Преглед НЦД 21 (2012), 107–116

Mojca Šavnik, Tine Musek
National and University Library of Slovenia
Saša Baždar
MFC.2, d.o.o., Ljubljana, Slovenia

DIGITISATION OF OLD NEWSPAPERS IN THE NATIONAL AND UNIVERSITY LIBRARY OF SLOVENIA (NUK)

Abstract: *National and University Library of Slovenia* (NUK) permanently stores and makes available to the public library material *slovenica* including newspapers. Old newspapers are available to users in the original format, copies on microfilm and in digital format on portal *Digital Library of Slovenia*. Article describes digitisation of printed newspapers, as well as the digitisation of copies on microfilm and workflow of digitisation in NUK and at the site of the outsourced contractor. There are specific procedures to select materials and its preparation for digitisation. Shown is an example of technical documentation as the parameters or instructions for contractor how to carry out the digitisation, implementation, and final quality control of digitised materials made by library experts. As the development of software and hardware in the 90 years of the 20th century increased, the digitisation of library materials has become involved also in the processes of protection of library materials. Digitisation of microfilm is also one way of preservation for microfilms from the extended usage and damaging environmental influences. Putting digital copies of newspapers and magazines online without any restrictions gives digitisation an additional value.

Key words: digitisation, old newspapers, microfilms, National Library of Slovenia (NUK), Digital library of Slovenia – dlib.si

1. Introduction

National and University Library of Slovenia (NUK) permanently holds the entire corpus of newspapers of *slovenica*. Printed newspapers and microfilmed copies are available to users without restriction. Frequent use of newspapers and exposure to undesirable environment influences have caused permanent damage, and to some extent even destroyed parts of the documents. Systematic microfilming of valuable, the most endangered and most borrowed materials enabled NUK to move originals to safe shelter vaults, while microfilm copy remained accessible to users. At the beginning of the first decade of the 21st century the first attempts of mass digitisation of newspapers began in NUK. At first, NUK digitised only material in analogue form (paper), but a growing share of the digitisation of newspapers is digitisation of microfilm. Digital copies of newspapers are widely available online on portal of Digital library of Slovenia – dlib.si without any restriction (Figure 1).

2. Selection and preparation of printed and microfilmed newspapers for digitisation

Selection of newspapers for digitisation consists of various criteria [1], such as image quality selection criteria, intellectual content selection criteria, refined selection criteria and identifying potential end users and their needs. The method of selection also depends on the project objectives, technical capabilities of outsourced company and NUK financial capabilities. There are also copyright-legal issues to be considered.

