Opening Books and the National Corpus of Graduate Research

William A. Ingram (VT Library), Edward A. Fox (VT CS), and Jian Wu (ODU CS)

Abstract

Virginia Tech University Libraries, in collaboration with Virginia Tech Department of Computer Science and Old Dominion University Department of Computer Science, request $505,214 in grant funding for a 3-year project, the goal of which is to **bring computational access to book-length documents, demonstrating that with Electronic Theses and Dissertations (ETDs)**. The project is motivated by the following library and community needs. (1) Despite huge volumes of book-length documents in digital libraries, there is a lack of models offering effective and efficient computational access to these long documents. (2) Nationwide open access services for ETDs generally function at the metadata level. Much important knowledge and scientific data lie hidden in ETDs, and we need better tools to mine the content and facilitate the identification, discovery, and reuse of these important components. (3) A wide range of audiences can potentially benefit from this research, including but not limited to Librarians, Students, Authors, Educators, Researchers, and other interested readers.

We will answer the following key research questions: (1) How can we effectively identify and extract key parts (chapters, sections, tables, figures, citations), in both born digital and page image formats? (2) How can we develop effective automatic classification as well as chapter summarization techniques? (3) How can our ETD digital library most effectively serve stakeholders? In response to these questions, we plan to first compile an ETD corpus consisting of at least 50,000 documents from multiple institutional repositories. We will make the corpus inclusive and diverse, covering a range of degrees (master's and doctoral), years, graduate programs (STEM and non-STEM), and authors (from HBCUs and non-HBCUs). Testing first with this sample, we will investigate three major research areas (**RAs**), outlined below.

RA 1: Document analysis and extraction, in which we experiment with machine/deep learning models for effective ETD segmentation and subsequent information extraction. Anticipated results of this research include new software tools that can be used and adapted by libraries for automatic extraction of structural metadata and document components (chapters, sections, figures, tables, citations, bibliographies) from ETDs — applied to both page image and born digital documents.

RA 2: Adding value, in which we investigate techniques and build machine/deep learning models to automatically summarize *and* classify ETD chapters. Anticipated results of this research include software implementations of a chapter-level text summarizer that generates paragraph-length summaries of ETD chapters, and a multi-label classifier that assigns subject categories to ETD chapters. Our aim is to develop software that can be adapted or replicated by libraries to add value to their existing ETD services.

RA 3: User services, in which we study users to identify and understand their information needs and information seeking behaviors, so that we may establish corresponding requirements for user interface and service components most useful for interacting with ETD content. Basing our design decisions on empirical evidence obtained from user analysis, we will construct a prototype system to demonstrate how these components can improve the user experience with ETD collections, and ultimately increase the capacity of libraries to provide access to ETDs and other long-form document content.

Our project brings to bear cutting-edge computer science and machine/deep learning technologies to advance discovery, use, and potential for reuse of the knowledge hidden in the text of books and book-length documents. In addition, by focusing on libraries' ETD collections (where legal restrictions from book publishers generally are not applicable), our research will open this rich corpus of graduate research and scholarship, leverage ETDs to advance further research and education, and allow libraries to achieve greater impact.

Opening Books and the National Corpus of Graduate Research

William A. Ingram (VT Library), Edward A. Fox (VT CS), and Jian Wu (ODU CS)

The University Libraries at Virginia Tech (VT), in partnership with VT Dept. of Computer Science and Old Dominion University (ODU) Dept. of Computer Science, requests $505,214 in grant funding for a three-year project to bring computational access to book-length documents, demonstrating that with Electronic Theses and Dissertations (ETDs). Our team has extensive experience in digital libraries, information retrieval, document analysis, and related methods in machine learning and deep learning. We will enhance libraries' ETD programs, devising effective and efficient methods, opening the knowledge currently hidden in this rich corpus of graduate research and scholarship. Our research aligns with (1) the IMLS Strategic Plan to increase public access to important content in a more readable and accessible manner and (2) the National Digital Infrastructures and Initiatives focal area of expanding digital cultural heritage capacities by enabling computational use of collections.

1 Statement of National Need

A fundamental problem in producing the national platform is improving access to book-length material. Despite the millions of digital books available online, theoretical and applied research (e.g., in text retrieval, extraction, categorization, and summarization) has focused on shorter documents. Access to scholarly material usually comes through online library catalogs and discovery services, abstracting and indexing databases, and increasingly through large search engines like Google or Google Scholar. However, the conceptual models, search algorithms, and techniques underlying such systems are not well-adapted to long documents.

Our research aims to enhance computational access to book-length documents, demonstrating that with ETDs, whose lengths range from 140 to 370 pages [89] compared with academic papers, whose medium length is 12 pages[1]. Our theoretical foundation is the 5S Framework for Digital Libraries (Societies, Scenarios, Spaces, Structures, Streams) [30, 22, 75, 21, 23, 31]. From the 5S perspective, the primary societies of interest for this work are librarians, the research community, and the community of library users: authors, students, researchers, educators, and readers. The set of scenarios, which describe the activities carried out by these societies, and the services to support those, include browsing, citing, classifying, describing, discovering, extracting, filtering, indexing, learning, measuring, reading, recommending, requesting, reviewing, searching, sharing, summarizing, visualizing, and writing. Spaces include 2D presentations of content, as well as internal feature and vector spaces. Structures of interest include citation networks, metadata, tables, taxonomies, and workflows, as well as the logical structure of the documents themselves. Streams relating to the collection of digital objects include chapter text, figures, summaries, citations, bibliographies, and the flow of content from/to users.

Our research will be informed by studying societies and scenarios, leveraging NSF I-Corps training on customer discovery [26, 58] and 30 years of work with ETDs [24, 81, 79, 80] that suggest scenarios and research library services like: "Graduate student Rachel Researcher has an idea for a thesis topic, so she visits the university institutional repository in search of ETDs in her field. She is pleased to find digital library services that provide her with an interface for browsing skimmable chapter summaries, a visualization of extracted figures and tables, and a citation graph for the collection. Likewise, Arthur Author uses these same summarization and classification services for preparing the abstract and keywords for his own research works. Behind the scene, Cathy Cataloger is tasked with indexing and describing a collection of book-length documents to be ingested into the digital library. She uses an application that automatically generates summaries for her documents and recommends hierarchical Library of Congress Classification categories using a machine learning algorithm." Figure 1 illustrates the main stakeholders, and ETD services of particular interest.

[1] Calculated with 2 million randomly sampled papers from CiteSeerX.

