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

Europeana Newspapers

Review Number: 1894
Publish date: Thursday, 18 February, 2016
Editor: Clemens Neudecker
Date of Publication: 2015
Publisher: Europeana Newspapers
Publisher url: http://www.europeana-newspapers.eu/
Place of Publication: Berlin
Reviewer: Bob Nicholson

It is generally assumed that the digital revolution will spell the end for print journalism. Newspaper sales are in terminal decline as an increasing number of readers turn to websites, smartphones, and social media for their news and entertainment. However, while the internet may eventually kill off modern-day newspapers, it has managed to breathe new life into their ancestors. Over the last decade, millions of pages from thousands of historical newspapers and periodicals have been digitised and made available via a rapidly expanding series of online archives. A few years ago, it was possible to list most of these databases in a single paragraph, but we have now reached a point where even the most dedicated press historians struggle to keep track of the riches that are now at our fingertips. In these times of plenty, researchers would be forgiven for greeting the launch of yet another newspaper archive with a bit less enthusiasm than in the past. However, *Europeana Newspapers* [2] is more than just another archive – it is an ambitious project that aims to tackle some of the biggest challenges involved in newspaper digitisation and has the potential to become a valuable new resource for the study of European history.

The project is an EU-funded initiative that aims to make the continent's historical newspapers and periodicals easier to find and explore. It is not a conventional digitisation project. Rather than scan new material from scratch, *Europeana Newspapers* aims to improve the quality and accessibility of *existing* digital collections. It has three key objectives:

1. To refine the quality of 10 million existing digitised newspaper pages, i.e. by applying Optical Character Recognition in order to enable full-text searches.
2. To develop a new search interface that allows researchers to explore newspapers from a range of European libraries in one place.
3. To develop/identify open-source tools and best practices that can be used for future digitisation projects.

The project launched in 2012 and officially ran until May 2015. A prototype of the new search interface [3] is complete and most of its promised features are in place. However, the archive is still in something of a transitional phase. At present, it is integrated into the website of The European Library. But this is just a temporary home. Over the course of 2016, the archive will be moved to a dedicated section of the Europeana [4] portal. It is too early to say how this migration might change the user experience. The search engine's core functions and content should remain intact, but it is possible that the interface described in this review will be altered in both positive and negative ways. More excitingly, it is also foreseen that Europeana's ambitious new cloud-based infrastructure will allow researchers to access the newspaper dataset via an API.

This would enable the use and development of advanced digital research tools that are not available via the basic search interface. Corpus analysis software, for example, could be used to 'distant read' thousands of newspapers and identify linguistic changes and patterns. Alternatively, the techniques used by Melodee Beals [5] and America's *Viral Texts* [6] project could be used to track the flow of news and map out European networks of reprinting. An open API would also encourage the creative 'remixing' of newspaper content for both academic and artistic purposes. But let's not get ahead of ourselves. For now, this review reflects on what the *Europeana Newspapers* project has achieved so far, and considers how historians might make use of its basic new search platform.

The archive is composed of newspapers and periodicals from nineteen different European countries, including particularly large collections from France, Austria, the Netherlands, and Germany. It covers an impressive timespan. The earliest available paper, the Dutch *Courante uyt Italien, Duytslandt, &c.* was published in 1618. At the other end of the spectrum, it is possible to access copies of several Icelandic newspapers that were published as recently as last month. Like most digital archives, the majority of titles come from the 19th and early 20th centuries – a period that has been particularly well-served by digitisation projects thanks to an abundance of material and the absence of copyright restrictions. That said, historians of the 20th century will be pleased to find a reasonable amount of content from the inter-war period, though the holdings for France, Austria, Germany and the Netherlands all cut off in the 1940s. There are a small number of post-war publications available from Luxemburg, Latvia, and Iceland, but scholars who are hoping to explore titles from the 1960s, 1970s, and 1980s will be disappointed. This is not the fault of the *Europeana* platform, but a wider problem with newspaper digitisation projects. The period between the 1960s and 1990s is something of a digital no man's land. From the the 90s onwards we have fairly good coverage using databases of born-digital news, but the digitisation of paper-based archives from these three preceding decades is encumbered by copyright. The *British Newspaper Archive* [7], for example, currently cuts off in 1959, while Australia's *Trove* [8] ends in 1954. Not long ago, this dividing line was drawn more cautiously at 1900 (*19th Century British Library Newspapers*) or the 1920s (*Chronicling America* [9]), so we're gradually making progress. *Europeana Newspapers* pushes the boundary a little bit further, but it will probably be several years before historians of the post-war decades have access to the same digital resources as their colleagues who study earlier periods.

