Digital Library of Poland-related Old Ephemeral Prints: Preserving Multilingual Cultural Heritage

Maciej Ogrodniczuk
Institute of Computer Science
Polish Academy of Sciences
maciej.ogrodniczuk@gmail.com

Wlodzimierz Gruszczynski
Warsaw School of Social Sciences
and Humanities
wgruszczynski@swps.edu.pl

Abstract

The article presents the results from the project of the thematic Digital Library of Polish and Poland-related Ephemeral Prints from the 16th, 17th and 18th Centuries, intending to preserve the unique multilingual material, make it available for education and extend it with the joint efforts of historians, philologists, librarians and computer scientists.

The paper also puts forward the idea of a living digital library, created with a closed set of resources which can form the basis for their further extension, thus turning traditional digital archives into collaboration platforms.

1 Introduction

Nowadays digital libraries most likely tend to mirror traditional libraries. They collect and display resources as traditional librarians would do, perfectly embracing the new archiving capabilities, but too often stopping at this border. The thematic Digital Library of Polish and Poland-related Ephemeral Prints from the 16th, 17th and 18th Centuries (PL. *Cyfrowa Biblioteka Drukow Ulotnych polskich i Polski dotyczacych z XVI, XVII i XVIII w.* – hence and from here: CBDU, http://cbdu.id.uw.edu.pl) offers a new approach to the idea of a modern digital library by extending the conventional paradigm with using consistent sets of materials as an object pool for new collaboration tasks. In our case the prints digitized in the first step of the project were further analyzed and enhanced by experts co-operating on a digital content platform.

The work had been carried out within the project financed by the Polish state with intention to provide public online access to all preserved and described in the literature pre-press documents. Some of them have been preserved in the single copy, so their availability had been evidently restricted. In the course of the project the resources were gathered basing on the list of 2,000 bibliographical entries from Konrad Zawadzki's publication (Zawadzki, 1990) extended with objects described or discovered after its issue. Selected prints have been commented by historians, media experts and linguists to explain less known background details or presently unintelligible metaphors or symbols. Links have been created between related documents (e.g. translations and their alleged sources or derivates) to facilitate comparisons of similar materials. For 70% of the prints their images were obtained and made available in DjVu format. At the end of the project the revised edition of Zawadzki's work was published in electronic form.

Apart from this particular work plan, the ulterior idea was to test the concept of using the digital library as a collaboration platform for experts from different backgrounds basing on the assumption that real opening resources to the public means not just providing them for viewing and download, but preparing the environment to extend them in the immediate or more distant future. The technical challenge was to select the computer system which could serve this purpose best by mostly configuration and without too much tedious low-level programming. This goal was achieved with EPrints free repository software.

At present the library is actively used by historians and linguists (not to mention the students) who seem to benefit the most from the cooperation and further development of the resources. Their recent interests include adding transliterations (which already started with a new project based on the library materials) and using the print texts as source data for the searchable text corpus. Multilinguality of texts should also be shortly addressed since the first stage of work concentrated mostly on Polish

Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop, pages 27–33,
Hissar, Bulgaria, 16 September 2011.

and German documents.

The aspect of making the digital library a collaboration platform seems the most important in our study, putting less light into other important issues of creating thematic libraries, commenting historical content or digital library-based teaching. As such, creation of a digital library may be in most cases turned into a development project, being of benefit to the scientific community.

2 Origin, scope and limitations of the project

The project has been initiated in response to the need of permanent access to ephemeral prints from 16th-18th centuries by researchers coming from the academic teams preparing the historical dictionaries of Polish (Institute of Polish Language and Institute of Literary Research at Polish Academy of Sciences) as well as journalism and translation historians (Institute of Journalism, Institute of Applied Linguistics and Institute of German Studies at the Warsaw University) and students of journalism.

Temporal range of selected materials results from the long history of Polish journalism with three significant dates: 1501 – when the first known press materials in Polish appeared in print (the report on the anti-Turkish treaty signed by the Holy See and several European countries, including Poland, in Buda), 1661 – when the first regularly issued Polish newspaper "Merkuriusz Polski Ordynaryjny" came out, and 1729 – when "Nowiny Polskie" (Polish news) started regular circulation. The period in between was filled by ephemeral prints – disposable and occasional informative publications, playing an important part in the development of Polish writing, serving as a the most influential media for important news.

