12-9-2021

ProQuest TDM Studio: A Text and Data Mining Solution

Anamika Megwalu

San Jose State University, anamika.megwalu@sjsu.edu

Anne Marie Engelsen

San Jose State University, annemarie.engelsen@sjsu.edu

Follow this and additional works at: https://scholarworks.sjsu.edu/faculty_rsca

Part of the [Library and Information Science Commons](#), and the [Other Computer Sciences Commons](#)

Recommended Citation

Anamika Megwalu and Anne Marie Engelsen. "ProQuest TDM Studio: A Text and Data Mining Solution" *The Charleston Advisor* (2021): 42-46. <https://doi.org/10.5260/chara.23.3.42>

This Article is brought to you for free and open access by SJSU ScholarWorks. It has been accepted for inclusion in Faculty Research, Scholarly, and Creative Activity by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

ADVISOR REVIEWS—STANDARD REVIEW

ProQuest TDM Studio: A Text and Data Mining Solution

doi:10.5260/chara.23.3.42

Date of Review: December 9, 2021

Composite Score: ★★ ★ 3/4

Reviewed by: **Anamika Megwalu**
Dr. Martin Luther King Jr. Library, San Jose State University

Anne Marie Engelsen
Dr. Martin Luther King Jr. Library, San Jose State University

<anamika.megwalu@sjsu.edu>

<annemarie.engelsen@sjsu.edu>

Abstract

TDM Studio is an integrated platform offered by ProQuest for data and text mining. TDM stands for text and data mining. This cloud-based, all-in-one innovative product is designed to offer researchers a clean interface with rights-cleared content, Jupyter notebook, and data visualization tools. As a result, researchers can now search ProQuest databases, create large datasets, import data to Jupyter notebook for analysis, and download results within a day.

Pricing Options

Initially, pricing was determined by the number of workbenches an institution purchased. However, ProQuest is currently offering institutions an unlimited number of work benches upon subscription. Considering this is a fairly new platform, the pricing model may evolve. According to ProQuest, pricing is determined by the size of the library or school, and by the number of prospective users. Interested libraries should contact ProQuest for pricing information specific to their institution.

Product Overview/Description

TDM Studio is a platform for data and text mining that was launched by ProQuest on April 14, 2020 (ProQuest, 2020). Researchers often face challenges at multiple levels when conducting studies that require data and text mining. These challenges include identifying

sources for data extraction, securing copyright permission, cleaning datasets, normalizing disparate data formats, dealing with gaps in coverage, finding appropriate software for analysis, as well as data storage and computing issues. TDM Studio was created to alleviate such issues by offering accessible full-text content to create datasets, embedded integration with Jupyter Notebook for data cleaning and data analysis, and a visualization studio for data geographic analysis and topic modelling. Jupyter Notebook, an online, open-source development notebook, is used to “create and share documents that contain live code, equations, visualizations, and narrative text” (Jupyter, 2021). TDM Studio is a cloud-based software that allows multiple users to have access to the workbench concurrently, making real-time collaboration easy. Additionally, there is no need for installing additional software and plug-ins.

Although current subscription to TDM Studio offers unlimited workbenches, institutions have to mediate access to the workbenches. That means institutions have to create a workflow and a permission protocol for users who are interested in having access to a TDM Studio workbench. The most common practice among institutions is to create a request form where users are asked to indicate name(s) and contact information of researcher(s), department affiliation, project scope, research questions, estimated research completion date, and the title of the workbench. Upon receiving a request, a designated employee of the institution asks ProQuest staff to create a workbench for the researcher or the research group. ProQuest representatives are swift in creating workbenches.

FIGURE 1 ProQuest TDM Studio User Interface

FIGURE 2 ProQuest TDM Studio Create Database Function

FIGURE 3 ProQuest TDM Geographic Visualization Analysis

User Interface/Navigation/Searching

TDM Studio has a clean and intuitive user interface that allows for novice text and data miners to easily navigate through the process. Once users log in, the TDM Studio landing page is the Workbench dashboard. From this dashboard, users can see a list of their existing datasets, create new datasets, and open a development environment through Jupyter Notebook (see Figure 1). At the top of the page, users can toggle between the Workbench dashboard and the Visualization dashboard.

