Published on *Reviews in History* (https://reviews.history.ac.uk)

# JISC Historical Texts

**Review Number:** 1770
**Publish date:** Thursday, 21 May, 2015
**Editor:** Scott Gibbens
**Date of Publication:** 2014
**Publisher:** JISC
**Publisher url:** http://www.jisc.ac.uk/historical-texts
**Place of Publication:** North Leigh
**Reviewer:** Judith Siefring

Jisc's *Historical Texts* [2] brings together for the first time three important collections of historical texts, spanning five centuries: *Early English Books Online (EEBO)*, *Eighteenth Century Collections Online (ECCO)*, and the British Library 19th-century collection. The platform can be accessed via subscription at UK Higher and Further Education institutions.

*Historical Texts* faces a significant challenge – how to bring together three collections (and more in the future) from different sources, containing both images and textual material, some of which is supported by transcribed full text, some by OCR full text and some with no full text at all, and to make them cross-searchable. It is an admirable aim. The difficulties in doing so, though, are formidable.

By allowing scholars to search across these collections, *Historical Texts* helpfully breaks the date-restrictive parameters of its source materials. *EEBO* (a collection owned and developed by ProQuest) has a cut-off date of 1700, while *ECCO* (owned and developed by Gale Cengage) covers 1701–1800. The British Library's 19th-century collection in fact contains materials from 1789 to 1914. For those researching on the cusp of centuries, or those interested in thematic development, for example, the ability to search for sources via a single platform will be a boon.

*Historical Texts* offers a simple search box, allowing users to keyword search across all three collections. Searches can also be refined by limiting, via a drop-down menu, by title, author, printer or publisher, place of publication, bibliographic number, language, illustration description, subject, and genre. An advanced search option allows for further refinement, within collections, for example, or within illustrated publications only. As with all resources of this type, if you know exactly what you are looking for, searching is a breeze. I want to look at Margaret Cavendish's *Poems and Fancies*, published in 1653. I search for 'Cavendish, Margaret' and get 749 results. I then filter by author (26 results) and by date (1653 – 2 results). I then click on the entry for *Poems and Fancies*. If I have no particular edition of a text in mind when searching, this is also quite simple. If I want to look at Shakespeare's *Venus and Adonis*, for example, I can search for 'Venus and Adonis' in the simple search box, which takes me to all examples across the collections, which can then easily be filtered by author, date, collection and so on. I can then whittle down my search results to focus on the particular editions that interest me.

Those with more general, or less well-defined, interests may find searching trickier. Keyword searching illustrates the value of careful consideration of the breadth of the resource (true of all digital resources of this type) and of good search strategies. A keyword search for 'parliament', for example, returns 120354 hits.

‘Witch’ returns 31580 hits. As with the search examples outlined above, these results can be filtered in various ways, including by genre and subject, but users would need to be prepared to put in considerable work to get any meaningful results out of such key word searches. Say I’m interested in historical travellers’ accounts of their visits to the Oxfordshire town of Abingdon. I search for ‘Abingdon’ in the main search box and it returns 7590 hits. I must also remember to use the fuzzy search and variant spelling options, which increases the number of hits for ‘Abingdon’ to 7701. I can refine such an unwieldy set of results: by subject (e.g. description and travel – 20 hits), by genre (e.g. non-fiction – ten hits), or by particular year (e.g. 1790 – 113 hits). In doing so, though, it is far from certain that I’m creating a meaningful subset of texts from which I can draw robust conclusions about visitors to Abingdon in the period. Furthermore, the search results pull up relevant volumes, but users can’t see the actual hits for ‘Abingdon’ within the texts concerned. This is functionality that users of ProQuest’s *EEBO* and Gale Cengage’s *ECCO* are used to and will miss.

Some users may prefer to browse rather than search the *Historical Texts* collections. Browsing can be done in a variety of useful ways: by author, printer or publisher, date, volume, and within the Thomason Tracts and the Thomason Tracts broadsides. As an example, if we browse by author, clicking on a particular writer pulls up thumbnails and short records for all works by that writer across the collections. When looking at the works of Hannah Woolley, for example, we can see that of the 29 works available, three come from *ECCO* and the remainder from *EEBO*, and that eight of the items have full text. The first result is an edition of Woolley’s *The accomplish’d lady’s delight*, published around 1720, from the ECCO collection. The source collection is clearly marked as is the availability of full text. When we click through to the text, the page images are delivered in a clear, easily navigable image viewer. A button to the left of the screen pulls up a convenient thumbnails view, while a button to the right pulls up the bibliographic record and a tab to the full text for that page. The interface here is nicely clean and easy to work with.

However, the juxtaposition of image and text here brings to light another problem with the source collections. The print and resulting image quality for this edition of *The accomplish’d lady’s delight* is very poor, and the *ECCO* full text accompanying it here is produced via OCR. The resulting full text is sadly useless. The first part of the title page reads (my transcription):

> The Accomplish’d LADY’S DELIGHT, IN PRESERVING, PHYSICK, BEAUTYFYING, COOKERY, AND GARDENING. CONTAINING I. The Art of *Preserving*, and *Candying*, Fruits and Flowers, and making all sorts of Conserves, Syrups, Jellies, and Pickles

The full text rendering of this passage in *Historical Texts* illustrates the perils of using OCR for early print materials:

> The Accomp!inh'd I A, D 'S DE :I, I s l N PRESERVING, PIiYSICK, BE/AUT'r ,-- i JC, COOKERx, and GAR1DEING. C 0 N 'F A I N 1 N G, I. The iArt of PrsIr.viN , and Cr?;'0 - i i;,, Fruitsand Flowers, andl mkinlg al l forts of Conserves, Syrups, Jelics, and Pic!l!s

There is significant ongoing work to improve OCR of historical materials, notably by the *Early Modern OCR Project* (*EMOP*) at Texas A&M University, but for now, the dreadful quality of the OCR underpinning the 18th-century materials included in *Historical Texts* is a real disadvantage.