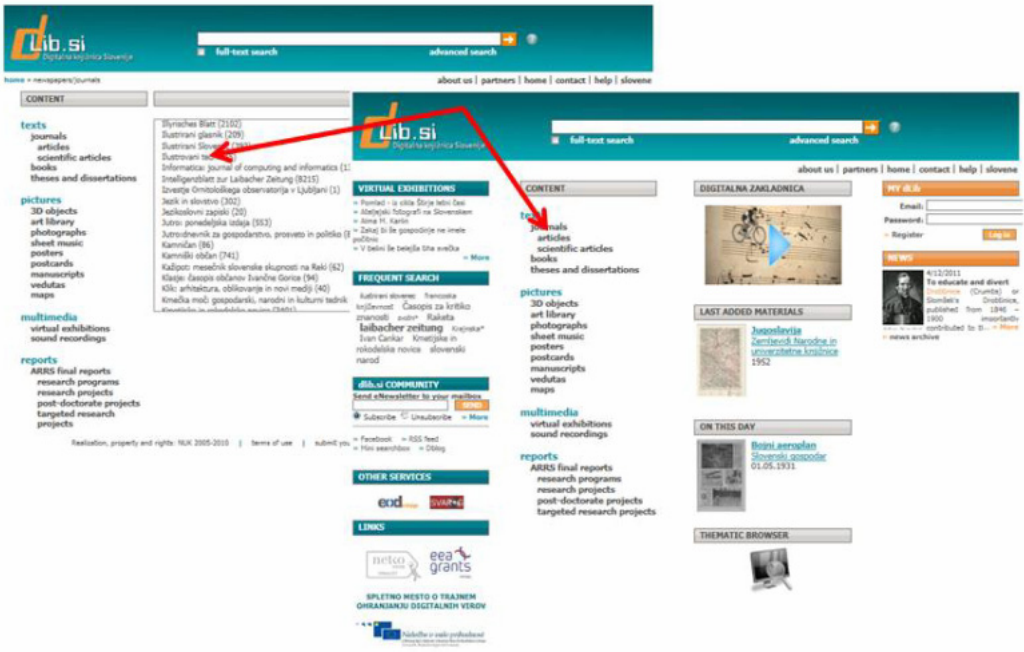


Figure 1: Digital library of Slovenia – dlib.si

NUK digitise selected newspapers to preserve originals, as they are in really poor condition due to frequent lending to end users and – by putting digital copies online – to meet users´ needs. Similar arguments prevailed while making decisions to digitise microfilmed newspapers. International recommendations and guidelines supported NUK´s decisions as correct ones, namely: while IFLA[1]´s *Guidelines for Newspaper Preservation Microfilming* [2] instructed what "Microfilming priorities [of newspapers] are in general", *Australian Newspaper Plan* [3] emphasised that since the end of 20th century "[...] at least two developments have led to a serious challenge to the previously pre-eminent preservation role of microfilming with regard to newspapers: The development of […] digital technology for capturing, organising and presenting content […]; […] increasing costs and difficulties of relying on microfilming as an adequate preservation and access path for newspaper content, […], [as it] is unpopular with users, and depends for access on equipment that is likely to become hard to maintain or acquire in the longer term.«

A refined selection criteria [1] appeared to be necessary as some of selected titles did not match to the recommended criteria [4] that »complete (or majority of) title run should be available on microfilm […]; [and] an effort should be made to deliver as complete a title run, within the prescribed date range, as possible. Locating and substituting a limited number of scanned images from paper may be necessary to complete the run." NUK holdings of old newspapers are incomplete – especially for newspapers published abroad (United States of America, Canada, South America, Australia etc.). Due to the two world wars and other historical issues in the Western Balkan region, some of old newspapers exist only as fragments. The dilemma was if the digitisation should take place if newspaper holdings were incomplete? NUK decided to digitise incomplete newspaper holdings, counting on wider community to complete missing volumes and issues and to meet users´ needs.

IFLA´s Newspaper Section pointed out in 2002 the importance of digitisation as a "digitisation has increased our means to newspapers more readily available by adding advanced search features to access text. [8]

[1] IFLA - International Federation of Library Associations and Institutions

2.1. Preparation of printed newspaper materials for digitisation. Preparation of printed newspaper material includes an overview of physical status of each title, volume and issue of newspaper: library employees count the number of pages, describe the condition of the newspaper (damage, the fragility of paper, fading of the press ...). Because the optical character recognition of a text is required, it is necessary to check the language of publications, font and the number of text columns. Bibliographic metadata are needed to prepare a general metadata scheme. Experts from the Conservation and Preservation centre of NUK inspect prepared newspapers to determine exact handling of material for digitisation process (Figure 2).

Figure 2: Example of published newspapers on portal dlib.si, digitised from paper

2.2. Preparing digitisation of microfilm. Each microfilm is reviewed on the microfilm reader. A well done microfilm inspection saves time and money, because with a thrill inspection library experts find "additional data about features in a volume or frames on a reel that will require special scanning procedures." [7] This information is valuable for library and digitisation company. Number of images is described and indicated, if the microfilm snapshot consists of more than one page of printed publications. For the optical character recognition purposes language, font and number of columns of a text are checked and described. It is well known that the quality of optical character recognition depends on

- »*The quality of original text and microfilm capture*. Poorly prepared original material, no matter how well microfilmed, yields poor results. Microfilm of bound material may have page curvature, gutter shadows, or out of focus pages that influence the digital image quality. Preference in selection should be given to titles on higher quality microfilm.
- *The reduction ratio used when microfilming the original newspaper*. This ratio directly influences image quality and optical character recognition results. The lower the reduction ratio (below 20x) the better.
- *Variations in density within images and between exposures*. Such variations require adjustment of scanning parameters within a reel. « [4]

NUK digitize archival microfilm copies (Figure 3). Archival microfilm copies are in better condition than copies used daily by users using microfilm readers. Bibliographic metadata is the same as for printed material.