Motivating our research are the millions of ETDs now in university collections, supported by digital library applications which do little to meet the needs of users beyond simple searching and browsing. Once discovered, an ETD usually is downloaded as a large PDF file. Sometimes an abstract or supplementary file may be downloaded separately, but these repositories only provide simple tools for accessing whole works, not their components. Full-text indexes, if available, do not allow searching for each of the tables, figures, images, or other data elements contained. Libraries lack the requisite local infrastructure for content mining of their ETDs. Their collections are not "computationally amenable." More broadly, nationwide open access services for ETDs [25] generally function at the metadata level, requiring users and content mining systems to go to each repository for download and analysis, one document at a time. Metadata is often limited to (some of) the simple Dublin Core elements. *Much important knowledge and scientific data lies hidden in ETDs, and we need better tools to mine the content and facilitate the identification, discovery, and reuse of these important components.*

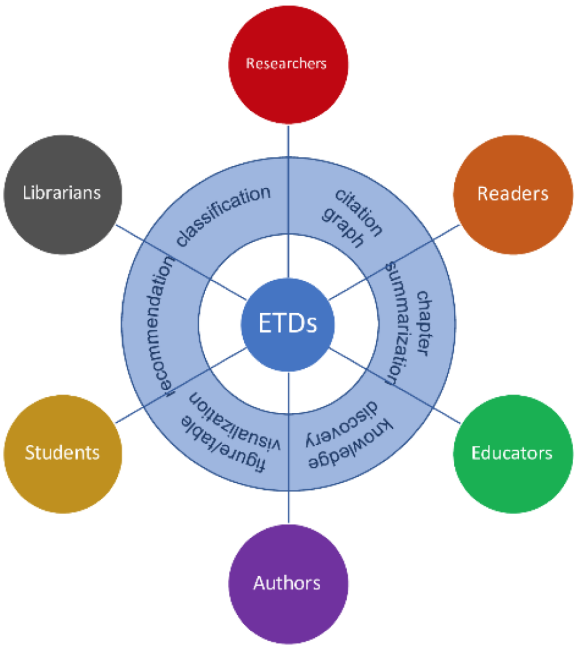

Figure 1: Stakeholders and ETD services.

Our research investigates, by looking at ETDs: Who are the users of long documents in digital libraries? What scenarios describe their work, what streams of information are useful to them, and what services best meet their needs? Moreover, what specific services can add value to library book collections by making them more computationally amenable? Our research questions are:

- **RQ1:** How can we effectively identify and extract key parts (chapters, sections, tables, figures, citations), in both born digital and page image formats?
- **RQ2:** How can we develop effective automatic classification as well as chapter summarization techniques?
- **RQ3:** How can our ETD digital library most effectively serve stakeholders?

Our submission falls into the *Research in Service to Practice* category because: (1) The project team (Section 2.4.1) will investigate key unsolved questions to overcome obstacles to better segmenting, extracting, and discovering knowledge from ETDs (Section 2.1). Our methods can potentially be applied to a vast amount of other book-length documents. The services can benefit many types of stakeholders (Figure 1). (2) Our proposed research is based on prior studies involving digital library theory, natural language processing, machine learning, and deep learning techniques (see the beginning portion of each part of Section 2.2). (3) Our proposal has clear research questions with detailed descriptions of promising solutions (Section 2.2). (4) The research products will be disseminated beyond publishing journal articles and presenting at academic conferences (Section 2.5).

2 Project Design

Addressing the key research questions are three goals of this project: (1) Research and enhance open source tools that automatically segment ETDs into chapters, and extract key parts; (2) Research and enhance effective automatic text classification and chapter summarization techniques for ETDs; (3) Develop and evaluate a proof-of-concept digital library prototype that supports searching, browsing, and discovering ETD *content* based on the techniques developed through (1) and (2). Figure 2 illustrate the goals, three research areas, tasks, and anticipated deliverable results.

Assumptions: This research project assumes that textual and non-textual information can be extracted from ETDs. We have observed a small fraction of ETDs that are not machine-readable due to permission restrictions.

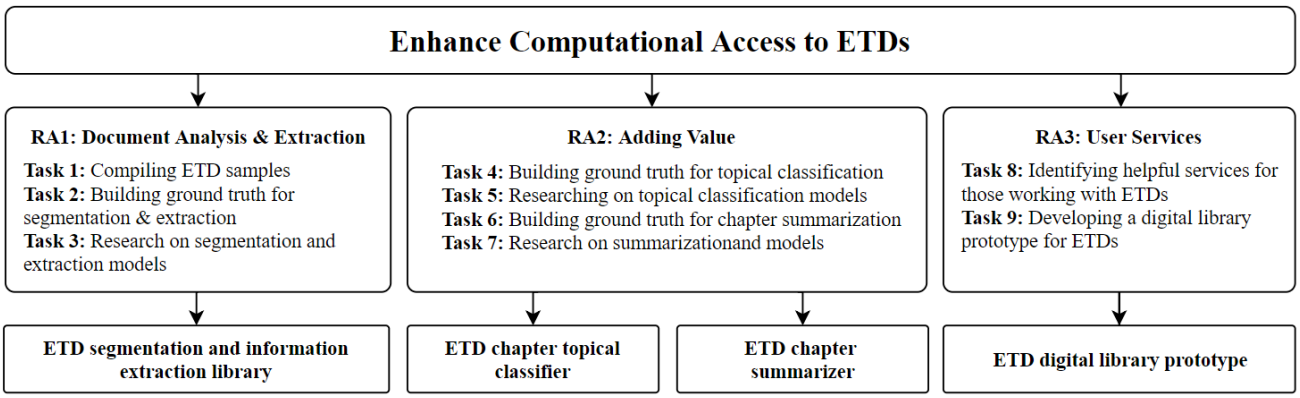

Figure 2: Goals, research areas (RAs), tasks, and results of the proposed research.

There are also non-digital-born ETDs that were scanned but not OCR'd (or poorly OCR'd) where extraction yields gibberish characters. These exceptions are usually not predictable, but can be easily identified, e.g., using extractor error messages emitted by extractors, or by output file sizes. Consequently, we assume access to English ETDs published in PDF, which is used for typical modern scholarly documents. **Potential risks:** Although we focus initial research on what we hope will be an unbiased sample of ETDs, it is inevitable to miss ETDs with unique styles (required by a department or institution or discipline). To mitigate this risk, evaluation will be on ETDs from a wide variety of graduate programs, including STEM and non-STEM disciplines. We also will design methods that allow adding new models for novel ETD styles.

2.1 ETD Corpus

We will collect an initial corpus of at least 50,000 ETDs from multiple institutional repositories, including but not limited to Virginia Tech, the first university to require ETD submissions, Old Dominion University (ODU), a minority serving institution, and Pennsylvania State University (PSU), one of the largest universities in the United States. The number of ETDs from these graduate programs, and lists of majors with the most ETD submissions, are tabulated in Table 1 in Appendix A (Supportingdoc1.pdf). Later we will scale to at least 200,000 ETDs.

