11th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2018. ISSN 2336-5021. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings/**

THE REFLECTION OF LITERARY ACTIVITIES IN DIGITAL SPACE

Pavla Hartmanová

hartmanova@ucl.cas.cz

Institute of Czech Literature, Czech Academy od Sciences

Paulina Czwordon-Lis

paulina.czwordon-lis@ibl.waw.pl

The Institute of Literary Research of the Polish Academy of Sciences, Poland

Abstract

The Czech Literary Bibliography comprises a set of bibliographical records which reflect cultural journalism and specialist texts on Czech literature. The aim of the contribution is introduction to a new project of the Institute of Czech Literature: The Czech Literary Internet. The project has extended our sources to excerpt platforms, web pages and electronic magazines whose content is not easily searchable through classic search engines. It turns out that this resource illustrates the professional debate on literary events and development and, in particular, brings new information on culture in regions and popular literature.

Keywords

Czech literary bibliography, Czech literary internet, databases, Webarchiv, archiving, search engine, Czech literary life, Polish literary bibliography, literary blogs

Introduction

Česká literární bibliografie (Czech Literary Bibliography – CLB) is a specialized analytical bibliography acquired at the Institute of Czech Literature of the Czech Academy of Sciences. Timewise, its database covers the period from the final third of the 18th century to the present and serves the needs of basic research of literature and literary life in the Czech lands. The bibliography holds almost 600,000 articles processed in standardized MARC21 format and around 1.6 million excerpts in the form of a digitised card index for the years 1770 to 1945.

The Polish Literary Bibliography (PBL), created by Pracownia Bibliografii Bieżącej (Department of Current Bibliography) operates at the Poznan department of the Institute of Literary Research of the Polish Academy of Sciences (Instytut Badań Literackich Polskiej Akademii Nauk). It has, for 70 years (since 1948), collected data from the sphere of Polish and foreign literature and literary theory and about Polish theatre and film. Processed data from 1944 to 1988 are made available through printed yearbooks (with an estimated total of 1.8 million records), data from 1989 to 2003 in the form of an electronic database which currently holds around 700,000 records. A card index which covers the 19th century and the 1st half of the 20th century is stored at the Warsaw branch of IBL and holds around 830,000 cards.

Both databases are, thanks to their broad scope, frequently used by the expert and lay public from the ranks of Bohemists, Polonists and researchers from related humanistic disciplines.

When processing the bibliography of contemporary production, both institutes come across an increasing number of e-sources. The "changing social and cultural status"[1] of electronic sources dealing with literature led both institutes to take the decision that the current literary bibliography cannot ignore such sources in their excerption. CLB therefore launched the "Czech Literary Internet"[2] project in 2017, the aim of which is to map out bibliographically literary life on the Czech Internet from the 1990s to the present day. PBL focuses on the same issue, among others, as part of the IBL.eu project, which also involves research into the issue of processing literary blogs. It is clear that the specific environment of the Internet holds texts similar to printed sources and materials that present those that process national literary science bibliographies with a methodological challenge.

In the paper that follows we endeavour to summarise the conceptual starting point and the initial experiences of both sister institutes in the bibliographic processing of Internet material. In addition to general information about both projects, we concentrate mainly on the general methodological questions associated with processing electronic sources for the needs of literary science bibliographic databases. We will therefore consider in more detail the issues connected to the collection of excerpted material and its archiving and will outline the fundamental problems associated with processing Polish literary blogs.

[1] KAŹMIERCZAK, Marek. Użytkownik, nadawca i odbiorca w Web 2.0. Uwagi o różnych sposobach odnoszenia się do literatury w serwisie Twitter. Teksty Drugie, 2012, 6, 217. ISSN 0867-0633.

[2] The study was established as part of implementation of the project *Czech Literary Bibliography – Czech Literary Internet: data, analyses, research*, CZ.02.1.01/0.0/0.0/16_013/0001743, which is cofinanced by the European Union through European Structural and Investment Funds within the Operational Programme Research, Development and Education.

