

A woman with dark hair, wearing a light-colored hoodie, is shown in profile from the chest up. She is looking down at a tablet device she is holding with both hands. The background is dark, and the lighting is soft, highlighting her face and the device. A large teal semi-transparent rectangle is overlaid on the left side of the image, containing white text.

*Gale Digital Scholar Lab*

Colloquium On Text & Data  
Mining in Libraries 2023

University of Toronto

Jess Ludwig, Director, Product  
Management

# Digital Humanities at Gale

Gale, here for **everyone**.

# Expand the Possibilities of Research

## GALE PRIMARY SOURCES

The screenshot shows the Gale Primary Sources website for The Times Digital Archive, 1785-2019. The header includes the Gale logo, navigation links for Library Menu, Gale International, and Web Demo, and language options for English, Spanish, and French. There are also links for signing in with Google and Microsoft. The main content area features a search bar with a search icon and a link to Advanced Search. Below the search bar, there are navigation icons for Home, About, Learning Center, Search History, and Get Link. The main content area is divided into two columns. The left column is titled 'ABOUT THIS RESOURCE' and contains a paragraph describing the archive as an online, full-text facsimile of more than 200 years of The Times, one of the most highly regarded resources for eighteenth-, nineteenth-, and twentieth-century news coverage. It mentions that the archive allows researchers an unparalleled opportunity to search and view the best-known and most cited newspaper in the world online in its original published context. A link 'Read more about this resource >' is provided. The right column is titled 'PRESENTED IN COLLABORATION WITH' and features the logo of The Times, which includes a crest with a crown and two lions, and the text 'THE TIMES' below it. At the bottom of the page, there is a 'SEARCH TOOLS' section with two options: 'Topic Finder' and 'Term Frequency'. The Topic Finder option includes a pie chart icon and a description: 'Visualize connections between search terms and topics and view relevant articles for those topics.' A link 'Go to Topic Finder >' is provided. The Term Frequency option includes a line graph icon and a description: 'Explores the coverage of your search terms, graphed over time and notes trends in coverage of multiple topics.' A link 'Go to Term Frequency >' is provided.

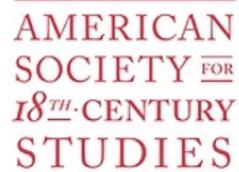
## GALE DIGITAL SCHOLAR LAB

The screenshot shows the Gale Digital Scholar Lab website. The header includes the Gale logo and the title 'GALE DIGITAL SCHOLAR LAB'. Below the title, there is a short introductory paragraph: 'Explore the depths of your institution's Gale Primary Sources using digital humanities methodologies to discover new insights. Get started by creating your account.' The main content area is divided into several sections. The first section is titled 'Digital Humanities Made Accessible to Everyone' and contains a paragraph: 'The Gale Digital Scholar Lab makes digital humanities accessible to everyone and approachable to those new to the field. Advanced features support researchers already deep into their digital humanities journey.' The second section is titled 'Build Your Content Sets' and includes a list of bullet points: 'Use Gale Primary Sources archives', 'Upload and use your own text files', and 'Download content sets for use elsewhere'. A link 'Learn more about the Build step >' is provided. The third section is titled 'Clean Texts and Prep for Analysis' and includes a list of bullet points: 'Apply stop words to your analysis', 'Use flexible options to target specific character removal', and 'Reuse configurations across Content Sets'. A link 'Learn more about Cleaning texts >' is provided. The fourth section is titled 'Analyze Content Sets with Powerful Tools' and includes a list of bullet points: 'Visualize data from up to 10,000 documents at a time', 'Explore individual documents overlaid with analysis data', and 'Download the raw data and your visualizations'. A link 'Learn more about the Analyze stage >' is provided. Below these sections, there is a video player titled 'What is the Gale Digital Scholar Lab?' with a play button icon and a video duration of 01:16. The video player is embedded in a dark blue frame. Below the video player, there are three more sections: 'What Texts are Available?' with a list icon, 'What Analyses Can I Run?' with a bar chart icon, and 'Explore Sample Projects' with a folder icon. Each of these sections includes a short description and a link to learn more.

