Humanum
Międzynarodowe Studia Społeczno-Humanistyczne

Wydawca / Publisher:
Instytut Studiów
Międzynarodowych
i Edukacji HUMANUM
www.humanum.org.pl

41 (2) 2021
ISSN 1898-8431

Przemysław Jatkiewicz
przemyslaw.jatkiewicz@ug.edu.pl
University of Gdańsk Poland

Obsolescence of the term deep web in the light of research on the activity of web users

Summary

The article addresses the issues of redefining the term Deep Web or introducing the term Inaccessible Web, which is a considerably better representation of the availability of web pages for Internet users. It presents the course and results of studying over 2 months of activity of employees in an institution, which indicate that they browse only a small group of internet portals, and this number depends on their age, gender, and education. Also, it has been observed that the browsed content was mainly in the user's native language, and less frequently in the English language. What is also addressed is the availability of web pages for people with disabilities and the obligation to adapt web pages to legally required accessibility standards.

Keywords: **Deep Web, Hidden Web, Inaccessibly Web, Darknet, WCAG.**

1. INTRODUCTION

Darknet, Deep Web, Invisible Net or Dark Internet are terms raising concern among average Internet users, and they are frequently, albeit improperly, used interchangeably. The origin of the term Deep Web is attributed to Mike Bergman, who has described a new, entirely different mechanism of browsing the resources of the Internet. He concluded that it has two distinct types. Surface – possible to find when using popular browsers – and deep, which among other things includes web pages generated dynamically in response to the users' queries, or resources available via protocols other than http. The research he performed indicated that the resources of the deep Internet are 400 to 550 times larger than that of the sur-face Internet and consist of approximately 550 billion documents [Bergman 2001].

C. Sherman and G. Price proposed another term - Invisible Web, covering web pages which, for technical or business reasons, are not indexed by popular browsers [ShermaN 2001:26]. Although this term is much broader than Deep Web, the aforementioned researchers stated that the invisible resources are only 2–50 times larger than the visible – indexed – ones. Both studies were performed in the year 2001. Since then, the capabilities of modern browsers have increased considerably.

The calculations of Bergman were also questioned by Lewandowski and Mayer, who pointed out numerous mistakes in his methodology. They themselves estimated the number of documents hidden within the Internet at no more than 100 billion [Lewandowski & Mayr 2006:7].

Although various authors present various estimations, it is clear that the hid-den re-sources are considerably larger than what is generally available, as depicted in figure 1.

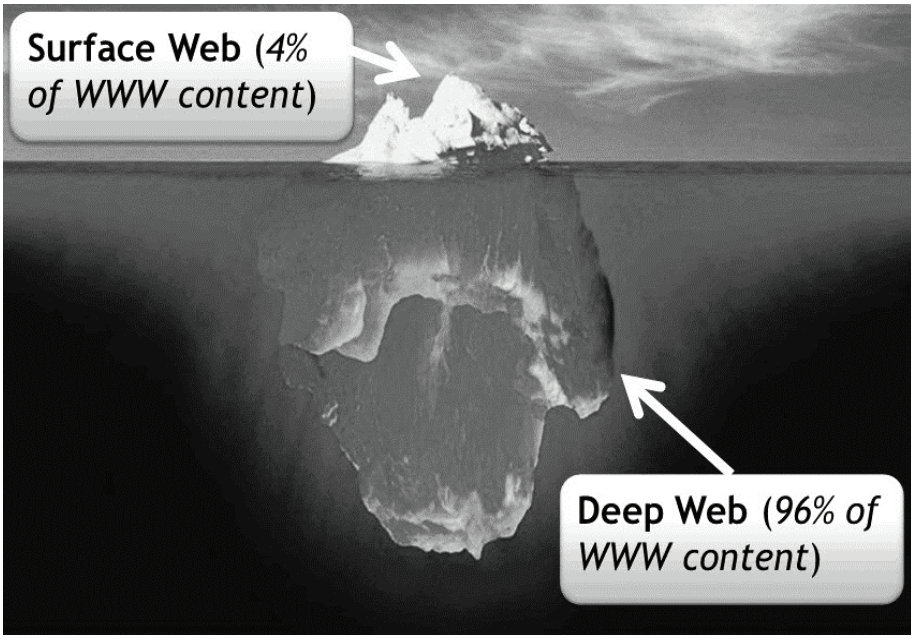

Figure 1. Graphical representation of the proportion of generally available to hid-den web pages (Source: https://gameplay.pl/news.asp?ID=86853).