All in all, the *Europeana Newspapers* platform provides varying degrees of access to 3,483 different titles. This is an impressive number, but it comes with some important provisos. Firstly, many of these papers are only available in limited runs. Some, such as the Dutch *Nymeegsche Courant*, only have a single issue online. Other holdings are much more substantial. France's *Le Journal des Débats*, for example, is available from 1814 to 1945, while copies of Germany's *Hamburger Nachrichten* can be accessed from 1792 to 1939. As with most newspaper archives, some of these longer runs still have missing issues. The interface simply lists the 'Start Date' and 'End Date' for each paper, which means that these gaps are not immediately obvious. Researchers should be cautious when exploring the database using keyword searches – a gap in results might simply be a gap in the archive. Using the 'Explore by Issue Date' tool it is possible to check for missing issues of a specific paper by identifying boxes on the calendar that are not shaded blue. In the example below, we can see that the 30 July 1845 issue of *Le Journal des Débats* is missing:

Image not found or type unknown

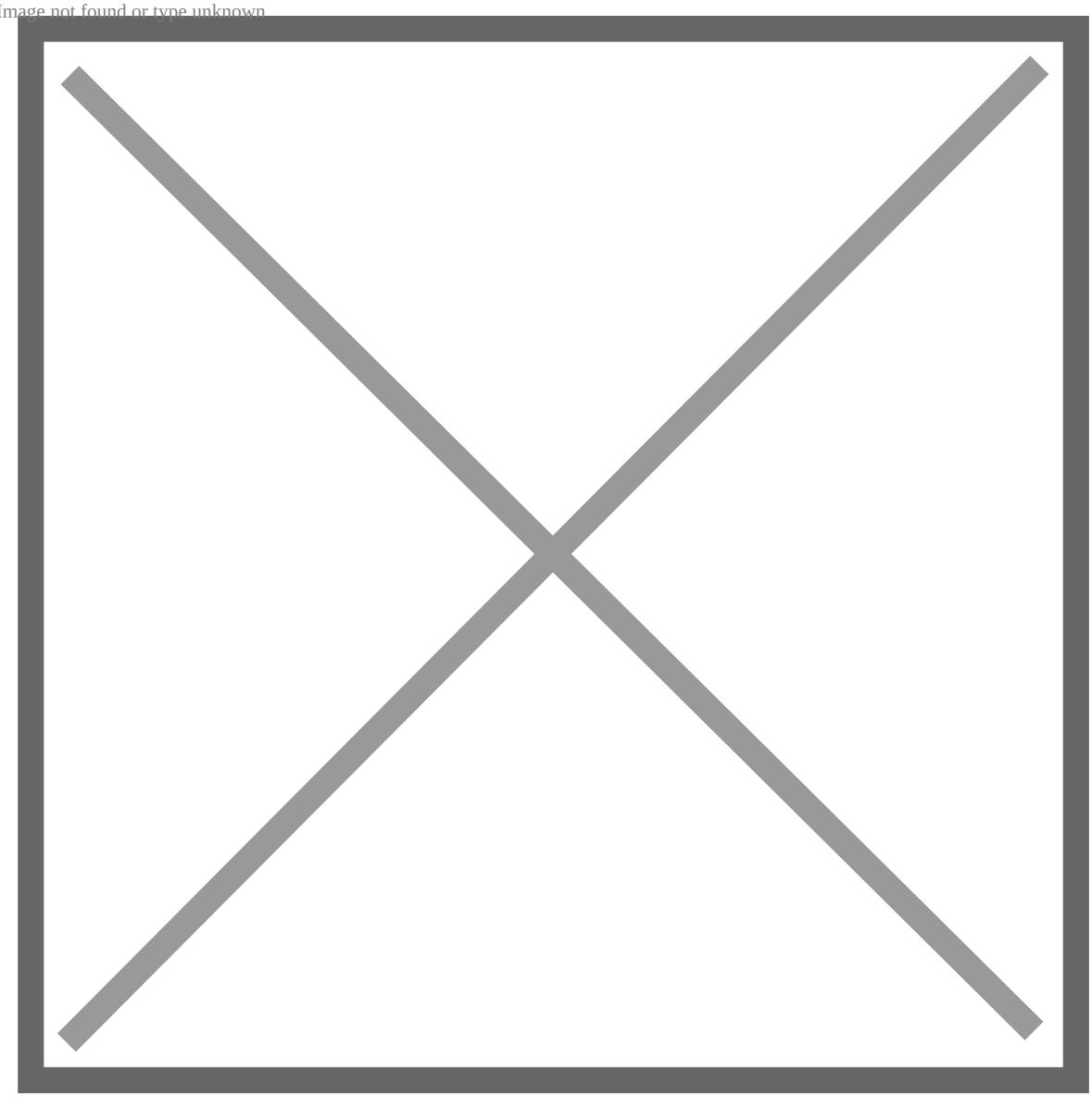

Fig 1. Using the 'Explore by Issue Date' tool to Identify missing issues in *Le Journal des Débats*.

This is a rather laborious process and is only viable for narrowly focused enquiries. It is important to stress that the *Europeana Newspapers* project is not to blame for these missing issues. Their search platform makes use of newspaper collections that have already been digitised by other European libraries and has inherited the gaps from these older datasets.