The Czech Literary Internet project

The Czech Literary Internet project will be handled at CLB from 2017 to 2021 and is supported by EU operational programmes. The aim of the project is to carry out comprehensive research of the Czech literary Internet. In addition to actual analytical processing of e-articles from Internet servers and electronic magazines, it also focuses on the development of software tools for the analyses of excerpted data and scientific research into this material. We are able to summarise the first, constituent results following the first year of work on the project, in particular the specifics of the bibliographic processing of e-sources.

During the first year of Internet excerption, the excerpt base was extended by 44 e-sources, which are processed retrospectively from the beginning of their existence to the present according to the CLB excerption criteria, i.e. we capture both primary texts (fiction) and secondary texts (reflections) in relation to Czech literature and literary culture. It was possible to increase the database to include 10,000 new records from the Internet environment by 31st August 2018. It became clear from preliminary analyses that the emerging database contains reference to articles which the user is often unable to find by merely using an Internet search engine and that they are ordered far to the back in lists of results of background research on Internet search engines due to their low page rank value. This value is primarily taken from the number of references from other pages to the relevant page. Periodicals published in PDF format are entirely invisible to search engines due to the impossibility of indexing their content.

The specifics of records of articles from the Internet

A separate base, given the internal label of INT (internet.ucl.cas.cz), was created for work with records of articles from the Internet as part of improving user services and more comfortable work with data, at the same time records of articles from the Internet are available from the main CLB database (biblio.ucl.cas.cz). Records from websites and electronic magazines can be recognised primarily by the supplement [online] stated after the name of the excerpted source (in MARC21 subfield 773t), for example Ikaros [online], and also by the systematically allocated URL links. Each record of an e-article contains a link to its online version, and if possible to the archival version in Webarchiv.

All excerpted titles, printed and electronic, are also kept on record in the base of excerpted sources (excas.ucl.cas.cz). Here the user finds information about whether the website is processed in full, or merely in part (part, section, etc.). Here we also have on record information about the numbers of records for one excerpted year, and thus volume, to allow the user to have a better idea of the extent to which the server in question deals with literature.

Criteria of selection of e-sources

The criteria for selecting appropriate servers for inclusion in the excerption base in many ways concur with the criteria set out for printed periodicals. We also took into consideration, in particular, the thematic focus of servers, the quality of content and a specific criterion for Internet sources, i.e. the archival possibilities of the selected servers. We consider the perspective of findability of texts through a classic search engine to be secondary. We have not yet included blogs in the exception selection for a number of reasons. There are few blogs

written by established Czech literary critics and those that do often contain texts already having been published in a printed periodical, and consequently captured in the article database. Blogs written by authors that publish only on the Internet, on the other hand, predominantly feature reviews of books written by foreign authors, which does not meet the criterion for inclusion in the CLB article database.

Archiving

The transience of electronic content relates on the one hand with the variability of links (impersistent links, transfers of articles to new sections) and, on the other, the lifetime of an actual e-source. The risk that a server and its entire content will disappear without replacement or back-up archiving is certainly not negligible. For these reasons archival possibilities have become an important criterion for the inclusion of individual servers among the excerpted platforms. Given the unforeseeable development of the Internet environment, we consider it necessary for each e-article to contain a link to its original location within the network and to the full text, backed-up at a trustworthy repository.

At first, we considered establishing our own repository in order to ensure the long-term archiving of online content. However, such a solution would be demanding on IT support and would slow down the work of the excerptors themselves. We would also have to deal with the issue of copyright to be able to make texts available to a third party, which would place considerably higher demands on the administrative assurance of the project.

For these reasons we eventually decided to use Webarchiv, the digital library of Czech electronic online sources, which is managed by Národní knihovna ČR (National Library of the Czech Republic) and which has been systematically involved in the archiving of the Czech web for a long time. Individual websites are archived by regular harvests and made available on a "wayback machine" platform, which makes it possible to view the concerned page in various historical versions. Webarchiv makes digital content available in accordance with copyright law and the contracts which it signs with the operators of individual websites, or based on the Creative Commons licence under which the pages are displayed. Thanks to Webarchiv, we always add a link to the oldest archived version of the document to the record so that the user has access to the text in its original form.

Of the 10,000 records we have in the base, we do not have an archival version connected to 11% of records. This result is, however, distorted by servers that do not yet have a contract in place with Webarchiv. Without them we arrive at only 6%. This portion is mainly made up of the latest records which have not yet been processed by Webarchiv.