The scope of the materials is Poland-related, which combines the Polish sources with prints published abroad, concerning the Republic of Poland, political, religious and military issues (e.g. the reports on the famous relief of Vienna in 1683), but also sensational facts or canards. The materials were prepared "live", mostly by participants or observers of the reported events and as such they are valuable sociological source of information on mechanisms of spreading information at that time, its reception, propaganda and readers' interests.

The list of bibliographical data of prints complying with all above-mentioned requirements has been collated by Konrad Zawadzki in the 1970s and 1980s and was published as a three-volume work with the first volume issued in 1979 and the last one in 1990. It covers 1967 prints dating from 1501 to 1725, each described with the title (in modern transcription), issue date and place, printer name, format, volume size, information on bibliographical sources and an exact description of the title page (with line breaks, font names, illustrations etc.).

The originals of prints remain in various libraries over Europe, but at the period of preparing the bibliography many were successfully borrowed from their mother institutions to produce microfilmed copies to be included into the resources of The Polish National Library (Zawadzki's place of employment). As microfilmed resources are not the most comfortable ones to operate on a daily basis, usually their photocopies were used for research and teaching. The online era created a new possibility to interact with such resources (e.g. broaden the scope of materials showed to students) and resulted in applying for funds to create a digital library of metadata and images of prints enhanced with information useful for understanding the context of a given object.

The priority of making prints available for the daily work (disparate with the need of preservation of original objects) resulted in the assumption of building on the quality of microfilms rather than archiving originals scattered over many libraries in many countries. This also helped minimize costs and speed up the overall process, mostly also thanks to the presence of vast majority of materials at the National Library.

The source of funding was constrained to national (ministry) level and not to the consortium of libraries (owners of microfilmed print copies) knowing the three contradictory factors:

1. vast resources of large libraries, such as the Polish National Library,
2. their limited funds for digitization,
3. a policy (telling the truth, a reasonable one) of digitizing the most valuable resources first.

This combination can make many interesting materials wait for years to be made available at their owner's institution. Providentially, in March 2009 the 12-month project obtained the support

Figure 1: A sample bibliographical entry

1012. A letter from the king of Poland to his queen, in which is inserted many particulars relating to the victories obtained against the Turks. London, R. Baldwin [po 19 X] 1683. 2°. K. 1, sygn. A.

E. — Hos. 497. Sturm. 1920.

[Tytuł nagł., ant.:] A || LETTER || FROM THE || King of Poland || TO HIS QUEEN. || In which is Incerted || [kurs.:] Many Particulars Relateing to the Victories obtained || against the Turks. With a Prayer of the [ant.:] Turks [kurs.:] against || the [ant.:] Christians. || [linia] || Translated from the [kurs.:] Colonne [ant.:] Gazette, Octobr. 19. 1683. [kurs.:] Numb. 84. || [linia] ||

[Kolofon na k. A v., kurs.:] London, [ant.:] Printed for [kurs.:] R. Baldwin, [ant.:] in the [kurs.:] Old-Bailey. 1683. ||

List króla Jana III Sobieskiego do królowej Marii Kazimiery o zwycięstwie wojsk sojuszniczych nad armią turecką pod Wiedniem 12 IX 1683.

Tekst polski zob. poz. 1005-1007, 1659.

Egz.: WStBibl. Wien
Mikrofilm: BN Mf. 62115

from the Ministry of Culture and National Heritage and the Foundation for the Development of Journalism Education.

3 Towards the thematic digital library

The project team was headed by Wlodzimierz Gruszczynski and joined forces of computer scientists (Maciej Ogrodniczuk, Jakub Wilk), historians (Adam Kozuchowski), philologists (Ewa Gruszczynska, Anna Just, Dorota Lewandowska-Jaros, Katarzyna Jasinska-Zdun) and librarians (the team of Maria Piber), coordinated by Grazyna Oblas.

The process started with scanning Zawadzki's bibliography in the format sufficient to automated OCR processing of the text (200 dpi, greyscale, lossless compression of the result files, see Fig. 1). The images have been read with ABBYY FineReader 9 with support for several modern languages turned on (French, German, Latin, Polish) since texts could contain transcriptions of names in various (modern) languages.