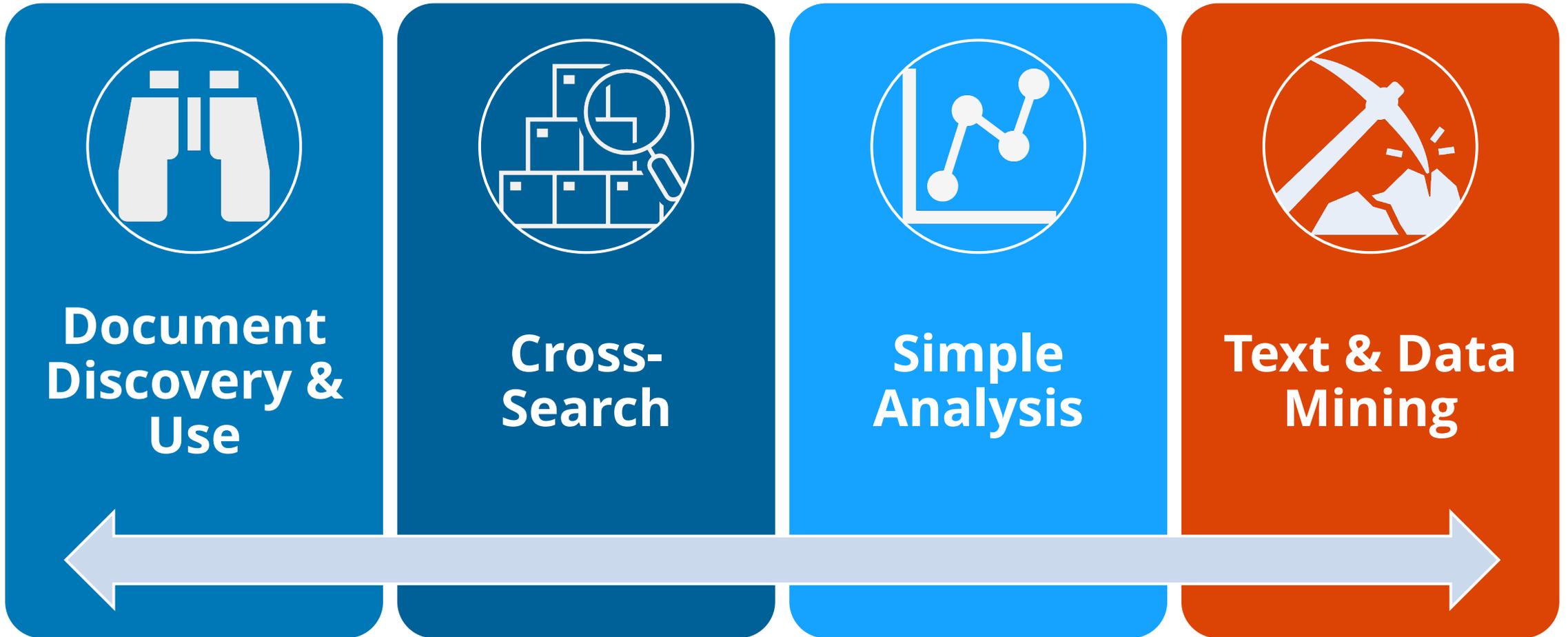
Gale, here for **everyone.**

# Global Partnerships

## Gale Fellowships



# Layers of Discovery



Boolean operators and multiple search fields provide users any number of search criteria combinations.

Basic Search Search by keyword Advanced Search

SEARCH OPTIONS  
**Advanced Search**

### Advanced Search

Search Terms

| Terms                            | Field                   | Finds results that...  |
|----------------------------------|-------------------------|--|
| Search for: <input type="text"/> | in <input type="text"/> | contain these terms in key fields; does not search entire document |
| And <input type="text"/>         | in <input type="text"/> | contain these terms in key fields; does not search entire document |
| And <input type="text"/>         | in <input type="text"/> | contain these terms in key fields; does not search entire document |

**Search Tips**  
Operators: AND, OR, NOT Proximity Nesting Special Characters: Quotation Marks Wildcards Ignored

**Selected Databases to Search (50/50)**

Select All Deselect All

- Amateur Newspapers from the American Antiquarian Society
- American Fiction, 1774-1920
- American Historical Periodicals from the American Antiquarian Society
- Archives of Sexuality and Gender
- Archives Unbound
- Associated Press Collections Online
- Brazilian and Portuguese History and Culture
- British Library Newspapers
- China and the Modern World
- Crime, Punishment, and Popular Culture, 1790-1920
- Daily Mail Historical Archive
- Declassified Documents Online: Twentieth-Century British Intelligence
- Eighteenth Century Collections Online
- Indigenous Peoples of North America
- International Herald Tribune Historical Archive, 1887-2013
- Liberty Magazine Historical Archive, 1924-1950
- Mirror Historical Archive, 1903-2000
- Nineteenth Century Collections Online
- Nineteenth Century U.S. Newspapers
- Nineteenth Century UK Periodicals
- Picture Post Historical Archive, 1938-1957
- Political Extremism and Radicalism
- Public Health Archives: Public Health in Modern America, 1890-1970
- Punch Historical Archive, 1841-1992
- Refugees, Relief, and Resettlement: Forced Migration and World War II
- Religions of America
- Sabin Americana: History of the Americas, 1500-1926
- Seventeenth and Eighteenth Century Burney Newspapers Collection
- Seventeenth and Eighteenth Century Nichols Newspapers Collection
- Slavery and Anti-Slavery: A Transnational

**Search Limiters**

by publication section:

by publication country:

by publication state/province:

by publication city:

by publication year(s):  All  Before  Within  After  Between  
 Include documents with no known publication date.

by content type:

by document type:

by publication title:

by language:

Source Library:

Illustrated Works:

OCR Confidence Range: Min  — Max   
Documents can range from 0 - 100 in OCR Confidence Score

Users will find many refining options to help limit the number of results to those that are most relevant to the definition of each corpus they create.



### Build Your Content Sets

- Use Gale Primary Sources archives
- Upload and use your own text files
- Download content sets for use elsewhere

[Learn more about the Build step »](#)



### Clean Texts for Computational Analysis

- Apply stop words to your analysis
- Use flexible options to target specific character removal
- Reuse configurations across Content Sets

[Learn more about the Clean step »](#)



### Analyze Content in Powerful New Ways

- Visualize data from up to 10,000 documents at a time
- Explore individual documents overlaid with analysis data
- Download the raw data and your visualizations

[Learn more about the Analyze step »](#)

Gale, here for **everyone.**

Library Menu: Demonstration Account

GALE DIGITAL SCHOLAR LAB

My Content Sets Build Clean Analyze

New Configuration Save As Test Configuration

## Clean

Select Cleaning Configuration to Edit  
Default Cleaning Configuration

**Video: Cleaning a Content Set**  
Additional topics:  

- Creating a Clean Configuration
- Applying During Analysis
- View All Clean Help »

### Stop words

Set the words you want the Analysis Tools to ignore.  
Choose a Starter List

- a
- about
- above
- across
- after
- afterwards
- again
- against
- all
- almost
- alone
- along
- already
- also
- although
- always
- am
- among
- amongst
- amongst
- amount
- an
- and
- another
- any
- anyhow
- anyone
- anything
- anyway
- anywhere
- are
- around
- as
- at
- back
- be
- became

Ignore stop words case  
[Clear List](#)

### Text Correction

Options for automatic text correction that will be applied before each Analysis

Turn on all options

**Text Modification**  
 All lower case

**Characters**  
 Remove all extended ASCII characters  
 Remove all number characters

**Special Characters**  
 Remove all special characters  
[Set specific characters >](#)

**Punctuation**  
 Remove all punctuation  
[Set specific punctuation >](#)

**Spacing**  
 Remove all tabs  
 Remove all line breaks  
 Reduce multiple spaces to one space (ex: "hello there" becomes "hello there")

**Document Sections**  
 Remove body text  
 Remove all non-body content  
[Set specific sections >](#)

**Replacements**  
 Replace this... With this...  
   
   
   
[Add a Row](#)

### Configuration Notes

Space to make notes or describe the purpose of this configuration.

