Create new possibilities for **ALL** Humanities Researchers with the Gale Digital Scholar Lab

Chris Houghton  
Head of Digital Scholarship, International  
chris.houghton@cengage.com  
@DHandDSatGale
Serendipity....

“making happy discoveries whilst looking for other things”
Collaborations and Partnerships: addressing the big digital challenges together
Good Data since 2002
Think about what existed before

- Limited number of users
- Difficult to use and easily damaged
- Slow to find things
Digital Archives solved the 3 challenges of microfilm

- Multiple Simultaneous Users
- Easy to Use, and a single user cannot destroy it!
- Find relevant material quickly
250 Million pages of content later...
Jump Ahead to 2010...
Requests to Access and Improve our Data

Text Creation Partnership (TCP)

- Manually keyed and text-encoded 2,231 of ECCO’s 150k texts
- Allows them to be used for purposes beyond the scope of the ECCO platform, including text mining
- TCP also worked with ProQuest’s EEBO
Requests for access beyond archives

Dr Kat Gupta
University of Nottingham
Unexpected Uses of Data

Prof Dallas Liddle
Augsburg College
A shift in workflows

Search text and Retrieve images

Gather data and Analyse content sets
In 2013 and 2014, we responded in 3 ways

Textual Analysis Tools include Term Clusters, Term Frequency Graphs, and More.

LEARN MORE ABOUT GALE ARTEMIS: PRIMARY SOURCES »
What we learned in 5 years...
What is Digital Humanities?

Digital Humanities is the critical study of how digital technologies and methods intersect with humanities scholarship and scholarly communication. It investigates the use of digital tools and software for interpretation and analysis of humanities research questions and how digital methodologies can be used to enhance disciplines such as Art History, Classical Studies, History, Literature, Music and many others.

Digital Humanities allows scholars to approach old problems with new means, or to ask new questions that could not have been asked with the traditional means of humanistic enquiry. Whatever the approach chosen, Digital Humanities remains grounded in humanities research and interests.

http://www.open.ac.uk/arts/research/digital-humanities/
Resource and Technical Support Limits in DH has resulted in a manifestation of the ‘1-9-90’ rule:

The 90-9-1 rule for participation in an online community
http://www.nngroup.com/articles/participation-inequality/
Challenge #1: Access to **relevant** data in an optimised format

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

Challenge #2: Hosting all of that data
Challenge #3: Tools are difficult to use
Our Solution....
Gale Digital Scholar Lab

TDM Research Environment

• Access to a broad range of texts from Gale Primary Source collections
• Access to powerful text mining tools
• Construct custom content sets across Gale’s collections
• Organise and manage research
• Integrated help and instructional materials
• Export OCR, statistical data and visualisations in standard formats
Developed with DH Scholars and experts
Digital Scholar Lab solves the 3 Challenges

- Access to relevant data in an optimised format
- Somewhere to host that data
- Familiar, Powerful tools
A Story of Exploration...
Pat Houghton
Shortening the "80%" of research time

<table>
<thead>
<tr>
<th>Name</th>
<th>Date modified</th>
<th>Type</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>OFFO-1785-APR01_issue.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>7 KB</td>
</tr>
<tr>
<td>OFFO-1785-APR01_Text.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>98 KB</td>
</tr>
<tr>
<td>OFFO-1785-APR02_issue.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>8 KB</td>
</tr>
<tr>
<td>OFFO-1785-APR02_Text.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>94 KB</td>
</tr>
<tr>
<td>OFFO-1785-APR04_issue.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>6 KB</td>
</tr>
<tr>
<td>OFFO-1785-APR04_Text.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>96 KB</td>
</tr>
<tr>
<td>OFFO-1785-APR05_issue.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>7 KB</td>
</tr>
<tr>
<td>OFFO-1785-APR05_Text.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>98 KB</td>
</tr>
<tr>
<td>OFFO-1785-APR06_issue.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>7 KB</td>
</tr>
<tr>
<td>OFFO-1785-APR06_Text.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>99 KB</td>
</tr>
<tr>
<td>OFFO-1785-APR07_issue.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>7 KB</td>
</tr>
<tr>
<td>OFFO-1785-APR07_Text.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>94 KB</td>
</tr>
<tr>
<td>OFFO-1785-APR08_issue.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>6 KB</td>
</tr>
<tr>
<td>OFFO-1785-APR08_Text.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>94 KB</td>
</tr>
<tr>
<td>OFFO-1785-APR09_issue.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>8 KB</td>
</tr>
<tr>
<td>OFFO-1785-APR09_Text.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>96 KB</td>
</tr>
<tr>
<td>OFFO-1785-APR11_issue.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>8 KB</td>
</tr>
<tr>
<td>OFFO-1785-APR11_Text.xml</td>
<td>23/12/2016 21:09</td>
<td>XML Document</td>
<td>96 KB</td>
</tr>
</tbody>
</table>
OCR Confidence ≠ OCR Accuracy
Exposing the OCR process, flaws and all

WHICH CAME FIRST: the business or the trip? A fortnight ago the gentleman on the right was happily combining a visit to the overseas office of a firm of solicitors in Australia with a tour of the P & O First Class Service to Australia. For this is the P & O First Class Service, the P & O fleet of the British businessman is not to be missed. Out of some six hundred and fifty thousand miles of track, a percentage of those will be travelled by men from the firm. The P & O service is a feature. You have the choice of rest or proper. You do more constructive work at home. Yet you arrive back refreshed and ready for a full working schedule. The P & O First Class Service to Australia is an investment.
Create Bespoke Content Sets

Adverts with 'Australia'. Telegraph 1950-1960

- ARCHIVES USED (1)
  - The Telegraph Historical Archive (3293)

- DOCUMENT TYPE (1)
  - Advertisement (3293)

- SOURCE LIBRARIES (2)
  - Telegraph Media Group (2464)
  - The Telegraph Media Group (828)

- TIMEFRAME
  - 1950 - 1960
Analysis Tools available in the Gale Lab

- Topic Modelling (Mallet)*
- Clustering (SciKit Learn)*
- Parts-of-Speech Tagger (spaCy)*
- Sentiment Analysis (OpenNLP)*
- Named Entity Recognition (spaCy)*
- Ngrams (Lucene)

*Denotes open source
What does this mean for Uncle Pat?

Daily Telegraph

Daily Mail
Topic Modelling for real Exploration
Gale’s DS Lab: for the ‘1%’, ‘9%’ and ‘90%’

- Facilitates the creation and use of data sets
- Can be used in teaching to analyse data collectively
- Allows everyone to build up data analysis and digital humanities skills