# Text and Data Mining with Dissertations

## The Evolution of Research. Right at your fingertips.

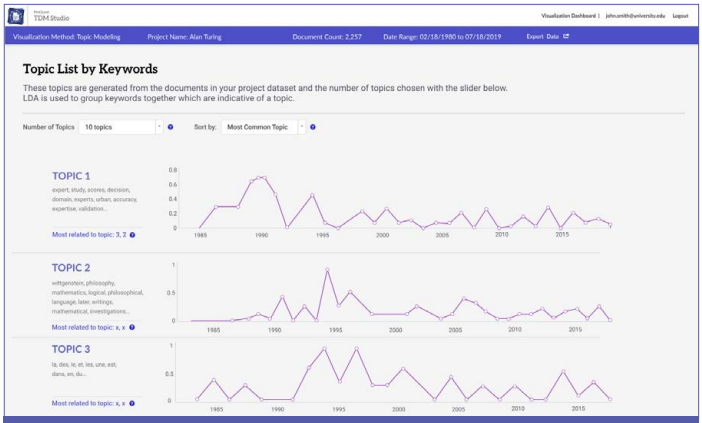
**ProQuest Dissertations & Theses Global | ProQuest TDM Studio**

Many of the most forward-thinking and world-changing discoveries come from the universities that set their sights high with a vision that is dedicated to the transformative power of ideas. ProQuest is dedicated to supporting the success of these institutions. We believe the value of individual research is enhanced exponentially when it is considered as part of a whole.

When historical and current research is analyzed together, researchers make discoveries that impact society and improve human lives. ProQuest is committed to connecting the graduate work from colleges and institutions from one side of the globe to the other. With cutting-edge content and data insights, ProQuest provides students and researchers with access to a full corpus of dissertations and theses content so they can see the evolution of research. Through the mining of the ProQuest Dissertation & These Global database, scholars can identify trends and generate groundbreaking hypotheses.

The ProQuest Dissertations & Theses (PQDT) Global database has over 5 million records, from 1637 to present day. In addition, nearly 200,000 new graduate works are added every year from all over the world, making it the perfect resource to use to understand relationships between research studies. PQDT Global is intended for use by scholars, institutions, corporations, and the government to help enhance their research. The ProQuest editorial team converts every full text record in PQDT to XML and carefully enhances and indexes the metadata, providing an unparalleled resource for text and data mining research.

The humanities and social sciences influence almost every aspect of daily life; from economic policy, to health and social policy, and from the environment and energy to technology and innovation. Understanding these trends and assessing the impacts on daily life takes skilled analysis — and given the diversity and complexities of the data, such analysis requires detailed modeling. By using text and data mining, you can extract large quantities of data and recombine them into patterns and topics to help the researcher further enhance the examination of the scholar's topic of choice.




*TDM Studio is a ProQuest tool that supports the analysis efforts researchers are already pursuing within the PQDT Global database.*

proquest.com

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

ProQuest®

# How Have Trends in Graduate Research Evolved?

**Research Case Study:** *Using TDM Studio to Explore Trends in Theses and Dissertations*, Virginia Polytechnic Institute and State University

**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.

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.

> ***"The TDM Studio enhanced our research output by allowing us access to the huge amount of data in the Dissertations & Theses Global collection. Our work benefits from both size and diversity of the data — the collection is a representative sample of the graduate research that's happening all over the country, all over the world. That's one of the reasons why I study ETDs: they're so diverse. The wide variety in style, formatting, and discipline-specific jargon makes mining ETDs enormously challenging, but that's what makes the work interesting."***
>
> — *William Ingram, Assistant Dean of University Libraries at Virginia Polytechnic Institute and State University*

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.

*William Ingram, Assistant Dean of University Libraries at Virginia Polytechnic Institute and State University. Photo credit: Liz McVoy*

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.