Figure 3: Examples of published newspapers on portal dlib.si, digitized from microfilm

3. Technical documentation – a manual for outsourced contractor

NUK prepares general manual [5] for outsourced contractors for digitisation of all library materials. General manual or technical documentation consists of three parts, or chapters:

- details about publication,
- the parameters of the digitisation,
- management of material.

3.1. Details about publication. NUK prepares short descriptive metadata about each newspaper title, exact number of volumes and estimated number of issues and pages. From these data, any contractor may assess the cost of digitisation and the time it will take to complete digitisation. NUK uses these data to prepare a timetable for the project. The timetable is an integral part of project documentation.

3.2 The parameters of digitisation. Specification of parameters for digitisation includes parameters for file creation [6] and optical character recognition, guidelines for structuring files and creating metadata. Parameters are divided into four subsections: digitisation parameters, optical character recognition, the structuring of individual numbers, linking files, scans and naming files.

naslov	št. letnikov	št. strani/št. skenogramov	format	potek besedila
Slovenec: političen list za slovenski narod	73	približno 197.000	mikrofilm	besedilo poteka v več stolpcih; vsebuje fotografije, risbe in podobno slikovno gradivo; besedilo je v latinici, v slovenskem jeziku.
Vestnik Slovenske krščansko-socijalne zveze	3	približno 300	mikrofilm	besedilo poteka v več stolpcih; vsebuje verze; vsebuje fotografije, risbe, tabele in podobno slikovno gradivo; besedilo je v latinici, v slovenskem jeziku.
Slovenski list	4	približno 2.450	mikrofilm	besedilo poteka v več stolpcih; vsebuje fotografije, risbe in podobno slikovno gradivo; besedilo je v latinici, v slovenskem jeziku.
Ponedeljski Slovenec	8	približno 4.900	mikrofilm	

Seznam gradiva

leto izida	letnik	približno št. strani v letniku	MF	priloge
1873 - 1874	1 - 2	1.070	1	
1875 - 1876	3 - 4	1.070	1	
1877 - 1878	5 - 6	1.070	1	
1879 - 1880	7 - 8	1.070	1	

Figure 4: Technical documentation: information about publication

3.2.1 Parameters for file creation. Digitisation parameters are determined for each publication separately and are in accordance with the guidelines for the digitisation of library materials [5]. Parameters determine the file format, scans resolution and colour depth, Parameters also determine all files formats which should be made during digitisation process (Figure 5).

Files format

jpg[2] files	The scans should be in the resolution set out by the parameters of digitisation. Name. jpg file scan may be the same name. pdf file. The same name must also be in xml schema. Files should be: - crop to include visible edge of page - resized on the same pixel size; - de-skew images with a skew greater than 3 deegres.
pdf[3] files	Files should be in resolution between 96 and 150 dpi[4]. The total number of publications (bibliographic unit) should be in a pdf file (multipage).
html[5] files	Text without formating. Encoding: UTF-8
txt[6] files	encoding: UTF – 8
xml[7] fiels	Encoding: UTF – 8

Figure 5: Files format specification

[2] JPG/JPEG – file format named after Joint Photographic Experts Group
[3] PDF – portable document format
[4] DPI – dots per inch
[5] HTML - Hyper Text Markup Language
[6] TXT – text file
[7] XML - Extensible Markup Language

3.2.2 Optical character recognition. Optical character recognition of texts should be implemented in a high-resolution scans (300 dpi or more). Recognition results should be in the file formats HTML, TXT, XML and PDF

