



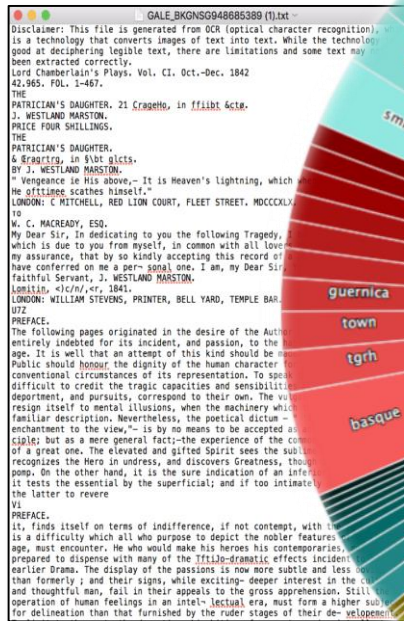
Gale Digital Scholar Lab

Gale数字学术实验室

DIGITAL SCHOLAR LAB

Explore your institution's holdings from Gale Primary Sources with new eyes,
using digital humanities to discover new insights on historical materials

内容 + 工具



在Gale数字学术实验室中，内容与工具彼此靠近，对许多使用GPS内容的研究者而言是翻天覆地的创新……

The Home Page and Search

主页和检索功能

Google or
Microsoft...and save
your content there.
谷歌或微软账号授权登陆.....
在此保存你所有的内容

Search is simple, but a
significant leap in
corpus building!
检索很简单，但却是构建语
料库的重要一步！

The screenshot shows the Digital Scholar Lab homepage. At the top, there is a navigation bar with links for 'Logout', 'My Library Links', 'English', and a 'Signed in: Marc' button highlighted with a red box. Below the navigation bar, the main header features the 'Digital Scholar Lab' logo and a search bar. The search bar is highlighted with a red box and contains the text 'Search using keywords and boolean operators (OR, AND, NOT)...'. Below the search bar, there are three main sections: 'Build your corpus', 'Analyze your documents', and 'Manage and share'. Each section has a brief description and a link to explore more. At the bottom, there is a footer with the 'GALE DIGITAL' logo and a list of links: 'About', 'Contact Us', 'Copyright', 'Terms of Use', and 'Privacy Policy'.

Goals of Lab are clearly
spelled out:

1. Content set building
2. Text analysis
3. Organisation

实验室的目标明确呈现:

1. 创建内容集
2. 文本分析
3. 组织研究

GDSL Features: Platform-Wide Features

GDSL特色：平台特色

FEATURE 特色	VALUE POINT 价值点
OMNI architecture OMNI架构	Leverages features from the same platform that hosts Gale Primary Sources 与Gale原始档案平台相同的平台架构
New features 新特色	Brings a host of new tools, workflows, and user interface to the platform 为平台带来一系列的新工具、流程和用户界面
Mobile / responsive design 移动设备响应式设计	Search, build content sets and visualise on mobile devices 在移动设备上检索、创建内容集和可视化结果
Google / Microsoft integration 谷歌/ 微软整合	Authenticate and share individual documents using Google and Microsoft services commonly used for collaboration 利用谷歌和微软常见的协作工具授权和分享单个文档
Scalable design and back end Architecture 扩展性设计和后端架构	The first 7 of many customer -requested open source tools are included at launch. 发布时包括7种很多客户要求的开源工具



