ARCHEION
2021, T. CXXII, s. 44–68
ISSN 0066-6041
e-ISSN 2658-1264

DOI 10.4467/26581264ARC.21.015.14495

Was this the real Web? Quantitative overview of the Polish ccTLD Internet Archive data (1996–2001)

Marcin Wilkowski

Uniwersytet Warszawski / University of Warsaw (Poland)
m.wilkowski@uw.edu.pl, ORCID 0000-0003-2924-268X

ABSTRACT
This article is an attempt to build a quantitative panorama of the Polish country code top-level domain (ccTLD) in the years 1996–2001 on the basis of data generously provided by the Internet Archive. The purpose of analyzing over 72 million captures is to show that these resources have limited potential in reconstructing the early Polish Web. The availability of historical Web resources and tools for their easy exploration in no way determines their potential value and usefulness in research, even if we do not have access to alternative sources.

KEYWORDS
Internet Archive, Polish Web, historical Web resources

Czy to był prawdziwy Web? Ilościowy przegląd polskiej domeny krajowej w zbiorach Internet Archive (1996–2001)

STRESZCZENIE
Artykuł przedstawia ilościowy opis zasobów polskiej domeny krajowej (*country code top-level domain*, ccTLD) z lat 1996–2001, dostępnych w zbiorach Wayback Machine, archiwum Webu prowadzonym przez Internet Archive. Celem analizy ponad 72 mln archiwizacji (captures) jest wykazanie, że zasoby te mają ograniczony potencjał w rekonstruowaniu polskiego wczesnego Webu. Dostępność historycznych zasobów WWW i narzędzi do ich łatwej eksploracji w żaden sposób nie przesądza o ich potencjalnej wartości i przydatności w badaniach, nawet jeśli nie mamy dostępu do alternatywnych źródeł.

SŁOWA KLUCZOWE
Internet Archive, polska domena krajowa, zasoby historyczne www

Introduction

This article is an attempt to build a quantitative panorama of the Polish ccTLD in the years 1996–2001 on the basis of data generously provided by the Internet Archive. A critical overview of archived content reveals issues with the data obtained in this way and forces a careful use of them as historical sources, as well as asking questions about the representativeness of the archived Web in relation to the live Web, actually available at a specific time in the past. Until now, there has never been a quantitative assessment of the earliest history of the Polish Web. Since Poland, unlike most European countries, does not have an institutional Web archival project, the Internet Archive resources, collected since 1996, are the only available means to have a broader view of the evolution of the Polish country domain. However, the limitations of this data identified during this study, make it impossible to fully use them to reconstruct the early Polish Web.

It is certain that working with *re-born digital*[1]resources forces us to expand the concepts of the *archive* and *archivisation*, which in the case of Web archives, are formed not only by the institution and its resources, acts of acquisition and preservation, or legal boundaries, but also the intermediation of software (through a black box effect[2]) or the randomness of seed lists, which for instance appears in the choice of the URLs to be archived. The Polish ccTLD collections in the Internet Archive are the result of archivisation formatted in this manner, which has a direct effect on their ability to reconstruct the live Web of the past.

The analysis of the Internet Archive data and an attempt at sketching a panorama of the Polish country domain between 1996–2001 are presented in the following sections. First, an analysis of the quality of the archived Web data and the limitations of big data within historical research. Secondly, some basic facts concerning the history and development of the Polish ccTLD. Then, a presentation of the general characteristics of the archival resources obtained from the Internet Archive, with a particular emphasis on the CDX metadata describing the contents the WARC (Web ARChive) source files, which form the basis of the Internet Archive collections. The next part of the article presents selected data concerning the reach and intensity of data for individual years between 1996 and 2001, as well as their structure over time. On the basis of this data, minimum numbers of URLs and hostnames from the PL domain in subsequent years are determined. On the basis of the number of hostnames in relation to the number of archivisations, the diversity of collections obtained for the studied period is analysed. HTTP response status code statistics are

[1] N. Brügger, *When the Present Web is Later the Past: Web Historiography, Digital History, and Internet Studies*, "Historical Social Research" / "Historische Sozialforschung" 2021, 37, 4 (142), p. 104.

[2] See: A. Ben-David, A. Amram, *The Internet Archive and the socio-technical construction of historical facts*, "Internet Histories: Digital Technology, Culture and Society" 2018, 2, p. 4–6, https://doi.org/10.1080/24701475.2018.1455412. Accessed 16.09.2021.

also analysed, which could indicate a change in crawling methods during the studied period and undermine the representativeness of the collections thereby obtained. Statistics for the types of archived resources are also presented. Finally, the last part of the article is a discussion of the problem of the representativeness of the archived Web data. Since these data are used for research aiming to illustrate trends and social phenomena, a proper critique of their provenance and background is necessary, since these could influence the image of the past reconstructed using them.

The time range for the data was chosen on the following basis: the Internet Archive foundation started archiving the Web in 1996, which is the first year with data for the PL domain, not taking into account the relatively few and unimportant captures from 1995, or those with a wrongly attributed date (e.g. 1980). We can therefore consider that there are no significant surviving historical resources for the Polish Web prior to 1996, but it should be added that sites archived in 1996 may have also been created earlier. The end of the range is based on statistical data: according to the World Bank, in 2001, 9.1 percent of the population in Poland used the Internet. In 2002, this was already over 20 percent[3]. The use of this demarcation line is also supported by the introduction in 2001 of Neostrada, a cheaper means of accessing the Internet for consumers, which supposedly led to an influx of new Internet users in subsequent years, outside of professional and academic circles. We could also use symbolic events from 2001 to justify this choice. In September of that year, the Polish version of Wikipedia was launched, signalling a new chapter in the evolution of the Web, while the first Polish portal, the 'garage-built' Wirtualna Polska, found itself on the brink of bankruptcy due to the stock market crisis and was sold to the Telekomunikacja Polska. Also in 2001, the controlling stake in the company publishing the most popular Polish portal (Onet.pl) was sold to the owners of the TVN television station – this takeover of Internet companies by outside actors exemplified the gradual passage of the Internet into the economic mainstream.

Between 1996 and 2001, the Polish Web changed qualitatively – from a niche academic and professional tool, to an open medium co-created by its new users, a space not only for exchanging information and opinions, but also for running businesses.

[3] *Individuals using the Internet (% of population)*, The World Bank Data, https://data.worldbank.org/indicator/IT.NET.USER.ZS?locations=PL. Accessed 16.09.2021.

