Getting Started – Logging In

Your library or institution will provide you with your login link.

You will have the option to sign in with a Google or Microsoft Account to ensure you have a personal account whilst you are working in the Lab.

To learn more about the Gale Digital Scholar Lab privacy policy, click the link in the Log In/Create Account screen.
After you have logged in, you will arrive at this page, where you can see the main components that make up a project workflow in the Lab.

This workflow begins with building your Corpus (personalised archive or content set) through the database-style search functionality. You can continue to the analysis of your content set using a variety of text-mining analysis methods, and finally move to the dashboard view where you can manage, organise and share the research accomplished in the Lab.

You can begin searching from the homepage with some **basic search** options here, including a keyword search and the ability to use Boolean search terms like ‘or’ ‘and’, ‘not’ etc.

This dropdown menu gives the option to search by subject or publication, or an entire document.

You can navigate to the **advanced search** page for more functionality.

It’s accessible from the link beneath the basic search bar, or from the header menu link.
Gale Digital Scholar Lab – Getting Started Walkthrough Guide

Search: Personalised Archive Building

Build Your Content Set
It’s easy to start building your own content sets. Gale Primary Source content is a cornerstone of the Lab and the ability to create personalised archives relatively quickly is what really sets it apart.

Type in a basic search term such as ‘global warming and climate change’ and then review your search results.

Search Results Page
The Search Results Page is where the process of building your own personalised archives begins.

You can see immediately how many documents have been returned in a search and scroll down to review this content.

In the search results, you can see a snapshot view of the high-level metadata about each document. The metadata includes many facets that you would expect to see about each document (such as the source archive, the publication date and the publisher).

You can also view the initial lines of OCR text for each document, as well as the OCR confidence rating.

OCR confidence is the OCR engine’s own confidence in its translation. OCR confidence is not the same as OCR accuracy; a document could have high accuracy but a low confidence score, skewed, for example, by images in the document.
From this search results page in this left column, I can further limit my search by selecting relevant databases, or subjects on which I’d like to focus for my specific content set.
# The Document Explorer

By clicking on the title of the document, the **Doc Explorer** page opens.

This puts up a side-by-side comparison of the original Primary Source next to the OCR text.

The Primary Source highlights the keywords used to perform the search and from where this document was derived.

Users can also see the OCR confidence level in this view, at the top of the Document Text.

By clicking on the ‘learn how this text was created’ link a pop-up window appears explaining what this means in more detail.

A close reading of the OCR Text and the Primary Source at this point is beneficial to researchers who can now quickly decide whether to include or exclude a document from a content set.
Creating Content Sets

Once happy with your choice of results you can **Add to Content Set** – to populate your bespoke corpus. The messaging on each document changes to indicate that it has been added to the content set.

I can choose to add a single document, a range of documents or ‘select all on a page’ to add up to 100 documents per results page. If your results are greater than 100 documents, you will be given the option to add up to 10,000 documents to your Content Set.

You have the option to create a new content set or add to an existing content set. I am going to add a new content set – and create a name for it here. This now forms your collection of documents that you have curated and saved.

Each time you log in, the list of content sets will be available to work on.

I can review and edit My Content Sets at any time.

One key benefit of this is that you don’t need to rebuild your corpus from the start – you can continually refine and update if you wish. You can make copies of your content set to facilitate versioning of your research project.
Editing My Content Sets

To edit a content set you can access your saved work from ‘My Content Sets’ here in the header.

By clicking on the edit button, you can change a content set name, and add a description. You can continue to curate your content set by adding or deleting documents through the search function, preparing the material to move forward to the analysis and visualisation process.

The Dashboard View

Clicking on the name of a Content Set will bring up this dashboard view.

The Overview tab gives a useful summary of the composition of your Content Set.

You also have the option to download your content set here.