The Results Page

检索结果页面

Toggle between content sets with a click.
点击即可在不同内容集间切换

OCR confidence level provides an indication of OCR quality
OCR置信水平表明OCR的质量

Select single documents or all 50 on a single page.
选择单个文档或一个页面上的所有50篇文章

The screenshot shows the Gale Scholar Lab interface. At the top, there's a navigation bar with 'Logout', 'My Library Links', 'English', and 'Signed in: Marc'. Below this is a search bar with the text 'Search using keywords and boolean operators (OR, AND, NOT)...'. The main content area is titled 'All Content (358)'. On the left, there's a sidebar with 'ACTIVE CONTENT SET' (19th C. Mining Editorials), 'MANAGE | ANALYZE', 'Search within results', 'Limit Search by:', 'Data bases', 'Content Type', 'Publication Title', 'Publication Date', 'Publication Language', 'Subjects', and 'Source Library'. The main results area shows a list of documents with columns for 'Select All', 'OCR CONFIDENCE', 'ARCHIVE DOCUMENT TYPE', 'CONTENT TYPE', 'DATE', 'PUBLICATION', 'PUBLISHER', 'SOURCE LIBRARY', and 'SUBJECTS'. The first document is 'Municipal Water Supply' with an OCR confidence of 50%. The second is 'To Correspondents' with an OCR confidence of 46%. The third is 'The Geological Survey of India' with an OCR confidence of 46%. The fourth is 'Prospects of the Harvest' with an OCR confidence of 35%. The fifth is 'The Rotherhithe Seizure' with an OCR confidence of 53%.

Extended metadata elements give users high-level details about individual documents that could impact whether or not it is chosen for a content set and/or analysis

扩展的元数据条目为用户提供每篇文档的更多细节，可能会影响它是否被选择加入到内容集和/或分析当中

文档视图

Context-centered guidance and explanation

背景信息指南与解释

文档浏览器：可选择查看内容、发现上下文中的关键词，以及在文档内检索

GDSL Features: Search and Results

GDSL特色：检索与结果

FEATURE 特色	VALUE POINT 价值点
Search result page 检索结果页面	Showcases the TDM relevant information without expanding the document View 无需打开文档就可查看文本数据挖掘的相关信息
Basic and advanced search options are included, similar to GPS 包括简单与高级检索选项，类似GPS	Early stage projects may use the basic search to explore. Developed projects Refine search through filtering options 项目早期可以利用简单检索发现内容。已开展的项目可利用筛选选项改进检索结果
Name / set Active Content Set 名称/有效内容集	Users define which project (or “bucket”) they add documents to 用户定义他们要将文档加入到哪个项目中
Contextual messaging (“Added”) 标识信息	Easily identify which documents are in use per content set 轻松识别每个内容集中使用了哪些文档
Document metadata at-a-glance 文档元数据速览	More robust metadata information about each document 关于每篇文档的更可靠的元数据信息
OCR blurb & confidence rating OCR信息和置信评级	See at a glance the quality of OCR and decide whether its valuable for analysis 速览OCR的质量，并决定它对分析过程的价值
The Doc Explorer and document view 文档浏览器和文档查看	Text & original document scan (side-by-side view), review metadata and quality of document pertaining to its OCR text 文本与原始文献资料并排显示，检查元数据和该文档OCR文本的质量



Managing Content Sets

管理内容集

Name and
annotate content
sets for
organisational
detail
内容集名称和注释，
便于组织管理

At-a-glance overview
provides students and
researchers valuable details
into the construction of a
content set, including:
快速浏览视图为学生和研究者提供内
容集架构的细节，包括：

- Source archives
来源档案库
- Document type
文档类型

Users can download a saved
content set for use outside
of the Digital Scholar Lab.
用户可以下载保存的内容集，在Gale
数字学术实验室以外使用。

The screenshot displays the Digital Scholar Lab interface. At the top, there's a navigation bar with 'Logout', 'My Library Links', 'English', and 'Signed in: Marc'. The main header reads 'DIGITAL SCHOLAR LAB From Gale'. Below this, the title 'Luftwaffe in the British Press 1938-1950' is shown with a description: 'Designed to support a number of tools that will extract key terms as well identify their frequency, etc.'.

