STUDIA POLITOLOGICZNE VOL. 54
DOI: 10.33896/SPolit.2019.54.8

Daniel Mider
ORCID: 0000-0003-2223-5997

Sztuka wyszukiwania w Internecie – autorski przegląd wybranych technik i narzędzi

Słowa kluczowe:
społeczeństwo informacyjne, biały wywiad, wywiad jawnoźródłowy, infobrokering

Wprowadzenie

Internet ma dualny charakter – można go rozpatrywać jako medium komunikacji[1] lub jako zbiornik danych[2], co wyznacza odrębne podejścia w zakresie pozyskiwania informacji[3]. W pierwszym wariancie źródło informacji stanowią osoby lub grupy osób, a platformy ich wyszukiwania to zasadniczo media społecznościowe oraz fora dyskusyjne. W takim ujęciu techniki eksploracji będą zogniskowane na kompetencjach efektywnej komunikacji interpersonalnej[4], mniejsze zaś znaczenie mają umiejętności

[1] Takie podejście ilustrowane jest np. przez: O.P. Ohiagu, *The Internet: The Medium of the Mass Media*, „Kiabara Journal of Humanities" 2011, nr 16(2).

[2] M. Bazzell, *Open Source Intelligence Techniques. Resources for Searching and Analyzing Online Information*, 6th ed., Charleston 2018.

[3] Bywają próby łączenia obu pojęć, jednakże w ograniczonym pod względem kanałów komunikacyjnych zakresie: F. Giglietto, L. Rossi, D. Bennato, *The Open Laboratory: Limits and Possibilities of Using Facebook, Twitter, and YouTube as a Research Data Source*, „Journal of Technology in Human Services" 2012, nr 30(3–4).

[4] Elementy werbalne, w tym znajomość – charakterystycznych dla poszczególnych grup – socjolektów, aspekty niewerbalne, w tym proksemiczne, kompetencje w zakresie stwarzania komfortu podczas rozmowy, a nawet umiejętności socjotechniczne.

techniczne, jednak i one odgrywają pewną rolę[5]. Niniejszy tekst zawiera analizę i próbę usystematyzowania narzędzi i technik eksploracji charakterystycznych dla drugiego z wymienionych sposobów podejścia do zjawiska – Internetu jako zbiornika danych. W tym kontekście kompetencje ogniskują się na informatyczno-technicznych elementach: biegłości w posługiwaniu się narzędziami wyszukiwawczymi (oprogramowanie), technikach eksploracji (szczegółowe sposoby formułowania zapytań) oraz taktykach (ogólne procedury wyszukiwania). Ich determinantami są kanały komunikacyjne (rozeznanie w różnorakich obszarach Internetu). Wysiłki poznawcze skupiono na tak zwanym powierzchniowym Internecie (*clearnet*)[6], a strukturę analiz wyznaczyły poszczególne techniki i narzędzia wyszukiwawcze.

Przegląd wybranych technik eksploracji Internetu zogniskowanych na wyszukiwarkach internetowych

Skuteczna procedura wyszukiwania w wyszukiwarkach każdorazowo wymaga opracowania i wdrożenia dwóch następujących elementów: haseł wyszukiwawczych oraz operatorów[7]. Hasła odnoszą się do wyszukiwanych treści, operatory zaś do zakresu zapytania (modyfikują je poprzez uszczegółowienie i dookreślenie).

Odnośnie do tworzenia haseł wyszukiwawczych można na podstawie doświadczeń własnych sformułować kilka ogólnych zasad. Pierwszorzędne znaczenie ma aspekt merytoryczny. Przede wszystkim należy utworzyć liczne, lecz adekwatne wersje wyszukiwanych haseł. W tym celu niezbędne jest wstępne rozeznanie w odniesieniu do wyszukiwanych treści. W toku tworzenia odrębnych, testowanych wedle przemyślanej kolejności haseł konieczne wydaje się uwzględnienie następujących elementów merytorycznych. Tworzenie haseł powinno się odbywać z uwzględnieniem refleksji o charakterze semantycznym, przede wszystkim sprawdzenia

[5] J.A. Benfield, W.J. Szlemko, *Internet-based Data Collection: Promises and Realities*, „Journal of Research Practice" 2006, nr 2(2).

[6] Pojęcie powierzchniowego (*Profound Web*, *Clearnet*) Internetu w zestawieniu z jego dopełnieniem logicznym, to jest Internetem głębokim/ukrytym (*Deep Web*, *Hidden Web*), zostało szerzej omówione w: D. Mider, *Mappa Mundi ukrytego Internetu. Próba kategoryzacji kanałów komunikacji i treści*, „EduAkcja. Magazyn Edukacji Elektronicznej" 2015, nr 2(10).

[7] W tym miejscu należałoby zaznaczyć istnienie jeszcze jednej istotnej dla wyszukiwania zmiennej – oceny jakości i wiarygodności źródeł informacji. Jest to jednak temat tak obszerny, iż wymaga odrębnego tekstu.

tożsamych nazw w różnych (najlepiej znanych wyszukującemu[8]) językach. Pomimo iż język angielski stanowi *lingua franca* Internetu, to liczne zasoby pozostają nietłumaczone na język angielski: z własnych doświadczeń wartymi polecenia wydają się treści publikowane w języku rosyjskim i ukraińskim, szczególnie, iż standardy i zasady obowiązywania własności intelektualnej różnią się w tych przypadkach od zachodnich. Konieczne wydaje się również przestudiowanie sposobów nazywania wyszukiwanych treści w socjolektach różnych grup społecznych. W Internecie rozmaite kanały komunikacyjne wymusiły specyficzne sposoby komunikacji (na przykład Snapchat, Twitter), powstały także społeczności wirtualne (subkultury) posługujące się rozmaitymi socjolektami (na przykład socjolekt używany przez użytkowników 4chan, 8chan czy grup JBWA w serwisie Facebook). Na te nowe podziały i sposoby komunikacji nakładają się odmienne w istotny sposób socjolekty różnych grup zawodowych czy zwolenników określonych opcji politycznych (różnice językowe wyznacza przede wszystkim oś prawica *versus* lewica). Różnice dotyczą nazewnictwa i odmiennego definiowania pojęć, uwzględniania kontekstów (sub) kulturowych i prowadzenia na ich podstawie werbalno-wizualnych gier słownych. Odnoszą się one również do sposobu kodowania informacji: zwielokrotniania, redukcji, zmiany znaków interpunkcyjnych i liter, stosowania lub nie zasad poprawnej pisowni, w tym znaków diakrytycznych danego języka, tworzenia form hybrydowych łączących tekst i grafikę (memy) oraz pisemnego oznaczania reakcji niewerbalnych i użycia emotikonów. Opracowanie listy haseł powinno również zawierać słownikowe ćwiczenie – zarówno odwołanie się do wokabularza synonimów i wyrazów bliskoznacznych, jak również uwzględnianie nazw pojęć w języku potocznym, publicystycznym i akademickim. Dla tworzenia listy haseł wyszukiwawczych istotna wydaje się również warstwa syntaktyczna, a więc szyk wyrazów w zdaniu (co warto każdorazowo testować *in vivo* w wyszukiwarkach), jak również alternatywne odmiany wyszukiwanych pojęć (o ile język, w którym wyszukujemy, je uwzględnia). Ważny okazuje się także techniczny aspekt tworzenia haseł. Niech egzemplifikacją będzie w tym zakresie wyszukiwarka Google. Wyszukuje ona maksymalnie 32 słowa w każdym zapytaniu, nadliczbowe słowa są przez nią ignorowane[9]. Wyszukiwarka Google nie odnotowuje różnic pomiędzy

[8] Niekiedy warto nawet skorzystać z automatycznego tłumaczenia treści z nieznanych wyszukującemu języków, gdyż treści takie mogą potencjalnie dostarczyć wartościowej informacji.

[9] Każda z wyszukiwarek ma odmienne charakterystyki.

słowami zapytań pisanymi wersalikami (majuskułą) a pismem zwykłym. Pozostaje również względnie niewrażliwa (własne testy ujawniają niewielkie różnice w wynikach wyszukiwania) na znaki diakrytyczne w językach narodowych, a także ignoruje szereg słów uznanych za zbyt krótkie, aby poddać je wyszukiwaniu – są to tak zwane *stop words*. Tym mianem określa się słowa uzupełniające tekst, lecz niemające samodzielnego sensu przenoszącego informację. Należą do nich przede wszystkim spójniki, zaimki, przyimki, jak również sformułowania typu: co to jest, jak rozumieć, na czym polega, jak definiować itd.[10] Warto również – szczególnie gdy przedmiotem poszukiwań są słowa używane potocznie lub gdy treści, w których wyszukujemy, są tworzone oddolnie – uwzględnić typowe błędy składniowe i językowe popełniane przez publikujących treści. Stworzoną listę haseł wyszukiwawczych wykorzystujemy w przemyślanej kolejności, dobrym zabiegiem wydaje się stosowanie „techniki lejka", według której podążamy od pojęć najogólniejszych do najbardziej wąskich.

Pracując z wyszukiwarką, należy zatem tworzyć kilka alternatywnych wersji hasła, biorąc pod uwagę zarówno warstwę semantyczną (nazwy w różnych językach, nazwy używane w socjolektach różnych grup społecznych, synonimy, uwzględnienie nazewnictwa potocznego, publicystycznego i akademickiego), jak i syntaktyczną (szyk wyrazów w zdaniu) oraz odmianę wyszukiwanych nazw. Niezbędna jest również opracowana strategia wyszukiwawcza i jednocześnie konsekwentne jej wdrażanie.

Kluczowe znaczenie w procesie wspomagania wyszukiwania mają operatory. Jest to pojęcie polisemiczne, które zakotwiczyło się również w informatyce, przynależąc do rodziny terminów z zakresu języków programowania oraz innych sposobów komunikowania się z komputerami, w szczególności języków zapytań (*query languages*). Operator jest to taka konstrukcja logiczna, której zadaniem jest zwracanie określonej wartości (wyniku działania, to jest transformaty) po wykonaniu działania na argumencie operatora (operandzie). W wyszukiwarkach internetowych najpowszechniej występują operatory przedrostkowe (prefiksowe), to jest takie, w których operand poprzedzany jest przez operator. W wyszukiwarkach internetowych operatory nie należą do złożonych, choć są liczne, różnorodne i niestandaryzowane. Charakteryzować je można przede wszystkim poprzez wykonywane działanie (ich funkcję). Ta właściwość stała się zatem przesłanką próby ich uporządkowania. Studium funkcjonowania operatorów w wyszukiwarkach internetowych stwarza przesłanki

[10] Z kompletną listą *stop words* w językach narodowych można zapoznać się w: *Stopwords*, https://www.ranks.nl/stopwords/ (dostęp: 24.01.2019).

do wyodrębnienia następujących klas (kryterium podziału są wykonywane przez nie działania): operatory logiczne, operatory lokalizacyjne i operatory kanałów komunikacyjnych, operatory chronometryczne, operatory eksploracji treści witryny oraz operatory wyszukiwania określonych typów treści.

Operatory logiczne służą do działań dokonywanych bezpośrednio na treści operandów. Rudyment i powszechnik w językach zapytań stanowią następujące trzy (spośród pięciu) elementy algebry stworzonej przez brytyjskiego matematyka George'a Boole'a: jednoargumentowy operator negacji (dopełnienia, „nie", zaprzeczenia logicznego), dwuargumentowa koniunkcja (iloczyn, logiczne „i") oraz alternatywa (suma, logiczne „lub").

Operator negacji działa tak, iż poprzedzone nim słowa nie pojawią się w wynikach wyszukiwania. Do zapisu negacji w wyszukiwarce Google (G), Yahoo! (Y!), Yandex (Y), Bing (B) i DuckDuckGo (DDG) służy dywiz/minus (znak: -). W ostatnich dwóch z wymienionych można alternatywnie stosować słowo „NOT". Operator ten występuje powszechnie w wyszukiwarkach oraz innych programach wykorzystujących języki zapytania. Sugeruje się, by używać go w formule „lejka", to znaczy dodawać kolejne słowa poprzedzone tym operatorem zapoznawszy się już z wynikami wyszukiwania i sekwencyjnie zawężać obszar wyszukiwań. Na przykład zapytanie „daniel mider" skierowane do wyszukiwarki Bing generuje w pierwszej dziesiątce również wyniki dla amerykańskiego operatora filmowego Daniela Richarda Modera. W celu eliminacji tych wyników zapytanie należy sformułować następująco: daniel mider -moder.

Zastosowanie operatora koniunkcji, zwanego również operatorem włączania, powoduje, iż każdy poprzedzony nim operand jest wymagany w wynikach wyszukiwania, a wyniki częściowo spełniające kryterium nie pojawią się. Najpopularniejszym zapisem tego operatora jest znak dodawania: +. W Bing można zastępować go znakiem et („etką", handlowym „i"), to jest: &, w Google zaś stosować słowo „AND". Operator ten może się pojawiać przed jednym, wieloma lub wszystkimi elementami sformułowanego zapytania. Zastosować można go jak następuje: +mider +cyberterroryzm, by wyświetlić wyniki wyszukiwania łączące nazwisko Mider z pojęciem cyberterroryzmu.