Types of processed servers

We broadened the existing excerption base to include electronic magazines in PDF format, which mainly take the form of paginated, regularly segmented documents, and to include integration sources (web servers) that do not have regular pagination. Their segmentation is, in contrast to electronic magazines, far more variable and content updates are irregular. The bibliographic data of e-articles from integration sources for these reasons usually only contain the date of publication and the ISSN of the relevant server.

We therefore newly included the following in excerption:

- Websites that supplement/broaden the content of printed excerpted magazines;
- Literary websites;
- Journalism websites;
- University magazines;
- Library magazines;
- Titles already having been excerpted that have moved from printed to electronic version.

Subject-matter

The thematic selection of websites reflects the criteria set for printed periodicals in the base. Even in the digital environment we are interested primarily in a reflection of literature and Czech culture from the perspective of other humanistic sciences. We therefore mainly monitor websites that focus on literature and culture, as well as specialised production published online. We are currently beginning to process large news servers without printed version (for example, Neviditelný pes, Aktuálně.cz, etc.) and in a further stage we plan to work with the online versions of national newspapers or public media websites (Czech Radio, Czech Television).

Quality of content

The web environment can be termed extremely liberal: to exaggerate a little, anyone can publish anything there, which is reflected in the varying quality of individual texts. This is influenced by the fact that there is often no editorial team and that articles are written by literary enthusiasts with varying degrees of experience in writing texts and with the genre that they are trying to achieve. We decided to confront this volatility of content using the criteria of the quality and information repleteness of texts. Whereas in printed periodicals, with the exception of the publishers' annotations and evidently commercial communications, we try to reflect any mention of a literary event, with materials from literary servers each excerptor must evaluate the information and content value of individual texts him/herself and define his/her own criteria. It is necessary to determine whether a particular text has any meaningful value at all and is worthwhile capturing or whether, on the contrary, it is merely a commercial communication or a text completely taken from another source.

Benefit

We consider capturing a reflection of popular literature[3], which appears only marginally in printed periodicals, to be a clear benefit of excerption of electronic sources for the CLB base.

[3] By this we primarily mean the shift of reflection on the genres of popular literature to the Internet environment. There are specialised websites here that deal with genres such as crime novels, horror, sci-fi or fantasy. More information about the genre system of popular literature can be found, for example, in the book by MOCNÁ, Dagmar a kolektiv. *Encyklopedie literárních žánrů*. Praha: Paseka, 2004. p. 503.

The move of publication platforms from printed periodicals to the web is perhaps most striking for popular literature. This step has allowed CLB to expand to include new names of creators of Czech popular literature and the names of critics that have concentrated on such creation for some time now. We are preparing authoritative records of newly-captured authors for a base of national authorities (AUT).

More information about happenings in regions is also making its way into the base - about readings by authors and literary competitions, which national periodicals understandably neglect. We see another positive in supplementing the publication activity of authors that already appear in our base. Before launching the project, we only had their publication activity in printed sources on record. Excerption of the literary Internet therefore made it possible, for example, to monitor whether an author publishes simultaneously on the Internet or whether he/she has opted to publish his/her work within the digital environment alone, for whatever reason. It is also possible to monitor whether an author acts differently on the Internet and in a printed periodical: for example, whether he/she focuses on different literary and journalistic genres, chooses different language, notices other issues, etc.

Problems associated with the processing of e-sources: archiving

As previously indicated, an important perspective in selecting electronic platforms for the Czech Literary Internet project was the inclusion of a webpage in the regular selection collections of Webarchiv[4]. It proved, however, that the inclusion of a server in regular gathering does not guarantee the one-hundred per cent existence of an archive link for an excerpted e-article, i.e. that Webarchiv does not always automatically gather all material from the concerned website.

Most of the problems involved in obtaining an archival version until now have arisen in conjunction with some historic change on the concerned server, such as moving an e-article to a different section, a section ceasing to exist or some change to the software settings of the relevant website which disables collection robots from saving their content in Webarchiv under the current format of URL link. We deal with this situation by manually searching the archival version of the whole website because there is a high chance that the relevant article was collected in the past. Thanks to a contract with CZ.NIC, the National Library archives the full Czech Internet at least once a year. If we are unsuccessful in our search, we contact Webarchiv and agree on the inclusion of specific URL links in selective collection, which is carried out every month.