The interface features several tabs: 'Overview' (selected), 'Documents (523)', 'Outputs', and 'Search History'. The 'Overview' tab is divided into three main sections:

- SOURCE ARCHIVES USED (10)**: A list of archives including The Telegraph Historical Archive (322), People Post Historical Archive (86), The Independent Digital Archive (53), The Sunday Times Digital Archive (44), International Herald Tribune Historical Archive 1887-2013 (6), The Listener Historical Archive, 1929-1991 (6), The Economist Historical Archive, 1843-2014 (3), Archives of Sexuality & Gender (1), British Library Newspapers (1), and The Illustrated London News Historical Archive, 1842-2003 (1).
- DOCUMENT TYPE (3)**: A list of document types including Article (523), Interview (3), and Report (1).
- AUTHORS (269)**: A list of authors including Marsha Dunstan (15), John Keegan (7), Maurice Weaver (6), and Andrew Gimson (6).

Below these sections is a 'SOURCE LIBRARIES (10)' section listing various libraries like Cengage Learning (322), Getty Images (86), The Independent (53), Times Newspapers Limited (44), The New York Times Company (6), British Broadcasting Corporation (6), The Economist Newspaper Limited (3), Lesbian Herstory Archives (1), Illustrated London News Ltd (1), and British Library (1).

A modal window titled 'Download Content Set' is open in the center. It explains that due to the size of the content sets, the downloadable file must first be generated, which can take up to a few hours. It states that once ready, an email with the download link will be received. The modal includes a 'CONTENT SET' section showing 'Unions + 19th Century' with '5,243 Documents'. At the bottom of the modal are two buttons: 'CANCEL' and 'GENERATE DOWNLOAD'. Below the buttons, a disclaimer states: 'By clicking Generate Download you agree to the terms of use for this content as agreed to by your licensed institution.'

Refining Your Content Set

完善你的内容集

Users can further refine their content set by using limiters and selecting articles/documents to be deleted.

用户可以通过使用限定选项或选择要删除的文章或文档，进一步完善他们的内容集

Users can also duplicate and/or version their content sets to analyse with similar or new tool configurations to create an array of outputs.

用户也可以复制和/或添加内容集的不同版本，利用类似或新的工具设置进行分析，形成一组结果。

Logout My Library Links English Signed in: Marc

DIGITAL SCHOLAR LAB
From Gale

Search Analyze My Content Sets

Luftwaffe in the British Press 1938-1950

Designed to support a number of tools that will extract key terms as well identify their frequency, etc.

Overview Documents (523) Outputs Search History

Documents

Document Type

Source Library

Author

Document Type

Source Library

Author

Select All	Document Title	OCR Confidence	Source Library	Content Type	Document Type	Date	Publication	Publisher
☐	Down Comes a Nazi Bomber	94%	Getty Images	Newspaper	Article	Apr. 19, 1941	Picture Post.	
☐	Continent owes a debt to Britain	80%	Getty Images	Newspaper	Article	Feb. 18, 1957	Picture Post.	
☐	Can Italy Be Knocked out of the War?	89%	Getty Images	Newspaper	Article	Nov. 21, 1942	Picture Post.	
☐	The Army's Airmen	92%	Getty Images	Newspaper	Article	Oct. 25, 1941	Picture Post.	

