Towards Aiding Research by Improving Access to Electronic Theses and Dissertations from Multiple Domains

CNI Fall 2021 Virtual Membership Meeting

Jian Wu (Old Dominion University Computer Science),
William A. Ingram (Virginia Tech University Libraries), Edward A. Fox (Virginia Tech Computer Science)

Opening Books and the National Corpus of Graduate Research

- IMLS National Leadership Grants for Libraries
- https://www.imls.gov/grants/awarded/lg-37-19-0078-19
- Investigating innovative ways machine learning and natural language processing can be applied to the national corpus of electronic theses and dissertations in order to extract knowledge, bibliographic and scientific data, and facilitate its identification, discovery, and reuse.
- Research Areas:
  1. Document analysis, information extraction
  2. Adding value through automatic classification and summarization
  3. User services – building better digital libraries

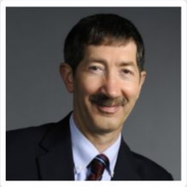
ETD Research Team at VT and ODU

Bill Ingram

Principal Investigator

Assistant Dean and Director of IT, University Libraries, Virginia Tech

Dr. Edward A. Fox



Co-PI

Professor, Computer Science, Virginia Tech

Dr. Jian Wu

Co-PI

Assistant Professor, Computer Science, Old Dominion University

Bipasha Banerjee

Gradudate Assistant

Ph.D. Candidate, Computer Science, Virginia Tech

Muntabir Choudhury

Gradudate Research Assistant

Ph.D. Student, Computer Science, Old Dominion University

And several notable students past and present (see some on next slide):

ODU: Himarsha Jayanetti, Md Sami Uddin, Lamia Salsabil, Neel C. Kawitkar, Richard Pates, Pooja Sonmale
VT: Adheesh Sunil Juvekar, Eman Abdelrahman, Fatimah Alotaibi, Palakh Mignonne Jude, Sampanna Kahu, John Aromando, Gunnar Reiske

Other Notable Students

Sami Uddin

Graduate Research Assistant

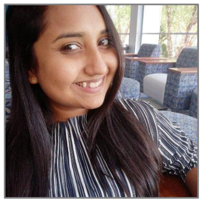
Master's student, Computer Science, Old Dominion University

Winston Shields

Graduate Research Assistant

Master's student, Computer Science, Old Dominion University

Himarsha Jayanetti



Graduate Research Assistant

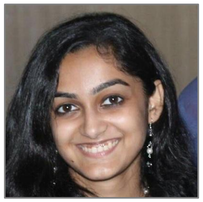
PhD student, Computer Science, Old Dominion University

Sampanna Kahu

Graduate Research Assistant

Master's student, Computer Engineering, Virginia Tech

Palakh Mignonne Jude



Graduate Research Assistant

Master's student, Computer Science, Virginia Tech

Eman Abdelrahman

Graduate Research Assistant

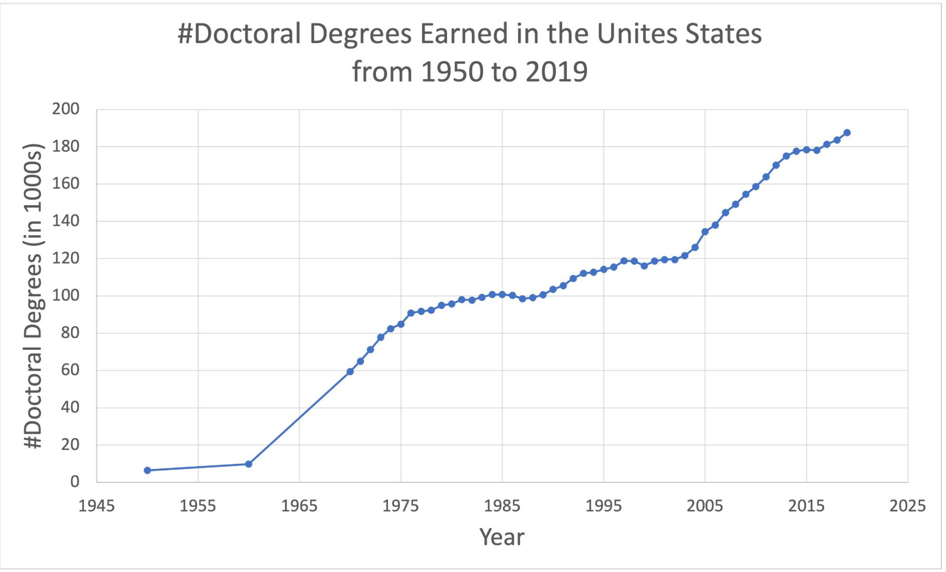
Master's student, Computer Science, Virginia Tech

Outline

- Introduction
- Data Acquisition and Characteristics
- Repository, Searching, Services
- Language Model, Evaluations
- Results
- Conclusions, Future Work