Most authors of publications related to the deep Internet point to the value and reliability of the information included thereon. The materials they list include, among other things [Derfert-Wolf 2007:3]:

- Scientific publications and reports, dissertations;
- Articles from newspapers and magazines;
- Government documents;
- Archives of source and reference materials;
- Library resources;

- Open Access repositories;
- Grey literature;
- Data, formulas, graphics;
- Dictionaries, encyclopedias, address databases.

As seen above, the terms Deep Web and Invisible Internet do not refer to any-thing dark. Rather, the utilisation of these difficult-to-access web resources re-quires the skills not of hackers, but those of information brokers who profession-ally search for information. The magnitude of the hidden resources also depends on the capabilities of browsers. During the several decades of their history, they went through many changes. Although their algorithms are strictly protected, the following types of them can be distinguished [Ledford 2009:19]:

- Linear search;
- Search trees;
- SQL search (Structured Query Language);
- Search based on solid information;
- Counter-search;
- Search based on limited satisfaction.

However, several algorithms and combinations thereof are usually used. What matters is not just the ability to find a web page containing the desired information, but also its rank, constituting a basis for the sequence of displaying the results. As indicated by the research performed in 2020 by Ignite Visibility [Lincoln 2021], most users do not go beyond checking the first two pages of results. Almost 45 % of them only use the first link, which is presented in figure 2. Therefore, it can be concluded that even websites which have been found and are ranked low remain invisible for an average user.

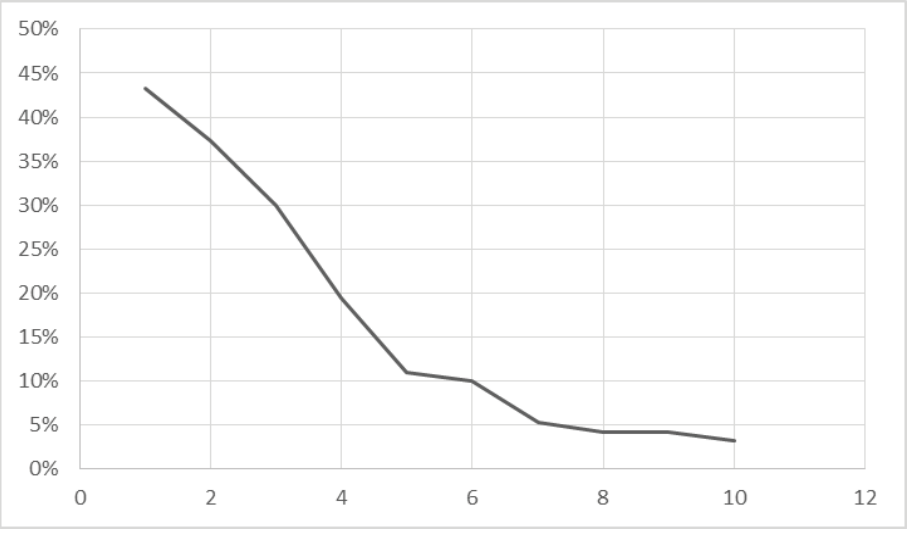

Figure 2. Graphical representation of the proportion of generally available to hidden web pages (Source: https://gameplay.pl/news.asp?ID=86853).

The situation is different in the case of the Dark Internet, or its equivalent term, dark address space. It means hosts which cannot be accessed at all or in a very limited manner, depending, e.g., on location in the web. This problem was investigated by Craig Labovitz in 2001 [Labovitz 2001:1]. As a result of 3 years of research involving analyses of routing tables of main Internet providers, it has been concluded that, due to the misunderstandings of operators, configuration errors, mal-functions, too-rigorous filtering rules, and malicious activity, there has been a permanent division of network topology and the isolation of a number of IP ad-dress ranges, the access to which is considerably hindered. It has been estimated that the problem concerns over 5 % of all addresses, and thus tens of millions of hosts. However, most of them, i.e., 78 %, can be reached. In-depth studies have proved that the aforementioned address spaces were usually assigned to DSL modems, or they represent unused, historical allocations for military networks of the United States.

The fact of the brief emergence of previously invisible hosts has also been observed. They vanished quickly after a short time of operation. These hosts, hid-den behind routers taken over by hackers, are used for unlawful activities.

