**Slide 1 - Data Types**

• The study utilizes two primary data types: full-text book content and thematic metadata.

• Full-text data is sourced from **Project Gutenberg**, providing the raw material for computational text analysis.

• Thematic metadata is collected from **Google Books**, including assigned themes, genres, and book descriptions.

• These data sources enable **statistical and machine learning-based text analysis** to examine censorship trends.

**Slide 2 - Data Description**

• The dataset consists of **20 banned books**, with 10 from Florida and 10 from Iowa.

• Florida’s selected titles include *Wuthering Heights*, *The Road*, *Redeemed*, and *Monster*.

• Iowa’s list features *The Picture of Dorian Gray*, *The Talisman*, *Glass*, and *Dead End*.

• The books cover **classic literature, contemporary fiction, and young adult novels**, allowing for diverse genre analysis.

**Slide 3&4 - Data Retrieval Process**

• Initial efforts to obtain a **random sample** were limited by full-text availability, so selections were made based on banned book lists.

• **Six out of twenty** books were directly retrieved via the *gutenbergr* package in R, while the remaining **fourteen** required web scraping from Project Gutenberg.

• Texts were stored in **data frames**: *fl\_book\_texts\_df* for Florida and *ia\_book\_texts\_df* for Iowa.

• Upcoming **data cleaning tasks** include removing HTML spacing, extra whitespace, chapter labels, stopwords, punctuation, and symbols.

• After cleaning, **trigram vectors, term frequency, and inverse document frequency (TF-IDF)** will be prepared for analysis.

Hello! I’m Conie O’Malley and I’m going to walk you through my project update for Understanding Book Banning Trends in the United States regarding data collection and cleaning.

The core of our analysis relies on two primary data types: full-text book content, which captures the complete text of the banned books, and thematic metadata, which includes categorization and contextual information about each book – often themes and back cover summaries. This fits into my project because I am comparing the full text themes that are deduced from textual analysis with the written back cover themes that many people will read before opening a book – an analysis of judging a book by its cover.

Project Gutenberg is my primary source for full-text data. This platform provides access to a number of free eBooks, allowing us to retrieve the complete texts of the selected banned books.

Thematic metadata is sourced from Google Books, which genre categorizations, book summaries, and thematic classifications of books for our comparative analysis.

The goal of collecting and processing this data is to enable statistical and machine learning-based analyses. This will help us uncover patterns and trends in censorship practices across different regions - primarily focusing on the connections of cover themes and between the cover themes.

The dataset includes a total of 20 banned books, with an equal representation of 10 books from Florida and 10 from Iowa. This balanced approach allows for a comparative analysis of censorship trends in these two states but also helps prevent the project from being too unwieldy.

One of the initial challenges encountered was the limited availability of full-text content for certain banned books. Out of the initial sample, I was only able to get one full text due to copyright issues. This caused me to pivot to sample texts available through Project Gutenberg from each state’s banned book list. I was only able to download 6 texts directly from the gutenbergr package. The other 14 I had to download directly from Project Gutenberg’s website using a function I wrote in r. This allowed me to download all the texts I needed; however I had to perform extra steps and cleaning for the additional 14 texts before storing them in data frames

The Google Books data will be downloaded in a similar fashion to pull the themes and summaries, to be stored in a data frame along with their related titles.

Before conducting any analysis, several data cleaning tasks are planned. These include removing HTML tags, extra whitespace, chapter labels, and irrelevant content such as stopwords and punctuation, ensuring the text is ready for analysis.

The final step in the data retrieval process involves preparing trigram vectors and calculating term frequency and inverse document frequency (TF-IDF). This will enable advanced text analysis to uncover patterns related to censorship.

That is the current update on my project – a description of the data, its sources, how I retrieved and cleaned them, and the next steps.