



Using Automatic Article Detection and Marking Software in Production of Newspaper Clippings of a Digitized Finnish Historical Journalistic Collection

Kimmo Kettunen, Pierrick Tranouez, Tuula Pääkkönen, Erno Samuli Liukkonen, Daniel Antelme, Yann Soullard

► To cite this version:

Kimmo Kettunen, Pierrick Tranouez, Tuula Pääkkönen, Erno Samuli Liukkonen, Daniel Antelme, et al.. Using Automatic Article Detection and Marking Software in Production of Newspaper Clippings of a Digitized Finnish Historical Journalistic Collection. *EuropeanaTech Insight*, 2019, 13. hal-04485545

HAL Id: hal-04485545

<https://hal.science/hal-04485545>

Submitted on 1 Mar 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Using Automatic Article Detection and Marking Software in Production of Newspaper Clippings of a Digitized Finnish Historical Journalistic Collection

Kimmo Kettunen*[0000-0003-2747-1382], Pierrick Tranouez#, Tuula Pääkkönen* [0000-0003-3958-9732], Erno Liukkonen*, Daniel Antelme# and Yann Soulard#

*University of Helsinki, The National Library of Finland, DH Research
Saimaankatu 6, Mikkeli, FI-50100 Finland

Firstname.lastname@helsinki.fi

#LITIS laboratory

University of Rouen Normandy
France

Pierrick.Tranouez@univ-rouen.fr

The need for Articles

It is a common practice that historical newspaper collections are digitized on page level: pages of the physical newspapers are scanned and OCRed and the page images serve as the basic browsing and searching unit of the collection. Searches to the collection are made on page level and results are shown on page level to the user. Page, however, is not any kind of basic informational unit of a newspaper, only a typographical or printing unit. Pages consist of articles or news items (and advertisements or notices of different kind, too), although length and form of them can be quite variable. Thus, separation of the article structure of digitized newspaper pages is an important step to improve usability of digital newspaper collections. As the amount of digitized historical journalistic information grows, also good search, browsing and exploration tools for harvesting the information are needed, as these affect usability of the collection. Contents of the collections are one of the key elements of usefulness of the collections, but also presentation of the contents for the user is important. Possibility to use article structure will also improve further analysis stages of the content, such as topic modeling or any other kind of content analysis. Several digitized historical newspaper collections have implemented article extraction on their pages. Good examples are for example Italian [La Stampa](#), [British Newspaper Archive](#), and Australian [Trove](#).

The historical digital newspaper archive environment of the National Library of Finland is based on commercial [docWorks](#) software. The software is capable of article detection and extraction, but our material does not seem to behave well in the system in this respect. We have not been able to produce good article segmentation with docWorks, although such work has been accomplished e.g. in the Europeana Newspaper framework. However, we have recently produced article separation and marking on pages of one newspaper, *Uusi Suometar*, by using article extraction software named PIVAJ developed in the LITIS laboratory of University of Rouen Normandy [1]. In this article we describe intended use of the extracted articles in our digital library presentation system, [digi.kansalliskirjasto.fi](#) (Digi), as newspaper clippings which can be collected by the user out of the markings of the article extraction software.

Article Extraction with PIVAJ

We have described results of article extraction using PIVAJ software in a recent article [2] at the DATeCH2019 conference. The results we achieved with our training and evaluation collection were at least decent, if not remarkable, and we believe that they provide a useful way to introduce articles for users, too. Figure 1 shows an example of PIVAJ’s graphical output. Different colors show different articles. This colored output is useful to have a quick idea of how the software behaved on a given material: in production PIVAJ outputs METS and ALTO files.

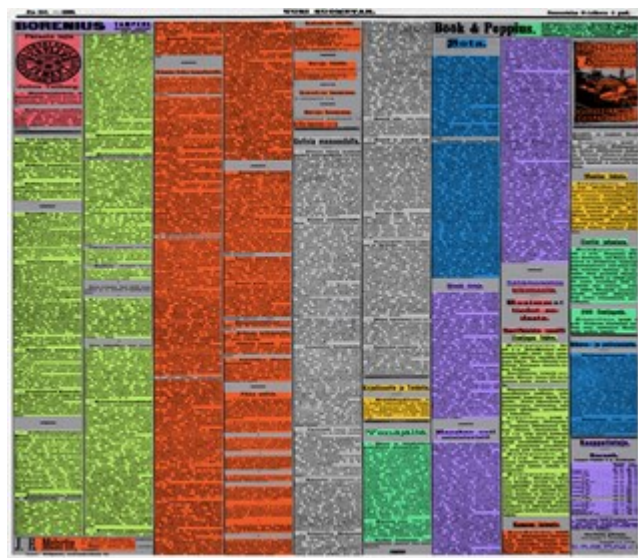


Figure 1. Example output of PIVAJ’s article extraction for a complex page

PIVAJ’s current release version (2.1) is centered on articles and thus it does not extract advertisements. PIVAJ’s development version investigates a totally different pipeline fully based on Deep Learning (more specifically FCN, Fully Convolutional Networks). Current page level extraction performs a semantic segmentation at the pixel level with satisfying results (around 90% accuracy) on Luxembourg National Library ML Starter pack dataset¹ and Gallica’s European Newspapers dataset². An example graphical result of the development version’s output is shown in Figure 2.

