Digital Library 2.0 — Source of Knowledge and Research Collaboration Platform

Włodzimierz Gruszczyński[1], **Maciej Ogrodniczuk**[2]
[1]Warsaw School of Social Sciences and Humanities
Chodakowska 19/31, Warsaw, Poland
wgruszczynski@swps.edu.pl

[2]Institute of Computer Science
Polish Academy of Sciences
Jana Kazimierza 5, Warsaw, Poland
maciej.ogrodniczuk@ipipan.waw.pl

Abstract

Digital libraries are frequently treated just as a new method of storage of digitized artifacts, with all consequences of transferring long-established ways of dealing with physical objects into the digital world. Such attitude improves availability, but often neglects other opportunities offered by global and immediate access, virtuality and linking — as easy as never before.

The article presents the idea of transforming a conventional digital library into knowledge source and research collaboration platform, facilitating content augmentation, interpretation and co-operation of geographically distributed researchers representing different academic fields. This concept has been verified by the process of extending descriptions stored in thematic *Digital Library of Polish and Poland-related Ephemeral Prints from the 16th, 17th and 18th Centuries* with extended item-associated information provided by historians, philologists, librarians and computer scientists. It resulted in associating the customary fixed metadata and digitized content with historical comments, glossaries of foreign interjections or explanation of less-known background details.

Keywords: digital libraries, old-Polish prints, historical corpus

1. Introduction

Apart from serving the needs of the linguistic community, *The Digital Library of Polish and Poland-related Ephemeral Prints from the 16th, 17th and 18th Centuries* project (PL. *Cyfrowa Biblioteka Druków Ulotnych polskich i Polski dotyczących z XVI, XVII i XVIII w.*, CBDU: `http://cbdu.id.uw.edu.pl`) was from the very beginning planned to be set up as a proof of concept validating the idea of making a thematic digital library a platform for co-operation of researchers with the possibility of storing the research results very closely related to the original digital object.