3.2.3 Structuring and naming files. Each file in formats PDF, TXT, HTML or XML must conform to a single number. A file of each issue of newspaper is primarily named with COBISS ID[8] of this publication, year of issuing and issue number. In the specific field of database is required specific record, for example: 3177015_1904_001.pdf, which is the identifier of the record but also information on how other files are named. COBISS ID was chosen for naming files, because each COBISS ID is unique identifier, consisting of just numbers. File names must not contain diacritical marks, or spaces. Instead of spaces underscore applies. NUK writes detailed instructions how to name all files because a precise description for a consistent file system, archive and permanent preservation of digitized material is needed. Example:

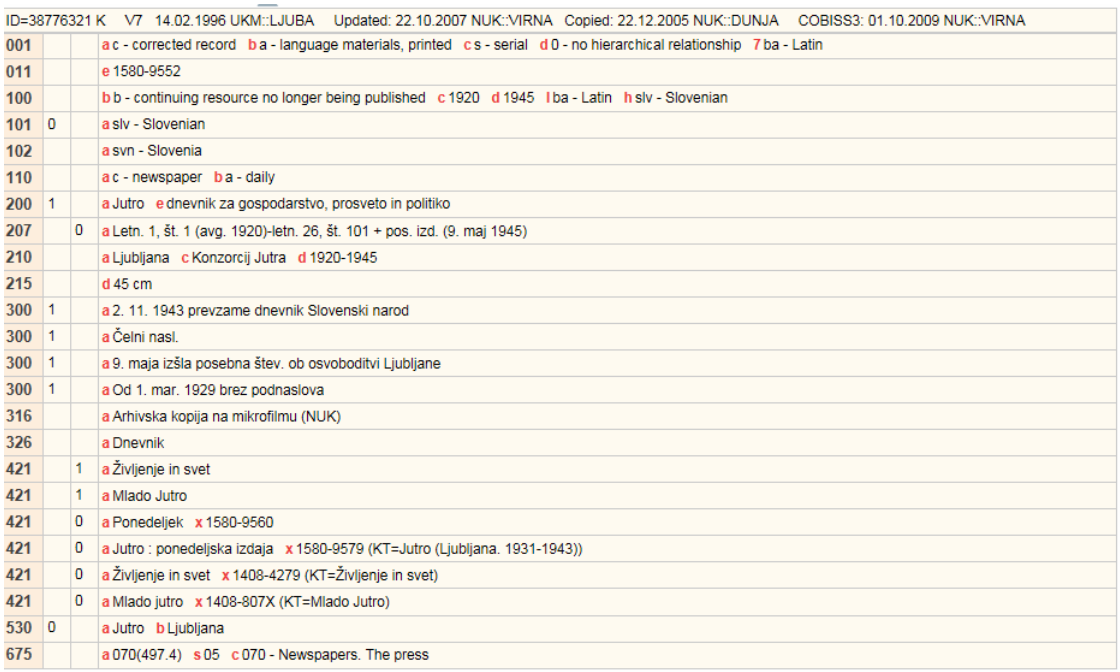
a. *Files names, structured by numbers:*
3177015_1904_001.pdf (FIRST number)
3177015_1904_001.html
3177015_1904_001.txt
3177015_1904_001.xml

b. *Scans (pages) names:*
3177015_1904_001_001.jpg (scan of FIRST page of first number)
3177015_1904_001_002.jpg (scan of SECOND page of first number)
3177015_1904_001_00….jpg

3.2.4 Creating metadata. NUK prepares a general metadata scheme as the basis for generating specific bibliographic metadata for each digitised unit of newspaper. Bibliographic metadata are prepared on the basis of bibliographic record in COBIBB.SI[9] cataloguing database. Metadata in COMARC[10] format (Figure 6) are mapped to Dublin Core[11] format and adapted for Dlib AP[12]. Model metadata scheme is also prepared by NUK and consist constant bibliographic metadata (title, subtitle, publisher, language ...) for each digitised newspaper unit. (Figure 7)

3.3 Handling of material. Conservation and Preservation centre of NUK prepared instructions [5] on how to handle library materials during the process of digitisation. The instructions are written for each type of library material and media. The contractor is handling the material during the process of digitisation as per instructions of NUK.

