IFLA WLIC 2013
SINGAPORE

Submitted on: July 31, 2013

The challenge of making digitised European newspaper content available online

Susan Reilly
Project Manager
LIBER
Den Haag, NL
Susan.reilly@kb.nl

Abstract:

Europeana Newspapers is a three year project with the aim of bringing Europeana Newspaper into Europeana, the main portal for digital cultural heritage in Europe.
The project is working to bring newspaper content into Europeana in several ways. As well as the content which will be delivered by the project partners, additional content will be brought in from 'associated partners'. These partners have been identified through a survey of 450 LIBER members, which was designed to provide an insight into the current state of digitised newspaper collections in Europe. What the survey revealed is that, though a large majority of libraries make their digitised content available for free, the extent of this availability is limited. Of the collections which had been digitised, 36% had received no OCR whatsoever. Of the remaining 64% that had OCR, only 36% had been made available in a way which exposed the full text to be searched or viewed by an end user. 13% of these were available in portals which allowed for faceted browsing or named entity extraction. The other challenge in terms of availability is the fear of breaching copyright. The most recent available content is over 70 years old, and for many libraries 140 years was the cut off point after which content was made fully available.
For Europeana Newspapers to succeed in its aim of making European newspaper content available in Europeana it seems that it must address several key challenges including:

1. Convincing libraries to provide their collections to Europeana
2. Helping to increase the accessibility and availability of content by sharing best practice in newspaper digitisation methods
3. Providing a content browser to help expose newspaper content more effectively
4. Opening up a dialogue between libraries and rights holder in order to lower copyright barriers to making digitised newspapers available

Introduction

LIBER, the Association of European Research Libraries, is leading the dissemination activity for Europeana Newspapers[i]. Leading such an activity is a huge investment of time and effort for our organisation, an investment we would only make if there is clear value in such an activity for our organisation, our members, and for researchers.

For LIBER, Europeana Newspapers is one of the mechanisms for achieving our mission to provide an information infrastructure to enable research in LIBER institutions to be world class[ii], in particular within the context of reshaping the research library. The research library is being reshaped by the constantly shifting digital information environment as well as the changing way in which research is performed in such an environment. Research libraries are in the midst of the data deluge, digitising collections, collection born digital objects, acting as repositories for electronic articles and, most recently, research data. Whether digitised material or raw research data, all of this content are data and it is up to libraries to ensure that this data is accessible and made available in a meaningful way. This is especially true for digitised newspaper collections, where to truly realise the value of newspapers and their unique attributes, further refinement must be carried out. Making digitised content available in this way can allow the researcher to place a research question in a wider context by exploring high quality images of the content in question and also by linking to related content[iii].

Through the Europeana Newspaper project we have been able to develop and share best practice in the digitisation of newspaper content. In particular, the project focuses on:

- The use of refinement methods for OCR, OLR/article segmentation, and named entity recognition (NER), and page
- class recognition to enhance search and presentation functionalities for Europeana customers
- quality evaluation for automatic refinement technologies
- transformation of local metadata to the Europeana Data Model (EDM)
- metadata standardization

Over the course of 3 years the project will bring over 13 million pages of newspaper content to Europe's cultural heritage portal, Europeana. There will be a special focus on newspapers published during the First World War, thus providing a meaningful addition to the resources aggregated by the Europeana 1914-1918 project[iv]. Newspapers provided within the project will also cover a wider time period, so that numerous historical events from before and after 1914-1918 linking directly to events within that time period will be accessible via the Europeana portal. This will allow people to easily research important local, national, or European events in a broad European context; something that has so far not been possible. As of June 2013, the project has almost reached its halfway point. Thus far, datasets from the 18 libraries providing newspaper content to the project have been selected for refinement. This has been done according to criteria such as availability (e.g. licence attribute of the metadata), quality of digitisation, document characteristics such as font and language, and technical considerations such as format and metadata type[v]. The technical requirements for processes such as Optical Character Recognition (OCR), Optical Layout Recognition (OLR) and Named Entities Recognition (NER) have also been defined to ensure that the content is fully integrated into the European ecosystem.

As this project is a best practice network it places a huge emphasis on engagement and the sharing of best practice outside of the network. Also, by establishing an overview of the state of Europe's digitised newspapers collections more generally, it provides an insight into the issues preventing these collections from becoming more accessible and available.
Making newspapers easy to search and presenting them attractively online is currently a challenge.The project has also set out to develop a content browser, which will be specially designed to take advantage of the nature of the newspaper content and the refinement processes that have been applied to it.