A really dark side of the Internet is represented by the darknet. They are net-works in-tended to store and transfer information, ensuring the anonymity of their users. One can-not generalise and claim that anyone who wishes to remain anonymous online has a tendency for antisocial behavior, but it is a fact that on the darknet it is possible to find web pages with pedophilic material and instructions for creating and using explosives or illicit drugs. Similar information can also be found in the remaining, public part of the Internet, but it is scarce and easy to track by law enforcement authorities. The two most popular virtual net-works providing a high level of anonymity, frequently used to spread illegal con-tents and services, are TOR (The Onion Router) [Dingledine et al. 2004:1] and FreeNet (Free Network) [Clarke et al 2001:46].

The vast majority of previous research divided the network into Deep Web and Surface Web, based on the technical capabilities of Internet browsers. Even when web pages can be found without the necessity to involve specialised browsers, this does not mean that every Internet user becomes their recipient. This is because of barriers, such as the language in which the content is presented, as well as the siloisation of access, the principle of which is that users become acquainted with information presented on a relatively small number of portals. They focus on individual portals aggregating their services and they use a single selected browser. Therefore, the current understanding of the term 'deep web' seems inadequate.

2. METHODOLOGY AND COURSE OF RESEARCH

The research was performed in a territorial government unit in northern Poland. The organisation hired 320 workers, including 60 manual labourers who do not use computers when performing their work. 1 million records containing information on the visited web pages were downloaded from the Palo Alto PA-500 firewall device protecting the network. The data were recorded during a period from 21 January 2021 to 7 April 2021, so they covered over 2 months of activity of both the employees of the unit and clients who used the available Wi-Fi network. They were subjected

to filtration, which allowed for the deletion of URL addresses (Uniform Resource Locator) which were undeniably related to the performed work, i.e., Intranet addresses of the organisation's web page, MS Teams and Office 365 pages. Data related to hardware and software updates were also deleted. After the filtration, 416 259 records were left remaining, each with the following structure:

- Login;
- URL;
- Category.

The record of URL addresses was standardised to forms indicating the home pages of portals, and the data were supplemented with the language of the web page. The login, combined with the data contained in the Active Directory and the data of the HR department, allowed for establishing the age bracket (under 30, between 30 and 50, above 50 years old) of the employees and their education (secondary or higher). Categorisation of the visited web pages was performed automatically by data-collecting devices. The numbers of visits of pages in the individual categories are presented in table 1. In one of the most frequently visited categories of pages, i.e., 'computer-and-internet-info', there can be data related to pages activated without the user's participation, e.g., due to the actions of scripts located on other pages.

Table 1. The numbers of visits of pages in the individual categories (Source: own study).

Category	Number	Category	Number
Adult	25	Unresolved	2
Alcohol and tobacco	416	Online storage and backup	750
Auctions	18	Personal sites and blogs	1483
Business and economy	72477	Philosophy and political advocacy	1049
Computer and internet info	66770	Proxy avoidance and anonymisers	28
Content delivery networks	22993	Real estate	549
Dating	167	Recreation and hobbies	3
Educational institutions	1472	Reference and research	2143
Entertainment and arts	3501	Religion	351
Financial services	8380	Search engines	49894
Gambling	1063	Shareware and freeware	74
GIS	323	Shopping	12688
Government	11561	Social networking	22373
Health and medicine	1145	Society	1238

Category	Number	Category	Number
Home and garden	412	Sports	140
Insufficient content	841	Stock advice and tools	204
Internet communications and telephony	4124	Streaming media	19452
Internet portals	923	Training and tools	475
Job search	260	Translation	606
Legal	333	Travel	3272
Malware	7	Unknown	1
Military	2	Vehicle	1
Motor vehicles	583	Web advertisements	51551
Music	1067	Web-based email	810
News	47500	Web hosting	759

An analysis of the data allowed for establishing that, during the studied period, the employees visited only 2130 web pages with unique addresses. An additional 237 pages were visited by clients. The median of the number of web pages visited by the employees is 28. However, for the news category it equals 4, and for the search-engines category it is 2. The numbers of visits to the most popular portals in the news category are presented in table 2, and figure 3 presents the most frequently used Internet browsers.

Table 2. The numbers of visits to the most popular Internet portals in the news category (Source: own study).

Service	Number of visits
interia.pl	3501
trojmiasto.pl	4357
msn.com	15151
wp.pl	16420

It should be noted that among portals listed in table 2, trójmiasto.pl has a completely different nature, since, unlike the others, it contains mainly regional information, i.e., related to the so-called Tri-City area of Gdańsk, Gdynia, and Sopot. Also, the high result of the msn.com portal results from the fact that directly after installing the Windows operating system, which constituted a standard in the studied organization, the default web search engine is bing.com, and the start-up page is msn.com.