Introduction

- Trend of increasing numbers of doctoral degrees in the United States
- Machine Learning (ML) and Deep Learning (DL): often data-driven
  - Pre-trained models in Computer Vision (CV), Natural Language Processing (NLP)
  - CV: VGG16, VGG19, ResNet-100, etc.: ImageNet (14 million), MS COCO (330k)
  - NLP: BERT, XLNet, etc.: BooksCorpus (800M), English Wikipedia (2500M), ClueWeb, Common Crawl
    - Self-supervised training: unlabeled data



Source: https://www.statista.com/statistics/185167/number-of-doctoral-degrees-by-gender-since-1950/

ETD Collections

- NDLTD (Networked Digital Library of Theses and Dissertations):
  - 6+ million ETDs, 1.6 million in English
  - records only
- PDTG (ProQuest Dissertations & Theses Global):
  - 5 million ETD records, 2.7 million full text (majority in English)
  - subscription required
- Our collection:
  - ~450,000 ETD full text and metadata records

Numbers obtained on 5 November 2021

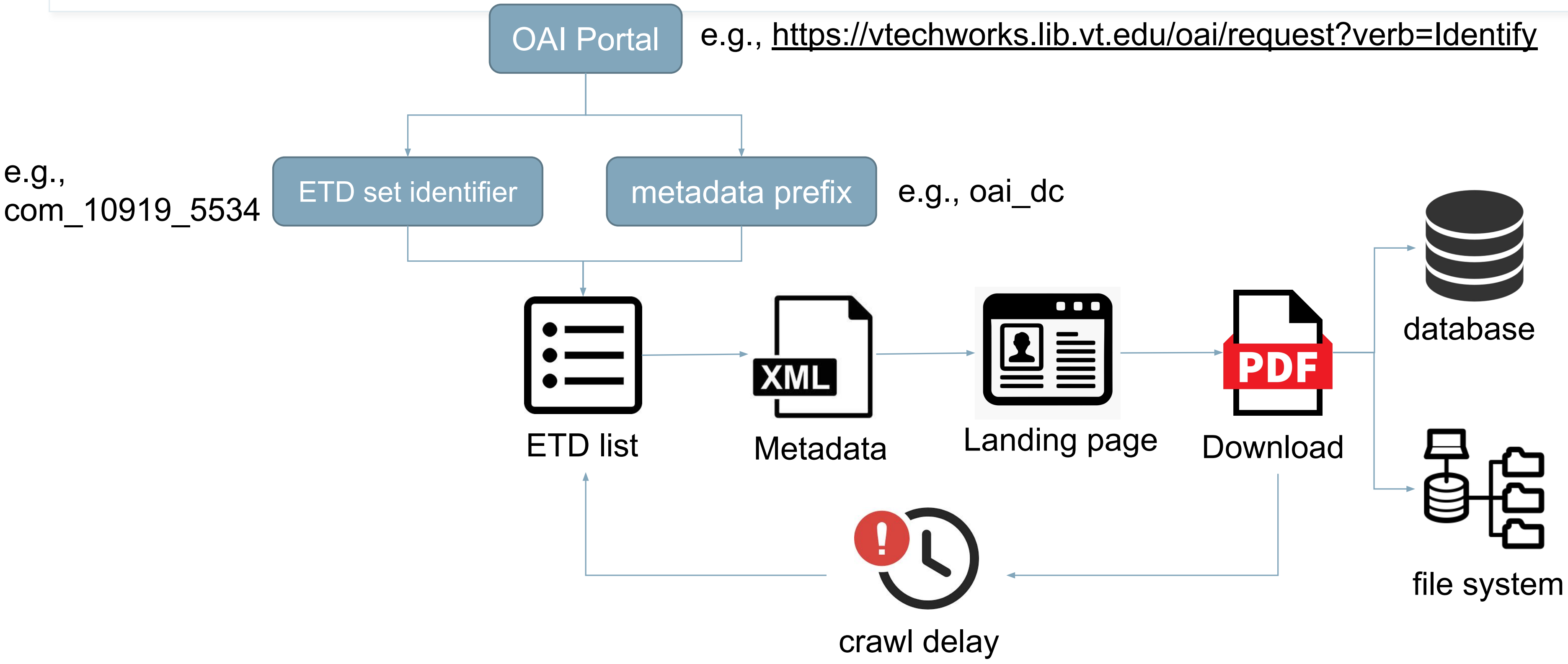
Building Our ETD Collection

- From over 40 university libraries; examples:

Access type	University	URL Example	ETD set identifier
OAI-PMH	Virginia Tech	https://vtechworks.lib.vt.edu/oai/request?verb=Identify	com_10919_5534
Sitemaps	UC system	https://escholarship.org/siteMapIndex.xml	

