Providing access to historical documents through digitization

Barbara Chmielewska
Acquisition and Collection Development Department,
University of Warsaw Library, Warsaw, Poland, and

Agnieszka Wróbel
Department for Development of Digital Resources,
University of Warsaw Library, Warsaw, Poland

Abstract

Purpose – The purpose of this paper is to show the beginnings and development of digitization processes in University of Warsaw Library by examining the digital library of e-bUW.

Design/methodology/approach – The main method is the comparison to other similar actions within the country of Poland. The subject scope of the paper also includes a description of methods of presenting digitized collections to patrons.

Findings – It turned out that fair visibility of catalogue records of digitized material is possible mainly because the Library of Warsaw University is a member of Digital Libraries Federation (Federacja Bibliotek Cyfrowych). Since 2007 the collection has been visited by over 5 million readers.

Originality/value – The original value of the paper is made by presenting efforts of one country on an international level to a broader audience with special attention to individual solutions.

Keywords Digitization, Presentation, Historical documents, Library federation, Historical research, Poland

Paper type Case study

Digital Library of Warsaw University, so called e-bUW, was established in 2007. Currently it hosts over 100,000 documents, the majority of which are XIXth century periodicals. One of the aims of establishing e-bUW was to enable access to less used documents or to the historical documents that can no longer be presented to the users easily. Documents chosen to be included into the collection should also be useful for the current research processes at the University, for the education processes as well, in particular distant learning.

Internet users, apart from accessing the content directly from the library’s home page, also have other possibilities of reaching the e-bUW collection via various access point, domestic or foreign.

Fair visibility of such catalogue records is possible while the Library of Warsaw University is a member of Digital Libraries Federation (Federacja Bibliotek Cyfrowych). Since 2007 the collection has been visited by over 5 million readers.

Brief history of Warsaw University Library and its characteristics
Warsaw University Library is currently the three biggest libraries in Poland. The collection has already exceeded three million items of library material. The library was...
established in the end of XIX-the century (1881) and nearly since the very beginning it was given the privilege and right to receive a legal deposit. This characteristic is definitely worth mentioning for the legal deposit copies are very often a core of our digitization projects. This fact is especially important in case of multivolume titles, series and periodicals. The library receives the legal deposit up to present day[1] and because of library’s flat (and decent) budget it makes the largest source of book and journals acquisition and holdings.

Brief history of e-bUW collection
A rapid development and success of presenting of electronic documents in the 90s by digital libraries in the USA and in Western Europe encouraged also the polish librarian and scientific environment to undertake similar activities towards building their own digital resources. First initiatives that allowed to put polish full-text resources, both digitized and of electronic origin, to the internet appeared in the beginning of this century: Virtual Library of Polish Literature (2001)[2], Digital Library of Wielkopolska (2002)[3] and Polish Internet Library (2001)[4].

Warsaw University Library started its works on presenting digitized documents (historical resources) in the beginning of the year 2003[5]. At that time, Library of Warsaw University in cooperation with Polish National Library decide to launch a project called “Digital Collection of Polish Periodicals”[6]. The main purpose of this initiative was to provide a vast (public) access to the most important Polish Periodicals (in very broad time context: from the beginning of XVII century to the first half of the XX century) through digitization of microfilm copies, originating from a great collection of microfilmed Polish periodicals owned by Polish National Library. The results of this project were to be presented on both libraries’ home pages. Since 2005 this project was financed and implemented only by BUW (Library of Warsaw University), what resulted in deceleration of the project. A year later, 2006, some other institutions joined the project, among them the Library of Polish Museum of Sport and Tourism and the Library of Main School of Merchandise. With their financial help it was possible to digitize some large titles and put them on the web: “Czasopismo Skarbowe” 1926-1939; “Gazeta Bankowa” 1921-1938; “Polityka Gospodarcza” 1935; “Przegląd Gospodarczy” 1920-1939; “Przegląd Sportowy” 1921-1939, 1945-1959; “Przemysł i Handel” 1920; “Tygodnik Handlowy” 1919-1936.