Exploring the state of Europe's digitised newspaper collections

One of the project's aims was to identify other digital newspaper collections in Europe. This was achieved through a survey of LIBER member libraries. The survey focused on newspaper title and time range, metadata in use, data distribution, capabilities, and quality of digitization including technology used for refinement. As well as providing an overview of other collections which could benefit from the best practice techniques developed through Europeana libraries, the survey uncovered some potential issues and gaps to be addressed in terms of making more of Europe's newspaper heritage available and accessible via Europeana.
The survey of LIBER libraries was conducted in Autumn 2012. It went out to 420 LIBER libraries and received 47 responses. The rate of response is not huge, but it may be an indication of the number of libraries in Europe that are actually digitising newspaper collections as the survey was aimed at only libraries that are engaged in newspaper digitisation. The survey description also included a definition of newspapers which is widely used by libraries such as the British Library, Library of Congress and the Australian National Library:
"a serial publication which contains news on current events of special or general interest. The individual parts are listed chronologically or numerically and appear usually at least once a week"
The respondents were asked 12 questions on topics such as the size, availability and accessibility of their digital newspaper collections. The response to these questions painted the following picture.

Less than 10% of newspaper collections in European libraries have been digitised.

Although the number of pages and newspaper titles is not insignificant, on average less than 10% of the respondents newspapers collections were actually digitised. There may be various reasons for this, some of which are mentioned below.
In terms of actual figures, libraries managed to identify nearly 130m pages of digitised content comprising of nearly 24,000 titles (129, 041, 663 of pages and 23,987 titles were the precise figures obtained). Sic of the respondent libraries could not provide a number of titles because of the large size of their collection and the cursory nature of their cataloguing. Seven other libraries could only give an estimate. This means that the actual figure for number of pages digitised could, in reality be much larger.
When the number of pages digitised were compared to the size of newspaper holding it was found that only 26% had digitised more than 10% of their collections. Only two of those had done more than 50% - the consortium of libraries represented by the Biblioteca Virtual de Prensa Histórica (58% of their pages were digitised) and the National Library of Turkey, unique for having digitised its entire collection of 800,000 pages and 845 titles. Large scale newspaper digitisation projects were most common.

Given that the Enumerate survey of digitised collections in cultural heritage institutions in Europe found that over 20% of all cultural heritage collections were digitised[vi]this comparatively lower figure of 10% may be indicative of the particular challenges, both in format and policy, that the digitisation of newspapers presents.

Access to digitised newspapers is nearly always free of charge.
One positive results of the survey was the discovery that access to the digitised newspaper collections of the respondent libraries was nearly always free of charge.85% of respondents provided free access to their digitised newspaper collections. Only one library used a pay per view model. Three libraries made their content available via a subscription service and four licensed their content to other institutions. Free access was not always universally available, some libraries only provided free access at national level or within their institution.