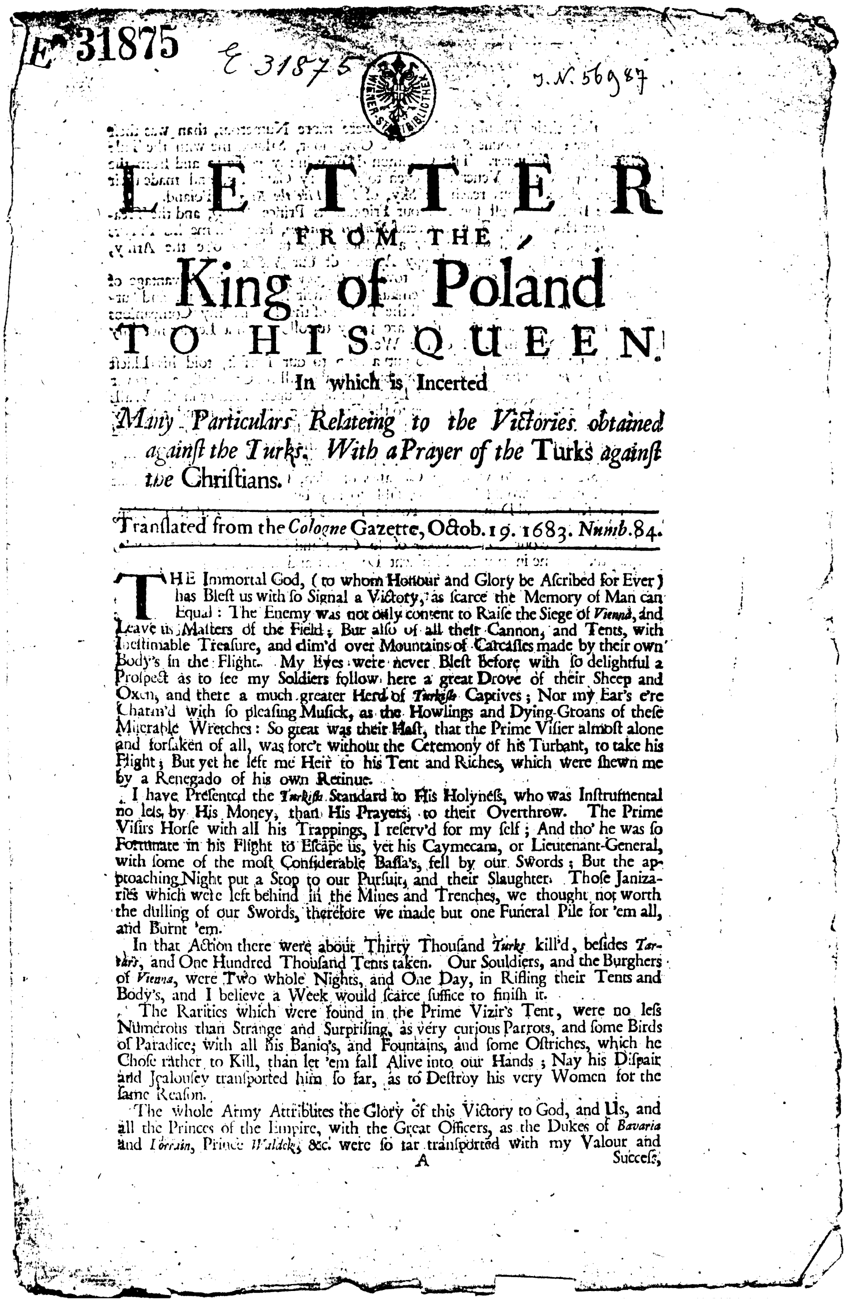
CBDU comprises about 2,000 pre-press documents dated between 1501 and 1729. At that time, before Polish newspapers started regular circulation, ephemeral prints — short, disposable and occasional informative publications — were playing a significant part in the development of Polish writing, serving as the most influential media of the time. 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 or letters on the famous relief of Vienna in 1683, see e.g. Figure 1), 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. This variety is reflected in diversity of topics, layouts and language variants.

The digital library was created with the intention of providing public online access to all preserved and described in the literature pre-press documents. The prints were identified, scanned from microfilms (since the original specimens are preserved in libraries all over Europe), converted into DjVu (Bottou et al., 1998) format and saved in a EPrints-based system (EPrints, 2010) (see (Ogrodniczuk and Gruszczyński, 2011) for a more detailed description of the process). Currently over 70% of the prints are available as DjVu images.

2. From Storage to Knowledge About Objects

Obviously, for researchers the most important goal of the digital library is making available scanned versions of objects (images, not text) with best possible quality — and this requirement is fulfilled by virtually all digital libraries, including CBDU. What is worthy of attention is that this goal is final for most thematic digital libraries managing manuscripts and old prints. This makes a very serious limitation, making it impossible to e.g. automatically search for textual content or reuse (copy-paste) fragments in further research. For less experienced users from the general public it is even more hermetic due to the sheer form of the old print with its fonts (Schwabacher, Fraktur), no longer used abbreviations etc. Language and time barriers create even more problems — our library comprises foreign-language prints, but even old Polish prints can contain many archaic words and syntactic constructs which hinder understandability of the text. Foreign-language quotations, particularly Latin, pose another challenge.

We believe that the thematic digital library should offer far



LETTER FROM THE King of Poland TO HIS QUEEN.

In which is Incerted

Many Particulars Relateing to the Victories obtained againſt the Turks. With a Prayer of the Turks *againſt the* Chriſtians.

Tranſlated from the *Cologne* Gazette, Octob. 19. 1683. *Numb.* 84.