For unknown technical reasons, then, articles from, for example, the interesting literary website "Opičí revue", are not available in Webarchiv, that site having contained many stimulating literary reviews and functioning from 2010 to 2017. Webarchiv did gather it until 2017, until it closed, but an unknown error on the website means that only articles to the year 2013 are visible with the Webarchiv environment. We consequently only excerpted articles through

[4] Webarchiv distinguishes three types of harvest: blanket, selective and thematic. "Selective collection only covers selected sources, but in contrast to blanket collection the emphasis is placed on capturing sources and changes to them to the full extent." In: KVASNICA, Jaroslav, RUDIŠINOVÁ, Barbora, HAŠKOVCOVÁ, Marie, HOLOUBKOVÁ, Monika and Markéta HRDLIČKOVÁ. *Strategie budování sbírky Webarchivu: aktualizované znění* [online]. Verze 2.0. Praha: Národní knihovna, 2017 [Accessed 12 October 2018]. Available from: **https://webarchiv.cz/static/www/download/collection-policy-2017.pdf**

the archival version of the server and in doing so enhanced our database with 207 articles that are no longer findable and displayable using an Internet search engine.

We assume that is not an isolated case and it is likely that non-functioning links to other e-articles will appear in our database in a few years. We will most likely deal with this problem in the future with an automatic script that will detect and mark such non-functioning links.

The Polish Literary Internet project[5]

One of the objectives of the IBL.eu project is to capture the diversity of the Polish literary blogosphere. Implementation was initiated at an important historical milestone for this publication channel: two blog platforms were closed at the beginning of 2018 - blog.pl and bloog.pl, arousing debate over the end of the "golden days of the Polish literary blogosphere"[6]. The truth is that a whole host of writers' blogs have not been updated for a number of years, were officially closed or have disappeared from the web and the literary discussion ongoing there has moved to social networks. These provide the required "immediate exchange of ideas, general access and simplicity of use"[7]. This trend, however, does not affect review and readers' blogs, which continue to maintain their status. Opinions do appear that favour blogger-reviewers over reviewers from printed and electronic magazines. At the very least, as far as the quantity of literature which they are able to read and appraise is concerned (special attention is devoted to "parental" blogs, the number of which has risen in recent years in line with the growth of literature for children).

Literary blogs

In contrast to the Czech environment, literary blogs play a more important role on the Polish Internet. As projects that are generally backed by a single person, they show a higher degree of instability than "regular" literary websites or e-magazines, which usually have an editorial team of several members to call on. It is now clear that part of the Polish literary blogosphere has probably been lost for good, as is the case for some of the content of Czech literary servers.

The project of processing these is in its infancy and thus stands at the stage of initial research, bringing with it the need to ask fundamental methodological questions. If literary blogs become the starting point for a future project of archiving (in its entirety?) the Polish literary blogosphere, it will be required to specify its boundaries. At this stage of the project, however, a basic typology of blogs is sufficient.

We distinguish the following types of blogs for its needs:

- the blogs of writers that can be termed "author sites". These are used to publicise writing activities, new publications or readings by authors (Przemysław Dakowicz);

[5] Thank you to Anie Gnot for the translation of the Polish text into Czech.
[6] WIŚNIEWSKI, Michał R. Ludzie, którzy piszą w internecie. Dwutygodnik [online]. 01/2018 [Accessed 26 August 2018]. Available from: http://www.dwutygodnik.com/artykul/7600-ludzie-ktorzy-pisza-w-internecie.html
[7] KAŹMIERCZAK, Marek.: op. cit., p. 274. Also WIŚNIEWSKI, Michał R: op. cit.

- blogs which contain literary formations[8]: novels to be continued, aphorisms, poems, etc., including blogs which are later published in book format; in this case publication raises their literary character to a certain extent (the Zorkownia blog written by Agnieszka Kaluga); (the notepad character of blogs as such requires special attention);

- blogs by writers written in the format that can be classified as the (literary) genre of "electronic diary"[9] (Jerzy Sosnowski);

- review blogs (Bernadetta Darska) and readers' blogs (Niezatapialna armada), which vary in terms of the level of professionalism and the authorial strategies.