Creating a dataset is relatively easy within TDM Studio. Once users click on the Create a Dataset button, a drop-down list allows them to create a dataset using either publication titles or ProQuest databases (see Figure 2).

If the user opts to search specific publications, they can either browse titles from the list provided or use a search to find the publications. Users should be careful while selecting publications as several titles have multiple entries that vary on source type, coverage years, and availability of full text. Once the publications have been selected, users can then perform a search for their desired content. ProQuest's familiar user interface ensures that searching for content is simple and straightforward for all users. As with other ProQuest databases, users can use Boolean terms, proximity operators, and other search

strategies to customize their search terms. A full list of search options is available from the ProQuest Platform LibGuide (ProQuest, 2021). Once the search is completed, users can then use several limiters to narrow down the results by date, full-text, document type, and more.

Once the dataset is compiled, users can assign a name and description before submitting the data for processing. The dataset will be listed on the Workbench dashboard as complete once it finishes processing, which can take several hours depending on the number of documents included in the dataset. TDM Studio can process one hundred thousand documents per hour; however, there have been instances where processing has been faster.

For those not familiar with using R or Python, the Visualization dashboard can provide data visualizations with limited features. Users can choose between geographic analysis (see Figure 3) or topic analysis (see Figure 4) for their projects. Geographic analysis provides an interactive map with instances displayed as bubbles of increasing size and users can export the data as GeoJSON or CSV files.

The topic analysis option finds common topics within the selected datasets by grouping key words with high instances within the dataset. TDM Studio's topic analysis displays the results using the Scikit-Learn's implementation of Latent Dirichlet Allocation (LDA). Users can create projects by searching terms of interest using allowed ProQuest search operators.

FIGURE 4 ProQuest TDM
Topical Visualization Analysis

TDM Studio is partially compliant with the Web Content Accessibility Guidelines v. 2.1 AA standard (ProQuest Support Center, 2020). Many parts of the website are accessible via a screen reader, though some parts, notably the limiters used to narrow search results, are missing the labels required for the screen reader to identify them. There are issues with color contrast in some places, but those can be fixed by using a high contrast extension for the user's preferred browser. The website can be magnified to 200% without losing usability. Most of the website can be navigated using speech recognition software or keyboard navigation.

Critical Evaluation

The following sections highlight some of the strengths and limitations of TDM Studio.

DATA SOURCES

The strength of TDM Studio lies in its integration of ProQuest database content cleared for data and text mining, workbench, and Jupyter Notebook. Researchers from various social science disciplines such as political science, history, sociology, communication studies, global studies, and business are already using the platform because their research relies on the content of ProQuest databases. ProQuest databases have a large collection of newspapers, financial reports, media coverage, scholarly articles, dissertations, theses, and primary sources for many of the researchers in social sciences. Search results may consist of sources with overlapping date ranges. Users should pay special attention to the date range column while selecting sources for data sets.

Users can import their own data from non-ProQuest sources into their workbenches for which they have obtained rights to text and data mining. It is important to note that non-ProQuest data may require significant cleaning and formatting.

Users of TDM Studio can access content from ProQuest databases that only their institutions subscribe to. As such, datasets may vary in content depending on the breadth of the subscription. However, e-books, select journal titles, as well as third-party databases such as GeoRef, TOXLINE, and Philosopher's Index are not included in TDM Studio regardless of the subscription status. ProQuest can create an institutional holdings list for institutions interested in subscribing to TDM Studio.

Any changes to the subscribed content can impact the coverage of the datasets. Generally, it is difficult and time consuming for researchers to obtain rights to content for text and data mining purposes. This platform removes that hurdle by providing content that has been cleared for text and data mining rights by ProQuest.

WORKBENCH LIMITATIONS

There are many advantages to using TDM Studio. However, there are limitations to workbenches that users and institutions might want to be aware of.