Literature and sources

Projects to archive parts of the Web have been implemented since the mid-1990s[4]. In many of the initiatives of this type, the only means (or at least one of the most important ones) to define the range of resources obtained is the top-level country domain, although the consensus within the relevant literature is that this is not a full representation of national Web resources, since these can also be available from other domains[5]. On the other hand, the country domain, unlike the Web sphere[6], is an easy to define filter for selecting addresses to be archived and can be precisely defined in national legislation which sets the boundaries of a Web archive. Since building Web collections solely based on the ccTLD can limit the representativeness of the collected data, interesting projects are being developed to use social media as a crawl index source – so as to allow a more efficient capture of resources and topics actually present in public debate[7].

Archived country domain resources are used in research using statistical methods and natural language processing, based on both HTTP header data and payload content. Research of this type requires direct access to the source files (and not via Web interface) as well as an appropriate processing infrastructure. Interpreting the effects of this research requires paying particular attention to the origins and background of the data used. The literature contains many studies which point to the limitations of the crawling process and their effect on the quality of the collected resources. As an example, Marc Spaniol et al. describe the problem of time shifts in archived collections: capturing a large Web site may span hours or even days, which increases the risk that the contents collected

[4] Eveline Vlassenroot et al. present the general characteristics of European Web archival projects, describing their technical bases and legal conditions, see: E. Vlassenroot, S. Chambers, E. Di Pretoro et al., *Web archives as a data resource for digital scholars*, "International Journal of Digital Humanities" 2019, 1, p. 85–111, https://doi.org/10.1007/s42803-019-00007-7. Accessed 16.09.2021.

[5] N. Brügger, L. Ditte, *Historical studies of National Web Domains*, [in:] *The SAGE Handbook of Web History*, N. Brügger, I. Milligan, eds., Sage, Los Angeles–London–New Delhi 2018, p. 417–419.

[6] "Web sphere is a collection of dynamically defined digital resources spanning multiple Web sites deemed relevant or related to a central theme or object", see: K. Foot, *Web sphere analysis and cybercultural studies*, [in:] *Critical Cyberculture Studies*, D. Silver, A. Massanari, eds., NYU Press, New York 2006, p.88.

[7] I. Milligan, N. Ruest, J. Lin, *Content Selection and Curation for Web Archiving: The Gatekeepers vs. the Masses,* [in:] *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 2016, p. 107–110.

so far are incoherent with the parts that are still to be crawled[8], while Dimitar Denev et al. even distinguish between *blurred* and *sharp* Web captures based on their coherency[9]. As Matthew S. Weber writes, "despite the differences in crawl scoping and crawl completeness, almost every Web crawl faces some degree of uncertainty with regards to the completeness of the crawl, and in turn, the validity and reliability of the data extracted from a given Web archive"[10]. Scott A. Hale et al. examine the low completeness of the data held in the Internet Archive and highlight the bias toward prominent web pages within its resources[11]. But collection bias does not result solely from the limitations of crawling. For instance, Nicola Jayne Bingham and Helena Byrne admit that, for legal reasons, the UK Web Archive collections lack social media discussions. Because the provenance of replies and comments is difficult to ascertain, the archive collects only social media accounts of named individuals or organisations, which can legally be included in the UK deposits. As they write, "this potentially limits the heritage preserved to those individuals and organisations that represent the more official voice"[12]. The shape of historical Web collections is also influenced by the need to respect the right to privacy and protect personal data, especially within projects collecting sensitive information, such as those archiving the websites and discussion boards of LGBTQ communities. Thus, any Web archive can be seen as a "memoryware", as Anat Ben-Davit writes – "specific forms of preservation techniques which involve both software and hardware, but also crawlers, bots, curators and users"[13].

[8] M. Spaniol et al., *Data quality in Web Archiving,* [in:] *Proceedings of the 3rd Workshop on Information Credibility on the Web*, Association for Computing Machinery, New York 2009, p. 19, https://doi.org/10.1145/1526993.1526999. Accessed 16.09.2021.

[9] D. Denev et al., *The SHARC framework for data quality in Web archiving*, "The VLDB Journal" 2011, 20, p. 184, https://doi.org/10.1007/s00778-011-0219-9. Accessed 16.09.2021.

[10] M.S. Weber, *Web Archives: A Critical Method for the Future of Digital Research*, WARCnet Papers, Aarhus 2020, p. 10–11, https://cc.au.dk/fileadmin/user_upload/WARCnet/Weber_Web_Archives_A_Critical_Method.pdf. Accessed 16.09.2021.

[11] S.A. Hale, G. Blank, V.D. Alexander, *Live versus archive: Comparing a web archive to a population of web pages,* [in:] *The Web as History. Using Web Archives to Understand the Past and the Present*, N. Brügger and R. Schroeder, eds., UCL Press, London 2017, p. 45–61.

[12] N.J. Bingham, H. Byrne, *Archival strategies for contemporary collecting in a world of big data: Challenges and opportunities with curating the UK web archive*, "Big Data & Society" 2021, 8, 1, p. 2, https://doi.org/10.1177/2053951721990409. Accessed 16.09.2021.

[13] A. Ben-David, *Critical Web Archive Research,* [in:] *The Past Web: Exploring Web Archives*, D. Gomes, E. Demidova, J. Winters, T. Risse, eds., Springer Nature Switzerland, Cham 2021, p. 181, https://doi.org/10.1007/978-3-030-63291-5_14. Accessed 16.09.2021.

Increased access to archived resources and tools which allow to explore large amounts of data, even without programming skills (see the SHINE project in the UK Web Archive[14]) can be seen as a new opportunity to use Web collections in research within the social sciences and the humanities. On the other hand, a lack of critical consideration as to the quality of the sources used, can lead to data biases and erroneous interpretations. As Ian Milligan writes, "It is possible to find almost anything you want within 38 million web pages. I can find evidence on any matter of topics that advances one particular argument or interpretation. Without the contextual information provided by the archive itself, we can be misled"[15].

The quality of archived Web data is a specific challenge in relation to the early Web. As Matthew S. Weber writes, "early Web archiving was sporadic in nature, as collections were often built from donated datasets or were constructed in an ad hoc nature by sampling across web domains"[16]. It is also known that the early Web resources collected in the Wayback Machine were a result of using link data from Alexa company, while Alexa used access patterns to determine the depth of crawl for each site – links and clicks were essentially votes on the value of a given page[17]. For this period, there is also a lack of alternative sources or data allowing us to compare the studied samples. Today, we have Common Crawl[18](since 2008/2009), HTTP Archive[19](2010) and collections created by national institutions. As a complement and a context to the Internet Archive data, we can also use the online Web directories popular early on (e.g. DMOZ), printed indexes and other data sources with a limited scope.

