DELIVERABLE

Project Acronym: Europeana Newspapers

Grant Agreement number: 297380

Project Title: A Gateway to European Newspapers Online

Deliverable 4.1: European Newspaper Survey Report

Revision: 1.0

Authors: **Alastair Dunning, The European Library / Europeana Foundation (alastair.dunning@kb.nl)**

Contributions: **Reviewed by Clemens Neudecker, Koninklijke Bibliotheek**

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	x
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Revision	Date	Author	Organisation	Description
1.0	Nov 2012	Alastair Dunning	TEL	

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of Contents

Executive Summary 4
Full Results 5
Survey Methodology 5
Definition of Newspapers 5
Extent of Digitisation 6
Percentage digitised against physical newspaper collections 7
Copyright 7
Enhancing Digitised Newspapers 9
Functionality 10
Access Conditions 11
Technical Standards 12
Common File Formats Used 12
Common Metadata Standards Used 13
Appendix A – Survey Questions 14
Appendix B – Responding Institutions 17

Survey of Library Newspaper Digitisation Projects in Europe (Version 1)

Alastair Dunning, Programme Manager, The European Library
(alastair.dunning@kb.nl)

November 2012

As part of the three year project Europeana Newspapers (http://www.europeana-newspapers.eu/), the European Library undertook a survey of the extent of newspaper digitisation within public, research and national libraries in Europe (ie excluding digitisation projects from collections held outside libraries). A series of 12 questions were circulated to libraries in the summer and autumn of 2012, and the results of the survey are below.

Executive Summary

Access to digitised newspapers is nearly always free of charge.

Of the 47 respondents, at least 40 (85%) offered free access to their digitised newspapers. Very few made use of the opportunity to charge users. One library had pay per view, whilst another three offered subscription services for users (ie paid access per day or per month). Only four libraries licensed their newspaper contents to other groups (e.g. school, universities).

Access to twentieth-century content remains problematic.

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% (11 out of 47) had an agreement with a rights organisation so that in-copyright digitised newspapers could be published. However, this tended to be restricted to individual titles rather than collective agreements for complete collections.

There is still much to be done to exploit the richness of digitised newspaper content

In scanning their newspapers, 36% (17 out of 47) of libraries have not used any form of Optical Character Recognition (OCR), meaning that searching through the full text of newspaper content is not possible. And while 64% have used OCR, only 17 of the libraries (36%) exposed the resulting full text to the viewer, indicating that they had reservations about the quality of the OCRd text. There were also low numbers (36%) for those that had undertaken zoning and segmentation and only six libraries (13%) had included features such as facetted browsing or extracting entities such as place or name.

Full Results

Survey Methodology

The survey was sent out in two parts. The first part was aimed at any European library, whether inside or outside the European Union, that had undertaken a newspaper digitisation project. It did not include private companies that have undertaken such work. The request for responses was circulated via the mailing lists of the The European Library, LIBER and various social media in May 2012 and the survey finally closed at the August 2012. 38 responses were received. Two libraries (The National Library of the Republic of Moldova and the Bodleian Libraries, University of Oxford) had newspapers but had not digitised any of them. A third respondent (Oslo and Akershus University College) licenced newspapers on behalf of its users but did not actually undertake digitisation. These three were discounted from the final survey results.

The second part of the survey went to the partners in the Europeana Newspapers project in October 2012. All 12 partners with newspaper content responded. In both cases, respondents were asked about the topics above, and invited to add extra comments where relevant.

Thus the survey is based on the feedback of 47 organisations that have been involved in newspaper digitisation.

The original questions are included in Appendix A. The respondents are listed in Appendix B.

The publication of these results may well uncover some other digitised collections. An updated version will be produced if there is sufficient additional data.

SBB will integrate the survey data in the ZDB in as far as possible, set up a dedicated website to present this information to the public, and will be responsible for keeping the data current over time. It is foreseen to extend the information offer with data from commercial content providers in future.