The overall number of periodicals in this collection was 27[7]. In the year 2008 it was decided that Digital Collection of Polish Periodicals (e-KCzP) should be moved to the collection of e-bUW and to the dLibra platform[8]. It was a pragmatic decision. It would make no sense to run two similar projects and a united the collection was also easier for the readers to browse. Another reason for moving the collection into one place was that the library system gave the user more searching possibilities for the transferred documents underwent the process of OCR. Until 2010 18 titles from the collection have been moved, which resulted in 3507 single volumes, and subsequently 127,752 scans. The migration process has been suspended in the year 2010 because of another project – NUKAT – the highway of digital information.

Construction of the e-bUW digital library began concurrently in December 2006. It lasted about 12 months. From the very beginning the digitized collection was meant to be a new unit within the library based on already existing library structure and organization. It was planned, that the “production line” would be submitted to the
usual way of any other document, that functions in the library. It was decided, as far as possible, to use already existing solutions. If a digital document required any adjustment, changes should be made in a way that not disturb the usual work. The official launch of the e-bUW took place on 21 December 2007 (www.buw.uw.edu.pl/index.php?option=com_content&task=view&id=465).

Today, many librarians from different departments are engaged in the work of the digital library. Every department has also its own task on different stages of accessing an electronic document to the collection. Special Collections – planning the development of the collection and process the special collections documents. Periodicals Department – processing of the periodicals. Cataloguing Department – processing of the monographs. Conservation and Preservation Department – scanning and processing the scans into the final version, suitable for presenting. Public Services Department – planning the collection, publishing in the digital library, modifying the bibliographic records and work coordination.

Warsaw University Library is the main coordinator and manager of the e-bUW. Nevertheless all 47 faculty libraries of Warsaw University are invited to participate in the project. The collective character of the digital library is expressed also in the name of the digital library. It is a small “b” not the capital “B” to differentiate between the main library and a system of all libraries.

An active participation in the project is done by six units of Warsaw University: The library of Law and Administration Faculty, Library of Biology Faculty, Library of Hungarian Philology Institute, library of Journalism and Political Sciences Faculty, University of Warsaw Archives and University of Warsaw Museum. Each of these units takes part according to their staff and work potential.

Selecting documents for the collection
One of the assumptions made when planning the digital library was that the core of the internet resources would be made from the documents from the public domain. In Poland, materials can enter public domain 70 years after the death of the last author, or when the author is unknown, and the first publication of the work was 70 years ago. One should mention in addition that the Polish copyright law does not protect normative acts or their official projects, official documents, materials, signs and symbols, published patent descriptions or simple press information. The digital library was also seen as a place for the workers of Warsaw University for their publications to be put. However, it was difficult to compete with a significant resistance of a part of the academic environment and eventually the main focus of the collection was moved towards the documents that are no longer copyrighted.

There are a few criteria to take into consideration when selecting documents for the collection. First of all, we try to pick material that has a value of national heritage. Second, the material included in the collection should promote knowledge about history and culture of Warsaw and the Mazovia region, and thirdly it should be valuable and useful for classes and lectures held on the university and for university research projects. We also try to avoid duplicating documents that are already accessible through other digital libraries. There are, however, exceptions from that rule. In some cases, when the condition of the document is very poor (old frail acid paper) or when a copy of a title is unique, for example with an inscription from the author. Some may say, that it is an obvious practice – to avoid duplication of the
documents. But on the very beginning of the project, when many different digital libraries existed separately, it was very hard to do that, also because of the lack of proper tools. The possibility to verify the digital duplicates appeared only with the birth of the Federation of Digital Libraries[9]. It is still a tool to be mastered, but the main advantage of it is its existence.

