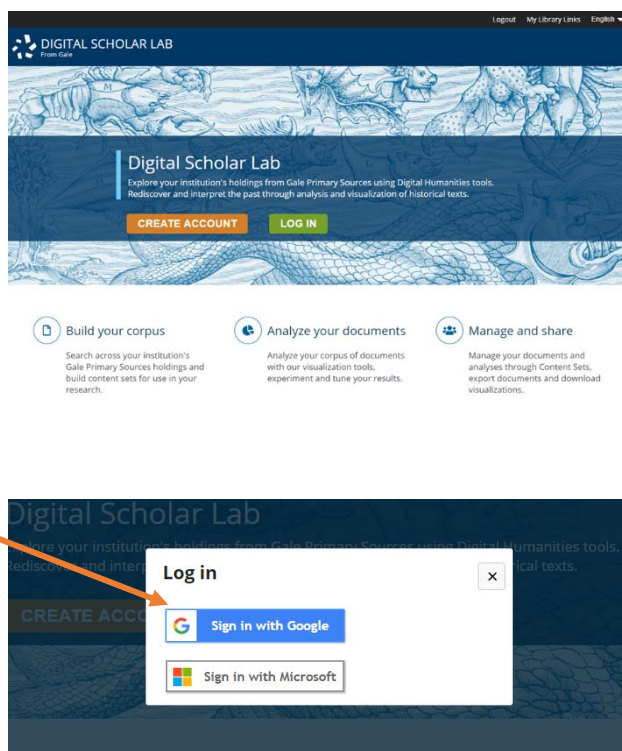# Gale Digital Scholar Lab – Getting Started Walkthrough Guide

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, this is so you have a personal account whilst you are working in the Lab.

Homepage – Overview

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 desktop view where you can manage, organise and share the research accomplished in the Lab.
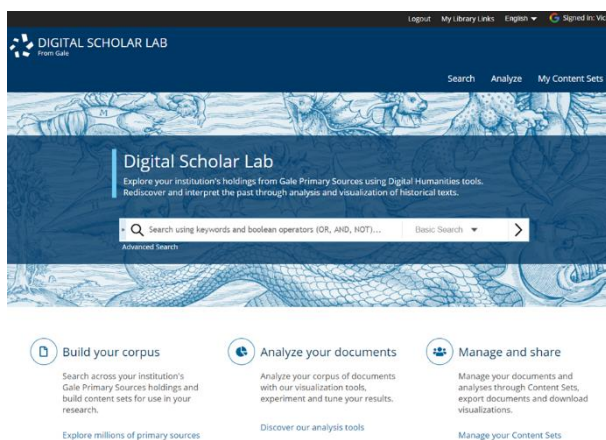
# Gale Digital Scholar Lab – Getting Started Walkthrough Guide

Searching

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.

