Breathing life into digital collections at the British Library

By Mia Ridge

Biography

Dr Mia Ridge is the British Library's Digital Curator for Western Heritage Collections. As part of the Library's Digital Scholarship team, she enables innovative research based on digital collections, providing guidance and training on computational methods for historical collections. Current projects include crowdsourcing work with historical playbills, and experimenting with machine learning. Her PhD was titled 'Making digital history: The impact of digitality on public participation and scholarly practices in historical research'. Formerly Lead Web Developer at the Science Museum Group, her career began in Australia with roles at Melbourne Museum and Vicnet at the State Library of Victoria.

Introduction

How are research libraries preparing to meet the needs of 21st century researchers? For the past decade, the British Library's Digital Scholarship team has worked to ensure that the Library's collections, systems, policies and processes meet the emerging needs of anyone who wants to conduct innovative research with the Library's digital collections and data. This article firstly provides some context for the British Library's investment in this area, then discusses how the team seeks to understand and encourage the use of collections in digital scholarship and, finally, addresses some of the challenges this entails.

Working at scale — the collections of the British Library

The British Library is the national library of the United Kingdom. Its purpose is to make the intellectual heritage represented by its collections accessible to everyone, for 'research, inspiration and enjoyment'. This work is supported by six core purposes, some of which — including working internationally to advance knowledge and mutual understanding; supporting and stimulating research of all kinds; inspiring learners of all ages and engaging everyone with memorable cultural experiences — are directly linked to the Library's support for digital access to collections for research and learning.

> *The British Library holds an estimated 180–200 million items, including over 14 million books; 8 million stamps; 310,000 manuscript volumes; 4 million maps; 60 million patents, 260,000 journal titles; sound files; pamphlets, magazines, sheet music and newspapers; television and radio recordings; and archived websites. Over 3 million new items are added every year ... within a few years the library expects to ingest 5 terabytes of data a day.*

One of the largest libraries in the world, the British Library holds an estimated 180–200 million items, including over 14 million books; 8 million stamps; 310,000 manuscript volumes; 4 million maps; 60 million patents; 260,000 journal titles; sound files; pamphlets, magazines, sheet music and newspapers; television and radio recordings; and archived websites. Over 3 million new items are added every year, and as digital publishing increases in volume, within a few years the library expects to ingest 5 terabytes of data a day.

Digital Scholarship at the British Library

The Digital Scholarship team was set up in 2010 to enable innovative research with the Library's digital collections and data. Four digital curators (including the author) are embedded within specific Collections and Curation departments, and provide advice, support and training in the creation and use of relevant digital collections for their staff. For example, my current focus is exploring the applications of data science-based research methods to digitised historical collections, by investigating algorithm-based metadata generation to make collections more discoverable, and seeking to understand how disciplines such as computer vision [http://blogs.bl.uk/digital-scholarship/2018/05/seeing-british-library-collections-through-a-digital-lens.html] or computational linguistics approach Library collections. Other digital curators work on specific digitisation projects, building capacity for digital scholarship within potential user groups through workshops, pilots and documentation. We want to help researchers think beyond reading a page at a time or manually compiling a database of records they're interested in, to thinking about 'reading' thousands of pages from hundreds of sources or using text and data-mining techniques to scale up their research.

We collaborate closely with the Mellon-funded British Library Labs [https://www.bl.uk/projects/british-library-labs] team, the Endangered Archives Programme [http://eap.bl.uk/], IT projects (such as the new, standards-based item viewers [http://blogs.bl.uk/digital-scholarship/2016/12/new-viewer-digitised-collections-british-library.html]), and other researcher-focused teams. Collectively we aim to share knowledge, expertise and experience; to connect scholars with the resources they need; and to experiment with digital methods to address barriers to collections access for users. Below, I outline key activities that help us meet those goals.

Building internal capacity — training Library staff

A key activity for the team is devising and running training in digital scholarship for other Library staff. Begun in 2012, our Digital Scholarship Training Programme [https://www.bl.uk/projects/digital-scholarship-training-programme] is the result of an extensive consultation exercise and survey of the digital scholarship landscape to understand the foundational concepts, methods and tools with which staff would need to be familiar (McGregor *et al.* 2016). Courses provide a mixture of hands-on, practical exercises, and time to explore and discuss innovative digital projects and case studies. Providing training in subjects such as crowdsourcing and data visualisation for cultural heritage collections, copyright and using online data sources helps Library staff understand how other scholars might apply new technologies and methods to digital collections, enabling better research collaborations.