In an effort to include ETDs from underrepresented institutes, we include a collection of ETDs from Historically Black Colleges and Universities (HBCUs). According to the National Center for Education Statistics [53], in academic year 2014-2015[2], there were in total 178,548 doctoral and 758,708 master's degrees conferred by post-secondary institutions, of which about 2,490 doctoral (1.4%) and 7,427 master's (1.0%) degrees were from HBCUs. A preliminary analysis of the top 10 HBCUs ranked by enrollment [45], indicates that most HBCUs use ProQuest to archive ETDs. By searching institution names, 8,948 full-text ETDs were found from 6 out of the 10 HBCUs we examined. We will manually download 1,000 randomly from ProQuest.

State-of-the-art research in machine learning for document segmentation, classification, and summarization is biased heavily toward STEM disciplines [6]. For example, CiteSeerX is a digital library of scientific literature, not focused on arts or the humanities. Several of the tools used for extracting header information, metadata, and bibliographic data from academic papers come pre-trained on computer or information science papers, due to their intended use. Most state-of-the-art abstractive summarization work is based on machine/deep learning models trained on news article collections. The arts and humanities are not well represented in any of these.

[2]More recent data is not available at the time of this proposal.

The ETD templates vary significantly across graduate programs. To increase the diversity, we will incorporate material from a wide cross selection of majors and disciplines. Our corpus will include ETDs from STEM, but also the social sciences, arts, and humanities.

The properties of ETD data to be collected are detailed in Table 1 in Appendix A (Supportingdoc1.pdf).

2.2 Research Areas

2.2.1 Research Area 1: Document analysis and extraction

Our first area of research is to explore how to effectively identify and extract key parts from ETDs: chapters, sections, tables, figures, and citations. We will evaluate our research in this area separately for each type of extracted content, as well as holistically with proof-of-concept applications. There has been much research on segmentation and information extraction from scholarly big data, e.g., [40]. Here, we briefly review existing techniques developed, and identify tools that might be adapted to ETDs.

Related works

Metadata Extraction: A comparison between 9 metadata extraction tools for academic documents indicates that GROBID [49] outperforms all other off-the-shelf tools, such as SVMHeaderParse [33]. GROBID applies Conditional Random Fields (CRF), a sequential machine learning model, to extract and tag metadata appearing in academic documents, including header information (e.g., title, authors, abstract) and references. The current release (version 0.5) trains 11 models (2 for patents), each using the same generic CRF-based framework, but having their own set of features, training data, and normalization. GROBID is able to resolve the hierarchical structures of a document. However, the data GROBID trained on do not include ETDs. Experiments indicate that running it directly on ETDs produces unexpected results. There are other extraction tools such as CERMINE [84] and ScienceParse [3], but they do not compare to GROBID in terms of extraction fields and quality.

Citation Extraction: ParsCit is a CRF based citation parser for academic papers [15]. It has been used for papers, [42], patents [50], books [91], and ETDs [67]. ParsCit works well when citations appear in one bibliography section. When applied to ETDs, in which citations are spread across different chapters, ParsCit only captures from the last chapter. It also fails to capture from reference sections titled "Notes" or "Literature Cited". In preliminary work, we demonstrated machine learning could aid parsing of references from a broad range of disciplines, such as those found at the end of ETDs [66, 67]. Extending this research, we are working with a team from the National University of Singapore to explore and develop Neural ParsCit, an open source application for accurately parsing and extracting references from scholarly works with deep learning. The team recently retrained Neural ParsCit [69] with a large volume of synthesized citation strings, and evaluated it against a corpus of citations extracted from ETDs [89]. Based on a manuscript we are preparing [89], we plan to integrate Neural ParsCit with other ETD segmentation and extraction processing.

Figure and Table Extraction: Figures and tables are ubiquitous in ETDs, to report experimental and analysis results. Early work on figure extraction focused on particular domains, such as chemistry [8, 10], computer science [9], and biomedical science [48]. Tables were extracted separately using rule-based methods, e.g., [47, 46], and machine learning models [19]. Recently, domain independent algorithms were developed to extract figures and tables together using rule-based [14], machine learning [13], and neural network methods [76]. To the best of our knowledge, *pdffigures2* and *deepfigures* are top-rated open source extraction tools for figures and tables [13, 76], with minimum assumptions regarding document lengths, so of potential value for ETDs.

Optical Layout Analysis: ETD collections contain documents from image scanning. Many tools mentioned above may not be applicable to such documents. In addition, ETDs may also have a variety of layout styles, the information of which is lost when converted to plain text. Optical document layout analysis is a promising solution to this issue, by identifying bounding boxes of designated text on given pages. Early work uses computer

vision algorithms to perform page segmentation and zone recognition [87, 61]. Many recent works focus on document image and text extraction using deep learning methods [78, 39, 29, 77]. Hybrid methods like [62] combine conventional page segmentation and deep learning for block content classification.

Research activities

The first step in this process will be to perform a detailed analysis of our corpus of ETDs. Our goal is to implement document segmentation and extraction techniques that can be applied to a diverse range of documents from multiple majors and disciplines. For example, the structure of an architecture ETD may be very different from the structure of a biology ETD. We also recognize the importance of analyzing a mix of old and new ETDs. Our corpus of ETDs contains both born digital and digital facsimiles of physical documents. For the latter, we will compare state-of-the-art OCR [37, 83, 93] with commonly available and widely used tools for text extraction.

Task 1: Compiling ETD sample: We will begin with at least 200 ETDs, covering multiple graduate programs and a range of years. To make the sample more representative of the nationwide population of graduate student authors, we impose 3 selection criteria based on the counts of doctoral and master's degrees between 1970 and 2015 in the *Digest of Education Statistics* [53]. (1) The STEM vs. non-STEM ETD ratio is roughly 1:1 for doctoral degrees. This ratio is about 1:0.7 for master's degrees. (2) The ratio between ETDs before 2000 and after 2000 is about 1:1.7 for doctoral degrees, and 1:2 for master's degrees. (3) The ETDs from HBCUs are about 1.4% for doctoral and 1.0% for master's degrees, so later we will over-sample from these.

Task 2: Building ground truth for ETD segmentation and extraction: From this diverse sample, we will construct a ground truth for segmentation and extraction by manually identifying and extracting key document parts. We will manually extract the full text for all chapters, sections containing related works or literature reviews, tables, figures, associated caption text, and references. We expect the layout and structure of these will vary widely across disciplines. This ground truth will be used later for training and evaluation.

Task 3: Researching on ETD segmentation and extraction models: Next, we will research on robust heuristics and machine learning models for automatic document segmentation and extraction. We will first research on the performance of state-of-the-art document segmentation and extraction tools GROBID, CERMINE, and SCIENCE PARSE for chapter and section extraction when applied to ETDs. Secondly, we will research on methods based on optical based document layout analysis, and then compare them with non-optical based methods.