Znajdującym najmniej zastosowań w praktycznych wyszukiwaniach jest operator alternatywy. W logice dwuwartościowych predykatów jego zastosowanie zwraca „prawda", jeśli co najmniej jedna z operand jest prawdą. A zatem służy on do wyszukiwania co najmniej jednego z operandów zawartych w zapytaniu do wyszukiwarki. Zapisuje się go w postaci znaku pisarskiego kreski pionowej (*pipe*) w Bing (w Bing także podwój-

nej) i Yahoo! albo słowa OR w Bing, Google, Yahoo! i DuckDuckGo: | (kod ASCII: 124, dostępny po naciśnięciu na klawiaturze klawiszy Shift + \)[11]. Nie występuje on w Yandex. Wpisanie: cyberterroryzm | cyberprzestępczość | cyberzagrożenia sprawi, iż wyszukane zostaną te strony, gdzie znajduje się choć jedno z trzech wymienionych pojęć.

Funkcjonują dwa operatory zastępowania znaków – operator zastępowania ciągu znaków reprezentowany przez asterysk (znak: *, uzyskiwany po wciśnięciu Shift + 8) oraz operator zastępowania pojedynczego znaku zapisywany jako kropka (znak: .). Operatory te pozwalają na wyszukiwanie pojęć lub fraz, których dokładnego (poprawnego) zapisu nie znamy. Przykładowy zapis: m.der winien zwrócić wszystkie słowa zawierające wskazane litery oraz dowolny znak pomiędzy pierwszą a pozostałymi (w tym na przykład spację lub kropkę). Zastosowanie asterysku wydłuża dowolnie odległość pomiędzy wpisanymi znakami, a zatem ten operator nadaje się raczej do fraz. Praktyka zastosowania powyższych operatorów dowodzi, iż w różnych wyszukiwarkach mają one umiarkowane wyniki wyszukiwania, ponadto są one wrażliwe na spacje. Operator zastępowania ciągu znaków jest powszechniejszy (B, DDG, G, Y, Y!) niż operator zastępowania pojedynczego znaku (tylko G i deklaratywnie w Y!).

Wyszukiwarki (B, DDG, Y, Y!, w mniejszym stopniu G) umożliwiają również wyszukiwanie danych numerycznych według podawanych przez użytkownika zakresów, na przykład cen czy rozmiarów. Służy do tego operator zakresu zapisywany za pomocą dwóch kropek (następująco: ..). Zapis wyszukiwania: +BMW +650GS +cc600..cc800 lub +BMW +650GS +cc +600..800 dla hipotetycznej giełdy motocykli zwróci jako wynik wskazane modele motocykli, lecz o pojemności zawartej pomiędzy 600 a 800 cm^3. Zapis cen oznaczamy jak następuje: $100..$400. Operator ten działa niesatysfakcjonująco, jego funkcjonowanie polepsza użycie go w parze z operatorem site:.

Zaimplementowane w wyszukiwarkach, a przede wszystkim w Google, mechanizmy ułatwiające wyszukiwanie poprzez rozszerzenie jego zakresu i zasugerowanie użytkownikowi „właściwego", to jest najczęściej wyszukiwanego wyniku lub mechanizmy sztucznej inteligencji rozpoznające składnię i znaczenie zapytań (na przykład Koliber w Google) paradoksalnie utrudniają wyszukiwanie. Istnieje jednak w nielicznych wyszukiwarkach (DDG, Y) mechanizm wyłączający działanie powyższych algorytmów i umożliwiający użytkownikowi wyszukanie dokładne – co

[11] Niekiedy na klawiaturze reprezentowany jest przez złamaną (przerywaną) pionową kreskę: ¦.

do znaku i bez zmian rozszerzających. Operator ten oznaczany jest za pomocą wykrzyknika (znak: !), a zapytanie: „!daniela !midera" oznacza, by wyszukać nazwisko dokładnie w takiej formie: „Daniela Midera" (nie jest to pomyłka). DuckDuckGo umożliwia stosowanie tego operatora dla innych wyszukiwarek jako pośrednik. W menu Google można zamiast powyższego operatora użyć zakładki Narzędzia → Dokładnie.

Występują dwa typy operatorów grupowania. Pierwszy z nich umożliwia wyszukiwanie dokładnie jak wprowadzono (we wpisanej kolejności) i nazywany jest operatorem grupowania uporządkowanego. Do jego zapisania używamy cudzysłowu (znaki: „"). Zapytanie: „volenti non fit iniuria" zwróci wyłącznie tak zapisaną łacińską sentencję (kolejność słów jak wprowadzono). Operator ten działa wadliwie lub nie w pełni (B, DDG, Y!). Drugi – operator grupowania nieuporządkowanego – reprezentowany jest przez parę nawiasów (znaki: ()). Wyszukuje wyrazy umieszczone w nawiasie, jednakże mogą one występować w kolejności dowolnej.

Operatory lokalizacyjne służą do filtrowania wyników wyszukiwania wedle całości lub części adresu internetowego (dowolnie zapisanego: w postaci adresu IP, nazwy domeny, URL, nazwy DNS, ale także lokalizacji geograficznej lub geograficzno-językowej), jest ich kilkanaście.

Kluczowym operatorem jest operator site: zawężający wyniki wyszukiwania do danej strony (i podstron). Zapytanie: mider site:www.inp.uw.edu.pl wyszukuje nazwisko Mider wyłącznie w witrynie Instytutu Nauk Politycznych UW. Operatora tego można również używać jako samodzielnego – wówczas efektem jego działania jest enumeratywna lista podstron danej witryny (prezentowana jednak w sposób nieuporządkowany z punktu widzenia jej struktury).

Odmienne jest działanie operatora URL[12]: wyszukującego adres danej strony, gdziekolwiek się on znajduje. Prawidłowy zapis zapytania jest następujący: url:http://www.inp.uw.edu.pl/. Można zastąpić go innymi operatorami, jego istnienie nie jest niezbędne. Jego pochodną są operatory inurl:, allinurl:, url: (działają tylko dla G i DDG, ten ostatni dla Y). Wyszukują one słowo/słowa użyte w adresie URL danej witryny. Na przykład zapytanie allinurl:inp uw spowoduje, iż wyszukane zostaną dokumenty mające w adresie URL zarówno słowo „inp", jak i „uw". Operatory te reagują wyłącznie na słowa, nie zaś na składniki adresu URL (na

[12] Nazwa jest abrewiaturą i w pełnej wersji brzmi: *Uniform Resource Locator*. Oznacza standaryzowany na podstawie dokumentu RFC 1738 format adresowania zasobów w sieci globalnej i sieciach lokalnych. Składa się z trzech elementów: protokołu (np. http, https, ftp, telnet, nntp, mailto), adresu serwera oraz (opcjonalnie) ścieżki do zasobu.

przykład kropka, dwukropek, prawy ukośnik) i nie istnieje możliwość ominięcia tego ograniczenia. Różnica pomiędzy inurl: oraz allinurl: jest następująca: allinurl:mider daniel spowoduje wyszukanie takich adresów, które zawierają słowa zarówno Daniel, jak i Mider. Z kolei inurl:mider daniel wyszuka adresy URL ze słowem Mider oraz słowem Daniel w dowolnym miejscu strony. Zapytania: inurl:mider inurl:daniel oraz allinurl:mider daniel są tożsame. Dobrym sposobem wyszukiwania subdomen (poddomen) przypisanych do danej domeny jest operator domain: (B, DDG, Y!). Następujące sformułowanie: domain: inp.uw.edu.pl lokalizuje przykładowo subdomeny takie jak gpss.inp.uw.edu.pl czy podyplomowe.inp.uw.edu.pl. Operator ip: (wyłącznie dla: B, DDG, Y!) zwraca adres DNS (*Domain Name System*), czyli nazwę mnemoniczną, łatwą do posługiwania się w komunikacji międzyludzkiej. Zapis: ip:86.111.240.162 spowoduje wyświetlenie się wyników dla adresu www.inp.uw.edu.pl. Operatory host: i rhost: działają w Yandex i wydają się działać w Bing, choć nie w pełni prawidłowo. Wyszukują one podstrony witryny, jednak wyłącznie w ramach danego hosta, to jest maszyny, na której zamieszczone są zasoby.

Możliwe jest również wyszukiwanie zasobów wedle lokalizacji stron (serwerów, na których się znajdują). Służą do tego celu location: i loc: dla Bing, region: i r: dla DuckDuckGo oraz cat: dla Yandex. Wymienione operatory zawężają wyniki wyszukiwania do stron znajdujących się w określonej lokalizacji geograficznej. Bing akceptuje dwuliterowe kody normy ISO 3166-1 (tzw. kod alfa-2)[13]. Wyszukiwarka Yandex miała możliwość wyszukiwania identyfikatorów tematycznych za pomocą operatora cat:[14]. Wyszukiwanie odbywało się w katalogu Yandex. Obecnie nie funkcjonuje. Google oferuje tę usługę w Wyszukiwaniu zaawansowanym. Wyszukiwarki Yandex i Bing wprowadziły dodatkowo możliwość wyszukiwania jednocześnie lokalizacji języka danej strony. Operandy powinny być formułowane (prawdopodobnie!) według normy ISO 639-1. Na przykład zapytanie: institute altloc:pl-en wyszukuje polskie strony (lokalizacja) w języku angielskim zawierające słowo „institute".

Blisko spokrewnione z wyżej analizowanymi są operatory kanałów komunikacyjnych w mediach społecznościowych – umożliwiają ograniczenie wyszukiwania do wybranych obszarów mediów społecznościowych. Hasztag (*hashtag*, w skrócie tag) jest to pojedyncze słowo lub

[13] *Lista kodów dla Bing*, https://pl.wikipedia.org/wiki/ISO_3166-1 (dostęp: 12.02.2019); *Lista kodów dla DuckDuckGo*, https://duckduckgo.com/params (dostęp: 12.02.2019).

[14] *Lista kodów regionalnych Yandex*, *Lista kodów tematycznych Yandex*, http://search.yaca.yandex.ru/cat.c2n (dostęp: 12.02.2019).

fraza nierozdzielona spacjami, poprzedzone symbolem # (*hash*, kratka, krzyżyk lub płotek). Jest to forma znacznika umożliwiająca w mediach społecznościowych oddolne, niehierarchiczne grupowanie wiadomości[15]. Usługa wyszukiwania w hasztagach mediów społecznościowych funkcjonuje w Google i Yahoo! oraz w mniejszym stopniu w DuckDuckGo i Bing (wadliwie). Wpisanie: #covfefe w wymienionych wyszukiwarkach spowoduje odszukanie w mediach społecznościowych wiadomości oznaczonych tym właśnie tagiem. Z kolei poprzedzenie nazwy własnej użytkownika znakiem @ (*at, commercial at*, małpa) pozwala na wyszukiwanie profili w social mediach (B, DDG, G, Y!). Operator blogurl: wyszukujący blogi w określonej domenie funkcjonuje wyłącznie w Google. Zapytanie: blogurl:inp.uw.edu.pl spowoduje wyszukanie blogów w domenie Instytutu Nauk Politycznych UW. Jedynie w Bing można skorzystać z operatora feed: odnajdującego kanały RSS (*Really Simple Sindication*) / Atom mające wskazaną nazwę. Przydatnym rozwiązaniem może okazać się również hasfeed: pozwalający na sprawdzanie, czy w danej domenie ulokowano kanały RSS. Operatory kanałów komunikacyjnych mediów społecznościowych to dość siermiężne narzędzie – istnieją zautomatyzowane, profesjonalne programy pozwalające na wielokrotnie bardziej efektywne i precyzyjne wyszukiwanie w social mediach (na przykład mechanizm wyszukiwania dla mikrobloga Twitter wbudowany w program Maltego).

Kluczowe znaczenie mają operatory chronometryczne pozwalające na zawężenie wyników wyszukiwania do określonych przedziałów lub punktów czasu. Możliwość takiego wyszukiwania jest ze względu na wygodę użytkowników realizowana w innej formule – wyszukiwania okienkowego i z użyciem menu. Operator date: pozwalający na wyszukiwanie punktowe (dla dnia, miesiąca, roku) pozostawiono wyłącznie w wyszukiwarce Yandex. W Google, a do niedawna i w Yandex, funkcjonuje operator daterange: oferujący wyszukiwanie w zakresach dat. W Google wymagał użycia daty juliańskiej[16] oraz rozdzielenia wskazywanych przez użytkownika dat dywizem, a w Yandex – dwiema kropkami. Na przykład 4 kwietnia 2018 to wedle zapisu daterange:2458212.500000. Operatory te są na tyle ważne, iż zaimplementowano je w trybie okienkowym (kalendarzowym), a w trybie linii poleceń – zmarginalizowano. Funkcjonującym i przydatnym spośród operatorów chronometrycznych jest cache: poka-

[15] Przede wszystkim dotyczy to Facebooka, Instagrama i Twittera.