[8] COBISS ID is the unique identifier of the catalog record in the database COBISS.SI (Cooperative online bibliographic system and services).
[9] COBIBB.SI - Union bibliographic/catalogue database
10 COMARC - Cooperative Machine-Readable Cataloging
[11] Dublin Core - The Dublin Core Metadata Initiative, or "DCMI", is an open organization supporting innovation in metadata design and best practices across the metadata ecology.
[12] Dlib AP – Digital library of Slovenija Aplication Profile



ID=38776321 K V7 14.02.1996 UKM::LJUBA Updated: 22.10.2007 NUK::VIRNA Copied: 22.12.2005 NUK::DUNJA COBISS3: 01.10.2009 NUK::VIRNA

001			a c - corrected record b a - language materials, printed c s - serial d 0 - no hierarchical relationship 7 ba - Latin
011			e 1580-9552
100			b b - continuing resource no longer being published c 1920 d 1945 l ba - Latin h slv - Slovenian
101	0		a slv - Slovenian
102			a svn - Slovenia
110			a c - newspaper b a - daily
200	1		a Jutro e dnevnik za gospodarstvo, prosveto in politiko
207		0	a Letn. 1, št. 1 (avg. 1920)-letn. 26, št. 101 + pos. izd. (9. maj 1945)
210			a Ljubljana c Konzorcij Jutra d 1920-1945
215			d 45 cm
300	1		a 2. 11. 1943 prevzame dnevnik Slovenski narod
300	1		a Čelni nasl.
300	1		a 9. maja izšla posebna štev. ob osvoboditvi Ljubljane
300	1		a Od 1. mar. 1929 brez podnaslova
316			a Arhivska kopija na mikrofilmu (NUK)
326			a Dnevnik
421		1	a Življenje in svet
421		1	a Mlado Jutro
421		0	a Ponedeljek x 1580-9560
421		0	a Jutro : ponedeljska izdaja x 1580-9579 (KT=Jutro (Ljubljana. 1931-1943))
421		0	a Življenje in svet x 1408-4279 (KT=Življenje in svet)
421		0	a Mlado jutro x 1408-807X (KT=Mlado Jutro)
530	0		a Jutro b Ljubljana
675			a 070(497.4) s 05 c 070 - Newspapers. The press

Figure 6: Catalogue record for serial publication Jutro: dnevnik za gospodarstvo, prosveto in politiko (Morning: journal for the economy, education and politics) in COMARC format

General metadata scheme:

Field in XML scheme	Description and notes
<title>**publication title**</title>	Prepared by NUK
<publisher>**publisher**</publisher>	Prepared by NUK
<date>**date**</date>	DD.MM.YYYY
<type>**text, printed**</type>	Prepared by NUK
<format>**volume x, issue x, xx page(s)**</format>	Only avaliable information sholud be written.
<identifier>**unique identifier**</identifier>	Prepared by contractor
<source>**publication title**</source>	Prepared by NUK
<language>**language**</language>	Prepared by NUK
<relation>**file.pdf**</relation>	File naming is determined in manual
<scans>**scan name**</scans>	File naming is determined in manual

Example for publication JUTRO: DNEVNIK ZA GOSPODARSTVO, PROSVETO IN POLITIKO:

Field in XML scheme	Description and notes
<title>**Jutro: dnevnik za gospodarstvo, prosveto in politiko**</title>	Prepared by NUK
<publisher>**Konzorcij Jutra**</publisher>	Prepared by NUK
<date>**01.08.1920**</date>	DD.MM.YYYY
<type>**text, printed**</type>	Prepared by NUK
<format>**volume 1, issue 1, 12 pages**</format>	Only avaliable information sholud be written.
<identifier>**unique identifikator**</identifier>	Prepared by contractor
<source>**Jutro: dnevnik za gospodarstvo, prosveto in politiko**</source>	Prepared by NUK
<language>**slv**</language>	Prepared by NUK
<relation>**file.pdf**</relation>	File naming is determined in manual
<scans>**ime scan datoteke**</scans>	File naming is determined in manual