The Polish country domain, as a whole, has not been studied until now from a historical or archivistic point of view. The quantitative data for the Polish country domain for 1996–2001 are also incomplete. The Polish National Research Institute (NASK) collects data on the number of domains registered within the Polish ccTLD, however, this data does not reference the number of pages or other resources available under these domains. My own research based on a printed

[14] SHINE, https://www.webarchive.org.uk/shine. Accessed 16.09.2021.

[15] I. Milligan, *Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives*, "International Journal of Humanities and Arts Computing" 2016, 10, 1, p. 81, https://doi.org/10.3366/ijhac.2016.0161. Accessed 16.09.2021.

[16] M.S. Weber, op.cit., p. 3.

[17] M. Kimpton, J. Ubois, *Year-by-Year: From an Archive of the Internet to an Archive on the Internet,* [in:] *Web Archiving*, Julien Masanes, ed., Springer, Berlin–Heidelberg–New York 2006, p. 203.

[18] Common Crawl, https://commoncrawl.org//. Accessed 16.09.2021.

[19] HTTP Archive, https://httparchive.org/. Accessed 16.09.2021.

index from 1997 has shown the existence of at least 558 URLs available online by the end of 1996 as part of the Polish ccTLD, out of which around 20 percent are still available today.[20] A global context for research on the PL domain between 1996–2001 can be found in the research by Ricardo Baeza-Yates et al., using national Web data from 2000 to 2005[21] as well as data for Austria from 2001–2002[22]. For the period which interests us, important data were collected as part of the Internet Domain Survey (IDS) project, which also contains estimates for the Polish ccTLD. IDS data was not gathered by crawling the Web but by querying the Domain Name System for the names assigned to IP addresses[23].

Data and methods

This analysis is based on data obtained directly from the Internet Archive foundation as part of the IDUB (Excellence Initiative – Research University) project of the University of Warsaw. We obtained 203.3 GB of resources in all, in which WARC and CDX files contained the metadata and content of more than 72 million captures from 1996 to the end of 2001. WARC is the standard format for archived Internet content and HTTP headers, while CDX is a metadata format describing each archivisation. CDX files allow to perform fast queries in relation to URLs of archived resources, the archivisation dates and numbers of archived versions for these resources, server response status codes or file sizes. WARC files allow for a deep analysis of the content of archived resources. For HTML files, which form the basis of Web pages, queries on WARCs give access to the various elements of the DOM tree (and so HTML tags) or the factual content of the pages.

The queries performed as part of this study were in the Scala language, using the Archive Spark library and interactive Jupyter Notebooks. Archive Spark

[20] M. Wilkowski, P*olish Web resources described in the "Polish World" directory (1997). Characteristics of domains and their conservation state*, "Archiwa – Kancelarie – Zbiory" 2020, 11, 13, pp. 119–140, https://doi.org/10.12775/AKZ.2020.005. Accessed 16.09.2021.

[21] R. Baeza-Yates, C. Castillo, E.N. Efthimiadis, *Characterization of national Web domains*, "ACM Transactions on Internet Technology" 2007, 7, 2, pp. 1–33, https://doi.org/10.1145/1239971.1239973. Accessed 16.09.2021.

[22] A. Rauber et al., *Uncovering Information Hidden in Web Archives. A Glimpse at Web Analysis building on Data Warehouses*, "D-Lib Magazine" 2002, 8, 12, https://www.doi.org/10.1045/december2002-rauber. Accessed 16.09.2021.

[23] Internet Domain Survey Background (2003), https://web.archive.org/web/20031002012504/http://www.isc.org/ds/new-survey.html. Accessed 16.09.2021.

allows to work efficiently on large sets of Web archive data using CDX metadata and offers functions to filter and extract the contents of archived Web pages[24]. The use of Jupyter Notebooks allows to document and share the data calculation process[25]. The analysis was performed on a six-core, 3.6 GHz computer with 32 GB RAM, using Archive Spark within a Docker container[26], which limited the need to install the various software components.

The study of the 1996–2001 PL domain collections was mostly based on examining quantities and the relationships between the subsequent values in each individual archivisation (capture), available through the CDX metadata[27]:

- originalUrl – unique Internet address for each resource available online. A single URL could have been archived multiple times during the selected period;
- timestamp – date and time for the creation of a specific capture. As part of this study, used as the basis to distinguish capture collections for individual years in the 1996–2001 range;
- digest – checksum (generated with the SHA-1 algorithm) computed from the content of the server response (*payload*), specifying its version. In theory unique, in practice this value is repeatable for responses with no content (e.g. empty pages). For this reason, the number of versions can be greater than the number of archived distinct URLs within a given year (this is the case for 1996). At the same time, each archived URL may have had many versions;
- status – HTTP server response status code[28]. Over the studied period, the archivisation of a selected URL may have had different status codes depending on the actual online availability of the resource. The HTTP 200

[24] H. Holzmann, V. Goel, A. Anand, *ArchiveSpark: Efficient Web Archive Access, Extraction and Derivation,* [in:] *16th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, Newark, New Jersey 2016, p. 83–92, https://doi.org/10.1145/2910896.2910902. Accessed 16.09.2021. Source code accessed on GitHub: https://github.com/helgeho/ArchiveSpark. Accessed 16.09.2021.

[25] GitHub. Polish_ccTLD-ia-data, https://github.com/mw0000/Polish_ccTLD-ia-data/. Accessed 18.10.2021.

[26] ArchiveSpark, https://github.com/ibnesayeed/docker-archivespark. Accessed 18.10.2021.

[27] K.R. Blumenthal, *Access Archive-It's Wayback index with the CDX/C API*, https://support.archive-it.org/hc/en-us/articles/115001790023-Access-Archive-It-s-Wayback-index-with-the-CDX-C-API. Accessed 16.09.2021.

[28] MDN Web Docs. HTTP response status codes, https://developer.mozilla.org/en-US/docs/Web/HTTP/Status. Accessed 16.09.2021.

response status code was chosen as confirmation of the availability of the resource, while the others were considered as errors;

- mime (*mimetype*) – type of archived resource (file type and format, e.g. text/html for a web page, image/jpeg for a JPG image, etc.).