[16] Dla ułatwienia zamiany daty gregoriańskiej na juliańską można było skorzystać z następującego konwertera: *Julian Date Converter*, http://aa.usno.navy.mil/data/docs/JulianDate.php (dostęp: 12.02.2019).

zujący wersję strony z pamięci podręcznej wyszukiwarki, a więc stronę (najprawdopodobniej) archiwalną. Funkcjonuje on tylko w Google, z kolei w Bing umożliwia wyszukiwanie poprzez funkcję Zbuforowano (strzałka skierowana w dół przy odnośniku).

Operatory wyszukiwania w treści strony umożliwiają systematyczne przeszukiwanie witryn internetowych pod kątem określonych wartości w podziale na tytuł witryny, treść witryny, jej metaznaczniki i elementy informacyjne oraz inne. Operatory wyszukiwania w treści strony to operatory elementarne, istnieją stosunkowo długo, powstały wraz z Web 1.0.

Wyszukiwanie odnośników w tekście strony jest możliwe za pomocą operatora inanchor:. W HTML *anchor* oznacza treść umieszczaną pomiędzy znacznikami <a> oraz . Taka oto treść jest wyświetlana na stronie www odnośnik (link):

Instytut Nauk Politycznych UW

Aby wyszukać taką treść na przykład na stronie Wydziału Nauk Politycznych i Studiów Międzynarodowych wpisujemy: site:wnpism.uw.edu.pl inanchor:instytut – uzyskujemy zwrot takich stron, dla których link stanowi słowo „instytut".

Potrzebne, lecz niedziałające są operatory linkfromdomain: (B) i link: (G, Y!) – służą uwidacznianiu URL, do których zawierają odnośniki. A zatem potencjalnie można byłoby odnajdywać wszystkie te strony, które zawierają odnośniki, linkują stronę będącą przedmiotem wyszukiwania.

Istnieje szereg operatorów umożliwiających precyzyjne, ukierunkowane wyszukiwanie zarówno w nagłówku strony www (informacje zawarte pomiędzy znacznikami nagłówka <head></head>), jak i jej treści (tzw. body, to jest zawartości z ulokowanej między znacznikami HTML <body></body>).

Stronę internetową opisują w sposób najbardziej ogólny tak zwane metaznaczniki informujące ogólnie o jej treści, standardach językowych i formatowaniu. Znacznik meta: pozwala na wyszukiwanie słów zamieszczonych w nagłówku strony w metaznaczniku „keywords", który zawiera ustalone przez autora strony rozdzielone przecinkami słowa lub frazy opisujące treść strony. W przypadku strony Instytutu Nauk Politycznych UW ów fragment kodu jest następujący:

<meta name="Keywords" content="nauki polityczne, politologia, bezpieczeństwo wewnętrzne, administracja rządowa, administracja publiczna, studia podyplomowe, studia bezpieczeństwa, zarządzanie systemami bezpieczeństwa, bezpieczeństwo narodowe, europeistyka, marketing polityczny, studia dzienne, studia wieczorowe, studia zaoczne,

studia licencjackie, studia magisterskie, rekrutacja uw, rekrutacja 2011, licencjat na uw, studia na uw, magister na uw" />

W celu wyszukania słowa kluczowego politologia w wyszukiwarce Bing (w innych operator nie działa, w Bing funkcjonuje – już lub jeszcze – wadliwie) użyjemy następująco sformułowanego zapisu: meta:politologia. Bing wprowadził analogiczny, lecz jeszcze bardziej restrykcyjny operator literalmeta:, jednak w chwili obecnej nie wydaje się on działać zgodnie z przeznaczeniem.

Informacje identyfikujące stronę www mogą być również przez autorów strony umieszczone w metaznaczniku „description":

<meta name="Description" content="Politologia i bezpieczeństwo wewnętrzne w Instytucie Nauk Politycznych Uniwersytetu Warszawskiego. Studia licencjackie i studia magisterskie." />

Wyszukiwanie w ramach wymienionych treści odbywa się za sprawą operatora info: (wyłącznie w Google). Jako operand wystąpić może zarówno słowo, jak i fraza. W drugim z przypadków należy ująć ją w cudzysłów. Ze względu na mechanizmy rozszerzające wyszukiwanie operator ten działa w Google niesatysfakcjonująco. Analogicznie funkcjonuje para operatorów intitle: (title: dla Yandex) i allintitle:. Operatory ograniczają wyniki do stron zawierających wszystkie wyrazy zapytania w tytule strony. Tytuł strony znajduje się w nagłówku źródła strony pomiędzy znacznikami HTML <title> i </title>. Zamieszczane są tam przez autorów/programistów stron internetowych. Z opcji tej można także skorzystać w Google na stronie Szukanie zaawansowane. Różnica pomiędzy rozważanymi operatorami jest następująca: intitle: jest operatorem jednowyrazowym, a allintitle: umieszczamy przed frazami. Są to operatory uniwersalne, dostępne we wszystkich pięciu analizowanych wyszukiwarkach.

W wyszukiwarkach Bing i Yahoo! i w nieco mniejszym stopniu w Yandex można wyszukiwać według języka strony www zadeklarowanego w nagłówku HTML. Przykładowy kod zawierający informację o języku strony wygląda jak następuje:

<html class="…" lang="en">

<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="pl">

Wyszukiwania dokonuje się za pomocą operatora language: (B, Y!) lub lang: (Y). Przykładowe zapytanie dla Bing: „politologia" lang:en. Kod języka pozyskuje się z predefiniowanych list[17].

[17] *Lista kodów języka dla Bing*, https://msdn.microsoft.com/en-us/library/dd250941.aspx (dostęp: 12.02.2019); *Lista kodów dla Yandex*, https://tech.yandex.com/translate/doc/dg/

Wyszukiwanie treści na stronach internetowych umożliwiają następujące operatory: intext: w DDG, allintext: w Google, Yahoo! i Yandex oraz inbody: w Bing. Najczęściej łączone są z operatorem site:.

Metaoperatory keyword: i instreamset: (Y, B) służą do wyszukiwania w ramach innych operatorów. Następujący rozkaz: keyword:(intitle inbody)infobrokering pozwoli wyszukać słowo „infobrokering" jednocześnie w tytule i w tekście dokumentu.

Przydatne rozwiązanie stanowi operator prefer:. W zamierzeniach twórców ma podkreślać dane słowo pośród innych wyszukiwanych, zwiększając umieszczenie na wyższych pozycjach w wyszukiwaniu stron zawierających to słowo. Działa wadliwie. Lepsze efekty uzyskujemy (sposób dla Google i Bing), jeśli wielokrotne powtórzymy dane słowo w zapytaniu.

Operatory near: oraz around: (tylko B i G, lecz i tam działają niesatysfakcjonująco) określają maksymalną odległość wyszukiwanych od siebie słów. Na przykład zapytanie w Google: daniel around(2) mider będzie oznaczało, że wyszukane mają zostać strony, na których słowa „daniel" i „mider" znajdują się maksymalnie w odległości dwóch wyrazów (oddzielone są maksymalnie dwoma wyrazami).

Operatory wyszukiwania określonych typów treści zapewniają odnajdywanie ściśle określonych informacji, na przykład ściśle zdefiniowanych typów plików (Word, Excel), informacji pogodowych czy definicji słownikowych i encyklopedycznych. Jest to zróżnicowana tematycznie grupa operatorów, jednak największa ich liczba występuje w Google.

Najbardziej przydatna i występująca we wszystkich wyszukiwarkach jest możliwość dookreślenia typu wyszukiwanych zasobów poprzez rozszerzenie pliku. Odbywa się to z użyciem operatora filetype: (G, DDG, Y!), mime: (Y), ext: (B – obecnie przestało działać). Operator wyszukuje określone typy plików, na przykład arkusze kalkulacyjne (.xls, .xlsx), dokumenty tekstowe (.doc, .docx, .odt) itd. Wstawiamy je bez kropki. A zatem zapytanie: site:inp.uw.edu.pl filetype:docx zwróci listę plików Word (Office 2007) dla www Instytutu Nauk Politycznych. W Bing i Yandex obecny jest operator contains:. Za jego pomocą odnajdujemy strony (a nie same dokumenty jak wyżej) zawierające linki do dokumentów o rozszerzeniach określonych w danym operatorze. Na przykład: contains:doc site:inp.uw.edu.pl zwraca wszystkie strony w INP, na których znalazły się odnośniki do dokumentów tekstowych edytora.

concepts/api-overview-docpage/ (dostęp: 12.02.2019); *Lista kodów dla Yahoo!*, https://developer.yahoo.com/search/languages.html (dostęp: 12.02.2019).

Google oraz Yahoo! umożliwiają przeszukiwanie treści encyklopedii, leksykonów i słowników. Operator define: umożliwia wyszukiwanie pośród tych zasobów według słowa lub frazy. Z kolei operator related: powoduje wyświetlenie listy stron „podobnych" do określonej strony internetowej. Na przykład zapytanie related:www.inp.uw.edu.pl spowoduje wyświetlenie stron internetowych, które są podobne do strony głównej Instytutu Nauk Politycznych UW. Z opcji tej można także skorzystać, wybierając Podobne strony na głównej stronie z wynikami wyszukiwania Google oraz na stronie Szukanie zaawansowane w sekcji Informacje o danej stronie internetowej → Podobne do. Z kolei znak ~ (tylda) to operator wyszukiwania synonimów. Działa jednak wadliwie lub nie działa wcale. Z wyszukiwarki Google można skorzystać również do pozyskania informacji liczbowych. Służy do tego operator convert: oferujący przeliczanie według kursów walut, a także rozmaitych miar (wagi, odległości)[18].

Możliwe jest również wyszukiwanie treści multimedialnych z użyciem operatorów, choć odchodzi się już od tego rozwiązania na rzecz prostszego, bardziej przejrzystego trybu okienkowego. Operator imagesize: umożliwia określenie wielkości wyszukiwanego obrazu. Do dyspozycji pozostają trzy wielkości: mała (*small*), średnia (*medium*) oraz duża (*large*). Zapytanie formułuje się następująco: „daniel mider" imagesize:small. Wbrew deklaracjom liczby (na przykład 600 dpi) w Bing nie działają.

W Bing funkcjonował operator msite:, ogniskujący wyniki wyszukiwania na stronach multimedialnych (fotografie i filmy). Przykładowe zapytanie: msite:mider.

Większość analizowanych wyszukiwarek (B, G, Y oraz do pewnego stopnia Y!) zapewnia możliwość wyszukiwania w mapach za pomocą operatora maps:. Zwraca uwagę fakt, iż wynik wyszukiwania jest odmienny od tego, jaki uzyskujemy otwierając zakładkę Mapy wyszukiwarki. Yahoo! ma wbudowaną funkcję wyszukiwania map we frazach (teren USA), jeśli padanie słowo *map*, na przykład: map of New York.

Wyszukiwarka Google oferuje możliwość wyszukiwania treści publikacji wedle tytułu i autora. Operator book: pozwala na zaawansowane wyszukiwanie książek według słów kluczowych zawartych w tytule. Przeszukuje bazę Google Books[19]. Z kolei operator author: wyszukuje autorów tekstów (artykułów).

---

[18] *Lista miar podlegających konwersji*, http://searchcommands.com/convert/ (dostęp: 12.02.2019). Obecnie składnia wymaga jedynie wpisania np.: 1 inch to cm.

[19] Wyszukiwanie można przeprowadzić również: https://books.google.com/advanced_book_search (dostęp: 12.02.2019).

W Google, dla terenu Stanów Zjednoczonych, przez krótki czas funkcjonowała usługa wyszukiwania numerów telefonów (operatory phonebook:, bphonebook:, rphonebook:). Usługa ta została zlikwidowana ze względu na zbyt duże zainteresowanie.

Warto zwrócić uwagę na operatory informacyjne. Operator movie: serwuje repertuar kin, informacje o filmach i recenzje. Przykładowe użycie: movie:Warszawa. Z kolei operator weather: – analogicznie – podaje informacje dotyczące aury.

Przegląd wybranych narzędzi eksploracji Internetu – wyszukiwarek internetowych

Przegląd nie ma charakteru wyczerpującego ani pogłębionego, służy raczej wstępnej orientacji zainteresowanych w uniwersum wyszukiwarek internetowych. Wskazano i przeanalizowano między innymi wyszukiwarki globalne, wyszukiwarki zogniskowane na prywatności użytkownika, meta- i multiwyszukiwarki, wyszukiwarki i katalogi lokalne, wyszukiwarki ludzi, wyszukiwarki szarej literatury i wyszukiwarki naukowe, a także wyszukiwanie w archiwach Internetu.

W tekście zrezygnowano z licznych wątków tematycznych – zabrakło na przykład wyszukiwarek multimediów oraz wyszukiwarek mediów społecznościowych, a także wielu branżowych wyszukiwarek. Pominięte zostały również wyszukiwarki i paradygmaty wyszukiwania w innych niż powierzchniowych obszarach Internetu, między innymi Usenecie, File Transfer Protocol, The Onion Router, Invisible Internet Project (I2P), OpenNIC, CesidianRoot, Freenet. Uzasadnieniem takiego działania jest ograniczona objętość tekstu oraz fakt, iż treści te mogą zostać wprowadzone po opanowaniu przedstawianych w tekście informacji, a także – co istotniejsze – dotarcie do wymienionych wyżej zasobów Internetu wymaga więcej niż podstawowej wiedzy na temat jego topografii i solidnego instruktażu związanego zarówno z instalacją i konfiguracją, jak również użyciem narzędzi.