Configuration description/notes

Use a default stop word configuration, customize it, or create your own and then test with a sample from your content set.

Further customize what you'd like to remove or find and replace in the content set.

# Primary Sources Analysis

## EXPANDING RESEARCH POSSIBILITIES

- Sentiment Analysis

- Parts of Speech



- Named Entity Recognition



- Ngrams

- Topic Modeling



- Document Clustering



Library Menu: Demonstration Accounts | English | Signed in: Jess

GALE DIGITAL SCHOLAR LAB | My Content Sets | Build | Clean | Analyze

Analyze

Content Set: 1650-1800 Tea

Video: Analyze a Content Set

- Additional topics:
- Selecting the Right Tool
- Setting up and Running
- View All Analyze Help >

Select All

Document Clustering: Document Similarity

Named Entity Recognition: Entities Found

Ngrams: Word Cloud, Term Frequency

Parts of Speech: Parts Comparison

Sentiment Analysis: Sentiment Scores, Sentiment Over Time

Topic Modeling: Topics, Topic Proportion

GALE | ABOUT | HELP | DICTIONARY | CONTACT US | TERMS OF USE | PRIVACY POLICY | ACCESSIBILITY | END SESSION | Gale, here for everyone.

Gale, here for **everyone**.



Back

Ngrams

1950-1980 Black America & The Law - no subject terms

Download Help

Legend

Legend

View

- Top 100 Ngrams
- Ngram Search

Enter terms to match ngrams

Color Coding

- Frequency
- Rank



### Sentiment Analysis

Refugees - Times Opinion & Editorial

- Legend
- Tool Setup
- Run History

#### Legend

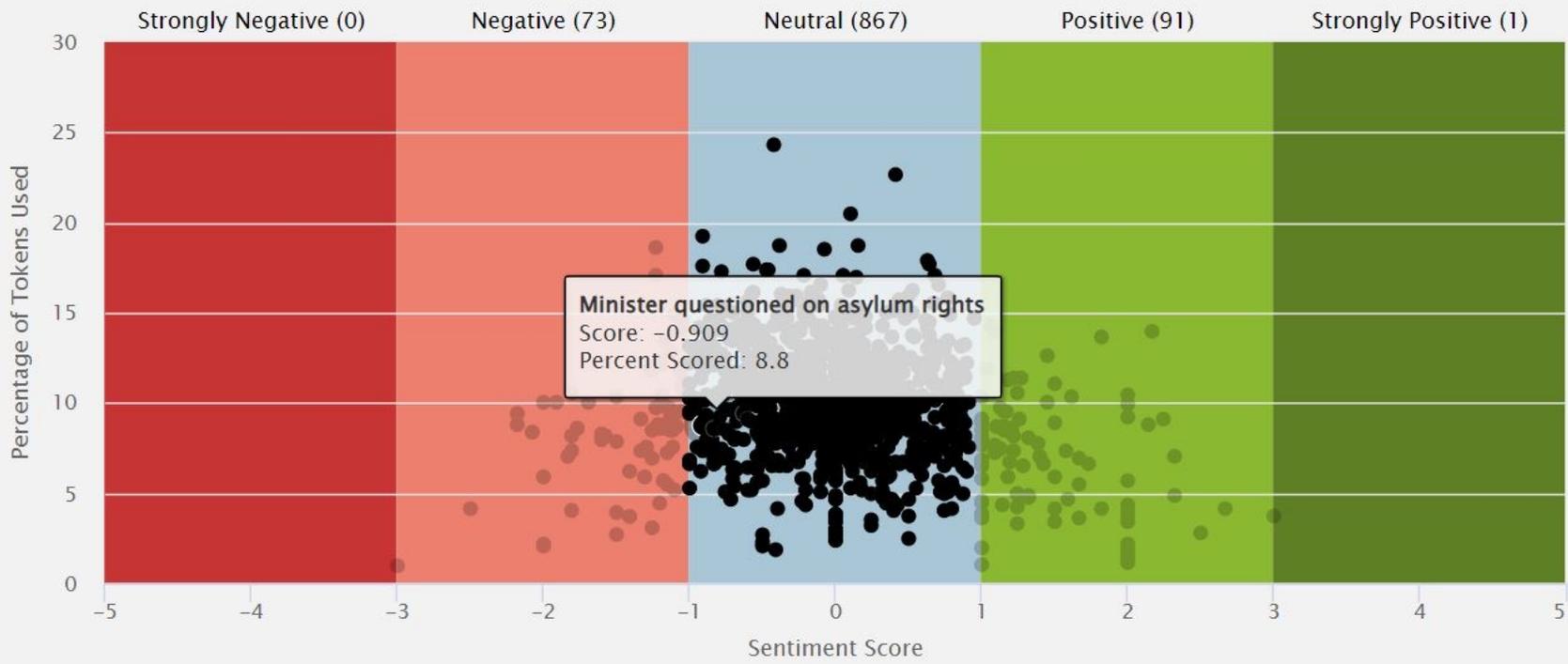
0.003 Overall Content Set Sentiment

#### Document Sentiment Scale

- Strongly Positive
- Positive
- Neutral
- Negative
- Strongly Negative

#### Document Issues

7 Unscored Documents



Back

Parts of Speech  
1650-1800 Coffee - Manners & Customs

Download Help

Legend

Legend

Parts of Speech

Click to toggle categories on and off

- NOUN
- INTJ
- PART
- PROPN
- CCONJ
- PUNCT
- PRON
- SCONJ
- NUM
- VERB
- DET
- SYM
- ADV
- ADP
- SPACE
- ADJ
- AUX
- OTHER

