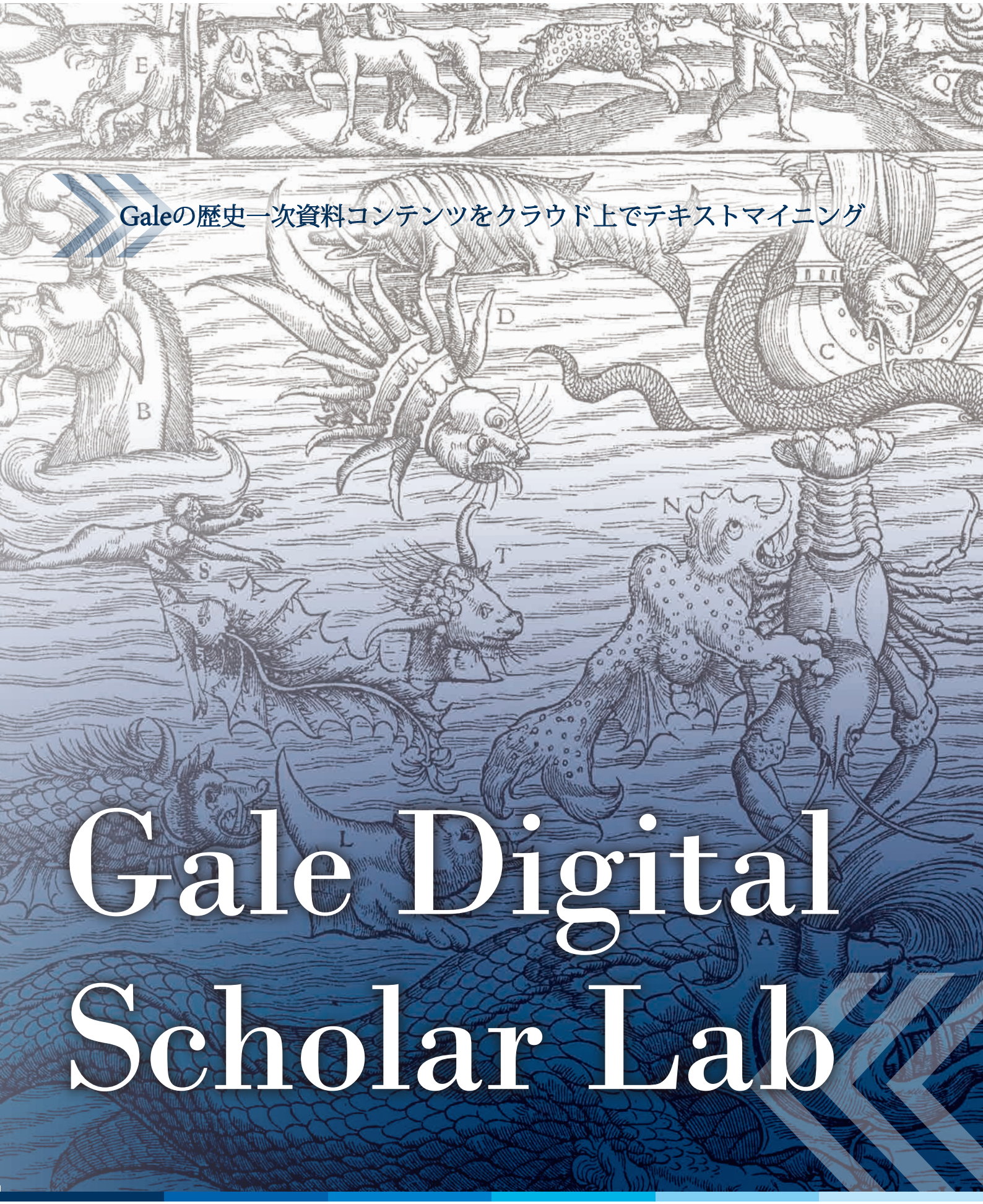
GALE
A Cengage Company



Galeの歴史一次資料コンテンツをクラウド上でテキストマイニング

Gale Digital Scholar Lab

EMPOWER™ DISCOVERY

Gale Digital Scho

新しい方法論としての「デジタル人文学」

近年、英米を中心に一次資料のデジタル化・データベース化がすすみ、歴史・文学・社会・法律・経済など人文・社会科学系の研究者は、従来からは考えられないほど多彩かつ膨大な一次資料に比較的簡便にアクセスできるようになりました。一方で、新たな研究手法として、膨大な一次資料データにテキストマイニングや自然言語処理など情報科学のノウハウを応用することにより、資料群をコーパスとして統計的・俯瞰的に分析する、いわゆる「デジタル人文学」や「人文情報学」も盛んとなってきています。