The data collected for each year result from multiple crawls, with different seed lists[29] and different scopes – IA did not perform archivisation of the resources of only the PL domain and did not perform annual crawls[30]. Because of this, the data come from the global collection of the Internet Archive foundation, created by multiple archival projects run by the foundation and its institutional partners. As an example, data for the hum.amu.edu.pl hostname (part of the resources of the Adam Mickiewicz University in Poznań), collected during the years 1996––2001, are available as part of 16 different collections of WARC and CDX file catalogues, generated for various time ranges[31]. This scattered data was integrated in the Internet Archive by filtering by domain (only Polish ccTLD resources) and time (1.01.1996–31.12.2001).

Results

1. Basic values and archiving intensity

Table 1 shows the basic values for the analysed collection. The analysis was performed on over 72 million captures. Each URL from the collection was archived at least once between 1996–2001. A comparison of mean and median numbers of captures per distinct URL for some years (1996–2001) shows a strong variation in the intensity of archiving resources for the Polish country domain. An illustration of this could be the fact that the most archived resource in the collection is the banner ad for the Rzeczpospolita newspaper – over the entire

[29] The seed list is a set of URLs from which the web crawler starts to archive Web resources.

[30] Within studies of the historical Web, only taking into account correct server responses (status code 200) is not the only correct approach. For instance, in their analysis of the resources of the historical Danish Web, Niels Brügger et al. used server responses with status codes from 200 to 599, i.e. those informing of errors or redirection to another address. See: N. Brügger, J. Nielsen, D. Laursen, *Big data experiments with the archived Web: Methodological reflections on studying the development of a nation's Web*, "First Monday" 2020, 25, 3, https://doi.org/10.5210/fm.v25i3.10384. Accessed 16.09.2021.

[31] https://web.archive.org/web/collections/20010601000000*/hum.amu.edu.pl. Accessed 16.09.2021.

period studied, it appears over 164 000 times[32]. Despite this, it was first archived only in July 2000[33]. At the same time, the varied and sometimes large number of captures does not translate into a large number of versions of the archived resource (Table 2).

Interpreting the low median values for captures per distinct URL and distinct digests per URL we should consider the problem of the variability of Web resources during the year: a low number of captures translates to a low number of archived versions. On the other hand, a low number of archived versions could suggest that the variability of archived content was not high, while the 'content drift' effect was limited[34]. Captures for the Polish country domain were certainly not performed regularly, nor with the need to register changes within the archived resources. This type of archivisation should certainly be performed by a national Web archive initiative. The low archiving intensity may call into question the validity of using the data collected by IA as a representation of the live Web of the past.

Table 1. Basic numerical values for the analysed collection. Internet Archive data for the Polish ccTLD.

Year	captures	distinct URLs	captures per distinct URL (mean)	captures per distinct URL (median)
1996	161318	72229	2.2	2
1997	2781718	1087617	2.5	2
1998	2250522	875773	2.5	2
1999	4578548	1454305	3.1	2
2000	15300386	3565614	4.2	2
2001	47162716	13918167	3.3	2

[32] https://github.com/mw0000/Polish_ccTLD-ia-data/blob/main/digestsPerUrl.ipynb. Accessed 18.10.2021.

[33] https://web.archive.org/web/20000815000000*/http://www.rzeczpospolita.pl:80/gifs/rek1.gif. Accessed 16.09.2021.

[34] Content drift is a term to indicate changes in the content of a resource continuously available at the same URL. See: S.M. Jones et al., *Scholarly context adrift: three out of four URI references lead to changed content*, "PLOS One" 2016, 11 (12), p. 1–32, https://doi.org/10.1371/journal.pone.0167475. Accessed 16.09.2021.

Table 2. Variability of archived resources over the year. Internet Archive data for the Polish ccTLD.

year	distinct digests per URL (mean)	distinct digests per URL (median)
1996	1	1
1997	1	1
1998	1	1
1999	1.1	1
2000	1.3	1
2001	1.1	1

The representativeness of the collected data, if it were to be treated as the whole, is also undermined by the time structure. Data from the Internet Archive do not represent the resources from the Polish country domain in 1996–2001 evenly: captures from the first four years (1996–1999) represent only 13.5 percent of captures, while captures from 2001 represent more than all archivisations from previous years, see Table 3.

Table 3. Time structure of analysed collection. Internet Archive data for the Polish ccTLD.

year	Captures
1996	0.2 percent
1997	3.9
1998	3.1
1999	6.3
2000	21.2
2001	65.3

Within the collection, resources from 2001 clearly dominate, while the 1990s are represented to a very limited extent. We should also consider that URLs archived in the early years could also have been successfully archived in the subsequent ones. Or that just because a resource was archived in a certain year, does not mean it was not available before (e.g. a page published in 1998 could have been archived only in 1999). Due to the fact that the time structure of the collection is dominated by resources from 2001, while those from 1996–1999 are relatively rare, the value of an analysis based on the entire collection is very limited – an analysis of the data by a specific year would seem to be a better solution.

2. Numerical estimates for the live Web

Despite the limitations described above, the number of distinct URLs by year can be used to estimate the **minimum** numbers for the live Web. In that case, it is not significant when the resources available at the time were produced, but simply that they were available during the studied period. One issue is the proper definition of the available resource – should it be solely a distinct URL with a 200 status code? Table 4 shows distinct URLs for each year, taking into account all HTTP 200 response status codes, the correct ones which allow content to display in a browser. For comparison, numerical values are also given for registered domains within the Polish ccTLD for each year, obtained from NASK and IDS.

Table 4. Distinct URLs with HTTP 200 response status code and number of domains registered within the Polish ccTLD, taken from NASK and IDS data. IDS data since 1998 takes into account both level 2 and level 3 domains of the Polish ccTLD.

year	distinct URLs (HTTP 200)	domains (NASK)	domains (IDS)
1996	72166	855	1408[35] (July)
1997	1087617	5309	4755[36] (July)
1998	859477	13947	12263[37] (July)
1999	1152859	35740	24171[38] (July)
2000	2951183	88958	57003[39] (July)
2001	12211565	134721	100846[40] (July)

The distinct URL numbers for each year can contain addresses created either earlier or later, e.g. the data for 1997 could contain pages published in 1996, and so on. Which is why Table 4 **cannot document the growth dynamics for**

[35] https://web.archive.org/web/20031209185659/http://www.isc.org/ds/WWW-9607/dist-bynum.html. Accessed 16.09.2021.

[36] https://web.archive.org/web/20031015210857/http://www.isc.org/ds/WWW-9707/dist-bynum.html. Accessed 16.09.2021.

[37] https://web.archive.org/web/20031015114419/http://www.isc.org/ds/WWW-9807/dist-bynum.html. Accessed 16.09.2021.