It seems, that duplicates within digitized collections would not be avoided fully for one more reason. In Poland there is no central unit to coordinate work of digital libraries. We may ask, if it is at all necessary? Especially, when we look at library collections from the historical perspective. Just to mention one example. A great number of periodicals titles are scattered within several libraries so that any library does not hold the whole of the title. Furthermore, the items are very often preserved in different, often tragic condition. In such cases duplicating – presenting more than one copy of a title is a very ergonomic strategy.

One more example from our library. Historical resources of periodicals in the Warsaw University Library represent quite a complete collection, especially among those titles that were published on the territory of the Polish Kingdom in Napoleonic times and on the territory of Russian annexation. Such situation allowed us to supplement our gaps and destructs in the periodicals’ collection with the resources of other libraries. A similar approach was applied when e-bUW started its work. An additional task was fulfilled while digitizing the periodicals. Our library decided to digitize all titles that are somehow connected with each other in order to enable the user as broad access as possible. This kind of strategy was not implemented by all digital libraries, what caused some problems for our library solution. Is the lack of full collection a sufficient argument to duplicate the whole resource? And maybe it is only worthy to fill the gaps, but who will be then able to check, how much has been digitized? Up to now, Polish digital libraries have not worked out unified (common) standards in the field of scanning and digitizing. They scan and process scanned documents according their own regulations. Unfortunately, in many cases it results in new dilemmas, for the quality of the end-documents is a great deal to be desired.

There are the same questions arising when proceeding multivolume monographs, but is much more easy to deal with such documents. In some cases we have started cooperation with other libraries and now we are presenting the whole title in the collections of more than one library but this has been a deliberate decision. An example of such multivolume monograph is Herbarz Polski Kaspra Niesieckiego… (Book of polish Coat-of Arms by Kasper Niesiecki), where three first volumes are published by the Digital Library of Zielona Góra (http://zbc.uz.zgora.pl/dlibra), and seven other volumes are published by e-bUW (http://ebuw.uw.edu.pl/dlibra/publication/166?tab = 1).

Warsaw University Library was able for three times to gain money for digitization from external funds, mostly governmental and union. We will describe shortly the three projects financed with help of those funds:

(1) Ephemera. The implementation of this project was possible thanks to the subsidy of Ministry of Culture and National Heritage, which the library was granted within the program “Cultural Heritage” in 2006. There was an overall of 4,400 documents digitized (8952) scans. From this number 241 documents are now presented in e-bUW.
(2) Book Collection of the Warsaw Society of the Friends of Science. The implementation of this project was possible thanks to the subsidy of Ministry of Culture and National Heritage, which the library was granted within the program “Cultural Heritage” in 2007. Within the project a small number of books from Warsaw Scientific Society Book Collection was digitized. All 387 titles are now presented in e-bUW with an overall of 45,753 scans.

(3) NUKAT – Highway of Digital Information. The implementation of this project was possible within an Innovative Economy Operative Program[10]. We are going to use the funds to digitize over 500 periodicals titles[11], published from the end of XVIII century to the first years of the XX century. A majority of those titles are XIX-th century journals (about 300,000 issues, 3,000,000 scans). We have already published 118 titles (82,412 issues, 755,232 scans)[12].

The users of the digital library also have power to decide, what is going to appear as next items. They have the possibility to place their requests via e-mail (http://ebuw.uw.edu.pl/dlibra/news?news=full). One of the later mentioned project was also provided with a short survey, where the users could share their interests towards digitized documents. The library makes efforts to engage the academic workers as well. This cooperation has already succeeded in creating a few thematic collections, such as: “Literature for pure reason” or “Charles Darwin and surroundings”.

In 2009 e-bUW took part in a guidance survey prepared within the project EuropeanaLocal. Three countries were chosen to fill a trial version of the survey: Sweden, Greece and Poland. In every country there were four collections chosen to be the object of the survey. Among the polish collections there was one collection from e-bUW “Science and Didactics”. The aim of the survey was to retrieve information about resources that could be further included into the Europeana Library.

