Web, Wayback Machine a Web Crawling

zpracování tématu ze sekce Digitální stopa
Matěj Ulrich

V dnešní době je internet zatížen velkým množstvím dat. Můžete zde nalézt jakoukoliv informaci, kterou zrovna potřebujete ke svému životu. Každá neznalost v našem světě ztrácí na relevanci, protože všechny informace, které naši předkové bedlivě schraňovali, nyní máme na dosah své ruky. Takové pohodlí má ovšem i svoji cenu.

Je všeobecně známým faktem, že cokoliv na internet nahrajeme, z něj nikdy zcela nedostaneme zpět. Jakýkoliv příspěvek na sociálních sítích může kdokoliv bez námahy stáhnout, což vytváří určitou zodpovědnost, která se přenáší na nás jako na uživatele, abychom pečlivě zvážili, jaké příspěvky budeme zveřejňovat.

Co mají všechna předchozí přirovnání společného? Dají se shrnout jedním souslovím: “digitální stopa”. Digitální stopou rozumíme cokoliv co o naší osobě lze dohledat na internetu. Může tím být naše jméno, příjmení, datum narození a jiné citlivé údaje, ale také účty na sociálních sítích, fotografie nebo webové stránky. My se zaměříme na poslední vyjmenovanou položku, a to sice webové stránky. Pro příklad a zjednodušení budu předpokládat, že čtenář zná webové stránky a jejich programování. Na našem blogu, který si pro ukázku vytvoříme, zveřejníme příspěvek. V tomto příspěvku je informace, kterou chceme sdílet s našimi návštěvníky. K názornosti této ukázky předpokládejme, že se jedná o něco citlivého nebo osobního. Po určité době se nám to, že tuto informaci sdílíme, přestane líbit, a tak ji ze stránek odstraníme. Uplyne určitá doba a my narazíme na tuto informaci na sociálních sítích. Jak je to možné?

Okno do binární minulosti

Wayback Machine, nástroj od roku 2001 zpřístupněný veřejnosti, je věc, kterou se dokáže kdokoli s připojením k internetu podívat na to, jak určitá stránka vypadala v minulosti. V současnosti (tedy v době psaní článku) má tento nástroj ve své databázi zaindexovaných 946 miliard webových stránek a nejstarší z nich se datují až do roku 1996. Wayback Machine funguje na principu takzvaného “web crawlingu”, tedy procházení odkazů na webové stránky a jejich archivaci. Jednotliví uživatelé Wayback Machine mohou požádat o archivaci stránek, které ještě zaindexovány nejsou.

Díky této stránce se v našem příkladu informace dostala do daného příspěvku na sociálních sítích. Někdo náš blog nechal zaindexovat v době, kdy se tam nacházela citlivá informace.

Web crawleři a jejich etika

Pozorný čtenář jistě již tuší, proč může být takový web crawler problematický. Sice má, tak jako každý nástroj, svá omezení, například, že neindexuje dynamické informace webu. To v praxi znamená, že nemůžete vyhledávat na Wayback Machine stránku Googlu z roku 2013 a aktivně ji používat k vyhledávání.

Wayback Machine není jediný způsob, kterým se dá stránka zarchivovat. Pochopitelně si ji nebo informaci z naší ukázky může kdokoliv stáhnout a uložit z prohlížeče do svého počítače. Je také pravděpodobné, že naše informace skončí v některém z datasetů, jež využívají marketingové společnosti k analýze trhu. To může být problematické, neboť dle zákona GDPR musí daná společnost obdržet náš souhlas k využití těchto dat pro komerční účely. Je tomu tak ovšem pouze v případě, že společnost chce použít naše osobní údaje, neboť GDPR se na jiný druh dat nevztahuje.

Nechceme na stránkách roboty

A nyní přichází otázka, jak takové problémy řešit? Jak zabránit tomu, aby se náš blog nebo jakákoliv jiná webová stránka zobrazovala v nástroji Wayback Machine nebo jiných nástrojích Wayback Machine podobných?

Existuje několik konkrétních řešení. Jestliže vyvíjíme webovou stránku pomocí HTML, lze v její hlavičce použít specifický meta tag, jenž je všeobecně známý jako “robots”. Tento HTML tag zabraňuje web crawlerům, kteří by stránku chtěli zaindexovat, aby na ní vstoupili a zaindexovali ji.

V případě, že webová stránka běží na aplikaci WordPress, obsahuje soubor se jménem “robots.txt”. V souboru se nachází informace o tom, zda mají web crawleři přístup na webovou stránku a pokud ano, na které její části. Dá se zde nastavit i jaké konkrétní web crawlery chceme vpustit na naši webovou stránku. Tato možnost je speciálně důležitá, chceme-li mít webovou stránku vyhledatelnou ve vyhledávači. Google totiž funguje také na principu web crawlingu.

Při programování v HTML lze nastavit u konkrétních odkazů, zda jimi může web crawler projít. K tomu se používá atribut “nofollow” při značce “rel”. V takovém případě web crawler zaregistruje, že daný odkaz existuje, nicméně na něj nemá přístup, takže nezaindexuje jeho obsah. Nedá se na tento tag ovšem spoléhat v každém případě. Je dobré jej zkombinovat s jinými metodami popsanými v tomto článku.

Soukromí nakonec

Na internetu nikdy nic zcela úplně nezmizí. Každá webová stránka, každý komentář, každá diskuze na fóru je součástí dokonalého ekosystému internetu. Ale dříve, než se soukromí stane součástí globální sítě počítačů, je potřeba podnikat patřičné kroky k jeho integraci. Při vytváření stránek je dobré využívat tagy, které zabraňují indexování stránek a v neposlední řadě také zvažovat, jaké informace zveřejňujeme. Protože soukromí je mnohem důležitější než být součástí internetové historie.

Zdroje:

https://www.techtarget.com/whatis/definition/Wayback-Machine

https://www.quora.com/How-does-the-Wayback-Machine-work-What-technology-does-it-use-to-save-old-versions-of-websites-and-when-did-it-start-operating

https://www.jakpsatweb.cz/robots-txt.html

https://3idatascrapingblog.medium.com/the-role-of-web-scraping-in-marketing-and-advertising-30daff029190

https://www.reddit.com/r/explainlikeimfive/comments/11mix3/what_is_so_bad_about_web_scraping/

https://iapp.org/news/a/the-state-of-web-scraping-in-the-eu