Data przesłania artykułu: 25 IX 2020 r.
Data przyjęcia artykułu do druku: 13 X 2020 r.
DOI: http://dx.doi.org/10.12775/AKZ.2020.005

Marcin Wilkowski
(Uniwersytet Warszawski, Instytut Badań Literackich PAN w Warszawie)

# POLISH WEB RESOURCES DESCRIBED IN THE *POLISH WORLD* DIRECTORY (1997). CHARACTERISTICS OF DOMAINS AND THEIR CONSERVATION STATE

**Słowa kluczowe**

Polish national domain, Web archiving, Web archives

**Keywords**

polska domena krajowa, archiwistyka Webu

## Summary

For the purposes of this study, the print version of the Polish World directory by Martin Miszczak (Helion, 1997) was used to create an index of historical URLs and verify their current availability and presence in Web archives. The quantitative analysis of the index

Marcin Wilkowski programista w Centrum Kompetencji Cyfrowych UW, doktorant w projekcie „Humanistyka cyfrowa. Studia doktoranckie Instytutu Badań Literackich Polskiej Akademii Nauk i Polsko-Japońskiej Akademii Technik Komputerowych”. Twórca portalu „Historia i Media” (2005–2016). Interesuje się historią cyfrową, archiwizacją Webu i innowacjami cyfrowymi w sektorze kultury. Programuje w R i językach webowych Autor m.in. Nowoczesna instytucja kultury w Internecie (2017).
E-mail: m.wilkowski@uw.edu.pl
ORCID ID: 0000-0003-2924-268X

was prepared to obtain the rank data on top-level domains (TLDs) and subdomains, while the language of pages published in domains other than .PL was also examined. This study uncovered a low current availability (21.77 per cent) of Polish World URIs with a 79.6 presence in Web archives (60.35 for addresses unreachable today). Forty-six per cent of the addresses from the directory were available on domains other than .PL, of which only 15.36 per cent had content in Polish. It would seem that in 1997, Polish Internet users were able to use Polish-centric resources, mostly already available through the Polish country domain. The 180 domain names with the .PL suffix uncovered during the study constitute around 20 per cent of .PL domain names active until at least the end of 1996 on the Web.

**Streszczenie**

**Zasoby polskiego Webu opisane w katalogu *Polish World* (1997). Charakterystyka domen i stan zachowania**

W ramach badania wykorzystano drukowaną wersję katalogu Polish World Martina Miszczaka (wyd. Helion, 1997) w celu stworzenia indeksu historycznych adresów URL i zbadania ich współczesnej dostępności oraz obecności w archiwach Webu. Zasoby katalogu poddano analizie ilościowej pod kątem statystyki domen najwyższego rzędu i subdomen oraz zbadano języki stron publikowanych w domenie innej niż PL. Badanie ujawniło niską współczesną dostępność tych adresów (21.77%) przy obecności kopii w archiwach Webu na poziomie 79.6% (dla nieosiągalnych dziś adresów – 60.35% ). 40.64% adresów z katalogu dostępnych było na domenach innych niż PL, przy czym tylko 15.36% z nich posiadało treść w języku polskim. Wydaje się, że w początkach 1997 roku polscy użytkownicy korzystać mogli z polskocentrycznych zasobów dostępnych już przede wszystkim w polskiej domenie krajowej. Wyodrębnione w trakcie badania 180 wspólnych nazw domenowych z domeny PL to około 20 proc. nazw domenowych PL aktywnych przynajmniej do końca 1996 roku w sieci WWW.

## Introduction

Published in January 1997 by Helion, Martin Miszczak's *Internet. Katalog Polish World*[1] is part of a series of printed Internet guides sold in Polish book stores in the second half of the 1990s.[2] These types of editions, which

[1] M. Miszczak, *Internet. Katalog Polish World*, Gliwice 1997.

[2] Besides Miszczak's book, we can also note: M. Lewandowski, A. Stec, *5000 użytecznych adresów Internetu*, Warszawa 1996; M. Sokołowski, *Internet w Polsce [podręcznik dla każdego]*, Gdańsk 1996 (2nd ed. 1998); P. Rygałło, *Internet bez tajemnic*, Warszawa 1997; A. J. Kennedy, *Internet*, trans. J. Białko, P. Fraś, P. Goraj, Bielsko-Biała 1999 (subsequent editions in 2000 and 2001); *WWW informator adresowy*, ed. S, Narożniak, R. Żurawski,

were not solely a Polish specificity[3], were part of the broader offer of guides for those wishing to learn how to use computers and the Internet. The major part of these guides consisted of a curated list of Internet resources, which distinguished them from other guides which mostly contained practical advice on how to use the Internet, e.g. step-by-step instructions on how to install software allowing one to connect to the Internet or use e-mail. Due to limited access to computers and the Internet in Poland at that time[4], the resources already available online could not perform an educational role by themselves – these print guides provided help to those accessing the Internet for the first time and who did not know what to expect. These guides ceased to be necessary as the level of digital skills and access to the Internet increased. Today however, they constitute a valuable resource for studying the early Polish Web.

Within this context, *Internet. Katalog Polish World* is an exceptional publication since it constitutes a copy or a variation of the Web directory of the same name, published from February 1996 in Polish and English and built by its users[5]. Martin Miszczak, who developed the directory and authored the book, was in the 1990s the owner of the Internet Polska firm in Washington, which offered website building to state institutions, Internet directory creation and business consulting[6].