THE Immortal God, (to whom Honour and Glory be Aſcribed for Ever) has Bleſt us with ſo Signal a Victory, as ſcarce the Memory of Man can Equal: The Enemy was not only content to Raiſe the Siege of *Vienna*, and Leave us Maſters of the Field; But alſo of all their Cannon, and Tents, with Ineſtimable Treaſure, and clim'd over Mountains of Carcaſſes made by their own Body's in the Flight. My Eyes were never Bleſt Before with ſo delightful a Proſpect as to ſee my Soldiers follow here a great Drove of their Sheep and Oxen, and there a much greater Herd of *Turkiſh* Captives; Nor my Ear's e're Charm'd with ſo pleaſing Muſick, as the Howlings and Dying-Groans of theſe Miſerable Wretches: So great was their Haſt, that the Prime Viſier almoſt alone and forſaken of all, was forc't without the Ceremony of his Turbant, to take his Flight; But yet he left me Heir to his Tent and Riches, which were ſhewn me by a Renegado of his own Retinue.

I have Preſented the *Turkiſh* Standard to His Holyneſs, who was Inſtrumental no leſs by His Money, than His Prayers, to their Overthrow. The Prime Viſirs Horſe with all his Trappings, I reſerv'd for my ſelf; And tho' he was ſo Fortunate in his Flight to Eſcape us, yet his Caymecam, or Lieutenant-General, with ſome of the moſt Conſiderable Baſſa's, fell by our Swords; But the approaching Night put a Stop to our Purſuit, and their Slaughter. Thoſe Janizaries which were left behind in the Mines and Trenches, we thought not worth the dulling of our Swords, therefore we made but one Funeral Pile for 'em all, and Burnt 'em.

In that Action there were about Thirty Thouſand *Turks* kill'd, beſides *Tartars*, and One Hundred Thouſand Tents taken. Our Souldiers, and the Burghers of *Vienna*, were Two whole Nights, and One Day, in Rifling their Tents and Body's, and I believe a Week would ſcarce ſuffice to finiſh it.

The Rarities which were found in the Prime Vizir's Tent, were no leſs Numerous than Strange and Surpriſing, as very curious Parrots, and ſome Birds of Paradice; with all his Baniq's, and Fountains, and ſome Oſtriches, which he Choſe rather to Kill, than let 'em fall Alive into our Hands; Nay his Diſpair and Jealouſey tranſported him ſo far, as to Deſtroy his very Women for the ſame Reaſon.

The whole Army Attributes the Glory of this Victory to God, and Us, and all the Princes of the Empire, with the Great Officers, as the Dukes of *Bavaria* and *Lorrain*, Prince *Waldek*, &c. were ſo far tranſported with my Valour and

A Succeſs,

Figure 1: An English abridged translation of a letter of John III Sobieski, King of Poland and Grand Duke of Lithuania, to his wife on the Battle of Vienna (see `http://cbdu.id.uw.edu.pl/10120/`).

more: different ways of accessing content, suitable for various groups of visitors with different abilities, knowledge and reasons for using the library.

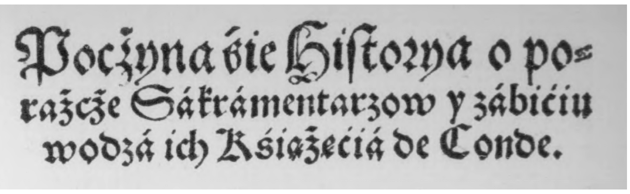
2.1. Augmented Content Representation

The above-mentioned reasons influenced our decision that CBDU prints will be available not only as scanned versions of original objects, but also in other, related forms such as transliterated or transcribed text. In the future we also plan to offer translations of foreign-language prints into Polish. Each of these versions will be synchronized with all other layers so that they can be viewed simultaneously when needed.

To clarify this scenario, let's assume that the user decided to see the scanned version first (see Figure 2). It apparently contains letters or words difficult for an inexperienced user. However, the system offers the transliteration possibility, which yields:

> *Poczyna śie Hiſtorya o porażcże Sákrámentarzow y zábićiu wodzá ich Kśiążęćiá de Conde.*

Such form is helpful, but still not very reader-friendly. For users unfamiliar with Polish who would like to get rough understanding of the content, it is also insufficient since e.g. automated translation into English does not recognize many words belonging to old Polish:



Poczyna śie Hiſtorya o po=
rażcże Sákrámentarzow y zábiciu
wodzá ich Kśiążęciá de Conde.

