Digital collections
New challenges for curation
New opportunities for data driven scholarship

Adam Farquhar, PhD
Head of Digital Scholarship
The British Library
15-Jan-2013
Growing content
Digitisation
The British Newspaper Archive

- Short-listed for 2012 Big-Data project of the year
- A partnership between the British Library and brightsolid
- Digitising up to 40 million pages over 10 years
- Collection includes runs of most newspapers published in the UK since 1800
- Over 7m pages added since launch
Google Books

- 250,000 Books, 1700-1870
- From the French Revolution to the end of slavery in the US
- Material in major European languages
- Focus on books that are not yet freely available in digital form online
- Access via Google Books and BL
- Storage at Google and BL
- Contract and terms available on the web!
Examining under the microscope
Improving access by converting to text

Broadcast News

- **Broadcast News**
  - Television and radio news programmes receivable in the UK, recorded by the British Library since May 2010
  - Currently record 37 hours per day from 15 TV channels and 2 radio channels such as Al Jazeera English, CNN, France 24, Russia Today
  - Innovative search across subtitles (where available)
  - Launch in reading rooms 2012

- **Opening up speech archives**
  - AHRC-funded project looking at speech-to-text technologies for opening up audio and video archives
  - Project will index 3,000 hours of TV news and 3,000 hours of radio content

IMPACT Historic Text

- **Improve the digital accessibility of printed text produced before 1900**
  - State-of-the-art OCR does not produce satisfactory results for old books, magazines and newspapers
  - Commercial OCR focuses on modern documents
  - Historic material have archaic fonts, complex layouts, warped or degraded pages
  - Manual post-correction is slow and expensive
Enhancing content
Viewing from a distance
Engaging with the community

IDP Wikipedia Editathon

AHRC Digital Transformations
Building our capacity

- Foundations in working with Digital Objects: From Images to A/V
- Behind the Screen: Basics of the Web
- What is Digital Scholarship?
- Digital Collections
- Digitisation
- Crowdsourcing in Libraries, Museums and Cultural Heritage Institutions
- Text Encoding Initiative & Annotation
- Data Visualisation for Analysis in Scholarly Research
- Geo-referencing and Digital Mapping
- Information Integration: Mash-ups, API’s and The Semantic Web
reading individual works is as irrelevant as describing the architecture of a building from a single brick, or the layout of a city from a single church – Franco Moretti, Stanford

Thank you!