Group Data By

Author

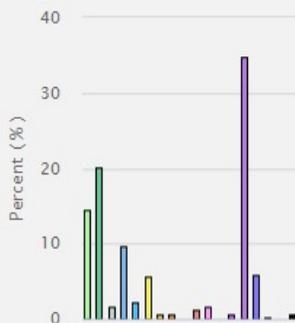
10 of 192 selected  
Change selection

Graphs Shown: 10

Grid Columns: 4

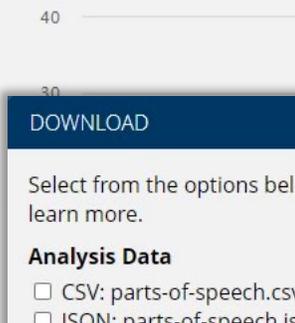
Adam Petrie

1 documents | Tagged  
26,243 Parts of Speech



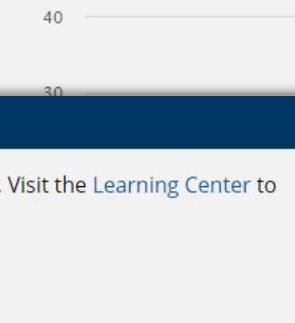
Alexander Pope

5 documents | Tagged  
305,224 Parts of Speech



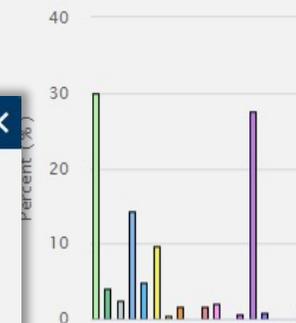
Althea Fanshawe

1 documents | Tagged  
34,760 Parts of Speech



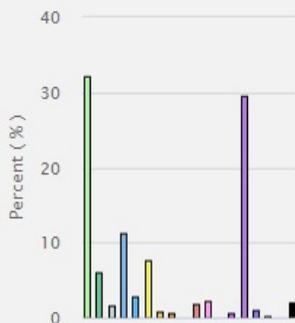
American in England

1 documents | Tagged  
25,357 Parts of Speech



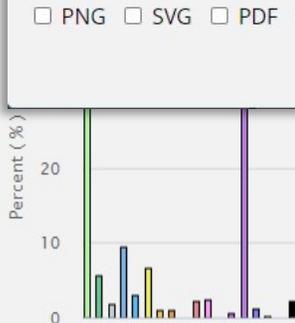
Anne Fisher

1 documents | Tagged  
111,368 Parts of Speech



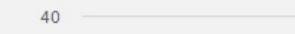
Arthur William Costigan

2 documents | Tagged  
87,586 Parts of Speech



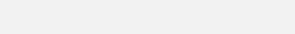
Arthur Murphy

4 documents | Tagged  
277,736 Parts of Speech



Argens, marquis d'

11 documents | Tagged  
814,165 Parts of Speech



**DOWNLOAD**

Select from the options below. Visit the [Learning Center](#) to learn more.

**Analysis Data**

- CSV: parts-of-speech.csv
- JSON: parts-of-speech.json

**Document Metadata**

- JSON: document-metadata.json

**Parts of Speech Visualization**

- PNG
- SVG
- PDF
- JPEG

Cancel Download

Inspect

Document

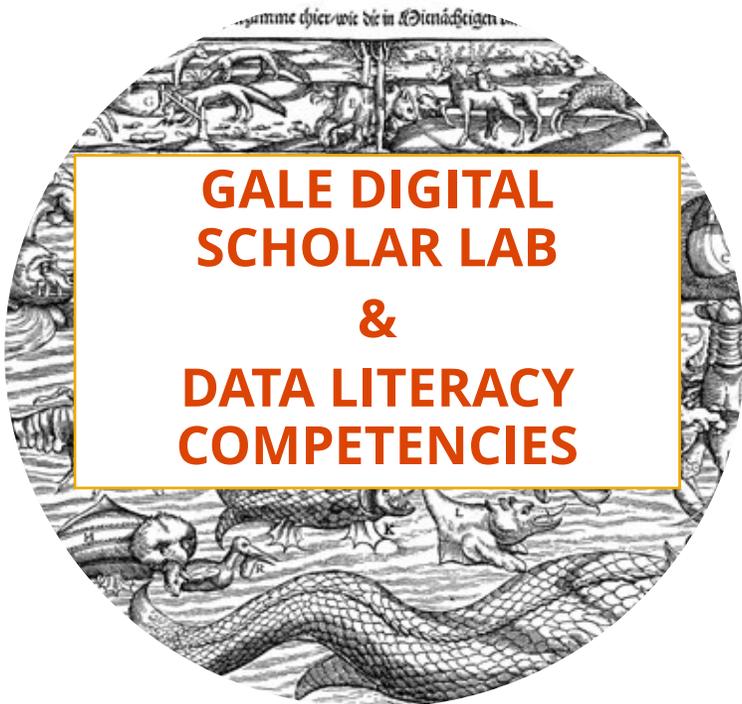
The works of Alexander Pope, Esq; vol. IV. Part II. Containing the second part of his letters

Open Document

Parts of Speech

| Category | Count  | Percentage |
|----------|--------|------------|
| NOUN     | 14,972 | 27.78%     |
| PROPN    | 4,901  | 9.09%      |
| PRON     | 1,339  | 2.48%      |
| VERB     | 6,044  | 11.21%     |
| ADV      | 1,789  | 3.32%      |
| ADJ      | 3,962  | 7.35%      |
| INTJ     | 399    | 0.74%      |
| CCONJ    | 349    | 0.65%      |
| SCONJ    | 0      | 0%         |
| DET      | 919    | 1.71%      |
| ADP      | 1,226  | 2.27%      |
| AUX      | 50     | 0.09%      |
| PART     | 368    | 0.68%      |
| PUNCT    | 16,038 | 29.76%     |
| NUM      | 652    | 1.21%      |





## Build

- Interpret and critically evaluate data and their sources
- Read/understand data types and formats
- Find, select, access, or create datasets in order to test a hypothesis or answer a research question
- Create metadata to meet data publication requirements