Figure 2: First sentence of a sample print in old Polish from 1569 (see `http://cbdu.id.uw.edu.pl/700/`).

> *The History conceives of porażcże Sákrámentarzow killing their leader s prince de Conde.*

Such problems are overcome when transcription of text is offered (by default — to contemporary Polish). The following version is almost perfectly understandable by an ordinary user of Polish:

> *Poczyna się historia o porażce sakramentarzów i zabiciu wodza ich Książęcia de Conde.*

and can be further automatically translated:

> *Begins the story of the failure sakramentarzów and killed their chief prince de Conde.*

The last persisting problem is the inflected outdated word *sakramentarz* (meaning *heretyk* — En. *heretic*), no longer used in contemporary Polish. Such words can be, on user's request, explained and presented in the form ready to be further processed, e.g. translated. Our example, after the new inflected form is applied to the transcribed text, makes the following:

> *Poczyna się historia o porażce heretyków i zabiciu wodza ich Książęcia de Conde.*

which can be understood not only by an ordinary Polish speaker, but also for a machine translation system:

> *Begins the story of the failure of heretics and killing their leader the prince de Conde.*

Such different textual versions synchronized with image is a valuable source of information for scientific work, e.g. creation of historical dictionaries (see e.g. (Mayenowa et al., 1966 2012)).

2.2. Interpretation and Collaboration Platform

The process of augmenting the representation has already been started in CBDU. Currently the 300 dpi scanned versions of prints have been supplemented with rich metadata (see Figure 3). Apart from fields considered standard (title transliteration, name of the font used, precise information about the original object and its known copies etc.), the library also contains descriptions going beyond the usual understanding of metadata, thus moving towards interpretation. These set of fields, also represented as metadata, are used to store e.g. additional information about the content, historical comments (concerning facts or people described by the print), relations between library objects (translations, adaptations, alterations of the base text, their alleged sources or derivates etc.) or glossaries providing translations of selected outdated terms or foreign interjections. Several prints have been also commented by media experts and linguists to explain less known background details or presently unintelligible metaphors or symbols.