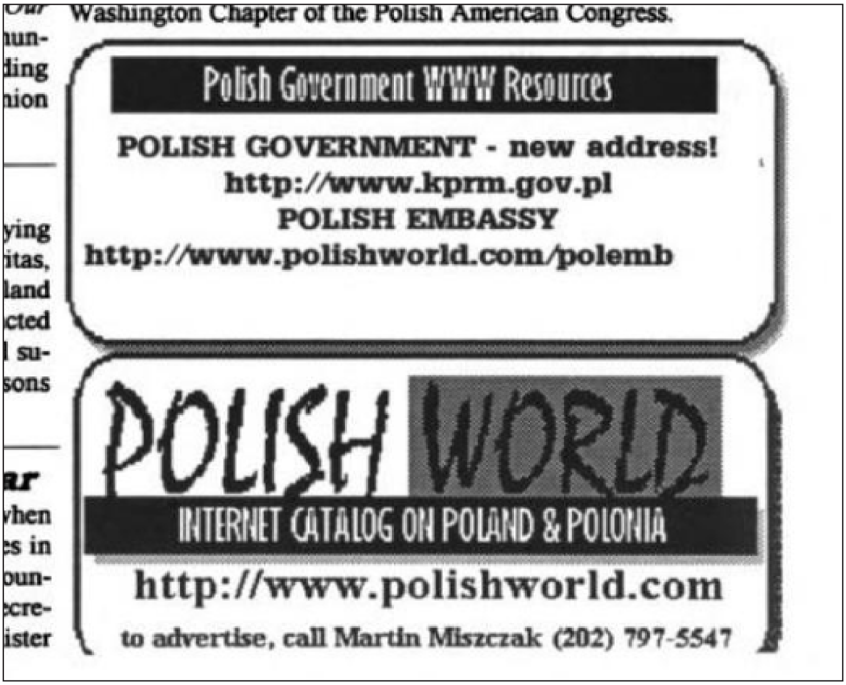
Warszawa 2000.; M. Rostek, P. Olejniczak, *Przewodnik po stronach internetowych polskich, angielskich, niemieckich*, Poznań 2002; M. Kaczmarczyk, *Najlepsze adresy WWW. Przewodnik*, Warszawa 2002.

[3] See. e.g. C. Maxwell, *McKinley Internet directory*, Indianapolis 1995; B. Cohen, *Social Studies Resources on the Internet: A Guide for Teachers*, Portsmouth 1998; B. J. Thomas, *The World Wide Web for Scientists & Engineers: A Complete Reference for Navigating, Researching & Publishing Online*, New York 1998; *Official Microsoft Bookshelf Internet Directory*, Redmond 1998; B. Hill, *Internet Directory for Dummies*, Foster City 1999.; *The History Highway 2000: A Guide to Internet Resources*, ed. D. A. Trinkle, S. A. Merriman, Amonk 2000.

[4] In 1997, around 8 per cent of Poles had a personal computer at home, while 5 per cent had Internet access in some form, see Centrum Badania Opinii Społecznej (CBOS), *Polacy i komputery.Komunikat z badań*, BS/69/69/97, Warszawa, 1997, pp. 4 and 6. The Internet was mostly used by those with a higher level of education, although the 1996/1997 period showed a certain opening towards students and schoolchildren, see *Wyniki badań zbiorowości użytkowników Internetu w Polsce,* "Internet: magazyn użytkowników sieci Internet" 1997, No 2, p. 11.

[5] *Polish World*, https://web.archive.org/web/19961221034200/http://www.polish-world.com/ (accessed 19.09.2020).

[6] M. Miszczak, *Miejskie serwisy w wyszukiwarce Polish World*, tekst of speech during 1st „Miasta w Internecie" conference (Tarnów 11–13 June 1997), http://arch.kmwi.pl/miejskie-serwisy-w-wyszukiwarce-polish-world/, (accessed 19.09.2020).



Our Washington Chapter of the Polish American Congress.

hun-
ding
nion

ying
itas,
land
cted
l su-
sons

ar
when
es in
oun-
ecre-
ister

Polish Government WWW Resources

POLISH GOVERNMENT - new address!
http://www.kprm.gov.pl
POLISH EMBASSY
http://www.polishworld.com/polemb

POLISH WORLD

INTERNET CATALOG ON POLAND & POLONIA

http://www.polishworld.com

to advertise, call Martin Miszczak (202) 797-5547

Fig . 1. Advertisement for *Polish World* and information concerning the web page of the Polish government and the Polish embassy in Washington, D.C.

Source: *News from Poland: Newsletter of the Embassy of the Republic of Poland*, Vol;. IV, March 1998, p. 5, Google Books project, https://books.google.pl/books?id=m6sjAQAAIAAJ&lpg=PA13-IA128&ots=ffhe7G0xKv&dq=Martin%20Miszczak%20polish%20world%20embassy&hl=pl&pg=PA13-IA128#v=onepage&q&f=false. (https://bit.ly/2ZRb48w)

Miszczak collaborated with the Polish embassy in Washington and created its official page on the polishworld.com domain[7], he also took part in conferences (e.g. *Polish Institute Of Arts And Sciences In America 54th Annual Meeting: A Multidisciplinary Conference*, Washington D.C. on 7–8 June 1996, *Miasta w Internecie*, Tarnów 11–13 June 1997). In 1997, the Polish-language version

[7] Embassy of Poland in Washington, D.C. http://www.polishworld.com/polemb/pindex.html (accessed 19.09.2020). At that time, the Polish Ministry of Foreign Affairs had a presence on the urm.gov.pl domain. The Embassy of Poland in Washington, D.C. did not provide me with any information on its collaboration with Martin Miszczak.

of *Polish World* was apparently used by 40 000 users monthly, although these estimates were based on dubious methodology[8]. According to Miszczak, at that time, the web directory contained around 3000 addresses in 12 categories. In comparison, the print edition is much more limited in scope and contains 958 unique addresses. The directory page is still available today[9], but in a version different from that in the 1990s. The links contained in the website do not have timestamps and for that reason cannot be used to study the early Polish Web. However, the print directory does allow this, as we can suppose that all the addresses it contains were accessible at the time it was published (January 1997).

## Research questions

The print version of the *Polish World* directory was used to create an index of URLs for web pages created no later than at the beginning of 1997. This index allowed to address two research questions.

The first relates to the dynamics of the evolution of the .PL country code top-level domain (ccTLD) in the latter part of the 1990s and concerns the habit of publishing Polish- language resources within foreign domains. Can *Polish World*, published in 1997, be perceived as evidence of the maturity of the Polish Web, able to provide resources to Polish users which are accessible under the national ccTLD? As part of the study, the part of resources contained in *Polish World* published outside of the Polish ccTLD was quantified, as well as the languages in which this part of the directory was published on the Web.

The second research questions relates to the contemporary availability of historical Web assets and is one of the most popular subjects in studies of Web history. As part of this research, the addresses contained in *Polish World* were verified in terms of their current availability in Web archives.

It is certain that the material used does not cover all the Polish resources available at the time. The issue of the representativity of the studied collection requires a separate discussion, which I shall attempt to initiate at the end of my analysis.

[8] M. Miszczak, *Miejskie serwisy.*

[9] *Polish World – Web portal to Poland in Polish and English*, http://www.polishworld.com/ (accessed 19.09.2020).