Figure 7: Metadata schemes examples

4. Process of digitisation

After tender for digitisation is completed, the outsourced contractor transports NUK´s newspapers to digitisation site. At the handover of the material, the contractor signs receipt, where

it is precisely written which material is to be transferred and the quantity of material. The contractor must transport material in containers which are protected with foam and each copy of the material wrapped in a special protective foil. Digitisation process follows the guidelines and recommendations and NUK technical documentation for creating digital copies and instructions for handling newspapers. With appropriate hardware (scanners) large size newspapers (up to A0 format[13]) can be digitised. Heritage institutions - including libraries - choose to digitise newspapers, which were previously recorded on microfilm. Digitisation of microfilm is faster than the digitisation of library materials in physical form - in this case newspapers. The quality processing of each image (scans), the contractor clips focus and lighting, and crops in a single size. In this way, the contractor prepares all the digital copies of material in the same quality. The same high quality scans provide higher quality of optical character recognition), as an indispensable part of the processing of digitised texts. Digitised material, together with metadata, contains a digital collection.

4.1 Hardware. Scanners are able to capture or digitise materials of different size formats. In practice digitisation of libraries and other heritage institutions can be implemented on digitisation scanners:

- up to A3 format[14] size documents
- documents from size A3 format to size A0 format.

Scanners for digitising large formats are larger devices, as heritage institutions digitise large volumes of library materials, maps, posters and similar materials in large formats. Due to the growing demand for digitisation of microfilm, digitising companies include microfilm scanners in their hardware.

4.2. Processing of digitised material (scans). Microfilm digitisation process is faster than the digitisation of printed library materials. Processing digitised scans varies - processing depends on digitised medium: paper or microfilm.

Processing scans, generated by the digitisation of microfilm is difficult and longer than processing digitised scans of printed materials. The main difference is that the digitisation of microfilm is digitising already modulated reproduction of the original. Images on microfilm lose a certain quality, especially sharpness and light. When the material is converted, either from the original material, or from the microfilm copies in digital form, contractor must make "all operations that change the image dimensions, spatial resolution, or orientation (e.g., cropping, de-skewing) […] to [high resolution scans] …" [4]. . All scans are cropped to the same size. Scans generated by the digitisation of microfilm are, after processing, usually lightened or darkened. All scans are in the same quality, before the optical recognition processes begun. Optical recognition processes are in the same quality - that ensures the uniform quality of the optical character recognition of texts.

4.3 Creating and naming files and metadata preparation. High resolution scans are processed as per NUK technical documentation. Results are files in pdf and txt format. Metadata are created and saved in xml format files.

Preparation of metadata is challenging and time-consuming. Process requires a lot of manual work and specific knowledge of bibliographic description of digitised material. Cooperation between NUK and contractor is needed while creating metadata and file naming. The result of this working method is a reduction of errors and corrections.

[13] Size of A0 format: 841 mm × 1189 mm
[14] Size of A3 format: 297 mm x 420 mm

5. Quality acquisition of digitised material

After digitisation is completed the contractor hands over digitised material to NUK in quality control and final acquisition. NUK counts all the files - the number of scans must correspond to the number of pages, number of files in PDF format, XML format, and HTML format must be the same. NUK controls high resolution scans colour depth, resolution and uniform size. Html files should not contain tables, pictures and graphs. Files in txt format, html format and xml format must be in encoding to be UTF-8 encoding schema. Pdf file format should be in the required resolution.

NUK pays a lot of attention to metadata quality control and file naming control. The random files checking find errors in xml files or in naming files. The contractor is obligated to correct mistakes according to NUK´s instructions.

