UKSG

Europeana Newspapers: searching digitized historical newspapers from 23 European countries

Europeana Newspapers is a European Commission-funded project which is refining, aggregating and giving researchers online access to historical newspaper content from 23 European libraries. It also offers free, open source tools which individual libraries can use to assess refinement quality and metadata standards in relation to their own digital newspaper collections. The content made available by the project is available both through Europeana and a dedicated newspaper browser. Europeana is a way to see the newspapers in the context of over 30 million other items of cultural heritage, while the dedicated browser hosted on The European Library site gives researchers the possibility to refine their search terms (e.g. by date, publishing country, issue) and to search the full text of many newspapers. So far, the browser, which has been online for approximately one year, has proven popular with researchers, and further improvements to its functionality are planned in 2015. This could include the ability for users to tag articles or to edit, and help correct, the full text of individual articles.

Introduction to Europeana Newspapers

The Europeana Newspapers Project[1] is a best practice network funded by the European Commission (EC) under the Competitiveness and Innovation Framework Programme (CIP) 2007-2013[2]. The project aims to improve access to digitized historical newspapers by refining and aggregating historical newspaper content from 23 European libraries. This rich collection, with newspaper content dating as far back as 1618, is made accessible via two prominent cultural heritage websites: The European Library[3] and Europeana[4] itself.

The project creates full text from 10 million already digitized historical newspaper pages (originating from 12 partner libraries in the project) through optical character recognition (OCR). Some of these pages have been further enhanced with optical layout recognition (OLR) to create article-level records and enriched with named entity recognition (NER) in Dutch, French and German to make it possible for users to search for names of persons, places and organizations.

In addition to these 10 million refined historical newspaper pages, the metadata for millions of historical newspaper pages is also being made available. As a best practice network, Europeana Newspapers has also developed a number of freely available tools[5] which individual libraries can use to assess refinement quality and metadata standards in relation to their own digital newspaper collections.

MARIEKE WILLEMS
(formerly of LIBER)

ROSSITZA ATANASSOVA

Digital Curator, Digitization The British Library

The Europeana Newspapers browser

A key part of the Europeana Newspapers project was the development of a browser, which enables users to perform full-text searches across millions of historical newspaper pages. The prototype of this browser has been available since January 2014. Since that time, small continual improvements have been implemented and the final interface is now available.

The prototype version of the browser has all the functionalities expected of a digitized newspaper site. Users can search by keyword, across the full text or the newspaper titles.

Various filters are available to refine search results, for example by content provider, language or date ranges. Users can also browse the content by selecting a specific date or newspaper title, or via a global map.

The search results display the issue-level metadata of 25 content-providing libraries. For eleven of these libraries, the full text and newspaper page image are displayed, with links to view the original in the source library, if available. Since the interface provides access to multinational content, the aggregation and browser design needed to take into account any restrictions imposed by contributing libraries due to national copyright laws and the libraries' business models for digitized newspapers. Therefore, the contributing partners were presented with four options to help them decide how their content should appear on the newspapers interface:

'the interface provides access to multinational content'

- Option 1: metadata, full text and full, zoomable images

- Option 2: metadata, full text and static images – either full size or snippets

- Option 3: metadata and full text only

- Option 4: metadata only.

The majority of libraries have accepted Option 1 and allow display of full-size, zoomable images that are either ingested by The European Library or delivered directly from the libraries' own image server. With a more consistent user experience and better search functionality in mind, The European Library has been able to convince some partners to offer a static full-size image rather than a snippet view, with some also considering zoomable images.

In order to inform the further development of the prototype interface, The European Library arranged for a round of usability testing in April 2014. The main objectives of the tests were to understand the needs and expectations of users of digitized historical newspapers, to evaluate their experience of a pan-European newspapers interface and to recommend any changes that would improve the site functionality and design.

Feedback from usability testing

Twelve participants from five countries represented in the project (UK, Latvia, Austria, Italy and Finland) took part in the first round of usability testing. The 60-minute long remote online test sessions were conducted by UserVision, a company of independent usability experts based in Edinburgh, Scotland. The user group for the test consisted of people with professional or personal research interests in historical newspaper collections.

The participants joined an online meeting with a UserVision moderator and were asked to complete six pre-determined tasks using the newspaper interface on The European Library portal. The scenarios included an exploration of the landing page, the performance of a basic keyword search for a place name, a refinement of the search results by country of publication, a search by date, by title and by region.