¹ data.bnfl.lu/data/historical-newspapers/

² api.bnf.fr/documents-de-presse-numerises-en-mode-article-du-projet-europeana-newspapers



Figure 2. Pixel segmentation of a page from Le Matin with PIVAJ development version. Advertisements in green.

Providing Articles for the User

Users of our digital presentation system Digi have been able to mark and collect so called clippings for several years [4]. This function has been quite popular and many users have collected hundreds and even thousands of clippings for their own collections on their user accounts. The clippings can also be seen by other users. Researchers have used the clipping function to collect their data, too. So far the function has been totally manual: the user has marked on the pdf representation of the page the textual area he/she is interested in and the image of the clipping has been stored with bibliographical information. The user can also add keywords, topic and title to the clipping. There has not been possibility of storing the OCRred text of the clipping so far, only an image file [4].

The procedure of creating articles automatically for the user utilizes the existing clipping functionality of Digi. PIVAJ uses the defined newspaper models of Uusi Suometar, and it provides as its output an XML file which contains the coordinates of the article regions for each recognized article on a page. In Digi's context these are the different parts of the clipping that are created in the order of the creation. After the regions have been entered to the presentation system, they are shown as individual clippings on the page.

Figure 3 illustrates the overall work flow of clipping production.

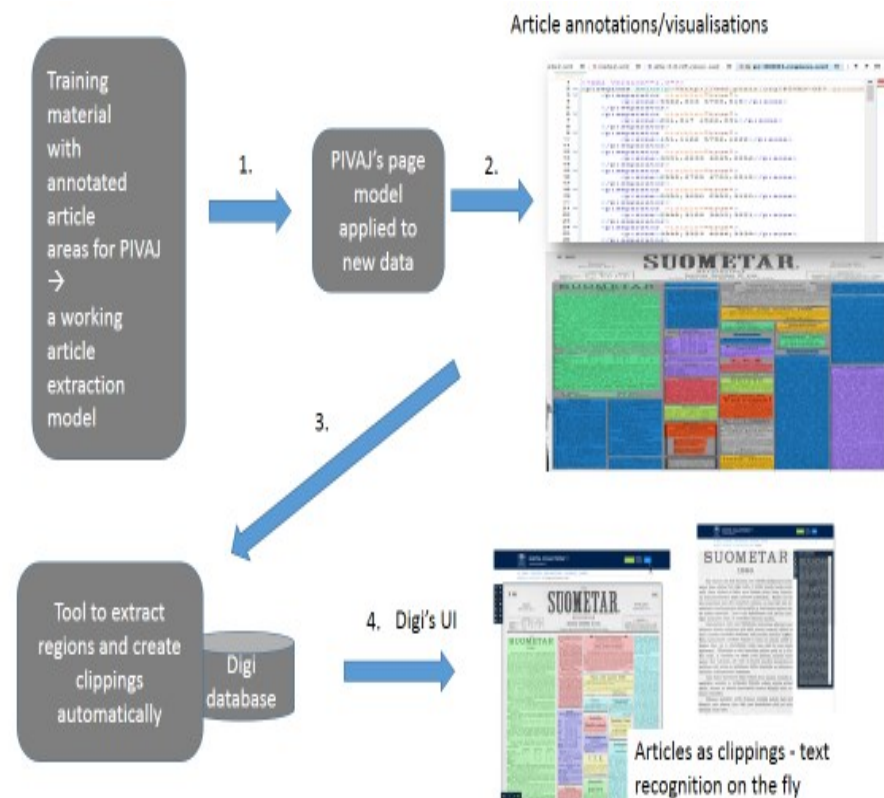


Figure 3. Flow of article and clipping creation

After choosing the article from the automatic pre-selection of PIVAJ, the user can store the article in his/her collection. The user is also able to store the OCR'd text along the clipping. This functionality is shown in Figure 4. The left part of the figure shows the clipping as an image, the right part shows the textual content.

References

1. D. Hebert, T. Palfray, T. Nicolas, P. Tranouez, T. Paquet (2014). PIVAJ: displaying and augmenting digitized newspapers on the Web Experimental feedback from the “Journal de Rouen” Collection. In Proceeding DATECH '14 Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, 173–178. <http://dl.acm.org/citation.cfm?id=2595217>
2. K. Kettunen, T. Ruokolainen, E. Liukkonen, P. Tranouez, D. Antelme, T. Paquet (2019). Detecting Articles in a Digitized Finnish Historical Newspaper Collection 1771–1929: Early Results Using the PIVAJ Software. DATECH 2019.
3. C. Clausner, S. Pletshacher, A. Antonacopoulos (2011). Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods. 2011 International Conference on Document Analysis and Recognition (ICDAR). DOI: 10.1109/ICDAR.2011.282
4. T. Pääkkönen (2015). Crowdsourcing metrics of digital collections. Liber Quarterly, <https://www.liberquarterly.eu/article/10.18352/lq.10090/>