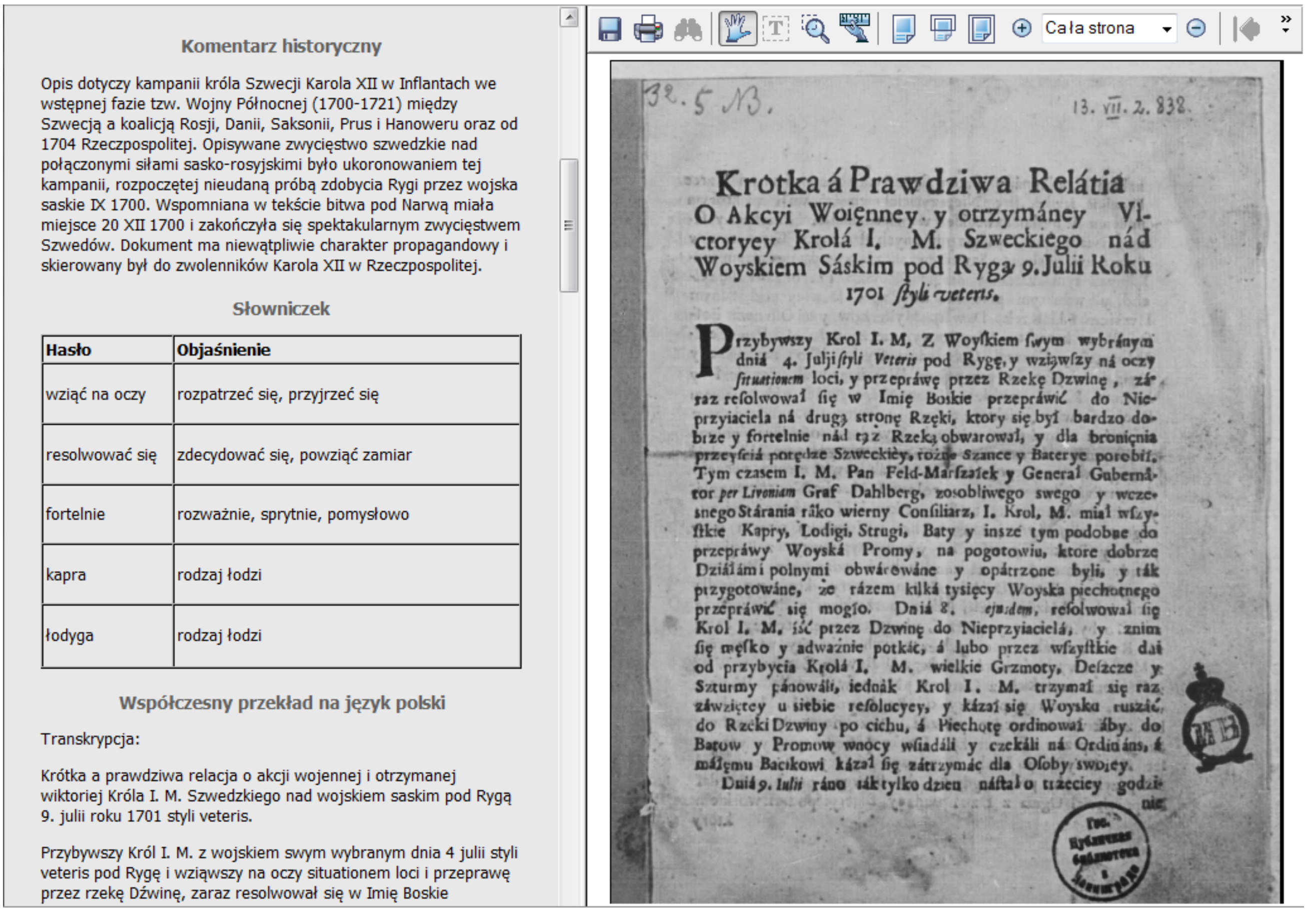
To achieve that, the library installation was from the very beginning used as a collaboration environment for historians, philologists, librarians and computer scientists, allowing them to jointly work out and extend metadata content and correct identified omissions on the fly. Proofreading of metadata, linking of objects and storing the newly created specialized print-related content were also carried out in the system, using the workflow defined for the project. What is more, additional fields were used to store expert comments, versioning information etc.

2.3. Current Work

Two new directions exploiting the cross-domain, constant-update approach of the library have been followed recently by establishing new project synergies. As for transliteration of the content, initially it was stored in a separate metadata field and was filled in for a single print, for verification only. After possibilities offered by DjVu format for similar content have been recently explored in the IMPACT project (see (Bień, 2011)), the hidden text layer, usually containing the results of Optical Character Recognition, is planned to be used to store the transliteration along with the coordinates of corresponding words on the scanned image. Creating synergy between CBDU and IMPACT, which used the library data for testing OCR algorithms, we plan to supplement the CBDU repository of prints with OCR-ed IMPACT data.

Transliterated text with word coordinates would also be used to improve user-friendliness of the interface. Glossaries explaining more difficult terms (see left window in Figure 3), currently stored as separate files and displayed in its entirety, will be used as sources of shorter explanatory notes shown on user demand in tooltips after pointing the term with a mouse (similarly to triggering thesaurus definitions in Microsoft Word or Open Office). All inflectional forms of a term will be automatically detected (cf. (Mykowiecka et al., 2012)). A similar paragraph-based mechanism will be used for presenting text translations. After pointing a foreign, e.g. Latin fragment on a scan, users will be able to display its transliteration and/or translation into Polish.