Most users restrict themselves to using a small number of 1 or 2 thematic services, i.e., belonging, e.g., to the categories of motor vehicles, dating, educational institutions,

health and medicine or financial services. A high diversity of addresses is observed in the computer and Internet info as well as business and economy categories.

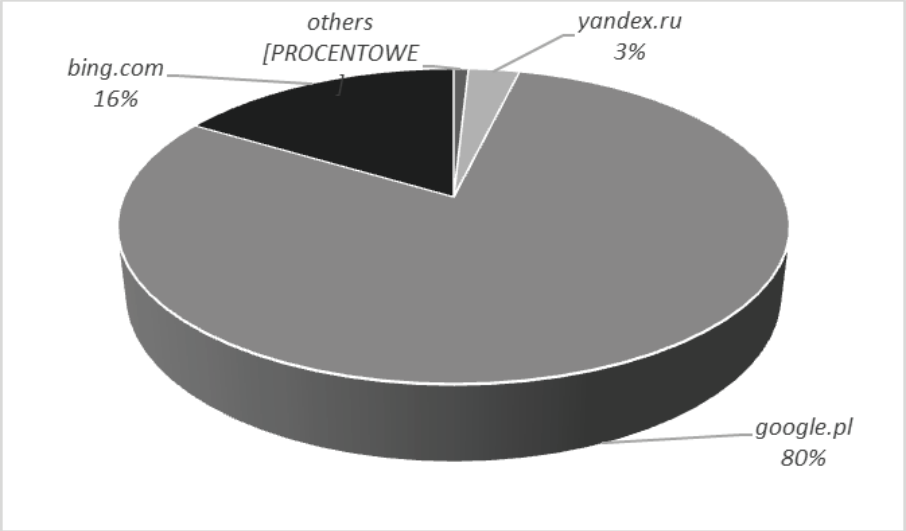

Figure 3. The most-frequently used Internet browsers (Source: Own study).

It was observed that women, having opened web pages only 213 times on average, were much less active during the studied period compared to men, who opened them as many as 1 760 times on average.

Spearman's rank correlation test was performed in order to study the relationship between the age of users and the number of pages visited by them. Spearman's coefficient was calculated using the following formula [Gauthier 2001:359]:

$$r_s = -\frac{6 \times \sum_{i=1}^{n} d_i^2}{N(N^2-1)} \quad (1)$$

where: d – difference between ranks

N = 263 – number of observations (employees)

The resulting value of 0.93 proves that there is a strong and positive correlation be-tween the age and the number of browsed web pages for the studied employees.

Spearman's rank correlation test was also used to determine the relationship of the education feature with the number of visited pages. The calculated value of 0.81 also proves positive correlation between the number of browsed pages and education, although this relationship is not as strong as that with respect to age.

Another barrier in accessing the contents of the Internet is the language in which they are presented. As indicated by the research of W3Techs [W3Techs 2021], the English language is used by 61.1 % of all portals. Portals using the Polish language constitute

only 0.6 %. The studies performed by the author resulted in the identification of 14 different languages of the web pages browsed by the employees, the predominant languages being Polish and English, as presented in figure 4.

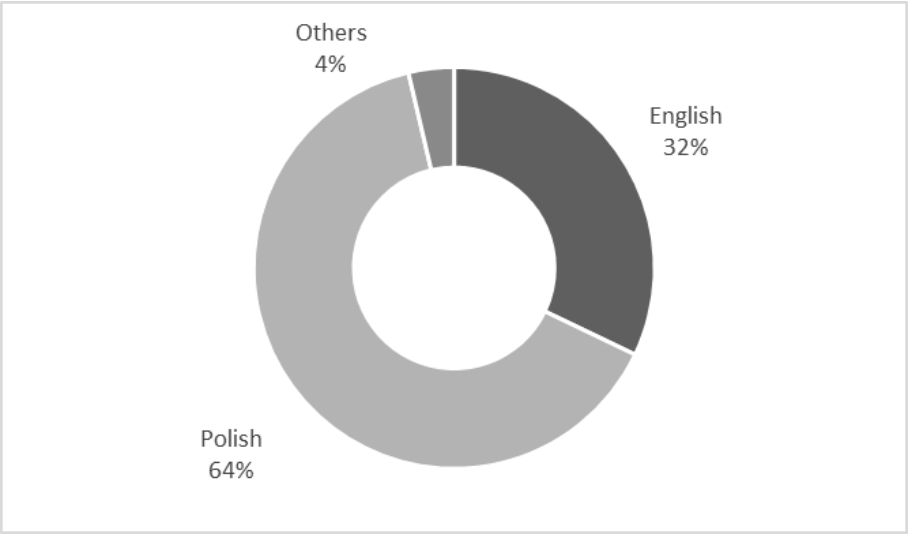

Figure 4. Languages of the browsed pages (Source: Own study).