Evaluation of blogs

Other criteria for the selection of blogs apply to blogs established by writers and critics (or those looking for recognition)[10] and others again to amateur blogs. In the first case, a significant role in the selection process is played by the writer him/herself, because every unreproducible, unprinted piece of material is part of his/her work or illustrates the content of his/her literary texts. In the case of novice or amateur writers, the risk again rises of loss of literary creation from the web, and with it data about the author. It is easy to set up a blog and for this reason the Internet abounds with such literary endeavours. It is therefore entirely obvious that the criterion of literary quality comes into play here. We have until now, in creating PBL, not assessed the quality of literary works because we respected the decisions of the editors of individual magazines in terms of putting these into print and we therefore excerpted all texts which met the thematic criteria for the inclusion of an article in the PBL database. In terms of the bibliographic processing of blogs, this criterion will have to be modified because the quality of published texts is extremely varied.

The availability and archiving of blogs

In all cases - updated, not updated and archival blog - we would like to make a copy of such data units[11] available and to back them up. However, we do not have access to any professional data storage site equivalent to Webarchiv in the Czech Republic. The database of internal sources operated within the bounds of the SYNAT/PASSIM project in place at the Polish National Library is not an archive of electronic sources and merely contains basic metadata and links. Only the incomplete, randomly archived version of blogs can be found at the Internet Archive server (archive.org), and within the scope of analogue initiatives. Full access to blogs which are no longer active is desirable both on the grounds of their literary value (for example, the blog, no longer existing, written by writer Inga Iwasiów) and for the possibility of the reconstruction of the original version of the Polish literary Internet (for example, the following blogs: kumple.blog.pl, mydziecisieci.blog.pl). The implementation

[8] MARYL, Maciej. Życie literackie w sieci: pisarze, instytucje i odbiorcy wobec przemian technologicznych. Warszawa: Instytut Badań Literackich PAN, 2015. ISBN 978-83-61750-61-1.

[9] MARYL, Maciej, op. cit., p. 263.

[10] MARYL, Maciej, op. cit., p. 139.

[11] We see the difference between a non-updated and an archival blog to be that the content of a non-updated blog is available on the Internet, but has not changed for two years or longer. The category of archival blogs includes such pages which are no longer publicly available, no longer exist or whose content is only available through an external Internet archive (for example, from the pages of archive.org).

of a project involving the archiving of the literary Internet would require the creation of a repository for such data or the use of existing online tools (only the second solution comes under consideration at the current stage of the project).

The content of blogs having been published in book form (the Zorkownia blog written by Agnieszka Kaluga), which predominantly contains the reprints of poems, columns, articles (the blog written by poet Wojciech Wencel) or which are located on the servers of publishing houses, magazines and other similar institutions (for example, the blog written by literary critic Justyna Sobolewska at the Polityka.pl journalism site) is obviously available.

Conclusion

In spite of the fact that both of the projects described above approach the issue of the literary Internet from different positions (mass retrospective excerption versus targeted narrower probe), similar methodological obstacles can certainly be seen. These are found in the selection of appropriate e-sources, their technical processing, the quality and transience of electronic content and their archiving. *The Czech Literary Internet builds on existing methods of processing printed periodicals and on the criteria for their selection which are set within the bounds of CLB, and can therefore base itself on material already having been processed.* For these reasons the Czech project included in excerption professional and specialised e-magazines and websites which focus on literature in which the prevailing portion of reflection on Czech literature is concentrated.

The bibliographic processing of materials from the literary Internet therefore presents both national literary science bibliographic projects with new methodological challenges. The bibliographer is no longer faced with the task of simply describing the concerned document - he/she must now, as stated above, become a more active evaluator of the quality of the analysed text, mainly as a result of the absence of editorial work on certain servers, which sometimes leads to texts being published which do not satisfy the conditions for inclusion in the database of articles (commercial communications, not fitting in with the genre, poor language).

The excerptor must in this way deal with the issue of archiving the concerned document: without the existence of an archived version of the document, the work of the bibliographer might be wasted at any time in the future. In this regard, the inclusion of Internet materials in a portfolio of processed documents is a more demanding task that it might appear at first glance.