[38] https://web.archive.org/web/20030626203013/http://isc.org/ds/WWW-9907/dist-bynum.html. Accessed 16.09.2021.

[39] https://web.archive.org/web/20031009175444/http://www.isc.org/ds/WWW-200007/dist-bynum.html. Accessed 16.09.2021.

[40] https://web.archive.org/web/20031209090138/http://www.isc.org/ds/WWW-200107/dist-bynum.html. Accessed 16.09.2021.

Internet resources within the PL domain year on year. For instance, it is not possible to state that between 1996 and 1997, the number of URLs within the PL domain increased by 1400 percent, since we do not know which part of the resources archived in 1997 were created in 1996. The number of distinct URLs for each year is however able to show the **minimum numerical state of the Polish Web for that period**. Since for 1996, various crawling projects were able to successfully (HTTP 200) gather copies of resources published under 72 166 addresses within the PL domain, we can state that in 1996, the Polish country domain contained not less than that number of URLs. In 1997, not less than 1 087 617 addresses, etc.

Additionally, the drop in the number of URLs in 1999, compared to 1998, does not mean that the resources within the PL domain had shrunk. It is impossible to demonstrate a downward trend, if only because we are using minimum values, while addresses archived in a specific year could have been created and available in previous ones, even if they were not archived. Secondly, shrinkage of the resources within the country domain would go against both the trend of an increase in the number of registered domains within the Polish ccTLD and that of a global increase in Web resources[41]. Finally, it should simply be considered that as part of the various crawling projects which transferred their data to the IA, less resources were collected for 1998 for the Polish domain than for 1997.

How can we use minimum numbers of distinct URLs to interpret the history of the Polish country domain over the whole period being studied? Firstly, we can carefully point to a general upward trend, yet without specifying its dynamics – this requires us to ignore the decreases between 1997 and 1998 and to refer to the upward trend in the number of registered domains within the Polish ccTLD. Between 1996 and 2001, it is certain the PL domain expanded, but the dynamics of this growth cannot be precisely determined. Secondly, the number of distinct URLs can be compared to the data collected for the old Web from other top-level domains. However, this requires further analysis of the collected data and filtering by hostname and specific mime types (see next sections).

[41] Data collected by Matthew Gray shows a constant increase in the number of sites between 1993 and 1997, https://www.mit.edu/people/mkgray/net/web-growth-summary.html. Accessed 16.09.2021. Further growth is confirmed by the research of Bharat and Broder (1998) i Lawrence and Giles (1999), see: A. Trotman, J. Zhang, *Future Web Growth and its Consequences for Web Search Architectures*, arXiv.org, p. 4, https://arxiv.org/abs/1307.1179. Accessed 16.09.2021.

3. Hostnames and domains

A hostname (host) is part of the URL, complemented by a higher-level domain name or TLD, and can be regarded as a common denominator for a set of URLs (a web site[42]). For many sites, the hostname will designate the home page or index page for resources and be part of the address of subpages. Access to the URL list allows one to freely delineate hostnames and examine their statistics[43]. These data provide information on the structure of the captures for each year, and allow to interpret their representativeness in terms of the live Web and also to study the evolution of individual web sites.

Tables 5 and 6 show the hostnames statistics for the collected URLs. The absolute number of identified hostnames does not say much about the structure of the annual data, unless you compare it to the number of all captures (Table 5.). A higher value of this quotient indicates greater diversity and a lower concentration of captures. Table 6 provides the number of web pages (understood as text/html documents) for a single hostname. The data for 1997 stand out: there is little variation (small number of hostnames relative to the number of captures), and the mean and median numbers of pages per hostname are the highest for the entire 1996–2001 set. They can hardly be considered representative of the live Web.

Similarly to distinct URLs with HTTP 200, the number of hostnames allows to estimate the minimal reach of the live Web. Irrespectively of how the individual archiving actions were profiled and how little they varied, we could conclude that there were at least 876 hostnames available in the Polish ccTLD in 1996, 2107 in 1997, etc., and that the minimum numbers of hostnames increased throughout the period studied.

Table 5. Captures and distinct hostnames. Internet Archive data for the Polish ccTLD.

year	captures	distinct hostnames	hostnames / captures
1996	161318	876	0.005
1997	2781718	2107	0.0007
1998	2250522	10978	0.0048

[42] Baeza-Yates et al. use the terms *Web sites* and *hosts* interchangeably. R. Baeza-Yates, op.cit., p. 3.

[43] For this paper, hosts have been extracted using generic function from Spark SQL, https://spark.apache.org/docs/latest/api/sql/#parse_url. Accessed 16.09.2021.

year	captures	distinct hostnames	hostnames / captures
1999	4578548	28795	0.006
2000	15300386	39406	0.002
2001	47162716	392965	0.008

Table 6. Distinct web pages per hostname, Internet Archive data for the Polish ccTLD.

year	distinct pages per hostname (mean)	distinct pages per hostname (median)
1996	43.48	1
1997	317.19	21
1998	52.79	3
1999	46.24	2
2000	73.9	3
2001	27.9	4

Table 6 also shows mean and median numbers of web pages per hostname, although pages returning HTTP 200 as well as errors and redirects were counted. These data cannot document changes in the mean and median sizes of the web sites on the live Web, rather, they show changes in the depth and concentration of archivisation. Thus, 1997 would be the year not of the largest websites within the PL domain, but of the most focused and deepest archivisation. Baeza-Yates et al. present median numbers for sub-pages within websites (hostnames) and demonstrate their wide variation within the analysed data: Brazil 66; Chile 58; Indochina 549; Italy 410; Greece 150; South Korea 224; Spain 52; United Kingdom 248[44]. These data were collected from varied queries performed between 2000 and 2004[45] and ignore single-page websites – their usefulness as context for the data in Table 6 is limited, with the exception that they document a wide variation in the number of pages per hostname (from 52 to 549). As it is indicated, the seemingly large number of pages within Indochinese, Italian and British websites results from dynamically-generated URLs (e.g. through session IDs). The use of hostnames in studying the historical Web therefore requires care and manual inspection, which in the case of tens or hundreds of thousands of hostnames is hardly practical.

[44] R. Baeza-Yates, op.cit., p. 10

[45] In Table II, they even describe the limits of the depth of site crawling (p. 6).

