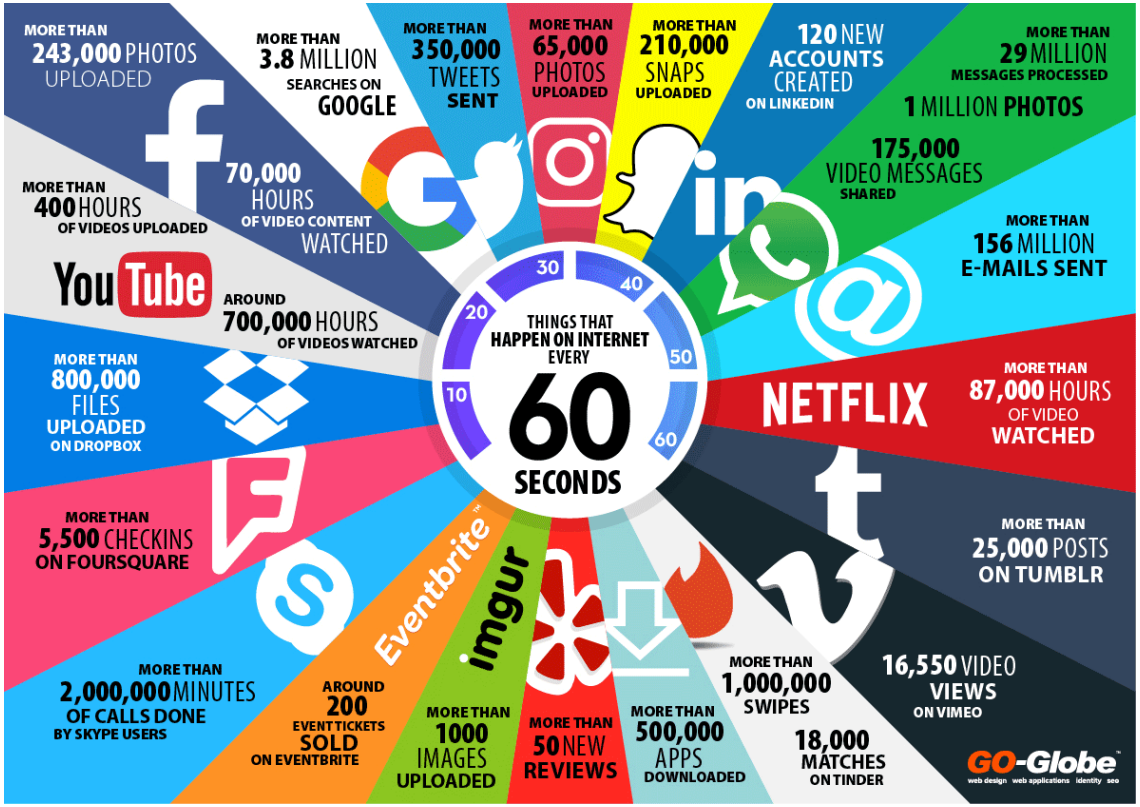
„Lessons Learned“ získané během projektu z pohledu NK ČR, vize do budoucna, strategie open-dat v NK ČR a vliv na práci s webovým archivem

Mgr. Tomáš Foltýn, Národní knihovna ČR, Klementinum, 18/10/2022



Webové zdroje jsou jedinečnou a dynamicky se rozvíjející se datovou základnou pro výzkum řady společenských jevů.

Dynamika rozšiřování zdrojů webového archivu

- v roce 2018 385 TB dat, v roce 2022 440 TB dat
- v roce 2018 9,5 miliardy unikátních objektů, v roce 2022 12 miliard objektů

Cílem správců „webového“ kulturního bohatství je otevřít data webových archivů analytickému výzkumu a pracovat na jejich propojování s dalšími sety dat.

Webarchiv uchovává český web

Pro první seznámení s webarchivem pokračujte zde

Zobrazení: vizuální, textové

Tematické sbírky

Tematické sbírky jsou monotematické soubory webových dokumentů. V rámci tematických sklizní sledujeme především takové děje, které doprovází celospolečenská debata a je u nich tedy předpoklad, že zaujmou významnější místo z hlediska dějin České republiky. Cíleně vybíráme události, které mají širší ohlas v prostředí internetu. Monitorujeme očekávané akce (např. volby), ale i nenadálé situace (např. povodně).

Válka na Ukrajině

Miroslav „Meky“ Žbirka

Miloš Zeman - hospitalizace prezidenta

Volby do Poslanecké sněmovny Parlamentu ČR 2021

Ničivá bouře / tornádo - jižní Morava, Lounsko

Vrbětice

Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů

- Mise projektu: **Řešení problematiky zpřístupnění dat z českého webového archivu a jejich poskytnutí badatelské obci pro vědecké a výzkumné využití.**
- Umožnit široké odborné veřejnosti využívat potenciálu dlouhodobě shromažďovaných a dosud z velké části nezpracovaných dat.
- Propojení a interpretace dat uložených v prostředí webového archivu.

Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů

Výstupy

- **1x Ztech :** integrace samostatných softwarových nástrojů do centralizovaného uživatelského rozhraní
- **4x R:** Software pro analýzu tématu dokumentu; Software pro analýzu audiosouborů; Databáze digitálních objektů; Software pro export datových setů
- **2x W:** prezentace projektu na počátku a konci realizace projektu
- **10 x J či D:** odborné články a studie vč. výstupů z účasti na vědeckých akcích

Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů

Lessons Learned:

- zástupci participujících oborů (knihovny, univerzity, vědecká pracoviště, soukromé společnosti) umějí spolupracovat směrem k dosažení jasného cíle
- největším limitem smysluplného zpřístupnění obsahu webové archivu je stále ještě legislativa
- moderní technologie využitelné pro oblast správy a zpřístupnění webových archivů jsou k dispozici, ale jsou mnohdy proměnlivé (riziko open-source)

Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů

Lessons Learned:

- zpracování velkých dat je ovlivněno výpočetními možnostmi HW vybavení a postupnou implementací nových technologií do praxe
- vědecká komunita nepotřebuje pouze koncová data, ale i možnost aktivně obsah webového archivu prohledávat a „dokonce“ pracovat i s metadaty (například daty o vzniku datových setů apod.)
- základem každého projektu a nositeli pokroku jsou vždy lidé

Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů

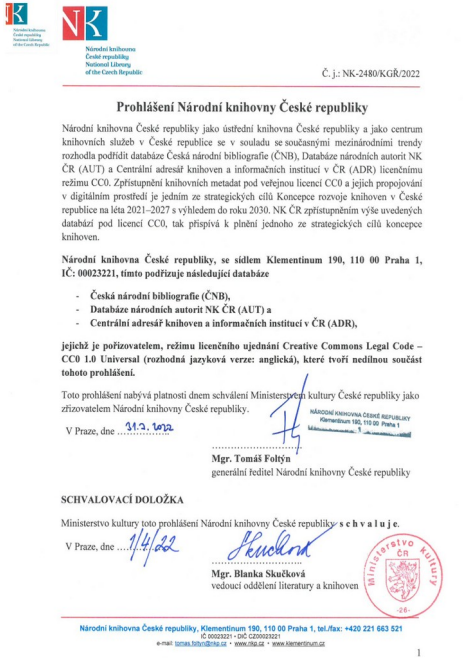
Vize do budoucna

- **optimalizace procesu ukládání data a škálovatelné úložiště**
- **implementace postupů strojového učení a automatizovaných analýz obsahu webového archivu** (například obrazové materiály)
- **tematika quality assurance** (například automatická detekce zaniklých webů atd.)
- **nové nástroje podporující kurátorskou práci** (nástroje umožňující lepší prohledávání webu na základě zadaných parametrů)

Strategie open-dat v prostředí NK ČR

Open-data?:

- Otevřená data jsou **informace a data zveřejněná na internetu, která jsou úplná, snadno dostupná, strojově čitelná, používající standardy s volně dostupnou specifikací, zpřístupněná za jasně definovaných podmínek užití dat s minimem omezení.**
- bibliografické databáze, autoritní databáze, faktografické



Národní knihovna České republiky
National Library of the Czech Republic

Č. j.: NK-2480/KGŘ/2022

Prohlášení Národní knihovny České republiky

Národní knihovna České republiky jako ústřední knihovna České republiky a jako centrum knihovních služeb v České republice se v souladu se současnými mezinárodními trendy rozhodla podřídit databáze Česká národní bibliografie (ČNB), Databáze národních autorit NK ČR (AUT) a Centrální adresář knihoven a informačních institucí v ČR (ADR) licenčnímu režimu CC0. Zpřístupnění knihovních metadat pod veřejnou licencí CC0 a jejich propojování v digitálním prostředí je jedním ze strategických cílů Koncepce rozvoje knihoven v České republice na léta 2021–2027 s výhledem do roku 2030. NK ČR zpřístupněním výše uvedených databází pod licencí CC0, tak přispívá k plnění jednoho ze strategických cílů koncepce knihoven.

Národní knihovna České republiky, se sídlem Klementinum 190, 110 00 Praha 1, IČ: 00023221, tímto podřizuje následující databáze

- **Česká národní bibliografie (ČNB),**
- **Databáze národních autorit NK ČR (AUT) a**
- **Centrální adresář knihoven a informačních institucí v ČR (ADR),**

jejichž je pořizovatelem, režimu licenčního ujednání Creative Commons Legal Code – CC0 1.0 Universal (rozhodná jazyková verze: anglická), které tvoří nedílnou součást tohoto prohlášení.

Toto prohlášení nabývá platnosti dnem schválení Ministerstvem kultury České republiky jako zřizovatelem Národní knihovny České republiky.

V Praze, dne 31.2.2022

Mgr. Tomáš Foltýn
generální ředitel Národní knihovny České republiky

SCHVALOVACÍ DOLOŽKA

Ministerstvo kultury toto prohlášení Národní knihovny České republiky **s c h v a l u j e**.

V Praze, dne 1/4/22

Mgr. Blanka Skučková
vedoucí oddělení literatury a knihoven

Národní knihovna České republiky, Klementinum 190, 110 00 Praha 1, tel./fax: +420 221 663 521
IČ 00023221 • DIC CZ00023221
e-mail: tomas.foltyn@nkp.cz • www.nkp.cz • www.klementinum.cz

1

Strategie open-dat v prostředí NK ČR

- letos proběhlo vystavení několika databází:
  - Česká národní bibliografie
  - Centrální adresář knihoven ČR
  - Databáze národních autorit
- ve formátu MARCXML, JSON, RDF/XML
- pod licencí CC0
- statická data s pravidelnou aktualizací katalogizovaná v Národním katalogu otevřených dat
  - ideálně jako propojená data
- info na webu “data.nkp.cz”

Sociologický ústav
Akademie věd ČR

ZÁPADOČESKÁ
UNIVERZITA
V PLZNI

Děkuji za pozornost!

Mgr. Tomáš Foltýn
tomas.foltyn@nkp.cz
+420739570956