Viewing Existing Outputs

查看现有结果

Output dashboard includes task name as well as visualisation and output options

结果面板包括任务名称以及可视化和结果选项。

The screenshot shows the SCHOLAR LAB interface. At the top, there's a navigation bar with 'Logout', 'My Library Links', 'English', and 'Signed in: Marc'. Below this is a search bar and 'Search', 'Analyze', and 'My Content Sets' buttons. The main header area displays the task name 'the British Press 1938-1950' and a brief description: 'A number of tools that will extract key terms as well identify their frequency, etc.' Below the header, there are three tabs: 'Documents (523)', 'Outputs' (which is selected), and 'Search History'. The 'Outputs' tab is divided into six panels, each representing a different analysis tool: 'Clustering', 'Named Entity Recognition', 'Ngram', 'Parts Of Speech', 'Term Frequency', and 'Topic Modelling'. Each panel contains icons for different visualization or output options. For example, the 'Clustering' panel has 'Tabular Data' and 'Scatter Plot' options. The 'Parts Of Speech' panel has 'Tabular Data' and 'Series View' options. The 'Term Frequency' panel has 'Tabular Data', 'Word Cloud', and 'Bar Chart' options. The 'Topic Modelling' panel has 'Tabular Data', 'Tree Map', 'Sunburst Wheel', and 'Network Diagram' options. An orange line points from the text on the left to the 'Tabular Data' icon in the 'Clustering' panel. Another orange line points from the text on the right to the 'Tabular Data' icon in the 'Ngram' panel.

Clicking any Output option takes users to the visualisation page, including configuration and visualisation options as well as options to for image download and tabular data download

点击任意结果选项，用户将进入可视化页面，包括设置和可视化选项，以及图片下载和表格数据下载

GDSL Features: Content Sets

GDSL特色：内容集

FEATURE 特色	VALUE POINT 价值点
Curate custom content sets on the fly 随时管理自定义内容集	Users select and organize documents relevant to their research 用户选择和组织与他们的研究相关的文档
“Select all” 全选	Users can add an entire page’s results to a content set at once 用户可以将整页检索结果一次性添加到内容集中
Dashboard view 面板视图	Makes high_level details about content sets, outputs, and search history available at a glance 可快速浏览内容集、结果和检索历史的众多细节
Content set search history 内容集检索历史	Search history is important for reproducibility, as well at the peer -review process 检索历史对于重现性和同行评审均很重要
Content set download (OCR) 内容集下载（OCR）	Users can download her content set for use Elsewhere 用户可以下载他的内容集，在其他地方使用



Tools

工具

The screenshot displays the Digital Scholar Lab interface. At the top, there's a navigation bar with 'Logout', 'My Library Links', 'English', and 'Signed in: Marc'. Below this is a search bar and links for 'Search', 'Analyze', and 'My Content Sets'. The main content area is titled 'Analysis Tools' and lists four tools: Ngram, Topic Modelling, Named Entity Recognition, and Parts Of Speech. Each tool has a description, example outputs (like Tabular Data, Bar Chart, Tree Map, Sunburst Wheel, Network Diagram, Tree View), and an 'ADD' button. The interface is clean and professional, with a blue header and a white background for the content area.

Logout My Library Links English Signed in: Marc

DIGITAL SCHOLAR LAB
From Gale

Search Analyze My Content Sets

Back to Analyze

Analysis Tools

Ngram

Ngrams are a type of collocation where words appear next to or in the proximity of others. When computing n-grams you typically move one word forward to show co-occurring words. Therefore, an n-gram is a contiguous sequence of n items from a given word within a document. An ngram of size 1 is called a unigram, size 2 is a bigram, size 3 is a trigram, size 4 is a four-gram, and so on. [LEARN MORE](#)

ADD

EXAMPLE OUTPUTS
Tabular Data Bar Chart

Topic Modelling

Topic modelling allows you to analyze a large corpus of unstructured (OCR) text. A "topic," often referred to as a "bag of words," is a collection of terms that frequently co-occur in your collection of documents. Mallet uses latent Dirichlet allocation (LDA) models to extract contextual clues in order to connect words with similar meanings, as well as differentiate between words that are spelled similarly but have differing meanings. Mallet will provide you not only with a list of possible "topics," but a ranked listing of words for each topic, which documents they appear in and their importance within a text as well as the overall corpus, and the matching statistical measures. [LEARN MORE](#)

ADD

EXAMPLE OUTPUTS
Tabular Data Tree Map Sunburst Wheel Network Diagram

Named Entity Recognition

Counts the specified terms/phrases and returns their distribution across a collection. Stemming and lemmatization are performed to handle variants. The NLTK library is used for creating term/phrase tokens. [LEARN MORE](#)

