Zpráva ze zahraniční služební cesty

|  |  |  |
| --- | --- | --- |
| Jméno a příjmení účastníka cesty | Zuzana Kvašová | |
| Pracoviště – dle organizační struktury | Oddělení LTP úložiště, 2.4.3 | |
| Pracoviště – zařazení | Vedoucí | |
| Důvod cesty | **Návštěva odborníků Národní knihovny Rakouska a Technické univerzity ve Vídni** | |
| Místo – město | **Vídeň** | |
| Místo – země | **Rakousko** | |
| Datum (od-do) | **20. 11. 2013 – 23. 11. 2013** | |
| Podrobný časový harmonogram | 20. 11. 2013 – odjezd z ČR,  21. 11. 2013 – 22. 11. 2013 – návštěvy institucí,  23. 11. 2013 – návrat do ČR | |
| Spolucestující z NK | Jaroslav Kvasnica, Rudolf Kreibich | |
| Finanční zajištění | VaV 0139 | |
| Cíle cesty | Cílem cesty návštěva odborníků v Národní knihovně Rakouska a na Technické univerzitě ve Vídni. Byly domluveny schůzky s jednotlivými lidmi, odborníky v oblasti sklízení internetu a dlouhodobé ochrany digitálních dat.  Cílem cesty je zjistit informace o možnostech ukládání dat z WebArchivu do digitálního repozitáře, zjistit nejnovější poznatky v oblasti dlouhodobé ochrany tohoto typu dat (projekt SCAPE) a zkušenosti s ukládáním tohoto typu dat v Národní knihovně Rakouska. Z hlediska VaV se jedná o oblast 3b. | |
| Plnění cílů cesty (konkrétně) | Cíle cesty byly splněny. Byly navázány kontakty v Národní knihovně Rakouska, na Technické univerzitě ve Vídni a na projekt SCAPE, který se zabývá dlouhodobou ochranou dat na úrovni partnerství mnoha institucí, především v Evropě. Byly zjištěny poznatky, které budou moci být využity pro ukládání dat z projektu Webarchiv do dlouhodobého úložiště a do budoucna pro vytvoření strategie dlouhodobé ochrany. | |
| Program a další podrobnější informace | 20. 11. 2013 – příjezd do Vídně  21. 11. 2013 – 22. 11. 2013 – domluvené návštěvy institucí  23. 11. 2013 – návrat do ČR | |
| Přivezené materiály |  | |
| Datum předložení zprávy | 27. 11. 2013 | |
| Podpis předkladatele zprávy |  | |
| Podpis nadřízeného | Datum: | Podpis: |
| Vloženo na Intranet | Datum: | Podpis: |
| Přijato v mezinárodním oddělení | Datum: | Podpis: |

Andreas Rauber, Technická univerzita ve Vídni

* Spolupráce na projektu SCAPE – projekty C3PO a Plato
* Pro určení strategie dlouhodobé ochrany je nutné znát jednat obsah, který má být ochraňován, tak budoucí využitelnost
* Možnost zpřístupnění sklizených webových stránek pomocí linkovaných PDF souborů
* Pro analýzu úspěšné ARC->WARC migrace navrhuje porovnání sklizní Waybacku původních a migrovaných dat

Bettina Kann, Michaela Mayr, Národní knihovna Rakouska

* Odbornice na dlouhodobou ochranu dat a webarchiv
* Data sklizejí do kontejneru ARC, WARC zatím neplánují využít
* Poskytují ochranu na úrovni bit-stream, strategii dlouhodobé ochrany plánují do budoucna společně s novým systémem pro dlouhodobou ochranu, který je nyní ve fázi tenderu na dodavatele
* Zpřístupnění dat sklizeného webu – pouze v lokální síti instituce a partnerských institucí, mají minimum uživatelů – ovšem uživatelské využití dat je důležité pro určení strategie dlouhodobé ochrany
* Metadata a data jsou v současném systému ukládána odděleně, to plánují změnit
* K datům z web archivu ukládají pouze základní technická metadata
* Mají Hadoop infrastrukturu, jsou schopni zpracovat 8 000 záznamů za hodinu při formátové analýze.

Sven Schlarb, Národní knihovna Rakouska

* Zkušenosti s migrací ARC do WARC, kolega, který je jedním z autorů migračního nástroje je dlouhodobě nemocný
* Implementace Hadoop pro zpracování velkého množství dat z web archivu – možnosti analýzy obsahu
* Pro analýzu úspěšné ARC->WARC migrace navrhuje vygenerování obrázků webů pomocí phantom.js z původních i migrovaných dat a automatizované porovnání rozdílů např. podle obrázkového hashe.

Kresimir Duretec, Artur Kulmukhametov, Technická univerzita ve Vídni

* Spolupráce na projektu SCAPE – projekty C3PO a Plato
* Zkušenost z analýzou výstupních metadat charakterizace digitální objektů
* Spolupracují s Dánskou národní knihovnou na collection profiling jejich sklizených dat
* Nabídka spolupráce pro NK ČR výstupních metadat charakterizace digitální objektů metadata z formátové analýzy dat WA NK ČR
* Vývojáři analyzačního nástroje C3PO – podle nich Droid nevhodný, doporučení FITS
* Zkušenosti s nástroji FITS a Apache Tika
* C3PO: <https://github.com/peshkira/c3po>: zpracovává a vizualizuje metadata z FITS a Apache TIKA
* Metadata mají v NoSQL databázi mongoDB v níž provádějí Map/Reduce operace
* Doporučil používat pro analýzu formátů namísto extrakce balíčku, streamline přístup. Spousta nástrojů se musí znova zavádět do paměti, čímž se snižuje jejich efektivita.