- Two access approaches
  - OAI-PMH -> metadata -> landing page -> PDF
  - Sitemaps -> landing page -> PDF
- Metadata
  - OAI-PMH
  - Scraping webpages, parsing HTML documents (obeying robots.txt)

Crawling Pipeline using OAI-PMH



OAI Portal

e.g., https://vtechworks.lib.vt.edu/oai/request?verb=Identify

e.g., com_10919_5534

ETD set identifier

metadata prefix

e.g., oai_dc

ETD list

Metadata

Landing page

Download

database

file system

crawl delay

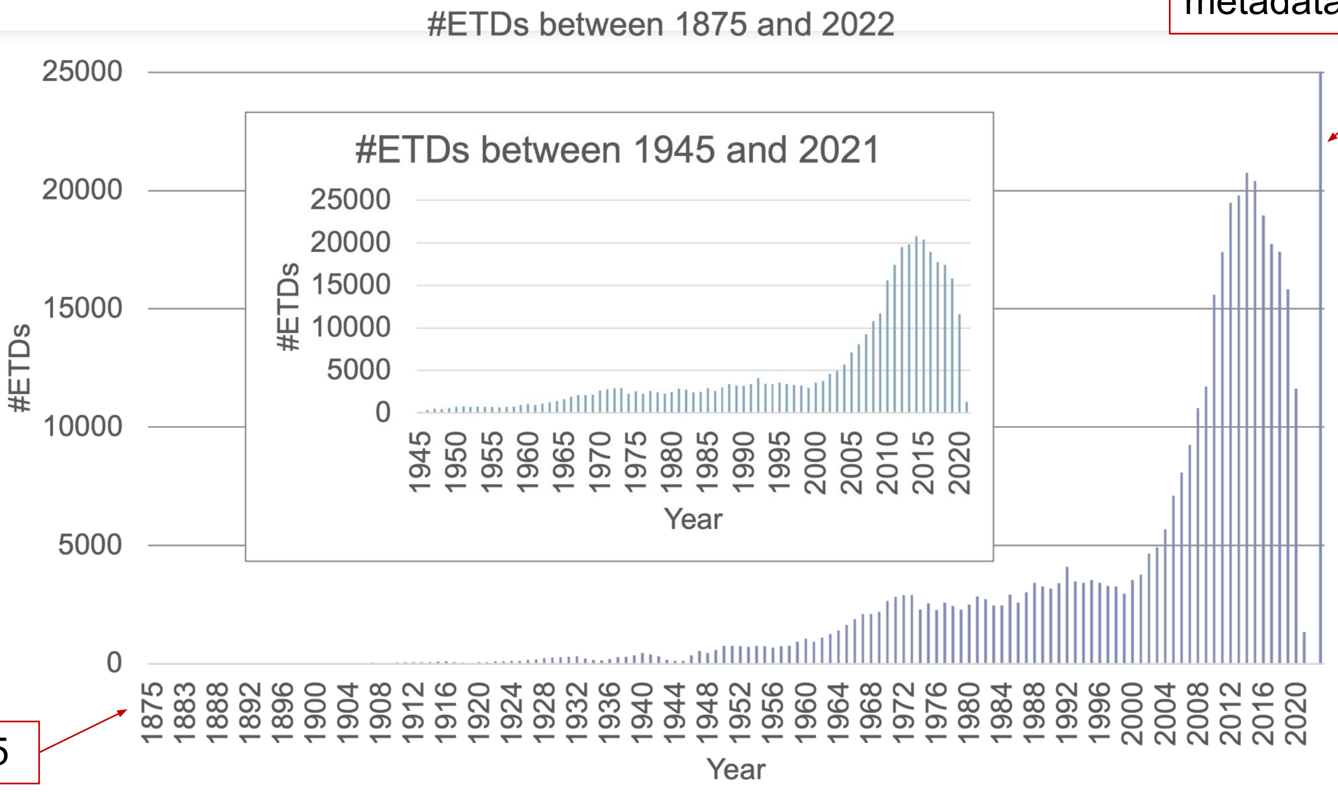
Challenges and Lessons

- Not all PDFs are downloadable.
  - Restricted access
  - HTML DOM structure varies across repositories.
- Not all metadata has same fields of information.
  - Often missing: department, discipline, subjects, year issued
- Inconsistent formats, e.g., “date-issued”
  - 'mm-dd-yyyy’
  - 'before yyyy’
  - 'after yyyy’
  - 'yyyy strings’
  - missing
- Requests blocked even when obeying crawl-delay in robots.txt

Our Top ETD Sources

Top 10 Universities	PDF
The Ohio State University	55780
Virginia Tech	29597
Georgia Institute of Technology	22400
Texas State University	21702
Kansas State University	19299
The University of Texas at Austin	18283
Oklahoma State University	17746
North Carolina State University	15365
University of Illinois at Urbana-Champaign	14281
Rice University	13151
Total (42)	**451,358**

