DIGITIZATION AND PRESERVATION AS AN EDITOR

Marc Cormier
Director, Humanities Publishing
Gale
Farmington Hills, MI

A test case for preservation and digitization:
The Daily Mail Historical Archive

Sourcing archival material
Image Capture
OCR
Metadata
XML
QA

Which Edition?

- There is no single edition!
- Content is usually similar, but “local” advertising, story selection will differentiate each edition.
- Digitizing all editions, while valuable is expensive
- Bringing all editions to a research environment can be messy
- Alternate editions can often be used to fill gaps in coverage of your chosen edition

GALE
A Cengage Company

Balancing priorities of preservation and access

- Unique version of the newspaper published on board the Atlantic liners that carried passengers between the UK and New York.
- The only known set of these issues is held in the Daily Mail offices, and it had long been forgotten and subject to water and rodent damage
- We decided to include these in the project for preservation purposes – otherwise they would have been genuinely lost to history.

GALE
A Cengage Company

Working with the source material: challenges and opportunities

SCANNING FROM ORIGINAL

SCANNING FROM MICROFILM

Working with the source material: challenges and opportunities

SCANNING FROM MICROFILM

- Many newspapers have been microfilmed, as part of preservation efforts from the 1960s onwards.
- Digitising from microfilm is cheaper and logistically easier than scanning physical copies
- Although it does mean that you will only get black & white images.

SCANNING FROM ORIGINAL

- Allow you to create high-quality colour images using the latest scanning technology
- it requires a VERY understanding source library to allow you to handle the bound volumes.
- For optimal results, you need to be able to dis-bind the volumes to avoid getting curvature on the image.

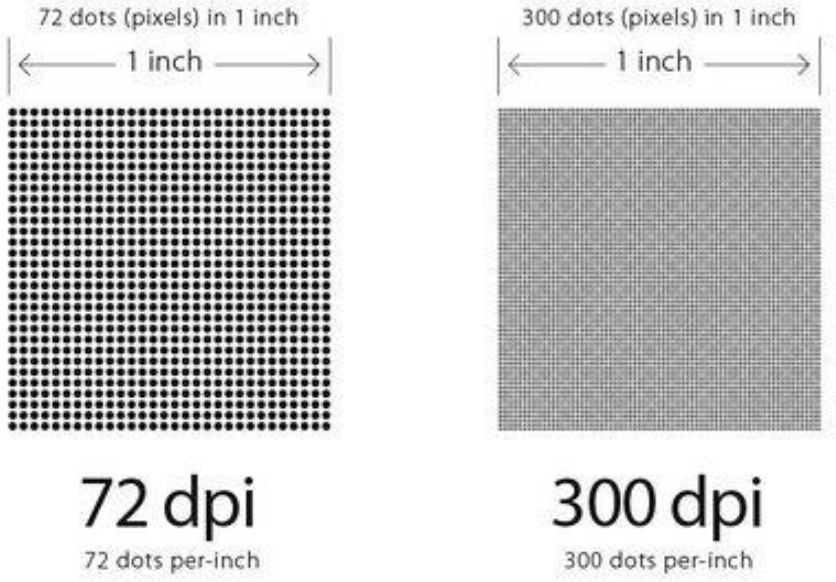


Image courtesy of www.zenfolio.com

Image Capture

- Higher DPI = higher clarity, but it also means large file sizes and slow load times.
- For Daily Mail, we used the LOC guidelines of 400dpi, a balance between readability and file size.
- For more intricate documents like mediaeval or scientific manuscripts, we may scan as high at 1,200dpi in order to permit close viewing of marginalia, ink and script details, illustrations, etc.

GALE
A Cengage Company

Image options: focusing on the reading experience

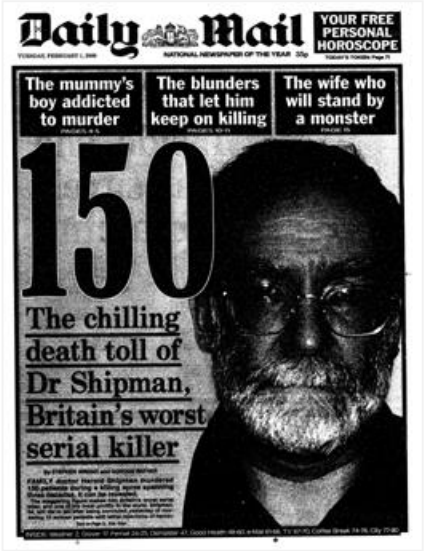
BITONAL

- Bitonal: B/W scan provides high contrast but can hinder images and illustrations
- Grayscale: permits subtlety of tone and shades but can pick up grain from paper
- Bitonal is regularly used for historic newspapers but the Daily Mail project forced us to reconsider, based on when photographs first appear.

GREYSCALE

Image options: focusing on the reading experience

BITONAL



- The Daily Mail becomes increasingly more illustrated throughout the 20th C.
- Bitonal is used up to 1962, when color and true halftone photographs began appearing more regularly.
- Users will notice a contrast in the way the images appear either side of this date.

GREYSCALE

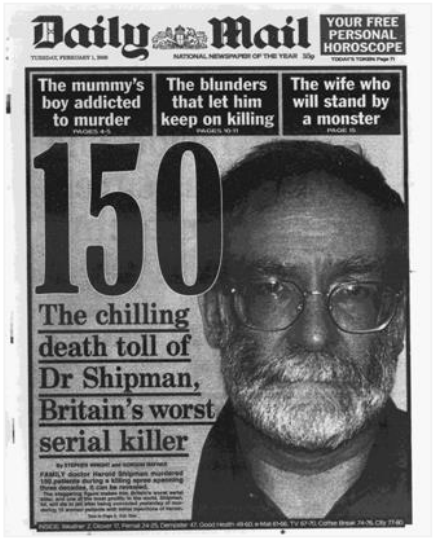


Image courtesy of www.explainthatstuff.com

Daily Mail

THURSDAY, DECEMBER 31, 1914 LONDON MANCHESTER PARIS NO 5,848 ONE HALFPENNY

CHRISTMAS TRUCE AT FRONT

STORY TOLD IN LETTERS FROM THE TRENCHES TO DAILY MAIL

SNOWBALLS AND JOKES WITH THE FOE

FRIENDLY "CHIVWAG" WITH THE GERMANS"

CAN ANY OF YOU CHAPS MEND A GRAMOPHONE?