Definition of Newspapers

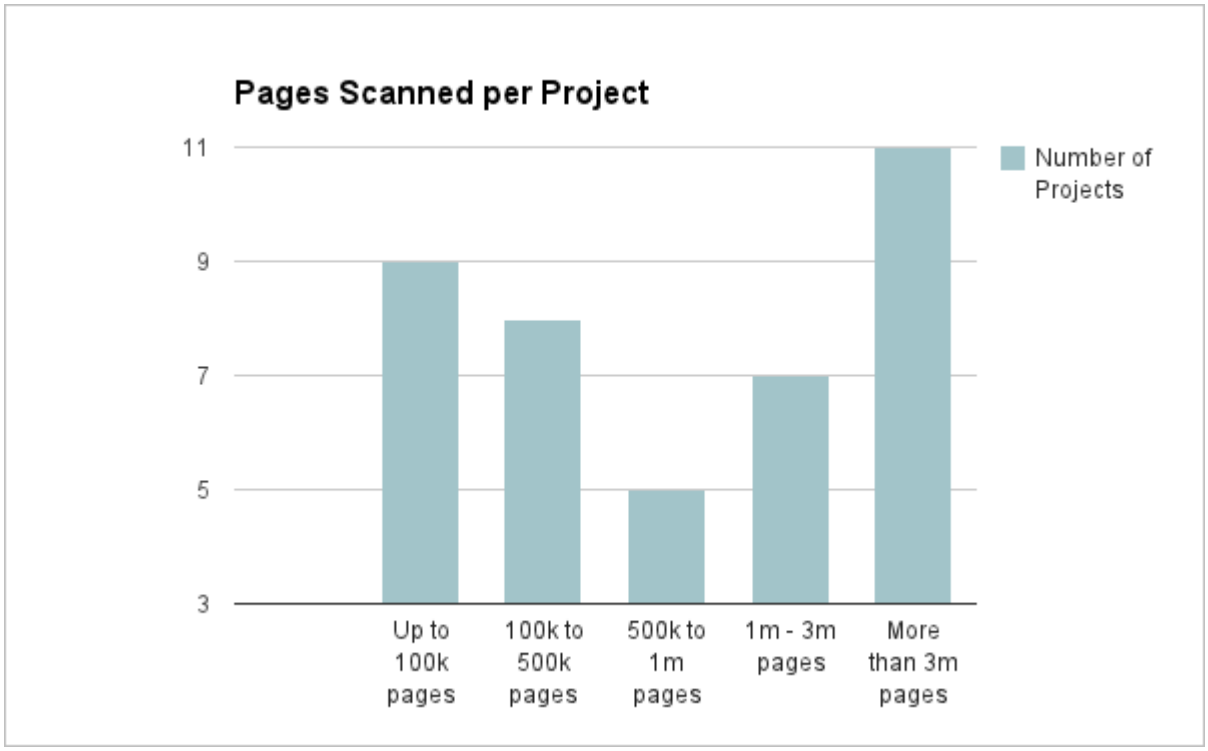
When a library requested clarification about the definition of a newspaper (as opposed to a journal or periodical) the definition supplied by the British Library was used: “a serial publication which

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"[1]

Extent of Digitisation

There can be no denying the extent of newspaper digitisation undertaken in Europe. Libraries managed to identify nearly 130m pages of digitised content comprising nearly 24,000 titles (129, 041, 663 of pages and 23,987 titles were the precise figures obtained). The number is likely to be much higher in reality. Because of the vast size of their collections and the cursory nature of their cataloguing, there were six libraries unable to give a number of titles, and seven who could not give a definite number of pages.

In terms of sizes of the digitisation projects, it was the large-scale projects that were actually the most common. 11 of the respondents had digitised over 3m pages.



[1] Email from Aly Conteh, Digitisation Programme Manager, British Library

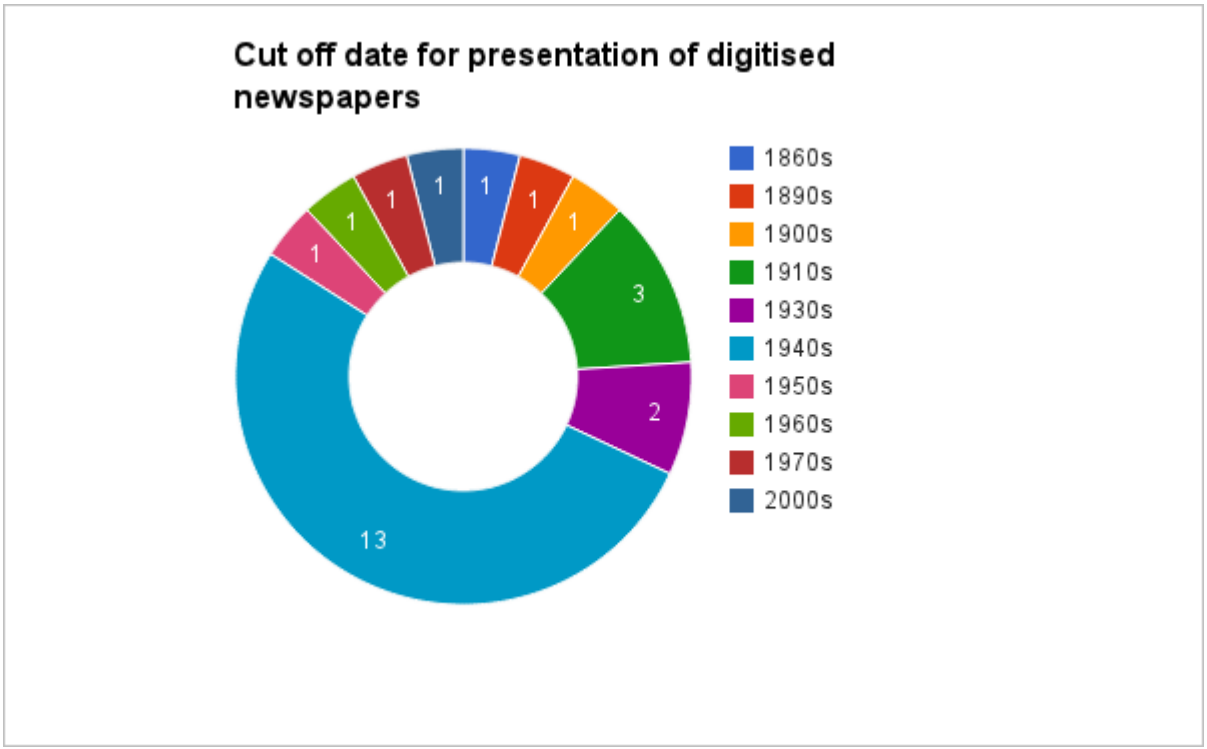
Percentage digitised against physical newspaper collections