This research will involve retraining the underlying machine-learning models, since they were originally trained on short documents and on a narrow range of disciplines, mainly science and engineering. The segmentation problem can be transformed into a text classification problem in which we seek the correct text boundary between two chapters. The extraction problem also can be solved with text classification tasks in which we first classify text on the line level and then on the token level. We will apply sequential tagging methods such as CRF [43] and Bi-LSTM [36]. In addition to lexical features, we also will incorporate positional and font features.

In preliminary research we extracted some images and captions from ETDs [81]. That study made use of *pdfimages* for figure and table extraction and *pdf2xml* for identifying and extracting captions. The preliminary work demonstrated some successful extraction of images and captions from born digital documents, but more research is needed, including for effective image extraction from scanned documents.

2.2.2 Research Area 2: Adding value

Our second area of research focuses on how to develop effective automatic classification and chapter summarization techniques for ETDs. These techniques provide a distinct added value to ETD collections, aiding discovery and selection of relevant ETDs, through advanced computational processing of these collections.

Related works

Text Classification: Text classification is a fundamental problem in text mining. With shorter documents, it has been extensively studied, e.g., [1]. Recent algorithms utilize machine learning (e.g., [38]) and recently deep neural networks (e.g., [73, 51, 94]). Word embeddings are found to be very effective representations when classifying sentences and news articles, e.g., [32, 44]. However, there are relatively less works on academic document classification. In a preliminary study, we compared supervised learning models against multilayer perceptron (MLP) models on classifying abstracts into six subject categories [88], namely, chemistry, physics, computer science, biology, materials science, and others, using a Web of Science database as the training set [12]. Our work indicates that given a sufficiently large training sample (>150k), MLP achieves comparable performance to traditional machine learning models, e.g., Random Forest. The best F_1 achieved ranges from 65% (others) to 94% (computer science). In another work, we researched topical categorization of ETDs [79]. This work uses machine learning methods to classify ETDs using the Library of Congress Classification (LCC) system.

Text Summarization: Text summarization has been long studied. Early research focused on extractive ways by selecting important sentences ranked by scores [56, 55]. In the DUC 2006 [59] and TAC 2011 [60] competitions, query-focused document summarization and guided summarization started to draw more attention [20, 64]. In the past decade much progress has been made [92], focusing on improving information coverage, coherence, and non-redundancy [35, 86]. In particular, machine learning models are injected into traditional models to learn weights of sentences. Recently, progress on abstractive summarization, yielding more readable summaries, has accelerated. However, most of these methods are trained and/or tested on news reports such as *Associated Press*, *New York Times*, *Xinhua News*, and *WSJ* articles, which are much shorter than ETDs [16, 17, 27, 5]. For instance, the articles in the *20newsgroup* dataset contain less than 100 words on average. We plan to fill the gap by researching effective extractive and abstractive methods to summarize ETD chapters. This will be particularly challenging, due to the length of chapters, the diversity of disciplines, the variation in writing styles, and the fact that ETDs report about the discovery of new knowledge, often introducing new vocabulary.

Research Activities

Task 4: Building ground truth for ETD (chapters) topical classification: In this task, we manually label ETDs, as well as their chapters, with one or multiple subject categories using the LCC taxonomy. We do this for the samples selected above, and use the results as ground truth for text classification. The labeling will be performed by at least two people and reviewed by a third person. Metadata information may be useful for classification; it can be harvested from institutional repositories via OAI-PMH [82, 54]. In some cases, this metadata is created by the author of the work being deposited. In other cases the metadata has have been created by professional library catalogers or by graduate program staff. We will use the labeled subject categories for training and evaluating the quality of our automated classification methods. The labels later can be used for browsing and searching, and when available per chapter, could be invaluable for interdisciplinary ETDs.

Task 5: Building topical classification models for ETDs and their chapters: This task will build on the works above and investigate how to develop scalable methods for automated *multi-label* classification at the ETD and chapter levels. In the *multi-label* case [57, 85], an ETD chapter can be classified as both "biology" and "computer science", with different likelihoods. Traditional solutions train N supervised classifiers, in which N is the number of classes. This is not efficient when the number of classes is high (e.g., $N > 100$). We will investigate neural network based methods and compare them with traditional solutions to classify ETD chapters in the LCC taxonomy. These methods represent key text blobs in an ETD chapter using dense high dimensional vectors. The challenge is to identify representative statements, and to represent the whole document using vector representations of individual words. Additional clues abound: In many ETDs, the introduction section includes short statements about high-level content descriptions of subsequent chapters.

Task 6: Building ground truth for ETD chapters summarization: One source to build the ground truth is to use abstracts of papers in publicly available sites such as arXiv. This is based on the observation that some

ETD chapters are similar to a preprint, or a portion of a published paper. At least 200 (chapter, abstract) pairs will be collected to generate a ground truth sample. We will create more ground truth samples as needed.

Task 7: Researching on summarization models for ETD chapters: In the fall of 2018 Co-PI Fox taught the course CS4984/5984: *Big Data Text Summarization*, which explored automatically constructing English language summaries of the important information in a large document collection. Three project teams focused on using state-of-the-art deep learning methods for generating abstractive summaries of ETD chapters. PI Ingram took the course and led one of the ETD teams. The ETD project teams made use of the large and diverse corpus of ETDs housed in University Libraries' institutional repository, VTechWorks[3]. Course projects from each of the three ETD teams are made public in VTechWorks [2, 41, 4]. Our research will build on this preliminary work. In particular, we will develop baselines using at least three types of summarization models: Sequence-to-Sequence (e.g., [52]), Pointer-Generator (e.g., [74]), and Reinforcement Learning (e.g., [7]). We will train models using a set of ground truth chapter summaries. We will also compare extractive and abstractive methods, and explore how results depend on document length.

To mitigate the shortage of training samples, we will leverage *transfer learning* to also summarize ETDs using bullet points. In this attempt, we will apply models, trained on a large volume of news summarization dataset, to ETDs. In particular, the Google DeepMind dataset contains 93k articles from the *CNN* and 220k articles from the *Daily Mail* websites. Both news providers supplement their articles with a number of bullet points summarizing aspects of the information contained in the article [34]. An alternative way to summarize ETD chapters is to use concept maps. In preliminary work, we extracted key concepts from ETDs, automatically generated concept maps, translated them into Spanish, and demonstrated their utility as alternatives to abstracts, to aid discovery of ETDs by Spanish speakers [70, 71]. We will compare this approach with the above mentioned methods and others that emerge in this fast moving research area.

2.2.3 Research Area 3: User services

Research Area 1 analyses ETDs to discover the properties and parts of these documents, the streams of information underlying the digital objects, and how they can be identified and extracted. Research Area 2 seeks to determine how computation can be applied to these collections and streams to add value. Our third research area builds on these findings with an investigation of how extracted information streams and value-added services can be presented to online library users. In particular, we investigate new ways to interact with ETD collections, and study which best support the needs of the user community.