23 European libraries are involved in the project. Combining content from so many different partners into a single platform comes with plenty of technical, legal, and economic challenges. Each of the libraries adopted different approaches when digitising their newspaper collections – their scans were captured at different resolutions, saved using different file formats, processed (or not) using different OCR and page-segmentation software, and then organised using their own metadata systems. Moreover, each library has its own business model. Some make their historic newspapers freely available anywhere online, others require users to visit the library's own website, and several are enclosed behind paywalls. Copyright on these texts is complicated and varies from country to country (and sometimes from newspaper to newspaper, depending on agreements with publishers). The *Europeana Newspapers* project found a way to circumvent these obstacles by offering libraries a degree of flexibility in how their data is presented to the public. Most of the

project's 12 'full partners' have granted permission for their newspapers to be searchable and viewable using the project's own interface. On the other hand, 'Associate Partners' have typically chosen a more restrictive approach. Some have opted into the full-text search engine, but require users to click through to their own website in order to view the newspaper pages. Other associate libraries have simply donated metadata for their newspaper holdings, which means that it is not possible to explore them using full-text search. The National Library of Spain, for example, has contributed more titles (1,107) and individual issues (479,370) than any other provider, but none of them are fully searchable using the *Europeana Newspapers* interface. Instead, users can search for the titles and publication dates of specific Spanish papers, and then click through to access individual issues using the library's own, rather basic, PDF viewer. Accommodating the needs of all of these libraries is an impressive achievement, but it results in a rather uneven user experience. Researchers will need to investigate the limitations of each section of the archive before trusting the accuracy of keyword search results.