## 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.

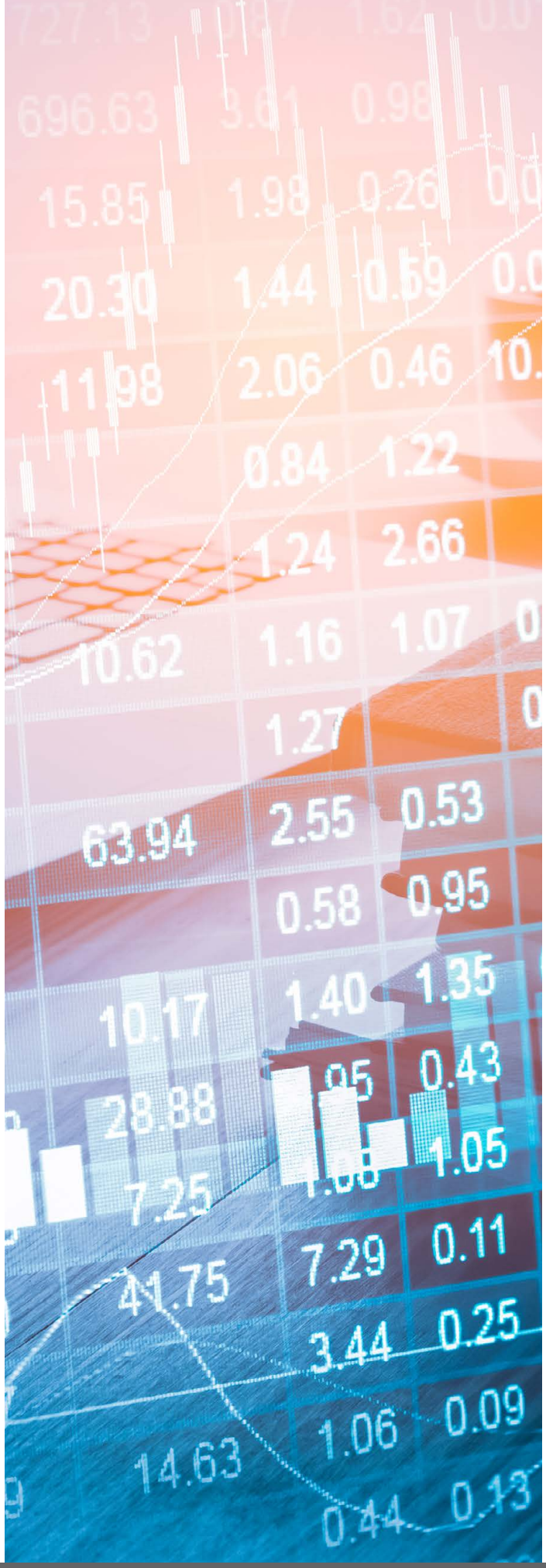
He also showed how particular topics that were more frequent in earlier dissertations became less popular with researchers over time, while emerging topics like "social networks," "machine learning" and "big data" appeared and gained in frequency.

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.

> ***"I would definitely recommend the 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."***
>
> — *William Ingram, Assistant Dean of University Libraries at Virginia Polytechnic Institute and State University*

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.



# Analyzing the Past to Understand the Future

**Research Case Study:** *Sociology of Science, Diversity Effects, and Research Innovation,* Stanford University

In this day and age, researchers need rich longitudinal datasets that accurately reflect a population of subjects. With the digitalization, many research professors can now tap into datasets like these that weren't available 20-30 years ago. Professors, like Daniel McFarland, a professor of Education at Stanford University, see resources like ProQuest Dissertations & Theses Global as a nice example of such high quality data and use it in their research.

McFarland studies the social and organizational dynamics of educational systems like universities and disciplines. McFarland is currently engaged on several different projects including writing a textbook on Social Network Analysis in R with Craig Rawlings, Jeff Smith, and James Moody. He also has generated a line of articles on the sociology of science, diversity effects, and research innovation. That's where using a product like PQDT Global has helped McFarland. Using PQDT Global and other resources, he's able to make his research as well-rounded as possible for publication.