In the past year we have responded to the need for a more flexible training programme by breaking day-long workshops into modules delivered over 'seasons'. Over a season, staff can learn about topics such as 'text and data mining for cultural heritage collections' through a mixture of talks, practical workshops, tutorials, and guest lectures from visiting specialists. This format has several advantages: staff find it easier to attend hour-long modules, staff can try out methods on their own collections between sessions, the ability to pick and choose sessions means that attendees for each module are more engaged, and new topics can be introduced on a 'just in time' basis as the technology changes. The modular format also means we can invite international experts and collaborators to give talks on their specialisms with relatively low organisational overhead.

The team needs to keep apace of changes in the field, so we run a monthly reading group [http://blogs.bl.uk/digital-scholarship/2018/05/what-do-deep-learning-community-archives-livy-and-the-politics-of-artefacts-have-in-common.html] and hands-on 'hack and yak' sessions. Both are open to anyone in the Library interested in a topic, activity or tool featured in that session.

Collaborating with external researchers

Many of our external research collaborations are based on PhD studentships, devised with the Library's Research Collaboration team [https://www.bl.uk/research-collaboration], and funded through research councils, with academic partners recruited through an open call. They provide access to Library collections and expertise for PhD students, while we learn from their in-depth explorations of specific research questions or methods. Students can attend our Training Programme, and are invited to give staff talks or run workshops based on their research, further strengthening the Training Programme.

We also take part in the Library's programme for three-month PhD placements, which provide valuable experience for students while helping deliver useful outcomes. We have also supervised undergraduate and master's dissertation projects, working with printed heritage, manuscripts and archives colleagues to shape their research projects around specific collections.

Challenges for internal and external collaboration

Building digital collections and scholarship into traditional structures can be challenging. For example, if a staff member is inspired to try text mining after attending a training session, they must first navigate the various permissions needed to access digitised sources, install software on their work computer and find the time to experiment. For the Library, the scale of the collections means that tools that work at a local scale may not be suitable for larger or more complex collections. Turning *ad hoc* pilots or experiments into larger, integrated projects is a challenge. On a more positive note, this provides some insight into the challenges that external researchers face when incorporating digital scholarship methods into their work.

> *We want to help researchers think beyond reading a page at a time or manually compiling a database of records they're interested in, to thinking about 'reading' thousands of pages from hundreds of sources or using text and data-mining techniques to scale up their research.*

Assuming they can find the right skills or collaborators to get started, academics may face challenges finding suitable outlets for publishing work based on digital scholarship. If they publish in traditional disciplinary journals, they may have to minimise computational aspects of their research, while journals in digital fields may only be looking for 'new' or 'innovative' work.

Opening access to data

Publishing well-documented digital and digitised collections online, under licences that encourage scholarly, creative and commercial reuse, is vital. The Library has published items on a range of platforms. Over 1 million digitised images from 19th century books [http://britishlibrary.typepad.co.uk/digital-scholarship/2013/12/a-million-first-steps.html] are freely available from Flickr Commons [https://www.flickr.com/photos/britishlibrary/], while the text is available via JISC's Historical Texts site [https://historicaltexts.jisc.ac.uk/home]. Library content — from maps to images of book bindings — also appears on Wikimedia Commons [https://commons.wikimedia.org/wiki/Category:Images_from_the_British_Library]. The Library's Metadata team has published a range of catalogues as datasets [http://www.bl.uk/bibliographic/datafree.html] in formats including linked open data (SPARQL, basic RDF/XML), 'Researcher Format' (CSV), MARC21 via Z39.50 and PDF. Some data from the UK Web Archive [https://www.webarchive.org.uk] is available for reuse. We also published linked open data descriptions of learning resources in collaboration with the BBC's Research and Education Space project [http://blogs.bl.uk/digital-scholarship/2017/05/how-can-a-turtle-and-the-bbc-connect-learners-with-literature.html].

Building on the work of BL Labs and digitisation colleagues in collecting files from legacy digitisation projects, the Library launched an open data portal [https://data.bl.uk] in 2016. We have found that publishing academic datasets built on British Library collections can give them a new lease of life, encouraging their use by early career scholars, and by established researchers looking for 'challenge datasets' they can test their tools with.