As a result, a recognized plain text of the bibliography has been obtained, covering not only full bibliographical entries (with additional comments, location information etc.), but also all front and back matter data (volume introductions, errata, foreign-language abstracts etc.) to be reused in the electronic edition of the bibliography. Plain text version was used to expedite the task of extraction of individual fields of each entry regardless of its initial inconsistent formatting (which was easily introduced at a later stage, basing on the entry structure). The text has been saved in UTF-8 character encoding to seamlessly represent all diacritics.

Perl scripts have been implemented to split bibliographical entries into individual data records according to the description of fields retrieved from the bibliography introduction (full and short title, information about author, publisher, format etc.) In fact, the field list was designed to be more specific than the original to support verification of content (e.g. the model of a list of copies with subfields storing library name and a catalogue number was introduced against the original composite string value). The field set has been later used as a basis of the target model for the computer system storing the library data with new fields describing the project-specific properties going beyond the traditional bibliographical entry (e.g. commentaries of an object). The records were then verified with regular expression patterns testing their contents (e.g. publication date standard format) and content of fields used in further steps (e.g. microfilm catalogue numbers) extracted separately and additionally verified.

As the majority of microfilms were available at the National Library, the preparation of scans has been commissioned mainly to this institution, after obtaining formal permission to use the resulting electronic documents on the project site, ultimately available to the general public. Since

the project duration was relatively short, the scans were produced in batches to let contributors start work on the previous portion when the following one was still in preparation (which sometimes required cleaning the microfilm plates or seeking improperly catalogued collection). Similarly, all other project phases (metadata proofreading, import and conversion of scanned materials, preparation of commentaries and dictionaries etc.) were also being carried out simultaneously.

Among many available repository systems for digital libraries, including a prominent (in Poland) dLibra (dLibra, 2010), EPrints has been selected as the target storage and publication framework. EPrints (EPrints, 2010) is a free, GPL-licensed multiplatform repository software developed since 2000 at the University of Southampton. Its open source origin presents a major benefit for projects intending not only to deploy it "as is", but seeking possibilities to extend it with new solution-specific functionalities. The system has been installed, configured and in most respects translated into Polish (with translations made available to the community), setting up the print repository.

As already stated above, an important organizational assumption made at the beginning of the project was to use EPrints also as a collaboration platform, starting from the very first phases of the work. Following this statement, the bibliographical entries (henceforth, metadata descriptions) were imported into EPrints even before they were finally proofread. This stage has been carried out already in the system, using the workflow defined for the project. After metadata has been revised against the image (or paper) version of the bibliography, the scanned images of prints were converted into DjVu format and uploaded into the library (each object corresponding to one file).

Versions of objects (most likely translations or alterations of the base text) have been linked basing on information available in the bibliography or detected by the experts. For materials only reported in literature, when the original is unknown or not preserved, the information on the base language has been provided.

Going beyond simply making the objects available as electronic versions of the bibliographic entries with scanned copies, the repository model has been extended with new fields storing the new value created by the project: historical commentaries relating to facts and people described in the material, media-historical or language-historical observations, translations of then common Latin interjections and translations of foreign texts into Polish or local dictionaries. Such approach seems novel in the design of digital libraries. What is more, the system creates possibility to form a living thematic information exchange environment around the library site, making it possible to store expert comments, versions of materials etc.

To make the scope of description complete and up-to-date, a survey of the library holdings has been carried out. The annual volumes of library professional journals published by ten major scientific libraries in Poland has been investigated and 80 new objects (not included in Zawadzki's bibliography) discovered. Since the project was short of funds to generate their scanned versions, only their metadata were added to the library.

The last phase of the project was preparation of the supplemented electronic edition of Zawadzki's bibliography which illustrates how digital content can help maintain traditional publications. The electronic version was intended to be published, but cuts in the initial project budget forced the team to leave this additional step to future projects.