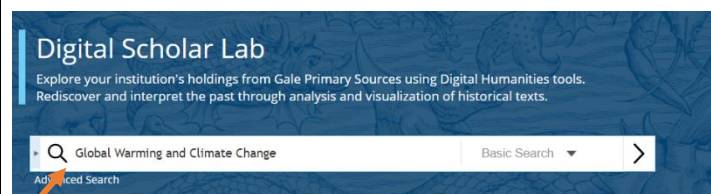It's accessible from the link beneath the basic search bar, or from the header menu link.
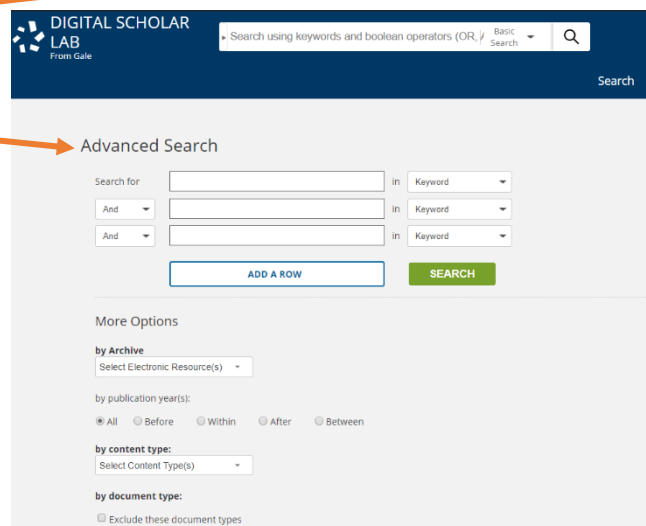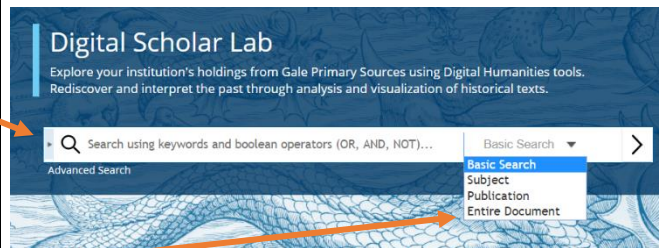
Search: Personalised Archive Building

Build Your Content Set
It's easy to start building your own content sets. Gale's Primary Sources 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 collection from where it came, the date of pub, the author)

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.

Doc 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 the OCR 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' to add up to 100 documents per results page.

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 scratch – you can continually refine and update if you wish. You can make copies of your content set to facilitate versioning of your research project.

Gale Digital Scholar Lab – Getting Started Walkthrough Guide

Editing My Content Sets

To edit a content set you can access your saved work from 'My Content Sets' here in the header.
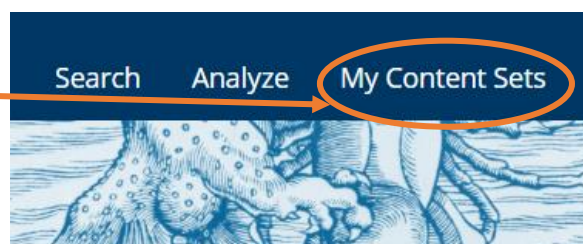
By clicking on the edit button, you can change a content set name, and add a description. You can continue to curate your content set by adding or deleting documents through the search function, preparing the material to move forward to the analysis and visualisation process.
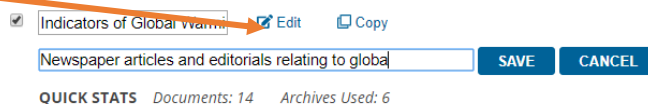
The Dashboard View

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

The **Overview** tab gives a useful summary of

Archives Used
Document Type
Authors
Source Libraries

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

The Dashboard View Continued

By clicking on the download button, you generate the process of this functionality. This pop-up window will open – which explains in more detail the parameters on downloading.

You 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.

Analyse

The Analyze stage is where you build and iterate on your visualizations. This is where you match your content set with the analytical tools you want to run against your archive.

You can navigate to the Analyze page from the search results page when you're building your content set, or by selecting the link in the navigation window.

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.

ANALYZE

CONTENT SET
Indicators of Global Warming in the Natural World ▾

Get started by adding an Analysis Tool

ADD TOOL

Analysis Tools

Clustering
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. **LEARN MORE** ⬏

ADD

EXAMPLE OUTPUTS
Tabular Data   Scatter Plot

Named Entity Recognition
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, religious, 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. **LEARN MORE** ⬏

ADD

EXAMPLE OUTPUTS
Hierarchical Tree   Tabular Data

GALE DIGITAL SCHOLAR LAB

Analyse Continued…

Once you've selected the tools by clicking the add button, you navigate back to the Analysis Page. I will now run my analysis by selecting all the tools and pressing the **Run Selected** icon.