## Research context and related literature

One of the key elements in constructing historical studies of the Web is the definition of the limits of the studied resources – such a delimiter can be the country code top-level domain (ccTLD). Research on the history of national web domains is highly developed within the framework of contemporary archival studies and Web historiography. It has been carried out at least since the early 2000s[10], and as at least part of the researchers wish, it forms a separate discipline (National Web Studies)[11]. A conceptualisation of the national Web – as Richard Rogers and co-authors write – is one of the methods which allow to pass from a "placeless" cyberspace, deprived of geographical markers, to an Internet of identifiable national domains and Web pages whose contents, ads and language are appropriate for their location. This allows to study the current state of the Web as delimited by national borders. The effect of such research is to characterise a country (country profiling) and to describe its history[12].

Defining research limits based on the national domain is not a decision without limitations. Above all, not all resources published in a national language or relating to the studied country are published under its country domain. This problem also concerns resources published by national institutions. A country domain does also not always guarantee that the publisher is local – some ccTLDs can be registered by foreign parties – for example, in 2019 this type of registration constituted over 11 per cent of new domain registrations for the .PL domain[13]. The literature also mentions the offshoring of country domains[14]. Fur-

[10] N. Brügger, *Probing a nation's web domain: a new approach to web history and a new kind of historical source*, [in:] *Routledge Companion to Global Internet Histories*, New York/Abington 2017, p. 62. A discussion of current research, with a copious bibliography, can be found in N. Brügger, D. Laursen, *Historical Studies of National Web Domains*, [in:] *The SAGE Handbook of Web History*, London 2019, pp. 413–427.

[11] R. Rogers *et al.*, *National Web Studies. The Case of Iran Online*, [in:] *A Companion to New Media Dynamics*, ed. by. J. Hartley, J. Burgess, A. Burns, Chichester 2013, pp. 142–143; R. Rogers, *Digital Methods*, Cambridge 2013, pp. 125–152.

[12] R. Rogers, *Digital Methods*, p. 142.

[13] *Rynek nazw domeny.pl. Szczegółowy raport NASK za drugi kwartał 2019 roku*, p. 10, https://www.dns.pl/NASK_Q2_2019_RAPORT.pdf (accessed 19.09.2020).

[14] See e.g. K. T. Nakahira et al., *Geographic locations of web servers*, [in:] *WWW '06: Proceedings of the 15th international conference on World Wide Web*, New York 2006, pp. 989–990. https://doi.org/10.1145/1135777.1135979.

thermore, some ccTLDs have a specific nature, e.g. Spanish-language resources are also available through other ccTLDs than the Spanish one (.ES), while English-language (and supranational) resources have a wide reach across country domains[15]. Due to these limitations, the country domain is sometimes defined in a more complex manner, e.g. on the basis of the location of the registering entities or the geolocation of server IP addresses. A separate problem pertains to using country domains for states which no longer exist (e.g. the ccTLDs for Czechoslovakia or Yugoslavia)[16]. The strategy of profiling Web resources on the basis of country domain is also pertinent for Web archival programmes carried out by numerous national libraries or archives[17]– sometimes this is even expressed directly through legal deposit requirements.

The country domain does not need to be the sole basis for indexing studied web pages. In numerous analyses of the issue of URLs ceasing to be accessible (link rot) the basis for the indexing are scientific journal articles or journal archives containing references to hyperlinks. A source for these indexes could also be web page directories available online, social media or newspaper or audio-video archives, reports, discussion groups and interviews (e.g. with web designers). Another source could be projects such as Common Crawl[18] or HTTP Archive[19] as well as Web archives through the appropriate programming interfaces.

Printed guides to Web resources have been previously used in archival studies of the Web. In 2000, Joel D. Kitchens Pixey Anne Moseley demonstrated, based on testing nearly 4000 addresses from nine printed directories, that after two

[15] R. Baeza-Yates, C. Castillo, E. N. Efthimiadis, *Characterization of national Web domains*, "ACM Transactions on Internet Technology" 2007, Vol. 7, No 2, p. 7. https://doi.org/10.1145/1239971.1239973.

[16] A. Ben-David, *What does the Web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain*, "New Media & Society", Vol. 18, No 7, 2016, pp. 1103–1119. https://doi.org/10.1177/1461444816643790

[17] The British Web archive collects resources from the .uk, .scot, .wales, .cymru and .london from servers whose physical location in the United Kingdom could be ascertained. The archival of Web resources documenting the French elections (2007) was based on the French country domain (36.11 per cent), but also generic top-level domains (28.93 .com, 25.12 .org) and other ccTLDs (e.g. 1.39 .de, 0.17 .us). F. Lasfargues, C. Oury, B. Wendland, *Legal deposit of the French Web: harvesting strategies for a national domain*, Aarhus, Denmark. Sep. 2008, p. 6. https://hal-bnf.archives-ouvertes.fr/hal-01098538 (accessed 19.09.2020).

[18] *Common Crawl*, https://commoncrawl.org/ (accessed 12.09.2020).

[19] *HTTP Archive*, https://httparchive.org/ (accessed 12.09.2020).

years, between 11 and 40 per cent of them ceased to be available[20]. However, the authors did not study the presence of copies of these resources in Web archives.

Until now, indexes for the archival study of the Polish Web had not been prepared and the current availability of historical resources had not been analysed, with the exception of studies by Marcin Roszkowski and Bartłomiej Włodarczyk, and partly, that by Karol Król, pertaining to link rot[21]. The crawl.pl project[22] in 2005–2006 created a corpus of around 20 million Web objects from the .PL country domain, of which around two per cent were found to be older than four years at the time of the study (so dating to before 2001)[23]. However, this study was not of a historical or archival nature. At the same time, the lack of an institutional archive of the Polish Web, as well as the limitations of Web archives in accessing indexes of TLDs create difficulties for the historical analysis of the entirety, or a large sample, of the resources of the Polish country domain. In this situation, the use of printed Web directories, at least for partial studies, seems justified.

As mentioned above, national Web resources cannot be deemed identical to the resources within the country domain. For studies of the Polish Web in the 1990s, it becomes necessary to determine the frequency and characteristics of publishing Polish-language content on foreign domains. This phenomenon should also be analysed through the prism of the influence of Polish emigrants (or more generally, the Polonia, the Polish diaspora) on the development of the

[20] J. D. Kitchens, P. A. Moseley, *Error 404: or, what is the shelf-life of printed Internet guides?*, "Library Collections, Acquisitions, & Technical Services" 2000, Vol 24, No. 4, pp. 467–478.