The number of pages digitised is impressive. Yet where it was possible to compare the number of titles or pages digitised against the actual size of the physical collections, the ongoing challenge of creating an entirely digital library of newspaper holdings was reinforced. Only 12 (26%) of the libraries had digitised more than 10% of their collection (either in terms of titles or page numbers), and 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.

Copyright

As with access to all digitised cultural heritage, access to recent twentieth-century content is deeply problematic. 27 of the libraries had specific dates beyond which they would not present digitised newspapers online. Of these 27, eleven had dates in the 1940s, with 1942 (ie a sliding scale of 70 years ago) being the most common.

Some smaller libraries that had specific collections tended to have a later date. Larger libraries with more extensive collections (which presumably include newspaper titles of considerable prestige) often had earlier dates for some titles. For example, the National Library of Portugal currently applies a date of 1860 for some titles and 1940 for others.

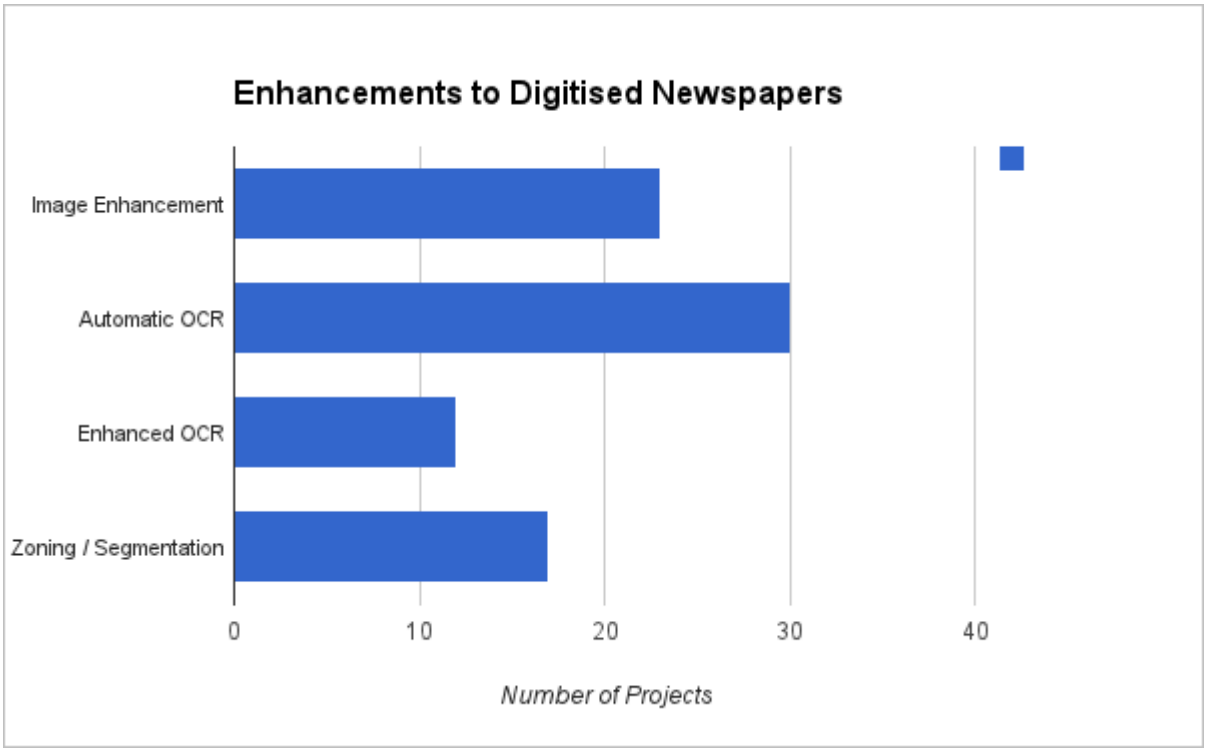


The cut off date was also influenced by where the user was accessing the newspapers. For example, the newspapers of the National Czech Library are “free and public[ly] available generally [until] 1890 (with some exceptions to 1939); other documents [are] available only in the library premises”.