ADD

EXAMPLE OUTPUTS
Tree View Tabular Data

Parts Of Speech

Counts the specified terms/phrases and returns their

EXAMPLE OUTPUTS
Tabular Data Bar Chart

QUANTITATIVE
Analysis methods
at release:
发布时的定量分析方法:

- Frequency 频数
- Ngrams
- Named Entity Recognition 命名实体识别

QUALITATIVE
Analysis methods
at release:
发布时的定性分析工具:

- Topic Modelling 主题建模
- Parts-of-Speech 词性标注
- Sentiment Analysis 情感分析
- Clustering 聚类

Analysis Method: Named Entity Recognition (NER)

分析方法：命名实体识别（NER）

- NER recognises and extracts Named Entities from documents within a content set, and output lists of entities.
NER从内容集的文档中识别并提取命名实体，然后输出实体列表。
- Not designed to analyse textual features, but to extract them, and create new datasets which can then be used to contextualize content sets.
并非设计为分析文本特征，而是提取它们，并创建新的数据集，随后用于分析内容集的背景
- Example “entity types”: “实体类型” 举例：
 - people (including fictional), 人物（包括虚构人物）
 - groups (nationalities, religious, or political), 团体（国籍、宗教或政治）
 - organisations (companies, agencies, institutions), 组织（公司、机构、协会）
 - locations (countries, states, cities), 地点（国家、州、城市）
 - products (objects, vehicles, foods, etc.), works of art (titles of books, songs, etc.), 产品（物品、车辆、食物等）、艺术作品（书名、歌曲名等）
 - dates (absolute or relative dates or periods) 日期（绝对或相对日期或时代）



Back to Analyze

EXPLORER

Named Entity Recognition

Counts the specified terms/phrases and returns their distribution across a collection. Stemming and lemmatization are performed to handle variants. The NLTK library is used for creating term/phrase tokens. [LEARN MORE](#)

CONTENT SET

Guernica

RUN STATUS

Complete

TOOL

VISUALIZATIONS

Named Entity Recognition

Tree View

Tabular Data

EXPLORER

Named Entity Recognition

Counts the specified terms/phrases and returns their distribution across a collection. Stemming and lemmatization are performed to handle variants. The NLTK library is used for creating term/phrase tokens. [LEARN MORE](#)

CONTENT SET

Guernica

RUN STATUS

Complete

TOOL

VISUALIZATIONS

Named Entity Recognition

Tree View

Tabular Data

- ORG
- PERSON
 - Shafrazi
 - Tony
 - Francis Bacon
 - John Edwards
 - Picasso
 - Bacon
 - Guernica
 - Manuel Vazquez
 - Francisco Franco
 - Vet Franco
 - Thev
 - Ramon Serrano Suno
 - Enrique Gonzalez Duro
 - Hitler
 - ovl
 - Franco
 - Sack
 - Elizabeth Jane Howard
 - Foley Street
 - Guernica Restaurant
 - Rioja
 - HILARI
 - Eduardo González Calleja
 - Pedro Sainz
 - SIAl
 - Julio Aróstegui
 - Mussolini
 - FZDWMU954345480
 - QHVEHI881370669
 - EL1800065945
 - ICMFHO446993421
 - PFJCYN755563817
 - GNRAVJ949229430
 - YYISRS361664034
 - LJOYOP986455907
 - Guglielmo Marconi
 - Adwa
 - Viñas



Analysis Method: Topic Modelling

分析方法：主题建模

- To analyse of a large corpus of unstructured (OCR) text. 分析大型非结构化文本（OCR）语料库

- A “topic,” is a collection of terms that frequently co-occur in your collection of documents.

