListSets>

#### **Theses and Dissertations, Electronic (ETDs) [com\_1774.2\_836] 2619**

##### The original code download all the xml, no pdf.

**KansasStateU: finished**

##### <https://krex.k-state.edu/dspace-oai/request>

##### records = sickle.ListRecords(metadataPrefix='dim', set='col\_2097\_4234')

##### the original record just download “Prior to 2004 “ 11577

##### K-State Electronic Theses, Dissertations, and Reports, Prior to 2004 [col\_2097\_4234]

##### Add the set:

##### K-State Electronic Theses, Dissertations, and Reports: 2004 - [col\_2097\_4] 7971

##### A screenshot of a computer program Description automatically generated

**MichiganTech: finished**

##### A screenshot of a computer Description automatically generated

##### NCStateU: harvesting right now

##### <https://repository.lib.ncsu.edu/oai/request> does not work

##### find the new one: <https://repository.lib.ncsu.edu/server/oai/request>

##### # records = sickle.ListRecords(metadataPrefix='dim', set='col\_1840.20\_25')

##### records = sickle.ListRecords(metadataPrefix='dim', set='com\_1840.20\_23') 17327

##### update pdf download

##### A computer screen shot of a program Description automatically generated

##### No shifter, write a new shifter

##### Naval Postgraduate School: Forbidden crawling

##### <https://calhoun.nps.edu/oai/request?verb=ListSets> worked

#### **Theses and Dissertations [com\_10945\_16] 35160**

##### requests.exceptions.HTTPError: 403 Client Error: Forbidden for url: https://calhoun.nps.edu/oai/request?metadataPrefix=dim&set=com\_10945\_16&verb=ListRecords

**OklahomaUniversities : finished crawling, and test shifter, and all the etds have been copied to hawking. Next step is to run shifter on hawking.**

##### Worked

##### https://shareok.org/oai/request

#### **OSU - Electronic Theses and Dissertations [com\_11244\_10460] 27783**

A screen shot of a computer

Description automatically generated

**Princeton: finished crawling, and test shifter, and all the etds have been copied to hawking. Next step is to run shifter on hawking.**

##### <https://dataspace.princeton.edu/oai/request>

#### **Princeton University Undergraduate Senior Theses, 1924-2023 [com\_88435\_dsp019c67wm88m] 74472**

##### Princeton University Doctoral Dissertations, 2011-2024 [com\_88435\_dsp01td96k251d] 4741

##### Harvest just get the xml, the pdf can not download. Some pdf can be downloaded, work it later.

##### A screenshot of a computer Description automatically generated

##### A computer screen shot of a number Description automatically generated

**RhodeIsland: finished crawling, and test shifter, and all the etds have been copied to hawking. Next step is to run shifter on hawking.**

##### The pdf can be downloaded by the web browser, but forbidden crawler

##### Download metadata first

##### 

##### Rice:

##### OAI does not work, need sitemap

**TexasA&M: finished crawling, and test shifter, and all the etds have been copied to hawking. Next step is to run shifter on hawking.**

##### Add new sets:

##### <https://oaktrust.library.tamu.edu/dspace-oai/request>

#### **Electronic Theses, Dissertations, and Records of Study (2002– ) [col\_1969.1\_2] 24564**

#### **Digitized Theses and Dissertations (1922–2004) [col\_1969.1\_129207] 25022**

##### Downloaded some pdfs, and then Permission denied: '/hdl.handle.net', so just download xml.

##### A screenshot of a computer program Description automatically generated

##### TexasChristianU:

##### No harvest, no shifter

##### OAI does not work

**TexasChristianUni：finished crawling, and test shifter, and all the etds have been copied to hawking. Next step is to run shifter on hawking.**

##### <https://repository.tcu.edu/oai/request>?

#### **Masters Theses [col\_116099117\_50528] 4150**

#### **Doctoral Dissertations [col\_116099117\_50527] 1486**

##### A screenshot of a computer program Description automatically generated

**TexasState：finished crawling, and test shifter, and all the etds have been copied to hawking. Next step is to run shifter on hawking.**

##### <https://digital.library.txst.edu/server/oai/request>

#### **Theses and Dissertations, Capstones, and Directed Research [com\_10877\_134] 7004**

##### A screen shot of a computer program Description automatically generated

##### TexasTech：crawling right now

##### <https://ttu-ir.tdl.org/server/oai/request>

#### **Electronic Theses and Dissertations [col\_2346\_521] 24669**

##### urllib3.exceptions.MaxRetryError: HTTPSConnectionPool(host='ttu-ir.tdl.org', port=443): Max retries exceeded with url: /server/oai/request?resumptionToken=dim%2F%2F%2Fcol\_2346\_521%2F1700&verb=ListRecords (Caused by NameResolutionError("<urllib3.connection.HTTPSConnection object at 0x7f8beb9e3970>: Failed to resolve 'ttu-ir.tdl.org' ([Errno -3] Temporary failure in name resolution)"))

##### uAlabama：

##### https://ir.ua.edu/oai/request

##### Not work

##### UC System：crawling right now

##### <https://escholarship.org/siteMapIndex.xml>

##### UGeorgia:

##### <https://athenaeum.libs.uga.edu/oai/request>? Not work

**uHawaii\_at\_Monoa: finished crawling, Next step is testing shifter, and transfer all the etds to hawking, and then run shifter on hawking.**

##### Anthropology Dissertations & Theses [com\_10125\_8002] 92

##### A screen shot of a computer Description automatically generated

##### uHouston:crawling right now

##### <https://uh-ir.tdl.org/oai/request?verb=ListSets> not work,change to :

##### <https://uh-ir.tdl.org/server/oai/request?verb=ListSets>

#### **Digitized Theses and Dissertations (1940 - 2009) [com\_10657\_3932] 4884**

#### Electronic Theses and Dissertations (2010 - Present) [com\_10657\_1] **5668**

**UKansas: finished crawling, Next step is testing shifter, and transfer all the etds to hawking, and then run shifter on hawking.**

##### <https://kuscholarworks.ku.edu/oai/request>?

#### **Dissertations and Theses [com\_1808\_1260] 9565**

**uMaryland: finished crawling, Next step is testing shifter, and transfer all the etds to hawking, and then run shifter on hawking.**

##### https://drum.lib.umd.edu/oai/request

#### Theses and Dissertations from UMD [com\_1903\_2] **18747**

##### No pdf

**UMichigan: finished crawling, Next step is testing shifter, and transfer all the etds to hawking, and then run shifter on hawking.**

##### <https://deepblue.lib.umich.edu/dspace-oai/request?verb=ListSets>

#### **Dissertations and Theses (Ph.D. and Master's) [col\_2027.42\_39366] 51894**

##### Pdf need to login

##### uMissouri\_at\_KansasCity：crawler does not work.

##### <https://mospace.umsystem.edu/oai/request>

#### **Theses and Dissertations (MU) [com\_10355\_3987] 11133**

##### Other universities not start yet:

##### uNevadaLasVegas

##### University of Kansas

##### uOregon

##### uTexas\_at\_Austin

##### Vanderbilt

##### Vermont

##### Wisconsin-Milwaukee