Respondents were also asked if they had any collective agreement with a rights organisation so that in-copyright digitised newspapers could be published. Eleven out of 47 libraries gave partial affirmation to this. However, such agreements tended not to be collective but specific to a limited range of publishers. For example, the Bibliothèque nationale de Luxembourg had agreements with individual publishers, although agreements with collective rights management organisations were being planned. Likewise, the Biblioteca Virtual de Prensa Histórica noted that “Agreements have been made with some current newspapers that started to be published at the end of XIXth century or at the beginning of XXth century and with publishers of current cultural magazines as well.” However, many of the other libraries remarked on the difficulties of tracing owners and negotiating with publishers of newspapers. Even putting an individual rights agreement with a single publisher into place, let alone a collective agreement, remains a difficult task.

Enhancing Digitised Newspapers

While there has been considerable work done in digitising newspapers, not all libraries have been able to accompany this with enhancement of the digital images of they have created.



The statistics suggest that while enhancement of the scanned images is not a standard part of all projects, it is reasonably common. For example, 23 out of 47 (49%) of libraries had undertaken basic image enhancement such as cropping or correcting skewing or other problems caused by the scanning process. Seventeen projects had done work on zoning their newspaper pages into separate articles. Eleven libraries indicated they had undertaken enhanced OCR, improving the quality of the text once the initial OCR process has taken place.

However, the accompanying comments demonstrated that extra work undertaken on the images is often more on an experimental or reduced level.

For example, most libraries indicated that work done on improving OCR quality is selective. The National Library of the Czech Republic or the Swiss National Library have concentrated only on improving a restricted number of titles. Most other libraries that have improved ‘textual quality’ by manually improving the actual fulltexts of newspapers and other descriptive metadata (such as image captions).

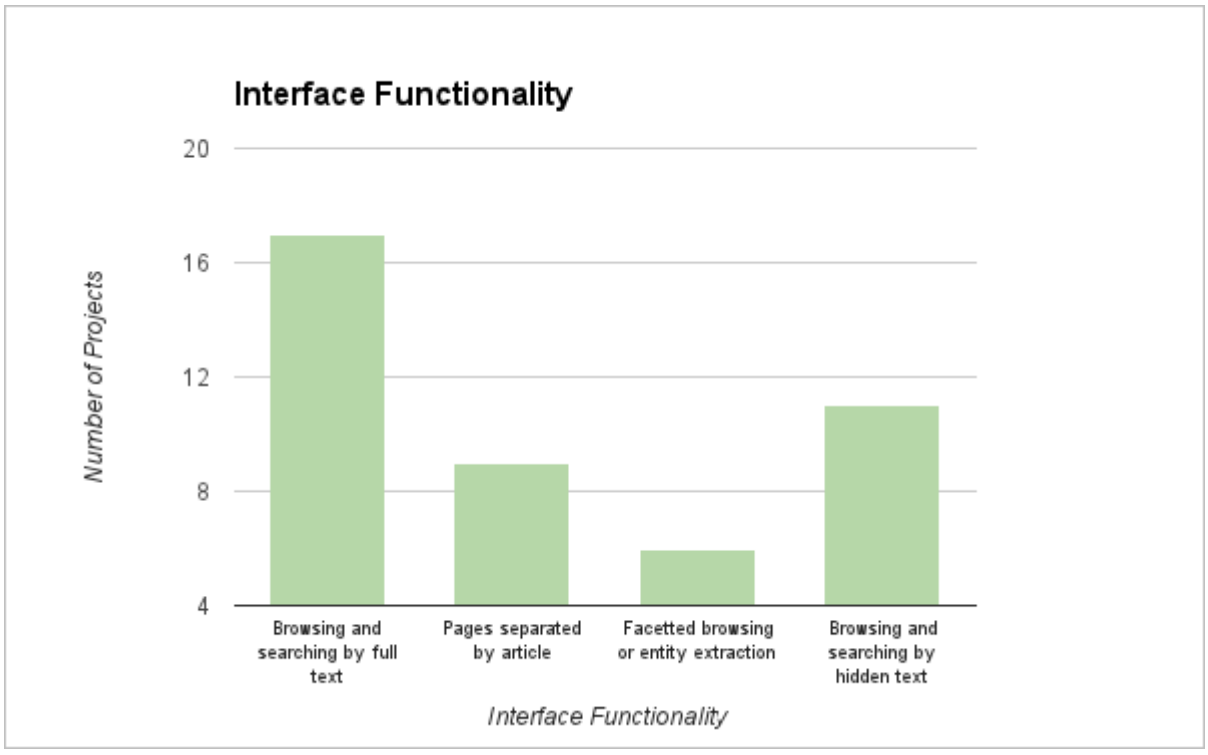
The survey did not question the success of automated OCR (ie the accuracy of the software in recognising the words / characters of the original document).