Distribution By Year

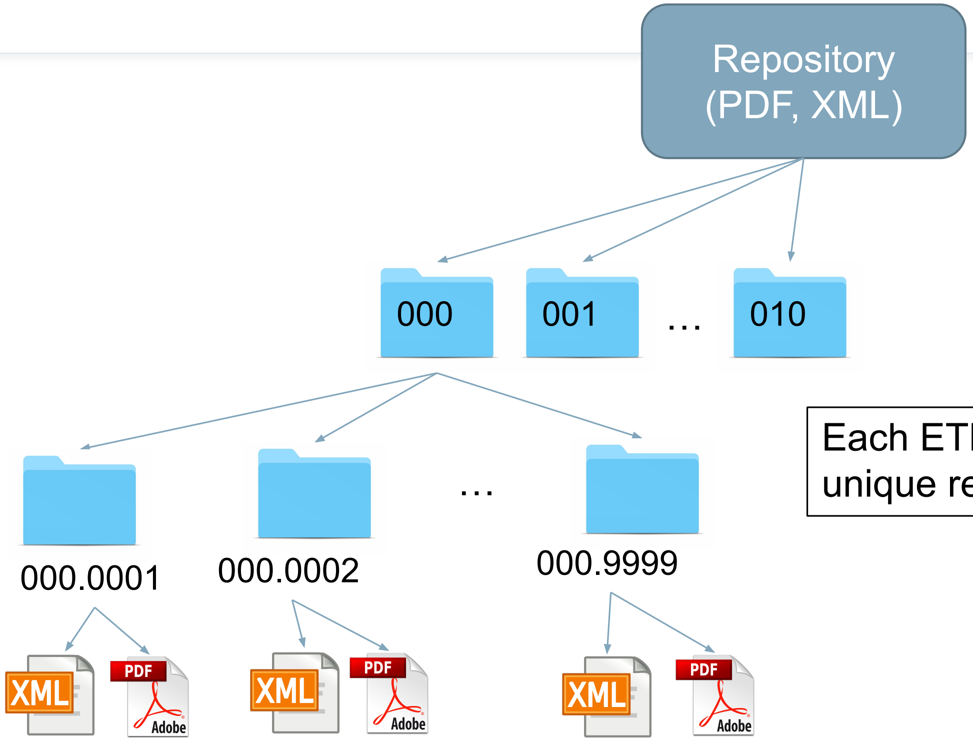


65955 ETDs dates not available in university-provided metadata

1875

ETD Repository

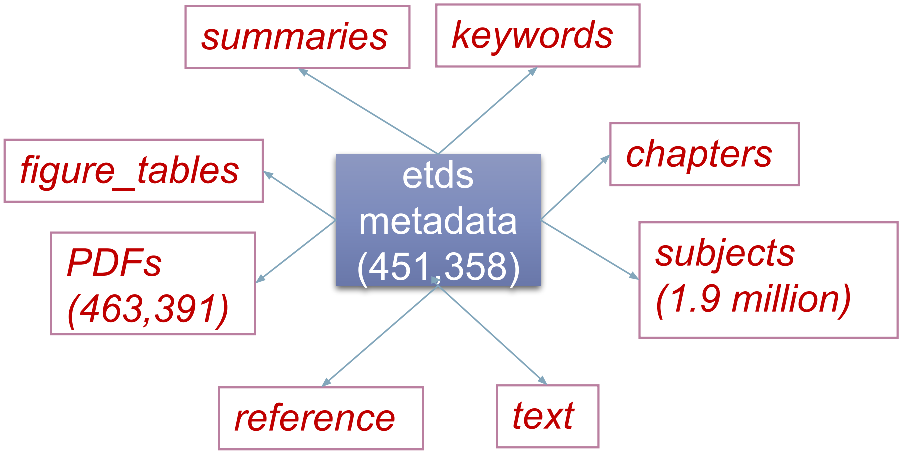
- ETDs have full text in PDF and XML.
- Total size: 3.4TB
- Hosted at ODU Computer Science data center
- Mirrored at VT University Libraries



Each ETD has a unique repository ID.