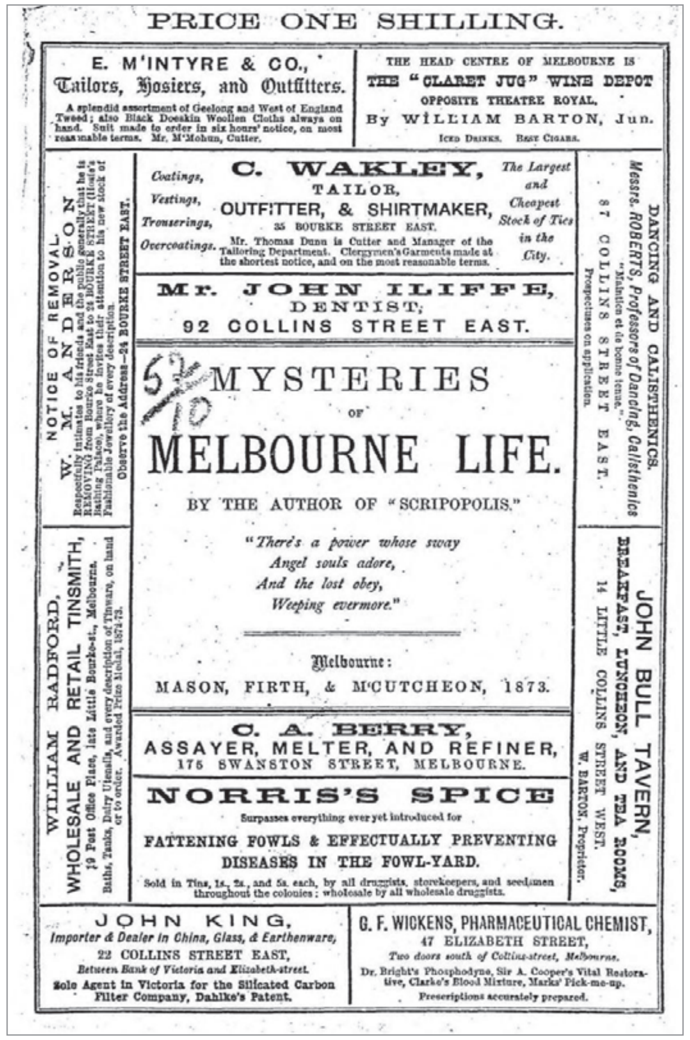
Challenges for publishing usable collections data

However, there are several reasons why collections and metadata published by the Library may be challenging for would-be digital scholars. Overall, the biggest challenge is the pace of cataloguing and digitisation in relation to the scale and variety of the collections. Our best estimates are that 1–4% of collections are digitised or born digital.

While ideally all digitised items should have detailed catalogue records and specialist metadata, and automatically transcribed text to enable full-text search and reuse, this is not always the case. Cataloguing and digitisation practices have varied over time and between projects, and the resulting variability in metadata quality increases the challenge in finding and using relevant collections in digital (or indeed, any) scholarship.

Entities recorded in metadata about historical collections, such as dates, names and places, may be uncertain, ambiguous, imprecise and generally 'messy' compared to modern data. This can cause problems for systems that expect modern, precise records about conventional books.

Researchers may initially have high expectations for digitised collections. The availability of accurately transcribed text is key for many digital methods, including text and data-mining techniques to extract the topics, people, places and other entities mentioned in the text, drawing network graphs of relationships between entities, or algorithmically compiling quantitative records for analysis. When published online, a single dataset of digitised texts can be used by multiple researchers. For example, the Library's 19th century newspaper collections have been studied to answer questions on the depictions of London in British newspapers [https://ihrdighist.blogs.sas.ac.uk/2015/12/14/tuesday-19-january-2016-tessa-hauswedell-european-or-imperial-metropolis-depictions-of-london-in-british-newspapers-1870-1900/], the locations of political meetings [https://ihrdighist.blogs.sas.ac.uk/2015/12/14/tuesday-2-february-katrina-navickas-political-meetings-mapper-with-british-library-labs-mapping-the-origins-of-british-democratic-movements-with-text-mining-nlp-geo-parsing-and-crowd-sourcing/], attitudes to immigrants and refugees [http://www.lancaster.ac.uk/people-profiles/ruth-byrne], and the temporal and spatial relationship to disease [http://blogs.bl.uk/digital-scholarship/2016/07/a-temporal-spatial-investigation-of-disease-in-19th-century-british-newspapers.html]. However, resources are rarely available for manually transcribing and marking-up records, and the quality of text transcribed with optical character recognition (OCR) tools can be poor, particularly for early digitisation projects. (Re-OCRing material can help, where resources allow.) The Library is exploring methods for handwritten text recognition [http://transkribus.eu/], which have the potential to transform access to manuscript and archive collections.