The small number of portals in languages other than Polish and English indicates with a high probability that they were not visited on purpose, and the access to them resulted from following links from other pages.

The problem of accessibility to contents published on the Internet for people with disabilities was addressed by Tim Berners-Lee [Berners-Lee 1994] during the second international World Wide Web conference in Chicago. Already in the year following, Trace Research & Development Centre at the University of Maryland prepared a document titled Unified Web Site Accessibility Guidelines, whose 8th version became a basis for the WCAG 1.0 guidelines [Paciello 2000:41], which were published on 5 May 1999 under the WAI (Web Accessibility Initiative) scheme managed by the W3C (World Wide Web Consortium).

The WCAG guidelines are intended to facilitate access to the contents of web pages for people with:

- Visual impairments;
- Hearing impairments;
- Difficulties in understanding;
- Physical disabilities.

As claimed by Vargas et al. [Vargas et al. 2019:19], people with disabilities are the largest social minor-ity in the world. According to the 2011 report of the World Health Organization, the world was inhabited by over a billion people with some form of disability, almost 200 million of whom were said to experience considerable

difficulty in functioning [WHO 2021]. This amounted to approximately 15 % of the global population. For comparison, the data from the census performed in Poland in 2011 indicated that during that time almost 4.7 million people with disabilities lived in the country, meaning 12.2 % of its entire population [Slany 2014:42]. There-fore, the implementation of the WCAG guidelines allows for limiting the phenomenon of the digital divide among a considerable part of the society.

The first version of the WCAG guidelines was already appreciated and applied for the design of the web pages of public institutions, e.g., in Taiwan [Li et al. 2012:87]. It was not until version 2.0, published on 11 December 2008 [W3C 2021.1], that the guidelines were adopted in nation-al regulations, in a fashion similar to the regulations of the United States, Great Britain and Australia. In Poland, the requirement of adhering to versions WCAG 2.0 and WCAG 2.1 of the guidelines [W3C 2021.2] applies to pages of institutions of the public finance sector or of those financed in over 50 % from public funds.

The data acquired in the course of the research do not allow for determining the sources of financing of the organisations whose websites were visited. However, the institutions whose portals were categorised as government belong to the public finance sector, their number being only 391. There is no certainty that even these pages are available for people with disabilities, since adherence to the WCAG guidelines does not constitute an answer to the problems of all people with disabilities, and moreover, institutions obliged to take them into account are not fulfilling their obligations properly. The studies of portals managed by Polish units of territorial government performed in 2016 indicated that only a few of them comply with the WCAG at the required AA level [Jatkiewicz 2016:39-52].

3. CONCLUSIONS

When focusing on technological issues, the terms Deep Web or Invisible Web take no account of the questions of availability related to the characteristics of Internet users themselves, i.e., their knowledge, experience, competences, curiosity or even their physiology.

A more curious user will probably become acquainted with a larger number of links than those included on the first two pages of results displayed by popular search engines. If they have proper competence, they will probably use specialised browsers or formulate their questions using operators available in the browsers.

A user willing to take risks, or an amoral one, will become acquainted with the contents available on the Darknet. Paid content will be available for a person having at their disposal proper financial resources. Likewise, scientists working at universities or re-search institutes can view documents included in the so-called Academic Invisible Web for free. Other employers also provide their workers with the possibility to browse information which is not generally available, and which is necessary when performing official duties.

For a polyglot, the multilingual character of Internet resources will constitute a smaller obstacle. A healthy person will definitely assimilate more content available online than will a person with disabilities.

The presented research has indicated that the true availability of web pages is difficult to determine unambiguously, and in practice impossible in isolation from the human factor.