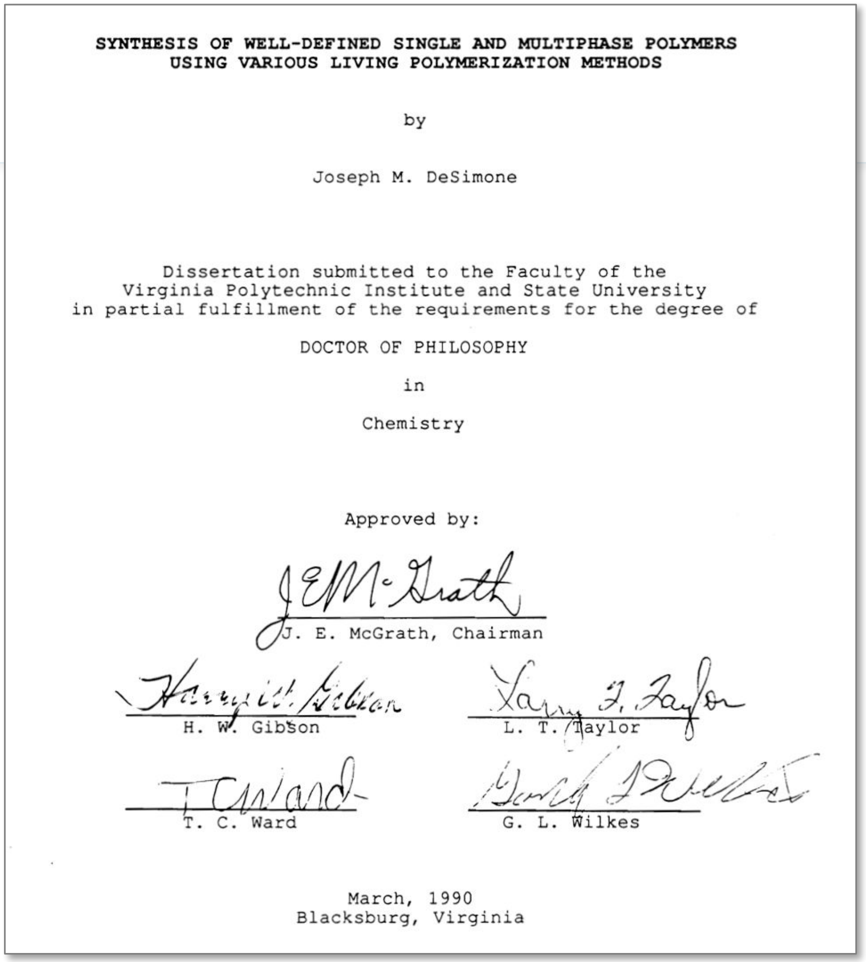
ETD Database

- Hosted in MySQL
- Born-digital vs. scanned
- Largest table: subject (1.9 million rows)
- Main table size: 451,358 rows



Database schema

Metadata Extraction from Scanned ETDs



SYNTHESIS OF WELL-DEFINED SINGLE AND MULTIPHASE POLYMERS
USING VARIOUS LIVING POLYMERIZATION METHODS

by

Joseph M. DeSimone

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Chemistry

Approved by:

J. E. McGrath, Chairman

H. W. Gibson

L. T. Taylor

T. C. Ward

G. L. Wilkes

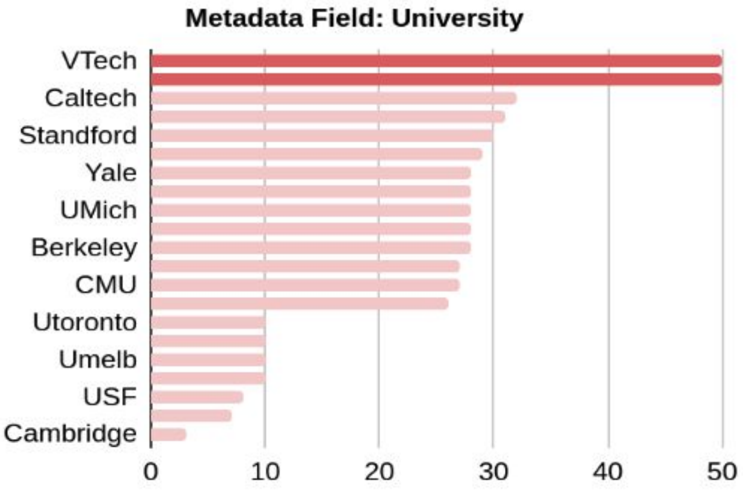
March, 1990
Blacksburg, Virginia

Field	Extracted Metadata
Title	synthesis of well-defined single and multiphase polymers using various living polymerization methods
Author	joseph m. desimone
University	virginia polytechnic institute and state university
Degree	doctor of philosophy
Program	chemistry
Advisor	j. e. mcgrath h. w. gibson l. t. taylor t. c. ward g. l. wilkes
Year	1990

