ЕЛЕКТРОНЕН АРХИВ НА НАУЧЕН АРХИВ – БАН

инж. Валентин Георгиев, Марио Развигоров, НА – БАН

В отговор на нарасналото търсене и използване на документи, съхранявани в Научен архив на Българската академия на науките (НА – БАН), и с оглед ограничаване неизбежните разрушителни въздействия при съответния физически достъп до тях, се извършва интегриране на модерните технологии и интелектуални комуникации в архивната работа. Целта е да се предоставят на изследователите, независимо къде се намират те в дадения момент, онлайн е-услуги относно интересуващите ги архивни справочници, документални свидетелства и ръкописи. Ето защо създаването на съвременен електронен архив на традиционния Научен архив на академията е приоритетна задача в условията на мрежовото общество.

Интересно е да се отбележи, че на този етап достъпът до електронния архив на НА – БАН вече е възможен до ниво опис. Самият архив е ситуиран на интернет страницата на НА – БАН[1], което означава, че в читалнята на Архива през локалната мрежа от интернет сайта потребителят вече има възможност да чете дигитализирани текстове от *Сбирка X – Старобългарски и славянски книги и ръкописи в Научния архив на БАН* (Вж. Приложение № 1).

Освен това електронният архив включва и значителен масив от дигитализирани документи от различни фондове и сбирки. Приблизителният обем на съответната информация е 1,5 TB (терабайта), и включва изображения от най-често ползваните архивни фондове, сред които на проф. Иван Шишманов, Константин Иречек, Българското книжовно дружество и др.[2]

Относно процеса на дигитализация: оригиналният ***Мастер-файл*** е с резолюция от 600 dpi, отговаряща на оптическата резолюция на скене-

ра. При самото сканиране на документа, електронния вариант се запаметява първо в TIFF формат[3]. Форматът е избран поради възможността му да съхранява информацията без промяна в качеството ѝ, като запазва и цветния профил на устройството, с който е получен, и се поддържа от голям брой приложения за обработка на *растерни изображения.*

Копията на оригиналния дигитален документ вече подлежат на обработка, преформатиране – JPG, PNG[4], и поставяне на защитни елементи. Така обработени те се предоставят за ползване от потребителите.

Друг тип файл, който намира широка употреба в НА – БАН, е DJVU[5]. Благодарение на сложни алгоритми за обработка, които разделят текста, като преден план в изображението, и фона – заден план, в различни слоеве със специфични методи за компресия (Вж. Приложение № 2), този формат успява да постигне уникални степени на *смачкване на информацията* – при съпоставимо с PDF[6] качество DJVU *файловете* са около 10 пъти по-малки по размер. Това прави DJVU оптимален вариант за съхраняване на големи обеми от цифровизирана техническа документация, за която са характерни много графики и илюстрации. Друго предимство е, че въпросният формат не позволява изменение на информацията, предоставяна от документа, след като той вече е създаден. Освен това поддържа и опция за вграждане на OCR[7] разпознат текст. За разлика от JPG, DJVU дава възможност за отваряне на многолистови файлове, което не само облекчава трафика *интернет/интранет среда,* респ. за изграждане и поддържане на информационния сайт, но и на *документните бази данни* – за съхраняване и обработка на информацията от архивните фондове и сбирки[8].

Научният архив на БАН, използвайки развитието на електронните технологии се е заел и със задача да предостави нови възможности на изследователите. Конкретно визираме опцията на *Optical character recognition* (OCR) за машинно разпознаване на буквените символи от дадено изображение, което в случая дава възможност на изследователя да търси, маркира и копира текст от архивния документ, както и да го постави в обработваем файл (Вж. Приложение № 3). Тази опция улеснява изследователите и в работата във връзка с написването на научните им трудове. Тя е полезна, спомагайки за разчитането на букви и думи, които са трудно различими в използваните източници.

Някои софтуерни продукти позволяват коригиране на неразпознатите думи и запаметяване на новия модел на разпознаване. Известно е, че

испанската дигитална библиотека *Мигел де Сервантес* отдавна използва посочената възможност за улеснение на потребителите си[9]. За съжаление все още липсва подобна опция относно ръкописните текстове.

Като продължение на идеята за получаване на пълния текст на сканирания документ, Научният архив на БАН обръща специално внимание на разработването на т. нар. *уики модул,* който да дава възможност на потребителите да подобряват информацията, извличана от документа по описания метод. Това ще позволи пълномащабно търсене във всички документи от базата данни.