[21] M. Roszkowski, B. Włodarczyk, *Cytowania zasobów sieciowych w polskich czasopismach z zakresu bibliotekoznawstwa i informatologii: analiza aktualności adresów URL*, „Zagadnienia Informacji Naukowej – Studia Informacyjne" 2016, t. 54, nr. 1, 2016, pp. 21–43, http://dx.doi.org/https%3A//doi.org/10.36702/zin.153; K. Król, *The Link Rot Phenomenon and its Influence on the Quality of the Websites of Rural Tourism Facilities in Poland*, „Economic and Regional Studies / Studia Ekonomiczne i Regionalne" 2019, t. 12, nr 1, 2019, pp. 68–79, https://doi.org/10.2478/ers-2019-0007. In the study by Roszkowski and Włodarczyk, resources from the Polish country domain constitute around 25 per cent of the analysed collection of links. In that of Król, around 40 per cent of web pages belonged to the .pl domain.

[22] *Crawl.pl project*, http://users.pja.edu.pl/~msyd/crawlPlProject.html (accessed 12.09.2020).

[23] C. Castillo, B. Starosta, M. Sydow, *"CRAWL.PL" Measuring Statistical and Structural Properties of the Polish Web: Technical Report*, "Studia Informatica: systems and information technology" 2007, 1/8, p. 46. See also *Grafy Polskiego WWW*, http://users.pja.edu.pl/~msyd/polskiWWWzbioryDanych.html#graf%20WWW (accessed 19.09.2020).

Polish Web. Due to limited access to the Internet in Poland in the 1990s, as well as limited access to computers in general, the lack of free hosting platforms and a low level of digital skills, the Polish diaspora – in particular in the West – may have had a certain technological advantage, permitting a faster reaction time to the need for Polish content on the emerging Web[24]. This advantage may have been particularly evident in the case of Poles working at foreign universities, since in the 1990s, the Internet still showed a certain academic, elitist nature. This question requires further study, which goes beyond a mere analysis of Web resources. However, it would seem that the *Polish World* index could be seen as documenting precisely this situation. In that case, the number of Polish-language resources present in foreign domains, catalogued in it, would have to be quite significant. Otherwise, the directory could be treated as an indication of a certain maturity of the Polish country domain, within which one could already find most of the content relevant to a Polish user in 1997. *Polish World* was a Polish-centric directory, even if a large portion of it was devoted to foreign web pages, yet still aimed at Polish users or those of Polish ancestry.

The ephemeral nature and state of conservation of Web resources is one of the main research themes in Web archival studies. However, there have been no studies on the availability of the resources within the Polish country domain as a whole. For instance, due to the differences in the research attempts to calculate the percentage of conserved resources, it is difficult to ascertain any unequivocal figure. At the time *Polish World* was published, a random analysis testing 361 URLs was performed by Wallace Koehler: between December 1996 and January 2001, only 34.4 per cent of the addresses were still reachable[25]. Jason Hennessey and Steven Xijin Ge studied the availability of 17 110 URLs from the footnotes in scientific articles held in the Web of Science collection and published between 1996 and 2010. Only around 40 per cent of the links in articles from 1996 (so existing no later than 1996), were still available in 2011. The authors also indicate that the historical availability of URLs depends above all on the time they were published and the TLD (some domains display greater stability than others). According to the model published by the authors,

[24] The literature on the Internet-related activities of emigrants is quite extensive, yet it lacks studies on the Polish diaspora in the 1990s. The phenomenon I describe is noted in K. Król *Latarnicy sieci. Kulturowe funkcje państwa spełniają internetowi emigranci*, „Wprost" 39/1998, https://www.wprost.pl/tygodnik/6172/Latarnicy-sieci.html.

[25] W. Koehler, *Web Page Change and Persistence – A Four-Year Longitudinal Study*, "Journal of the American society for information science and technology" 2002, Vol. 53, No. 2, p. 164.

the probability that a link published in a specific year will still be available later decreases by 3.7 per cent with each passing year[26]. The issue of the availability of Web resources over time is also related to the question of their presence in Web archives, as well as the stability/changeability of their contents[27].

## Developing the index and research methods

The 1997 print edition of the *Polish World* directory was scanned. The PDF files with the scanned images were then automatically converted through OCR into text documents using Google Drive. Next, using the R language, CSV source files were created for the index, with the columns corresponding to the layout of the descriptions for the various directory resources (link title, URL, description). The correctness of each URL was verified and errors due to the imperfections of the OCR process were corrected by hand. Of the 958 links in the index, those linking to services other than web servers were excluded[28], leaving 951 unique addresses. Using urtools-1.7.3[29], each URL was attributed a top-level domain and a generic/regional one (if available). In this manner, data was collected for 940 addresses[30]. For the 558 addresses within the .PL domain, 180 shared do-

[26] J. Hennessey, S. X. Ge, *A cross disciplinary study of link decay and the effectiveness of mitigation techniques*, "BMC Bioinformatics" 2013, 14, S5, p. 3, https://doi.org/10.1186/1471-2105-14-S14-S5.

[27] On the availability of pages in Web archives, see S. G. Ainsworth et al., *How much of the web is archived?*, [in:] *Proceeding of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries – JCDL '11, 133.* Ottawa 2011, p. 133-136, https://doi.org/10.1145/1998076.1998100; A. Alsum et al., *Profiling web archive coverage for top-level domain and content language*, "International Journal on Digital Libraries" 2014, Vol 14, pp. 149–166. On the changeability of web pages see e.g. D. Fetterly, D., Manasse, M. Najork, J. L. Wiener, *A large□scale study of the evolution of Web pages,* "Software: Practice and Experience" 2004, Vol. 34, No 2, pp. 213–237, doi:10.1002/spe.577; S. G. Ainsworth, M. L. Nelson, H. Van de Sompel, *Only one out of five archived web pages existed as presented*, in: *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, Cyprus 2015, pp. 257–266, https://doi.org/10.1145/2700171.2791044; H. Weinreich et al., *Not quite the average: An empirical study of Web use*, "ACM Transactions on the Web" 2008, Vol. 2, No 1, pp. 1–31, https://doi.org/10.1145/1326561.1326566.

[28] Usenet and Gopher addresses were removed.