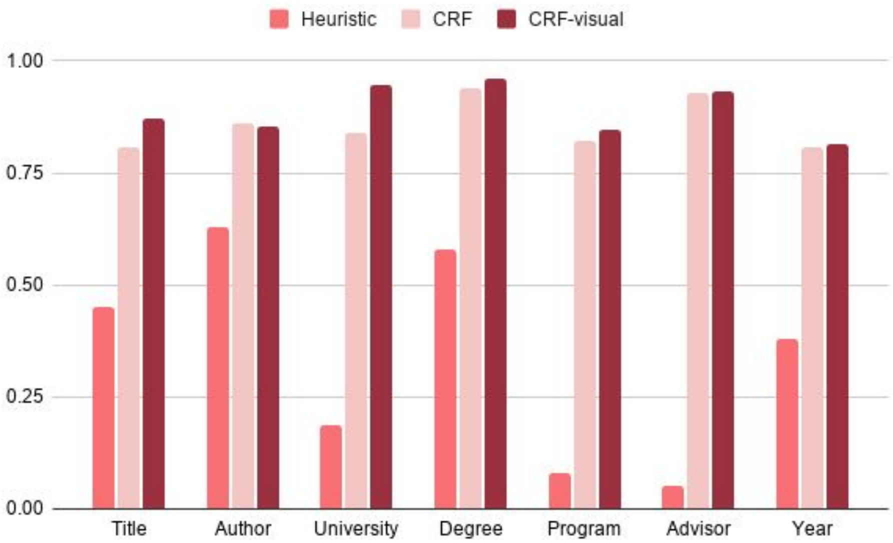
Choudhury et al. (2020 JCDL)
Choudhury et al. (2021 JCDL)

Metadata extraction from scanned ETDs

- ETD samples 1940-1990 (left figure).
- Conditional random field model with visual features (right figure)