The above-stated approach will be applied in a recently started project of an electronic corpus of 17th and 18th century Polish texts (see `http://clip.ipipan.waw.pl/KORBA` for details). It intends to be the first stage of supplementing the National Corpus of Polish (Pol: Narodowy Korpus Języka Polskiego, NKJP, see `http://nkjp.pl/` and (Przepiórkowski et al., 2012)) with old Polish content and covers the so-called middle Polish content, exactly from the years 1601–1772. The corpus will feature annotation for text structure and language (all foreign elements, e.g. Latin intrusions, will be distinguished), and a portion of it will also feature morphological annota-



Komentarz historyczny

Opis dotyczy kampanii króla Szwecji Karola XII w Inflantach we wstępnej fazie tzw. Wojny Północnej (1700-1721) między Szwecją a koalicją Rosji, Danii, Saksonii, Prus i Hanoweru oraz od 1704 Rzeczpospolitej. Opisywane zwycięstwo szwedzkie nad połączonymi siłami sasko-rosyjskimi było ukoronowaniem tej kampanii, rozpoczętej nieudaną próbą zdobycia Rygi przez wojska saskie IX 1700. Wspomniana w tekście bitwa pod Narwą miała miejsce 20 XII 1700 i zakończyła się spektakularnym zwycięstwem Szwedów. Dokument ma niewątpliwie charakter propagandowy i skierowany był do zwolenników Karola XII w Rzeczpospolitej.

Słowniczek

Hasło	Objaśnienie
wziąć na oczy	rozpatrzeć się, przyjrzeć się
resolwować się	zdecydować się, powziąć zamiar
fortelnie	rozważnie, sprytnie, pomysłowo
kapra	rodzaj łodzi
łodyga	rodzaj łodzi

Współczesny przekład na język polski

Transkrypcja:

Krótka a prawdziwa relacja o akcji wojennej i otrzymanej wiktoriej Króla I. M. Szwedzkiego nad wojskiem saskim pod Rygą 9. julii roku 1701 styli veteris.

Przybywszy Król I. M. z wojskiem swym wybranym dnia 4 julii styli veteris pod Rygę i wziąwszy na oczy situationem loci i przeprawę przez rzekę Dźwinę, zaraz resolwował się w Imię Boskie

32.5 N3.

13. VII. 2. 838.

Krotka á Prawdziwa Relátia
O Akcyi Woienney y otrzymáney Victoryey Krolá I. M. Szweckiego nád Woyskiem Sáskim pod Rygą 9. Julii Roku 1701 *ſtyli veteris.*

Przybywszy Krol I. M, Z Woyſkiem ſwym wybránym dniá 4. Julji *ſtyli Veteris* pod Rygę, y wziąwſzy ná oczy *ſituationem* loci, y przepráwę przez Rzekę Dzwinę, zá raz reſolwował ſię w Imię Boskie przepráwić do Nieprzyiaciela ná drugą stronę Rzęki, ktory się był bardzo dobrze y fortelnie nád tąż Rzeką obwarował, y dla bronięnia przeyſciá potędze Szweckiey, rożne Szance y Baterye porobił. Tym czasem I. M. Pan Feld-Marſzałek y General Gubernátor *per Livoniam* Graf Dahlberg, z osobliwego swego y wczesnego Stárania ráko wierny Conſiliarz, I. Krol, M. miał wſzyſtkie Kapry, Lodigi, Strugi, Baty y insze tym podobne do przepráwy Woyská Promy, na pogotowiu, ktore dobrze Działámi polnymi obwárowáne y opátrzone byli, y ták przygotowáne, że rázem kilká tysięcy Woyska piechotnego przepráwić się mogło. Dniá 8. *ejusdem*, reſolwował ſię Krol I. M. iść przez Dzwinę do Nieprzyiaciela, y znim ſię męſko y adważnie potkác, á lubo przez wſzyſtkie dni od przybycia Krolá I. M. wielkie Grzmoty, Deſzcze y Szturmy pánowáli, iednák Krol I. M. trzymał się raz záwziętey u siebie reſolucyey, y kázał się Woysku ruszáć do Rzeki Dzwiny po cichu, á Piechotę ordinował áby do Batow y Promow wnocy wſiadáli y czekáli ná Ordinans, á máłęmu Bacikowi kázał ſię zátrzymác dla Oſoby swoiey.