研究者の参入をはばむ技術的・実務的障壁

しかしながら、そうしたデジタル人文学の分析手法は多くの場合、既存のデータベースの枠組みの外で行われるため、多くの研究者にとっては、プログラミング技術やノウハウの習得、分析対象となる一次資料元データの探索・収集、版権所有者への許諾申請、データ形式の整理と統一、ローカル・サーバーへのホスティングと適切な管理、分析結果のビジュアル化、他の研究者との共有など、数多くのハードルが存在し、ごく一部の研究者やプロジェクトチームをのぞいて、多くの人文系研究者、とりわけ歴史系の研究者にとっては敷居が高く感じられる分野であることも事実です。

Gale コンテンツのテキストマイニング・プラットフォーム

そこで Gale では、18 世紀英語文献の「Eighteenth Century Collections Online (ECCO)」、イギリスを代表する新聞の「Times Digital Archive」、社会経済史文献の「The Making of the Modern World (MOMW)」など、これまで数多くのデジタル一次資料群を構築・リリースしてきた経験を生かし、デジタル人文学を実践する研究者の助言やフィードバックのもと、弊社データベースのコンテンツについて、オンライン上で直感的にテキストマイニングを行うことができる新プラットフォーム「Gale Digital Scholar Lab」をリリースしました。

より広いユーザー層にデジタル人文学の底辺を拡大

「Gale Digital Scholar Lab」により、歴史分野で定評のある Gale のアーカイブ・コンテンツをコーパスとするテキストマイニングがオンライン上で、手軽にできるようになるばかりでなく、作成した分析結果の共有、履歴の保存や修正などもクラウド上で行うことができるようになり、デジタル人文学的手法に興味がありながらも、技術面の抵抗から躊躇してきた研究者や、院生・学部生を含め、テキストマイニングの裾野がいっそう広がることが期待されます。また、データ形式がすでに統一されており、定評あるオープンソース・ツールを多く採用し、OCR テキストのダウンロードも可能にしているため、すでにデジタル人文学を実践している研究者も、Gale コンテンツを用いた分析への糸口として利用できます。

▶ デジタル人文学の典型的なワークフロー

分析対象となる一次資料を探す

↓

データ所有者に利用許可を申請し、データを入手する

↓

データの保管・ホスティング場所を確保し、
管理・バックアップ方式等を決める

↓

データ形式やメタデータを整理・統一し、
クリーンアップする

↓

データ分析用のアルゴリズムにかけ、
試行と修正を繰り返す

↓

分析結果を検証し、発表用に整理または視覚化する

▶ Gale Digital Scholar Lab のワークフロー

資料を Lab 内で検索し、
分析対象とするコンテンツセットを作る

↓

適用したい分析ツールを選び、分析にかける

↓

分析結果を検証し、ダウンロードする

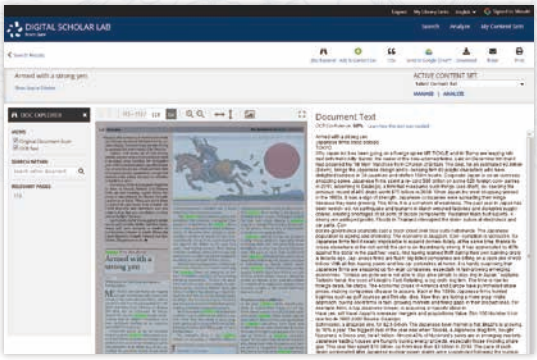
lar Lab

GALE
A Cengage Company

Gale Digital Scholar Lab の特色

- Gale 社の一次資料 1 億 6600 万ページ分のテキストデータを搭載可能
- プログラミング等の知識は不要
- OCR テキストをダウンロード可能
（1セッションあたり 1000 件まで）
- 各文献について、OCR テキストと元画像を対照表示
- トピック・モデリング、クラスタリング、N グラム頻度、固有表現抽出など、6 種類の分析ツールをオンラインで適用可能（次ページ参照）
- 分析結果のエクスポートも可能
- 検索履歴・分析ツール使用履歴等を保存、再現可能
- Google、Microsoft の各ログインが利用可能
- クラウドベースで管理は簡単