The Europeana Newspapers project is beginning to address this problem, undertaking all four of the enrichments listed in the chart above to improve the quality of digitised newspapers. But the evidence indicates how much further work there is to be done so that the advantages of having historic newspapers in digital form can be fully exploited.

Functionality

The extent to which OCRing takes place has a knock-on effect on the development of the interface for access to digitised newspapers. OCRing the text allows the user to undertake full text searching - although the success of this is obviously dependent on the quality of the original OCR.

The quantitative statistics indicate that some libraries offer more than simply browsing through images.



However, the additional comments again pointed out such features were rarely fully incorporated into an online service - rather, they applied to a limited part of the digitised corpus. As the National Library of Estonia pointed out, they could only allow full text searching where they had undertaken OCR. For the Bibliothèque nationale de France, the decision to offer full text searching to end users was dependent on the quality of the OCR process.

Access Conditions

One overwhelming conclusion from the survey was that digitised newspapers were being made freely available for online viewing. 40 out of 47 libraries made their content available for users to view immediately on the web.

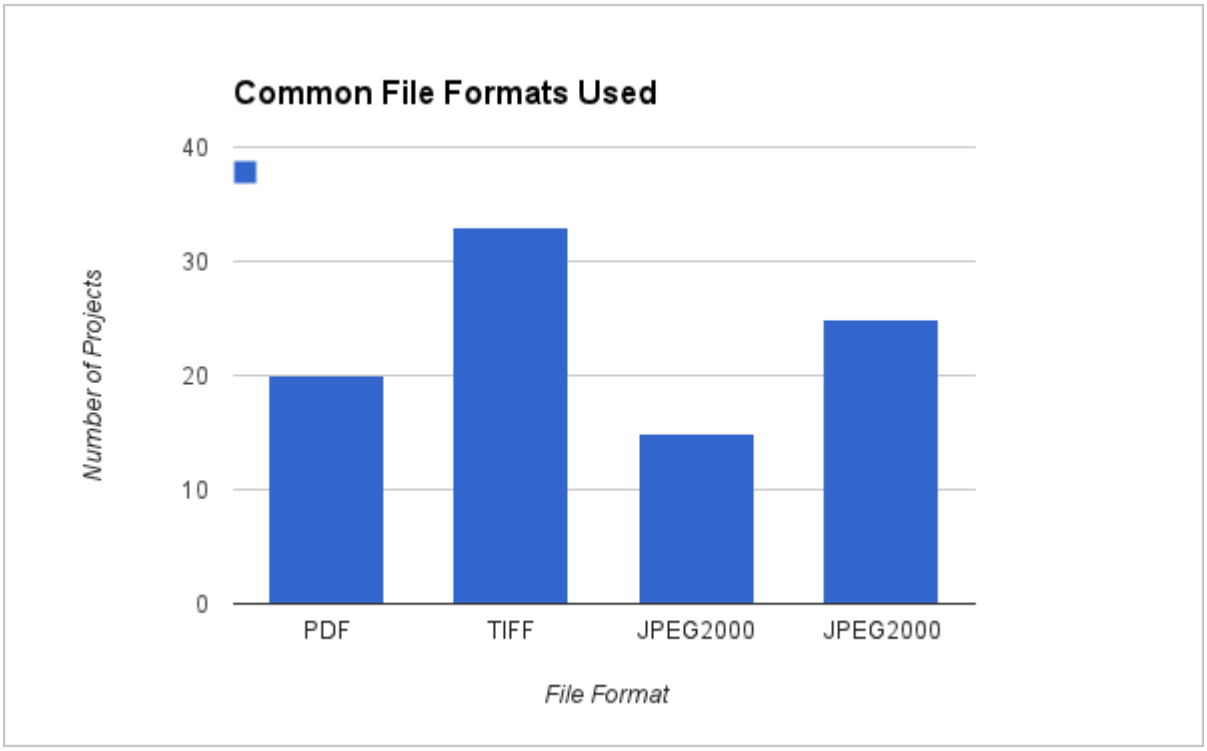
Very few libraries had explored alternative business models.