## Clean

- Clean / process / convert data

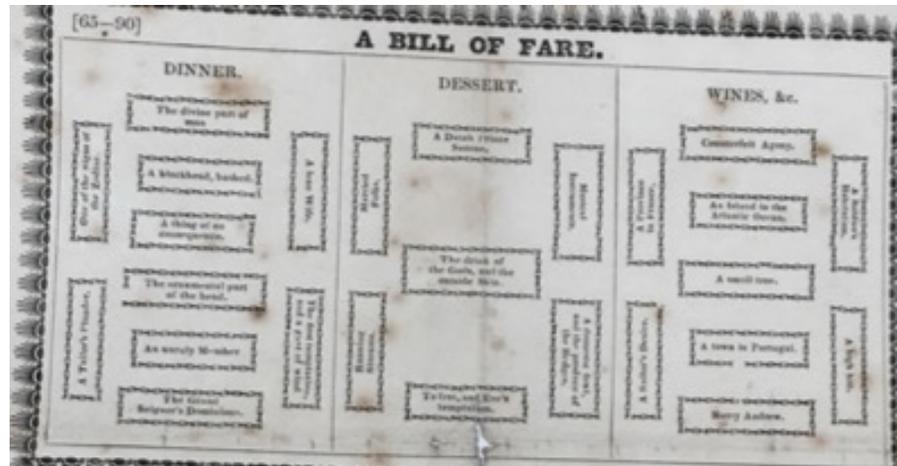
## Analyze

- Analyze data
- Communicate data effectively to different audiences, in part by using visualizations
- Ethically collect / use / cite data
- Integrate and synthesize data into different contexts with other sources and prior knowledge

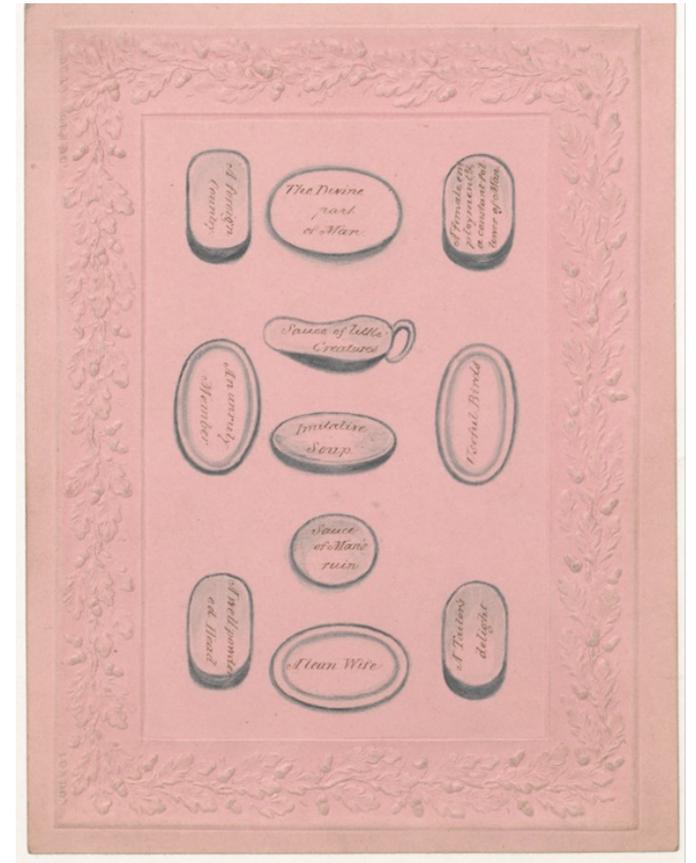
# Gale Digital Scholar Lab Case Studies

# McGill University: Riddle Culture in the 18<sup>th</sup> & 19<sup>th</sup> Centuries

Enigmatic Bills of Fare led to explorations of riddling culture in 18<sup>th</sup> and 19<sup>th</sup> centuries.



**Figure 1.** Printed Enigmatic Bill of Fare. Bodleian, John Johnson Collection



**Figure 2.** Manuscript Enigmatic Bill of Fare. Cadbury Research Library: Special Collections



# Loyola University: Anthropology 100: Globalization & Local Cultures

## ANTHROPOLOGY

SPRING 2021

100 and 200 level offerings



### ANTH 100 GLOBALIZATION AND LOCAL CULTURES

Multiple Sections—Check LOCUS for details  
Tier I Core Credit  
Counts toward: African Studies, Black World Studies,  
Asian Studies, GIST, Intl Business



### ANTH 102 CULTURE, SOCIETY, DIVERSITY

Prof. Ben Penglose  
Some synchronous sessions  
MWF 10:50 - 11:40am  
Tier II Core Credit



### ANTH 207 ECONOMIES, CULTURE, DEVELOPMENT

Prof. Noah Butler  
Synchronous  
TR 3-4:35pm  
Prerequisites : ANTH 100 or ANTH 102  
ANTH and GIST credit



### ANTH 244 HISTORICAL ARCHAEOLOGY

Prof. Vicky Kynourgiopoulou  
Asynchronous  
MWF 9:20 - 10:10am



### ANTH 231 LINGUISTIC ANTHROPOLOGY

Prof. Thea Strand  
Synchronous sessions on Wed only  
MWF 1:30 - 2:20pm  
Prereq: ANTH 100 or 102 or instructor permission

CULTURE, BIOLOGY, ARCHAEOLOGY, LINGUISTICS

### ANTH 101 HUMAN ORIGINS

Prof. Paula Tomczak  
Asynchronous  
TR 9:45 - 11am  
Tier II Core Credit

### ANTH 105 HUMAN BIOCULTURAL DIVERSITY

Prof. Kristin Krueger  
Asynchronous  
TR 11:30 - 12:45pm  
Tier II Core Credit

### ANTH 208 LANGUAGE AND IDENTITY

Prof. Thea Strand  
Tier II Core Credit  
Synchronous sessions on Wed only  
Multiple Sections—check LOCUS for details

### ANTH 217 MEXICAN CULTURE AND HERITAGE

Prof. Ruth Gomberg-Muñoz  
Asynchronous  
TR 11:30 - 12:45pm  
Counts for: GIST and LALS credit