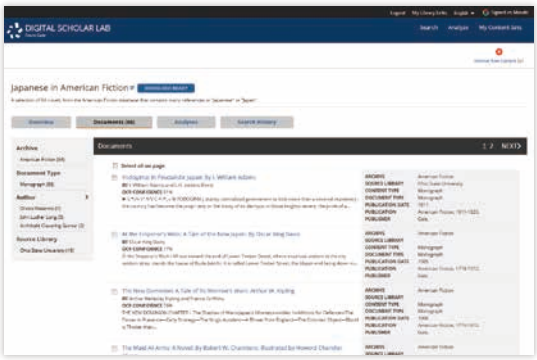
ホーム画面



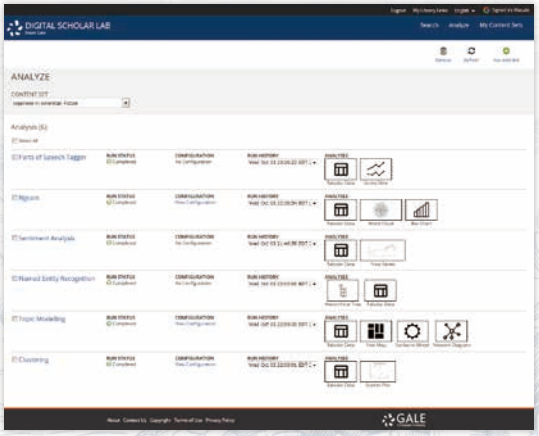
本文画像と OCR テキストの対照表示

Gale Digital Scholar Lab のメリット

- Gale の定評あるコンテンツ群のテキストを簡単に分析にかけることができる
- プログラミングなどの技能を必要としないため、幅広い研究者や学生が利用することが出来る
- 個々の文献の閲読からは見極めにくい、文書群全体に頻出する用語やテーマなどを抽出・分析することができる
- 分析結果を簡単に視覚化し、共有・発表できる
- 実行履歴から分析結果を繰り返し再現することができる
- 統一されたデータ形式でメタデータも揃っているため、データ最適化に時間を費やす必要がない
- クラウドベースのため、データの保管やメンテナンスにかかる手間がはぶける
- デジタル人文学の入門講座等でも利用できる
- OCR テキストのダウンロードにより、外部ツールを用いた分析ニーズにも対応できる
- 日本語文献や国内所蔵文献にかたよりがちなデジタル人文学プロジェクトを西欧文献に広げることができる
- 図書館によるサポートも簡単にできる
- 図書館で所蔵している Gale コンテンツを生かした新たな研究成果の発信が期待できる