“主题”就是频繁出现在你的文档集中的词语集合

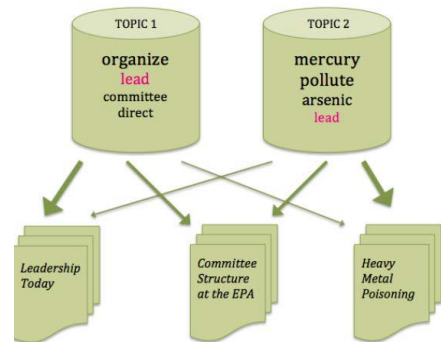
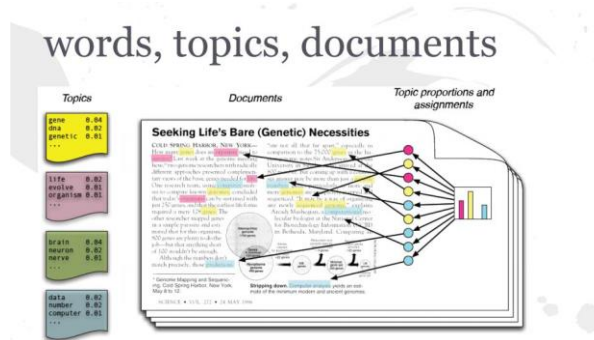
- Mallet extracts contextual clues to: Mallet提取上下文线索

- connect words with similar meanings (i.e. synonyms: Happy: content, joyful, mirthful, upbeat

关联类似含义的词语（即同义词，例如高兴：满意、快乐、欢乐、乐观）

- differentiate words that are spelled similarly but have differing meanings (i.e. homographs: Bat: piece of sports equipment OR an animal).

区别拼写相近但意思不同的词语（即同形异义词，例如Bat，一种体育用品或一种动物）

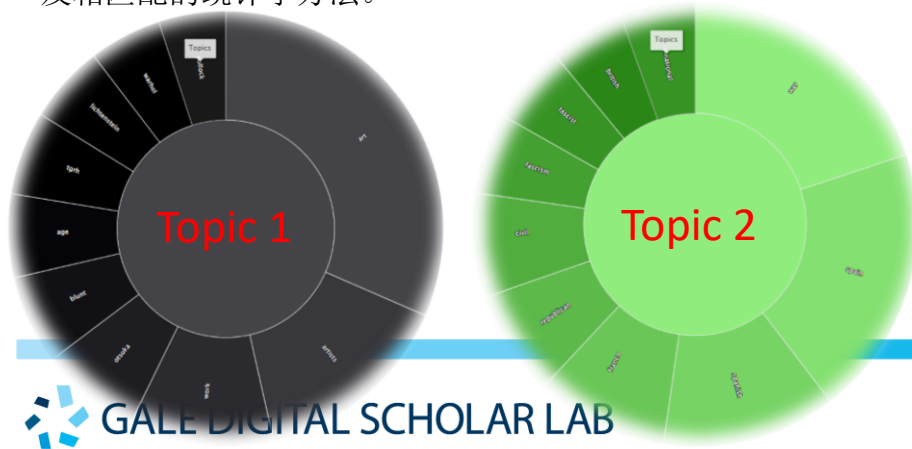


Analysis Method: Topic Modelling

分析方法：主题建模

Mallet provides a list of possible ‘topics’, and a ranked listing of words for each topic, which documents they appear in and their importance within a text as well as the overall corpus, and the matching statistical measures.

Mallet提供一系列可能的“主题”，以及每一主题下词语的排名，记录了它们在一个文本或整个语料库中出现的次数及重要性，以及相匹配的统计学方法。



DIGITAL SCHOLAR LAB
From Gale

Search Analyze My Content Sets

Back to Analyze

EXPLORER

Topic Modelling

Topic modelling allows you to analyze a large corpus of unstructured (OCR) text. A "topic" often referred to as a "bag of words," is a collection of terms that frequently co-occur in your collection of documents. Mallet uses latent Dirichlet allocation (LDA) models to extract contextual clues in order to connect words with similar meanings, as well as differentiate between words that are spelled similarly but have differing meanings. Mallet will provide you not only with a list of possible topics, but a ranked listing of words for each topic, which documents they appear in and their importance within a text as well as the overall corpus, and the matching statistical measures. [LEARN MORE](#)