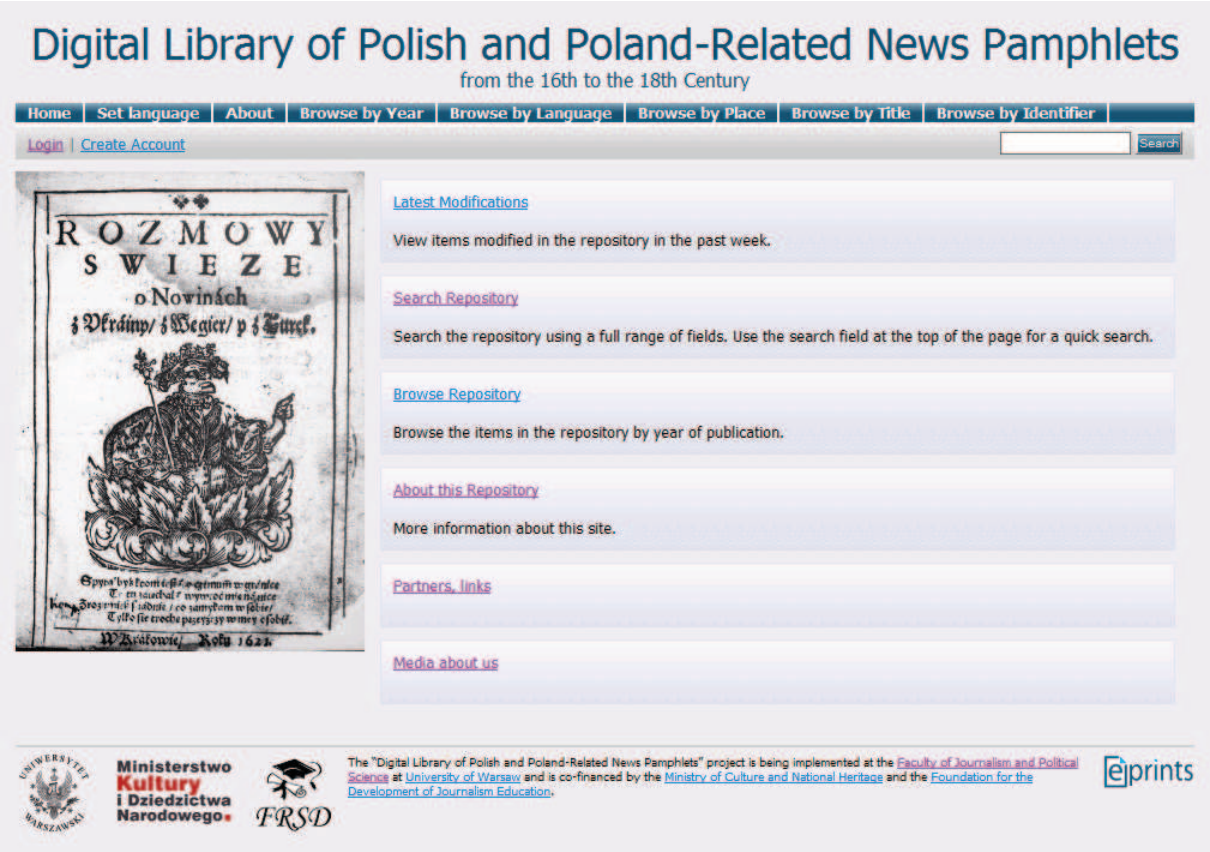
The new publication layout has been created in LateX. To facilitate usage, in contrast to the original work, separation into three volumes was not preserved and a single volume was produced. Introductions, lists of printers and print shops as well as back matter illustrations were taken over from the OCR-ed source and collated. All materials transferred to the digital library have been exported into XML format and were included in the final edition. This means that all supplements and errata which were merged with the library objects were automatically corrected. Back matter content such as lists of titles, people names and geographical names were not transferred from the original work but were regenerated from the library.

All scripts and transformations created throughout the project were preserved to facilitate generation of new electronic editions of the bibliography in case new errors are reported by the library users or new materials included. This also demonstrates new collaborative approach to preparation of electronic publications

4 Library interface, statistics and usage

The repository is located on the server of the Institute of Journalism of the University of Warsaw and

Figure 2: Library homepage



the library has been made available at `http://cbdu.id.uw.edu.pl`[1]. The site (see homepage on Fig. 2) allows typical browse and search interfaces in Polish or English. Objects can be browsed according to multiple criteria, including those normally used in bibliographical descriptions (mostly borrowed from Zawadzki), as well as those less typical: thematic, genre-related and other. The native EPrints advanced search is supplemented with a recently customized simple search with a Google-like single text field.