コンテンツセット編集画面



分析ツール選択画面

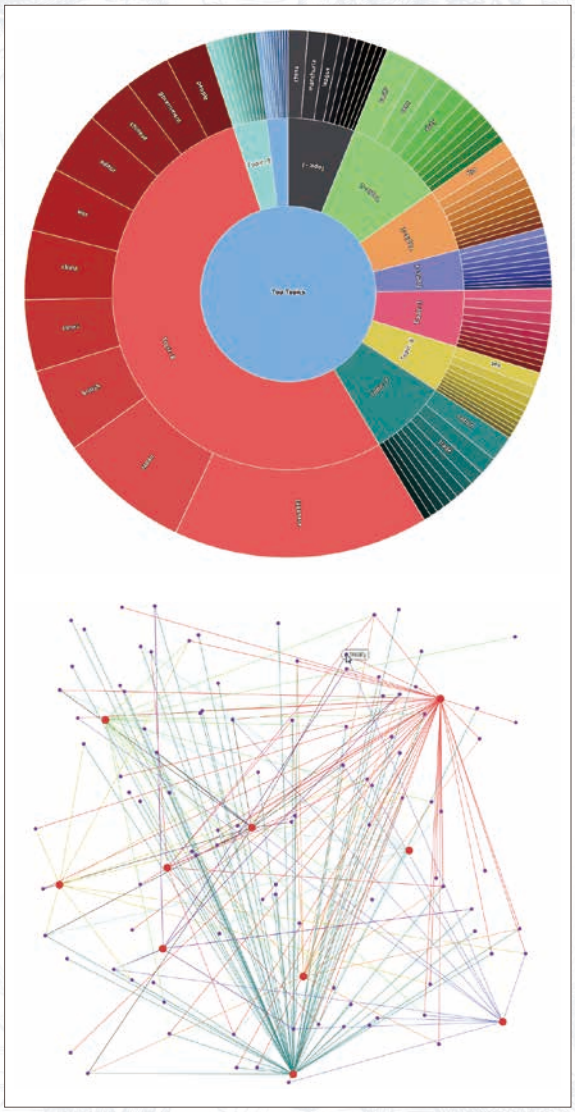
GALE DIGITAL SCHOLAR LAB

Gale Digital Scholar Lab で利用可能な分析ツール

デジタル人文学の研究者によるフィードバックをもとに、テキストマイニングで広く使われている、汎用性の高いツールを採用しました。Gale Digital Scholar Lab 内では、マウスのクリックや簡単な設定で実行できるように工夫されており、プログラミング等の知識は必要ありません。出力形式も視覚的なグラフ形式等だけでなく、表形式も備えています。ツールは今後も、ユーザーからのフィードバックをもとに、新たに追加・改良されていく予定です。