Related works

Digital Library Services: CiteSeerX was one of the first academic search engines connecting scholarly documents using automated citation indexing [28, 90]. There are ETDs indexed, but the median number of pages of a sample of 2 million documents is only 12, indicating that the majority are much shorter than ETDs. ProQuest is a commercial publishing system for ETDs. Their services are focused on submitting, storing, disseminating, and archiving ETDs. Google Books and HathiTrust are both digital libraries that focus on books, with about 15 million and 16 million volumes, respectively [11]. Among the more important services they provide are discovery and digital preservation [72, 63]. The HathiTrust Research Center (HRTC) has developed tools to do shallow analysis of documents present in the digital library, but our proposed work is fundamentally different. (1) The HTRC toolkit ignores the document structure, which is one of our research foci. (2) Their text analysis functions are at the phrase level, rather than on constructing meaningful sentences with rich semantic meanings, as is needed for summarization, which we will research at the chapter level.

Research Activities

[3] http://hdl.handle.net/10919/5534

Task 8: Identifying helpful services for those working with ETDs: Good user interface design is vital in providing access to digital content, collections, and services to a wide range of users. This task is aimed at building an understanding of the services best suited to ETD collections. By leveraging activities of Virginia Tech's Center for Human-Computer Interaction (HCI) and recent NSF I-Corps training on customer discovery [26, 58], we will study user problems and needs regarding digital libraries. Academic database platforms like ScienceDirect [18] present users with advanced visual features for interacting with journal articles and e-books, such as outlines, figures, tables, references, article highlights, and recommended works. By conducting Institutional Review Board approved user studies, we will identify and understand users' information needs and information seeking behaviors. We will identify user characteristics, develop user personas, and specify user tasks.

Task 9: Developing a digital library prototype for ETDs: Instead of presenting ETDs solely as whole documents, we will reduce readers' effort by generating readable summaries with the most relevant content. Addressing requirements established in the previous task, we will construct a prototype to demonstrate how appropriate services can improve the user experience, and ultimately increase the capacity of libraries to provide access to ETD content. An earlier research project at Virginia Tech used data obtained from the Networked Digital Library of Theses and Dissertations (NDLTD) Union Catalog to develop a prototype of an enhanced document browsing system called ETD-Enhance [81]. Its focus was to demonstrate the accuracy of extracting correct information from documents. Our focus will be on usability and the suitability of various services for addressing users' needs when interacting with ETD collections. We will base our design decisions on empirical evidence obtained from user analysis, and make these design decisions concrete by building them into a prototype.

2.3 Evaluation

Evaluations are performed for each group of tasks towards a deliverable. For ***Tasks 1–3***, the segmentation and extraction approaches will be evaluated on each field that is extracted from the ETDs (e.g., title, author, thesis committee members, chapters), based on the ground truth (compiled manually or from the metadata). For each field i, we will calculate the accuracy A_i, considering the number of ETDs in which field i is correctly extracted. The citation parser will be evaluated using an annotated corpus of about 1,000 manually curated representative citation strings using the precision, recall, and F_1-measures. For ***Tasks 4–5***, the evaluation will be performed using the manually labeled ground truth. Because each ETD chapter may have multiple labels (subject categories), we will calculate precision, recall, and F_1-measure for each label. ***Tasks 6–7*** will be evaluated using both human evaluators and commonly used metrics. We will have at least 2 human assessments of each generated summary. We will also use ROUGE-N factors (Recall-Oriented Understudy for Gisting Evaluation), defined as the overlap between the generated vs. ground truth summaries ($N = 1$ for unigrams and $N = 2$ for bigrams). If we view summarization as a translation problem, we can use the BLEU metric (Bilingual Evaluation Understudy Score) [65, 68]. ***Task 8*** evaluation will include a survey, that will be completed by a diversity of respondents, using Likert scales and asking for constructive comments. ***Task 9*** will be evaluated by user studies with graduate students, faculty researchers, and professional librarians to evaluate the usefulness of various methods of searching, browsing, visualizing, and interacting with ETDs. In our studies, we will use pre-screening to maximize diversity of samples to include participants of divergent age groups and gender. All planned user study interactions and questions will be reviewed and approved by the Virginia Tech IRB.

2.4 Project Management

2.4.1 Project Team

This project will proceed at Virginia Tech (VT) and Old Dominion University (ODU). Responsibility for management, research, and dissemination will be shared between PI Ingram and Co-PIs Fox and Wu. The project will hire one graduate research assistant at VT and one at ODU. The graduate assistants will undertake data

collection, labeling, programming, user studies, analysis, development, and experimentation. Working toward doctoral degrees, they also will assist with publishing and other dissemination activities.

Mr. Ingram will serve as Principal Investigator. Ingram is Assistant Dean of University Libraries at VT and Director of the Libraries' IT division. He received an M.S. in Library and Information Science from the University of Illinois at Urbana-Champaign in 2008. Since then, he has been involved in projects and services related to scholarly communication, digital preservation, and repository services, including managing the Library's ETD program at Illinois. Ingram also is pursuing a Ph.D. in Computer Science (with dissertation focused on ETDs) and is an active member of the Digital Library Research Laboratory (DLRL). He will direct this project, supervise a graduate research assistant, and lead the work on summarization as well as other services relevant to libraries.

Dr. Fox will serve as a Co-PI. At VT, Fox directs DLRL, and is active in both the Center for HCI and the Discovery Analytics Center. He is the Executive Director, Chairman of the Board, and founder of the Networked Digital Library of Theses and Dissertations (NDLTD); see its support letter. Fox is a Fellow of ACM and a Fellow of IEEE, cited for work on digital libraries and information retrieval. His papers have been cited more than 17K times (Google Scholar) with an h-index of 58. At VT since 1983, and having started work on ETDs in 1988, he has taught courses on information retrieval, big data, and computational linguistics (NLP). Dr. Fox's project responsibilities include ETD access, discovery, and summarization.

Dr. Wu will serve as a Co-PI for this research project. At ODU, Wu works with the Web Science Digital Library (WS-DL) group to research mining of scholarly big data, including millions of academic papers. Wu has been the tech leader of CiteSeerX since 2013, and published 30+ peer-reviewed papers on document classification, information extraction, and other related topics. Dr. Wu will be responsible for data acquisition and analysis. He will supervise a graduate research assistant to: collect data; populate ETD databases; perform analysis, information extraction, and classification; and prepare data for downstream training and searching.