WYSZUKIWARKI GLOBALNE[20]

Do globalnych wyszukiwarek internetowych zaliczymy „wielką trójkę": Google, Bing, Yahoo!, jednak wyliczenie takie ma charakter umowny: Google wyraźnie dystansuje pozostałe wymienione narzędzia. Jest to witryna najczęściej odwiedzana na świecie – dziennie odnotowuje ponad miliard unikatowych użytkowników (1 022 345 310), którzy otwierają ją ponad osiem miliardów razy (8 178 762 480)[21]. Wyszukiwarka Google to aż 90,28% wszystkich wyszukiwań, podczas gdy pozostałe wyszukiwarki globalne stanowią margines: Bing – 3,82% i Yahoo! – 2,76%[22]. Wyniki te od 2010 roku zmieniły się niewiele – Google stale utrzymuje przewagę, jedyna zmiana dotyczy popularności Bing, która od 2010 roku wzrosła aż dwukrotnie[23]. W przypadku dostępu za pomocą urządzeń mobilnych przewaga ta sięga aż 94,15%[24]. W Polsce wyszukiwarka Google uzyskała miażdżącą przewagę – 98,73%, podczas gdy Yahoo! – 0,56%, a Bing – 0,44%[25]. Nie wszędzie jednak Google ma dominującą pozycję – jest niepopularne w Chinach, gdzie na przykład w grudniu 2018 roku odwoływano się do niej zaledwie w 2,57% wyszukiwań. Na pierwszym miejscu lokuje się tam Baidu (z wynikiem 70,3%), Shenma (15,62%), Sogou (4,74%), Haosou (4,54%). Druga z wymienionych wyszukiwarek stara się konkurować z wiodącą Baidu – w lutym 2018 roku udział Baidu spadł do 58,55%, a w tym samym czasie wzrosło istotnie zainteresowanie wyszukiwarką Shenma – do 28,76%. W Federacji Rosyjskiej również rodzima wyszukiwarka ma znaczną przewagę – Yandex w grudniu 2018 miał 54,27% udziału w rynku wyszukiwarek, a Google – 42,42%. Na trzecim miejscu ulokowała się Mail.

---

[20] Pojęcie „globalne" zostało użyte w znaczeniu komercyjnym, to jest pokrycia rynku wyszukiwarek. Z akademickiego punktu widzenia słuszne jest rozróżnienie Sabiny Cisek na wyszukiwarki globalne (horyzontalne, uniwersalne) oraz wyszukiwarki specjalistyczne (wertykalne): S. Cisek, *Wyszukiwarki specjalistyczne*, http://sabinacisek.blogspot.com/2012/11/wyszukiwarki-specjalistyczne.html (dostęp: 12.02.2019).

[21] *Web Analysis and Statistics*, https://web2stat.com/w/google.com (dostęp: 24.01.2019).

[22] Są to najnowsze (na czas powstania tekstu) dostępne dane, za październik 2018 r.: *Worldwide Desktop Market Share of Leading Search Engines from January 2010 to October 2018, Statista. The Statistics Portal*, Statista, https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/ (dostęp: 24.01.2019).

[23] C. Mangles, *Search Engine Statistics 2018*, SmartInsights, 30.01.2018, http://www.smartinsights.com/search-engine-marketing/search-engine-statistics/ (dostęp: 24.01.2019).

[24] *Mobile Search Engine Market Share Worldwide*, StatCounter, http://gs.statcounter.com/search-engine-market-share/mobile/worldwide (dostęp: 24.01.2019).

[25] *Search Engine Market Share Poland*, StatCounter, http://gs.statcounter.com/search-engine-market-share/mobile/worldwide (dostęp: 24.01.2019).

ru (z wynikiem 2,24%)[26]. Przewaga Google jest – paradoksalnie – nieco mniejsza w Stanach Zjednoczonych Ameryki Północnej, Google ma tam 86,91% udziału, Yahoo! – 6,31%, Bing – 5,56%, a DuckDuckGo – 0,9%. Notowane są tam również: MSN – 0,16% i Baidu – 0,04%[27].

Kluczową informacją wydaje się nie częstość używania, lecz odpowiedź na pytanie o liczbę stron zindeksowanych przez każdą z wyszukiwarek. Poznanie, choć przybliżone, całkowitej liczby zindeksowanych przez wyszukiwarki stron możliwe jest dzięki stylistyce kwantytatywnej – dyscyplinie naukowej z pogranicza stylistyki i matematyki. Konkretnie, dokonanie pomiaru umożliwia prawidłowość odkryta przez Jeana-Baptiste'a Estoupa i George'a Kingsleya Zipfa (prawo Estoupa–Zipfa). Określenie wielkości zindeksowanej sieci opiera się na przybliżonych rachunkach wolumenu Google, Bing i Yahoo! Search, uwzględniając wzajemne nakładanie się wyników, co prowadzi do przeszacowania, odpowiednio pomniejszając uzyskane sumy[28]. Wielkość sumaryczną indeksu określa się następująco. Po pierwsze, konieczny jest zbiór referencyjny (wzorzec, zbiór odniesienia, korpus) dostępny offline (w praktyce jego zawartość stanowi milion stron internetowych z katalogu DMOZ, co może być – potencjalnie – uważane za reprezentatywną próbkę World Wide Web). Po wtóre, niezbędne jest sekwencyjne użycie głównych wyszukiwarek. Obliczenie wolumenu stron odbywa się na podstawie porównania proporcji słów wykazywanych przez korpus z wykazywanymi przez wyszukiwarki. Przykładowo: jeśli słowo x występuje w korpusie w 75% dokumentów, to jeśli zostało ono wykazane online w 15 mld dokumentów, wówczas orzekamy, iż istnieje 20 mld stron. W praktyce każdego dnia 50 reprezentatywnych słów (rozlokowanych równomiernie w logarytmicznych odstępach), których częstotliwość została obliczona w korpusie, jest wysyłanych do poddawanych pomiarowi wyszukiwarek. Liczba stron znalezionych dla tych słów jest rejestrowana i porównywana z ich względnymi częstotliwościami w korpusie[29]. Na tej podstawie oce-

[26] *Search Engine Market Share Russian Federation*, StatCounter, http://gs.statcounter.com/search-engine-market-share/all/russian-federation (dostęp: 24.01.2019).

[27] *Search Engine Market Share United States of America*, StatCounter, http://gs.statcounter.com/search-engine-market-share/all/united-states-of-america (dostęp: 24.01.2019).

[28] Rachunek nakładających się wyników obliczany jest w sekwencji, począwszy od jednej z czterech wyszukiwarek, a zatem możliwych jest kilka porządków, co prowadzi do różnych całkowitych sum oszacowań. Ważne jest również to, że taki algorytm obliczeń sprawia, że wolumen stron jest niedoszacowany.

[29] A. van den Bosch, T. Bogers, M. de Kunder, *Estimating Search Engine Index Size Variability: A 9-year Longitudinal Study*, http://www.dekunder.nl/Media/10.1007_s11192-016-

nia się, iż Google zindeksowało 62 mld stron (sekwencja Google–Bing) lub nieco ponad 6 mld stron (sekwencja Bing–Google)[30]. Obecnie nie podaje się wyników dla dwóch pozostałych, pierwotnie uczestniczących w pomiarze wyszukiwarek – Yahoo! Search oraz Ask, ponieważ ich właściciele zrezygnowali z informowania użytkowników o liczebności wyników wyszukiwania.

WYSZUKIWARKI ZOGNISKOWANE NA PRYWATNOŚCI UŻYTKOWNIKA

Wyszukiwarki zogniskowane na prywatności użytkownika powstały w odpowiedzi na nieobecność na rynku wyszukiwarek nieprofilujących, a więc takich, które nie dokonują rozpoznawania, analizowania i segmentacji użytkowników pod kątem płci, wieku, lokalizacji, preferencji politycznych czy zainteresowań w celu dostosowania treści reklamowych przez użytkownika (warto podkreślić, iż nawet tryb incognito w Google nie chroni użytkownika przed tak zdefiniowanym profilowaniem)[31]. Obok naruszenia prywatności, stanowiącej dobro samo w sobie, działania Google prowadzą do utrudnień w korzystaniu z wyszukiwarki, czyniąc jej użytkowanie ze względu na obecność treści reklamowych uciążliwe, bo nieprzejrzyste.

Spośród wyszukiwarek szanujących prywatność użytkownika na pierwszym miejscu należy bezwzględnie wymienić wyszukiwarkę DuckDuckGo (nazwa pochodzi od dziecięcej zabawy *Duck, duck, goose*) dostępną pod adresem https://duckduckgo.com lub z przekierowaniem https://duck.com. Narzędzie to powstało w 2008 roku i obecnie występuje w wersji zarówno desktopowej, jak i dla urządzeń mobilnych. DDG została zaimplementowana do licznych przeglądarek internetowych (przede wszystkim Firefox 33.1. oraz Tor Browser 6.0), a obecnie (dane za styczeń 2019) odnotowuje około 30 mln wyszukiwań w ciągu doby (194. pozycja w rankingu Alexa[32]).

DDG chroni prywatność użytkowników, nie zbierając o nich informacji w celu profilowania (w efekcie nie serwuje żadnych treści marketingowych), a także umożliwia korzystanie z narzędzi wyszukiwawczych przez

1863-z.pdf (dostęp: 24.01.2019).

30 *The Size of the World Wide Web (The Internet)*, WorldWideWebSize.com, http://www.worldwidewebsize.com (dostęp: 24.01.2019).

31 L. Vaas, *Google's Private Browsing Doesn't Keep Your Searches Anonymous*, 6.12.2018, https://nakedsecurity.sophos.com/2018/12/06/googles-private-browsing-doesnt-keep-your-searches-anonymous/ (dostęp: 15.02.2019).

32 *Duckduckgo.com Traffic Statistics*, http://www.alexa.com/siteinfo/duckduckgo.com (dostęp: 12.02.2019).

system anonimizujący TOR. *Votum separatum* jednego z użytkowników wskazywało na to, że jednak DDG dokonuje ograniczonego trackingu danych socjo-psycho-demograficznych[33], co DDG zdementowała, wyjaśniając sposób i zakres działania jednego z aktywowanych przez wyszukiwarkę elementów przeglądarki Firefox. Popularność DDG wzrosła istotnie po ujawnieniu przez Edwarda Snowdena szczegółów projektu PRISM amerykańskiej Agencji Bezpieczeństwa Narodowego (National Security Agency, NSA) oraz zaimplementowania DDG w produktach Apple. Polityka nieprofilowania uprawiana przez DDG nie tylko chroni użytkowników przez treściami reklamowymi i utratą prywatności, lecz przede wszystkim zapobiega najpoważniejszemu problemowi wyszukujących – bańce filtrującej (*filter bubble*) – a więc zjawisku często nieuświadamianej i niechcianej preselekcji informacji wyszukiwanych przez użytkownika pod kątem jego wcześniejszych wyszukiwań[34].

Indeksowane przez DDG dane pochodzą z licznych źródeł: ponad 400 źródeł indywidualnych (na przykład Search Boss, Wolfram Alpha), w tym crowdsourcingowych (na przykład Wikipedia), zoptymalizowanych wyników wyszukiwań pochodzących z Bing, Yahoo! oraz Yandex, a także działań własnego webcrawlera DuckDuckBot.

Drugą z wyszukiwarek szanujących prywatność jest Startpage.com (istniejąca od 1998 roku), dawniej IxQuick.com (do 2016 roku), reklamująca się jako „Najbardziej prywatna wyszukiwarka na świecie". Jej zasada działania jest prosta – jest ona pośrednikiem (*proxy*) między Google a użytkownikiem końcowym. W imieniu użytkownika składa zapytania serwerom Google, a wyniki, bez trackingu i gromadzenia danych osobowych, przekazuje użytkownikowi końcowemu. Warto zwrócić uwagę, iż wyszukiwarka ta nie notuje i nie przechowuje jakichkolwiek danych użytkownika, w tym jego adresu IP.

Kolejną wyszukiwarką przyjazną anonimowości użytkowników jest Searx (http://searx.me) – metawyszukiwarka, która korzysta z innych wyszukiwarek, nie udostępniając jakichkolwiek danych składającego zapytanie. Ponadto sama nie zapisuje żadnych danych wyszukującego – zapewnia użytkownikowi wyszukiwanie poprzez mechanizm HTTP

[33] D. Parrack, *DuckDuckGo Denies Using Browser Fingerprinting*, http://www.makeuseof.com/tag/duckduckgo-denies-browser-fingerprinting/ (dostęp: 12.02.2019).

[34] Więcej na ten temat: E. Pariser, *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*, Londyn 2012, a poglądy kwestionujące istnienie bańki filtrującej odnajdziemy w: P. Boutin, *Your Results May Vary*, http://web.archive.org/web/20151214060050/http://www.wsj.com/articles/SB10001424052748703421204576327414266287254, strona obecnie dostępna *via* IWM.