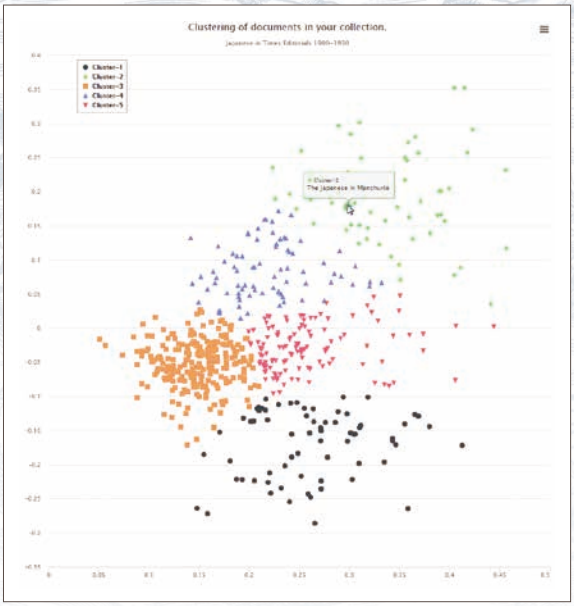
Nグラム頻度
Ngram

元となるライブラリ：Lucene
主な用途：頻出単語・フレーズの抽出
出力形式：ワードクラウド、棒グラフ、表形式



トピック・モデリング
Topic Modelling

元となるライブラリ：Mallet
主な用途：複数の文書に共通するトピック群の抽出
出力形式：円形表示、矩形表示、ネットワーク表示、表形式

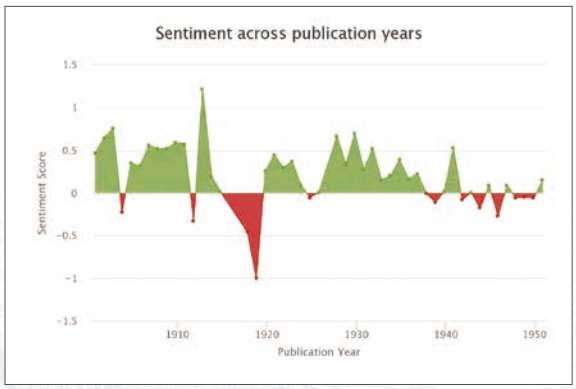


クラスタリング
Clustering

元となるライブラリ：SciKit Learn
主な用途：文書を類縁性で分類
出力形式：座標形式、表形式

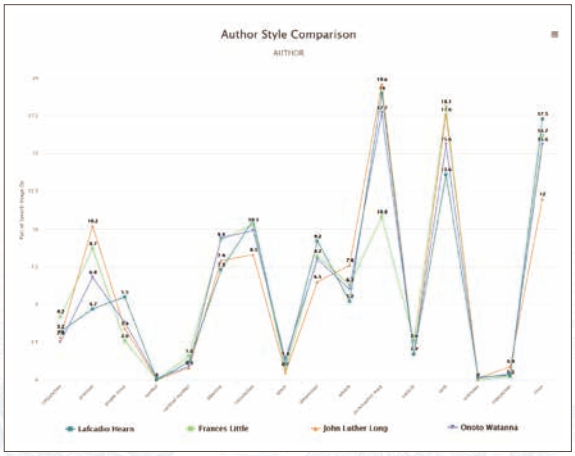
固有表現抽出
Named Entity Recognition

元となるライブラリ：spaCy
主な用途：固有名詞・数字・日付等の抽出
出力形式：階層表示形式、表形式



感情分析
Sentiment Analysis

元となるライブラリ：OpenNLP
主な用途：文書内容が肯定的か否定的かを分析
出力形式：年代別グラフ形式、表形式



品詞タグ付け
Parts-of-Speech Tagger

元となるライブラリ：spaCy
主な用途：著者ごとに品詞の使用頻度を比較
出力形式：著者別グラフ形式、表形式

GALE DIGITAL SCHOLAR LAB

Gale Digital Scholar Lab で利用可能なコンテンツ群

Gale Digital Scholar Lab では、以下の Gale Primary Sources コンテンツのデータを搭載することが可能です。
契約時にご利用いただけるのは、下記のうち、貴館でご導入いただいているコンテンツのみとなります。
あらかじめご了承ください。

総合・文学

American Fiction
Archives Unbound
Eighteenth Century Collections Online (ECCO)
Nineteenth Century Collections Online (NCCO)
The Making of the Modern World
Sabin Americana, 1500-1926

新聞・雑誌

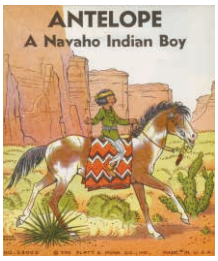
17th & 18th Century Burney Collection
17th & 18th Century Nichols Collection
19th Century UK Periodicals
Nineteenth Century U.S. Newspapers
American Amateur Newspapers
American Historical Periodicals from the American Antiquarian Society
British Library Newspapers
Daily Mail Historical Archive 1896-2004
The Economist Historical Archive 1843-2014
The Illustrated London News Historical Archive 1842-2003
The Independent Digital Archive 1986-2016
International Herald Tribune Historical Archive 1887-2013
Liberty Magazine Historical Archive 1924-1950
The Listener Historical Archive 1929-1991
Picture Post Historical Archive 1938-1957
Punch Historical Archive 1841-1992
The Sunday Times Digital Archive 1822-2016
The Telegraph Historical Archive 1855-2000
The Times Digital Archive 1785-2012
Times Literary Supplement Historical Archive 1902-2013

法律・政治・社会

American Civil Liberties Union Papers, 1912-1990
Archives of Sexuality & Gender
Associated Press Collections Online
Crime, Punishment, and Popular Culture 1790-1920
The Making of Modern Law: Foreign Primary Sources
The Making of Modern Law: Foreign, Comparative and International Law
The Making of Modern Law: Legal Treatises
The Making of Modern Law: Primary Sources
The Making of Modern Law: Trials
Political Extremism & Radicalism
U.S. Declassified Documents Online
U.S. Supreme Court Records and Briefs
Women's Studies Archive

地域研究

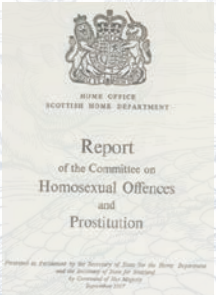
Brazilian and Portuguese History and Culture
China and the Modern World*
Indigenous Peoples of North America
Religions of America
Smithsonian Collections Online



* 旧名称：China from Empire to Republic

今後、新たにリリースされるコンテンツ群も順次追加される予定です。

以下のコンテンツ群は、技術上の障壁や権利者の意向などにより、現時点では搭載できません。
- British Literary Manuscripts Online
- Chatham House Online Archive
- Early Arabic Printed Books
- Financial Times Historical Archive
- National Geographic Magazine Archive
- State Papers Online
- Slavery & Anti-Slavery: A Transnational Archive
- World Scholar: Latin America & the Caribbean



GALE DIGITAL SCHOLAR LAB

Gale Digital Scholar Lab の今後

デジタル人文学の分野において次々と斬新な手法やアプローチが試されているように、Gale Digital Scholar Lab もユーザーからのフィードバックやニーズにあわせて、絶えず改良・改善されていく予定です。直近では、OCR のテキスト・クリーニングツールの追加や、学生向けの教育レイヤーの充実などが予定されています。
進化しつづける Gale Digital Scholar Lab に今後もご注目ください。

■ お問合わせ先

GALE
A Cengage Company

センゲージラーニング株式会社　Gale 部門

〒102-0073 東京都千代田区九段北1-11-11　第2フナトビル　5階
Tel : 03-3511-4135　Fax : 03-3511-4391
E-mail : GaleJapan@cengage.com
gale.com/scholarlab