References

KAŹMIERCZAK, Marek. Użytkownik, nadawca i odbiorca w Web 2.0. Uwagi o różnych sposobach odnoszenia się do literatury w serwisie Twitter. *Teksty Drugie*, 2012, **6**, 217. ISSN 0867-0633.

KVASNICA, Jaroslav, RUDIŠINOVÁ, Barbora, HAŠKOVCOVÁ, Marie, HOLOUBKOVÁ, Monika and Markéta HRDLIČKOVÁ. *Strategie budování sbírky Webarchivu: aktualizované znění* [online]. Verze 2.0. Praha: Národní knihovna, 2017 [Accessed 12 October 2018]. Available from: **https://webarchiv.cz/static/www/download/collection-policy-2017.pdf**

MARYL, Maciej. *Życie literackie w sieci: pisarze, instytucje i odbiorcy wobec przemian technologicznych*. Warszawa: Instytut Badań Literackich PAN, 2015. ISBN 978-83-61750-61-1.

MOCNÁ, Dagmar and Josef PETERKA. *Encyklopedie literárních žánrů*. Praha: Paseka, 2004. ISBN 80-7185-669-X.

WIŚNIEWSKI, Michał R. Ludzie, którzy piszą w internecie. *Dwutygodnik* [online]. 01/2018 [Accessed 26 August 2018]. Available from: **http://www.dwutygodnik.com/artykul/7600-ludzie-ktorzy-pisza-w-internecie.html**

Databases

Česká literární bibliografie [online]. ÚČL AV ČR, 2011 [Accessed 17 October 2018]. Available from: **http://biblio.ucl.cas.cz**

Databáze excerpovaných časopisů [online]. ÚČL AV ČR, 2011 [Accessed 17 October 2018]. Available from: **http://excas.ucl.cas.cz**

Český literární internet [online]. ÚČL AV ČR, 2011 [Accessed 17 October 2018]. Available from: **http://internet.ucl.cas.cz**

Polska Bibliografia Literacka [online]. IBL PAN, 2018 [Accessed 17 October 2018]. Available from: **http://pbl.ibl.poznan.pl/dostep/**

Internet Archive [online]. The Internet Archive, 2018 [Accessed 17 October 2018]. Available from: **https://archive.org/**

Polish blogs

Brukowiec literacki kumple [online]. 2002-2015 [Accessed 17 October 2018]. Available from: **https://web.archive.org/web/20080715000000*/http://kumple.blog.pl/**

DAKOWICZ, Przemysław. *Dakowicz* [online]. 2010- [Accessed 17 October 2018]. Available from: **http://dakowicz.blogspot.com/**

DARSKA, Bernadetta. *Nowości książkowe - Blog Bernadetty Darskiej* [online]. 2015- [Accessed 17 October 2018]. Available from: **http://bernadettadarska.blogspot.com/**

IWASIÓW, Inga. *Świat książki* [online]. 2010-2016 [Accessed 17 October 2018]. Available from: **https://web.archive.org/web/20100801000000*/http://ingaiwasiow.pl/**

KALUGA, Agnieszka. *Zorkownia* [online]. 2010- [Accessed 17 October 2018]. Available from: **http://www.zorkownia.pl/**

Mydziecisieci [online]. 2001-2016 [Accessed 17 October 2018]. Available from: **https://web.archive.org/web/20070915000000*/http://mydziecisieci.blog.pl/**

Niezatapialna Armada Kolonasa Waazona [online]. 2011- [Accessed 17 October 2018]. Available from: **https://niezatapialna-armada.blogspot.com/**

SOBOLEWSKA, Justyna. *Oczytany | Blog Justyny Sobolewskiej* [online]. 2009- [Accessed 17 October 2018]. Available from: **https://sobolewska.blog.polityka.pl/**

SOSNOWSKI, Jerzy. *Jerzy Sosnowski* [online]. 2006- [Accessed 17 October 2018]. Available from: **http://jerzysosnowski.pl/**

WENCEL, Wojciech. *Wojciech Wencel* [online]. 2009- [Accessed 17 October 2018]. Available from: **http://wojciechwencel.blogspot.com/**