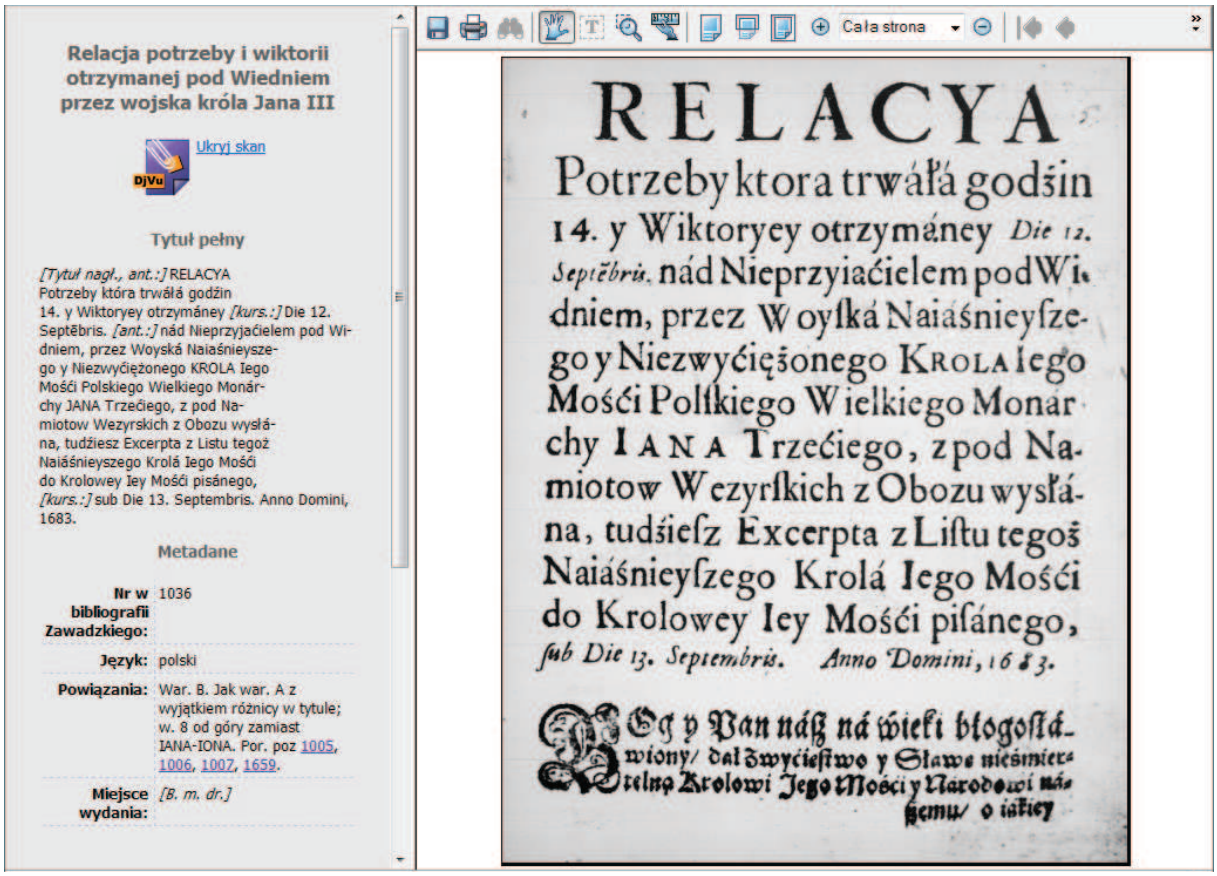
Prints relating to the same facts are hyperlinked, especially those that are most likely or certainly translations from other texts available in the library. Taking into account the number of identified dependencies and benefits resulting from possibility of their visual comparison, the system offers a function to open related documents in a split window. Moreover, the new implementations include enhanced inter-metadata linking capabilities, new facets for repository browsing and a list of recently revised objects replacing the standard list of recently added.

Currently the library holds 2009 objects. 1404 objects have scans attached (with 11 585 pages in total). 11 languages are represented. The languages with widest coverage (over 50 objects) are: German (797 objects), Polish (325 objects – see Fig. 3), Italian (180) and Latin (69); the remaining ones are Swedish, French, Spanish, English, Dutch, Czech and Danish. Around 200 prints in Polish and 50 in German have attached dictionaries explaining currently unused words or phrases (giving explanations in contemporary Polish or German). Latin dictionaries are also included.

The final version of the library has been made official early 2010, so it has been more than half a year since its resources can be used for interested parties and some initial findings from observed usage of the library resources can be obtained. Data from March-November 2010 show that the library hosted 34 unique visitors daily (47 visits) and is stabilizing; approx. 40% of the open pages are DjVu files (not indexed by popular search engines which should imply human visitors).