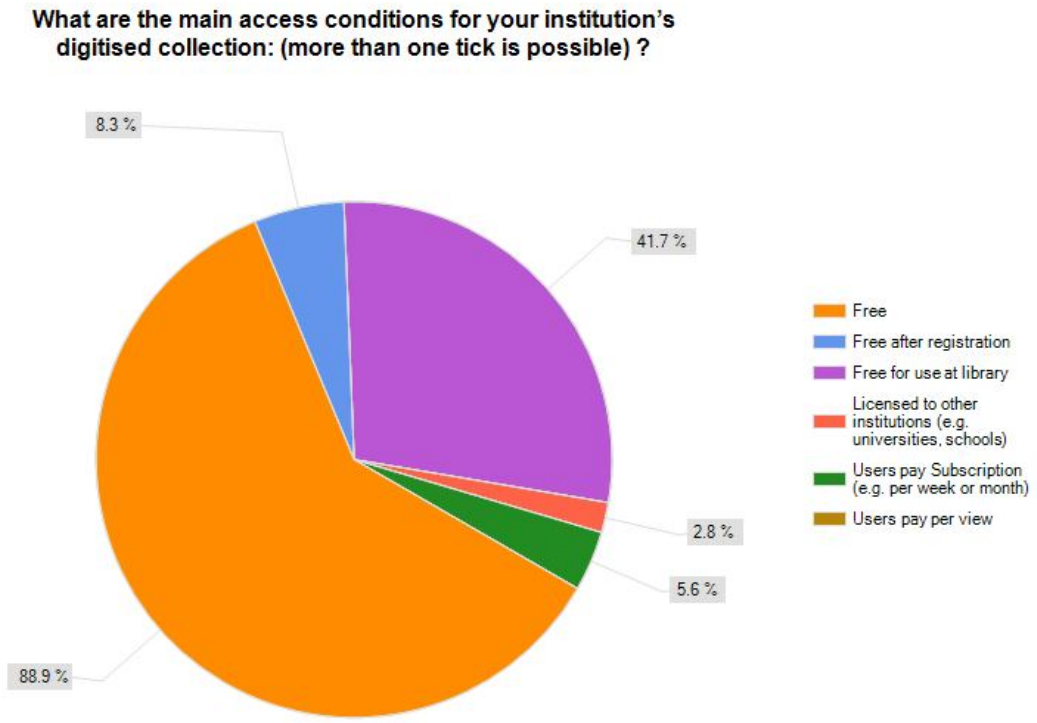

Figure 1. Access conditions

Access to twentieth-century content remains problematic.
Digitising 20th century content is problematic, primarily for copyright reasons. Newspapers are particularly problematic as each individual article or news story in a newspaper is a separate copyright work. In addition, the contents of the newspaper as a whole enjoy copyright.[vii] This means that negotiating the right to digitised in-copyright newspaper content can be far more complex than for monographs. The survey also revealed that some libraries have a particularly conservative approach when it comes to interpreting copyright, with one library having a cut-off date of 1863, from after which they would not make content available.

Over half of the libraries (27 out of 47, 57%) have a cut off date beyond which they will not publish digitised newspapers on the web. Most frequently, this is based on a 70 year sliding scale,
meaning that content after 1942 is inaccessible in digital form. 23% of respondents had an arrangement with a specific range of publishers so that in-copyright digitised newspapers could be published. Several respondents reported difficulties with negotiating rights agreements with individual publishers, indicating that collective rights agreements would be even more difficult and complex.

The richness of our digitised newspaper content remains to be exploited
There is huge potential for improvement when it comes to enhancement of the newspaper content. 36% of respondent applied no form of OCR what so ever, meaning that the full text of the digitised newspaper content could not be searched. Half of those respondents who did emply OCR were not confident enough in the results to expose them via a search interface. 36% had employed zoning and segmentation and only 6% used named entity recognition. There was a huge amount of variance in metadata work. Many libraries used Dublin Core only, whilst other had developed their own standards.

Future work and areas for action
Identifying other digitised newspaper collections in Europe has enabled the project to extend it's network and bring in new content. Following the survey analysis a further eleven libraries were invited to join the project as networking partners. These were:

- National Library of Wales
- St. Cyril and Methodius National Library (The National Library of Bulgaria)
- National Library of Czech Republic
- National and University Library, Ljubljana, Slovenia
- Lucian Blaga Central University Library, Cluj-Napoca, Romania
- National and University Library of Iceland
- National Library of Luxembourg
- National and University Library in Zagreb, Croatia
- National Library of Belgium
- National Library of Portugal
- Spanish National Library

These libraries will provide additional newspaper content to Europeana and will benefit from attendance at the best practice workshops run by the project on topics such as refinement and aggregation, as well as on policy issues. By extending the network in this way it is hoped that the best practice developed during the project, e.g. in areas such metadata and refinement, will have not only a broader reach but will also have a deeper and long lasting impact on raising the standard of digitised newspaper collections in Europe.
The project is also working to promote the value of making digitised newspaper collections available and accessible online by hosting national information days in partner countries. This means that the value of the project can be promoted to those who will ultimately fund future digitisation projects, decision makers at national level and to potential user communities.
Engagement with communities beyond Europe is occurring through two international competitions (hosted by the University of Salford, a technical project partner) – one on the Layout Analysis of Historical Newspapers and one on the Recognition of Historical Books with Distortions.

The work of the project so far has highlighted that there is much work to be done in order to bring all of Europe's newspaper heritage online. There is a clear need for the development of best practices and standards. It is also important that we have the right tools in place in order to ensure that we can properly expose the richness of enhanced digitised newspaper content. There is also much work to be done on the policy and engagement front, not only in order to encourage the uptake of best practice and secure funding for digitisation, but also in order to negotiate rights issues and improve clarity around these issues. The first step should be to start a discussion with newspaper publishers about the potential benefits of, and conditions under which, publishers and libraries can works together to make more 20th century newspaper content available online.

---

[i] http://www.europeana-newspapers.eu/

[ii] http://www.libereurope.eu/strategy

[iii] COLLINS, E., JUBB, M.. How do Researchers in the Humanities Use Information Resources? LIBER Quarterly, North America, 21, jan. 2012. Available at: <http://liber.library.uu.nl/index.php/lq/article/view/URN%3ANBN%3ANL%3AUI%3A10-1-113621/8364>. Date accessed: 12 Jun. 2013.

[iv] http://www.europeana1914-1918.eu/en

[v] Neudecker et al.(2013) Europeana Newspapers Report on datasets for refinement, http://www.europeana-newspapers.eu/wp-content/uploads/2012/04/D-2-1_Dataset_for_refinement.pdf, Access June 12th 2013.

[vi] http://www.enumerate.eu/en/statistics/

[vii] Oppenheim, C. (2003). Newspaper copyright developments: A European union and United Kingdom perspective. *IFLA journal*, *29*(4), 317-320.