TOMMY'S MARVELLOUS CHRISTMAS CONCERTS

Bill-posting gun: Inventors' ideas in 1914

Give your appetite a "silver lining"

'Silver Shred' Lemon Marmalade

will shift that "can't eat" feeling.

Made by ROBERTSON— of "Golden Shred" fame

Optical Character Recognition

- The quality of the OCR text usually says more about the condition of the original materials than it does about the performance of the OCR software.
- Typically, older newspapers produce much less satisfactory results than modern ones.
- Wartime newspapers noticeably produce poorer results than adjacent periods of history; especially 'bleed-through' of text from the other side of the page.

GALE
A Cengage Company

How accurate is *your* OCR?

WHAT DOES "ACCURACY" REALLY MEAN?

- The word 'accuracy' is misleading; OCR software works from a *confidence* rating, and NOT true accuracy
- The software calculates a confidence level from 0-9 for each character it detects, but does not know whether a character has been converted correctly or not.
- The *Daily Mail Historical Archive* had a team of over 400 people creating and reviewing the data for the archive, but with over 1.2 million pages to digitise and convert, it was not physically possible to clean up the OCR for every article.
- Typically, only small-scale or crowd-sourced digitisation projects have a realistic opportunity of producing 100% perfect OCR.

METADATA

High quality metadata makes it much easier to find the specific types of information a researcher is looking for, and to place effective parameters and filters around a search query, such as date ranges limited to specific article types.

Daily Mail 28p

FEMAIL Magazine SEE PAGES 23-34

5AM: BUSH STOPS WAR

Iraqis 'accept all the UN terms for peace'

From GEORGE GORDON in Washington

PRESIDENT Bush declared an end to fighting in the Gulf War early today.

He announced a ceasefire to begin at midnight (5am British time).

And within hours a number of reports from the United Nations said that Iraq's foreign minister Tariq Aziz had sent a letter to the UN — prepared before President Bush's announcement — which stated: 'The government of Iraq agrees to fully comply with the 12 UN resolutions.'

In a dramatic televised speech, six weeks to the minute since his declaration of war, Mr Bush said Kuwait is liberated, Iraq's army is defeated.

The 'suspension of all offensive combat operations' was contingent on Iraq's compliance with four points, he said.

It must immediately release all captives POWs and the remains of war victims, withdraw its troops and must inform Kuwaiti authorities of the location and nature of all land and sea mines. And it must comply fully with all relevant UN resolutions, including rescinding its claim on Kuwait and acceptance of responsibility to pay war damages. It is up to Iraq whether this becomes a permanent ceasefire, he said.

Violated

The decision was made in close consultation with Downing Street after reports of massive destruction of Saddam's war machine. Prime Minister John Major will make a full Commons statement today.

The announcement came only four days, exactly 100 hours to Mr Bush and it, after the start of the largest land offensive since World War II, which quickly turned into a spectacular rout of disillusioned Iraqi troops.

Pentagon officials said last night the fighting was almost at an end and Iraq's Republican Guard was in its death throes after a three battle lasting all day.

Mr Bush said if Iraq violated his conditions by firing on any coalition forces or launching Scud missile attacks, military operations would be resumed.

Your military objectives are met. Kuwait is once more in the hands of Kuwaitis, who control their own destiny. The Kuwait flag flies above the capital of a free and

Turn To Page 2, Col. 1

VICTIMS OF A TERRIBLE BATTLE ERROR

Youngest victim ... Fusilier Conrad Cole, aged 17

Nine Desert Rats killed by American 'friendly fire'

NINE young Desert Rats have been killed by American 'friendly fire', it was learned yesterday.

An A-10 aircraft's mission and cloth hit the troops in their two Warrior fighting cars. And last night Pentagon chiefs were questioning the role of the tank-buster plane, which is said to give its pilot poor visibility.

Six of those who died in Tuesday night's accident were from the 1st Battalion, included one of the Royal guard soldiers fighting in the Desert Corps' first. It also joined the Royal Regiment of Fusiliers, which suffered its first casualties in the war.

Daily Mail Reporter

Another victim, Fusilier Kevin Lewis, 20, was also one his youngest friends. Mark went to the

They were with the 3rd Battalion the Royal Regiment of Fusiliers as part of

Strafed

The brigade was fighting its way through an Iraqi armoured division guarding the western flank of the Kuwait border when it was attacked by A-10. Mistaken for Iraqi armoured vehicles, they were strafed by

Turn To Page 4, Col. 2

Who?
(George Gordon)

What?
(Article Title: 5AM: Bush Stops War)
(Article Type: News)
(679 words)

When?
(28 February 1991)

