Zpráva ze služební cesty v ČR

|  |  |  |
| --- | --- | --- |
| Jméno a příjmení účastníka cesty | Mgr. Tomáš Foltýn | |
| Pracoviště – dle organizační struktury | 1.3. Odbor správy fondů | |
| Pracoviště – zařazení | Ředitel Odboru správy fondů NK ČR | |
| Důvod cesty | **Konference QQML 2015** | |
| Místo – město | **Paříž** | |
| Místo – země | **Francie** | |
| Datum (od-do) | 26. – 29. . června 2015 | |
| Podrobný časový harmonogram | 26. 5. – přílet do Paříže, účast na konferenci  27. 5. – 28. 5. účast na konferenci  29. 5. – účast na konferenci, odlet | |
| Spolucestující z NK | --- | |
| Finanční zajištění | NAKI 0142 | |
| Cíle cesty | Účast na mezinárodní konferenci QQML 2015 – prohloubení znalostí v dané oblasti, navázání kontaktů. Vystoupení s příspěvkem „How to improve OCR results? Tools for Digitizaton Process Optimizing“. Vedení odborné sekce Library and Information Science and Research“. | |
| Plnění cílů cesty (konkrétně) | Viz níže | |
| Program a další podrobnější informace | http://elag2015.org/ | |
| Přivezené materiály |  | |
| Datum předložení zprávy |  | |
| Podpis předkladatele zprávy |  | |
| Podpis nadřízeného | Datum: | Podpis: |
| Vloženo na Intranet | Datum: | Podpis: |
| Přijato v domácím oddělení | Datum: | Podpis: |

**Během účasti na konferenci byly přednesen příspěvek:**

How to improve OCR results? Tools for Digitizaton Process Optimizing“

* abstrakt příspěvku:

**How to improve OCR results?**

**Tools for Digitization Process Optimizing.**

**Tomas Foltyn1**

1 Collection Management Division Director, National Library of Czech Republic, Klementinum 190, Prague 1, Czech Republic, Phone Number +420739570956, email: tomas.foltyn@nkp.cz

**1. Text of the abstract**

The quality of image files has a direct affect to the OCR results and subsequent user possibilities of the access to the digital content via full-text searching. National Library of Czech Republic solves quite a long time the development of the tools, which can increase the OCR accuracy of the old Czech language, how it was used especially in the 19th century. The first step is to measure and improve the quality of every single scan. Next step is to choose the right documents according to the language periods of the old Czech language and thematic groups (eg. agriculture, mathematic, history etc.), which will constitute the ground of knowledge database of old Czech terms. This database is filled by number of external employees, whose are responsible for the OCRed texts recorrection and parallel segmentation. When "groundtruthed" results are available, language and IT experts start the deduplication and delete useless words as conjuctions or preposions. When these specialized processes will finish, "clean" terms are uploaded to the database and reused during next OCR production. This improves the results. All the mentioned processed are made on server-client base in HITECO tool, which was developed in the frame of Czech NAKI research funding. Its quality was verified also during the IMPACT project, when all the Czech "groundtruthed" texts were processed via this software. The presentation will describe not only the software perspectives of the project but also mentioned linguistic background in detail - namely the reuse of the content for the construction of dictionary databases and subsequent lemmatization.

**Organizační zabezpečení sekce:**

Library and Information Science and Research

* účastníci:
  + Marwin Britto – Planning a Cohesive and Informative Library Assessment Framework
  + Raluce Trifu – Quantitative Methods Used in the Evaluation Process of the Library Services
  + Olga Einasto – Conceptualization of User Participation in the Context of Library E-Services Quality
  + Milica Cvetkovic Stosic – Freedom of Information and the Library
  + Cathy Weng – Towards Sustainable Partnership : Examining Cross Perceptions of Public and Technical Services Librarians