LOYOLA UNIVERSITY CHICAGO | 150

A-Z CONTACT DIRECTORIES LOCUS SUPPORT LOC

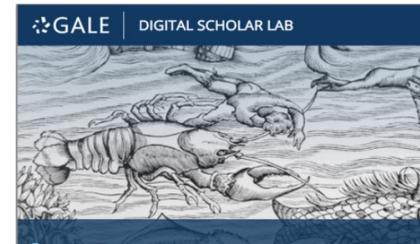
LINKS  Q

## CENTER FOR TEXTUAL STUDIES AND DIGITAL HUMANITIES

ABOUT RESEARCH PROJECTS ACADEMICS EVENTS RESOURCES GIRLS WHO CODE

ABOUT  
RESEARCH PROJECTS  
ACADEMICS  
EVENTS  
RESOURCES  
GIRLS WHO CODE

### Workshop: Gale Digital Scholar Lab



The University Libraries is initiating a trial of the *Gale Digital Scholar Lab*, a platform that enables students and researchers to do text mining and analysis on materials from *Gale Primary Sources*, which includes digitized texts from the 16th century through the 20th century. This workshop will include an interactive introduction to using the lab and a discussion of best practices. Attendees will learn how to perform searches, manage data sets, apply analysis tools, and review results. This workshop will be appropriate for researchers from a range of disciplines, and requires no prior experience with digital research methods.

The availability of digital humanities data and the tools to analyze it has allowed me to truly engage ALL my students in research. – Catherine Nichols, Advanced Lecturer in Cultural Anthropology and Museum Studies, Loyola University

Gale, here for **everyone.**



# University of Washington: Applied Digital Humanities

**Worlds Co**

- About the Team
- Selections from the Collection
- StorymapJS Items
- Data Curation Journey
- Data Visualizations
- Sentiment Analysis
- Visualization
- Topic Modeling
- Visualizations
- StorymapJS
- Visualization

**Research Question**

What major themes emerged from the Chicago World's Fair, as...

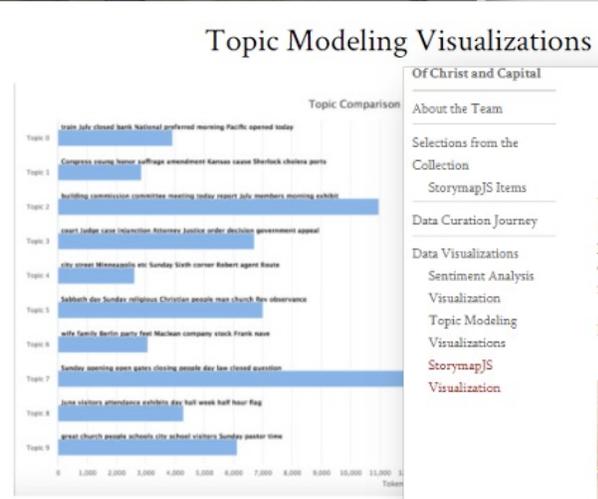
**Project Overview**

When selecting a topic, Chicago and were familiar...

**About the "Sunday"**

The identification of the content set's creation issues. Sabbatarianism, as an act of religious group. On the other Sundays, citing the not only one of economic...

**About the "Sunday"**



This Topic Modeling analysis utilized Elise's final cleaning of these results, I began with the Topic Comparison graphs. I app graph, giving me a bit more information and an overview of the immediately that the topic with the most tokens was Topic 7. I four of the ten topics the analysis identified. From here, there topic and the fifth, so I chose to focus specifically on these top

| Topic     | TERM       | COUNT  | PERCENTAGE | BIAS |
|-----------|------------|--------|------------|------|
| Topic 7   | open       | 100    | 0.0001     | 0.0  |
|           | gates      | 100    | 0.0001     | 0.0  |
|           | close      | 100    | 0.0001     | 0.0  |
|           | law        | 100    | 0.0001     | 0.0  |
|           | commission | 100    | 0.0001     | 0.0  |
|           | meeting    | 100    | 0.0001     | 0.0  |
|           | committee  | 100    | 0.0001     | 0.0  |
|           | committee  | 100    | 0.0001     | 0.0  |
|           | committee  | 100    | 0.0001     | 0.0  |
|           | committee  | 100    | 0.0001     | 0.0  |
| committee | 100        | 0.0001 | 0.0        |      |

themes I was expecting to see reflected in these results. I was u token word in this largest topic category, showing up a total of between the most and second most occurring words. The wor "Open", "gates", and "closing" were the next tokens on this list. there was some conversation happening in 1893 regarding the appearance of the word "law" is interesting in that unlike the o this conversation as I was imagining it beyond the bounds of a

| Topic   | TERM       | COUNT | PERCENTAGE | BIAS |
|---------|------------|-------|------------|------|
| Topic 2 | open       | 100   | 0.0001     | 0.0  |
|         | gates      | 100   | 0.0001     | 0.0  |
|         | close      | 100   | 0.0001     | 0.0  |
|         | law        | 100   | 0.0001     | 0.0  |
|         | commission | 100   | 0.0001     | 0.0  |
|         | meeting    | 100   | 0.0001     | 0.0  |
|         | committee  | 100   | 0.0001     | 0.0  |
|         | committee  | 100   | 0.0001     | 0.0  |
|         | committee  | 100   | 0.0001     | 0.0  |
|         | committee  | 100   | 0.0001     | 0.0  |

## THE SUNDAY QUESTION: CLOSURE DURING THE 1893 COLUMBIAN EXPOSITION: StorymapJS

Using Named Entity Recognition (NER) analysis, we determined areas of significance throughout the fairgrounds (buildings, lawns, gates, streets, etc.) pertaining to the controversy surrounding the Sunday Question. Place entities were plotted manually based on souvenir maps of the fairgrounds from 1893 held by the Library of Congress. Each slide has an explanatory note beneath the map which describes the significance of the excerpted document and the location it references.