POST, co uniemożliwia zapis w logach serwera. Każdy z wyszukanych odnośników jest bezpośredni (a nie jak w Google – podawany w postaci linku przekierowania i śledzony). Wyszukane strony można przeglądać jako zbuforowane, a więc nie pozostawiając swojego cyfrowego odcisku palca na ich serwerach.

Do wyszukiwarek szanujących prywatność użytkownika zaliczana jest również Qwant (http://www.qwant.com), której twórcy podkreślają (niezgodnie z prawdą), że jest jedyną z europejskich wyszukiwarek. Deklarują również, iż nie zapisuje ona historii wyszukiwania użytkownika oraz plików *cookies*.

METAWYSZUKIWARKI I MULTIWYSZUKIWARKI

Pojęcia metawyszukiwarki i multiwyszukiwarki są zazwyczaj utożsamiane[35]. Na ogół rozumie się je jako narzędzia (programy) działające jednocześnie na wielu serwisach wyszukiwawczych. Na potrzeby dydaktyczne można zaproponować, by pojęciem multiwyszukiwarki określać te narzędzia, które wyłącznie grupują wyniki wyszukiwania wielu wyszukiwarek, a pojęciem metawyszukiwarki takie narzędzia, które dodatkowo zawierają algorytm indeksowania i rangowania, a w efekcie prezentowania użytkownikowi odnalezionych zasobów. Takie rozróżnienie wydaje się mieć istotne znaczenie w ocenie i walidacji pozyskiwanych informacji[36]. Przesłanką korzystania z więcej niż jednej wyszukiwarki jest fakt, że żadna z istniejących nie indeksuje zasobów w identyczny sposób (istnieje kilka algorytmów indeksowania, na przykład binarny, PageRank, klikohit), w tym samym czasie (obieg robota sieciowego wyszukiwarki trwa kilka–kilkanaście dni) i w efekcie nie pokrywa całości zasobów.

Klasyczną multiwyszukiwarką w zdefiniowanym wyżej znaczeniu był Bjorgul (http://www.bjorgul.com) umożliwiający wyszukiwanie w każdej z parudziesięciu wyszukiwarek razem i osobno. Obecnie (tymczasowo

[35] Patrz na przykład: S. Cisek, *Warsztat infobrokera – poszukiwanie informacji*, http://www.academia.edu/32396257/Warsztat_infobrokera_-_poszukiwanie_informacji (dostęp: 12.02.2019).

[36] Jeszcze inne rozumienie pojęcia „metawyszukiwarka" proponuje Dominika Paleczna – uznaje ona, że metawyszukiwarki to takie systemy wyszukiwań, które odpytują systemy zdalne w czasie rzeczywistym, co w zwykłym użytkowaniu przekłada się na długi czas oczekiwania na wynik. Zalicza do nich m.in. Katalog Rozproszony Bibliotek Polskich (KaRo), Bazy Biblioteki Narodowej oraz Mazowiecki System Informacji Bibliotecznej (EHIS). Takie rozumienie nie koliduje jednak z przyjętym – rozważania autorki dotyczą specjalistycznych systemów bibliotecznych, nie zaś wyszukiwarek globalnych. D. Paleczna, *Systemy discovery vs. metawyszukiwarki*, http://nowetrendy.bibliosfera.net/2014/08.systemy_discovery.pdf (dostęp: 12.02.2019).

lub stale) nie funkcjonuje. Z kolei Etools (http://www.etools.ch/) pełni jednocześnie funkcję metawyszukiwarki i multiwyszukiwarki. Umożliwia jednocześnie wyszukiwanie zsyntetyzowane, jak również eksplorację wedle pojedynczych wyszukiwarek na jednej stronie. Wyszukiwarka ta konsumuje wyniki 17 innych, a algorytm syntetycznego wyszukiwania pozostaje nieujawniony. Podany ranking narzędzi daje jednak orientację o rangach nadawanych poszczególnym indeksom cząstkowym wyszukiwarek[37]. Wyszukiwarka Izito (http://www.izito.com) jest metawyszukiwarką, korzysta bowiem jednocześnie z Yahoo!, Bing, zasobów – jak Wikipedia, a także YouTube oraz Entireweb, a algorytmy selekcji i rangowania pozostają nieujawniane przez twórców. Również metaEureka (https://www.metaeureka.com/) może być uznana za metawyszukiwarkę, stanowiąc wyjątek pod tym względem, że jej twórcy opisują indeks punktowy przyznawany określonym zasobom[38]. Wartościowa wydaje się wyszukiwarka Dogpile, pomimo faktu, że algorytm, którym się posługuje, pozostaje tajemnicą. Jej twórcy wdrożyli ją na podstawie interesujących studiów odnoszących się do pokrycia obszaru wyszukiwań przez poszczególne wiodące wyszukiwarki. Okazuje się, że poszczególne wyniki wyszukiwań w różnych wyszukiwarkach pokrywają się w mniej niż jednym procencie[39]. Warto również wspomnieć o innych wyszukiwarkach funkcjonujących długo, lecz nieujawniających swoich algorytmów działania, na przykład Entireweb (http://www.entireweb.com/), Gigablast (http://www.gigablast.com/), Lycos (http://www.lycos.com/).

WYSZUKIWARKI OFERUJĄCE AGREGOWANIE TREŚCI (TEMATYCZNIE, CHRONOMETRYCZNIE, LOKALIZACYJNIE)

Wyniki wyszukiwania we wszystkich większych wyszukiwarkach podawane są w postaci jednolitego, ciągłego rankingu, którego porządek wyznaczany jest przez liczbę punktów przyznanych według ustalonych algorytmów. Wyszukiwarki takie jak Carrot2 oraz Yippy (http://yippy.com/search) umożliwiają pozyskiwanie treści w postaci zagregowanej.

Wyszukiwarka Yippy umożliwia wyszukiwanie według źródeł, to jest miejsca zmieszczenia informacji (wówczas treści prezentowane są na przykład w podziale na te zamieszczone w prasie, zamieszczone na stro-

[37] *How Does eTools.ch Work?*, http://www.etools.ch/searchInfo.do (dostęp: 12.02.2019).

[38] *How is this Working?*, https://www.metaeureka.com/help.shtml (dostęp: 12.02.2019).

[39] *Different Engines, Different Results Web Searchers Not Always Finding What They're Looking for Online A Research Study by Dogpile.com*, 2007, http://cdn1.inspsearchapi.com/dogpile/11.6.0.452/content/downloads/overlap-differentenginesdifferentresults.pdf (dostęp: 12.02.2019).

nach agencji informacyjnych, stacji telewizyjnych itd.), według domen (wówczas możemy spodziewać się jako wyników wyszukiwania katalogów z nazwami domen, jak .com, .org, .edu, .net), czasu (zakreślane przez użytkownika wyszukiwarki interwałowo), tematyki (algorytm ustalany oddolnie, na podstawie analizy treści stron internetowych). Z kolei Carrot2 oferuje przeszukiwanie w podziale na eksplorację stron internetowych z użyciem metawyszukiwarki eTools, przeszukiwanie Wikipedii, wyszukiwanie w PubMed (baza danych obejmująca artykuły z dziedziny medycyny i nauk biologicznych, zawiera ponad 26 mln rekordów, w tym publikacji pełnotekstowych w wolnym dostępie) oraz PUT (narzędzie Politechniki Poznańskiej wykorzystujące eTools). Wyniki wyszukiwania prezentowane są w katalogach tematycznych. Niekiedy element usługi świadczonej przez wyszukiwarkę stanowi również wizualizacja treści, jak w przypadku Carrot2.

WYSZUKIWARKI I KATALOGI LOKALNE

Do najbardziej udanych projektów wyszukiwarek lokalnych należą rosyjski (obecnie rosyjsko-holenderski) Yandex[40] oraz czeski Seznam. Wyszukiwarka Yandex wdrożona została w 1997 roku jako wyszukiwarka internetowa, obecnie znajduje się globalnie w pierwszej dziesiątce wyszukiwarek, a w Rosji kontroluje ponad połowę rynku wyszukiwarek. Oferuje – oprócz wyszukiwania – ponad 70 różnego rodzaju usług online, między innymi nawigacji (Yandex.Navigator), tłumaczenia (Yandex. Translate), zamawiania taksówek (Yandex.Taxi), poczty elektronicznej (YandexMail), dysku w chmurze (Yandex.Disk). Wyszukiwarka Seznam[41] według rankingu Alexa zajmuje w Czechach trzecie miejsce spośród najczęściej otwieranych stron, mając aż 62% udziału w rynku wyszukiwarek – dystansuje Google z zaledwie 29% udziału w rynku[42]. Katalogi, zarówno globalne (Jasmine Directory, https://www.jasminedirectory.com) jak i lokalne, są technologią schyłkową, wyparły je bowiem wyszukiwarki. Pierwotnie zasoby Internetu usiłowano katalogować, lecz wraz z lawinowym wzrostem tych zasobów taki sposób agregowania treści stał się nieopłacalny i niemożliwy (na przykład Yahoo!, a i wiele innych wyszukiwarek, pierwotnie powstawały jako katalogi). Aktualnie sens funkcjonowaniu katalogów nadaje pozycjonowanie stron (notabene wiele katalogów

[40] *Yandex*, http://www.yandex.ru, www.yandex.com (dostęp: 12.02.2019).

[41] *Seznam*, http://seznam.cz (dostęp: 12.02.2019).

[42] G. Marczak, *Czeska wyszukiwarka seznam warta miliard dolarów!*, Antyweb, 18.08.2008, http://antyweb.pl/czeska-wyszukiwakra-seznam-warta-miliard-dolarow/ (dostęp: 12.02.2019).

wprost nawiązuje do tego procederu, na przykład Katalog SEO, https://katalogseo.net.pl). Wyczerpujący przegląd katalogów lokalnych zawiera OpenKontakt (http://www.openkontakt.com/pl/wyszukiwarki).

WYSZUKIWARKI LUDZI

Wyszukiwarki ludzi, w szczególności w wolnym dostępie, nie są zbyt powszechne. Do najbardziej udanych wydaje się należeć wyszukiwarka Yasni wyświetlająca na jednej stronie wszystkie ogólnodostępne informacje oraz pogrupowane wyniki wyszukiwana dla wybranego nazwiska: teksty, fotografie, inne dane, artykuły w mediach, profile społecznościowe, wypowiedzi na forach. Po stworzeniu konta i zalogowaniu się można doprecyzować informacje na swój temat[43]. Kolejna wyszukiwarka – Pipl jest przedsięwzięciem komercyjnym, jednak dostarczającym nieodpłatnie usługi informacyjne w wersji podstawowej. Jej twórcy twierdzą, iż jest to największa na świecie wyszukiwarka osób łącząca publicznie dostępne informacje online i offline z wielu źródeł[44]. Do innych tego typu wyszukiwarek należą między innymi Snitch (http://snitch.name) oraz PeekYou (https://www.peekyou.com/). Należy podkreślić, że wyszukiwarki ludzi nie stanowią narzędzia pierwszego wyboru przy ich poszukiwaniu. Istnieją liczne, bardziej profesjonalne i skuteczne metody, zarówno zautomatyzowane (vide: Maltego, Oryon OSINT Browser), jak i ręczne (na przykład algorytmy przeszukiwania mediów społecznościowych)[45].

Od 2016 roku oferowane są w Internecie nieodpłatne usługi wyszukiwania imion i nazwisk na podstawie numeru telefonicznego. Dostępne są one pod adresami: http://www.truecaller.com oraz http://www.sync.com. Warunek skorzystania z usługi stanowi zgoda na pobranie całości swojej książki adresowej z konta Google. Książka teleadresowa Google stanowi źródło informacji o numerach telefonicznych. Ominięcie tej niedogodności jest banalne i polega na założeniu nowego konta, które nie zawiera tego typu informacji. Pierwsza z wymienionych baz zawiera jakoby ponad trzy miliardy numerów telefonicznych, druga zaś – prawie miliard[46].

---

[43] *Yasni*, http://www.yasni.com (dostęp: 12.02.2019).

[44] *Pipl*, https://pipl.com/ (dostęp: 12.02.2019).

[45] Warto zdecydowanie polecić zestawienia Marcusa P. Zillmana uzupełniane, uaktualniane i publikowane przezeń systematycznie: M.P. Zillman, *Finding People Resources and Sites 2019*, http://whitepapers.virtualprivatelibrary.net/Finding%20People.pdf (dostęp: 12.02.2019).

[46] A. Haertle, *Jak ustalić nazwisko posiadacza numeru telefonu – Twoje pewnie też*, Zaufana Trzecia Strona, 26.11.2016, https://zaufanatrzeciastrona.pl/post/jak-ustalic-nazwisko-posiadacza-numeru-telefonu-twoje-pewnie-tez/ (dostęp: 20.01.2019).