2.4.2 Advisory Board

To aid us in evaluation and performance management, we have assembled an advisory board to meet with the project team regularly, evaluate our progress, and keep us on track. Advisory board members are Dr. Daniel Alemneh, Head, Digital Curation Unit, University of North Texas Libraries; Dr. Suzie Allard, Professor, School of Information Sciences, University of Tennessee Knoxville; Dr. Cornelia Caragea, Associate Professor at the University of Illinois at Chicago; Dr. George Fowler, University Librarian at ODU; Dr. C. Lee Giles, Professor at Penn State and PI of the CiteSeerX Project; Gail McMillan, Director, Scholarly Communication and Professor, University Libraries, Virginia Tech; Gabrielle V. Michalek, Program Director, Connected Scholarship, Carnegie Mellon University Libraries; Roxanne Shirazi, Dissertation Research Librarian, The Graduate Center, CUNY; Karen Vaughan, Digital Initiatives Librarian at ODU; and Dr. Zhiwu Xie, Chief Strategy Officer and Professor, University Libraries, Virginia Tech. The advisory board brings to the project expertise in ETDs, information science, library science, artificial intelligence, and digital library infrastructure. The advisory board will play an active role in our project. We will consult with them regularly. We plan to hold annual plenary meetings for all members. Otherwise, we will hold smaller monthly meetings, targeted on a particular topic.

2.4.3 Project Support

The project team will work closely with the Statistical Applications and Innovations Group at Virginia Tech, regarding experimental design, data analytics, interpretation of results, and drawing appropriate conclusions.

2.5 Project Dissemination and Sustainability

We plan to release all software created by this project as open-source, hosted in VT University Libraries' Git repositories. We will share our research findings through appropriate conferences and journals, such as ALA,

ACRL, ASIS&T, JAIST, CRL, JCDL, IJDL, the Open Repositories Conference, and the International Symposium on ETDs. We will also introduce interesting data and software to institutional libraries by hosting workshops and tutorials co-located with conferences. Co-PIs Fox and Wu will incorporate segmentation, summarization, and other components into graduate and undergraduate courses, such as Information Retrieval (Wu) and Digital Libraries (Fox). Other research output generated by this project, including articles, white papers, reports, presentations, derived data sets, software, and other digital products, will be preserved and made publicly available through the VTechWorks and VTechData repositories at VT.

3 National Impact

The proposed research project will have a national impact by investigating ETDs: (1) to determine how to identify and extract key knowledge, (2) to add value to ETD libraries using techniques from natural language processing (NLP) and machine/deep learning, and (3) to enhance services for interacting with ETD content. Our research findings will generate new tools, models, and practices that will be widely used, adapted, or replicated by libraries engaged in providing access to book-length documents, especially ETDs. By investigating key questions related to ETD services, we aim to have a significant national impact on the capacity of libraries to provide access to the knowledge buried in these long documents. As such, we aim to provide libraries with:

- techniques, best practices, and software for segmenting long documents into constituent parts — chapters, sections, figures, tables, citations, and reference lists — from both born digital and page image documents;
- an automated summarization tool for producing a short, abstractive summary for a given ETD chapter by employing computational linguistics (NLP) and deep learning;
- an implementation of supervised learning models for classifying ETDs using the LCC taxonomy;
- a digital library prototype, improving the user experience with ETD collections and ultimately increasing the capacity of libraries to provide access to ETD content; and
- valuable derived datasets, positioned to stimulate other research studies focused on long documents.

IMLS has recognized the need to equip libraries and librarians with skills and tools for intensive computational use of their digital collections.[4] Shifts in thinking are emerging about the potential of full-text analysis of library collections, in response to recent advances in natural language processing/computational linguistics and deep learning. Our project brings to bear cutting-edge computer science and machine/deep learning technologies to advance discovery, use, and potential for reuse of the knowledge hidden in the text of books and book-length documents. By focusing on libraries' ETD collections (where legal restrictions from book publishers generally are not applicable), our research will open this rich corpus of graduate research and scholarship, leverage ETDs to advance further research and education, and allow libraries to achieve greater impact.

Equipping libraries with the capability to automatically identify, segment, extract, and summarize key parts from ETDs would lead to a systemic change in the way ETDs and the knowledge they contain are discovered, used, and reused. Facilitating expanded access to, and use of, this important corpus of scholarly content is strongly in the national interest. The ability to computationally process prior ETDs and leverage them to advance further research and education will increase the use of ETD collections and allow libraries to achieve greater impact. Further, methods developed for ETDs can be extended to aid related work with other types of long documents.

[4] https://www.imls.gov/news-events/upnext-blog/2016/10/computational-librarianship-and-data-driven-library-practice

Schedule of Completion

	Project Year 1											
Activities and Milestones	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul
Administrative work												
Launch project												
Plan plenary advisory board meetings												
Advisory board teleconference (monthly, most with different subgroups)												
Hire graduate research assistants												
Create and maintain a project website												
Evaluate project progress												
Review data management plan												
Annual reporting												
Research Area 1: Document analysis and extraction (lead by Wu*)												
Task 1: Compiling ETD sample												
Task 2: Building ground truth for ETD segmentation and extraction												
Task 3: Researching on ETD segmentation and extraction models												
Research Area 2: Adding value (lead by Ingram and Fox*)												
Task 4: Building ground truth for ETD (chapters) topical classification												
Task 5: Building topical classification models for ETDs and their chapters												

*All investigators will collaborate and make contributions.

	Project Year 2											
Activities and Milestones	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul
Administrative work												
Advisory board teleconference (monthly, most with different subgroups)												
Hire graduate research assistants												
Create and maintain a project website												
Evaluate project progress												
Review data management plan												
Annual reporting												
Research Area 1: Document analysis and extraction (lead by Wu*)												
Task 3: Researching on ETD segmentation and extraction models												
Research Area 2: Adding value (lead by Ingram and Fox*)												
Task 4: Building ground truth for ETD (chapters) topical classification												
Task 5: Building topical classification models for ETDs and their chapters												
Task 6: Building ground truth for ETD chapters summarization												
Task 7: Researching on summarization models for ETD chapters												
Research Area 3: User services (lead by Ingram and Fox*)												
Task 8: Identifying helpful services for those working with ETDs												

*All investigators will collaborate and make contributions.

	Project Year 3											
Activities and Milestones	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul
Administrative work												
Advisory board teleconference (monthly, most with different subgroups)												
Hire graduate research assistants												
Create and maintain a project website												
Evaluate project progress												
Review data management plan												
Annual reporting												
Research Area 2: Adding value (lead by Ingram and Fox*)												
Task 6: Building ground truth for ETD chapters summarization												
Task 7: Researching on summarization models for ETD chapters												
Research Area 3: User services (lead by Ingram and Fox*)												
Task 8: Identifying helpful services for those working with ETDs												
Task 9: Developing an ETD digital library prototype												

*All investigators will collaborate and make contributions.

DIGITAL PRODUCT FORM

Introduction

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (e.g., digital content, resources, assets, software, and datasets). The products you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions

All applications must include a Digital Product Form.

☐ Please check here if you have reviewed Parts I, II, III, and IV below and you have determined that your proposal does NOT involve the creation of digital products (i.e., digital content, resources, assets, software, or datasets). You must still submit this Digital Product Form with your proposal even if you check this box, because this Digital Product Form is a Required Document.