Dniá *9. Iulii* ráno ják tylko dzien náſtał o trzeciey godzinie

Figure 3: Example of augmented content.

tion. Mutually, the texts transcribed in the course of this project will bring new content to CBDU.

We are aware that CBDU resources are still far from completeness. Zawadzki's bibliography (Zawadzki, 1990) used as a source for object selection in the first phase of our original project does not cover all preserved Polish and Poland-related ephemeral prints. Thanks to development of digital libraries and common access platforms (represented by Digital Libraries Federation[1] in Poland and EUROPEANA[2] in Europe) supplementing the material became relatively easy. All newly identified objects are successively acquired by our library by means of storing their metadata and references to external sources. Mirroring content still (unfortunately) requires separate arrangement with libraries preserving the originals.

3. Conclusion

Although extending descriptions of stored objects with non-standard interpretation layers goes beyond the present understanding of a "classical" digital library, we have proved that this is the step in the right direction. Whenever a digital library is not intended to purposefully reflect the physical one, the "new type" library can be created, ready to store content closely related to objects (such as relevant metadata-rich references to other digital libraries, also offered by our project), making partial critical editions and offering fragment translations.

4. References

Bień, J. S. (2011). Efficient search in hidden text of large DjVu documents. In *Advanced Language Technologies for Digital Libraries. Lecture Notes in Computer Science (Theoretical Computer Science and General Issues)*, volume 6699, pages 1–14. Springer.

Bottou, L., Haffner, P., Howard, P. G., Simard, P., Bengio, Y., and Lecun, Y. (1998). High Quality Document Image Compression with DjVu. *Journal of Electronic Imaging*, 7:410–425.

EPrints. (2010). *EPrints Manual*. School of Electronics and Computer Science at the University of Southampton. http://wiki.eprints.org/w/EPrints_Manual.

Mayenowa, M. R., Pepłowski, F., and Mrowcewicz, K., editors. (1966–2012). *Dictionary of 16th century Polish (Pol. Słownik polszczyzny XVI w.), in Polish*, volume 1–36. National Ossoliński Institute and The Institute of Literary Research of the Polish Academy of Sciences, Wrocław and Kraków.

Mykowiecka, A., Rychlik, P., and Waszczuk, J. (2012). Building an electronic dictionary of old polish on the base of the paper resource. In *Proceedings of the Work-*

[1] Pol. Federacja Bibliotek Cyfrowych, see http://fbc.pionier.net.pl/owoc.

[2] http://www.europeana.eu/

shop on Adaptation of Language Resources and Tools for Processing Cultural Heritage at LREC 2012, pages 16–21. European Language Resources Association.

Ogrodniczuk, M. and Gruszczyński, W. (2011). Digital Library of Poland-related Old Ephemeral Prints: Preserving Multilingual Cultural Heritage. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 27–33, Hissar, Bulgaria. http://www.aclweb.org/anthology/W11-4105.

Przepiórkowski, A., Bańko, M., Górski, R. L., and Lewandowska-Tomaszczyk, B., editors. (2012). *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish].* Wydawnictwo Naukowe PWN, Warsaw.

Zawadzki, K. (1990). *Polish and Poland-related Ephemeral Prints from the 16th-18th Centuries (Pol. Gazety ulotne polskie i Polski dotyczące z XVI, XVII i XVIII wieku), in Polish.* National Ossoliński Institute, Polish Academy of Sciences, Wrocław.