The main limitation of the research is the selection of the research sample. This is because the studies of Internet activities involved employees who were performing their official duties. They did not manage their time freely, and they were probably more careful about visiting web pages which could constitute a threat to the organisation or a risk to their reputation among co-workers. This is reflected by the very small numbers of pages from the adult (25 visits) or malware categories (7 visits).

However, it should be noted that an entirely random selection of the sample is not possible. This is because, as far as private individuals are concerned, it would be necessary to obtain their consent for tracing their activities, and the very awareness of the fact would probably affect the results.

According to the author of the studies, they confirm the assumption made at the be-ginning, that the users of the Internet move within a relatively small group of Internet portals, which considerably limits their access to the entirety of contents presented online. This number depends on the gender, age and education. It presumably also depends on numerous other features, which however were not objects of research.

Due to the fact that the term Deep Web has existed for many years in the subject literature and is well established among researchers, it is suggested that it should not be redefined, instead using the new term Inaccessible Web, which would mean resources of the Internet which are inaccessible for a person with specific physical and intellectual features, having at their disposal specified technical measures and knowledge, and functioning in a particular environment. Analogically, it is also reasonable to adopt the term Accessible Web, which would correspond to the term Surface Web when taking the abovementioned features into account. Therefore, it can be assumed that the Deep Web is only a subset of the Inaccessible Web.

REFERENCES

1. Bergman, M.K. The Deep Web: surfacing hidden value. The Journal of Electronic Publishing 2001, vol. 7, issue 1.
2. Berners-Lee, T. The World Wide Web and W3C. Invited plenary, Second International World Wide Web Conference, Chicago 1994.
3. Clarke I., et al. Freenet: A distributed anonymous information storage and retrieval system. Designing Privacy Enhancing Technologies Springer Berlin Heidelberg, 2001.
4. Derfert-Wolf, L. Odkrywanie niewidzialnych zasobów sieci. II seminarium z cyklu INFOBROKER, Wyszukiwanie i przetwarzanie cyfrowych informacji, Centrum Promocji Informatyki, 2007.
5. Dingledine, R. Mathewson N., Syverson P., Tor: The second-generation onion router, Naval Research Lab Washington DC0 (2004).
6. Gauthier, T. D. Detecting trends using Spearman's rank correlation coefficient. Environmental forensics 2001, vol. 2(4), 359–362.
7. Jatkiewicz, P. Presentation of information resources in the information systems of local government. Collegium of Economic Analysis Annals 2016, vol. (42), 39–52.

8. Ledford, J. L. Search engine optimization bible. (Vol. 584), John Wiley & Sons, 2015.
9. Lewandowski, D., Mayr,P. Exploring the academic invisible web. Library hi tech, 2006.
10. Li, S. H., Yen, D. C., Lu, W. H., Lin, T. L. Migrating from WCAG 1.0 to WCAG 2.0 – A comparative study based on Web Content Accessibility Guidelines in Taiwan. Computers in Human Behavior 2012, vol. 28(1), 87–96.
11. Lincoln J. E. Google Click-Through Rates (CTR) By Ranking Position [2020]. Available online https://ignitevisibility.com/google-ctr-by-ranking-position (accessed on 19-04-2021).
12. Labovitz, C., Ahuja, A., Bailey, M. Shining Light on Dark Address Space. Arbor Networks, 2001.
13. Paciello, M. Web accessibility for people with disabilities. CRC Press, 2000.
14. Sherman, C., Price, G. The Invisible Web Uncovering Information Sources Search Engines Can't See. Information Today, Inc (2001).
15. Slany, K. Osoby niepełnosprawne w świetle Narodowego Spisu Powszechnego Ludności i Mieszkań z 2011 r.–wybrane aspekty. Niepełnosprawność-zagadnienia, problemy, rozwiązania 2014, 2, 44–62.
16. Vargas, S., Szarota, M., Mazur R. M. O osobach z niepełnosprawnościami. Fundacja Humanity in Action, 2019.
17. W3Techs. Usage statistics of content languages for websites. Available online: https://w3techs.com/technologies/overview/content_language (accessed on 19-04-2021).
18. W3C. Web Content Accessibility Guidelines (WCAG) 2.0. Available online: https://www.w3.org/TR/WCAG20 (accessed on 19-04-2021).
19. W3C. Web Content Accessibility Guidelines (WCAG) 2.1. Available online: https://www.w3.org/TR/WCAG21 (accessed on 19-04-2021).
20. WHO. World Report on Disability. Available online: https://www.who.int/disabilities/world_report/2011/report.pdf (accessed on 19-04-2021).