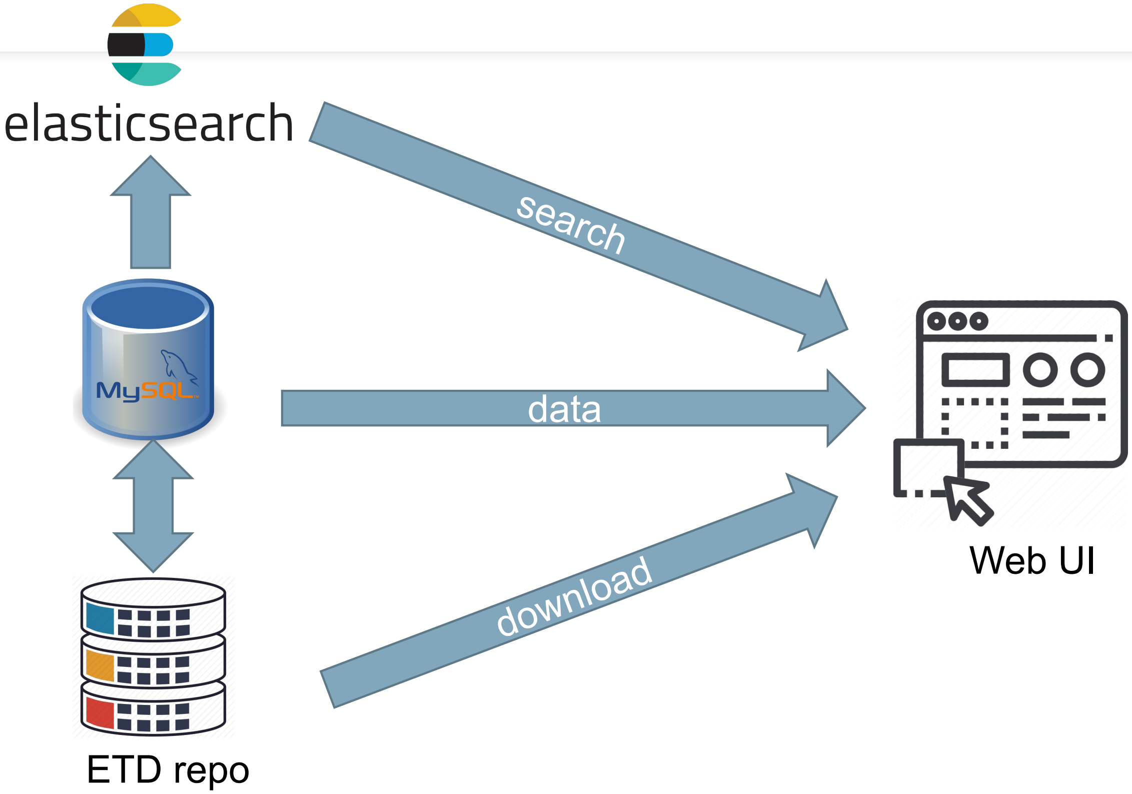


Distribution of 500 scanned ETD samples over universities



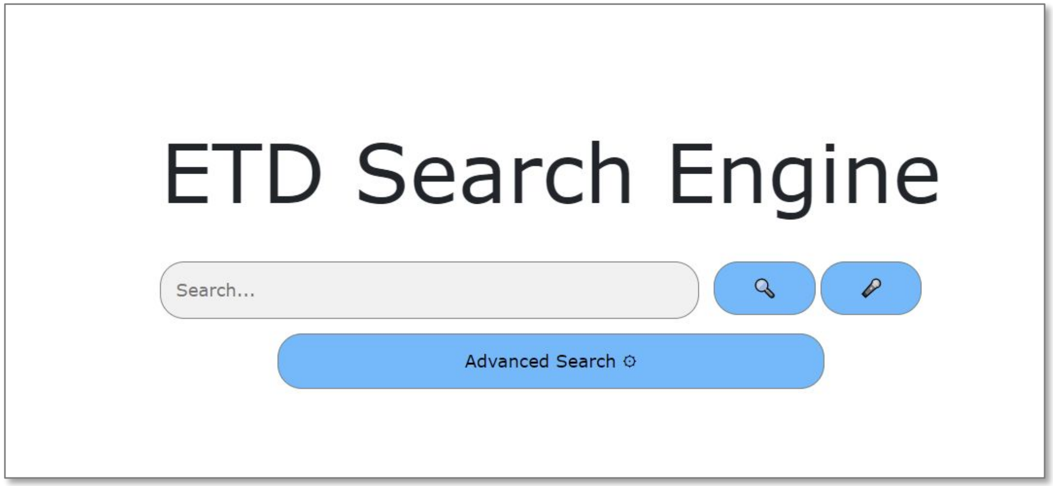
Model Performance (**81.3% - 97%**) to extract 7 metadata fields.

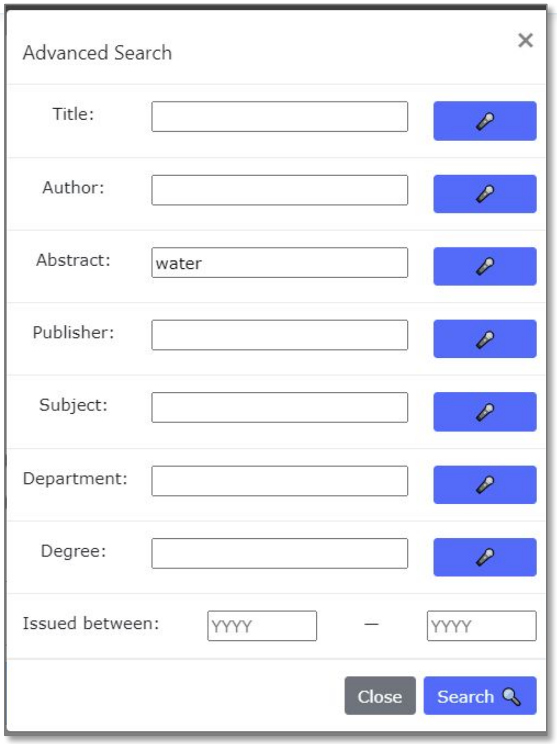
ETD Search Engine



Basic ETD Search Interface

- Single text box search
- Advanced search
- Autocomplete
- Spell check, Voice queries
- Voting
- RESTful search API




Advanced Search

Title:

Author:

Abstract: water

Publisher:

Subject:

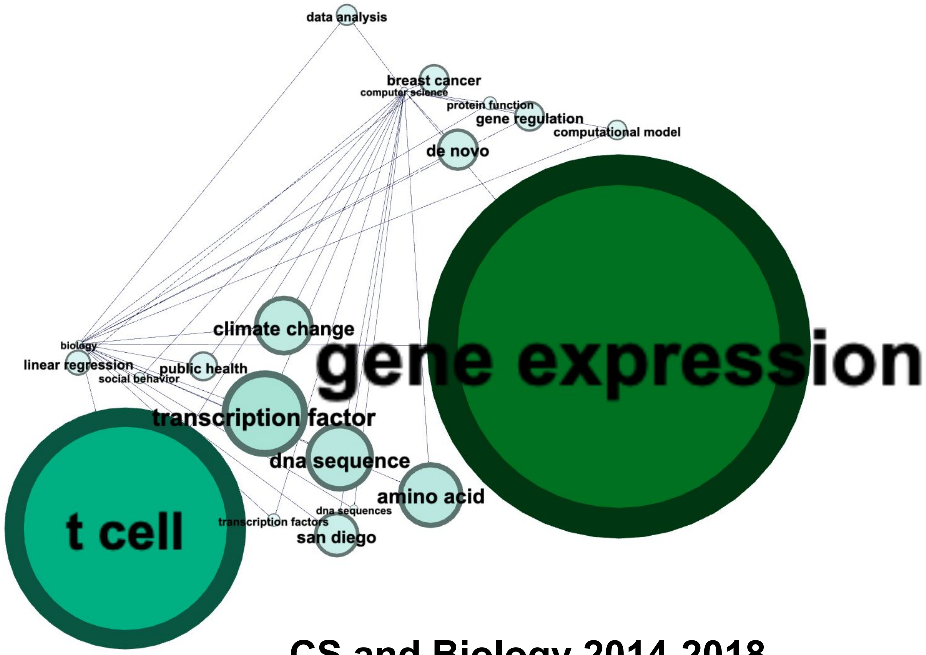
Department:

Degree:

Issued between: YYYY – YYYY

Services with Collected ETDs

- Segmentation
- Summarization
- Subject classification
- Figure and table search (Kahu et al. 2021 JCDL)
- Research topic analysis (right diagram)



CS and Biology 2014-2018

Courtesy: Ingram 2020 CNI presentation.