At present, the search interface is rather basic and lacks some of the features that we've come to expect from commercial newspaper archives. There is a single box in which to enter search terms, which will probably be sufficient for most simple queries. Entering multiple search terms returns articles that contain *at least one* of your chosen words, rather than limiting itself to those that contain *all* of them. So, a search for 'America, England' will return many articles that only feature the word 'England.' As far as I can tell, the search engine doesn't appear to support standard Boolean operators, so a search for 'America AND England' won't solve the problem – in fact, it returns thousands of articles containing variations on the word 'and'. This example highlights another problem. The search engine automatically makes use of 'fuzzy' searches, which return variations on your search terms. Used correctly, this offers a useful way to work around OCR errors (sometimes 'America' is automatically transcribed as 'Amorica'), pluralised words (American/Americans) and variations in spelling (Humor/Humour). However, other archives give users the ability to turn this feature off and control how far it is permitted to deviate from the original search terms. *Europeana Newspapers* has fuzzy searching permanently turned on, which means that some search terms return a lot of irrelevant results. The archive tries to side-step these problems by listing results in order of 'relevance' (i.e. how often your precise search terms appear in each article), but as soon as you sort results by date this effect is undone and users are left to wade through thousands of articles that bear little relation to their search. This lack of precision makes it difficult to construct focused searches. Commercial newspaper archives usually support the use of multiple search boxes, which make it easier to construct complex enquiries using multiple keywords. For example, Gale Cengage's *19th Century British Library Newspapers* archive would allow users to search for all articles that:

- feature **one** of the following words: 'America', 'American', 'United States', 'U.S.A', 'Yankee', or 'New York'
- **in combination with** one of these words: 'new', 'modern', 'fresh'
- and **exclude** any articles that feature the phrase 'South America'.

These kind of searches are essential when dealing with broad topics (such as press coverage of an entire country) and when attempting to navigate large archives. It is a shame that the current version of the *Europeana Newspapers* portal doesn't support these searches, and I hope that they'll consider implementing a fully-featured Advanced Search page once they have migrated to the main *Europeana Collections* platform.

On the plus side, the *Europeana Newspapers* search engine does have some useful search features. Enclosing a phrase within double quotation marks has the expected effect and will only return articles featuring that precise phrase. The search box also has a useful predictive feature that suggests alternative terms that are similar to the ones being entered. Here, for example, we see alternatives to a search for 'America':