Processing the documents

All the materials for the digitization coming from the collections of the Warsaw University Library are scanned within the library. Here are some technical regulations according the digitization of the documents. The assumption was that documents smaller than A4 format would be scanned double pages on one scan/file (2:1). Since February 2012 in order to make the copy as close as possible to the original we thought out a different way of scanning, that is one page in on scan/file (1:1). The digital object are saved in TIFF format, one-page in 24bit color, no compression, lossless compression is permissible. In case of documents containing text only with good quality of print there are other features acceptable: gray scale, white-black scale. A set resolution for monographs and serials is: 300, 400, 450 dpi, for selected objects, especially special collections, the resolution is 600 dpi.

Source files before the conversion into the viewing format could be modified and adjusted. (e.g. sharpening or color correction). At this stage we also pay attention to the right pages numeration, for it happens sometimes that there is a failure in binding. Only after all those preparations the files undergo the conversion into viewing files. Digital library of e-bUW decided, similar to the majority of digital libraries in Poland, that the documents presented in the Internet would be in DjVu format. During the process of DjVu conversion each file is provided with metadata description.
If the condition of the object is fair enough all of the documents published in e-bUW are also automatically processed with OCR. However due to restricted technical possibilities OCR is not fully controlled. In order to facilitate the user to navigate through the text, every item with number of pages greater than 21 are also “equipped” with interactive tables of content (internal hyperlinks). Unfortunately because of the bad condition of paper publications (different kind and size of font, paper condition, etc.) not every document has the mentioned facilities.

Creating the catalogue records (metadata) and links to various catalogues
During all those years, in Poland there are still no national regulations for creating bibliographic descriptions in digital libraries. Only some attempts were made in order to establish such regulations (see: http://ebuw.uw.edu.pl/dlibra/news?news=full). These, however, are still regulations with no obligation for the library to be implement. They are only rank-and-file initiatives, with their main focus of local institution.

There are two main factors that influence the descriptive metadata in e-bUW collection: firstly, it is the software, that has been chosen to create the digital library and second, it is the route (the way) of the document in the library. It has been assumed, that before a document enters e-bUW, it undergoes following stages: technical preparation – cataloguing process – digitization – publishing in e-bUW – additional data complementation.

The e-bUW digital library uses the dLibra software[13]. It is a domestic product, launched by Poznan ´ Supercomputing and Networking Center[14]. To describe documents within this system we chose the standard of Dublin Core Metadata Element Set ver. 1.1[15]. This software allows to extend the schema with additional attributes and sub-attributes, as well as placing the elements in a random sequence. dLibra repositories are able to work online with other systems by using open communication standards, such as OAI-OMH protocols or RSS channels. dLibra system is supported by the latest version of the OAI-PHM protocol (www.openarchives.org/pmh/). Every digital item in dLibra system has its own unique access link and OAI tag.

Thanks to the open communication standards e-bUW resources are present in the Europeana collection. The users have the possibility to find e-bUW resources through various access points, such as, e.g.: Europeana (www.europeana.eu/portal/), KaRo (Distributed Catalogue of Polish Libraries) (http://karod.ukm.pl), or through BESE website (www.base-search.net), or last but not least through National Universal Central Catalogue NUKAT (www.nukat.edu.pl).

The central catalogue of scientific and academic libraries NUKAT is based on the method of co-cataloguing. It is the place where bibliographic description are created (according to ISBD in MARC 21) and as well the records of subject headings. Library of Warsaw University cooperates in cataloguing with NUKAT from its very beginning[16]. Still some of the bibliographic records are located only in the local catalogue[17, 18] while the library introduced the electronic catalogue in 1994 already and NUKAT was launched only in 2002[19].

The basis for the bibliographic description in the e-bUW is the record of the original document located in the online catalogue of the library. It provides the required date coherence for the catalogue is being supervised by the subject headings – a tool that the digital library does not use. We should mention, that part of the documents, for instance manuscripts, due to the formal reasons are not catalogued in NUKAT. If such
a document shall be directed into the digital library, it is being processed and catalogued in the local catalogue according to the rules accepted internally for each type of documents, basing mostly on book cataloguing standards.