Language Model

- LM power depends on training text.
  - BERT/RoBERTa/DistilBERT/XLNet/ELECTRA – Wikipedia, BooksCorpus, ClueWeb, CommonCrawl, Gigaword
  - SciBERT – PubMed
  - Specter – S2ORC (Semantic Scholar)
- Pre-trained LMs may not work well for ETDs on certain tasks.
  - The low vocabulary overlapping problem (e.g., discipline specific jargon) is seen in existing studies (e.g., Kandimalla et al. 2020)
- Training from scratch is expensive. –> Fine-tune an existing LM
  - BERT (3.3 billion tokens, 52 hrs with a TPU; 7 days with a GPU)
  - SciBERT (3.2 billion tokens, 5 days with a TPU)

Building a Language Model for ETDs

- Start with SciBERT as the base model.
- Extract text from 8606 born-digital PDFs.
- 300 million tokens
- Fine-tune the LM using PyTorch transformers.
- Time taken ~20 hours on 1 Nvidia Quadro RTX 4000 GPU

- **Experiment 1**: 8606 documents (195 departments)
- **Experiment 2**: Remove the front matter (e.g., title page, table of contents, acknowledgements)

Evaluation

- Training with 8606 documents from 195 departments
- Measure: Perplexity
- Evaluation result: front matter confuses model -> excluded

Model Name	**Training Text**	**Perplexity Score (the lower the better)**
Language model 1	Full text	17.26
Language model 2	Full text but excluding the front matter	7.32

- Further evaluation regarding LM effect on classification

Conclusions

- OAI-PMH metadata gathering, crawling university libraries -> Collection of ~450k ETDs: full text PDF + metadata
- Analysis shows ETDs have inconsistent metadata information.
- Language model trained on 300 million tokens from ETDs
- Training uses fewer resources than general purpose (e.g., BERT) and scientific LMs (SciBERT) -> investigate whether it achieves comparable performance on subject classification

Ongoing and Future Work

- Improve metadata quality
  - Missing metadata (Choudhury et al. 2021 JCDL)
  - Fixing inconsistencies between metadata and data (table below)
  - Resolving university names: acronyms, full names, typos, e.g.,
    - Johns Hopkins University vs. jhu
    - Texas A&M University vs. Texas A & M University (additional space around "&")

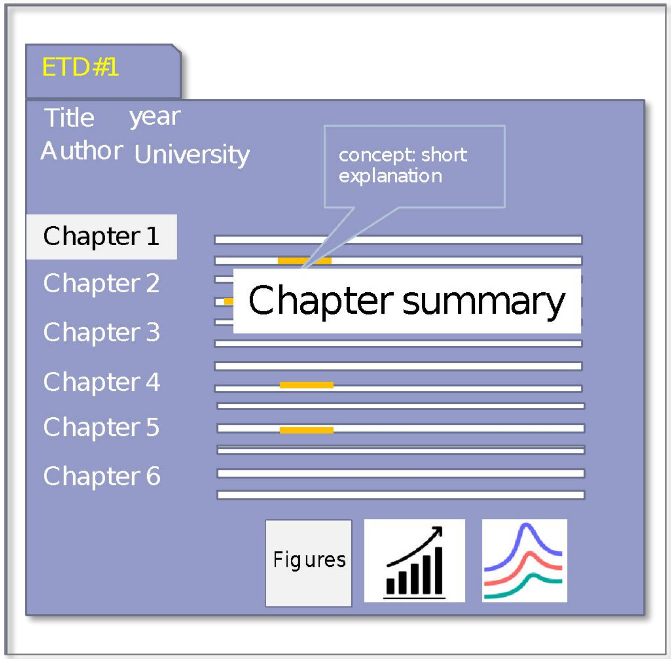
ETD Cover Page	Library Metadata
chemistry	polymers
administration and supervision of special education	special education
educational administration	school administration

Ongoing and Future Work

- Improving LM
  - Increase number of training documents
  - Removing traces of tables and figures from text
  - Department names and disciplines in metadata -> unified schema (e.g., Microsoft Academic Graph) for subject classification
- Challenges
  - Conversion from PDF to text is not perfect.
  - Obtaining clean sentences

Ongoing and Future Work

- New features to be added to UI:
  - Multi-modality search results (chapters, figures, and tables)
  - Improved document summary page (right figure)
  - Improved services: metadata extraction, segmentation, classification, entity extraction, knowledge graph generation



Document Summary Page