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

You can **configure** each tool as well. Most of the analysis tools are customisable, allowing users to set the tool up to their own specifications. All tools have a default configuration to complement the method, so all users can generate an analysis, no matter their level of experience.

A **run history** allows you to regenerate any analysis previously run

If you want to toggle between different configuration settings to see how the results differ, you can do so. Finally, you are presented with a thumbnail icon that represents the output types you can expect from each analysis.

Gale Digital Scholar Lab – Getting Started Walkthrough Guide

Tools

By clicking on the name of each tool, you can see the visualisations rendered by your analysis.

Clicking the tool name for each analysis method opens the Explorer view for that specific tool. The Explorer view allows you to see what types of visual outputs have been rendered by each analysis.

Analysis (6)

☐ Select All

☐ Parts of Speech Tagger

☐ Ngram

☐ Sentiment Analysis

☐ Named Entity Recognition

☐ Topic Modelling

Explorer View

All tools contain a link to a Tool Description, which in itself contains a link to the underlying analysis software.

As an example, if you click on the **Ngram** tool you will see the visualisation options that are available for the tool, on the LH side, as well as accessing the tabular data for each.

Options to configure the tool are in the main pane.

In this instance I have amended my **Ngram Size** to a min of 5 and have decided to select the Word Cloud visualisation option.

EXPLORER ✕

Ngram

Tool Description

CONTENT SET
Indicators of Global Warming in the Natural World

RUN HISTORY
Mon Oct 22 08:46:56 EDT 2018 ▾

RUN STATUS
⊘ Completed

TOOL
Configuration

VISUALIZATIONS
Ngram
◯ Word Cloud
◯ Bar Chart
◯ Tabular Data

TOOL CONFIGURATION

Ngram Size

Sets the minimum and maximum size of the ngrams captured in the output (i.e. unigram, bigram, trigram, etc.). The ngram size must be within the range of 1 - 6.

MIN 5   Default: 1
MAX 6   Default: 4

Stop words

A list of words that are excluded from analysis and will not appear in any outputs. All stop words are normalized for case, so different capitalizations do not need to be included (Eg: "above" appears in the list only once, because "Above" would be treated the same as "above").

Reset To Default   Clear All

a
about
above

Explorer View & Tools Continued…

You can return to your Analyze page to view a different set of visualisations using another tool.

If you select the **Topic Modelling** tool you will see that there are various layers of visualisation available.

For example, the visualisation on the right shows the **Tree Map** under the **Topic** level

You can also select visualisations at the Topic-Documents, Document-Topics, Topic Word Count and Network level
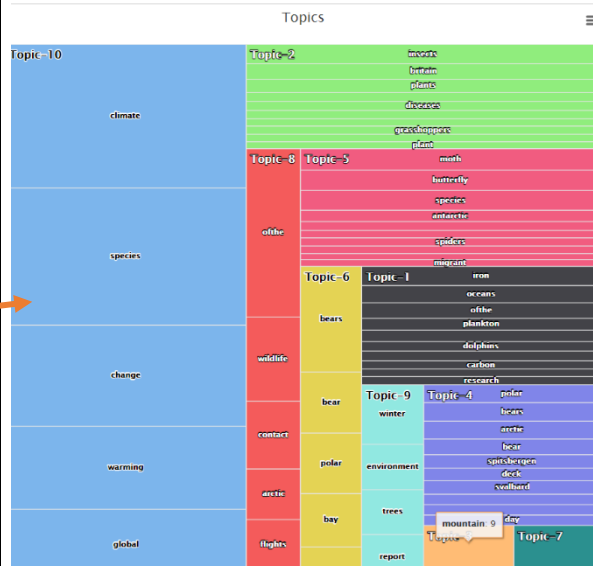
The Topic-Documents Tree Map visualisation has the added benefit of linking you back to a specific document in your content set – so if you wanted to read the primary source or check the OCR text of a document you can do so.