Научният архив на БАН има амбицията да представя дигитализираните документи и ръкописи в интернет използвайки предимствата, както на релационните БД (бази от данни) за изграждане и поддръжка на информационния сайт, така и на документните бази данни за съхраняване и обработка на информацията от архивните фондове и сбирки. За целта се предвижда новият вариант на сайта на Архива да се базира на система за управление на съдържанието работеща върху MySQL *БД*. Така ще се улесни внедряването на система за контрол на достъпа до информацията и ще се подсигури възможността за въвеждане на електронни разплащания.

Изграждането на електронното хранилище обаче ще трябва да се базира на *документна БД* от типа на MongoDB, за да се преодолее основния недостатък на релационните системи за управление на БД, а именно: когато се стремим към тяхното нормализиране, информацията да не се дублира, т.е. да се намира само на едно място в БД, тъй като при обекти за описване с разностранен характер и параметри, броят на таблиците значително се увеличава. Работейки обаче NOSQL *БД*, ще се използват само необходимите за описване на дадения документ полета (Вж. Приложение № 4).

Бележки

[1] <http://archiv.cl.bas.bg/>, 10:10, 21. 08. 2015 г.

[2] <http://archiv.cl.bas.bg/kolekc1.htm>, 10:11, 21. 08. 2015 г.

[3] *Tagged Image File Format* – формат за запис на растерни изображения, разработен като универсален формат за съхранение на сканирани изображения.

[4] *Joint Photographic Experts Group; Portable Network Graphics.*

[5] Произнася се *дежавю* – файлов формат разработен преди всичко за съхранение на сканирани изображения с текст и рисунки <http://djvu.org/>, 10:21, 21. 08. 2015 г.

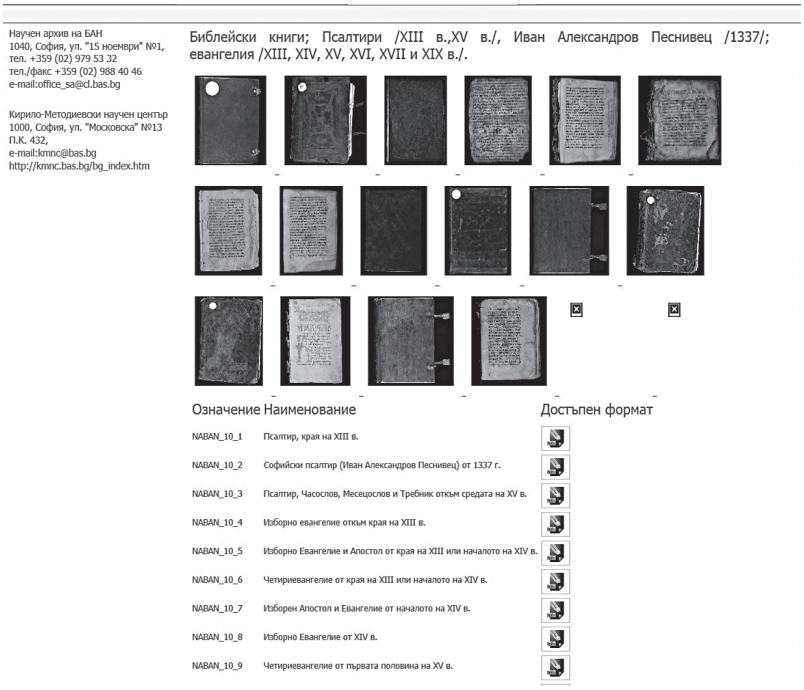
[6] *Portable Document Format* – файлов формат предназначен за обмен на документи, независимо от хардуера, операционната система или приложния софтуер. <https://get.adobe.com/reader/>, 10:31, 21. 08. 2015 г.

[7] *Optical Character Recognition* – Оптично разпознаване на текст.

[8] <http://djvu.org/gallery/documents/books/frameset.php?item=lacy/index.djvu&param=zoom=200%20scrollbars=no%20navpane=thumbnail%2Cleft%20toolbar=bottom,fixed-ruler%2Ccalibrate%2Ctextsel%2C,search%2Clizard%2Cdoublepage%2Cfore%2Cback%2Ccolor%2Cbw>, 10:26, 21. 08. 2015 г.

[9] <http://www.cervantesvirtual.com/partes/234931/la-ilustracion-espanola-y-americana/81>, 10:38, 21. 08. 2015 г.