Image not found or type unknown

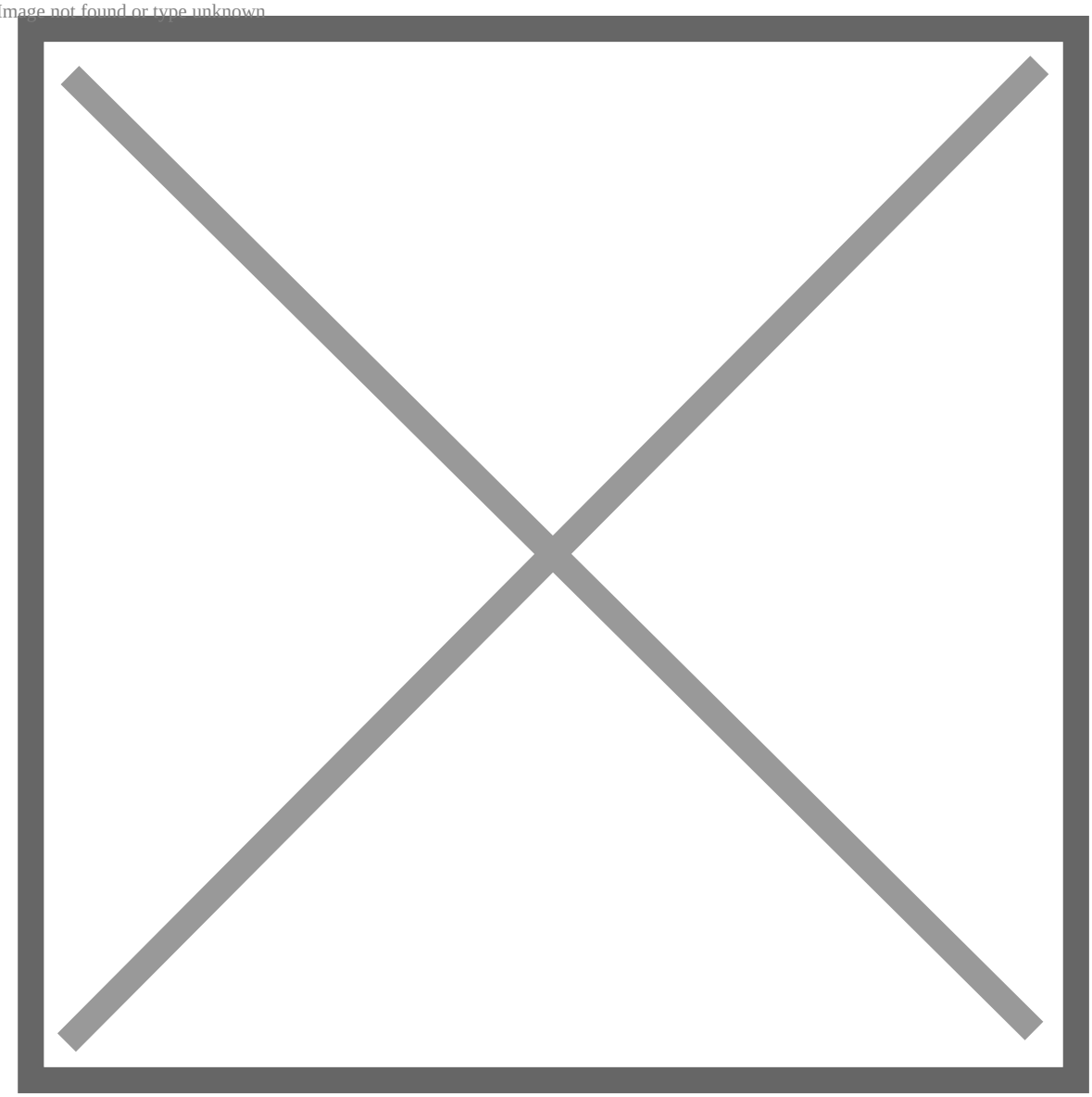

Fig 2. *Europeana Newspapers'* predictive search feature.

These predictive searches are a particularly useful feature for a pan-European, multi-lingual archive like *Europeana*. In order to construct searches that transcend national boundaries, it is necessary to identify viable keywords in a range of different languages. While the predictive search doesn't do all of this work for us, it often provides a helpful starting point and can highlight useful alternatives to basic search terms found using dictionaries and online translators.

The initial search window provides some basic filtering options. Below the search box are options that allow users to focus their enquiries on specific libraries, languages, and dates. The language option is a particularly useful addition for an archive of this nature, but it is important to recognise that the languages it lists are determined at publication-level. For example, focusing your search on the English-language will reveal some interesting titles from the National Library of the Netherlands, but will not find individual articles or passages that were written in English but printed in predominantly French or German speaking newspapers. In other words, there is more English-language content in the archive than the search filter suggests.

Once you have entered your search and clicked 'Go', the results page provides a second opportunity to filter, sort, and refine your enquiry. This time, the interface displays the number of results that fit into each particular category. This makes it easier to make an informed decision about how to filter a search and is

particularly useful for focusing on specific publications – an option that is not available via the initial search window. Results are also visualised on a small map of Europe which users can click on in order to focus on newspapers from a particular country. This is a neat idea and could be developed further by using a more detailed and interactive map, and by experimenting with ways to automatically geo-locate newspapers and articles to specific European cities or districts.

Search results are displayed in a fairly conventional list. As mentioned above, articles are initially sorted by 'relevance' rather than by date. So, it is important to remember that the article at the top of *Europeana Newspaper's* results page is not the earliest to make use of your search term. It is easy enough to re-order results so that they appear chronologically, but doing this each time you enter a search can become rather irritating – it would be helpful if the archive remembered users' preferences. Where full-text search is available, results are presented with a brief preview of the text that gives some indication as to an article's contents. This means that you can skip over obviously irrelevant articles without going through the trouble of loading them, which makes it much easier to wade through searches with thousands of results. This has become a fairly standard feature in most recent newspaper archives, but the *Europeana Newspapers* team have implemented it very effectively.