The three libraries to charge for access are the National Library of Turkey, the British Library and the Mediacenter Sarajevo, which deals with contemporary as well as historic newspapers.

However, a slightly higher number had explored the possibility of licensing their content to specific audiences - the national libraries of Britain, Latvia and Finland and the Portuguese Biblioteca Municipal de Faro. The British Library had made the greatest explorations of the possibilities here, licensing their content to different audiences (such as universities and public libraries), as well as building a large-scale public-private partnership with a publisher to digitise large amounts of their newspapers collections.

Technical Standards

Common File Formats Used



The TIFF file format was the most commonly used, followed by the JPEG format.

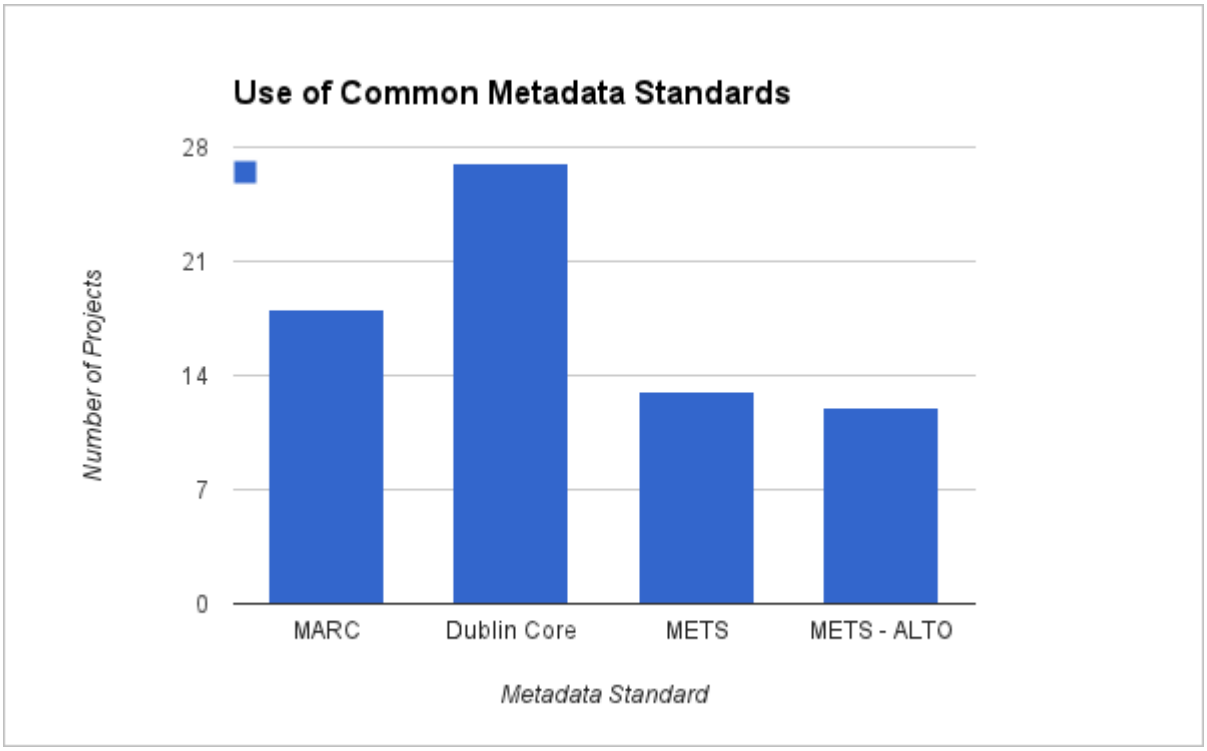
Most of the uses of the PDF format were in tandem with TIFF and JPEG - indeed many respondents added comments to indicate that they used TIFF as the master format and PDF and or JPEG for delivering the images.

The Biblioteca Municipal José Régio in Portugal, The National Library of Norway, The British Library and the National Library of the Netherlands were the four respondents who used JPEG2000 without using TIFF.

There were four instances of libraries using PDF as the sole format for newspaper digitisation.

Other file formats being used (normally by a handful of libraries only) included PNG, TXT files for the full text; ePUB for cultural magazines; and the DjVu format for document delivery.

Common Metadata Standards Used



In line with the other findings within the survey, many libraries have taken the first step in newspaper digitisation projects but not had the opportunity to fully exploit the potential. Related to this, it is not a surprise to find the basic Dublin Core standard being the most common standard used. Indeed, even some large scale projects used Dublin Core only - for example five of the partners in the Europeana Newspapers project itself use Dublin Core as the only established metadata standard.

However, it should also be noted that this was the area of work with the greatest variance. At least 7 libraries developed their own standards; the same number also used variants of METS.