Istnieją liczne możliwości wyszukiwania użytkowników (za pomocą ich nazw własnych, to jest *nicknames*) w mediach społecznościowych, na przykład CheckUsernames (https://checkusernames.com/), UserSherlock (http://www.usersherlock.com/), KnowEm? (https://knowem.com/) oraz UserSearch (https://usersearch.org). Istotnym identyfikatorem użytkownika w sieci jest adres poczty elektronicznej. Powstały liczne usługi walidacji i permutacji adresów poczty elektronicznej. Funkcjonowanie serwisów walidujących (na przykład Bulk Validation – http://leopathu.com/verif-email, Hunter – http://hunter.io/email-verifier) polega na weryfikacji adresu poczty elektronicznej (prawidłowość formatu adresu, istnienie i odpowiadanie serwera w danej domenie, akceptacja danego adresu e-mail przez ten serwer). Działanie permutatorów polega na systematycznym generowaniu i sprawdzaniu potencjalnych adresów poczty elektronicznej po wprowadzeniu danych wejściowych (na przykład imienia, nazwiska, potencjalnego *nickname*, ewentualnie domeny itd.). Tego typu programy istnieją zarówno online (na przykład http://inteltechniques.com/OSINT/email.html), jak i w postaci oprogramowania służącego do tego celu (na przykład narzędzie Querytool w Oryon OSINT Browser).

Odrębną kategorię stanowią agregatory treści zamieszczanych w mediach społecznościowych – umożliwiające wyszukiwanie treści zamieszczanych przez użytkowników w przystępnej, przejrzystej formule. Do tego typu narzędzi należy StalkScan (https://stalkscan.com). Za jego pomocą można przeglądać zasoby każdego z użytkowników medium społecznościowego Facebook w podziale na znajomych, fotografie, zainteresowania i inne. Informacje te nie różnią się od tych możliwych do pozyskania przy bezpośrednim przeglądaniu profilu, jednak przewagą tego narzędzia jest wygoda jego użytkowania.

Warto wskazać wyszukiwarkę zawierającą konta, do których hasła wyciekły (są to głównie dane kont usług poczty elektronicznej, dostępu do serwisów społecznościowych oraz sklepów internetowych). Usługa została nosi nazwę HaveIBeenPwned? i udostępniono ją pod adresem https://haveibeenpwned.com. Jest to serwis prowadzony przez pracownika Microsoft Troya Hunta. Umożliwia ocenę bezpieczeństwa własnych kont na podstawie pokaźnej bazy zawierającej blisko 7 mld rekordów. CERT Polska oraz Orange Polska przygotowali polską wersję tej strony: https://www.cert.orange.pl/haveibeenpwned. Ze znacznie szerszego zakresu usług możemy skorzystać w serwisie WeLeakInfo (https://weleakinfo.com/) zawierającego blisko 9 mld rekordów. Wyszukiwać wycieki możemy według następujących kryteriów: nazwy użytkownika, adresu

poczty elektronicznej, hasła, adresu IP, a nawet nazwiska lub numeru telefonu. Serwis świadczy również odpłatną usługę udostępniania danych (na przykład dostęp do całodziennego wyszukiwania to koszt zaledwie 2 dolarów, natomiast uiszczenie 666 dolarów pozwala uzyskać dostęp dożywotni). Odpłatną usługę wyszukiwania świadczy również serwis Citadel (http://citadel.pw) funkcjonujący od 2017 roku i zawierający – według szczegółowej deklaracji – ponad 9,6 mld rekordów. Z kolei serwis LeakedSource (http://www.leakedsource.ru, dawniej znajdujący się w domenie .com) oferuje możliwość sprawdzenia, czy konkretny adres e-mail znajduje się w wyciekach informacji, ogniskując się na serwisach Tumblr i LinkedIn. Usługi dostępne są w wariancie darmowym i odpłatnym – rozszerzonym. Aktualne wycieki (nazwę strony/usługi, datę oraz zakres wycieku) można zweryfikować w Hacked Emails (https://hacked–emails.com/) dokumentującym wycieki 14 mld rekordów.

WYSZUKIWARKI OPARTE NA MAPACH I GEOLOKALIZACYJNE

Właściwości, możliwości i ograniczenia map takich jak Google (http://maps.google.com), Bing (https://www.bing.com/maps) czy OpenStreetMap (http://www.openstreetmap.org/) są raczej dobrze rozpoznawane. Warto omówić kilka mniej znanych.

Wikimapia (http://wikimapia.org/) to serwis umożliwiający oddolne, zbiorowe opisywanie obiektów geograficznych. Zawiera wiele nie tylko ciekawych, lecz również praktycznych informacji umożliwiających dobrą orientację w terenie. Informacje nie są standaryzowane, zróżnicowana jest ich liczba, jakość i stopień szczegółowości w poszczególnych miejscach mapy. Serwis ten zawiera liczne informacje niedostępne na innych mapach: w tym elementy mikrotopografii, nazw lokalnych – nawet mikrotoponimów, ciekawostki historyczne i współczesne.

Serwis Mapillary (http://www.mapillary.com/) jest usługą tworzoną oddolnie, przez użytkowników. Zawiera geotagowane i naniesione na mapę fotografie. Omija ograniczenia Google, które zamazuje między innymi numery rejestracyjne pojazdów czy twarze uwiecznione na fotografiach. Fotografie wyposażone w znaczniki czasowe często zamieszczane są w dłuższych seriach. Z tego powodu (dla najbardziej uczęszczanych miejsc) usługa ta wydaje się przydatna jako narzędzie śledcze odsłaniające, choć wybiórczo, *status quo ante*. Analogiczne informacje pozyskamy również dzięki OpenStreetCam (https://openstreetcam.org).

Przydatne mogą okazać się mapy dostarczające odwzorowań historycznych – na przykład Historic Aerials (http://historicaerials.com),

która zawiera mapy z lat 1957–2015, jednak nie mają one charakteru globalnego i wyczerpującego. Podobne możliwości wyszukiwania dostępne są w serwisach: Terra Server (http://terraserver.com) – od 1997 roku oraz Land Viewer (http://eos.com) – fotografie od 1982 roku.

Infobroker Michael Bazzell stworzył funkcjonalność agregującą różnorakie usługi związane z mapami na jednej stronie – multiwyszukiwarkę geolokalizacyjną[47]. Jest to wartościowe narzędzie – niemal pełny zbiór usług związanych z mapami umożliwiający bezpośrednie w nich wyszukiwanie. Znajdują się tam między innymi: komplety map Bing, Google i Yandex, a także liczne mapy satelitarne (między innymi Zoom Earth Satellite, Land Viewer Satellite, Descartes Satellite).

Wyszukiwarki geolokalizacyjne przeznaczone są do ekstrakcji metadanych znajdujących się w serwisach społecznościowych w celu wydobycia geotagowania zamieszczanych wpisów, fotografii, filmów i innych materiałów. Zazwyczaj usługi tego typu są odpłatne, wymagając comiesięcznej subskrypcji (na przykład https://app.echosec.net).

Wyszukiwarka GeoSocialFootprint umożliwia pozyskiwanie metadanych geolokalizacyjnych pochodzących z wpisów w mikroblogu Twitter (http://geosocialfootprint.com/). Użyteczne informacje na temat różnego rodzaju zdarzeń nadzwyczajnych wraz z ich lokalizacją i krótkim opisem można pozyskać z mapy GlobalIncidentMap (http://www.globalincidentmap.com). Analogiczne narzędzie stanowi Keitharm (https://keitharm.me/). Są to następujące zróżnicowane kategorie zdarzeń zarówno endo-, jak i egzogennych: pożary lasów, epidemie, aktywność grup przestępczych (gangów), incydenty graniczne, akty terrorystyczne, trzęsienia ziemi, incydenty na statkach powietrznych niestanowiące aktów terrorystycznych, incydenty w obszarze medycyny i żywności, handel ludźmi.

WYSZUKIWARKI „SZAREJ LITERATURY" (*GREY LITERATURE*) I WYSZUKIWARKI NAUKOWE

Szara literatura to pewna forma magazynowania wiedzy akademickiej lokująca się między literaturą białą, to jest co najmniej opublikowanymi, a najlepiej zrecenzowanymi wedle określonych standardów książkami, artykułami w czasopismach i publikacjach pokonferencyjnych, a tzw. czarną literaturą rozumianą jako idee, pomysły, myśli, czyli zmaterializowaną i upowszechnioną w stopniu nikłym. Do szarej literatury zalicza się preprinty, wydania elektroniczne, raporty techniczne, treść

[47] Została przezeń udostępniona pod adresem: https://inteltechniques.com/osint/maps.html (dostęp: 11.02.2019).

wykładów, zbiory danych, multimedia (nagrania dźwiękowe i wizualne, fotografie), a także niektóre zasoby Internetu, jak zasoby blogów i mikroblogów, forów, podcasty, wiki i inne media społecznościowe. Należy podkreślić, że drogi pozyskiwania szarej literatury są liczne – odnajdywać ją można w bazach tradycyjnych, w tym archiwach, bazach elektronicznych (na przykład w repozytoriach prac promocyjnych, zasobach bibliotecznych, zasobach przedsiębiorstw i organizacji), pozyskiwać z komunikacji osobistej lub zdalnej[48]. Zalety i wady posługiwania się takimi źródłami w pracy naukowej widoczne są natychmiast: zyskujemy skrócenie obiegu myśli naukowej (o miesiące, a niekiedy o lata, tyle bowiem może trwać cykl publikacyjny), z drugiej strony nakładany jest na badacza bardziej rygorystyczny obowiązek weryfikacji źródłoznawczej pozyskanego materiału. Dostrzegalne wydają się tendencje zmierzające ku centralizacji tych zasobów – w sensie dyscyplinowym, narodowym, a nawet globalnym. Rozmaite organizacje gromadzą i udostępniają na różnych warunkach drukowaną i cyfrową szarą literaturę, jednakże kosztochłonność odnalezienia i skatalogowania sprawia, iż nie są to zbiory obszerne. Zasoby te jednak wciąż pozostają względnie rozproszone, a ich zbiory wymagają podjęcia prób dalszej eksploracji i usystematyzowania. Rozpoznawalną globalną inicjatywą jest projekt OpenGrey[49] zawierający wyszukiwarkę (choć jeszcze niezbyt zasobną) szarej literatury (na razie głównie europejskiej i w języku angielskim) oraz założony w 1992 roku Grey Literature Network Service (GreyNet)[50], którego celem jest animacja dialogu, badań i komunikacji między osobami i organizacjami w dziedzinie szarej literatury. Spośród narodowych zbiorów pokaźnym zasobem szarej literatury dysponuje The British Library[51], która zainteresowała się tą formą źródeł tuż po II wojnie światowej[52]. Odnajdziemy również na przykład

[48] Szerzej na temat kategoryzacji biała–szara–czarna literatura oraz subtypów szarej: D. Giustini, *Finding the Hard to Finds: Searching for Grey Literature*, 2010, http://www.slideshare.net/giustinid/finding-the-hard-to-findssearching-for-grey-gray-literature-2010 (dostęp: 12.02.2019); *Document Types in Grey Literature*, http://www.greynet.org/greysourceindex/documenttypes.html (dostęp: 12.02.2019). Systematyczny i – wydaje się – wyczerpujący przegląd typów szarej literatury zawiera: *Vocabulary of the Types Of Grey Literature*, http://repozitar.techlib.cz/record/3/files/idr-3_1.pdf (dostęp: 12.02.2019).

[49] *OpenGrey*, http://www.opengrey.eu/ (dostęp: 12.02.2019).

[50] *GreyNet*, http://www.greynet.org/home.html (dostęp: 12.02.2019).

[51] Wyszukiwarka znajduje się tutaj: https://ondemand.bl.uk/onDemand/search/index (dostęp: 12.02.2019). Należy nadmienić, iż większość zbiorów udostępnia się odpłatnie.

[52] S. Tillett, E. Newbold, *Grey Literature at The British Library: Revealing a Hidden Resource*, „Interlending & Document Supply" 2006, nr 34.

zasoby włoskie[53] i holenderskie[54]. Wiele zasobów narodowych czy dyscyplinowych zostało zamieszczonych na liście GreyNetu[55] – najbardziej rozpoznawalnej organizacji zajmującej się szarą literaturą oraz zostało zindeksowane przez Wikipedię[56]. Ostatnia z wymienionych list zawiera również wyszukiwarki naukowe o różnych zasadach dostępu. Korzystającą z licznych źródeł darmową wyszukiwarką literatury akademickiej, głównie białej, jest Bielfeld Academic Search Engine (BASE)[57]. Spośród baz zasobów naukowych spełniających warunki multidyscyplinarności (różnych dyscyplin, w szczególności społecznych i humanistycznych) i wolnego (nieodpłatnego) dostępu wymienić można:

- WorldWideScience.org – obejmuje dzięki wielostronnemu partnerstwu krajowe naukowe bazy danych i portale w ponad 70 krajach świata[58];
- GoogleScholar – funkcjonuje już od 2004 roku i obejmuje blisko 400 mln dokumentów[59];
- Microsoft Academic – darmowa publiczna wyszukiwarka internetowa dla publikacji naukowych i literatury opracowana przez Microsoft Research, uruchomiona ponownie w 2016 roku i wykorzystująca technologie semantycznego wyszukiwania. Jak deklarują jej twórcy, posiada ponad 375 mln dokumentów, z czego 170 mln to dokumenty naukowe, grupuje ponad ćwierć miliona autorów, blisko 50 tys. czasopism i ponad 25 tys. instytucji[60];
- ScienceOpen – stanowi nie tylko repozytorium zasobów, lecz również platformę wspierającą publikację, recenzowanie i dyskusję nad zamieszczanymi tekstami. Posiada blisko 40 mln artykułów, a dla każdego z tekstów umożliwiono śledzenie metadanych dotyczących recenzji, zmian i cytowań[61];

---

53 *Consiglio Nazionale delle Ricerche*, http://polarcnr.area.ge.cnr.it/cataloghi/bice/index.php?type=Grigia (dostęp: 12.02.2019).