[29] O. Keyes, J. Jacobs, *urltools: A package for elegantly handling and parsing URLs from within R*, https://cran.r-project.org/web/packages/urltools/index.html (accessed 19.09.2020).

[30] Links without a domain name (listed as IP addresses) and malformed ones (preventing the identification of the TLD) were removed.

main names were found by analysing the strings preceding the top-level domain or generic/regional domain and the top-level domain[31]. The determination of shared domain names allowed to compare them with the historical statistics for active .PL domains, kept by the operator of the Polish country domain registry, NASK (Research and Academic Computer Network).

The current availability of the resources in the index was examined with httr-1.4.2[32] and through analysis of the server response codes in the HTTP header[33]. The presence of URLs from the index in Web archives was determined through the Time Travel API[34]. Thanks to this service, archived copies of 951 pages were discovered through their URLs listed in the *Polish World* directory, with a date as close to the publication date (1997) as possible. Where copies existed in multiple Web archives, the copy from the first archive returned by the API was chosen. Copies which were HTML files were cleaned of tags with the Literary Exploration Machine (LEM) tool, provided by the CLARIN-PL consortium[35]. The contents of the pages fetched were automatically analysed linguistically with the help of the cid2 tool[36]. In the case of bilingual pages (Polish-English), the language was marked as Polish, since the goal of the analysis was above all to determine Polish content. The data used in the study were made available as open research data[37].

[31] For example, for hum.amu.edu.pl and ia.amu.edu.pl, the shared domain is amu.edu.pl, and the administrators for the domain (Adam Mickiewicz University) can create further subdomains independently of the national registry. While tpsa.pl, terner.pl and ata.com.pl have different domain names (tpsa, terner, ata).

[32] H. Wickham, *httr: Tools for Working with URLs and HTTP*, https://cran.r-project.org/web/packages/httr/index.html (accessed 19.09.2020).

[33] R. Fielding, J. Reschke, *Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content* (RFC 7231), https://tools.ietf.org/html/rfc7231#section-6.1 (accessed 19.09.2020).

[34] *Time Travel APIs*, http://timetravel.mementoweb.org/guide/api/ (accessed 19.09.2020).

[35] M. Piasecki, T. Walkowiak, M. Maryl, *Literary Exploration Machine: A New Tool for Distant Readers of Polish Literature*, "Digital Humanities" 2017, p. 1–5, https://www.semanticscholar.org/paper/Literary-Exploration-Machine.-New-Tool-for-Distant-Piasecki-Walkowiak/d65405b77d38a008eab5667eca8687d2b3d48007 (accessed 19.09.2020).

[36] J. Ooms, D. Sites, *cld2: Google's Compact Language Detector 2*, https://cran.r-project.org/web/packages/cld2/index.html (accessed 19.09.2020).

[37] M. Wilkowski, *mw0000/polish-world-index*, https://github.com/mw0000/polish-world-index. Web page URLs were anonymised due to the copyright restrictions.

## Index of domains and languages of resources not in Polish country domain

*Polish World* was not a directory of solely Polish web pages, but it nevertheless concentrated on the needs and interests of Polish users, both those living in Poland and abroad. According to Miszczak, *Polish World* is a guide to the Polish Internet and that of its diaspora[38], and in it, information on Silesian firms or universities in Warsaw coexists with that for services in New York aimed at the Polish diaspora, that for Polish collections in American libraries or simply foreign web pages concerning Central and Eastern Europe (its economy, monuments, numismatics, etc.). The preparation of top-level domain statistics, both national (ccTLDs, such as .PL, .SE, etc.) and generic (gTLD, such as .COM, .EDU, etc.) allowed the study of 'country-centricity' of the Miszczak directory and estimate the potential of the .PL domain at the time. To what level were its resources in 1997 sufficient to fill the print version of the Web directory aimed at *Poles and the diaspora*?

Analysis showed that 59.5 per cent of the links in the Miszczak directory linked to resources within the .PL domain. (see Table 1). The second most common domain was .com (12.9 per cent), then .NET (9.1) and .EDU (6.4). The directory contained some exotic domains, such as .KR (South Korea)[39] or .SG (Singapore)[40].

Table 1. URLs from the print edition of the *Polish World* directory, subdivided into top-level domains (n = 940)

| **Domain suffix** | **%** |
|---|---|
| pl | 59.36 |
| com | 13.19 |

[38] M. Miszczak, *Miejskie serwisy.*

[39] The directory contained a web page published in 1994 relating to the stay of a South Korean in Warsaw, see B. Chung. *Poland Diary*, 1994, 30. December, https://web.archive.org/web/19961202232216/http://poppy.kaist.ac.kr:80/monica/live/warsaw/paper01.html

[40] An English-language web page on stamps from Central and Eastern Europe, see *Tan Wee Cheng's Central & Eastern European Philatelic Resources*, https://web.archive.org/web/19981205201430/http://sunflower.singnet.com.sg/~tanwc2/stamps/stamps.htm (accessed 19.09.2020).

Table 1. URLs from the print edition of the *Polish World* directory

| **Domain suffix** | **%** |
|---|---|
| net | 9.15 |
| edu | 6.38 |
| ca | 1.91 |
| org | 1.91 |
| au | 1.81 |
| uk | 1.7 |
| de | 1.28 |
| remaining (< 1 per cent) | 3.31 |

In parallel, the languages of web pages whose copies, if possible from the time the directory was published, have remained in Web archives were studied. Language statistics for 306 web sites published under another domain than .PL are shown in Table 2. The dominant language is English. Polish or Polish-English resources constitute less than 17 per cent of the sites for which the language could be determined. The presence of Polish-language resources in web sites not hosted on the .pl domain and catalogued in *Polish World* is therefore quite small.

Table 2. Languages of web pages outside the .pl domain, catalogued in *Polish World*

| **language** | **%** |
|---|---|
| en | 73.20 |
| pl | 15.36 |
| not applicable | 9.48 |
| de | 0.98 |
| sv | 0.65 |
| fr | 0.33 |

Data is shown solely for copies maintained in Web archives (306 sites). For Polish-English pages, language was marked as Polish. Some copies were lacking content allowing to determine the language (9.48 per cent)

## Availability and conservation state

In the analysis of current availability of the resources from the *Polish World* directory, the server response codes, sent in response to each URL request, were studied. The HTTP 200 code confirms the availability of the resource, which can be served to the client. Other response codes can be considered as signalling the unavailability of the requested resource. However, the HTTP 200 response code cannot be seen as a guarantee that the content currently available at the URL is identical to the historical content, which was available to the author of *Polish World*.