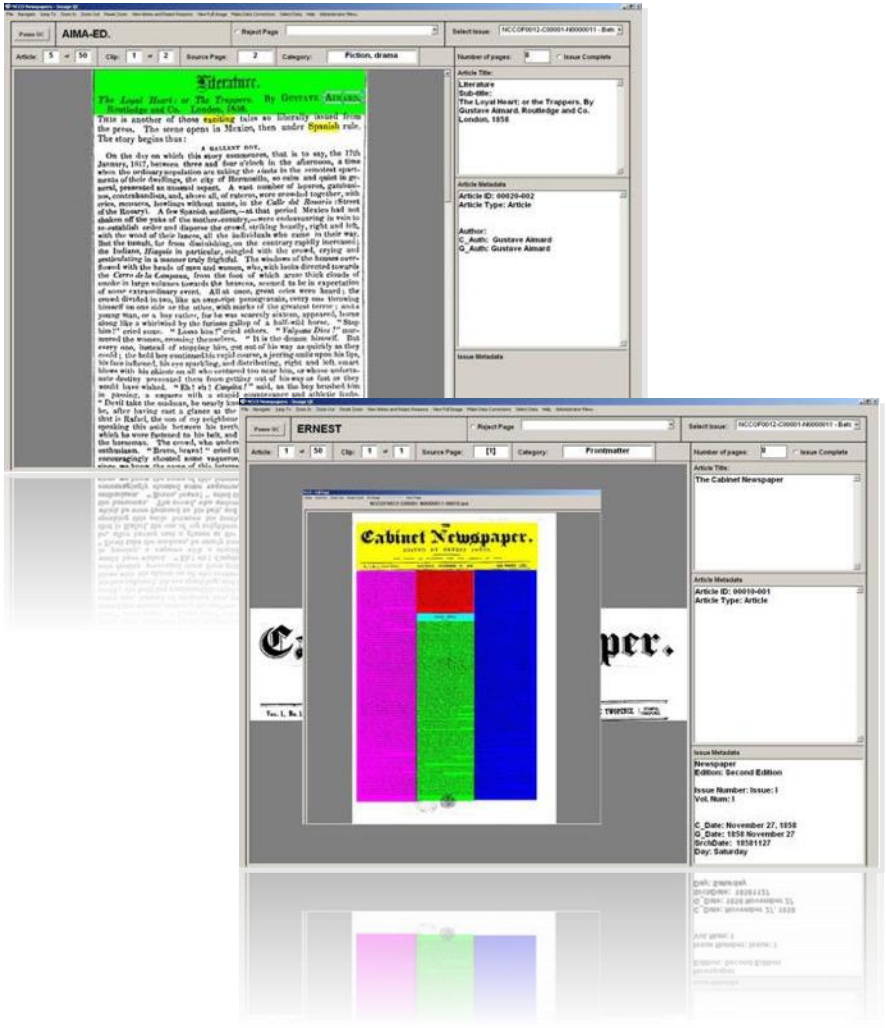
Where?
(Daily Mail, London, England)
(Page 1)
(Issue 29, 453)

GALE
A Cengage Company

- Creating XML is by far the costliest and most labour-intensive part of any digitisation project.
- Word coordinates are essential to hit-term highlighting and zoning of article copy.
- DTD defines the data structure for the archive, outlining a list of permissible legal elements and attributes.
- All data that is captured for a project must fit the DTD, or it does not pass verification – there are no exceptions.

XML

```
<wd pos="547,1273,622,1298">From</wd>
<wd pos="638,1271,784,1297">WALTER</wd>
<wd pos="799,1272,890,1299">FARR</wd>
<wd pos="906,1273,989,1303">(Daily</wd>
<wd pos="1004,1273,1065,1299">Mail</wd>
<wd pos="1080,1272,1182,1303">Special</wd>
<wd pos="1199,1273,1412,1304">Correspondent)</wd>
<wd pos="1446,1286,1649,1318">Washington,</wd>
<wd pos="1668,1288,1766,1319">Friday.</wd>
<wd pos="574,1325,802,1367">BRITAIN,</wd>
<wd pos="831,1328,894,1362">the</wd>
<wd pos="925,1326,1058,1362">United</wd>
<wd pos="1087,1328,1214,1369">States,</wd>
<wd pos="1239,1328,1366,1372">China,</wd>
<wd pos="1394,1331,1529,1373">Russia,</wd>
<wd pos="1553,1333,1624,1368">and</wd>
```

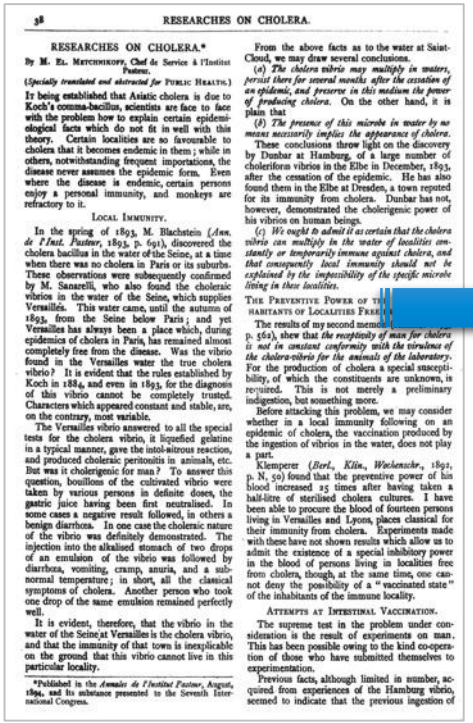


Quality Assurance

- Article zoning and segmentation defines the borders of each article, page property, section property, etc.
- “Clip” view allows users to see article segmentation through highlighting
- Manual QA is used to ensure segmentation is accurate and that article split points are correct.

GALE
A Cengage Company

TEXT DATA MINING – driving new insights from traditional content



38 RESEARCHES ON CHOLERA.

RESEARCHES ON CHOLERA.*

By M. El. Metchnikoff, Chef de Service à l'Institut Pasteur.

(Specially translated and abstracted for Public Health.)

It being established that Asiatic cholera is due to Koch's comma-bacillus, scientists are face to face with the problem how to explain certain epidemiological facts which do not fit in well with this theory. Certain localities are so favourable to cholera that it becomes endemic in them; while in others, notwithstanding frequent importations, the disease never assumes the epidemic form. Even where the disease is endemic, certain persons enjoy a personal immunity, and monkeys are refractory to it.

Local Immunity.

In the spring of 1893, M. Blachstein (*Ann. de l'Inst. Pasteur*, 1893, p. 691), discovered the cholera bacillus in the water of the Seine, at a time when there was no cholera in Paris or its suburbs. These observations were subsequently confirmed by M. Sanarelli, who also found the choleraic vibrios in the water of the Seine, which supplies Versailles. This water came, until the autumn of 1893, from the Seine below Paris; and yet Versailles has always been a place which, during epidemics of cholera in Paris, has remained almost completely free from the disease. Was the vibrio found in the Versailles water the true cholera vibrio? It is evident that the rules established by Koch in 1884, and even in 1893, for the diagnosis of this vibrio cannot be completely trusted. Characters which appeared constant and stable, are, on the contrary, most variable.

The Versailles vibrio answered to all the special tests for the cholera vibrio, it liquefied gelatine in a typical manner, gave the intol-nitrous reaction, and produced choleraic peritonitis in animals, etc. But was it cholerigenic for man? To answer this question, bouillons of the cultivated vibrio were taken by various persons in definite doses, the gastric juice having been first neutralised. In some cases a negative result followed, in others a benign diarrhœa. In one case the choleraic nature of the vibrio was definitely demonstrated. The injection into the alkalised stomach of two drops of an emulsion of the vibrio was followed by diarrhœa, vomiting, cramp, anuria, and a subnormal temperature; in short, all the classical symptoms of cholera. Another person who took one drop of the same emulsion remained perfectly well.