Table 7. Presence of hostnames over analysed period. Internet Archive data for the Polish ccTLD.

since year	Hostnames
1996	0.11 percent
1997	0.24
1998	1.64
1999	3.36
2000	5.31
2001	89.33

The structure of the data available within the Internet Archive does not allow for a proper analysis of the history of hostnames during the studied period – only 0.1 percent of the hostnames from 1996 to 2001 have data on the number of sub-pages, while 89 percent have this data only for a single year (Table 7). Since the age of the vast majority of hostnames is just one year, we can determine that crawl indexes for the following year nearly never use those from the previous one – generally, crawl requests were not repeated over the years.

Table 8. Share of functional domains within all Polish ccTLD web pages per year. Internet Archive data for the Polish ccTLD.

year	COM.PL	EDU.PL	GOV.PL	ORG.PL	PL
1996	9.3	24.1	2.8	0.3	22.1
1997	13.9	19.4	4.9	0.8	24.9
1998	25.1	16.4	2.5	3.1	22.6
1999	21.1	15.5	7.3	1.9	35
2000	16.7	10.7	2.9	1.7	48.1
2001	17.3	3.9	1.3	1.7	60.8

With access to hostnames and the number of their sub-pages, it is possible to calculate the number of sub-pages for regional and functional domains within the Polish ccTLD, as well as their proportion within the whole set for each year (Table 8). Again, not in order to calculate the size of the ‘commercial’ Web (COM.PL) or the government domain (GOV.PL), but rather to distinguish changes in the structure of the data and to reflect upon their significance in reconstructing the historical Web. For example, do changes to the proportion of the educational

Web sphere within the collected data (decrease from 24 to 3.9 percent, despite a constant increase in the number of web pages) suggest that between 1996 and 2001, the content available within the Polish country domain was aimed less at professionals and academics and became more varied? Confirming this hypothesis would require obtaining different data, e.g. from printed Internet guides, Web directories or contextual information such as the social profile of Internet users. The rising number of domains registered within the Polish ccTLD would suggest an increased differentiation of the Web. Unfortunately, for the period which interests us, we only have information about the number of Polish domains, but not their types (first level, secondary, etc.). In the analysed data, the most complete representation of hostnames diversity can be found for 2001 (see Table 5) – in that year, domains registered directly under Polish ccTLD accounted for 60 percent of all web pages (see Table 8)[46].

Identifying the hostnames also allows to clearly show the degree of formatting of the obtained data introduced by the configuration of the crawling projects. Web pages assigned to the largest web hostname in a given year represent from 0.4 (1998) to up to 9.6 percent (2000) of all web pages. If in 1999 and 2000 the web pages of the www.rzeczpospolita.pl website represented one-tenth of all web pages it is difficult not to speak of concentrated, directional data. Their representativeness in terms of the entire Web must be even more limited because of this.

Thus, it is obvious that Internet Archive data cannot be a replacement for a national Web archival effort, focused not only on the national domain, but also on collecting longitudinal data for hostnames. When the data are fragmentary and formatted to such an extent by the crawler configuration, and also represent a small number of hostnames, the evolution of a specific web site cannot be described properly. These data allow to at least show the increasing differentiation of the resources within the PL domain, but it is impossible to describe the dynamic of this growth.

[46] Analysis performed using the httr R package with suffix extraction. O. Keyes et al., *urltools: Vectorised Tools for URL Handling and Parsing*, https://cran.r-project.org/package=urltools. Accessed 16.09.2021.

4. HTTP status codes

During crawling, server responses with different status codes are archived, informing not only of correct responses (200–299), but also of server-side (500–599) and client-side (400–499) errors, including Not Found (404) and redirects (300–399). Table 9 shows the proportion of status codes other than HTTP 200 in all captures for a given year. Response status code 200 means that the requested resource has been fetched and transmitted in the message body.

Table 9. Proportion of status codes other than HTTP 200 in all captures for a given year. Internet Archive data for the Polish ccTLD.

Year	status codes other than 200
1996	0.08
1997	0
1998	1.5
1999	19.27
2000	16.59
2001	22.94

Table documents a formatting problem in Internet Archive data. The lack of any errors, or a low proportion, in the annual data (visible for 1996–1998), suggests that the crawling processes performed at that time could have been specifically configured to archive only correct server responses. Unfortunately, it was unable to obtain any comments about the data from the Internet Archive. As a comparison point, the data collected for national collections by Baeza-Yates et al. suggest a proportion of correct responses in the 80–85 percent range (2000–2004)[47]. Within the Common Crawl project, which collected information about the contemporary Web, the proportion of correct HTTP 200 status code responses is around 70 percent (data from June 2001)[48].

The data for HTTP status codes of archived collections can be used to study the link rot phenomenon. Due to the specific crawl settings used to generate data, it is impossible to do this for 1996–1998. Data from 1999 onwards indicate a problem in reaching some Web resources – more than 20 percent of the requests sent in 2001 received statuses indicating some errors or

[47] https://github.com/commoncrawl/cc-crawl-statistics/blob/master/stats/crawler/CC-MAIN-2021-25.json. Accessed 16.09.2021.

[48] R. Baeza-Yates, op.cit., p. 21.

redirections – thanks to the presence of these errors, it is possible to study the change in availability of specific URLs. They may suggest changes in the structure of specific websites and serve as an inspiration to perform qualitative analysis of individual captures.

5. Mime-types

CDX metadata allows us to count URLs by their format (mime type) without looking into the contents of the archived file. It would be tempting to study changes in the quantity of image or multimedia files, for instance to demonstrate that during the studied period, the Polish Web was becoming more visual and image-based. Unfortunately, due to the structure of the data, attempting such an analysis would be risky.

Firstly, data obtained from the IA contain significant imprecisions concerning mime types, in particular those for 1998–2000. These data were generated from specific crawling projects, which ignored traditional mime types such as image/jpeg or image/gif and classified available resources of this type as im. Secondly, the collection profile for individual years can result in a greater or lesser presence of image files in the collection – the scholarly Web sphere will be visually different from hostnames in the news portal category. Thirdly, the quantity of image files itself does not say much about the visual nature of the Web because of a 'spacer GIF' effect (tiny transparent files used to control the visual layout[49]) or the widespread use of small image files as icons on pages. In effect, a web page may contain many references to image files, but still have a textual nature[50]. To study the visual nature of the Web would require calculating the surface areas of image files published on pages more than the number of IMG tags (which is possible through WARC files). This data can be obtained from the IMG tag attributes (width, height) or from a visual snapshot of the page[51].

[49] Wikipedia. Spacer GIF, https://en.wikipedia.org/wiki/Spacer_GIF. Accessed 16.09.2021.