[1]The default language of the interface is Polish; it can be changed to English by selecting the second menu option ("Ustaw jezyk") and then the first entry on the language list (English – "angielski").

Figure 3: A sample object



The library has been included into the network of Polish Digital Libraries Federation by means of The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH, 2002) with metadata represented in Dublin Core (ISO, 2009). OAI-PMH is natively supported by EPrints, so no additional implementation was necessary.

To advertise the library among the general public a series of dissemination events have been organized. They included radio broadcasts (university radio Kampus, Polish news radio TOK FM), newspaper articles and historical portal news.

5 Further steps and conclusions

Recently the library contents found new applications: it is currently used in the EU-co-funded IMPACT project for training Gothic script OCR software by ABBYY (the producer of FineReader with XIX module for recognition of Fraktur or "black letter" texts). Simultaneously the transcribed versions are being prepared and are planned to be included in the library (respective metadata fields were defined in the project, but only one transcription was filled in since it exceeded the scope of the project). Other project which can benefit from the library resources, currently being carried out by the Institute of Polish Language at the Polish Academy of Sciences is the dictionary of the historical Polish from 17th and the first half of 18th century (see SXVII, 2010).

Another interesting direction is broadening of the scope of materials the library covers. Since the project has been closely related to the resources of The Polish National Library, it can be extended to cover materials coming from other libraries, most likely from abroad, both preserving originals and storing their microfilmed copies. With the possibility of preparation of copies on site, the costs should not significantly differ from the costs of the national project. Starting with Zawadzki's bibliography, there are still around 200-300 objects to be acquired this way. The geographical key seems an important factor here: the limitations imposed on Zawadzki before 1989 resulted in underrepresentation of resources from the libraries of the countries belonging to the former Soviet Union. Last but not least, Vatican archives can prove to be one of the most important source of materials of the

described type, with so far limited availability.

Apart from obvious development ideas such as widening the scope of description of gathered objects or deepening the analysis, the idea of creating a collaboration platform of the digital library site can be followed. For instance, the system can be extended with new functionalities provided to the library users (not just experts with editing rights) such as adding comments to materials or an integrated forum. Such add-ons can prove similarly efficient in development of the materials and in related projects.

Despite its small scale, we trust that our library will prove equally useful for old-Polish researchers as much larger heritage accessibility projects such as Europeana (see `http://www.europeana.eu`) or ENRICH (European Networking Resources and Information concerning Cultural Heritage, see `http://enrich.manuscriptorium.com/`).

References

dLibra (2010). *Digital Library Framework*. Poznan Supercomputing and Networking Center affiliated to the Institute of Bioorganic Chemistry of the Polish Academy of Sciences.

EPrints (2010). *EPrints free software*. Southampton: School of Electronics and Computer Science at the University of Southampton.

Gruszczynski W., Ogrodniczuk M. (2010). *Cyfrowa Biblioteka Drukow Ulotnych Polskich i Polski dotyczacych z XVI, XVII i XVIII w. w nauce i dydaktyce* (*Digital Library of Poland-related Old Ephemeral Prints in research and teaching*, in Polish). In Proceedings of "Polskie Biblioteki Cyfrowe 2010" (*Polish Digital Libraries 2010*) conference, Poznan, 18-22 October 2010.

ISO 2009: *ISO 15836:2009. Information and documentation – The Dublin Core metadata element set*.

OAI-PMH (2002). *The Open Archives Initiative Protocol for Metadata Harvesting*. Protocol Version 2.0 of 2002-06-14, Document Version 2008-12-07. http://www.openarchives.org/OAI/openarchivesprotocol.html.

SVXI (1987). *Slownik polszczyzny XVI w.* (*Dictionary of 16 century Polish*, in Polish). Krakow: Instytut Badan Literackich PAN.

SVXII (2010). *Slownik jezyka polskiego XVII i 1. polowy XVIII w.* (*Dictionary of 17 century and 1st half of 18 century Polish*, in Polish). Warszawa: Polska Akademia Nauk, Instytut Jezyka Polskiego.

Zawadzki, K. (1990). *Gazety ulotne polskie i Polski dotyczace z XVI, XVII i XVIII wieku* (*Polish and Poland-related Ephemeral Prints from the 16th-18th Centuries*, in Polish). Wroclaw: Zaklad Narodowy im. Ossolinskich, Wydawnictwo Polskiej Akademii Nauk.