It is evident, therefore, that the vibrio in the water of the Seine at Versailles is the cholera vibrio, and that the immunity of that town is inexplicable on the ground that this vibrio cannot live in this particular locality.

*Published in the *Annales de l'Institut Pasteur*, August, 1894, and its substance presented to the Seventh International Congress.

From the above facts as to the water at Saint-Cloud, we may draw several conclusions.

(*a*) *The cholera vibrio may multiply in waters, persist there for several months after the cessation of an epidemic, and preserve in this medium the power of producing cholera.* On the other hand, it is plain that

(*b*) *The presence of this microbe in water by no means necessarily implies the appearance of cholera.*

These conclusions throw light on the discovery by Dunbar at Hamburg, of a large number of choleriform vibrios in the Elbe in December, 1893, after the cessation of the epidemic. He has also found them in the Elbe at Dresden, a town reputed for its immunity from cholera. Dunbar has not, however, demonstrated the cholerigenic power of his vibrios on human beings.

(*c*) *We ought to admit it as certain that the cholera vibrio can multiply in the water of localities constantly or temporarily immune against cholera, and that consequently local immunity should not be explained by the impossibility of the specific microbe living in these localities.*

The Preventive Power of the Inhabitants of Localities Free [illegible]

The results of my second memoir [illegible] p. 562), shew that *the receptivity of man for cholera is not in constant conformity with the virulence of the cholera-vibrio for the animals of the laboratory.* For the production of cholera a special susceptibility, of which the constituents are unknown, is required. This is not merely a preliminary indigestion, but something more.

Before attacking this problem, we may consider whether in a local immunity following on an epidemic of cholera, the vaccination produced by the ingestion of vibrios in the water, does not play a part.

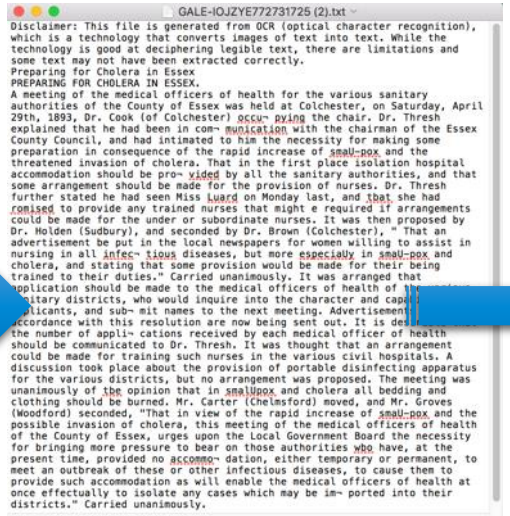
Klemperer (*Berl., Klin., Wochenschr.*, 1892, p. N, 50) found that the preventive power of his blood increased 25 times after having taken a half-litre of sterilised cholera cultures. I have been able to procure the blood of fourteen persons living in Versailles and Lyons, places classical for their immunity from cholera. Experiments made with these have not shown results which allow us to admit the existence of a special inhibitory power in the blood of persons living in localities free from cholera, though, at the same time, one cannot deny the possibility of a "vaccinated state" of the inhabitants of the immune locality.

Attempts at Intestinal Vaccination.

The supreme test in the problem under consideration is the result of experiments on man. This has been possible owing to the kind co-operation of those who have submitted themselves to experimentation.

Previous facts, although limited in number, acquired from experiences of the Hamburg vibrio, seemed to indicate that the previous ingestion of

Metchnikoff, M. El. "Researches on Cholera." Public Health, The Journal of the Society of Medical Officers of Health, vol. [VII], no. [78], 1894, p. 38+. Nineteenth Century Collections Online,



GALE-IOJZYE772731725 (2).txt

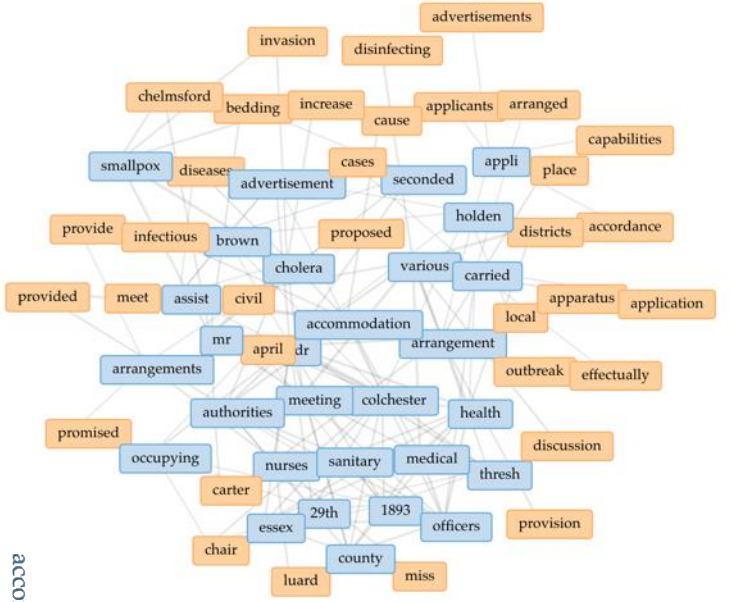
Disclaimer: This file is generated from OCR (optical character recognition), which is a technology that converts images of text into text. While the technology is good at deciphering legible text, there are limitations and some text may not have been extracted correctly.