Here you can click on the title of the document and then click on the white pop up box that returns you to the **Doc Explorer** view
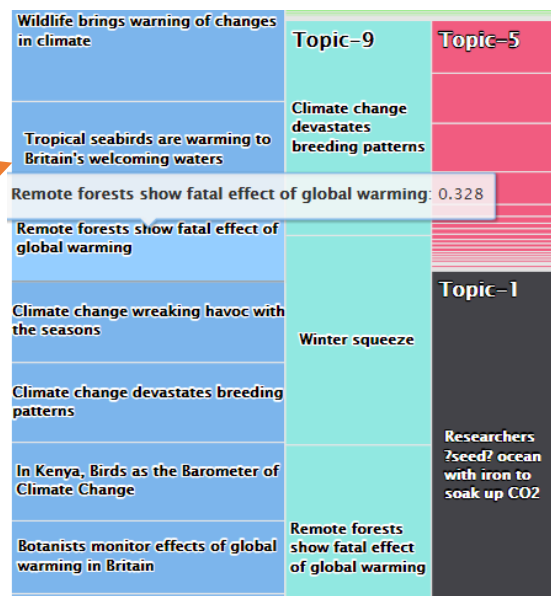
The Dashboard View Continued

The **Analyses** part of the Dashboard View makes more sense now you have used the Analysis Tools.

Indicators of Global Warming in the Natural World
Newspaper articles and editorials relating to global warming.

| Overview | Documents (124) | Analyses | Search History |

ARCHIVES USED (12)
The Independent Digital Archive (59)
The Times Digital Archive (33)
The Sunday Times Digital Archive (6)
The Telegraph Historical Archive (5)
Daily Mail Historical Archive, 1896-2004 (4)
View More

DOCUMENT TYPE (7)
Article (112)
Essay (5)
Editorial (3)
Letter to the editor (2)
Conference news (1)
View More

AUTHORS (76)
Steve Connor, Science Editor (7)
Lewis Smith (7)
Michael McCarthy, Environment Editor (7)
Michael McCarthy (5)
Nick Nuttall (2)
View More

SOURCE LIBRARIES (12)
The Independent (59)
Times Newspapers Limited (40)
Telegraph Media Group (5)
Associated Newspapers Limited (4)
The New York Times Company (4)
View More

By clicking here users will return to the results of their data analyses using the Lab's text mining tools.

Indicators of Global Warming in the Natural World
Newspaper articles and editorials relating to global warming.

| Overview | Documents (124) | Analyses | Search History |

Clustering
Tabular Data   Scatter Plot

Clustering
Tabular Data   Scatter Plot

Named Entity Recognition
Hierarchical Tree   Tabular Data

Ngram
Tabular Data   Word Cloud   Bar Chart

Ngram
Tabular Data   Word Cloud   Bar Chart

Parts of Speech Tagger
Tabular Data   Series View

Finally, the Dashboard view contains a **Search History** tab. This lists all the searches that have been made in order to create your content set. You can link to these search results pages with a single click, to continually refine and update **your content set**.

SEARCH	RESULTS	DOCS ADDED	TIMESTAMP
Basic Search ("global warming and climate change")	65	5	Thu Oct 18 09:46:11 EDT 2018
Basic Search (insects And climate change)	10,870	88	Mon Oct 22 09:40:28 EDT 2018
Basic Search (insects And climate change) LIMITS: Publication Date (Jan 01, 1990 - Sep 30, 2018)	831	15	Mon Oct 22 09:43:54 EDT 2018
Basic Search ("effects of global warming") LIMITS: Entire Document (creatures)	26	7	Thu Oct 18 09:50:05 EDT 2018
Basic Search (insects And climate change)	10,870	7	Mon Oct 22 05:55:42 EDT 2018
Basic Search ("effects of global warming")	621	1	Thu Oct 18 09:48:27 EDT 2018

GALE DIGITAL SCHOLAR LAB