[50] As an example, the page at https://web.archive.org/web/20001217103500/ https://sigma.wsb-nlu.nowy-sacz.pl:80/local/stats/accesswatch/report/page.html is text-based, even though it contains over 2000 image files (small GIFs).

[51] For example, by using the Stroke Width Transform (SWT) algorithm. See: A. Cocciolo, *Quantitative Web History Methods,* [in:] *The SAGE Handbook of Web History*, op.cit., p. 145 and following.

Although the numbers of specific file types would seem to be of limited significance, the date of their first appearance in a collection can be interesting. It can be used to define ranges which show the introduction of new technologies within the Polish Web. For example, Cascading Style Sheets (CSS) first appear in the 1997 collections for the PL domain (their initial release was in December 1996). Portable Network Graphics (PNG) first appear in 1998 (initially released in October 1996). Flash files first appear in 2001, but had established themselves globally by early 2000. On the other hand, it is not only the presence of archived copies of certain file types which demonstrates their presence, HTML tags within web pages also provide this information.

Tables 10 and 11 point to a formatting of archived Polish country domain data. The low proportion of HTML files within the collections stands out (slightly above 50 percent in 1996). In the data collected by Rauber et al. for the Austrian Web, HTML files represented over 95 percent[52]. In the data collected by Baeza-Yates et al., this was also above 95 percent[53]. Currently, based on Common Crawl data, this proportion stands at over 83 percent[54]. Table 11 shows the variation in the annual collected data in terms of the number of different file types. As mentioned, the 1998–2000 data are specific in this respect, especially 1999. The more file types, the more varied the resources which were archived.

The relatively low proportion of HTML in the archived collections suggests that crawls archived hostnames to a low depth which resulted in a low number of archived sub-pages (see Table 6), that the hostnames were of a small size or linked to resources which were not Web pages (e.g. ZIP, TXT, etc. files) or had embedded images (like the ‘spacer GIF’ effect). A better overall picture would be provided by an analysis of the links and tags available in the source code of archived pages. This would also allow to step beyond the archived resources, since links and tags could point to files which were not secured but which existed on the historical live Web.

[52] A. Rauber, op.cit., p. 6.

[53] R. Baeza-Yates, op.cit., p. 22.

[54] https://commoncrawl.github.io/cc-crawl-statistics/plots/mimetypes. Accessed 16.09.2021.

Table 10. HTML pages among distinct URLs. Internet Archive data for the Polish ccTLD.

Year	HTML
1996	52.6
1997	60.5
1998	66.1
1999	90.2
2000	79.6
2001	76.9

Table 11. Number of mime types among distinct URLs. Internet Archive data for the Polish ccTLD.

Year	mime types
1996	42
1997	111
1998	102
1999	30
2000	109
2001	253

5. Conclusions

The availability of historical Web resources, and the tools to easily explore them, in no way prejudges their potential value and usefulness in research, even if we do not have access to alternative sources. The analysis of Internet Archive collections for the Polish ccTLD showed their limited usefulness in building a historical picture of the Polish Web sphere, as defined by domain. We are able to determine the minimum size of the Web defined in this way for successive years from 1996–2001. We can prove that in a specific year, files of a specific type were already available, which allows us to compare the technical evolution of the Polish Web to other national Webs. By demonstrating the presence of HTTP redirects and errors, we can document the existence of the link rot problem, which would be best illustrated by showing the history of selected URLs and hostnames. The resources acquired from the IA allow to study selected hostnames, perform linguistic or thematic analysis of selected content, yet without expecting high representativeness. However, the way the Polish ccTLD data was gathered in the Internet Archive, as well as its structure, make quantitative analyses of the evolution of the entire Polish domain, or documenting changes in its diversity, impossible. Such analyses require regularly collected, stable sets of seeds, allowing for the study of changes in addresses and hostnames over time – only a Web archive, built by a national remembrance institution as part of its mission, can provide them.

Zeynep Tufekci points out several fundamental limitations of research using big data resources[55], her classification of the issues involved can be reused to build appropriate methods for analysing Polish ccTLD resources within Internet Archive: 1) platform bias – the Internet Archive and Wayback Machine are an obvious, and sometimes only, source of historical Web resources, but they do not collect them in an unadulterated state. The published collections are influenced by the methods used to acquire them; 2) selecting for dependent variables without taking the requisite precautions – none of the elements present in the analysed resources possess any objective meaning. For example, the analysis of changes in the number of image files on specific sites must take into account the issue of spacer GIFs, while the interpretation of links must take into account their various possible meanings[56]; 3) the denominator problem created by vague, unclear or unrepresentative sampling – we do not know what part of the Polish ccTLD resources available in 1996–2001 in the live Web were collected by the Internet Archive, therefore any knowledge built using them may not apply to the whole national Web. Tufekci also refers to the problem of the actual visibility of the analysed resources (in her study, this concerns tweets on Twitter) – we do not know which of the sites of the early Polish Web were popular and those which were not, or what their viewership levels were. Should academic websites be seen as representative of the national domain at the time, or should the first Web portals? 4) the prevalence of single-platform studies which overlook the wider social ecology of interaction and diffusion – paradoxically, in research on the early Web, historical Web resources do not have to be the most important source, in particular if they are only available in a partial form. A historical perspective on the medium requires an appreciation of its ecology, the forms and ways in

[55] Z. Tufekci, *Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls*, [in:] *Eighth International AAAI Conference on Weblogs and Social Media*, Association for the Advancement of Artificial Intelligence, 2014, p. 505–514.

[56] A. Helmond, *A Historiography of the Hyperlink: Periodizing the Web Through the Changing Role of the Hyperlink*, [in:] *The SAGE Handbook of Web History*, op.cit., p. 227–241.

which it exists beyond the Internet – newspaper commentaries, websites mirrors distributed on compact discs with printed computer magazines, printed indexes or guides to the Web. But it also requires knowledge of social, material realities: the level of Internet access and the quality of connections, the availability of computers in schools or households, the class and professional profile of early users. After all, the Web is not just a collection of sites, but a social practice.

Bibliography

Baeza-Yates R., Castillo C., Efthimiadis E.N., *Characterization of national Web domains*, "ACM Transactions on Internet Technology" 2007, 7, 2, p. 1–32, https://doi.org/10.1145/1239971.1239973. Accessed 16.09.2021.

Ben-David A., *Critical Web Archive Research,* [in:] *The Past Web: Exploring Web Archives*, D. Gomes, E. Demidova, J. Winters, T. Risse, eds., Springer Nature Switzerland, Cham 2021, p. 181–188, https://doi.org/10.1007/978-3-030-63291-5_14. Accessed 16.09.2021.