The consultants observed the participants' performance and offered assistance with the tasks where necessary. Before, during and after the sessions, participants were asked questions to determine their expectations, identify and explore areas of concern and formulate recommendations as to how issues they encountered can be addressed.

The user feedback from the usability test was broadly positive. Participants reacted well to the layout and functionality of the site. Their initial expectations of the Europeana Newspapers interface were high due to the quality of The European Library website and the scope of the content made available by the project. Overall, participants gave the site a rating of 'slightly above expectations'.[6]

'Participants reacted well to the layout and functionality of the site'

Requirements for enhanced search functionality

Participants in the usability testing confirmed that search results displayed well and the information provided in the link, the short description and image thumbnail helped them assess easily the relevance of the result.

At the time of the usability test, however, the order of search results was not configurable and this prevented participants from understanding the order of results or exercising any control over it. The updated interface displays results by relevance and also allows users to sort by ascending and descending date and configure the number of results displayed on a web page. Another change that has already been implemented is the improved management of filters through the addition of an 'x' icon next to a selected filter.

The addition of a full screen mode option (relating to the newspaper image and full text pane presentation) is an important enhancement and allows users to see complete lines of text without the need for scrolling. This addresses the navigation difficulties experienced during the usability testing, due to the restricted horizontal width of the content pane window.

The browser interface also features navigation controls to allow back and forth movement between search results, and a 'back to search results' button control.

The usability test also strongly recommended the addition of a search input box on the results page 'populated with the original search term(s) entered', to allow users to modify their search terms or perform a new search without having to go back to the landing page and thus resetting the already selected filters.

The lack of an option to download or save locally images and associated metadata was seen as an obvious shortcoming of the interface, and this needs to be addressed. The development of this option, and the related print option, will also need to meet the expectations of contributing libraries.

User research practices and expectations

Judging from the participants' performance of the tasks set in the usability testing and the feedback from the interviews, it seems their preferred method of using the historical newspapers archive is through controlled and refined search options, rather than through browsing. This is explained by the difficulty of browsing such a vast newspaper archive and by the participants' strong research interest in historical newspapers and well-established search strategies.

This particular group of users would like to see more advanced search options or facets to help them filter through and manipulate the search results. Such functionalities would be consistent with the advanced search options (Boolean search, article type facets, multiple layers of filters) implemented on other interfaces to digitized historical newspapers, such as the British Newspaper Archive[7], Chronicling America[8] and Trove[9]. Further options that participants mentioned include searching by 'subject area' and 'historical period', which points to a very specific rather than a general interest in the site.

In addition, the participants in the test expected to be able to create a user account offering a research space to save search histories. They would also like to be notified of newly published content on the site and to have the option to submit feedback. As already mentioned, their research needs require the ability to download a local copy of an image, text or metadata for the purpose of building their own personal archives.

A unique value of the Europeana Newspapers interface is that it offers cross-searchability of content published in over 20 languages, but this could also present a barrier for users. Users will need to know different languages to be able to make the most of the site. Even if one can confidently perform a basic search across all content, a deeper knowledge of languages is necessary to interpret the results. In the first round of usability testing,

'a deeper knowledge of languages is necessary to interpret the results'

participants did not comment on this as a problem, as they mainly searched content in their own language and from their own country. However, it has been suggested by researchers and librarians that the interface should embed tools to assist users with translation. This is a challenge because tools such as Google Translate are unable to effectively translate the type of content found on the Europeana Newspapers interface.

Researchers as a primary target group

As confirmed by the first round of usability testing, there is a great demand for making digitized historical newspapers available and the academic research community has therefore received the Europeana Newspapers project with great enthusiasm. LIBER[10], which leads the promotional work on the project, and other partners, including the British Library[11], have been actively engaging this user group to promote the resource and to better understand what use researchers will make of the archive, and in particular the new research possibilities that this content opens up.

A series of Q&A interviews with newspaper researchers from different European countries is published on the project website and highlights some of the ideas researchers have for mining the content on the Europeana Newspapers interface. The researchers interviewed turn to newspapers to study a range of subjects and topics: 19th-century popular culture and humour, history, literature, evolution of language, public discourse, reference cultures and professional careers.