PRICE ONE SHILLING.

E. M'INTYRE & CO.,
Tailors, Hosiers, and Outfitters.
A splendid assortment of Geelong and West of England Tweed; also Black Doeskin Woollen Cloths always on hand. Suit made to order in six hours' notice, on most reasonable terms. Mr. M'Mohun, Cutter.

THE HEAD CENTRE OF MELBOURNE IS
THE "CLARET JUG" WINE DEPOT
OPPOSITE THEATRE ROYAL,
By WILLIAM BARTON, Jun.
Iced Drinks. Best Cigars.

NOTICE OF REMOVAL.
W. M. ANDERSON
Respectfully intimates to his friends and the public generally that he is REMOVING from Bourke Street East to 24 BOURKE STREET (Hosie's Bathing Palace), where he invites their attention to his new stock of Fashionable Jewellery of every description.
Observe the Address—24 BOURKE STREET EAST.

C. WAKLEY,
TAILOR,
OUTFITTER, & SHIRTMAKER,
35 BOURKE STREET EAST.
Coatings, Vestings, Trouserings, Overcoatings.
Mr. Thomas Dunn is Cutter and Manager of the Tailoring Department. Clergymen's Garments made at the shortest notice, and on the most reasonable terms.
The Largest and Cheapest Stock of Ties in the City.

DANCING AND CALISTHENICS.
Messrs. ROBERTS, Professors of Dancing, Calisthenics
"Maintien et de bonne tenue."
87 COLLINS STREET EAST.
Prospectuses on application.

Mr. JOHN ILIFFE,
DENTIST,
92 COLLINS STREET EAST.

MYSTERIES
OF
MELBOURNE LIFE.

BY THE AUTHOR OF "SCRIPOPOLIS."

"There's a power whose sway
Angel souls adore,
And the lost obey,
Weeping evermore."

Melbourne:
MASON, FIRTH, & M'CUTCHEON, 1873.

WILLIAM RADFORD,
WHOLESALE AND RETAIL TINSMITH,
19 Post Office Place, late Little Bourke-st., Melbourne.
Baths, Tanks, Dairy Utensils, and every description of Tinware, on hand or to order. Awarded Prize Medal, 1872-73.

JOHN BULL TAVERN,
BREAKFAST, LUNCHEON, AND TEA ROOMS,
14 LITTLE COLLINS STREET WEST.
W. BARTON, Proprietor.

C. A. BERRY,
ASSAYER, MELTER, AND REFINER,
175 SWANSTON STREET, MELBOURNE.

NORRIS'S SPICE
Surpasses everything ever yet introduced for
FATTENING FOWLS & EFFECTUALLY PREVENTING DISEASES IN THE FOWL-YARD.
Sold in Tins, 1s., 2s., and 5s. each, by all druggists, storekeepers, and seedsmen throughout the colonies; wholesale by all wholesale druggists.

JOHN KING,
Importer & Dealer in China, Glass, & Earthenware,
22 COLLINS STREET EAST,
Between Bank of Victoria and Elizabeth-street.
Sole Agent in Victoria for the Silicated Carbon Filter Company, Dahlke's Patent.

G. F. WICKENS, PHARMACEUTICAL CHEMIST,
47 ELIZABETH STREET,
Two doors south of Collins-street, Melbourne.
Dr. Bright's Phosphodyne, Sir A. Cooper's Vital Restorative, Clarke's Blood Mixture, Marks' Pick-me-up.
Prescriptions accurately prepared.

Figure 1: In an ideal world, this digitised page would be tagged with a linked data identifier to clarify whether 'Melbourne' refers to Victoria or Florida. Source: https://archive.org/details/MysteriesOfMelbourneLife

While ideally all digitised items should have detailed catalogue records and specialist metadata, and automatically transcribed text to enable full-text search and reuse, this is not always the case. Cataloguing and digitisation practices have varied over time and between projects, and the resulting variability in metadata quality increases the challenge in finding and using relevant collections in digital (or indeed, any) scholarship.