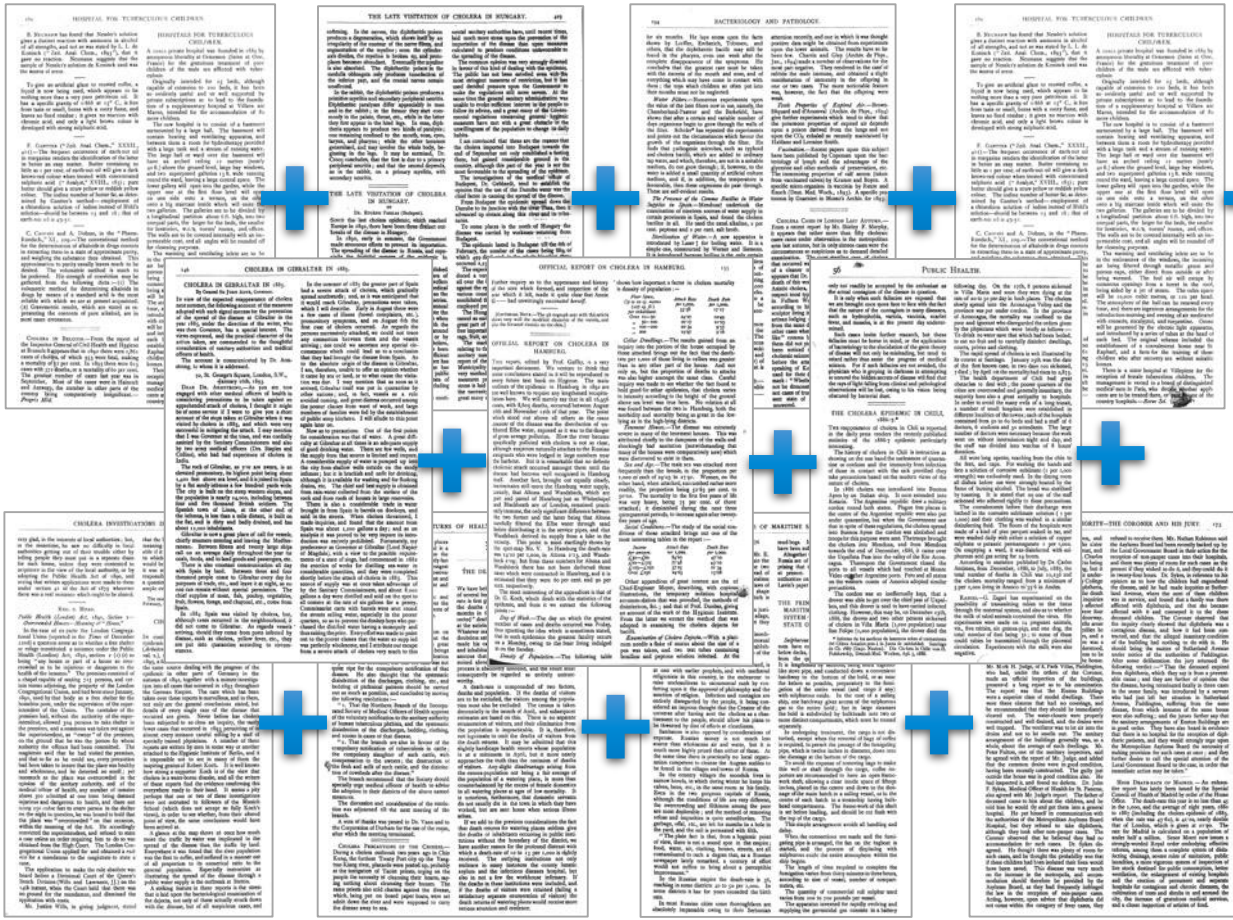
Preparing for Cholera in Essex

PREPARING FOR CHOLERA IN ESSEX.

A meeting of the medical officers of health for the various sanitary authorities of the County of Essex was held at Colchester, on Saturday, April 29th, 1893, Dr. Cook (of Colchester) occu- pying the chair. Dr. Thresh explained that he had been in com- munication with the chairman of the Essex County Council, and had intimated to him the necessity for making some preparation in consequence of the rapid increase of smaU-pox and the threatened invasion of cholera. That in the first place isolation hospital accommodation should be pro- vided by all the sanitary authorities, and that some arrangement should be made for the provision of nurses. Dr. Thresh further stated he had seen Miss Luard on Monday last, and that she had romised to provide any trained nurses that might e required if arrangements could be made for the under or subordinate nurses. It was then proposed by Dr. Holden (Sudbury), and seconded by Dr. Brown (Colchester), " That an advertisement be put in the local newspapers for women willing to assist in nursing in all infec- tious diseases, but more especiaUy in smaU-pox and cholera, and stating that some provision would be made for their being trained to their duties." Carried unanimously. It was arranged that application should be made to the medical officers of health of [illegible] sanitary districts, who would inquire into the character and capa[illegible] applicants, and sub- mit names to the next meeting. Advertisements [illegible] accordance with this resolution are now being sent out. It is des[illegible] the number of appli- cations received by each medical officer of health should be communicated to Dr. Thresh. It was thought that an arrangement could be made for training such nurses in the various civil hospitals. A discussion took place about the provision of portable disinfecting apparatus for the various districts, but no arrangement was proposed. The meeting was unanimously of the opinion that in smaUpox and cholera all bedding and clothing should be burned. Mr. Carter (Chelmsford) moved, and Mr. Groves (Woodford) seconded, "That in view of the rapid increase of smaU-pox and the possible invasion of cholera, this meeting of the medical officers of health of the County of Essex, urges upon the Local Government Board the necessity for bringing more pressure to bear on those authorities who have, at the present time, provided no accommo- dation, either temporary or permanent, to meet an outbreak of these or other infectious diseases, to cause them to provide such accommodation as will enable the medical officers of health at once effectually to isolate any cases which may be im- ported into their districts." Carried unanimously.

- Users can search for and find individual documents, create a bibliography, save, send, etc.
- OCR text of each document can be analyzed with open source tools like Voyant, etc.