For them, such an aggregation of millions of pages of European newspapers offers exciting new opportunities for transnational comparative research and computational analysis of the data. One of the interviewees, Professor Toine Pieters of Utrecht, sees the multilingualism of the archive not as a barrier but rather as a challenge that needs to be overcome and is already being addressed by a project which will explore reference cultures in Europe with the help of multilingual text mining techniques.[12]

'an aggregation of millions of pages of European newspapers offers exciting new opportunities for transnational comparative research'

The Europeana Newspapers Information Days, organized by project partners, have been another vehicle for engaging the researcher community. In total, ten Information days have taken place, in Turkey[13], Latvia[14], Poland[15], Germany[16], UK[17], Italy[18], Austria[19], France[20], The Netherlands[21] and Estonia[22]. These events days brought together researchers in the fields of history, literature, print culture, media and social science, as well as digital humanities scholars. This latter group is particularly excited about the possibilities of exploiting historical newspaper content in new ways and applying digital humanities research methods to the data aggregated by Europeana Newspapers.

Newspapers contain a plethora of illustrations, maps, photographs, and one idea would be to extract through algorithm the illustrations found in the newspaper pages in the archive and invite users to tag, organize thematically or link back to captions and descriptions found in the newspapers. The inspiration for this idea comes from the British Library Labs project which extracted one million images from 65,000 digitized 19th-century books and released them on Flickr Commons for users to tag and reuse. This project enabled users to describe the images, create thematic albums, reuse creatively for commercial and educational purposes and use them for research in the areas of image recognition and automated classification of historical images.[23,24] The new 'Victorian Meme Machine' project, conducted by Bob Nicholson of Edge Hill University and supported by BL Labs, is creating a database of Victorian jokes extracted from digitized 19th-century British newspapers and will semi-automatically pair them with appropriate images from the Library's digital collections to create new context and reuse for these Victorian jokes.[25]

Such innovative approaches to digitized historical newspapers could be applied to the corpus created by Europeana Newspapers and would help attract new professional and amateur audiences and engage the wider user community. The value of Europeana Newspapers is not limited to academic researchers, but also genealogists, local historians,

the teaching and learning community and all citizens of Europe and outside Europe. The content aggregated by the project would be a valuable resource for the study of European history, society, culture, languages, publishing, literature, art, design and much more.

Next steps for the interface

Further changes relating to search results are to be implemented to the 'browser' interface before its final release in early 2015. These include the ability of users to locate 'named entities', i.e. specific mentions of people and places even if they have different spellings (for example, Deen Haag, La Haya, S' Gravenhage).

Time and resources permitting, The European Library aspires to implement additional features, such as an option for users to correct the OCRed text. This functionality has been successfully implemented by other newspaper archives, like the British Newspaper Archive and Trove, and there are many good reasons why it should be added on the Europeana Newspapers interface. The ability to edit and even tag articles would be appreciated by many user groups and would both improve the level of text accuracy in the corpus and increase user engagement with this digital archive. The European Library is also looking into the possibility of creating an application programming interface (API) to provide access to large sets of data via machine harvesting and analysis.

The complexity of providing access to digitized historical newspapers is not unique to Europeana Newspapers, but the challenges are augmented by the size and range of the data set involved. To create a fulfilling online experience, the project interface has to effectively display not only the content overall but also the individual characteristics of each paper. It is important to show what is available as well as gaps in the coverage, and to manage expectations with regard to the quality of the images and full text. Some of these challenges are shared and addressed by other digitized historical newspaper interfaces[26], whilst others are more specific to the Europeana Newspapers project and are affected by political, economic and legal policy issues.[27] The development of the project interface will also need to be sustainable after the project ends in March 2015.

Many of the challenges of improving access to digitized historical newspapers and the opportunities that would result from successfully overcoming them were discussed at the project's final public workshop, entitled 'Newspapers in Europe and the Digital Agenda in Europe' (held in London, UK, on 29-30 September 2014). During this workshop a wall-sized image was made on the basis of the discussions and presentations. The image displayed all policy issues that affect improving access to digitized historical newspapers, and its digital version is available to all.[28]

Competing interests: The authors have not declared any competing interests.

> 'content aggregated by the project would be a valuable resource for the study of European history, society, culture, languages, publishing, literature, art, design and much more'

References

1. Europeana Newspapers:
   http://www.europeana-newspapers.eu (accessed 17 February 2015).

2. Competitiveness and Innovation Framework Programme (CIP):
   http://europa.eu/legislation_summaries/information_society/strategies/n26104_en.htm (accessed 17 February 2015).

3. The European Library's browser for historic newspapers:
   http://www.theeuropeanlibrary.org/tel4/newspapers (accessed 17 February 2015).