Приложение № 1



Научен архив на БАН
1040, София, ул. "15 ноември" №1,
тел. +359 (02) 979 53 32
тел./факс +359 (02) 988 40 46
e-mail:office_sa@cl.bas.bg

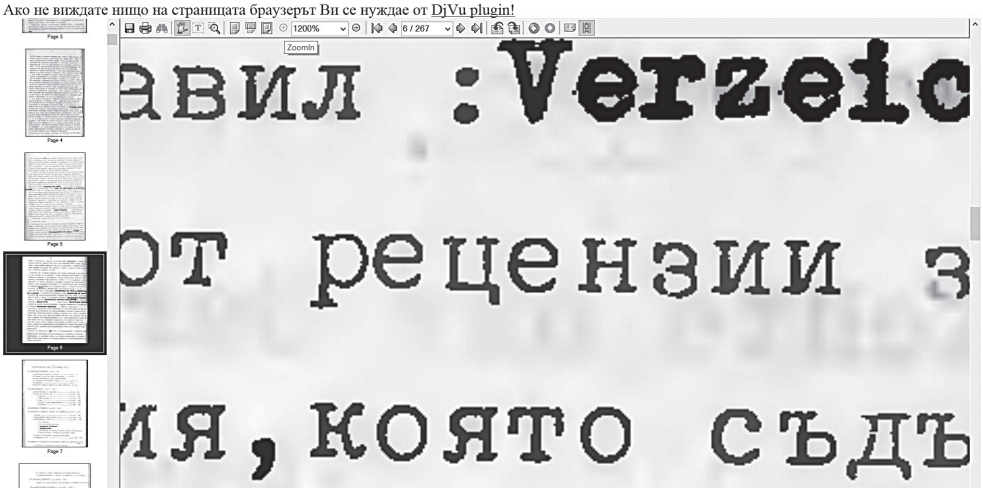
Кирило-Методиевски научен център
1000, София, ул. "Московска" №13
П.К. 432,
e-mail:kmnc@bas.bg
http://kmnc.bas.bg/bg_index.htm

Библейски книги; Псалтири /XIII в.,XV в./, Иван Александров Песнивец /1337/; евангелия /XIII, XIV, XV, XVI, XVII и XIX в./.

Означение	Наименование	Достъпен формат
NABAN_10_1	Псалтир, края на XIII в.	
NABAN_10_2	Софийски псалтир (Иван Александров Песнивец) от 1337 г.	
NABAN_10_3	Псалтир, Часослов, Месецослов и Требник откъм средата на XV в.	
NABAN_10_4	Изборно евангелие откъм края на XIII в.	
NABAN_10_5	Изборно Евангелие и Апостол от края на XIII или началото на XIV в.	
NABAN_10_6	Четириевангелие от края на XIII или началото на XIV в.	
NABAN_10_7	Изборен Апостол и Евангелие от началото на XIV в.	
NABAN_10_8	Изборно Евангелие от XIV в.	
NABAN_10_9	Четириевангелие от първата половина на XV в.	

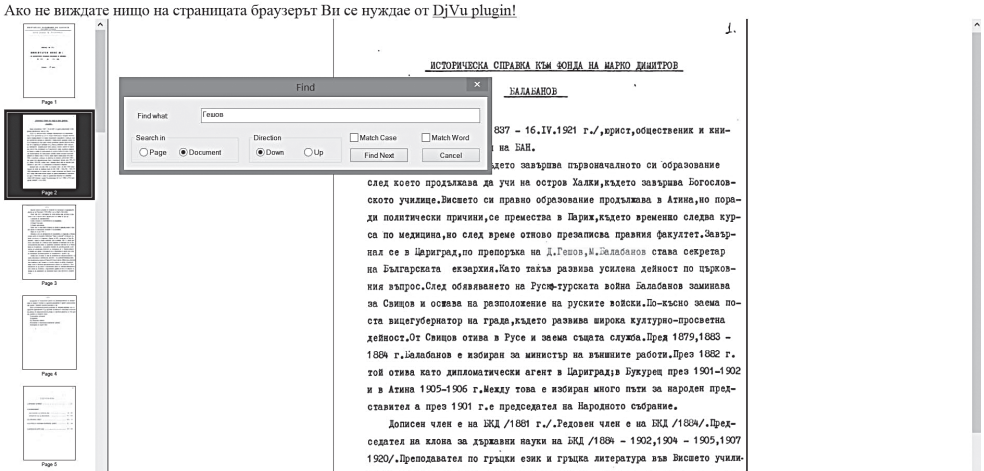
Приложение № 2

Ако не виждате нищо на страницата браузерът Ви се нуждае от DjVu plugin!



Приложение № 3

Ако не виждате нищо на страницата браузерът Ви се нуждае от DjVu plugin!



Приложение № 4

Колекция - Феликс Каниц

Обратно

Троянски манастир.

Експонат:	Живопис
Година/Век:	[1885]
Техника:	акварели
Материал:	хартия
Размери(см.):	20,5 x 29,8 см
Произход/Локализация/:	Виена
Състояние:	много добро
Местонахождение:	Научен архив на БАН
Забележки:	диптих
Анотация (бълг.):	Изгледи към Троянския манастир.