After clicking on a result, the experience of using the archive becomes more erratic. In some cases, you'll be whisked away to the website of one of the participating libraries and see your chosen article displayed within their own viewing platform. Results from associate libraries in Spain, Wales, Austria, Iceland, Serbia, Croatia, and Slovenia are all redirected to different library websites. Some offer fully-featured and user-friendly viewing experiences, while others are less intuitive and rather laborious to browse. The worst cases are simply unusable. Clicking on a result from Slovenia, for example, currently returns a tiny, low resolution thumbnail of a newspaper page that is impossible to read. Some archives, such as the National Libraries of the Netherlands and Austria, offer different viewing options depending on the newspaper in question. In the case of Dutch newspapers, material from before 1900 can be viewed on the Europeana site, while more recent publications require users to navigate to the library's own newspaper viewer. A similar dividing line for Austrian newspapers has been implemented in the 1870s. The Teismann Library has taken a slightly different approach and allows users to see a medium quality, full-page preview of all its newspaper pages on the Europeana site. However, in order to read articles comfortably it is necessary to click a button towards the top-left of the screen reading 'See Original At Library…'. The Europeana Newspapers team have done a remarkably good job of meeting each library's demands whilst ensuring that the core search interface remains useable. The process of navigating from one site to another is usually fairly painless – in most cases, users are automatically re-directed to the relevant site as soon as they click on a result, and are often and taken straight to the specific page/issue that they identified using Europeana's search engine.

The best experience comes when using *Europeana Newspaper's* own viewing platform. After you click on a search result, the relevant newspaper page is displayed in a zoomable HTML5 viewer. Keywords are highlighted in yellow, so users simply have to zoom in to the relevant section of the page in order to read the article that contains their search terms. There are pros and cons with this approach. On the one hand, it allows researchers to see an article in its original context and consider its position in the newspaper. Older archives often lead users straight to disembodied articles, which made them harder to contextualise and easier to misinterpret. However, zooming in each time you load a result can start to become rather tedious – particularly if you're wading through hundreds of articles. I also found the viewer's zoom function to be fiddly and overly sensitive, particularly when using my laptop's trackpad or a mouse's scroll wheel. Zooming in and out of the full page also requires the archive to load a lot more high-resolution images, which sometimes requires users to wait a for a few seconds while sections of the page load. More frustratingly, presenting newspapers in a zoomable viewer makes it very difficult for users to download them to their own computers or copy them over to a PowerPoint presentation. At present, it is not possible to simply right click the image and save it your computer, and nor does the Europeana archive provide a download button. This problem is by no means unique to Europeana's interface; most commercial newspaper archives have embraced similar, zoomable viewers in recent years, many of which are a far more clunky and restrictive than this one. Fortunately, the Europeana team have promised that a download feature

will be implemented at the end of 2016 once their new API is in place. This system should also allow users to download newspaper data in bulk and explore it using their own software. Indeed, the project has already released some of their newspaper datasets (OCR text and metadata, but not images) for download via the Europeana Research website [10]. Using these datasets will require some technical expertise, which means that they will mostly be of interest to scholars working in the Digital Humanities. Nevertheless, the project and its partners should be applauded for making large datasets available in such an open and accessible format. As a stop-gap, researchers can also download images of newspaper pages from the websites of partner libraries. For example, France's Bibliothèque Nationale provides a high-resolution download button if you view the newspaper on its own website. If this option is not available, the only remaining solution is to take screenshots using the Europeana viewer, though it will be very difficult to capture an entire page at a readable resolution.

The newspaper viewer initially appears to be enclosed in a fairly cramped window, but fortunately it has a 'full-screen' option that extends to fill the whole screen. This makes it possible to view the entire width of a newspaper page at once, providing you are not working on a particularly small laptop. Click on the newspaper and press the 'f' key on your keyboard to enable this feature, then press the 'escape' key to exit the full-screen viewer when you're finished. Newspapers that have been provided by the project's full partners also allow an advanced full-screen view that includes the digitised text alongside the image. This is usually the optimal way to browse a newspaper page, so look out for the blue arrows at the top left of the newspaper viewer in order to make use of it.