Table 3. Server responses to URL requests for addresses indexed in *Polish World* (n = 951)

| **Server response code** | **%** |
|---|---|
| 404 | 38.8 |
| Lack of server response | 35.96 |
| 200 | 21.77 |
| 403 | 2.73 |
| 406 | 0.21 |
| 500 | 0.21 |
| 400 | 0.11 |
| 405 | 0.11 |
| 503 | 0.11 |

Of the 951 URLs catalogued by Martin Miszczak, slightly more than 21 per cent are still reachable (return HTTP response code 200, see Table 3). For the .PL domain, that figure is 23.66 per cent (see Table 4). Availability does not necessarily mean that we can access the same content as in early 1997.

Table 4. Server responses to URL requests for addresses for the .pl domain, indexed in *Polish World* (n = 558)

| **Server response code** | **%** |
|---|---|
| Lack of server response | 39.25 |
| 404 | 34.77 |

Table 4. Server responses to URL requests for addresses for the .pl domain

| **Server response code** | **%** |
|---|---|
| 200 | 23.66 |
| 403 | 1.79 |
| 500 | 0.36 |
| 503 | 0.18 |

The conservation state for sites can be studied not only based on their current availability, but also their presence within Web archives. Data from the Memento API show that 79.6 per cent of the URLs in the *Polish World* directory are present in Web archives in some form. For sites within the .PL domain, this result is higher, with 80.47 percent. Some sites have copies stored in multiple Web archives. The presence of copies in Web archives is particularly significant in the case of sites which are no longer reachable. Of the entire set of addresses from *Polish World* which are unreachable today, 60.35 per cent are present within Web archives. While for .PL domain addresses, this figure reaches 77.46. Naturally, this does not imply that we have full access to the versions of sites available at the time of publication of Miszczak's guide. Copies created before the end of 1996 constitute just 14.18 per cent of all copies. For the .PL domain, that figure is 6.99. It is important to note that 1996 is an important inflection point for Web archival. It was then, on the initiative of the Internet Archive foundation, that a global Web archive was started. Analysis of the first archival date for all sites from *Polish World* shows that no copies were created earlier than 1996.[41]

Table 5. Server responses to requests for shared domain names (n = 180)

| **Server response code** | **%** |
|---|---|
| 200 | 60 |
| Lack of server response | 36.11 |
| 403 | 2.78 |
| 404 | 1.11 |

[41] The Memento project contains data on resources from over 20 public Web archives, none of which were created before 1996.

As part of the study, 180 shared domain names were selected from the 558 URLs within the .PL domain, and their current availability and presence in Web archives was checked. Of the domain names which existed in 1997, 60 per cent are still accessible (see Table 5).

The overwhelming majority (97.2 per cent) of these 180 shared domain names has some type of copy stored in Web archives, while this does not automatically imply that their original content at the time *Polish World* was published can be accessed. At the same time, 94.4 of the original domains from the Miszczak directory, which are themselves inaccessible today, have at least one copy stored in Web archives. Shared domain names can represent the home pages of individual sites, or networks of sites, published for a single institution (e.g. amu.edu.pl). But this is not necessarily so. Their possible availability within Web archives does not automatically translate into availability of archived copies of resources published within subdomains (e.g. hum.amu.edu.pl and ia.amu.edu.pl) or on separate pages.

Table 6. First archival date for shared .PL domain names (n =175)

| **First archival date** | **%** |
|---|---|
| 1996 | 21.71 |
| 1997 | 54.29 |
| 1998 | 13.71 |
| 1999–2019 | 10.29 |

Over 70 per cent of shared domain names within the country domain, extracted from the addresses in *Polish World* and having any copy in Web archives, has a first copy from the years 1996–1997 (see Table 6).

## Discussion

An analysis of the domains and the language of websites indexed by Martin Miszczak could indicate that at the start of 1997 the .PL domain was already sufficiently developed that a guide containing links to “useful and interesting web pages”[42] could be based mostly on its resources. Web sites within the .PL domain

[42] Description from back cover of print version of *Polish World* directory.

constitute nearly 60 percent of the *Polish World* directory. This is certainly not an overwhelming majority, but the nature of the work being discussed should be taken into account. It was not addressed just to users in Poland, but also to the Polish diaspora, which could explain the strong presence of sites from other country domains. At the same time, less than 17 per cent of these sites have content in Polish. This could indicate a lack of "diaspora-centric" effect in online publishing at the time, or exemplify its limited reach. Defining this effect would require further study, with the inclusion of older and newer indexes. However, it is also certain that when studying the early Polish Web, resources published outside of the country domain cannot be ignored.

The conservation level of the *Polish World* resources (21.77 per cent accessible today) cannot be easily compared to other studies on the availability of Web resources due to the specificity of this attempt, as well as the lack of other studies of the conservation of the Polish Web. Yet it does not stray far from the previously-mentioned studies by Koehler (general study) or Hennessey and Ge (URLs cited in scientific article footnotes). Subsequent studies of the *Polish World* collection should describe the process of the disappearance of addresses, listed in *Polish World* but which are inaccessible today, presenting data year by year and also checking if and when the "stabilisation"[43] process described by Koehler took place. It would then be possible to compare the data obtained in this way to those from other countries, as well as note similarities in the processes involved. Simultaneously, the current availability of 80 per cent of shared domain names indicates that despite the loss of individual URLs, the principal domains existing in 1997 are still active, or have been updated. Although this does not necessarily imply the availability of historical resources published on those domains. The presence of nearly 80 per cent of the URLs from *Polish World* in Web archives allows to better study these resources in their historical aspects. This result fits the estimates made for the global Web by S. G. Ainsworth et al[44]. When analysing the presence of copies in Web archives, it should be remembered that *Polish World* was based on a directory which was published online. The fact that so many URLs were collected in a single place may have contributed to better indexing of these resources by Web archives, which built their initial collections through web crawling.

[43] W. Koehler, *A longitudinal study of Web pages continued: a consideration of document persistence*, "IR Information Research" 2004, Vol. 9, No 2. http://informationr.net/ir/9-2/paper174.html.

[44] Between 35 and 90 per cent of URLs have at least one copy in a Web archive.