Next, the metadata are imported from the local OPAC database to the Dublin Core base. The fundamental idea is to collect of such data that require little modification after importing them to a new base. Once the copy of the original document has been placed in the digital library, the former bibliographic record of the analogue is complemented with an URL address in NUKAT catalogue.

Every document published in e-bUW has its subject characteristic in Polish and in English language. The only exceptions are those documents that of the formal reasons are not being given subject headings. In Polish the characteristic is created according the subject headings language of KABA[20]. All the date given in the formal description are transferred into the integrated headings. The English subject headings are created according to the Library of Congress Subject Headings[21]. e-bUW has also set its own subject headings requirements for new kinds of documents, like rare prints, that hasn’t been given the subject heading before.

As a result of the whole digitization project the library decided that the original documents, that are now presented in digitized form, would be withdrawn from the public use. There are different strategies for different kinds of documents. What concerns periodicals, if the library owns several resources, none of those resources are presented. In case of the monographs only the digitized copy is presented. Subsequently, after placing a digital copy of a document in e – bUW and necessary modifications in the bibliographic record of the analogue document the user is given a suitable information in the local records of holdings/copy.

e-bUW among other digital collections
Digitizing processes in Poland are more and more popular. More and more scientific and cultural institutions are joining this current: universities, libraries, archives, museums and research centers. The greatest number among those institutions are libraries. The net of scattered digital libraries and repositories finds its common place in the Federation of Digital Libraries (FBC)[22]. At present, the federation consists of 89 initiatives[23]. The biggest institution is the Digital Library of Wielkopolska (WBC)[24], which has published 164,270 documents so far. This library is created in cooperation with several units. Some of the members of the federation are also smaller libraries, such as City and County Public Library in Kolbuszowa. The Warsaw University Library joined the federation in 2007 (www.buw.uw.edu.pl/index.php?option=com_content&task=view&id=465), as the number 17. Currently, our resources are ranked on the third place according to the size of the collection. The core of the collection, nearly 90 percent of e-bUW, are XIXth century periodicals. A similar tendency, that is digitizing periodicals, can be observed also in other digital libraries, especially in the bigger ones, such as Digital Library of Wielkopolska (WBC), Jagiellonian Digital Library (JBC), Digital Library of Malopolska (MBC) (see Figure 1).

Usage statistics and future plans
One could think, that the interest towards historical resources could be restricted to the narrow group of scientists. Still the survey done among the users of Polish digital libraries at the break of 2008/2009 denies this attitude. The result of the survey show
that 31 percent of the visitors of the digital libraries are people at the age between 20
and 30 year of life. Moreover it turned out that every fifth user was also a student, and
nearly the same number makes the group of academic workers.

The growing interest in resources placed in digital libraries is mirrored also in the
Osage statistics: Digital Library of Wielkopolska (WBC; online services since
1.10.2002) – 21,557,441 entries, Jagiellonian Digital Library (JBC; online services since
19.07.2010) – 2,173,181 entries, Digital Library of Warsaw University (e-bUW; online
service since 1.06.2007) – 6,339,461 entries, Digital Library of Malopolska (MBC; online
service since 1.12.2006) – 6,081,722 entries[25].

Till the 31 May 2012 e-bUW has published 100,233 documents (1,112,071 scans).
The whole Federation of Digital Libraries allows access to 986,237 documents. In 2011
e-bUW webpage was visited 2,934,959 times (in 2010 – 924,272 times).

