A CASE STUDY

How Have Trends in Graduate Research Evolved?

Using TDM Studio to Explore Trends in Theses and Dissertations,
Virginia Polytechnic Institute and State University

about.proquest.com

To talk to the sales department, contact us at
1-800-779-0137 or **sales@proquest.com**.

ProQuest®
Part of **Clarivate**

William Ingram is the Assistant Dean of University Libraries at Virginia Polytechnic Institute and State University. His research areas of focus are building better digital libraries, machine learning and looking at collections as data.

Introduction

Virginia Tech University Libraries, in collaboration with Virginia Tech Department of Computer Science and Old Dominion University Department of Computer Science, received grant funding from IMLS for a 3-year project, the goal is to bring computational access to book-length documents, demonstrating that with electronic theses and dissertations (ETDs). The project is motivated by library and community needs to access the valuable content and scientific data found in dissertations and theses. There is currently a lack of models that provide computational access to these long documents and nationwide open access services for ETDs generally function at the metadata level. Much important knowledge and scientific data lie hidden in ETDs, and the goal is to develop better tools to mine the content, facilitate the discovery, and reuse of these important components. This case study outlines one specific output from the full work outlined in the grant.

The Project

Ingram and his team received an IMLS grant to analyze the evolution of graduate research topics over time, the ways different topics and disciplines overlap and how interdisciplinarity has developed in graduate research. One of the problems Ingram looks to solve through his research is to effectively extract and analyze book-length documents such as dissertations and define methods for summarizing them to open the knowledge hidden in this form of scholarship. According to Ingram, the ProQuest Dissertations & Theses Global "corpus is special" and "possesses unique properties" particularly suitable for his research inquiry.

Through his relationship with ProQuest's Dissertations team, Ingram became acquainted with a new tool recently launched by ProQuest that would simplify his ability to plumb the database – spanning 1.3 million dissertations during a designated time frame (2000-2018) – and efficiently analyze his findings. A 3-month pilot program of ProQuest's Text and Data Mining (TDM) Studio was arranged and along with his team, Ingram set out to explore the evolution of topics in graduate research.

Methodology

For this project, Ingram determined they would need to narrow the 1.3 million documents to full-text XML files with department metadata. Within the remaining 600,000 documents, they focused on the top 20 departments/majors with the most robust quantities of dissertations and organized them into batches by years and department. Top terms found in titles and abstracts were used to intuit research topics.

First, term frequency-inverse document frequency (TF-IDF) was used to calculate 2-and 3-word phrases to identify terms and determine paper topics within the corpus; however, this technique yielded too many irrelevant results. They switched to an entity recognition tool, Wikifier, to disambiguate terms using Wikipedia.

Once research topics were determined in each topic or major, Ingram and his team set out to determine how frequently these terms were used during different time intervals (2001-2005, 2006-2009, 2010-2013, 2014-2018). They plotted the highest terms from each department and time interval, then plotted terms across multiple departments to explore how research topics overlap and evolve over time.

> *"I would definitely recommend TDM Studio to my peers interested in mining academic documents. Frankly, the best part of the TDM Studio service is the people and working with the ProQuest team was a joy. The customer service was amazing."*

Results

As an example, Ingram demonstrated his team's findings in the computer science and biology departments. Using bubble graphs, he showed how fewer research topics were the focus of dissertations in computer science during the earlier time periods than in more recent years, when the variety of topics expanded exponentially.

He also showed how particular topics that were more frequent in earlier dissertations became less popular with researchers over time, while emerging topics like "social networks," "machine learning" and "big data" appeared and gained in frequency.

Ingram then revealed how research topics for dissertations in biology evolved throughout each time interval, noting with surprise that the term "climate change" didn't appear on his graph until 2010-2013. Additionally, when he overlaid search terms for both computer science and biology, he discovered "climate change," as well as "gene expression" and "t cell" among those that appeared as research topics in both departments in the later time intervals. Ingram noted, contrary to his expectations, that the majority of interdisciplinary research topics spanning the two majors related to biology, illustrating the influence of the discipline in computer science.

Likewise, Ingram mentioned examples of overlaps in other interdisciplinary areas, such as math and economics, pointing out how topics spanning both departments gained in frequency, proving how interdisciplinarity in graduate research has increased over time.

The findings Ingram and his team discovered "would not have been possible without ProQuest" he concluded. ProQuest's corpus of digitized dissertations provided more data than could have been collected from individual repositories, he explained, and TDM Studio facilitated thorough investigation and analysis to draw such compelling conclusions.

About TDM Studio

ProQuest's workflow solution for text and data mining is designed for research, teaching and learning. TDM Studio provides access to sought-after content including current and historical newspapers, primary sources, scholarly journals, and dissertations and theses. It empowers researchers, students and faculty to analyze documents by uncovering connections and patterns that lead to career-defining discoveries.

References

Ingram, W. A. (n.d.). *Institute of Museum and Library Services*. Opening Books and the National Corpus of Graduate Research. Retrieved February 6, 2023, from https://www.imls.gov/sites/default/files/project-proposals/lg-37-19-0078-19-full-proposal.pdf

Learn more at www.proquest.com/go/tdm-studio
or contact your ProQuest representative today.

about.proquest.com

To talk to the sales department, contact us at
1-800-779-0137 or **sales@proquest.com**.

LTR/LB-KS/02-23

ProQuest® Part of **Clarivate**