Applying content mining techniques to process items at scale has massive potential for digital scholarship and the discoverability of collection items. This, in turn, brings new challenges, including integrating tools for post-digitisation semantic enhancement into existing workflows, negotiating the role of curators and cataloguers in relation to these new tools, and finding software for processing non-Western materials and non-textual digital collections such as the UK Web and Sound Archives. Newer forms of digitisation, such as 3D modelling, which create complex digital objects, put further pressure on internal data systems but offer new possibilities for accessing objects, as explored by digital curator Dr Adi Keinan-Schoonbaert [http://britishlibrary.typepad.co.uk/asian-and-african/2016/05/cant-judge-a-book-by-its-cover-perhaps-you-can.html].

Publishing collections as datasets creates practical issues, too. When individual collection items are combined into datasets, their sheer size can create challenges for researchers. For example, one dataset available for download from the Library's data portal [https://data.bl.uk/] is over 400 GB in size. Smaller datasets may still be over 1 GB in size, making them difficult to download, uncompress, store, and computationally process for all but the most well-resourced researchers. Copyright and data protection laws can further limit immediate access to collections.

We also face more subtle issues. The Library's catalogues are traditionally based around the 'deliverable unit', the physical codex, bound volume or archive box that can be ordered to the reading room. However, emergent practices such as crowdsourced tagging and transcription, machine learning-led classification and content mining target single pages, or even regions of a page, and this has changed expectations about what a catalogue record represents. The mismatch in granularity between catalogues that describe the deliverable unit and technologies that describe the images and text on specific regions of manuscript, sheet or page must be resolved for us to take full advantage of newer technologies.

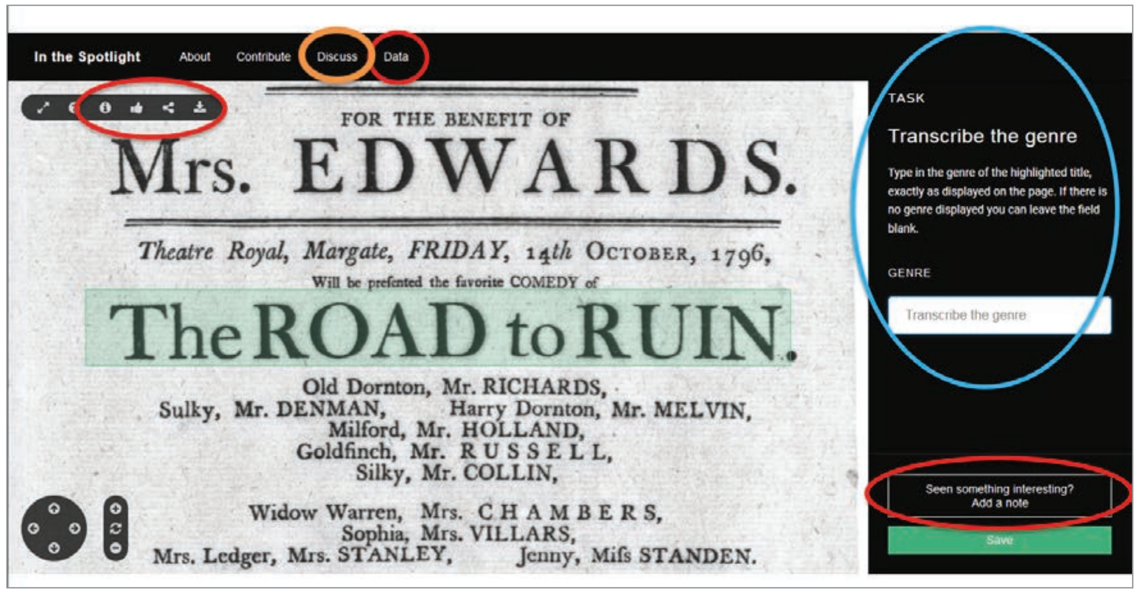
The role of outreach

Publishing data online and hoping that people will find it is not enough — an active programme of outreach activities is key for encouraging the use of digital collections. The BL Labs [https://www.bl.uk/projects/british-library-labs] team has taken digitised collections out to universities on 'roadshows'. These workshops are an opportunity to highlight innovative uses of digital collections and encourage academics to think creatively about including resources in their research and teaching [http://britishlibrary.typepad.co.uk/digital-scholarship/2016/05/success-story-the-bl_labs-roadshow-2016.html]. These events are also popular with university library staff curious to learn how we've faced some of the challenges, as well as academics who are considering digital scholarship projects.