The backdrop for the analysis of the conservation state of the URLs indexed in *Polish World* could be formed by the historical data on the number of active .PL domain names provided by NASK. Between 1995 and the end of 1996, this number increased from 91 to 885[45], so by nearly 870 per cent, although these were still low numbers. The 180 shared domain names within the .PL domain, which must have been active until at least the end of 1996, therefore represent around 20 per cent of all active domains with the Polish ccTLD at the time. This index would therefore seem to be of primary importance in the study of the history of the Polish country domain and allow for its further analysis, this time exploring resources stored in Web archives more extensively. It should also be noted that the *Polish World* directory documents the Polish domain before a significant breakthrough, which could be seen as the end of 1997, when the number of registered domain names had increased to 5309. The earliest history of the .PL domain requires further study. Seeing 1997 as an inflection point in the history of the Polish Web is also an argument made by an article summing up the state of Polish Web resources at that time. The journalists of *Internet* magazine, concluding a cycle presenting selected websites, underline that continuing to describe them in a print magazine is becoming ineffective:

> currently, a proper and detailed description of the Polish Web is simply impossible – this subject could be treated endlessly, yet at some point (which seems to be now) this would become meaningless. We could not manage to keep up with the dynamic development of the Internet anyway, even if we could reduce the publishing cycle to a single day.

They also recommend using printed Web guides, which are nothing other than the equivalent of telephone books[46]. The dynamic development of the Polish Web at the time leaves all printed directories behind.

## Bibliogaphy

Ainsworth, Scott G., Ahmed Alsum, Hany SalahEldeen, Michele C. Weigle, and Michael L. Nelson. „How Much of the Web Is Archived?" In *Proceeding of the 11th Annual*

[45] All historical data for active .pl domain names obtained through NASK access to public records, 20 February 2020.

[46] K. G. [Krystian Grzenkowicz], *Polskie WWW (cz. 7). Podsumowanie*, „Internet: magazyn użytkowników sieci Internet" 1997, No. 2, p. 16.

*International ACM/IEEE Joint Conference on Digital Libraries – JCDL '11*, 133. Ottawa, Ontario, Canada: ACM Press, 2011. https://doi.org/10.1145/1998076.1998100.

Ainsworth, Scott G., Michael L. Nelson, and Herbert Van de Sompel. „Only One Out of Five Archived Web Pages Existed as Presented". In *Proceedings of the 26th ACM Conference on Hypertext & Social Media – HT '15*, 257–66. Guzelyurt, Northern Cyprus: ACM Press, 2015. https://doi.org/10.1145/2700171.2791044.

AlSum, Ahmed, Michele C. Weigle, Michael L. Nelson, and Herbert Van de Sompel. „Profiling Web Archive Coverage for Top-Level Domain and Content Language". *International Journal on Digital Libraries* 14 (2014): 149–166.

„Ambasada RP w Waszyngtonie". Accessed 19.09.2020. http://www.polishworld.com/polemb/pindex.html.

Baeza-Yates, Ricardo, Carlos Castillo, and Efthimis N. Efthimiadis. „Characterization of National Web Domains". *ACM Transactions on Internet Technology* 7, No 2 (2007): 9–41. https://doi.org/10.1145/1239971.1239973.

Ben-David, Anat. „What Does the Web Remember of Its Deleted Past? An Archival Reconstruction of the Former Yugoslav Top-Level Domain". *New Media & Society* 18, No 7 (2016): 1103–19. https://doi.org/10.1177/1461444816643790.

Brügger, Niels, and Ditte Laursen. „Historical Studies of National Web Domains". In *The SAGE Handbook of Web History*, edited by Niels Brügger and Ian Milligan, 413–27. London: SAGE Publications, 2018.

Brügger, Niels. „Probing a nation's web domain : a new approach to web history and a new kind of historical source". In *The Routledge Companion to Global Internet Histories*, edited by Gerard Goggin and Mark McLelland, 61–73. New York/Abington: Routledge, 2017.

Castillo, Carlos, Bartłomiej Starosta, and Marcin Sydow. „"CRAWL.PL" Measuring Statistical and Structural Properties of the Polish Web : Technical Report". *Studia Informatica : Systems and Information Technology* 1(8) (2007): 43–73.

Centrum Badania Opinii Społecznej. „Polacy i komputery", 1997. https://www.cbos.pl/SPISKOM.POL/1997/K_069_97.PDF.

Chung, Brian. „Poland Diary". 1994, 30. December. https://web.archive.org/web/19961202232216/http:/poppy.kaist.ac.kr:80/monica/live/warsaw/paper01.html.

Cohen, Barbara. *Social Studies Resources on the Internet: a guide for teachers*. Portsmouth, NH: Heinemann, 1998.

„Common Crawl". Accessed 12.09.2020. https://commoncrawl.org/.

Fetterly, Dennis, Mark Manasse, Marc Najork, i Janet L. Wiener. „A Large-Scale Study of the Evolution of Web Pages". *Software: Practice and Experience* 34, No 2 (2004): 213–37. https://doi.org/10.1002/spe.577.

Fielding, Roy, and Julian Reschke. „Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content". Accessed 19.09.2020. https://tools.ietf.org/html/rfc7231#section-6.1.

G., K. „Polskie WWW (cz. 7). Podsumowanie". *Internet: magazyn użytkowników sieci Internet*, nr 2 (1997): 16.

„Grafy Polskiego WWW". Accessed 12.09.2020. http://users.pja.edu.pl/~msyd/polskiWWWzbioryDanych.html#graf%20WWW.

Hennessey, Jason, and Steven Xijin Ge. „A cross disciplinary study of link decay and the effectiveness of mitigation techniques". *BMC Bioinformatics* 14 (2013): S5. https://doi.org/10.1186/1471-2105-14-S14-S5.

Hill, Brad, and Lee Musick. *Internet Directory for Dummies*. Foster City, CA: IDG Books Worldwide, 1999.

„HTTP Archive". Accessed 12.09.2020. https://httparchive.org/.

Kaczmarczyk, Marcin. *Najlepsze adresy www: przewodnik*. Warszawa: Axel Springer Polska, 2002.

Kennedy, Angus J. *Internet*. Translated by Joanna Białko, Piotr Fraś, and Piotr Goraj. Bielsko-Biała: Pascal, 1999.

Keyes, Os, Jay Jacobs, Drew Schmidt, Mark Greenaway, Bob Rudis, Alex Pinto, Maryam Khezrzadeh, et al. *urltools: Vectorised Tools for URL Handling and Parsing* (ver. 1.7.3), 2019. https://CRAN.R-project.org/package=urltools.