The Search History tab will detail all the steps taken to create your Content Set, including the time stamp for the searches, how many results and how many documents you added to the Content Set per search.
By clicking on the download button, you begin the process of this functionality. This pop-up window will open—which explains in more detail the parameters for downloading.

You have the option to add a Cleaning Configuration to the data that you download to clean the OCR at scale.

You may need to refresh your browser for a status update. Once ready the download button will change to say 'download ready'.

By clicking on this you will get a similar message as before, saying that you can download up to 1,000 documents per session, but also have the option to click on a green download button. This will download the OCR text for each document as a zip file.

The Document tab lists each document in the content set, with links to the Doc Explorer view, and detailed metadata about each piece of work.

It's possible to narrow the range of displayed documents by selecting from a range of limiters.
The Analyze stage is where you build and iterate on your data analyses. Here, you can match your content set with the analytical tools you want to run against your archive.

When you’re first working to analyse data, select the **Analyze** link in the top bar.

Once you’ve selected the content set to analyse, you begin by selecting the analysis methods you want to run against your content set by clicking on the green **Add Tool** button.

This brings up a list of tools that are in the Lab. We have provided a user-friendly tool description for each analysis method. It explains what the tool does and why it might be useful.

The **Learn More** link takes you to the source documentation to learn more about each specific algorithm.

<table>
<thead>
<tr>
<th>Analysis Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clustering</strong></td>
</tr>
<tr>
<td>Clustering analyzes the documents from a content set using statistical measures and methods to group them around particular features or attributes. This implementation of clustering leverages the k-means algorithm to create clusters of documents according to similar words contained within each document of your content set. <a href="#">LEARN MORE</a></td>
</tr>
<tr>
<td><strong>Example Outputs</strong></td>
</tr>
<tr>
<td>Tabular Data</td>
</tr>
<tr>
<td><strong>Named Entity Recognition</strong></td>
</tr>
<tr>
<td>Named Entity Recognition (NER) recognizes and extracts proper and common nouns from documents using a Parts of Speech tagging method, and outputs them as lists of grouped by entity type. Some entity types available for extraction are people (including fictional), groups (nationalities, religions, or political organizations (companies, agencies, institutions, etc.), locations (countries, states, cities), products (objects, vehicles, foods, etc.), works of art (titles of books, songs, etc.), dates (absolute or relative dates or periods), among others. This implementation uses Spacy’s Named Entity Recognition model. <a href="#">LEARN MORE</a></td>
</tr>
<tr>
<td><strong>Example Outputs</strong></td>
</tr>
<tr>
<td>Hierarchical Tree</td>
</tr>
</tbody>
</table>
Once you’ve selected all the tools you want to run against your Content Set by clicking the **Add** button, select **Done**.

All tools can be easily configured by selecting **View** under Tool Setup, including adding a cleaning configuration of your choice to the tool prior to run.

All tools have a default configuration to complement the method, so all users can generate an analysis, no matter their level of experience.

Each tool has an **About** icon that will summarise the tool’s function.

You can **Run** each tool from the **Tool Setup** page or navigate back to the **Analyze** screen to select all tools you want to run before pressing the **Run Selected** icon.

You can see the **run status** for each tool, which will change as it processes.

The **Expected Results** icons will go green when the tool has finished running. You can click on the icons to view the results.

A time-stamped **run history** allows you to see the results of any analysis previously run.

If you want to toggle between different configuration settings to see how the results differ, you can do so.
Tools

The Tools used and the visualisations and analyses generated will all appear under the Analyses tab of the Dashboard View.

Tools Continued...

You can view the different analyses of the tools applied to your Content Set in the Dashboard View or view the Analyze page.

Each tool is used for a different purpose, and all have different output visualisations, from word clouds to scatterplots.

If you select the Topic Modelling tool you will see that some of the tools have various layers of visualisation available, with which you can interact.

For example, in Topic Modelling you can critically examine the various terms to identify and rename the thematic topics generated, while in Named Entity Recognition, you can create new Content Sets based on entities identified by your analysis.