MODS also occasionally appeared as a means of structuring descriptive data. Other standards highlighted included PREMIS, OAI-ORE (RDF), local variations of MARC, TEI, NISOIMG and ISBD (G).

Appendix A – Survey Questions

The Europeana Newspapers (http://www.europeana-newspapers.eu/) project has now finished its survey of projects outside the project and is now getting some final details on the partners within the project

The information below will be used as part of the report on the project and will also contribute to published articles on the various newspaper collections involved in the project.

We need results by October 15 so the report can be finalised by the end of the month

Many thanks !

Alastair Dunning and the Europeana Newspapers team
alastair.dunning@kb.nl

1. What is your institution ?

2. Approximately, how many pages and titles are there in your institution's physical collection of newspapers?

3. Approximately, how many pages and titles has your institution digitised?

4. What is the homepage URL for your digitised collection?

5. Copyright: Does your institution have a cut off date beyond which it will not publish digitised newspapers on the web?

Yes / No ?

And if yes, what is this date?

6. And does your institution have any collective agreement with a rights organisation so that in-copyright digitised newspapers can be published ?

Yes / No

Any comments?

7. What aspects of digitisation have been undertaken on your collection?

Simple capture of images

Image Enhancement

Automated OCR

Enhanced OCR (ie further improvement of the OCRd text)

Zoning and Segmentation

Any comments?

8. How are the your institution's digitised newspapers made available?

Images only via browsing

Images only via browsing and searching by hidden text

Images and full text via browsing and searching

Images and full texts separated by article

Images and full texts separated by article; other features such facetted browsing and searching over entities as well as keywords

Any comments?

9. What are the main access conditions for your institution's digitised collection: (more than one tick is possible) ?

Free

Free after registration

Free for use at library

Licensed to other institutions (e.g. universities, schools)

Users pay Subscription (e.g. per week or month)

Users pay per view

Any comments?

10. What standards for digital images do your institution's digitised newspapers use ?

PDF
TIFF
JPEG2000
JPEG

Do you use any others ?

11. What metadata standards do your institution's digitised newspapers use ?

MARC
Dublin Core
METS
METS-ALTO
TEI

Do you use any others ?

12. Any further information or comments that you wish to add ?

Appendix B – Responding Institutions

With URLs for newspaper collections and comments as given by respondents

Erfgoedcel CO7
(Intermediate partner for the online presentation of digitised newspapers of the municipal archives of Ieper and Poperinge, Flanders region, Belgium.)
http://www.historischekranten.be

National and University Library "St. Kliment Ohridski"-Skopje, R. Macedonia
Digital Library website: http://www.dlib.mk
Rare Periodicals: http://www.dlib.mk:8080/jspui/handle/68275/48

Mediacenter Sarajevo, digital archive INFOBIRO
http://www.infobiro.ba

National Library of Romania
http://www.bibnat.ro/

National Library of Wales
Digitised newspapers will be available online early in 2013 from
http://www.llgc.org.uk

National Parliamentary Library of Georgia
http://dspace.nplg.gov.ge/handle/1234/94

St. St. Cyril and Methodius National Library (The National Library of Bulgaria)
The digitalised Bulgarian newspapers are divided in three separate collections, presented in the library's website in Bulgarian and English versions: In the English version: Serials from 1844 to 1878 The collection Bulgarian Periodicals from the Revival (1806-1878) includes almost all the periodicals, published during the Bulgarian Revival.
http://193.200.14.178/DWWebClient/Integration.aspx?i=General&lc=VXNlcj1mcmVlXG5Qd2Q9ZnJlZQ&p=SRLV&sed=62c15340-c077-4f7a-9252-bb142f5db2f9&culture=en

Serials from 1878 to 1944
The collection of Bulgarian serials contains newspapers and journals, published in Bulgaria from 1878 to 1944. The selected titles are related to the historical trace in the social, political and cultural development of Bulgaria.
http://193.200.14.178/DWWebClient/Integration.aspx?i=General&lc=VXNlcj1mcmVlXG5Qd2Q9ZnJlZQ&p=SRLV&sed=fe0d8b5f-4a23-445a-b58c-155196d404e3&culture=en

United Bulgaria
In this collection are gathered the serials, published from 1879 to 1940 in Bulgaria and abroad. These serials present several generations of Bulgarians in connection with the historical, ethnical, religious and cultural rights of political union of Trakia, Macedonia, Moesia and Dobruja. This subject was from the leading ones in the periodicals, published during the decades between the decisions from the Congress of Berlin and the sign of the Treaty of Craiova.
http://193.200.14.178/DWWebClient/Integration.aspx?i=General&lc=VXNlcj1mcmVlXG5Qd2Q9ZnJlZQ&p=SRLV&sed=e497bb06-23b5-42ab-9d21-89cbb53f90b5&culture=en