- Although ProQuest advertises TDM Studio as a platform for non-coders, users do need to have knowledge of Python and R programming languages for data analysis.
- Users with no coding experience do have the option to use predefined data analysis; however, that would limit the scope of data wrangling. Also, predefined codes do not have sufficient documentation for users to comprehend the purpose of them.
- The platform only allows up to five simultaneous users per workbench.
- Users are limited to creating a maximum of ten datasets per workbench; each dataset can have a maximum of two million documents.
- Users might experience delays when importing dataset to the Jupyter Notebook.
- Users are not allowed to download the content of their datasets but can use them within the platform. However, users can download their scripts and results for later use.
- Users or institutions do not have the ability to customize the workbench except the workbench name.
- Users have temporary access to workbenches. Therefore, after the indicated project end date, ProQuest assumes that the project has been completed and therefore deletes the workbenches.

JUPYTER NOTEBOOK LIMITATIONS

- Users are not allowed to copy and paste scripts and queries in the Jupyter Notebook. They are only allowed to type their queries.
- Users are only allowed to export results that are 15 megabytes, which can sometimes be too small for exporting analysis. Any file

ProQuest TDM Studio Review Scores Composite: ★★★★★ 3/4

The maximum number of stars in each category is 5.

Content: ★★★★★

ProQuest has content rich databases and the content of the datasets is dependent on the types of ProQuest databases the institution subscribes to. The Workbench dashboard allows users to search all subscribed databases. However, Visualization dashboard only allows geographic analysis, topic modeling, and sentiment analysis on contents obtained from ten preselected publications.

User Interface/Searchability: ★★★★★

The interface of the Workbench dashboard is very similar to ProQuest databases. Interface is user friendly and easy to navigate. There might be search results with overlapping data ranges.

Pricing: ★★★★★

The pricing is dependent on institution type and size. This is not a good metric for determining the cost because the number of potential users for this product would be much less than the number of users of ProQuest databases.

Purchase/Contract Options: ★★★★★

The contract provisions seem to follow standard industry practice similar to database subscription. The contract does not highlight limitations such as the ones described in Jupyter Notebook Limitations that are assumingly in place to prevent misuse of the product and unauthorized downloading of proprietary data.

bigger than 15 megabytes has to be reviewed by a ProQuest staff prior to exporting from the Jupyter Notebook.

- Users have a rolling 7-day export limitation.

VISUALIZATION DASHBOARD LIMITATIONS

The data visualization option in TDM studio is helpful for users who are unfamiliar with using R or Python and can be easily integrated into a classroom environment to support student learning. The quick turnaround time for the visualization features coupled with a simple user interface make this tool ideal for introducing large data sets in many different courses. However, like any software, there are a few limitations to using the Visualization dashboard:

- The Visualization dashboard feature only allows users to create up to five projects.
- Unlike the Workbench dashboard, the Visualization dashboard will only search up to ten preselected publications and ProQuest's Dissertations and Theses databases (dependent on institutional subscriptions) and datasets are limited to ten thousand documents.

Once completed, projects can take up to fifteen minutes to process.

- Due to the nature of algorithmic text mining, there can be instances where common place names are misidentified and the article is represented in the wrong location. Additionally, articles with multiple locations mentioned will be counted in multiple locations, although not always included in the article list associated with every location.
- Users cannots download images from the Visualization dashboard. Therefore, the only way to obtain images is by taking screenshots.

Competitive Products

There are similar integrated data mining platforms such as Nexis Data Lab by LexisNexis, Constellate by JSTOR, and Gale Digital Scholar Lab by Gale on the market. They are not necessarily competitors of TDM Studio but have adopted the concept of an integrated platform for their database users. The primary difference between these products is the content within the respective databases. Therefore, institutions intending to invest in such integrated platforms should investigate the databases that are in demand by researchers involved in data and text mining projects.

Purchase & Contract Provisions

ProQuest was able to provide a sample contract. The contract provisions seem to follow standard industry practice similar to database subscription. Authorized users are allowed to export and use derived data (summaries, factual extracts, editorial content, and metadata) solely for teaching, learning, research and analysis, including publication of research results and peer review. ProQuest offers training to institutional affiliates. They also have user manuals that institutions can further develop for their users.