CONTENT SET

Guernica

RUN STATUS

Complete

TOOL

Configuration

VISUALIZATIONS

Topics

Sunburst Wheel

Tree Map

Tabular Data

Topic-Documents

Sunburst Wheel

Tree Map

Tabular Data

Document-Topics

Sunburst Wheel

Tree Map

Tabular Data

Topic Word Count

Tree Map

Tabular Data

Network

Network Diagram

Tabular Data

	A	B	C	D	E	F
	topic_id	distributio	tokenCour	word		word count
2	1	0.041265	2809	art		76
3	1	0.041265	2809	artists		36
4	1	0.041265	2809	work		26
5	1	0.041265	2809	otsuka		18
6	1	0.041265	2809	blunt		16
7	1	0.041265	2809	age		15
8	1	0.041265	2809	tgrh		15
9	1	0.041265	2809	lichtenstein		14
10	1	0.041265	2809	warhol		13
11	1	0.041265	2809	pollock		12
12	2	0.231985	8190	war		165
13	2	0.231985	8190	spain		163
14	2	0.231985	8190	spanish		105
15	2	0.231985	8190	franco		80
16	2	0.231985	8190	republican		64
17	2	0.231985	8190	civil		63
18	2	0.231985	8190	fascism		49
19	2	0.231985	8190	fascist		48
20	2	0.231985	8190	british		46
21	2	0.231985	8190	international		45