Press and Information Office Ministry of the Interior, Republic of Cyprus
Due to complicate copyright law, the digitised collection is not on line
http://www.moi.gov.cy/moi/pio/pio.nsf/index_en/index_en?opendocument

National Library of Czech Republic
New version of digital library Kramerius system with new data is now available only in-house
http://kramerius.nkp.cz/kramerius/Welcome.do

Nicolas Copernicus University
http://kpbc.umk.pl/dlibra

National Library of Scotland
(Only sample newspapers digitised)
http://www.nls.uk/

National and University Library in Zagreb
http://dnc.nsk.hr/newspapers/

Institut für Ost- und Südosteuropaforschung
http://www.difmoe.eu/
http://www.ios-regensburg.de/bibliothek/digitale-bibliothek.html
http://www.ios-regensburg.de/bibliothek/projekte/deutschsprachige-periodika.html

Koninklijke Bibliotheek van België - Bibliothèque royale de Belgique (KBR)
For now, digitised newspapers are not available online because of copyright issues. The corpus is only available inside the library
http://www.kbr.be/

Verein der Freunde der BF
http://www.bf-archiv.at

Narodna in univerzitetna knjižnica / National and University LIbrary, Ljubljana, Slovenia
http://www.dlib.si

Universitätsbibliothek Heidelberg
http://hd-historische-bestaende-digital.uni-hd.de
Newspapers and journals: http://www.ub.uni-heidelberg.de/helios/digi/zeitung.html

National Library of Portugal
Digital collection: http://purl.pt
Newspapers: http://purl.pt/index/per/PT/index.html

"Lucian Blaga" Central University Library, Cluj-Napoca, Romania
http://dspace.bcucluj.ro/
http://documente.bcucluj.ro/

Swiss National Library
The digitized collections are spread accross different sites, depending on the partner; the site http://www.pressesuissearchives.ch gives an overview and links to all the titles currently digitized; the platform http://www.digicoord.ch gives an overview of major digitization projects in Switzerland, including newspapers

Hemeroteca Digital. Biblioteca Nacional de España
http://hemerotecadigital.bne.es

ZBW German National Library of Economics - Leibniz Information Centre for Economics
http://zbw.eu/beta/p20

Azerbaijan National Library
http://www.anl.az/arxiv_neshr.php
(This site is for the internet use and contains only information about names and dates of the newspapers.)
web.anl.az/arxiv_neshr.php
(This site is for local use, readers can read content of the digitised newspapers.)

National Library of Norway
http://nb.no/aviser

Library of University of Latvia
http://www.lu.lv
(Project currently not finished)

Landsbókasafn Íslands - Háskólabókasafn National and Univeristy Library of Iceland
http://timarit.is/

Biblioteca Municipal de Faro
http://www.cm-faro.pt/menu/148/na-biblioteca.aspx

Secretariat of State for Culture (Spain)
http://prensahistorica.mcu.es

Moravian Library
http://kramerius.mzk.cz

Bibliothèque Cantonale et Universitaire - Lausanne
http://www.letempsarchives.ch
(for titles belonging to Le Temps)
http://www.rerodoc.ch
(for some small titles)
http://scriptorium.bcu-lausanne.ch
(for all Edipresse titles and other upcoming digitizations - to be launched in 2012).

University of Nish (Serbia), Nish Univerzity Library "Nikola Tesla"
http://www.ni.ac.rs/

Municipal Archives of Torres Novas City
Not available online

Bibliothèque nationale de Luxembourg - National Library of Luxembourg
http://www.eluxemburgensia.lu

Biblioteca Municipal José Régio
http://periodicos.bm-joseregio.com/geadopac/

Austrian National Library
http://anno.onb.ac.at/
http://www.onb.ac.at/bibliothek/digitaler_lesesaal.htm

Staatsbibliothek zu Berlin - Preußischer Kulturbesitz (SBB)
http://zefys.staatsbibliothek-berlin.de/

National Library of Poland
http://www.polona.pl
(no separate URL for the collection of newspapers)

Hamburg State and University Library
http://www.sub.uni-hamburg.de/bibliotheken/projekte/digitalisierte-bestaende.html
(with out newspaper collection which is still in process of development for web presentation)

Koninklijke Bibliotheek
http://kranten.kb.nl

National Library of Estonia
http://dea.nlib.ee

Dr. Friedrich Teßman Library, Italy
http://dza.tessmann.it

Bibliothèque nationale de France (BnF) - National Library of France
http://gallica.bnf.fr/

National Library of Latvia
http://periodika.lv/

National Library of Finland
http://digi.kansalliskirjasto.fi

National Library of Turkey (NLT)
http://sureli.mkutup.gov.tr/

British Library
http://newspapers.bl.uk

National library of Serbia
http://digital.nb.rs