Contact Information

ProQuest

789 E. Eisenhower Pkwy

Ann Arbor, MI 48108

Phone: (877) 779-6768 (sales)

E-mail: <sales@proquest.com>

Producer URL: <<https://about.proquest.com/>>

Product URL: <<https://tdmstudio.proquest.com/>> or <<https://about.proquest.com/en/products-services/TDM-Studio/>>

Free Text Keywords: Data Mining | Text Mining | Data Visualization

Primary Category: Multidisciplinary (or interdisciplinary)

Secondary Categories: Library and Information Science (LIS); Other - Text and Data Mining Platform

Type of product being reviewed: Library tool; Software; Other - Text and Data Mining Platform

Target Audience: Graduate/Faculty/Researcher

Access: Subscription

Authentication

As TDM Studio does not require IP authentication, users are able to log in and access their account without routing through a library website. It uses e-mail addresses to authenticate users. ProQuest recognizes the practice of research collaborations across institutions. Therefore, all members of a research group are given access to the workbench as long as one of the members is from the institution that subscribes to TDM Studio. Authentications for TDM Studio can be used to access TDM Visualizations.

Author's References

Jupyter. "The Jupyter Notebook." Accessed October 11, 2021. <<https://jupyter.org/>>s

ProQuest. 2020. "TDM Studio Launches Today. Learn How It's Already Inspiring Career-Defining Discoveries." Accessed October 11, 2021. <<https://about.proquest.com/en/blog/2020/TDM-Studio-Launches-Today-Learn-How-its-Already-Inspiring-Career-Defining-Discoveries-/>>

ProQuest. 2021. "ProQuest Platform." Search Tips. Last Updated September 17, 2021. <<https://proquest.libguides.com/proquestplatform/tips>>

ProQuest Support Center. 2020. "TDM Studio Accessibility Statement." Accessed October 11, 2021. <<https://support.proquest.com/article/detail?id=kA13r000000FyLNCA0>>

About the Author

Anamika Megwalu, Ph.D., M.L.I.S., is the Faculty Director of Library Instruction & Assessment at San Jose State University. Megwalu earned her doctoral degree in Information Studies from Long Island University, specializing in online scholarly communication. Her responsibilities include collaboratively planning, developing, and implementing information literacy programs and assessment, among other liaison responsibilities as an Engineering Librarian. Her research interests include agile library services for diverse user groups, online scholarly communication, and effective instructional strategies. She has published articles in *Reference Services Review*, *Against the Grain*, *Advances in Librarianship*, *The Reference Librarian*, *The Charleston Advisor*, *ASEE Conference Proceeding*, and *Science & Children*. She published a book titled, *Profiles of Academic Library Efforts to Develop Information Literacy Tutorials*. She is the subject editor of ACRL's Resource for College Libraries, *Computer and Information Technology Collection*.

Anne Marie Engelsen, MLIS, is a STEM Librarian at San Jose State University. Engelsen earned her MLIS from the San Jose State University iSchool and her Bachelor of Arts in Linguistics from the University of Arizona. Her research interests include supporting student success and self-efficacy in STEM students, especially those from underrepresented groups. She has published reviews with *The Library Quarterly* and an entry in ACRL's *The Scholarly Communications Cookbook* (in press). ■

FROM YOUR MANAGING EDITOR, *continued from page 3*

Controlled Digital Lending (CDL) at the Tipping Point

10. Determining parameters for Controlled Digital Lending for Interlibrary Loan is within the rights, powers, obligations, and judgments which librarians make every day.

To effectively support CDL practices, a variety of vendors and groups are working on software solutions to help support CDL in actual practice. As these tools develop and mature, they will provide a practical solution for libraries to implement the service. A CDL Implementors

group (<<https://sites.google.com/view/cdl-implementers>>) has been operational for some time where different opportunities and solutions are being discussed.

Although it is too early to tell whether widescale adoption of controlled digital lending will take place in academic libraries, if a legal framework can be established, libraries are anxious to provide this new service. ■