Co je webarchivace?

Proč se archivuje web

Archivovaný internet bude základním zdrojem informací pro budoucí badatele. Obrovské množství vědeckých a kulturních informací dnes vzniká výhradně v digitální podobě. Webový obsah je efemérní – obsahy na webu se velmi rychle mění, odkazy vyhnívají, informace, které byly online ještě včera, nenávratně mizí. Proto se světové paměťové instituce věnují vedle budování sbírek fyzických nosičů informací také sklízení a archivaci obsahu internetu.

Technologie archivace webu

K vlastnímu sklízení obsahu internetových stránek používá Webarchiv Národní knihovny ČR, podobně jako mnoho dalších institucí, open source crawler Heritrix . Pro hladké sklízení jsou potřeba další rozšíření nebo skripty. Crawler prochází web, stahuje obsah a vytváří obraz stránky v určitém okamžiku. Také vytváří index, který se pak používá při emulaci archivovaných stránek pro zpřístupnění.
Archivovaný obsah je ukládá do XML kontejnerů ARC nebo WARC , které pak slouží zachycení obsahu webu a přidávají také technická a administrativní metadata k uloženému obsahu.

Co lze archivovat pomocí technologie webarchivace?

V principu webarchivace znamená stažení html a css souborů, obrázků, objektů pdf, doc apod. a audio a video souborů, případně javascriptu.
Technologie webarchivace umožňují sklízet jen zlomek internetu. Nedostupná je velká část deep webu, obsahy vyžadující zaplacení nebo přihlášení, obsahy databází, problémy jsou se sklízením obsahu sociálních sítí nebo se streamovaným obsahem. Není také například snadné sklízet obsahy digitálního knihoven nebo dalších podobných aplikací.
Kromě technických omezení má sklízení i limity organizační a finanční. Webarchiv NK ČR nemá neomezené zdroje, takže například plošné sklizně českého internetu může provádět jen několikrát ročně. Další omezení je třeba nastavit na počet odkazů, které crawler sleduje, maximální velikost a počet stahovaných objektů apod.

Plošné vs. tematické a výběrové sklizně

Vlastní sklízení probíhá pomocí plošných nebo tematických a výběrových sklizní. Pravidelné plošné sklizně vytvářejí snapshot českého internetu v určitém okamžiku. Tematické sklizně se zaměřují na dokumentování dopadu konkrétní události v informačním prostoru internetu. Některé významné zdroje Webarchiv Národní knihovny ČR archivuje také výběrově, nad rámec plošných sklizní.
Strategie budování sbírky Webarchivu NK ČR

Jak se liší archivace webu od zálohování webové stránky a databáze lokálně

Z výše uvedeného popisu technologií webarchivace je snad dostatečně jasné, že webarchiv nemůže nahradit zálohování souborů tvořících webovou stránku, jejího CMS systému a databáze. Umožňuje ale i po zániku webu zpřístupnit obraz internetové stránky v určitém okamžiku.

Pro webmastery

Pro správce internetových stránek nepředstavuje sklízení crawlerem NK ČR obvykle žádné riziko. Robot Webarchivu Národní knihovny ČR se dá identifikovat v přístupových lozích a jeho přístup na některé objekty webu lze zakázat v robots.txt.

Webarchiv Národní knihovny ČR můžete na své zajímavé stránky upozornit prostřednictvím formuláře na http://webarchiv.cz/cs/pridat-web .

Technologie Webarchivace

<iframe src="//www.slideshare.net/slideshow/embed_code/key/A6nf28D8K8FUPl" width="595" height="485" frameborder="0" marginwidth="0" marginheight="0" scrolling="no" style="border:1px solid #CCC; border-width:1px; margin-bottom:5px; max-width: 100%;" allowfullscreen> </iframe>