> ***"What I like about PQDT is that it affords us a sizeable sample of beginning researcher's careers. Using National Center of Education Statistics census of all US PhDs, we can weight the PQDT sample and make inferences about all US PhDs. That's pretty great as most datasets on research reflect found data, and we can only guess how representative those samples are. With PQDT, we can argue they reflect a known population."***
>
> *— Daniel McFarland, Professor of Education and Sociology, Stanford University*

According to McFarland, the standards for being published have risen so researchers have to check their models and inferences many different ways. McFarland also stated that "compared to 20 years ago, research standards for publication are higher. That's a good thing, but it also means we have to find rich longitudinal information that can be anchored in known populations – then our inferences make sense."

McFarland and his research team turned to PQDT Global because they wanted to study scholar careers and they wanted data that accurately reflected the pool of potential scholars. "What I liked about PQDT Global is that it wasn't biased like some journals are and it's evenly representative across disciplines. We don't get only journal science, or conference proceedings, or books, but rather see people of all fields who write theses. PQDT also doesn't select on the outcome. By this I mean most studies look at faculty or persons who got PhDs and succeeded in becoming faculty. With PQDT we have recourse to following all PhDs into various jobs and out of academia – so the entire pool of potential faculty." McFarland and his team needed a data set of at least 30 years and all potential future faculty to develop models of "successful" academic careers. This is something that PQDT can offer with content that dates back to the 1700s.

"ProQuest Dissertations & Theses Global is a wonderful corpus for scholars to work with, and especially if they are interested in following the pool of potential researchers and their careers," McFarland stated.

# ProQuest TDM Studio

**TDM Studio** is a pathway to discovery for users of all levels to quickly spot trends and generate insights. It provides two analysis tools, Visualization for discoveries without coding and a Workbench for researchers that want to use their own coding methodologies.

### The Workbench

For those accustomed to using coding for text analysis, it provides programmatic access to ProQuest content.

### Visualizations

For those wishing to see trends and make connections quickly, it interrogates ProQuest content using data visualizations.

TDM Studio leverages the power of a library's content, such as ProQuest Dissertations & Theses Global to help researchers at all levels make new connections.

### Current users are saying...

- *"The interesting thing is that students are learning data literacies from lots of different everyday sources, so they were able to pick up TDM Studio almost on the fly."*

  *— Marco Duranit, Sr. Lecturer, The University of Sydney*

- *"TDM Studio provides streamlined access for projects similar in scope, allowing researchers like me to delve into historical text data in ways that would have been otherwise impossible."*

  *— Yuyang Zhong, Undergraduate Student Researcher, University of California, Berkeley*

# ProQuest Dissertations & Theses Global

**PQDT Global** is a singular repository of over 5 million works that the world's most prestigious universities contribute to each year, creating an ever-growing resource of emerging research to fuel innovation and new insights.

As the official repository of the Library of Congress, ProQuest aligns its editorial processing of each record to strict metadata quality guidelines. Every title is manually checked for completeness and accuracy to enhance the discoverability and text and data mining effectiveness.

### More Research Examples of Text and Data Mining with ProQuest Dissertations & Theses Global

- *David Zeitlyn, Daniel W. Hook used PQDT Global to explore and understand the dynamics of prestige in the academic hierarchy.* Perception, prestige, and PageRank, *PLOSONE, May 2019.*

- *Catherine Buffington, et.al used the PQDT Global data to verify dissertation publication and graduation against UMETRICS data and then analyze how STEM degrees correspond to career outcomes.* STEM Training and Early Career Outcomes of Female and Male Graduate Students: Evidence from UMETRICS Data Linked to the 2010 Census, *The American Economic Review, May 2016.*

- *The Committee on Institutional Cooperation used PQDT Global to study employment and earning outcomes of Ph.Ds.* Examining the Employment and Earning Outcomes of Ph.Ds. *Science magazine, 2016.*

The evolution of research is at your fingertips. Enhance your studies with a dataset you can't find anywhere else by using **ProQuest Dissertations & Theses Global** with **TDM Studio**. Reach out to one of our ProQuest specialists to learn how to trial both together!

## To sign up for a TDM Studio trial visit:

## https://go.proquest.com/TDMStudio-Trial/

proquest.com

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

7028 -LTR/JB-CS-KS/05-21