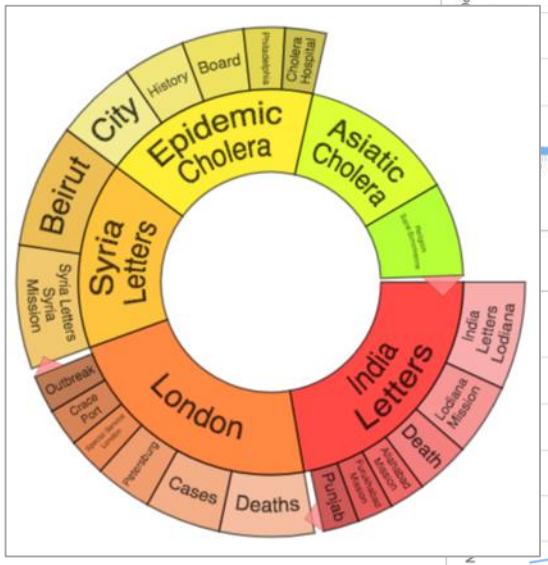


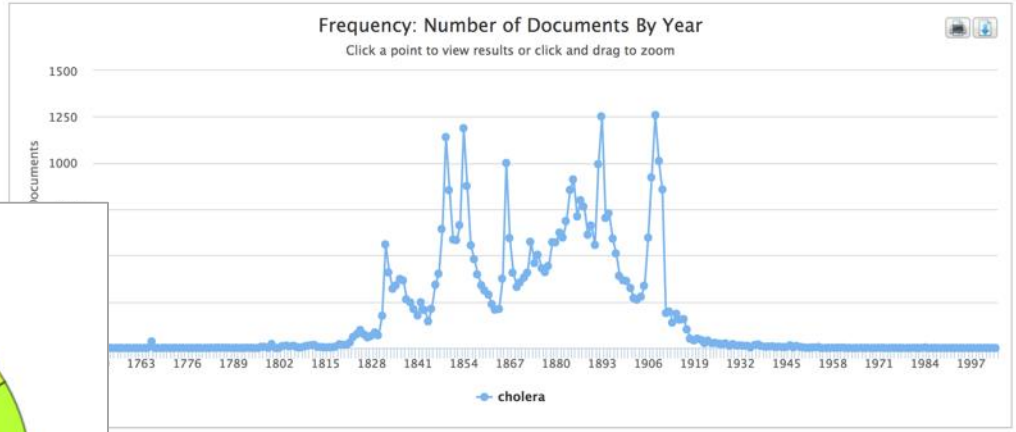


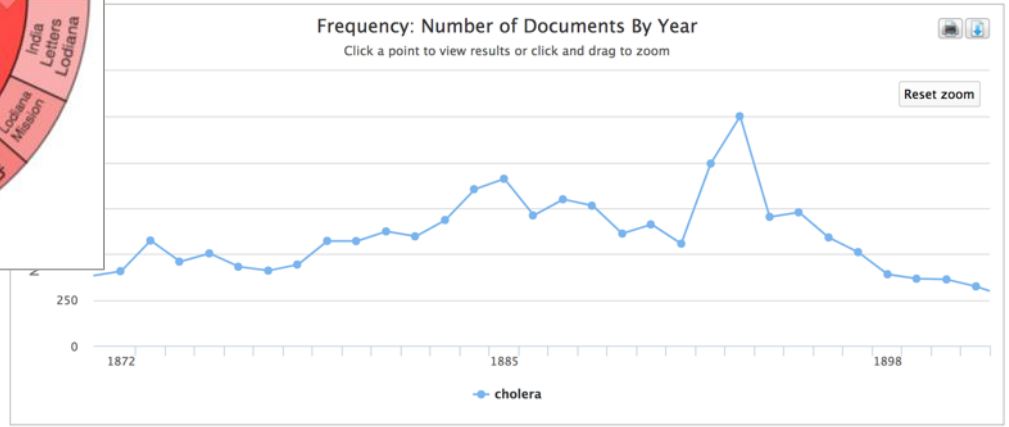
- Italy
- Budapest
- Russia
- London
- Chile
- Cagliari
- Gibraltar
- Japan
- Madrid
- Hungary
- Switzerland
- Belgium