topics

分析方法：主题建模

Topics

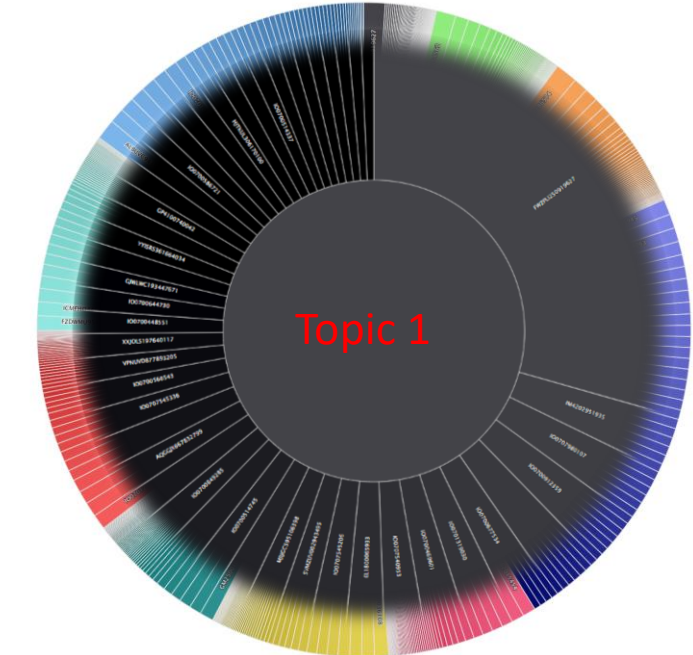
- ## Topic-Documents

- ## Document-Topics

- ### Topic Word Count

- ## Network

- Network Diagram
- Tabular Data



Analysis Method: Topic Modelling

分析方法：主题建模

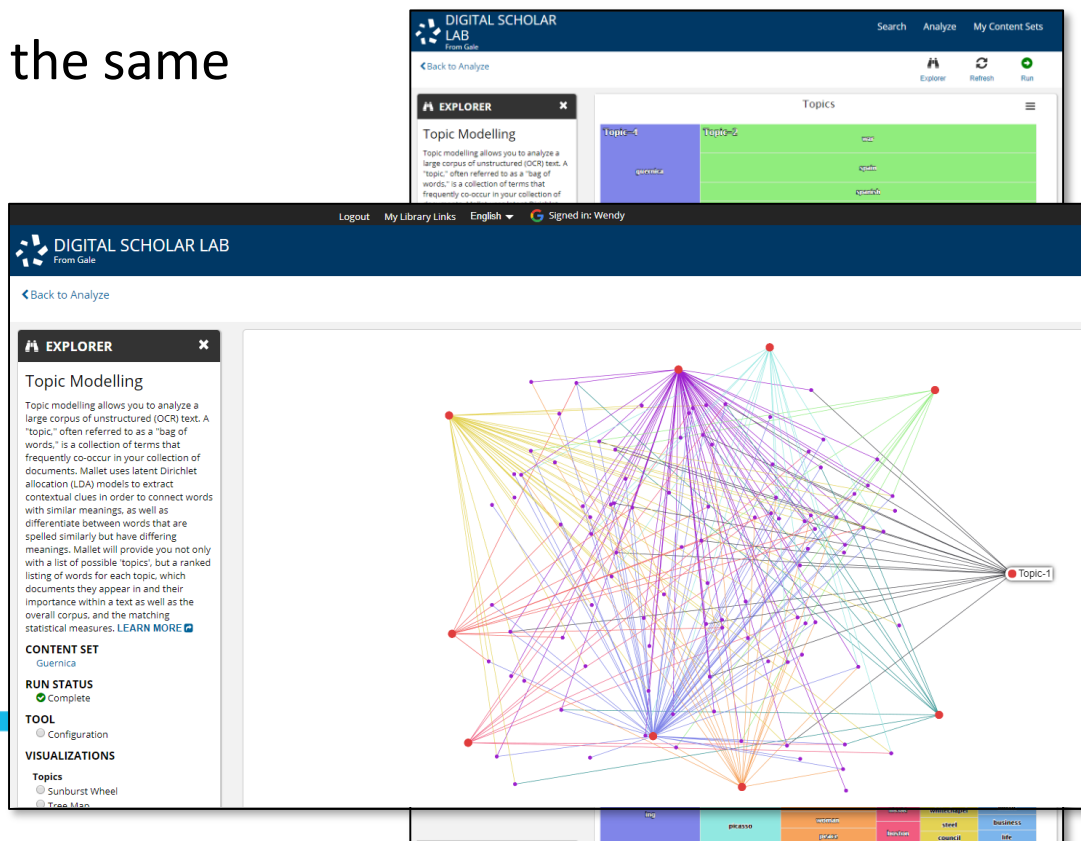
Different visual outputs for the same analysis tool.

同一分析工具的不同视觉输出。

- Sunburst Wheel
旭日图
- Tree Map
树形图
- Network Diagram
网络图



GALE DIGITAL SCHOLAR LAB



GDSL Features: Tools and Visualisations

GDSL特色：工具与可视化

FEATURE	VALUE POINT
Analysis tool configurations 分析工具配置	Users decide to use predefined, tool -specific default configurations or custom configurations 用户决定使用每个工具预设的默认配置，或自定义配置
User -friendly tool descriptions 易懂的工具说明	Users select which tool will be most helpful for addressing research Questions 用户选择哪种工具对解决研究问题最有帮助
Links to documentation 文档链接	Users can review open -source tool documentation to learn specifics about the various tools available 用户可以检查开源工具文件，了解现有各种工具的详情
Contextual error messaging 上下文错误信息	Show errors that prevent the successful run of a job 显示运算不能顺利完成的错误信息
Multiple visualization outputs per analysis tool 每种分析工具的多种可视化输出	Provides multiple views to facilitate interpretation of the analysis Outputs 提供多种视图，帮助解读分析结果
Access to tabular data output 访问表格数据结果	Images as well as tabular data generated by the analysis and used to derive a visualization are available for download 由分析过程产生的以及用于生成可视化图形的图像和表格数据都可以下载