6. Completion of the project and publishing on portal dlib.si

Digitisation project is successfully completed when the digitised material is published on the portal of the Digital Library of Slovenia - dlib.si (Figure 8) and securely stored in a digital archive. Archived digitised materials are managed in accordance to guidelines and recommendations of permanent preservation.

Digitisation of newspapers follows the principles of the NUK guidelines [5]. Cooperation between the client and the contractor during the project is reflected in the quality of implemented and published projects

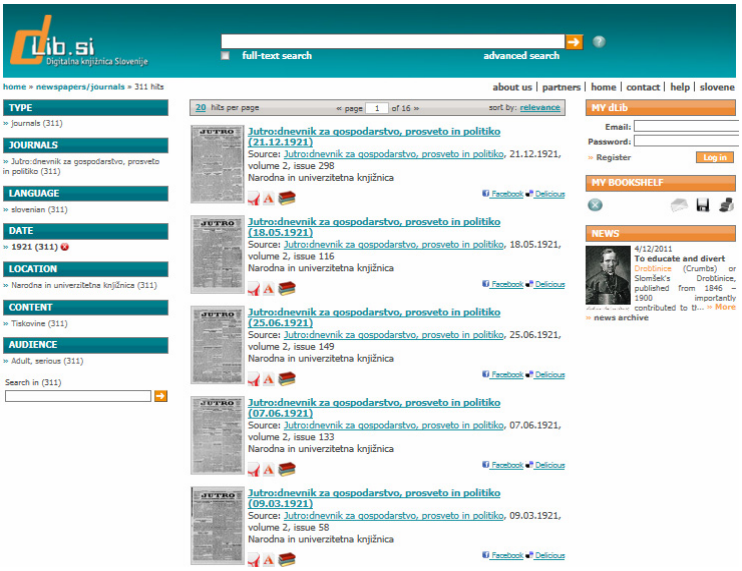


Figure 8: Digitised newspaper on portal dlib.si

References

[1] National Digital Newspaper Program: Guidelines and Resources. Available: http://www.loc.gov/ndnp/guidelines/selection.html. [Accessed: 2011-3-12]
[2] Guidelines for Newspaper Preservation Microfilming. Available: http://archive.ifla.org/VII/s39/broch/pr49-e.pdf. [Accessed: 2011-04-17]
[3] Australian Newspaper Plan. Available: http://www.nla.gov.au/anplan/about/preserve.html. [Accessed: 2011-03-31]
[4] The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants Available: http://www.neh.gov/. [Accessed: 2011-12-22]

[5] Smernice za digitalizacijo knjižničnega gradiva. Available: http://www.mk.gov.si/fileadmin/mk.gov.si/-pageuploads/Ministrstvo/Drugo/aktualno/2010/smernice_za_digitalizacijo_koncna_TL_ZK_MS.pdf . [Accessed 2011-04-17]
[6] Data Dictionary – Technical Metadata for Digital Still Images. Available: http://www.niso.org/kst/-reports/standards/kfile_download?id%3Austring%3Aiso-8859-1=Z39-87-2006.pdf&pt=RkGKiXzW643-YeUaYUqZ1BFwDhIG4-24RJbcZBWg8uE4vWdpZsJDs4RjLz0t90_d5_ym-Gsj_IKVaGZww13HuDlSn6cvwjex0ejiIKSaTYlErPbfamndQa6zkS6rLL3oIr [accessed 2011-03-01]
[7] Meyer, L., Janet Gertz: RGL Guidelines for Microfilming to Support Digitization. Available: http://www.oclc.org/research/publiacations/library/Pres_Micro_Supplement.pdf [accessed: 2011-12-23]
[8] Microfilming for Digitisaton and Optical Character Recognition: Supplement to guidelines. Avaliable: http://archive.ifla.org/VII/s39/broch/microfilming.htm. [accessed: 2011-12-23]

Mojca.Savnik@nuk.uni-lj.si
Tine.Musek@nuk.uni-lj.si
sasa.bazdar@gmail.com