Image not found or type unknown

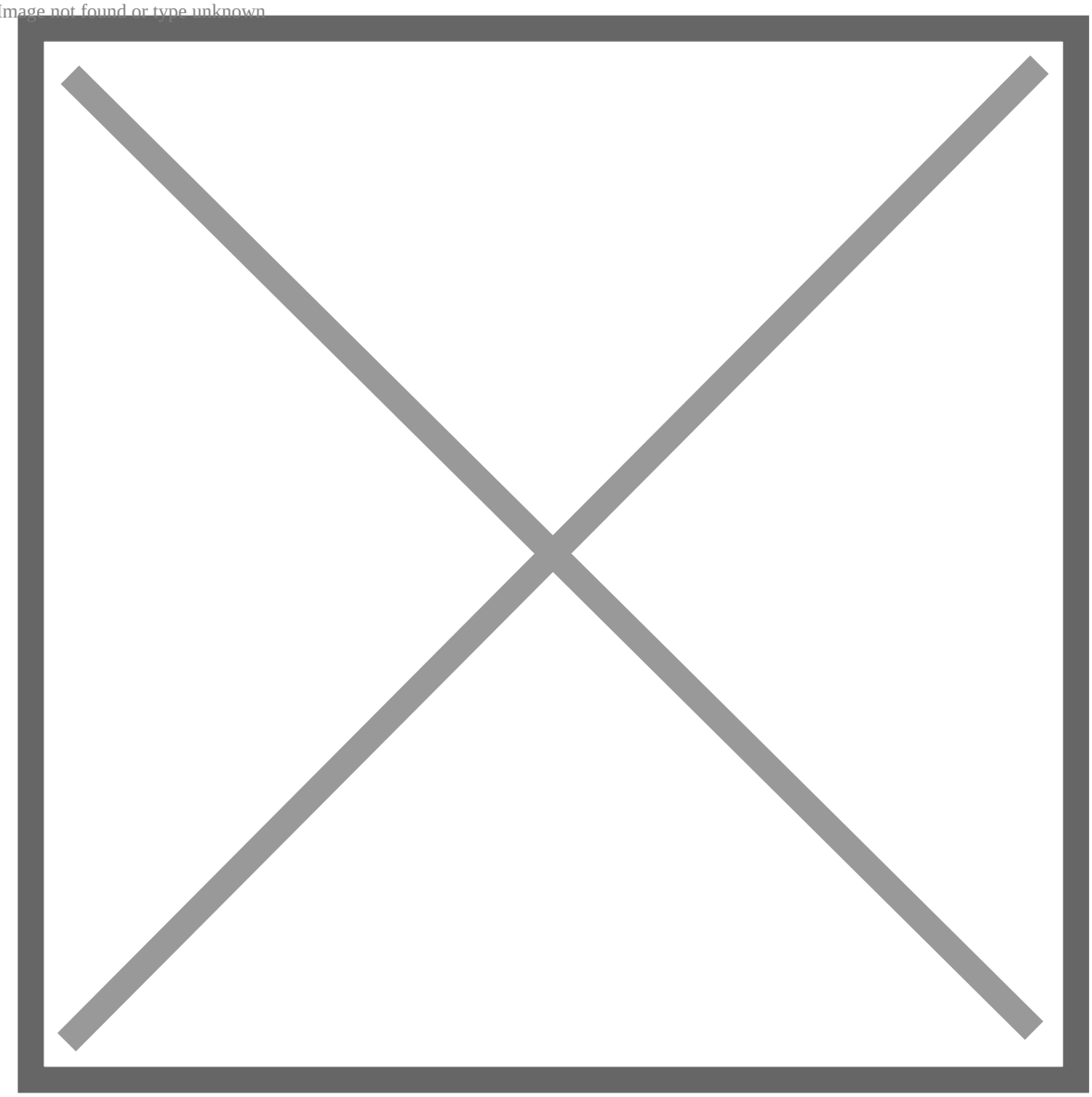

Fig 3. *Europeana Newspapers'* full screen viewer.