Once they have identified the text that they are interested in, users of *Historical Texts* can easily download the individual page image as a JPEG or the whole publication as a PDF. They can also download a citation in RIS format. Making it as easy as possible for users of digital resources to properly cite those resources is absolutely vital both for users and for digital content creators. It is extremely positive that Jisc have chosen to include this functionality, but the RIS-only format may be confusing to users and useless to those without the correct bibliographic software installed on their computer. Future improvements to functionality could include the ability to download a citation in multiple formats. Users can, however, click the 'Share this publication' button, which generates a permanent URL for the page.

Some additional functionality within the site is still in development, including the ability to save groups of texts as 'My Texts', which will prove particularly user-friendly. While the site is clear and navigable in general, it will benefit from ongoing improvement planned by Jisc. When I had finished looking at *The Accomplish'd Lady's delight*, for example, I could see no obvious way to return to my browsing list of Hannah Woolley's works and had to resort to my browser's back button. This took me back to my list of all authors beginning with Wo-, and I had to re-expand Woolley – not terrible, just more laborious than I'd like.

*Historical Texts*' help pages are detailed, constructive, and also very necessary. Bringing together such a range of materials inevitably generates many questions, and Jisc have gone some way to answering the most practical ones in their help pages. Detailed information is included about searching and browsing functionality, including Boolean operators, regular expressions, and fuzzy searching. However, some areas could do with enhancement (although we must bear in mind the universal problem for digital content creators – users rarely read the help pages). Information on the full text supporting the images, for example, is buried in the help page, but could usefully be expanded upon and explained. *Historical Texts* does state clearly that full text is available for the entire *ECCO* and British Library 19th-century collections and that this textual data is generated by OCR software. Full text is available for many, but not all, of the *EEBO* materials. This is because full text for the *EEBO* materials, in contrast to that for *ECCO* and the BL collection, has been hand-transcribed and encoded by the Text Creation Partnership. This distinction is important and is only the beginning of any engagement with the textual aspects of the resource. As illustrated above, OCR is extremely problematic when applied to historical materials and has very variable results. Hand-transcribed materials fare better but nonetheless are not perfect – Text Creation Partnership materials are indeed hand-transcribed, but the sheer volume of production meant only partial proofreading and correction of volumes was possible. It is vital that users of digital resources properly engage with the editorial processes which governed their creation, in order to fully understand the assumptions that can be made about research based on their use. This is of particular importance for *Historical Texts* because the platform brings together disparate collections, created in different ways. Jisc could do users of the data sets a real service by highlighting and exploring these areas more fully – although the difficulty of ensuring that supporting documentation is actually read would remain a headache.

Jisc continue to plan improvements to the *Historical Texts* platform. They are working to expand content by adding more *EEBO* Text Creation Partnership full texts (which will be downloadable in various formats), and are collaborating with the Wellcome Trust to include the UK 19th-century medical collection. They also plan to include the Burney Collection of 17th- and 18th-century newspapers and pamphlets. These will be wonderful additions to an already impressive spread of collections.

Jisc are working hard behind the scenes on particular areas of technical development. Work is ongoing in collaboration with researchers at the University of Lancaster on a new variant spelling engine, whilst a collaboration with experts at the University of Oxford is looking at the mechanics of searching images. Such work may be trialled initially through the planned 'Labs' section of the interface, which will allow in-development work to be tested.

More obviously for users of the resource, Jisc are also looking to improve interface functionality. They are focusing on making the collection easy to use, including improved performance via mobile devices. Other

planned improvements include clickable tables of contents for individual texts and a notes feature, which will allow users to make and save notes on the texts they are using. Jisc are commendably responsive to user feedback on their experience of the resource. What is less clear is how they can cope with corrections and improvements to the content. Both the *EEBO-TCP* full texts and the OCR-created *ECCO* and 19th-century British Library content could be considerably improved. How might such improvements be made? Jisc's resource could prove the ideal locus for collaborative curation and crowd-correction of these key historical collections. Involvement of the user community in this way would allow for significant improvement to the core content, and would also help to build a community of scholars around the resource.

The continued improvement of the resource, including the addition of much more content, and the potential for correction and updating does raise its own problems. If a resource is continually updated, as it must be to be relevant sustainable, how can users keep abreast of these changes? Will it be obvious that content has been updated since they last used it? Will their search results have changed (and indeed the conclusions that they have drawn from such searches)? Raising awareness of the issues around using digital resources would seem to be key – a kind of digital basic training, which ought to be part of undergraduate courses in all subjects.

Jisc's desire to provide access to large amounts of content and their commitment to meeting user needs are both laudable, and many of the problems with the *Historical Texts* site result from their bringing together large collections with distinct identities of their own, and trying to integrate them. What could improve the site further would be an even greater focus on user journeys. How do users navigate the site? What are their expectations? What kinds of searches do they want to do? How does the current interface encourage or prevent users getting to where they need to be? How can users be made more aware of the nature of the resources they are using? All of these questions ought to be asked by any creators of digital resources – only by truly understanding how users are interacting with our collections can we adequately serve their needs. Users of digital resources are (rightly) demanding: they want lots of content and they want to be able to search it and read it and download it in multiple formats, and they want to do it all intuitively and without reference to help pages. It's a tough challenge, but one that all of us involved in digital content delivery must meet.

---



**Links**
[1] https://reviews.history.ac.uk/item/136116
[2] http://historicaltexts.jisc.ac.uk/