54 *Grijze Literaturr in Nederland (GLIN)*, http://www.publiekwijzer.nl/bestanden.php?id=zoeknaar&db=3.2 (dostęp: 12.02.2019).

55 *GreySource A Selection of Web-based Resources in Grey Literature*, http://www.greynet.org/greysourceindex.html (dostęp: 12.02.2019).

56 *List of Academic Databases and Search Engines*, https://en.wikipedia.org/wiki/List_of_academic_databases_and_search_engines (dostęp: 12.02.2019).

57 *BASE*, http://www.base-search.net (dostęp: 12.02.2019).

58 *WorldWideScience. The Global Science Gateway*, https://worldwidescience.org (dostęp: 12.02.2019).

59 *GoogleScholar*, https://scholar.google.pl (dostęp: 12.02.2019).

60 *Microsoft Academic*, http:// academic.microsoft.com (dostęp: 12.02.2019).

61 *ScienceOpen*, https://www.scienceopen.com/ (dostęp: 12.02.2019).

- OAIster – jest to katalog z wyszukiwarką zawierający blisko 50 mln rekordów zasobów otwartego dostępu[62];
- Paperity – to agregator wolnodostępowych czasopism i dokumentów, który w swoich zasobach posiada blisko 2 mln dokumentów i ponad 4 tys. zarejestrowanych czasopism[63];
- SSRN (*Social Science Research Network*) – to repozytorium zogniskowane na naukach społecznych, humanistycznych, a także – od niedawna – biologii, chemii, inżynierii, medycynie, informatyce i innych. Autorzy mogą swobodnie zamieszczać i hostować swoje teksty, użytkownicy zaś mogą subskrybować spersonalizowane e-maile podawane w postaci abstraktów wraz z odnośnikami. Do pewnego stopnia można uznać portal za źródło szarej literatury, ponieważ publikowane są tam często materiały przed drukiem i w celu przedyskutowania ze społecznością użytkowników[64];
- Jurn – stanowi bezpłatne narzędzie do wyszukiwania pełnotekstowych prac naukowych w zakresie nauk społecznych, humanistycznych i ścisłych, współpracuje z licznymi bibliotekami akademickimi i rządowymi, w tym Centralną Biblioteką Komisji Europejskiej, University of Cambridge, University of California i Princeton University Library[65].

ARCHIWA INTERNETU

Archiwa Internetu powszechnie utożsamiane są z usługą Internet Wayback Machine[66] znajdującą się pod adresem http://archive.org/web/web.php, a funkcjonującą od 1996 roku. Celem powstania tej organizacji była próba zapobieżenia bezpośredniemu ulotowi treści stron internetowych, które z czasem są modyfikowane lub zamykane. Dostęp do archiwum jest bezpłatny, a historia witryn sięga 1996 roku (indeksowane z różnymi częstotliwościami, w zależności od ich popularności). Strony indeksowane zapisywane są nie tylko w sposób automatyczny – IWBM współpracuje z ponad 450 bibliotekami i innymi instytucjami za pośrednictwem pro-

[62] *OAIster*, https://www.oclc.org/en/oaister.html (dostęp: 12.02.2019) oraz wyszukiwarka: https://oaister.worldcat.org (dostęp: 12.02.2019).

[63] *Paperity*, http://paperity.org/ (dostęp: 18.02.2019).

[64] SSRN, https://www.ssrn.com/index.cfm/en/ (dostęp: 12.02.2019).

[65] *Jurn*, http://www.jurn.org (dostęp: 12.02.2019).

[66] Na temat tego przedsięwzięcia powstały ponad trzy setki artykułów naukowych, głównie w dyscyplinach takich jak nauki społeczne, bibliotekoznawstwo, technologie informacyjne. S.K. Arora, Y. Li, J. Youtie, P. Shapira, *Using the Wayback Machine to Mine Websites in the Social Sciences: A Methodological Resource*, „Journal of the Association for Information Science and Technology" 2015, nr 67.

gramu Archive-It, by zweryfikować ważne strony internetowe. Jak podają twórcy, obecnie archiwum zawiera 330 mld stron internetowych, 20 mln książek[67] i tekstów, blisko 5 mln nagrań audio (w tym 180 tys. koncertów na żywo), 4 mln filmów (w tym ponad 1,5 mln programów telewizyjnych[68]), 3 mln obrazów oraz 200 tys. programów komputerowych[69]. Po założeniu bezpłatnego konta można przesyłać własne multimedia do archiwum internetowego. Aktualnie archiwum zajmuje ponad 30 petabajtów. Od 2004 roku istnieje również brytyjska inicjatywa stworzenia archiwum Internetu – UK Web Archive (UKWA7, http://www.webarchive.org.uk/ukwa/), wspierana przez takie instytucje, jak Bodleian Libraries, Oxford University, British Library, Cambridge University Libraries, National Library of Scotland, National Library of Wales, Trinity College, Dublin, Bodleian Libraries. Inicjatywa ograniczona jest do brytyjskich stron WWW i zakłada kopiowanie ich co najmniej raz do roku. Zbieranie stron odbywa się automatycznie. Ważne strony internetowe (za takie zostały uznane głównie strony informacyjne) indeksowane są częściej (nawet codziennie). Uzupełnieniem automatycznego indeksowania jest praca ekspertów systematycznie zbierających strony internetowe[70].

Znacznie skromniejszym i pełniącym inne funkcje archiwum Internetu jest CachedPages (http://www.cachedpages.com/). Umożliwia korzystanie z wersji stron zapisywanych przez serwery jako kopie zapasowe. Usługa ta jest szczególnie przydatna, gdy aktualnie znajdująca się w Internecie wersja strony internetowej została zmodyfikowana lub gdy strona została usunięta albo gdy serwer, na której się znajduje, jest przeciążony lub doznał awarii. Buforowania stron internetowych dokonują między innymi Google i Coral, także CachedPages korzysta z nich. Ponadto interfejs wyszukiwarki umożliwia wyszukiwanie również w Archive.org. Warto podkreślić różnice pomiędzy buforowaniem w Google, które przechowuje najnowszą kopię strony sprzed od 1 do

---

[67] Program digitalizacji książek został rozpoczęty przez IWBM w 2005 r., a obecnie skanowanych jest tysiąc książek dziennie w 28 lokalizacjach na całym świecie. Książki wydane przed 1923 r. są dostępne do pobrania, a te później wydane można wypożyczać za pośrednictwem witryny Open Library.

[68] Archiwizację programów telewizyjnych rozpoczęto pod koniec 2000 r., a w 2009 r. rozpoczęto tworzenie wybranych wiadomości telewizyjnych w USA, które można było wyszukać przez napisy w archiwum TV News. Usługa ta pozwala badaczom używać telewizji jako cytowalnego i udostępnianego odniesienia.

[69] *About the Internet Archive*, https://archive.org/about/ (dostęp: 12.02.2019).

[70] Więcej na ten temat: K. Gmerek, *Archiwa internetowe po obu stronach Atlantyku – Internet Archive, Wayback Machine oraz UK Web Archive*, „Biuletyn EBIB" 2012, nr 1(128).

15 dni, a Coral, gdzie przechowywane strony są starsze (rzadziej bowiem indeksowane).

Podsumowanie

Należy podkreślić, iż ujawnienie poszukiwanych materiałów w Internecie stanowi preludium – nieodzowna jest wielostronna ewaluacja znaleziska (w zależności od rangi danych i wymaganej precyzji). W pierwszej kolejności prowadzi się krytykę zewnętrzną źródła (niższą). Jest to badanie jego cech zewnętrznych z wyłączeniem treści tego źródła. Istotą tego etapu jest ustalenie autentyczności źródła, to jest wykrycie potencjalnego falsyfikatu. W tym celu należy ustalić czas, miejsce pochodzenia (w szerokim sensie, w tym instytucjonalne) oraz autorstwo źródła. Dopełnieniem tego procesu jest wewnętrzna krytyka (wyższa) zmierzająca do ustalenia stopnia wiarygodności autora źródła lub samego źródła. Aby osiągnąć ten cel, należy dokonać interpretacji źródła, co czyni się w następujących wymiarach: syntaktycznym (formy językowe, struktura tekstu), semantycznym (znaczenie tekstu, rozumienie go) oraz pragmatycznym (podmiotowy sens tekstu, a więc interesy i poglądy autora, wpływ środowiska oraz możliwości i ograniczeń autora na ostateczny kształt źródła).

Wyłącznie techniczna, mechaniczna znajomość operatorów i wyszukiwarek analizowanych w artykule wydaje się całkowicie niewystarczająca, jeśli ich stosowanie nie podlega przemyślanym i wdrażanym w systematyczny sposób regułom heurystycznym. Jest to postępowanie żmudne, czasochłonne i powtarzalne, zatem w drodze oddolnych inicjatyw powstały liczne rozwiązania – mniej i bardziej zaawansowane pod względem informatycznym, a służące do automatyzacji wyżej analizowanych zapytań. Istnieje szereg narzędzi specjalistycznych, których element stanowi również możliwość zautomatyzowanego i uproszczonego wyszukiwania.

Jednym z najprostszych narzędzi oferujących na wpół zautomatyzowane wspomaganie wyszukiwania w Google i Bing jest program dostępny wyłącznie online o nazwie Advangle (http://advangle.com). Pozwala na wybór operatorów podanych w formie listy (na przykład *Page text*, *Domain*, *Country*, *Language*) z przejrzystego menu, a następnie zaopatrzenie ich w operandy z użyciem wygodnego okna. Efekty pracy otrzymujemy w postaci sformułowanego zapytania (operatory i operandy zestawione wedle reguł formułowania zapytań), gotowego do użycia w wymienionych wyszukiwarkach poprzez przycisk umożliwiający przekierowanie.

Istnieje również szereg bardziej zaawansowanych narzędzi dedykowanych szeroko pojętemu wywiadowi – nie tylko pozyskujemy za ich pomocą informacje odnośnie do osób, grup czy organizacji, ale przede wszystkim umożliwiają one badanie innych zasobów i typów struktur, ogniskując się na aspektach informatycznych (służyć mogą do testów penetracyjnych stron walidujących cyberbezpieczeństwo określonych podmiotów). Zaliczymy do nich narzędzia takie jak: Oryon OSINT Browser, Maltego SpiderFoot oraz FOCA (*Fingerprinting Organizations with Collected Archives*)[71].

Oryon OSINT Browser (dawniej: Oryon C Portable, Oryon Environment) stanowi aktualnie najpotężniejsze i najlepsze nieodpłatne narzędzie do prowadzenia wywiadu jawnoźródłowego. W pakiecie znajduje się 18 grup narzędzi, ponad 70 specjalistycznych programów przydatnych do codziennej pracy specjalisty pozyskującego dane w Internecie (odzyskiwanie danych, informatyka śledcza, *metadata harvesting* i inne), więcej niż 600 linków do rozmaitych wyszukiwarek i innych narzędzi. Wyposażony jest między innymi w rozwiązania zapewniające anonimowość (anonimowe surfowanie po Internecie, anonimowe maile i komunikatory, generatory fałszywych tożsamości), narzędzia analizy domen, wyszukiwarki osób i firm (rozmaite kryteria, wyszukiwanie w mediach społecznościowych), wyszukiwarki w DeepWeb, DarkWeb oraz HistoricalWeb, programy do analizy i wyszukiwania fotografii i filmów oraz wyszukiwania i analizy metadanych, mechanizmy Fake News Detection (ciekawe, lecz jeszcze słabo funkcjonujące rozwiązania). Dysponuje także Query Tool – narzędziem wspomagającym wyszukiwanie autorstwa polskiego infobrokera Marcina Mellera. Jego działanie polega na agregacji i automatyzacji wyników wyszukiwań różnych wyszukiwarek. Zawiera moduł ochrony anonimowości oraz komfortu użytkownika (adblock, proxy, brak rejestracji zapytań oraz cookies, ssl). Działa w systemie operacyjnym Windows, Mac, ponadto można je uruchomić w systemach Linux, wykorzystując na przykład Wine (umożliwiający zaimplementowanie środowiska Windows w systemach Linux).

Maltego, produkt informatyków z Republiki Południowej Afryki, pozwala na agregację informacji zamieszczanych w Internecie, ich przystępną i atrakcyjną wizualizację, przydatną również do wygodnej pracy z informacją. Jest narzędziem odpłatnym, choć darmowa (ograniczona czasowo) wersja programu dostępna jest pod adresem: https://www.

[71] Zastosowanie i reguły działania narzędzia zostały przybliżone przez Wojciecha Mincewicza w niniejszym tomie, dlatego tu zrezygnowano z opisu tego programu.

paterva.com/web7/downloads.php. Stanowi potężne narzędzie służące do pozyskiwania informacji metodą tzw. białego wywiadu (OSINT), choć do pewnego stopnia jest również narzędziem autoinwigilacji – przetwarzanie zapytań odbywa się (poza niektórymi wysokopłatnymi ofertami) wyłącznie na serwerach producenta programu.