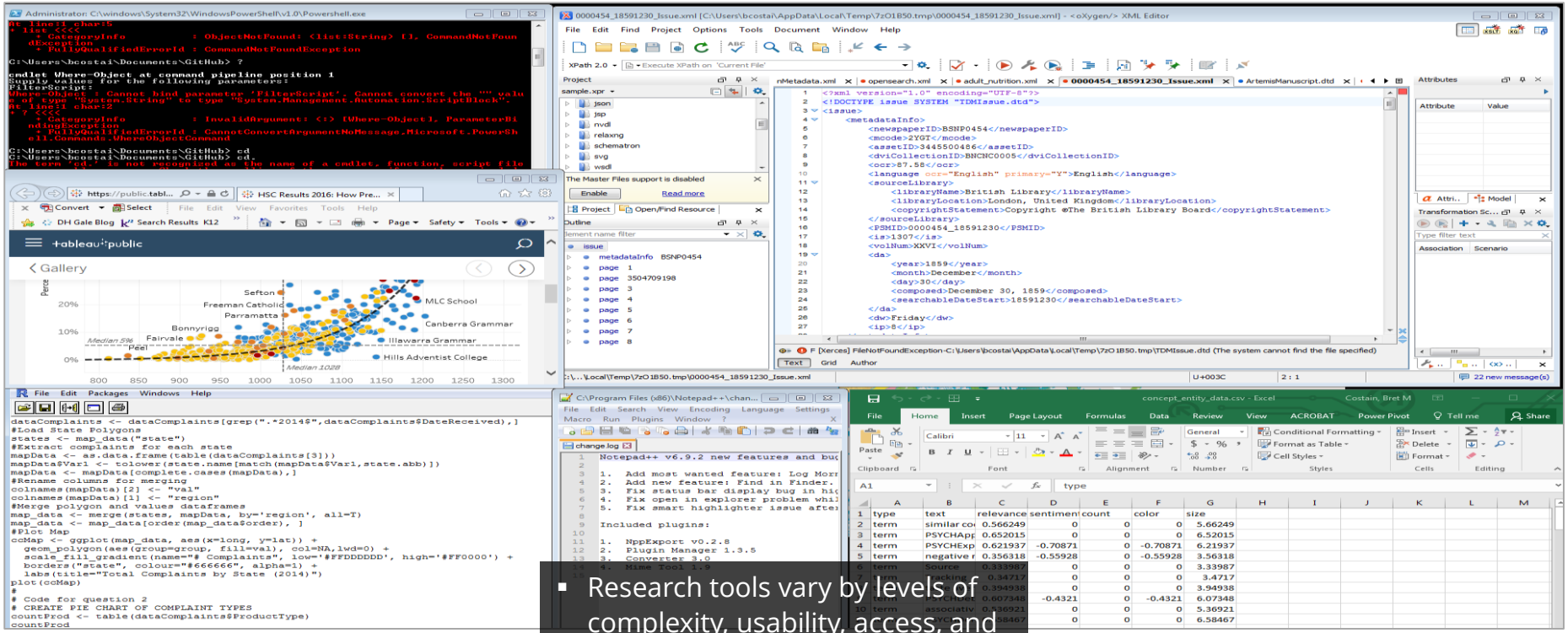
Articles

- Public Health, the Journal of t…
- Subscribers' Copies of Public…
- Notices (p.1)
- The Incorporated Society of …
- Annual Dinner (p.1)
- The Organisation of Meat Ins…
- The Incorporated Society of …
- International Congress of Hy…
- Bacteriological Investigations…
- The Statistics of Diphtheria in…
- On the Present State of Kno…
- On the Serum Therapeutics o…
- The Bearing of Experimental …
- Purification by Fire (p.20)
- Report on the Epidemic of En…
- Preservation of Fresh Milk (p.…
- The Corrected Death Rates o…
- Diagnosis of Influenza (p.32)
- Public Health (p.32)
- The Budapesth Congress (p.…
- The Corrected Death Rate of …
- Epidemic of Enteric Fever, Tr…
- On an Outbreak of Anomalou…
- A Case of Hyperpyrexia (p.38)



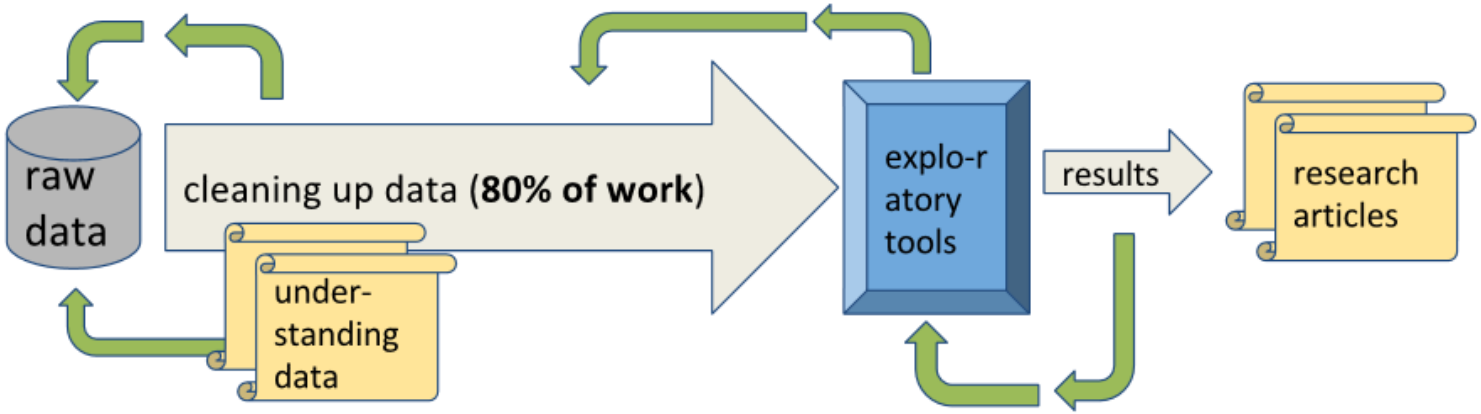






- Research tools vary by levels of complexity, usability, access, and support.
- Clean data ensures that each of these algorithms are starting with the least number of errors.

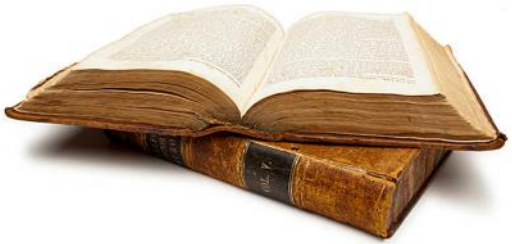
Text Data Mining (TDM) format: start with clean content

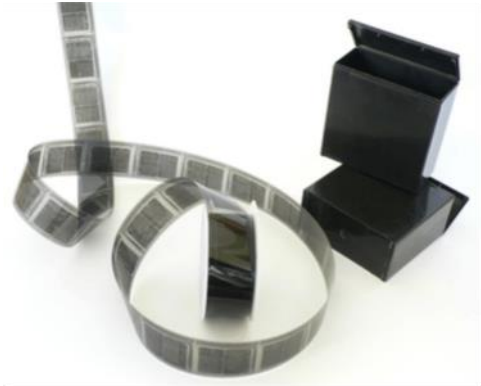


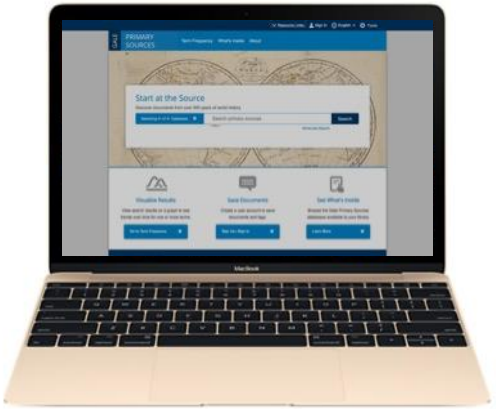
80% of your time for data cleanup, another 80% for algorithms, …

Text and data mining Eighteenth Century based on ESTC & ECCO
COMHIS Collective
BSECS Conference 2017, Oxford
http://j.mp/comhis-bsecs

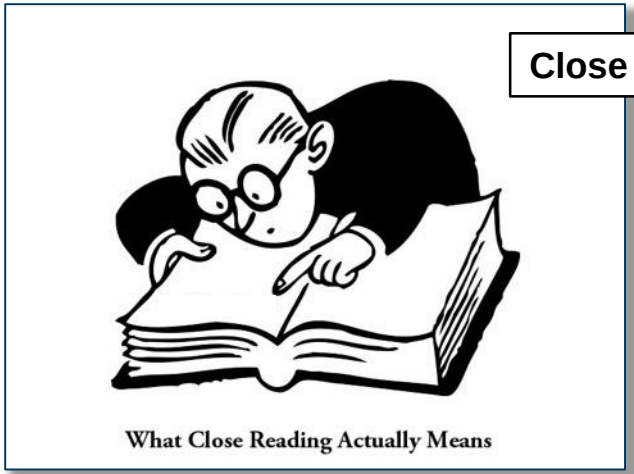
Digital Humanities ask different things from a text