To the left of the main viewing window is a space for viewing the digitised text of the article. This provides another way to control the newspaper viewer; clicking on a particular paragraph immediately zooms the viewer into that particular section of the page. As is often the case with Optical Character Recognition, the accuracy of this data varies considerably and is determined by the quality and formatting of both the original newspaper and the scanning process. Other projects, such as Australia's *Trove*, allow users to manually correct these errors and gradually improve the quality of the database. The *Europeana* project aims to introduce a similar feature in the future, but this is not available in the current prototype. For now, however, displaying this text serves another interesting purpose. Google's Chrome web browser has the ability to automatically translate text from one language to another. I was pleasantly surprised to discover that this tool works with the text displayed by Europeana's interface, which means that it is possible to automatically translate articles into other languages. Anyone who has used online translators will attest to the fact that the results are very crude, but even a very rough translation is often enough to get the gist of an article and determine whether it is worth translating more accurately. For example, I wanted to track reports of a Victorian Englishwoman who was found dead in a hotel room in Paris. Unfortunately, I only speak English and a little bit of half-remembered GCSE French, which made it difficult to find and consult reports that were published in Paris itself. However, by searching for the woman's name in Europeana and putting the French results through Google's translation tool, I was able to glean some interesting new details about her death. For scholars like me who do not possess the linguistic skills to independently conduct pan-European research, this kind of automatic translation (however rudimentary it might be) is tremendously useful and could encourage more scholars to engage with sources in unfamiliar languages.

In sum, it still feels a little too early to pass judgement on the Europeana Newspapers project. They have certainly achieved some impressive results. Combining and refining millions of pages of content from so many different European libraries is a remarkable achievement. Meeting the demands of these institutions has required the team to make some compromises, which in turn make the experience of browsing the archive rather uneven. Nevertheless, integrating all of these archives into a single search interface means that Europeana must now be regarded as the first port of call for anybody who wishes to explore the continent's newspapers. Of course, the archive is by no means exhaustive. Only a fraction of the historic newspaper archive has been digitised, and many of those that have been scanned will remain trapped behind paywalls for years to come. English newspapers, for example, are conspicuously absent from the Europeana portal, despite the fact that they have been digitised in great numbers. The British Library is listed as a one of the project's full partners, but their digitised newspapers are currently tied up by long-standing agreements with commercial publishers. Unfortunately, there is also work to be done on Europeana's search interface. At present, the prototype's search options don't offer the precision or flexibility required to work with such a large and complex dataset. Finally, I have also encountered a few technical problems when using the archive over the last year; it seems to be working well at the moment but, at times, loading speeds were slow and some features were unresponsive. If you encounter similar difficulties when exploring the archive, it might be necessary to return a few days later. But these need not be lasting problems. The *Europeana Newspapers* team have made an impressive start on building what promises to become an exciting platform for newspaper research. What's more, they have an admirable commitment to open access and are working to support researchers who want to make creative re-use of newspaper datasets. Indeed, it's a shame that not all digitisation projects are informed by a similarly open ethos. Once the dataset is fully integrated into Europeana and the search interface is (hopefully) enhanced, we will have a better sense of how the archive might be used. For now, it is still worth playing with the prototype – despite some of its limitations, it is still a fantastic new tool for exploring the past.

Links

[1] https://reviews.history.ac.uk/item/121174
[2] http://www.europeana-newspapers.eu/
[3] http://www.theeuropeanlibrary.org/tel4/newspapers
[4] http://www.europeana.eu/portal/
[5] http://mhbeals.com/multimodal/
[6] http://viraltexts.org/
[7] http://www.britishnewspaperarchive.co.uk/
[8] http://trove.nla.gov.au/
[9] http://chroniclingamerica.loc.gov/
[10] http://research.europeana.eu/itemtype/newspapers