SpiderFoot (aktualnie w wersji 2.12) jest jawnoźródłowym narzędziem stworzonym przez Steve'a Micallefa. Został zaprojektowany tak, aby był łatwy w użyciu, szybki i rozszerzalny. Jest to narzędzie służące do zdalnego rekonesansu (również w opcji pasywnej – niezauważalnej dla rozpoznawanego podmiotu). Prowadzi automatycznie wyszukiwanie, wykorzystując ponad setkę publicznych źródeł danych w celu zebrania informacji, między innymi na temat adresów IP, nazw domen, adresów e-mail i innych elementów.

Skuteczne użycie wymienionych narzędzi jest pozornie łatwe (i tak jest faktycznie na płaszczyźnie ich obsługi). Jednakże bez opanowania klasycznego wyszukiwania z użyciem operatorów oraz pozyskania wiedzy z zakresu działania i struktury zasobów Internetu ich wykorzystanie będzie nie w pełni efektywne, a nawet całkowicie nieefektywne. Sztuka i nauka wyszukiwania, ewaluacji, analizy i przetwarzania informacji pozyskiwanej z Internetu wciąż jest doskonalona, niemniej efekty tych udoskonaleń nie są znane ani powszechnie używane, dlatego niniejszy tekst zawiera silny rys dydaktyczny, mający na celu również popularyzację technik i narzędzi wyszukiwawczych.

Streszczenie

Tekst składa się z dwóch części tematycznych. W pierwszej przeanalizowano techniki wyszukiwania, w sensie ogólnym, to jest heurystyki, oraz szczegółowym, czyli konkretne techniki należące do rodziny języków zapytań – operatory (w podziale na operatory logiczne, lokalizacyjne, wyszukiwania kanałów komunikacyjnych w mediach społecznościowych, chronometryczne, wyszukiwania w treści strony www oraz operatory wyszukiwania określonych typów treści). Ich zasadniczą funkcję stanowi doprecyzowanie zapytań dla wyszukiwarek. Druga część tekstu zawiera przegląd i analizę wybranych narzędzi eksploracji Internetu – wyszukiwarek internetowych (wyszukiwarek globalnych, wyszukiwarek zogniskowanych na prywatności użytkownika, metawyszukiwarek i multiwyszukiwarek, wyszukiwarek i katalogów lokalnych, wyszukiwarek ludzi, wyszukiwarek szarej literatury i wyszukiwarek naukowych, archiwów Internetu). Dokonany przegląd nie ma

charakteru wyczerpującego, jest autorski i służy raczej do wstępnej orientacji zainteresowanym w uniwersum wyszukiwarek internetowych.

Daniel Mider

**THE ART OF SEARCHINIG ON THE INTERNET.
REVIEW OF SELECTED TECHNIQUES AND TOOLS**

The text consists of two parts. The first analyzed the internet search techniques – in a general heuristic sense and detailed i.e. specific techniques belonging to the family of query languages – so called operators (logical operators, localization operators, operators for communication channels in social media, chronometric operators, search operators in the content of the www and search operators for specific types of content). Their main function is to clarify search queries. The second part of the text contains a review and analysis of selected internet exploration tools – search engines (global search engines, search engines focusing on user privacy, metasearch engines and multisearch engines, regional search engines and catalogues, people search engines, search engines of gray literature and internet archives). Preview is not exhaustive or deepened, it serves rather the initial orientation of those interested in the search engine universe.

KEY WORDS: *information society, open source intelligence, infobrokering, search engine hacking*

Załącznik

Zestawienie operatorów dla wyszukiwarek Bing, DuckDuckGo, Google, Yahoo! i Yandex

Typ operatora	Operator	Google	DuckDuckGo	Yahoo!	Yandex	Bing
Operatory logiczne (w tym G. Boole'a)	- Google, Yahoo!, Bing, DDG, Yandex NOT Bing, DDG	+	+	+	+	+
	+ Google, Yahoo!, Yandex, DDG & Bing && Bing AND Bing, Google	+/-	+	+	+	+
	| Bing, Yahoo! || Bing OR Google, Bing, DDG, Yahoo!	+	+	+	- (?)	+
	*	+	+	+	+	+
	.	+	-	-/+	-	-
	..	+/-	+	+	+	+
	!	-	+	-	+	-
	„"	+	-/+	-/+	+	-/+
	()	+	+	+	+	+
Operatory lokalizacyjne	\	-	+	-	-	-
	site:[url]	+	+	+	+	+
	url:[url]	+/-	+	+	+	+

Typ operatora	Operator	Google	DuckDuckGo	Yahoo!	Yandex	Bing
Operatory lokalizacyjne	inurl:[tekst] allinurl:[tekst] url:[tekst] dla Yandex	+	+	-	+/-	-
	location:[kod iso] dla Bing loc:[kod iso] dla Bing region:[kod regionu] dla DDG r:[kod regionu] dla DDG cat:[kod regionu] dla Yandex	-	+	-	-/+	+
	altloc:[iso code]	-	-	-	+	+
	domain:[url]	-	+	+	-	+
	ip:[adres IP]	-	+	+	-	+
	host:[url dokładnie]	-	-	-	+	-/+
	rhost:[odwrócony url + operator zastępowania ciągu]	-	-	-	+	-/+
Operatory kanałów komunikacyjnych w mediach społecznościowych	#	+	+/-	+	-	+/-
	@	+	+	+	-	+
	feed:[nazwa kanału]	-	-	-	-	+
	blogurl:	+	-	-	-	-
	hasfeed:[nazwa kanału]	-	-	-	+	+
Operatory chronometryczne	cache:[url]	+	-	-	-	+/-
	date:				+	
	daterange:[data juliańska]-[data juliańska] dla Google daterange:[data]..[data] dla Yandex	-/+	-	-	-/+	-

Typ operatora	Operator	Google	DuckDuckGo	Yahoo!	Yandex	Bing
Operatory wyszukiwania w treści strony	linkfromdomain:[url] dla Bing link:[url] dla Google, Yahoo!	–/+	–	–/+	–	–/+
	inanchor:[tekst]	+/–	–	–	+	+/–
	info:[url]	+	–	–	–	–
	intitle:[tekst] allintitle:[tekst] title:[tekst] dla Yandex	+	+	+	+	+
	meta:[tekst]	–	–	–	–	–/+
	intext:[tekst] dla DDG allintext:[tekst] dla Google inbody:[tekst] dla Bing	+	+	+	+	+
	keyword:[tekst] instreamset:[tekst]	–	–	–	–/+	+/–
	language:[kod języka] dla Bing lang:[kod języka] dla Yandex	–	–	+	+/–	+
	prefer:[tekst]	–	–	–	–	–/+
	literalmeta:[tekst]	–	–	–	–	+
	near:[liczba – maksimum] around([liczba – maksimum])	+/–	–	–	–	+/–
Operatory wyszukiwania określonych typów treści	~	+/–	–	–	–	–/+
	msite:[tekst]	–	–	–	–	–/+

Typ operatora	Operator	Google	DuckDuckGo	Yahoo!	Yandex	Bing
Operatory wyszukiwania określonych typów treści	filetype:[rozszerzenie pliku] mime:[rozszerzenie pliku] dla Yandex ext: dla Bing	+	+	+	+	+
	contains:[rozszerzenie pliku]	–	–	–	–/?	+
	imagesize:[słowa: small, medium lub large]	–	–	–	–	+
	related:[url]	+	–	–	–	–
	book:[tytuł]	+	–	–	–	–
	author:[nazwisko]	+	–	–	–	–
	maps:[lokalizacja]	+	–	–/+	+	+
	define: [tekst]	+	–	+	–	–
	movie:[tekst]	+/–	+/–	+/–	–	+/–
	weather:[tekst]	+/–	+/–	+/–	–	+/–
	phonebook:[tekst] bphonebook:[tekst] rphonebook:[tekst]	–/+	–	–	–	–
	convert:	+	–	–	–	–

Legenda:
+ Operator występuje i funkcjonuje w danej wyszukiwarce
– Operator nie występuje i nie funkcjonuje w danej wyszukiwarce
+/– Operator występuje i z pewnymi ograniczeniami i błędami (wyszukiwania) funkcjonuje w danej wyszukiwarce
–/+ Operator występuje, jednak ze znacznymi ograniczeniami i błędami (wyszukiwania) funkcjonuje w danej wyszukiwarce
? Brak jednoznacznego potwierdzenia prawidłowego funkcjonowania operatora
[] W nawiasach kwadratowych znajdują się sugestie odnośnie do formatu operandu

Źródło: opracowanie własne.

Bibliografia

Arora S.K., Li Y., Youtie J., Shapira P., *Using the Wayback Machine to Mine Websites in the Social Sciences: A Methodological Resource*, „Journal of the Association for Information Science and Technology" 2015, nr 67.

Baran M., Cichocka E., Maranowski P., Pander W., *Cybernauci – diagnoza wiedzy, umiejętności i kompetencji dzieci i młodzieży, rodziców i opiekunów oraz nauczycieli w zakresie bezpiecznego korzystania z internetu. Raport podsumowujący badanie ex-ante*, Warszawa 2016, http://cybernauci.edu.pl/wp-content/uploads/2016/06/Cybernauci-diagnoza-wiedzy-umiejętności-i-kompetencji.-Raport.pdf (dostęp: 24.01.2019).

Bazzell M., *Open Source Intelligence Techniques. Resources for Searching and Analyzing Online Information*, 6th ed., Charleston 2018.

Benfield J.A., Szlemko W.J., *Internet-based Data Collection: Promises and Realities*, „Journal of Research Practice" 2006, nr 2(2).

Bosch A. van den, Bogers T., Kunder M. de, *Estimating Search Engine Index Size Variability: A 9-year Longitudinal Study*, http://www.dekunder.nl/Media/10.1007_s11192-016-1863-z.pdf (dostęp: 24.01.2019).

Boutin P., *Your Results May Vary*, http://web.archive.org/web/20151214060050/http://www.wsj.com/articles/SB10001424052748703421204576327414266287254, strona obecnie niedostępna.

Cisek S., *Warsztat infobrokera – poszukiwanie informacji*, http://www.academia.edu/32396257/Warsztat_infobrokera_-_poszukiwanie_informacji (dostęp: 12.02.2019).

Cisek S., *Wyszukiwarki specjalistyczne*, http://sabinacisek.blogspot.com/2012/11/wyszukiwarki-specjalistyczne.html (dostęp: 12.02.2019).

Giglietto F., Rossi L., Bennato D., *The Open Laboratory: Limits and Possibilities of Using Facebook, Twitter, and YouTube as a Research Data Source*, „Journal of Technology in Human Services" 2012, nr 30(3-4).

Giustini D., *Finding the Hard to Finds: Searching for Grey Literature*, 2010, http://www.slideshare.net/giustinid/finding-the-hard-to-findssearching-for-grey-gray-literature-2010 (dostęp: 12.02.2019).

Gmerek K., *Archiwa internetowe po obu stronach Atlantyku – Internet Archive, Wayback Machine oraz UK Web Archive*, „Biuletyn EBIB" 2012, nr 1(128).

Mangles C., *Search Engine Statistics 2018*, SmartInsights, 30.01.2018, http://www.smartinsights.com/search-engine-marketing/search-engine-statistics/ (dostęp: 24.01.2019).

Marczak G., *Czeska wyszukiwarka seznam warta miliard dolarów!*, Antyweb, 18.08.2008, http://antyweb.pl/czeska-wyszukiwakra-seznam-warta-miliard-dolarow/ (dostęp: 12.02.2019).

Mider D., *Mappa Mundi ukrytego Internetu. Próba kategoryzacji kanałów komunikacji i treści*, „EduAkcja. Magazyn Edukacji Elektronicznej" 2015, nr 2(10).

Ohiagu O.P., *The Internet: The Medium of the Mass Media*, „Kiabara Journal of Humanities" 2011, nr 16(2).

Paleczna D., *Systemy discovery vs. metawyszukiwarki*, http://nowetrendy.bibliosfera.net/2014/08.systemy_discovery.pdf (dostęp: 12.02.2019).

Pariser E., *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*, Londyn 2012.

Parrack D., *DuckDuckGo Denies Using Browser Fingerprinting*, http://www.makeuseof.com/tag/duckduckgo-denies-browser-fingerprinting/ (dostęp: 12.02.2019).

Shenk D., *Data Smog. Survuving the Information Glut*, Nowy Jork 1998.

Tillett S., Newbold E., *Grey literature at The British Library: revealing a hidden resource,* „Interlending & Document Supply" 2006, nr 34.

Vaas L., *Google's Private Browsing Doesn't Keep your Searches Anonymous*, https://nakedsecurity.sophos.com/2018/12/06/googles-private-browsing-doesnt-keep-your-searches-anonymous/ (dostęp: 15.02.2019).

Zillman M.P., *Finding People Resources and Sites 2019*, http://whitepapers.virtualprivatelibrary.net/Finding%20People.pdf (dostęp: 12.02.2019).