[Link to Full Page Version](#)

Map Overview | [Back To Beginning](#)

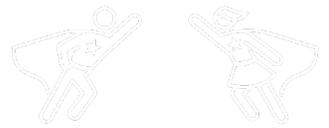
**THE SUNDAY QUESTION: CLOSURE DURING THE 1893 COLUMBIAN EXPOSITION**

[Start Exploring](#)

StoryMapJS | Leaflet | Map data by OpenStreetMap contributors, Imagery © Mapbox, © Mapbox, © OpenStreetMap contributors, CC BY 3.0

# Gale Digital Scholar Lab Roadmap

# User Engagement Group



**Catherine  
Nichols**  
Loyola  
University



**Kathrina  
Perry**  
University of  
Northampton



**Jose Intriago  
Suarez**  
Marquette  
University



**Qiuyang  
Chen**  
University of  
Warwick



**Oihane  
Etayo**  
University of  
Warwick



**Caitlin  
Bagley**  
Gonzaga  
University

# User Engagement Program

## PURPOSE

- General Platform Feedback
- Validation & Usability feedback for in progress projects
- Discovery feedback for upcoming projects

# New Data Sets

- Watergate in the News
- Watergate Declassified Docs
- Suffragettes
- Stonewall Riots
- The Murder of Olof Palme
- The Boxer Uprising
- Roberto Calvi Trial
- Sojourner Truth

## Dataset 1: Watergate in the News

### Get a copy of the dataset

Get a Copy

The Watergate in the News Dataset was compiled using Gale Primary Sources and Gale Digital Scholar Lab.

**Archives Used:** The International Herald Tribune Digital Archive; The Daily Mail; The Telegraph; The Sunday Times; The Times Digital Archive.

**Scope:** 1972 - 1975. 4741 documents.

**Source Libraries:** The New York Times Company (2194); Telegraph Media Group (1091); Times Newspapers Limited (1049); Associated Newspapers Limited (406).

**Document Types:** Article (4257); Editorial (251); Letter to the editor (144); Back matter (60); Front matter (29).



# Group Collaboration



## WHAT

A collaborative space in which students can work together to create content sets, analyse their materials and build their project as a group.

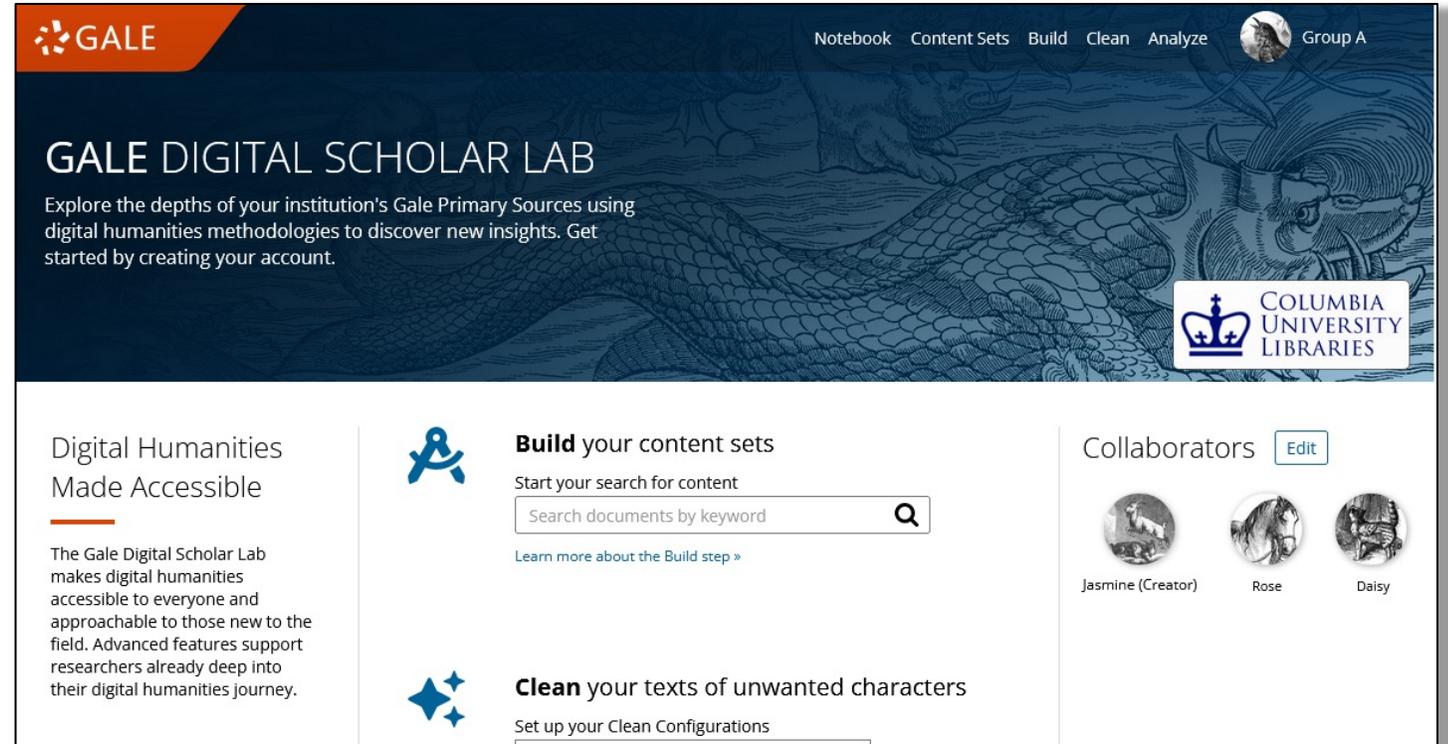
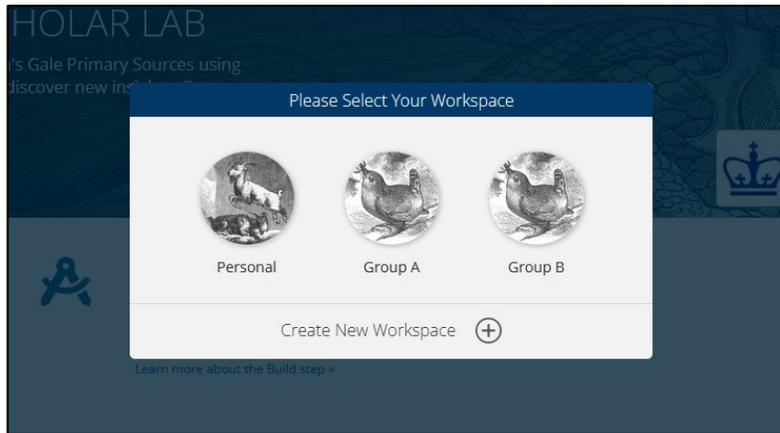
## WHY

47% of survey respondents (Spring 2021) identified collaboration tools as Extremely or Very Important. Integrating the Lab in the classroom increases usage and the likelihood of renewal.