Kitchens, Joel D., and Pixey Anne Mosley. „Error 404: Or, What Is the Shelf-Life of Printed Internet Guides?" *Library Collections, Acquisitions, & Technical Services* 24, No 4 (2000): 467–78. https://doi.org/10.1080/14649055.2000.10765711.

Koehler, Wallace. „A Longitudinal Study of Web Pages Continued: A Consideration of Document Persistence". *IR Information Research* 9, nr 2 (2004). http://informationr.net/ir/9-2/paper174.html.

Koehler, Wallace. „Web Page Change and Persistence – A Four-Year Longitudinal Study". *Journal of the American Society for Information Science and Technology* 53, No 2 (2002): 162–71. https://doi.org/10.1002/asi.10018.

Król, Karol. „The Link Rot Phenomenon and its Influence on the Quality of the Websites of Rural Tourism Facilities in Poland". *Economic and Regional Studies / Studia Ekonomiczne i Regionalne* 12, No 1 (2019): 68–79. https://doi.org/10.2478/ers-2019-0007.

Król, Krzysztof. „Latarnicy sieci. Kulturowe funkcje państwa spełniają internetowi emigranci". *Wprost*, 39 (1998). https://www.wprost.pl/tygodnik/6172/Latarnicy-sieci.html.

Lasfargues, France, Clément Oury, and Bert Wendland. „Legal Deposit of the French Web: Harvesting Strategies for a National Domain". Aarhus, Denmark. Sep. 2008. https://hal-bnf.archives-ouvertes.fr/hal-01098538.

Lewandowski, Mariusz, and Anna Stec. *5000 użytecznych adresów Internetu*. Warszawa: Wydawnictwo PLJ, 1996.

Maxwell, Christine. *McKinley Internet Directory*. Indianapolis, Ind.: New Riders Publ., 1995.

Miszczak, Martin. „Miejskie serwisy w wyszukiwarce Polish World – archiwum.kmwi.pl". Accessed 19.09.2020. http://arch.kmwi.pl/miejskie-serwisy-w-wyszukiwarce-polish-world/.

Miszczak, Martin. *Internet: katalog Polish World*. Gliwice: Helion, 1997.

Nakahira, Katsuko T., Tetsuya Hoshino, and Yoshiki Mikami. „Geographic Locations of Web Servers". In *Proceedings of the 15th International Conference on World*

*Wide Web – WWW ’06*. Edinburgh, Scotland: ACM Press, 2006. https://doi.org/10.1145/1135777.1135979.

Narożniak, Szymon, and Robert Żurawski, ed. *WWW informator adresowy*. Warszawa: Infor, 2000.

*Official Microsoft Bookshelf Internet Directory*. Redmond, Wash.: Microsoft Press, 1998.

Ooms, Jeroen, and Dirk Sites. *cld2: Google’s Compact Language Detector 2* (ver. 1.2.1), 2020. https://CRAN.R-project.org/package=cld2.

Piasecki, Maciej, Tomasz Walkowiak, i Maciej Maryl. „Literary Exploration Machine. New Tool for Distant Readers of Polish Literature”. *Digital Humanities*, 2017, 1–5. https://www.semanticscholar.org/paper/Literary-Exploration-Machine.-New-Tool-for-Distant-Piasecki-Walkowiak/d65405b77d38a008eab5667eca8687d2b3d48007.

„PJIIT crawl.pl Project: Crawling the Polish Web”. Accessed 12.09.2020. http://users.pja.edu.pl/~msyd/crawlPlProject.html.

„Polish World – Web portal to Poland in Polish and English”. Accessed 19.09.2020. http://www.polishworld.com/.

Rostek, Marta, and Piotr Olejniczak. *Przewodnik po stronach internetowych polskich, angielskich, niemieckich*. Poznań: Wagros, 2002.

Rogers, Richard, Esther Weltevrede, Erik Borra, and S. Niederer. „National Web Studies: The Case of Iran Online”. In *A companion to new media dynamics*, edited by John Hartley, Jean Burgess, i Axel Bruns, 142–66. Chichester; Malden, MA: John Wiley & Sons, 2013.

Rogers, Richard. *Digital Methods*. Cambridge, MA: The MIT Press, 2013.

Roszkowski, Marcin, and Bartłomiej Włodarczyk. „Cytowania zasobów sieciowych w polskich czasopismach z zakresu bibliotekoznawstwa i informatologii: analiza aktualności adresów URL”. *Zagadnienia Informacji Naukowej – Studia Informacyjne* 54, No 1(107) (2016): 21–43. https://doi.org/10.36702/zin.153.

„Rynek nazw domeny.pl. Szczegółowy raport NASK za drugi kwartał 2019 roku”. Accessed 19.09.2020. https://www.dns.pl/NASK_Q2_2019_RAPORT.pdf.

Sokołowski, Maciej. *Internet w Polsce [podręcznik dla każdego]*. Gdańsk: Investpol-Consulting, 1996.

“Tan Wee Cheng’s Central & Eastern European Philatelic Resources.” https://web.archive.org/web/19981205201430/http://sunflower.singnet.com.sg/~tanwc2/stamps/stamps.htm.

Thomas, Brian J. *The World Wide Web for Scientists & Engineers: a Complete Reference for Navigating, Researching & Publishing Online*. Bellingham, Wash.: New York: SPIE Press ; IEEE Press, 1998.

„Time Travel APIs”. Accessed 19.09.2020. http://timetravel.mementoweb.org/guide/api/.

Trinkle, Dennis A., and Scott A. Merriman, ed. *The History Highway 2000: A Guide to Internet Resources*. Armonk, N.Y.: M. E. Sharpe, 2000.

Weinreich, Harald, Hartmut Obendorf, Eelco Herder, i Matthias Mayer. „Not Quite the Average: An Empirical Study of Web Use”. *ACM Transactions on the Web* 2, No 1 (2008): 1–31. https://doi.org/10.1145/1326561.1326566.

Wickham, Hadley, and RStudio. *httr: Tools for Working with URLs and HTTP* (ver. 1.4.2), 2020. https://CRAN.R-project.org/package=httr.

Wilkowski, Marcin. *mw0000/polish-world-index*, 2020. https://github.com/mw0000/polish-world-index.

„Wyniki badań zbiorowości użytkowników Internetu w Polsce”. *Internet: magazyn użytkowników sieci Internet*, nr 2 (1997): 11.