Running competitions (or preparing material for use in other competitions) is an effective way to motivate the use of collections. From 2013 to 2016, the BL Labs team invited researchers, developers and artists to submit their important research question or creative idea leveraging the Library's digital content and data to their annual competition, and supported the winners in working on their idea. Digital Curator Stella Wisdom has also run *Off the Map* competitions [https://www.bl.uk/projects/off-the-map], a videogame design competition for UK students. Students use digitised British Library 'assets' including maps, views, texts, book illustrations and recorded sounds as creative inspiration. Efforts by colleagues including Nora McGregor to include historical Arabic manuscripts in technical competitions will help improve automatic text transcription for non-English items. In defining time-limited projects with clear expectations about what to submit and which rewards are possible, the competition format has encouraged creative uses of digital collections.

The annual British Library Labs Awards recognise outstanding work using the Library's digital collections and data in four categories: research, artistic, commercial and teaching/learning. The award format encourages people to nominate work with collections that would otherwise be difficult to track, and provides material for case studies.

Crowdsourcing tasks related to collections metadata is another form of outreach, engaging new audiences while making our collections more discoverable (Ridge 2013). In our most recent project, *In the Spotlight* [http://playbills.libcrowds.com/], was designed for both engagement and productivity. We added elements to the task interface to encourage participants to download images, view the full item on the main website, add their own tags to describe playbill sheets, comment on a sheet or discuss their findings on a forum. This approach appears to be working, as participants have shared interesting finds with us, and we recently celebrated our 100,000th contribution.



FOR THE BENEFIT OF
Mrs. EDWARDS.
Theatre Royal, Margate, FRIDAY, 14th OCTOBER, 1796,
Will be preſented the favorite COMEDY of
The ROAD to RUIN.
Old Dornton, Mr. RICHARDS,
Sulky, Mr. DENMAN, Harry Dornton, Mr. MELVIN,
Milford, Mr. HOLLAND,
Goldfinch, Mr. RUSSELL,
Silky, Mr. COLLIN,
Widow Warren, Mrs. CHAMBERS,
Sophia, Mrs. VILLARS,
Mrs. Ledger, Mrs. STANLEY, Jenny, Miſs STANDEN.

TASK

Transcribe the genre

Type in the genre of the highlighted title, exactly as displayed on the page. If there is no genre displayed you can leave the field blank.

GENRE

Transcribe the genre

Seen something interesting? Add a note

Figure 2: Screenshot of the In the Spotlight interface, with interface elements designed to encourage exploration highlighted in red and orange.

In addition to the activities outlined above, members of the team present at conferences and summer schools. We publish articles, case studies [http://bl.uk/digital] and blog posts [http://britishlibrary.typepad.co.uk/digital-scholarship/] on digital scholarship with the Library's collections. Case studies published on the Digital Scholarship website [http://bl.uk/digital] help scholars understand how their work could benefit from new and emerging methods for working with digitised collections. We also deliver versions of our Training Programme courses for PhD students and academic departments, and run evening events on topics related to Digital Scholarship for the public.

Conclusion

Describing this work at the British Library is to write from a position of privilege. The Library's investment in digitisation and digital scholarship is unusual, as are the hundreds of years of collecting collections at this scale. However, many of the activities described above can be scaled up or down for use in different contexts, or adapted in collaboration with other departments. Technology underlies many of the methods referenced but the real difference is in our investment in outreach, and in the Library's commitment to make collections accessible to everyone, for 'research, inspiration and enjoyment'.

References

McGregor, N, Ridge, M, Wisdom S & Alencar-Brayner A 2016, 'The Digital Scholarship Training Programme at British Library: Concluding Report & Future Developments', Text. Available at: http://dh2016.adho.org/abstracts/static/data/133.html.

Ridge, M, 2013, 'From Tagging to Theorizing: Deepening Engagement with Cultural Heritage through Crowdsourcing', *Curator: The Museum Journal* Vol. 56, No. 4.