4. Europeana:
   http://www.europeana.eu (accessed 17 February 2015).

5. Free tools for digitized historic newspapers:
   http://www.europeana-newspapers.eu/public-materials/tools (accessed 17 February 2015).

6. Blackwood, A, *The European Library: Newspaper Archive – Usability Testing*, April 2014:
   http://www.europeana-newspapers.eu/wp-content/uploads/2014/05/The-European-Library-Newspaper-Archive-Usability-testing-Report-April-2014.pdf (accessed 17 February 2015).

7. British Newspaper Archive:
   http://www.britishnewspaperarchive.co.uk/search/advanced (accessed 17 February 2015).

8. Chronicling America:
    http://chroniclingamerica.loc.gov/ (accessed 17 February 2015).

9. Trove:
    http://trove.nla.gov.au/ (accessed 17 February 2015).

10. LIBER:
    http://www.libereurope.eu (accessed 17 February 2015).

11. British Library:
    http://www.bl.uk (accessed 17 February 2015).

12. Europeana Newspapers interview with Toine Peters:
    http://www.europeana-newspapers.eu/qa-with-newspapers-researchers-toine-pieters (accessed 17 February 2015).

13. Turkish Information Day report:
    http://www.europeana-newspapers.eu/wp-content/uploads/2012/04/ENP_Report_Infoday_NLT.pdf (accessed 17 February 2015).

14. Latvian Information Day report:
    http://www.europeana-newspapers.eu/wp-content/uploads/2012/04/ENP_Report_Infoday_NLL.pdf (accessed 17 February 2015).

15. Polish Information Day report:
    http://www.europeana-newspapers.eu/wp-content/uploads/2012/04/ENP_Report_Infoday_NLP.pdf (accessed 17 February 2015).

16. German Information Day report:
    http://www.europeana-newspapers.eu/wp-content/uploads/2012/04/ENP_Report_Infoday_SBB_final.pdf (accessed 17 February 2015).

17. British Information Day report:
    http://www.europeana-newspapers.eu/wp-content/uploads/2014/08/D6.31-ENP-Report-InfoDay-BL.pdf (accessed 17 February 2015).

18. Italian Information Day report:
    http://www.europeana-newspapers.eu/wp-content/uploads/2012/04/ENP_Report_Infoday_LFT_1.0.pdf (accessed 17 February 2015).

19. Austrian Information Day report:
    http://www.europeana-newspapers.eu/wp-content/uploads/2012/04/ENP_Report_Infoday_ONB_1.0.pdf (accessed 17 February 2015).

20. French Information Day report:
    http://www.europeana-newspapers.eu/wp-content/uploads/2014/12/ENP_Report_Infoday_BnF_1.0.pdf (accessed 17 February 2015).

21. Dutch Information Day report:
    http://www.europeana-newspapers.eu/wp-content/uploads/2012/04/ENP_Report_Infoday_KB_1.0.pdf (accessed 17 February 2015).

22. Estonian Information Day report:
    http://www.europeana-newspapers.eu/wp-content/uploads/2012/04/ENP_Report_Infoday_NLE_1.0.pdf (accessed 17 February 2015).

23. British Library Labs on Flickr: for more information about the release of the images, see
    http://britishlibrary.typepad.co.uk/digital-scholarship/2013/12/a-million-first-steps.html (accessed 17 February 2015).

24. For examples of a creative re-use of the images, see
    http://blpublicdomain.wikispaces.com (accessed 17 February 2015).

25. Victorian Meme Machine:
    https://www.youtube.com/watch?v=FN1ZSAz2vMg (accessed 17 February 2015).

26. Diving into newspaper archives: Chronicling America. Interview with Deborah Thomas:
    http://www.europeana-newspapers.eu/diving-into-newspaper-archives-chronicling-america/ (accessed 17 February 2015).

27. Dunning, A and Neudecker, C, Representation and Absence in Digital Resources: The Case of Europeana Newspapers, *Digital Humanities*, 2014:
    http://dharchive.org/paper/DH2014/Paper-773.xml (accessed 17 February 2015).

28. Illustrating the Issues Related to Digital Newspapers, Europeana Newspapers:
    http://www.europeana-newspapers.eu/digital-newspapers-illustration/ (accessed 17 February 2015).

Corresponding author: Rossitza Atanassova
Digital Curator, Digitization
The British Library, 96 Euston Road, London NW1 2DB, UK
E-mail: Rossitza.Atanassova@bl.uk

To cite this article: