Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů Webarchiv

představení projektu

Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů

- projekt je financován ze zdrojů programu NAKI II
- rozpočet projektu: 25 526 000,- Kč
- 2018 - 2022
- Národní knihovna ČR| Sociologický ústav AV ČR | Fakulta aplikovaných věd Západočeské univerzity v Plzni

Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů

- hlavní řešitel projektu: Mgr. Tomáš Foltýn
- odpovědný řešitel projektu na straně ZČR: doc. Ing. Pavel Ircing, Ph.D.
- odpovědný řešitel projektu na straně SOÚ: Mgr. et Mgr. Paulína Tabery
- technologický partner projektu: INQOOL, a. s.

Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů

- cíl: **řešení problematiky zpřístupnění dat z českého webového archivu a jejich poskytnutí badatelské obci pro vědecké a výzkumné využití**
- propojení a interpretace dat uložených v prostředí webového archivu
- umožnit široké odborné veřejnosti využívat potenciálu dlouhodobě shromažďovaných a dosud z velké části nezpracovaných dat

Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů

Datová část řešení projektu:

- **příprava dat pro badatelskou analýzu** v oblasti relevantních dat pro specifikované badatelské záměry
- **vytvoření** plně **integrovaného fasetového a full-textového vyhledávače**, který umožní badatelům jasně definovat část dat, kterou potřebují pro realizaci svého výzkumu
- **exportní aplikace**, která umožní badatelům získat datové sety pro jejich výzkum

Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů

Procesní část řešení projektu:

- **vytvoření a aplikace automatických analytických nástrojů** na relevantní data
- automatické přiřazení metadat jednotlivým dokumentům s využitím metod rozpoznávání řeči a metod sémantické analýzy textu
- **exportní aplikace**, která umožní badatelům získat datové sety pro jejich výzkum

Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů

Badatelská část řešení projektu:

- **vytvoření sociologických výzkumů**, které konkrétně aplikují zpracovaná data, ale zároveň **definují, jak mají vypadat výstupy cílů** z předchozích částí
- zájem badatelské obce o archivovaná webová data roste s možnostmi práce s nimi – nutná **postupná obměna technologické infrastruktury**, která již nemá sloužit pouze k uchování fondů, ale i k možnostem **dalšího analytického zpracování**

Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů

I. etapa projektu = analytická (2018)

- proběhla **vstupní hloubková analýza dat** z webového archivu za účelem zjištění současného stavu a prozkoumání dat
- úprava parametrů pro další výzkumné a vývojové práce
- **definice výzkumných otázek sociálně vědního výzkumu** vč. stanovení základních metodologických postupů

Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů

II. etapa projektu = analytické zpracování dat webového archivu (2019-2020)

- **extrakce a analytické zpracování dat** z webového archivu
- **vytvoření strukturovaného indexu**
- **vývoj nástrojů pro sémantickou analýzu textů** z webových archivů a pro zpracování audio souborů
- zahájení **strojového zpracování dat**

Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů

III. etapa projektu = implementace softwarových nástrojů a jejich použití (2021-2022)

- **vývoj centralizovaného rozhraní** pro vytěžování velkých dat z webových archivů
- **integrace vytvořených aplikací** v předchozích etapách do jednoho systému
- **vytvoření grafického uživatelského rozhraní** pro vyhledávání
- **vývoj exportní aplikace**, která bude sloužit ke zpřístupnění dat pro koncové uživatele v různých formátech

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

„Nepsané“ výstupy:

- právní analýza popisující postavení webového archivu a možnosti využití dat
- analýza vývojových prací včetně napojení na UX postupy
- vytvoření aktualizovaného popisu současného technologického stavu webového archivu NK ČR včetně indexu
- doplnění hardwarového prostředí NK ČR o nové technologie – např. APACHE HADOOP

Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů

Všechny dílčí cíle projektu spolu úzce souvisí a vzájemně se ovlivňují. Bez jasného badatelského záměru není možné vytvořit nebo vybrat vhodný nástroj pro analýzu dat. A bez zpřístupnění a analýzy dat není možné naplnit případný badatelský záměr.

Děkuji za pozornost!

Mgr. Tomáš Foltýn

tomas.foltyn@nkp.cz

+420739570956