Ben-David A., Amram A., *The Internet Archive and the socio-technical construction of historical facts*, "Internet Histories: Digital Technology, Culture and Society" 2018, 2, p. 1–23, https://doi.org/10.1080/24701475.2018.1455412. Accessed 16.09.2021.

Bingham N.J., Byrne H., *Archival strategies for contemporary collecting in a world of big data: Challenges and opportunities with curating the UK web archive*, "Big Data & Society" 2021, 8, 1, p. 1–6, https://doi.org/10.1177/2053951721990409. Accessed 16.09.2021.

Brügger N., *When the Present Web is Later the Past: Web Historiography, Digital History, and Internet Studies*, "Historical Social Research" / "Historische Sozialforschung" 2012, 37, 4 (142), s. 102–117.

Brügger N., Ditte L., *Historical studies of National Web Domains*, [in:] *The SAGE Handbook of Web History*, N. Brügger, I. Milligan, eds., Sage, Los Angeles–London–New Delhi 2018, p. 413–427.

Brügger N., Nielsen J., Laursen D., *Big data experiments with the archived Web: Methodological reflections on studying the development of a nation's Web*, "First Monday" 2020, 25, 3, https://doi.org/10.5210/fm.v25i3.10384. Accessed 16.09.2021.

Cocciolo A., *Quantitative Web History Methods*, [in:] *The SAGE Handbook of Web History*, N. Brügger, I. Milligan, eds., Sage, Los Angeles–London–New Delhi 2018, p. 138–152.

Denev D. et al., *The SHARC framework for data quality in Web archiving*, "The VLDB Journal" 2011, 20, p. 184–207, https://doi.org/10.1007/s00778-011-0219-9. Accessed 16.09.2021.

Foot K., *Web Sphere Analysis and Cybercultural Studies*, [in:] *Critical Cyberculture Studies*, D. Silver, A. Massanari, eds, NYU Press, New York 2006, p. 88–96.

Hale S.A., Blank G., Alexander V.D., *Live versus archive: Comparing a web archive to a population of web pages*, [in:] *The Web as History. Using Web Archives to Understand the Past and the Present*, N. Brügger and R. Schroeder, eds., UCL Press, London 2017, p. 45–61.

Helmond A., *A Historiography of the Hyperlink: Periodizing the Web Through the Changing Role of the Hyperlink*, [in:] *The SAGE Handbook of Web History*, N. Brügger, I. Milligan, eds., Sage, Los Angeles–London–New Delhi 2018, p. 227–241.

Holzmann H., Goel V., Anand A., *ArchiveSpark: Efficient Web Archive Access, Extraction and Derivation*, [in:] *16th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, Newark, New Jersey, p. 83–92, https://doi.org/10.1145/2910896.2910902. Accessed 16.09.2021.

Jones S.M. et al., *Scholarly context adrift: three out of four URI references lead to changed content*, "PLOS One" 2016, 11 (12), p. 1–32, https://doi.org/10.1371/journal.pone.0167475. Accessed 16.09.2021.

Kimpton M., Ubois J., *Year-by-Year: From an Archive of the Internet to an Archive on the Internet*, [in:] *Web Archiving*, Julien Masanes, ed., Springer, Berlin–Heidelberg–New York 2006, p. 201–212.

Milligan I., *Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives*, "International Journal of Humanities and Arts Computing" 2016, 10, 1, p. 78–94, https://doi.org/10.3366/ijhac.2016.0161. Accessed 16.09.2021.

Milligan I., Ruest N., Lin J., *Content Selection and Curation for Web Archiving: The Gatekeepers vs. the Masses*, [in:] *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 2016, p. 107–110.

Rauber A. et al., *Uncovering Information Hidden in Web Archives. A Glimpse at Web Analysis building on Data Warehouses*, "D-Lib Magazine" 2002, 8, 12, https://www.doi.org/10.1045/december2002-rauber. Accessed 16.09.2021.

Spaniol M. et al., *Data quality in Web Archiving*, [in:] *Proceedings of the 3rd Workshop on Information Credibility on the Web*, Association for Computing Machinery, New York 2009, p. 19–26, https://doi.org/10.1145/1526993.1526999. Accessed 16.09.2021.

Trotman A., Zhang J., *Future Web Growth and its Consequences for Web Search Architectures*, arXiv.org, p. 1–41, https://arxiv.org/abs/1307.1179. Accessed 16.09.2021.

Tufekci Z., *Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls*, [in:] *Eighth International AAAI Conference on Weblogs and Social Media*, Association for the Advancement of Artificial Intelligence, 2014, p. 505–514.

Vlassenroot E., Chambers S., Di Pretoro E. et al., *Web archives as a data resource for digital scholars*, "International Journal of Digital Humanities" 2019, 1, p. 85–111, https://doi.org/10.1007/s42803-019-00007-7. Accessed 16.09.2021.

Weber M.S., *Web Archives: A Critical Method for the Future of Digital Research*, WARCnet Papers, Aarhus 2020, p. 1–17, https://cc.au.dk/fileadmin/user_upload/WARCnet/Weber_Web_Archives_A_Critical_Method.pdf. Accessed 16.09.2021.

Wilkowski W., *Polish Web resources described in the "Polish World" directory (1997). Characteristics of domains and their conservation state*, "Archiwa – Kancelarie – Zbiory" 2020, 11, 13, p. 119–140, https://doi.org/10.12775/AKZ.2020.005. Accessed 16.09.2021.

Internet Resources

Common Crawl, https://commoncrawl.org//. Accessed 16.09.2021.

GitHub, https://github.com. Accessed 18.10.2021.

Internet Archive, https://archive.org. Accessed 16.09.2021.

Internet Domain Survey Background (2003), https://web.archive.org/web/20031002012504/http://www.isc.org/ds/new-survey.html. Accessed 16.09.2021.

HTTP Archive, https://httparchive.org/. Accessed 16.09.2021.

MDN Web Docs. HTTP response status codes, https://developer.mozilla.org/en-US/docs/Web/HTTP/Status. Accessed 16.09.2021.

SHINE, https://www.webarchive.org.uk/shine. Accessed 16.09.2021.

Spark SQL, https://spark.apache.org. Accessed 16.09.2021.

The World Bank Data, https://data.worldbank.org. Accessed 16.09.2021.

Wikipedia. Spacer GIF, https://en.wikipedia.org/wiki/Spacer_GIF. Accessed 16.09.2021.