If you ARE creating digital products, you must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

Part I: Intellectual Property Rights and Permissions

A.1 What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? Who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

Intellectual property of digital products will remain with the authors, in accordance with Virginia Tech University Policy #13015 (http://www.policies.vt.edu/13015.pdf) which states that the project leader is expected to manage the university's ownership of research results and digital products in the ways that best advance the standard routes of dissemination for that particular field. We will work with the University Libraries to ensure research material and results, including software, digital notebooks, and files that are germane to the veracity and validity of the research claims are preserved. In doing so, we will grant Virginia Tech a permanent non-exclusive license to host and preserve and disseminate digital products with a Creative Commons license.

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

According to University Policy 13000 and 13015 (http://www.policies.vt.edu/13000.pdf; http://www.policies.vt.edu/13015.pdf), Virginia Tech asserts its rights to the results of research funded in any part with university resources. For traditional results of academic scholarship (e.g., white papers, reports), the presumption of ownership is to the author(s). All dissemination of the traditional digital products will be open access by default. All research datasets will be made openly available as well. Finally, it is our intention to release all software we create as open source. However, the university may claim ownership in accordance with policy 13000. If such a case arises, we will advocate for open access.

A. 3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

Although we only plan to use publicly available electronic theses and dissertations (ETDs) for our study, we will consult with repository managers before using any of their content. We will not redistribute these ETDs or provide access to their content except by linking to their home repositories.

If user studies are not deemed exempt by the Institutional Review Board (IRB), will collect PII only on physical consent forms, and this information will be kept in a locked file drawer in a locked office accessible only to the PI. All data collected will be deidentified at collection. PII will only be linkable to the data collected by a codebook that will be kept with the consent forms.

Part II: Projects Creating or Collecting Digital Content, Resources, or Assets

A. Creating or Collecting New Digital Content, Resources, or Assets

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and the format(s) you will use.

We will manually label/annotate at least 200 ETDs to create ground truth for segmentation, extraction, classification. We will manually create a corpus of summarizations of at least 100 ETD chapters, in addition to around 300k news reports and summaries (for transfer learning) adopted from Google DeepMind. Those files are in TXT with UTF-8 encoding. We anticipate generating descriptive metadata for ETDs (XML, JSON), citations (RIS, BibText), and prototype user interfaces (HTML, JavaScript). We will create source code for software implementation in Python, Java, and Scala. We will use MS Office and Latex to write papers (Word, PDF) and make presentations (PPT).

A.2 List the equipment, software, and supplies that you will use to create the content, resources, or assets, or the name of the service provider that will perform the work.

For the majority of the day-to-day work, we will use desktop and laptop computers provided by our universities. We will use the Virginia Tech DLRL deep learning server and Hadoop cluster, a hardware cluster provided by the University Libraries, as well as high-performance computing cyberinfrascture provided by VT's Advanced Research Computing group. Software used for this project will include standard desktop computing applications (word processors, spreadsheets, and presentation software), text editors, and integrated development environments (IDEs) for software and web development. We will primarily use Overleaf for LaTex editing and collaboration. We will use VT and ODU institutional Google Drive for storage and collaboration while we work.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to use, along with the relevant information about the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

We will use ETDs for training, testing, and evaluating the algorithms and prototype software we create. We anticipate the majority of these will be in PDF, but we expect the quality will vary widely. We will harvest metadata which describe these ETDs as XML and in some cases JSON or YAML. Likewise, we expect the metadata quality to vary. Our intention is to see how effective machine learning algorithms are for a wide variety of ETD and metadata quality. The software we create will be mainly Python, Java, or Scala, and will be written in plain text. Reports, papers, and presentations will be created with Microsoft Office tools, Google Docs, and/or LaTex. We plan to convert them to PDF/A for preservation and dissemination. We plan to extract figures from PDF files and store them in PNG format, the dimensions of which are determined by their sizes in the original PDFs.

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan. How will you monitor and evaluate your workflow and products?

Segmentation and extraction approaches will be evaluated on each field that is extracted from the ETDs, based on the ground truth compiled (manually or from the ETD metadata). We will calculate the accuracy, defined as the number of ETDs in which fields are correctly extracted over the sample size. Citation extraction will be evaluated using standard precision, recall, and F1 score (or f-measure). For evaluating summaries, we will use traditional ROUGE-N factors, defined as the overlap between the generated vs. ground truth summaries. We will also use human experts to evaluate summary quality. User services will be evaluated by conducting a user interaction study.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

Datasets selected for sharing will be made accessible through the VTechData repository (https://data.lib.vt.edu/) managed by the University Libraries at Virginia Tech. VTechData highlights, preserves, and provides access to data generated at Virginia Tech. The system relies on item and dataset level metadata as the primary building block to data discovery, access, and reuse. Traditional research output (e.g., white papers, reports) will be preserved and made accessible through VTechWorks (https://vtechworks.lib.vt.edu/) managed by the University Libraries at Virginia Tech. In both repositories, we plan to make our research Open Access with a Creative Commons Attribution License.

C. Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata. Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

We will use Dublin Core, METS, MODS, and PREMIS metadata schemas, as supported by the repositories (VTechWorks, VTechData). As part of our report/white paper, we will include project documentation, a user manual, and a developer manual explaining how to use our software as well as information for those who will maintain or enhance our work (PDF, Word/LaTex, HTML).

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

Metadata stored in the institutional repositories (VTechWorks, VTechData) is maintained by repository administrators and preserved according to Virginia Tech University Libraries preservation protocols and procedures.The University Libraries is dedicated to maintaining three preservation copies for each object for a minimum of five years, following digital preservation best practices. Preservation copies are deposited into local storage VT Archive and one subscription storage destination, either APTrust (http://aptrust.org/)) or the MetaArchive Cooperative (https://metaarchive.org/) both of which employ proven digital preservation strategies to monitor and maintain content.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

VTechWorks is OAI-PMH and OAI-ORE compliant. Additionally, VTechData pushes metatdata to DataCite for discoverability when minting DOIs. We will announce our work with the use of social media, professional listservs and forums such as CNI, ACRL, LITA, and CLIR.

D. Access and Use

D.1 Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

Papers and reports will be openly available online in non-proprietary standard formats in repositories that do not require authentication, and in formats that are ADA compliant. We plan to make examples of ETD summaries (both manually and automatically made) accessible via VTechData, and reports will be made accessible via VTechWorks. The repositores are indexed by all the major search engines so digital products hosted there will be easily discoverable.

D.2 Provide the name(s) and URL(s) (Uniform Resource Locator) for any examples of previous digital content, resources, or assets your organization has created.

Electronic Theses and Dissertations: Progress, Issues, and Prospects (http://hdl.handle.net/10919/9198)
Integration of VT ETD-db with Banner (http://hdl.handle.net/10919/19829)
OAI and ODL: Building Digital Libraries from Components (http://hdl.handle.net/10919/52811)
Viral Networks: Connecting Digital Humanities and Medical History (V2) (https://doi.org/10.7294/284t-bf10)

Part III. Projects Developing Software

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) it will serve.