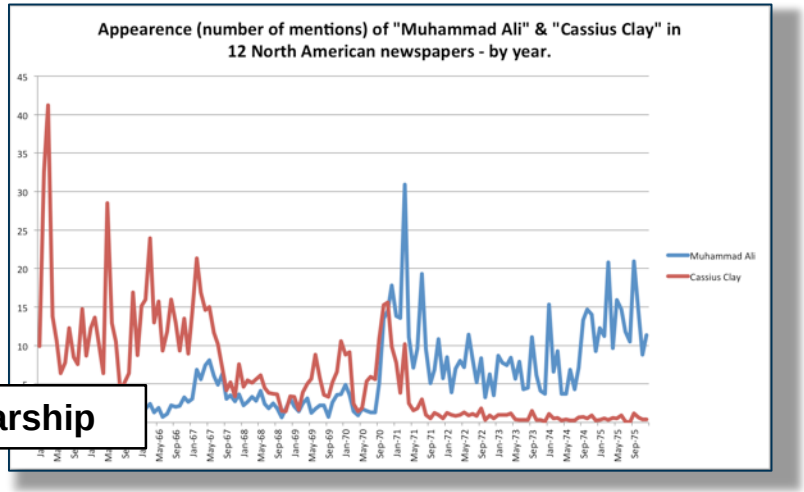




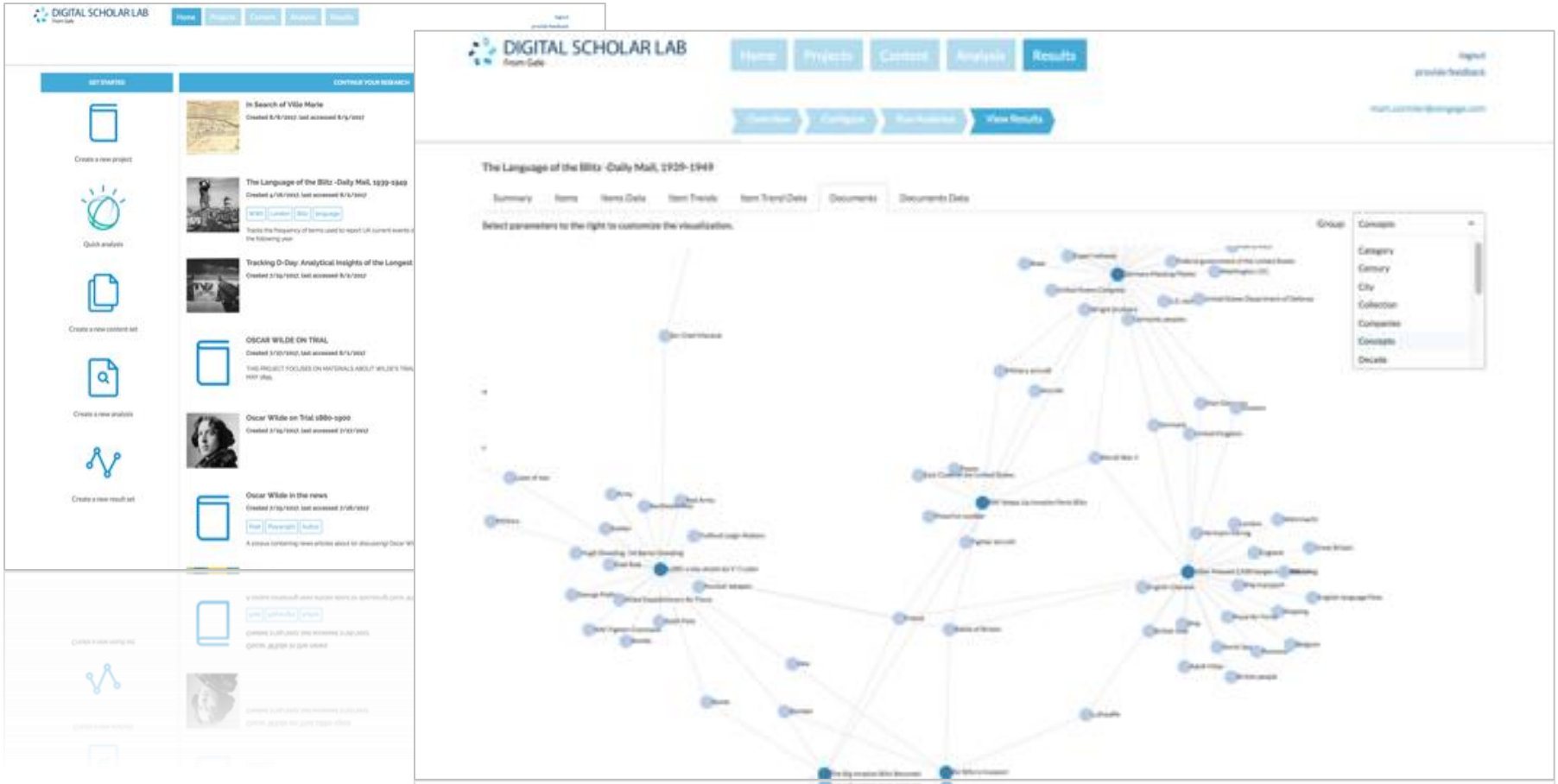
Close Reading vs. Distant Reading



Close Reading – Traditional Scholarship



Distant Reading – Digital Scholarship


DIGITAL SCHOLAR LAB
From Gale
Home
Projects
Content
Analysis
Results
Overview
Compare
Run Analytical
View Results
The Language of the Blitz -Daily Mail, 1939-1949
Summary
Items
Items Data
Item Trends
Item Trend Data
Documents
Documents Data
Select parameters to the right to customize the visualization.
Group
Concepts
Category
Century
City
Collection
Companies
Concepts
Decade