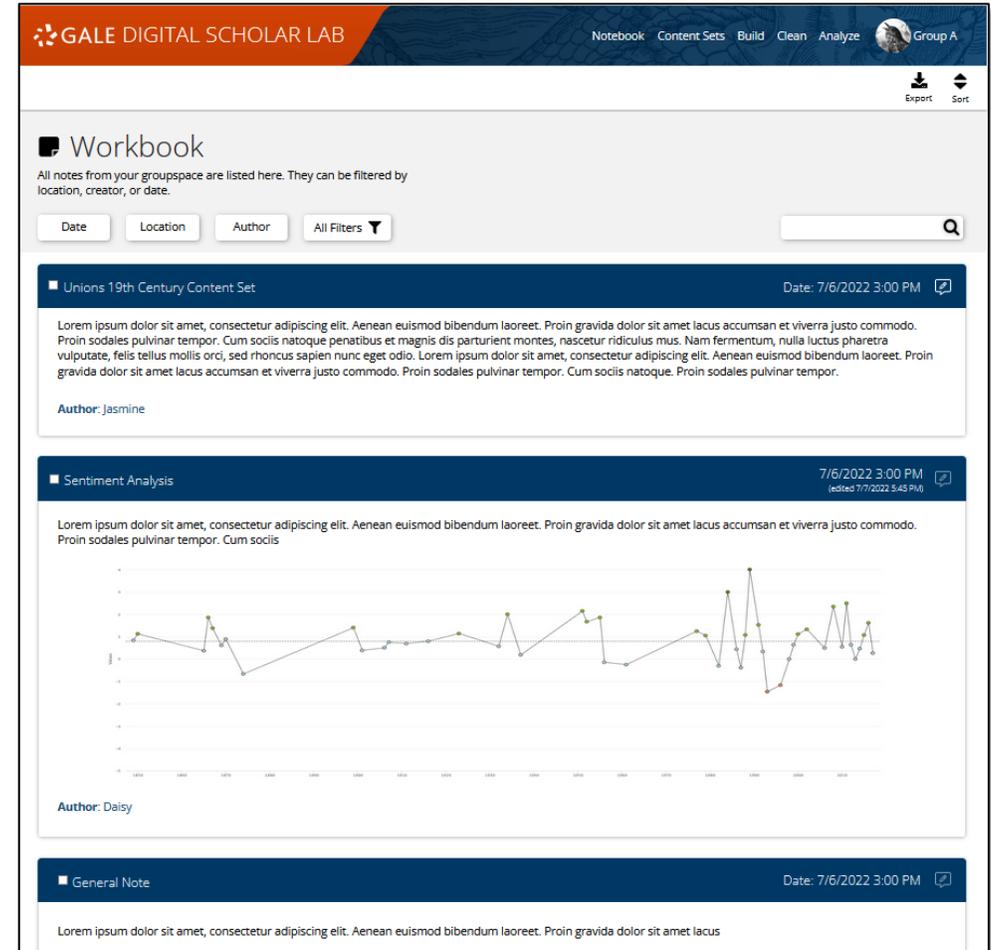
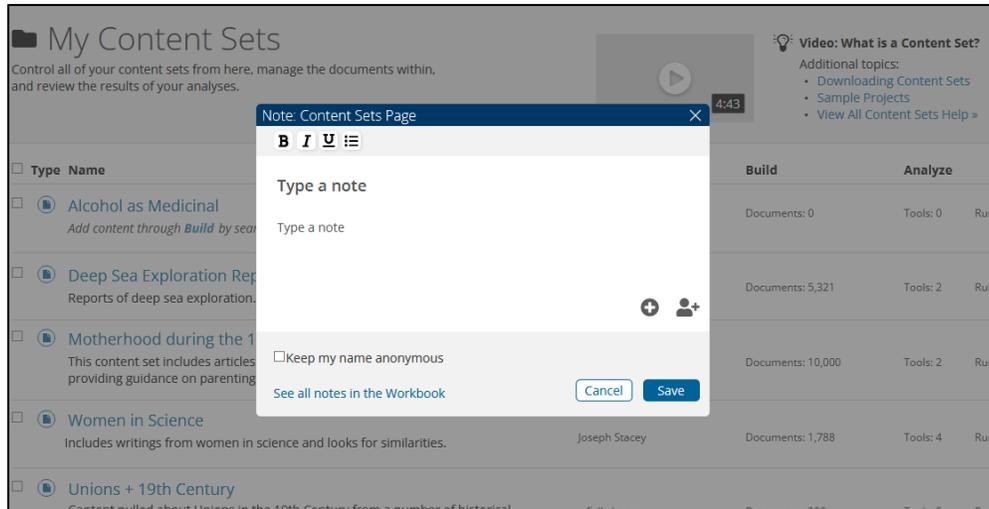
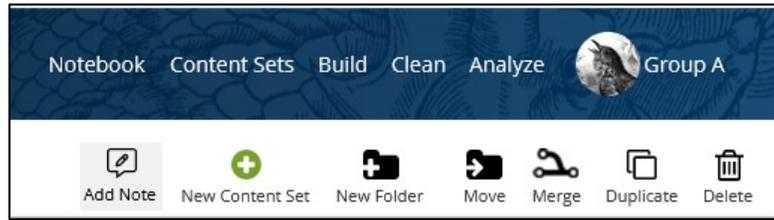
## HOW

Working closely with our users to define requirements, build and validate prototypes and undertake extensive usability testing to ensure the we are meeting user needs.

# Group Workspace



# Group Collaboration Notebook



Gale, here for **everyone.**



**Thank you!**

Gale, here for **everyone.**