We plan to create proof-of-concept software for 1) segmenting ETD documents, and identifying and extracting key parts, 2) an automated summarization tool for producing abstractive summaries of ETD chapter text, 3) supervised learning models for classifying ETDs, and 4) prototype digital library for demonstrating and evaluating the prior items. The primary audience will be academic researchers, teachers and learners, librarians and library patrons.

A.2 List other existing software that wholly or partially performs the same functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

Existing software for academic document segmentation and extraction: GROBID, CERMINE, Science Parse; citation extraction: ParsCit, Neural ParsCit; figure/table extraction: pdffigures, deepfigures, and pdfimages; text classification: VDCNN; text summarization: NATS; repository: DSpace, Fedora. The key difference in what we intend to create is that it will be designed specificailly for book-length documents, focusing on ETDs. We anticipate our proof-of-concept software will lead to improvements in the tools and infrastructure used by libraries for their ETD programs, and the software we intend to create could be used by researchers in ETD meta-analyses (across IRs).

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software and explain why you chose them.

We use Python, Java, and Scala since they are the most widely used programming languages for machine learning. We plan to use deep learning frameworks such as PyTorch and TensorFlow, as well as popular software and libraries such as Gensim, Keras, scikit-learn, Pandas, NumPy, and pythonrouge among others.

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

The segmentation and extraction software we intend to create may encapsulate or extend existing software like GROBID, Neural ParsCit, an Deepfigures. The classification software we intend to create will be based on existing ones (e.g., VDCNN) extended to be suitable for ETDs chapters.The digital library service we plan to create will be proof-of-concept to illustrate our techniques. When applicable, we will include in our documentation suggestions for how libraries can implement our methods and techniques into their existing repositories or digital libraries for ETDs.

B.3 Describe any underlying additional software or system dependencies necessary to run the software you intend to create.

We anticipate the software we create will be run on a Linux/UNIX system and may be run in a virtual machine environment or in the cloud. Python, Java and/or Scala will be required to run the software. Documentation of the software stack used in our research will be included with the software.

B.4 Describe the processes you will use for development, documentation, and for maintaining and updating documentation for users of the software.

We plan to use a Git repository (we have institutional accounts for GitHub or GitLab) for authoring, publishing, and maintaining documentation. We also plan to include project documentation, a user manual, and a developer manual in a report which will be preserved and made accessible through VTechWorks (https://vtechworks.lib.vt.edu/) managed by the University Libraries at Virginia Tech.

B.5 Provide the name(s) and URL(s) for examples of any previous software your organization has created.

The University Libraries at Virginia Tech's public GitHub group (https://github.com/vtul) contains about 60 repositories for software projects developed at the Libraries. Likewise, the ODU Web Science and Digital Libraries Research Group's GitHub (https://github.com/oduwsdl) contains about 37 repositories.

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you intend to create, and what conditions will you impose on its access and use? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

Unless prohibited by the Virginia Tech Intellectual Properties, Inc. (VTIP), we plan to release source code for the software we develop under the BSD 3-Clause "New" or "Revised" License, in accordance with IMLS expectations. However, the university may claim ownership in accordance with university policy 13000 (http://www.policies.vt.edu/13000.pdf). If such a case arises, we will advocate for open access with VTIP.

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

We plan to host make our software and source code available to the public on the University Libraries at Virginia Tech's public GitHub (https://github.com/vtul). Additionally, we plan to archive our software and documentation in Virginia Tech institutional repositories (VTechWorks, VTechData).

C.3 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:

GitHub: Virginia Polytechnic Institute and State University Libraries

URL:

https://github.com/vtul

Part IV: Projects Creating Datasets

A.1 Identify the type of data you plan to collect or generate, and the purpose or intended use to which you expect it to be put. Describe the method(s) you will use and the approximate dates or intervals at which you will collect or generate it.

At the outset of our project, we plan to collect over 50,000 publicly available ETDs, but we will not share, redistribute, or store them long-term. We will use the ETDs to create, train, and evaluate computational models, techniques, and proof-of-concept software. We plan to collect, create, and store metadata for the ETDs we use in our research along with persistent identifiers for locating ETDs in their home repositories. For user studies, we plan to gather demographic info from participants, but no personally identifiable information. We will gather data about the user experience of a test web portal. We also anticipate asking user study participants to conduct qualitative comparison of summary texts.

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

All planned user study interactions and questions will be reviewed and approved by the Virginia Tech IRB.

A.3 Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

By default, we do not plan to collect any PII, confidential information, or proprietary information, with exceptions explained below.

A.4 If you will collect additional documentation, such as consent agreements, along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

If required by IRB, we will collect consent agreements from user study participants. We will follow the Virginia Tech Procedures for Handling Confidential Information at Virginia Tech (https://www.research.vt.edu/compliance/procedures-handling-confidential-information-virginia-tech.html). We will collect PII only on physical consent forms, and this information will be kept in a locked file drawer in a locked office accessible only to the PI. All data collected will be deidentified at collection. PII will only be linkable to the data collected by a codebook that will be kept with the consent forms.

A.5 What methods will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

We plan to harvest metadata for ETDs via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) using free and open software. We plan to collect user study data using online form creation software such as Google Forms or Qualtrics. Datasets will be collected or generated using computer software, stored in computer file systems, and accessible using common image viewing (for figures/tables) and text editing (for textual files) software.

A.6 What documentation (e.g., data documentation, codebooks) will you capture or create along with the dataset(s)? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

We plan to issue an interal identifier for cross referencing ETDs, their metadata, and any other assocated files. We will store this data in a spreadsheet. If signed concent forms are required, PII will only be linkable to the data collected by a codebook that will be kept with the consent forms. User study data will be stored CSV format, exported from either Google Form or Qualtrics.

A.7 What is your plan for archiving, managing, and disseminating data after the completion of the award-funded project?

Datasets selected for sharing will be made accessible through VTechData managed by the University Libraries at Virginia Tech. The system relies on item and dataset level metadata as the primary building block to data discovery, access, and reuse. Libraries' personnel provide advice and some assistance on organizing, documenting and otherwise curating research data to improve its discoverability and reusability. The original and curated datasets are archived according to best practices developed by the Libraries (including conversion of file formats to non-proprietary options where appropriate) and accepted by the disciplinary communities. VTechData also provides researchers persistent digital object identifiers and data citations for published datasets. Researchers can assign licenses according to their interests.

A.8 Identify where you will deposit the dataset(s):

Name of repository:

VTechData

URL:

https://data.lib.vt.edu/

A.9 When and how frequently will you review this data management plan? How will the implementation be monitored?

This data management plan will be integral to the project. As such, it will be reviewed at the completion of every project milestone. The project director will be responsible for execution and monitoring of this data management plan.