The resources collected in e-bUW digital library are used mostly by people living in
Poland. But the interest among foreigners is also increasing. As a conclusion we would
like to quote two short abstracts from e-mails, received by the library:
I am extremely grateful for the publication on line of several directories, and even more for the publication of Sprawozdanie... Towarzystwa Wzajemnej Pomocy Kupców Wyznania Mojżeszowego Stol. M. Warszawy. I found my grand-father Berek (Bernard) ROTENSZTEJN, who died in Treblinka with the entire family, except my mother who went to France in 1938 as a student. There are very few traces of the existence of my grand-parents, so it was very precious to find his name in your publications. I hope more will follow. You are doing a great work.

And another letter:

Efforts like yours to put digital resources on the internet are very valuable to researchers like me sitting in Canada (and anywhere in the world). It is quite amazing that I can find so many useful resources in digital libraries in Poland, without leaving my house! Thanks for the advice and good luck with your work.

The words of gratitude from our users are very encouraging and, of course, allows us to strengthen our efforts in the future in order to enrich the collection continuously.

Notes
1. There are 17 libraries in Poland that receive the legal deposit. Most of them are academic libraries: National Library in Warsaw, Jagiellonian Library in Cracow, (both of them receive two copies of legal deposit), Maria Curie-Skłodowska University in Lublin, Łódź University Library, Toruń University Library, Adam Mickiewicz University Library in Poznań, Warsaw University Library, Wrocław University Library, Gdańska University Library, Catholic University in Lublin, University Library in Opole, University Library in Białystok, Pomeranian University Library in Szczecin, Silesian University Library in Katowice, Main Public Library in Warsaw, The National Cinematographic Library (it holds the legal deposit of films), Parliamentary Library.

2. Wirtualna Biblioteka Literatury Polskiej: http://literat.ug.edu.pl/books.htm#books


5. Library of Warsaw University has started the digitization work already in 2001. However it was mainly the material presented on CDs and only in the internal library web of ULTRANET.

6. The name was change afterwards to e-Collection of Polish Periodicals.


11. The list of the titles that are being digitized: http://autostrada.buw.uw.edu.pl/index.php/aktualnoci/lista-czasopism
12. The exact number of the scans was defined in the project – 3 000 000, the number of the publications is estimated. The numbers date to the end of May 2012.
15. The latest version of the software allows to switch to the meta data scheme PLMET: https://confluence.man.poznan.pl/community/display/FBCMETGUIDE/Wprowadzenie
17. The catalogue of the Libraries of Warsaw University http://opac.buw.uw.pl is currently created in cooperation of 39 library units.
18. One of the main goals of the NUKAT project (http://autostrada.buw.uw.edu.pl/), sponsored in frames of Innovative Economy Operative Program (2009-2013) is the consolidation of the local catalogue resources with NUKAT catalogue.
19. The first bibliographic records in NUKAT were created by Library of Warsaw University.
21. LCSH headings are given to all the documents located in the open stacks. All other documents with other localization in the library are not usually given the subject headings in English.
22. Federacja Bibliotek Cyfrowych: http://fbc.pionier.net.pl/
24. Wielkopolska Biblioteka Cyfrowa: www.wbc.poznan.pl/dlibra
25. Stand for 3 June 2012

Further reading
Lewandowska, A., Mazurek, C. and Werla, M. (2007), Federacja Bibliotek Cyfrowych w sieci PIONIER – Dostęp do otwartych bibliotek cyfrowych i repozytoriów, IV Ogólnopolska Konferencja EBIB Internet w bibliotekach Open Access, Toruń, 7-8 grudnia 2007 roku,


About the author
Barbara Chmielewska has been at the Library of Warsaw University since 2008. Her current position is Head of the Acquisition and Collection Development Department. She was born in 1978, and graduated from Warsaw University, with a Master’s degree in Russian Philology. Barbara Chmielewska is the corresponding author and can be contacted at: b.chmielewska@uw.edu.pl

Agnieszka Wrobel has been at the Library of Warsaw University since 2000. For the first six years she worked as a help-desk librarian in the sciences section. Since 2007 she has been working as an editor of the Digital Library of Warsaw University. She has also worked as a library information literacy tutor taking part in on-line courses for librarians.