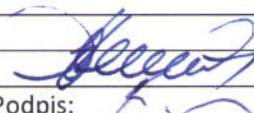


Zpráva ze zahraniční služební cesty

Jméno a příjmení účastníka cesty	Ing. Petr Knížek	
Pracoviště – dle organizační struktury	Náměstek sekce digitalizace a technologie	
Pracoviště – zařazení	---	
Důvod cesty	Konference TPDL 2018	
Místo – město	Porto	
Místo – země	Portugalsko	
Datum (od-do)	10. – 13. září 2018	
Podrobný časový harmonogram	<p>09. září 2018 – odlet do Portugalska</p> <p>09. září 2018 – přilet do Portugalska, ubytování,</p> <p>10. září 2018 – začátek konference, jednotlivé workshopy, semináře a odborné diskusní příspěvky</p> <p>11. září 2018 – pokračování konference, jednotlivé workshopy, semináře odborné diskuse účastníků</p> <p>12. září 2018 –workshopy, semináře odborné diskuse a ukončení konference</p> <p>13. září 2018 –specializované workshopy, odborné diskuse</p> <p>14. září 2018 –odlet a přilet do Prahy</p>	
Spolucestující z NK	-----	
Finanční zajištění	NK (letenky, ubytování a strava)	
Cíle cesty	Účast na konferenci, která je odborností zaměřena na digitální fondy, digitalizaci a práci na digitálními daty/fondy	
Plnění cílů cesty (konkrétně)	<p>Workshopy a prezentace :</p> <ul style="list-style-type: none"> • Linked Data Generation from Digital Libraries • Europeana hands-on session • Open Science in a Connected Society • Digital humanities <ul style="list-style-type: none"> ○ Towards better Understanding Researcher Strategies in Cross-lingual Event Analytics • A Web-Centric Pipeline for Archiving Scholarly Artifacts • Information Extraction • Information Retrieval <ul style="list-style-type: none"> ○ Scientific Claims Characterization for Claim-Based Analysis in Digital Libraries ○ Automatic Segmentation and Semantic Annotation of Verbose Queries in Digital Library 	
Program a další podrobnější informace	http://www.tpd.eu/tpdl2018/program-overview/	
Přivezené materiály	-----	
Datum předložení zprávy	23. září 2018	
Podpis předkladatele zprávy		
Podpis nadřízeného	Datum:	Podpis:
Vloženo na Intranet	Datum:	Podpis:
Přijato v mezinárodním oddělení	Datum:	Podpis:

Zpráva je pracovníkem do mezinárodního oddělení předložena nejpozději při vyúčtování cesty do 2 týdnů po jejím ukončení. Bez cestovní zprávy nebude provedeno vyúčtování. Při výjezdu více pracovníků na tutéž služební cestu s týmž programem lze odevzdat společnou cestovní zprávu.

Anotace jednotlivých aktivit :

1. Linked Data Generation from Digital Libraries

Knowledge acquisition, modeling and publishing are important in digital libraries with large heterogeneous data sources. The process of extracting, structuring, and organizing knowledge from one or multiple data sources is required to construct knowledge-intensive systems and services for the Semantic Web. This way, the processing of large and originally semantically heterogeneous data sources is enabled and new knowledge is captured. Thus, offering existing data as Linked Data increases its shareability, extensibility and reusability. However, using Linking Data, as a means to represent knowledge, has proven to be easier said than done!

2. Europeana hands-on session

The Europeana REST API allows you to build applications that use the wealth of Europeana collections drawn from the major libraries, museums, archives, and galleries across Europe. The Europeana collections contain over 54 million cultural heritage items, from books and paintings to 3D objects and audiovisual material, that celebrate over 3,500 cultural institutions across Europe.

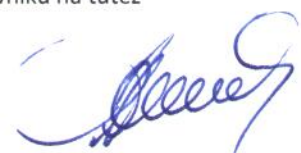
3. Open Science in a Connected Society

Open science comes on the heels of the fourth paradigm of science, which is based on data-intensive scientific discovery, and represents a new paradigm shift, affecting the entire research lifecycle and all aspects of science execution, collaboration, communication, innovation. From supporting and using (big) data infrastructures for data archiving and analysis, to continuously sharing with peers all types of research results at any stage of the research endeavor and to communicating them to the broad public or commercial audiences, openness moves science away from being a concern exclusively of researchers and research performing organisations and brings it to center stage of our connected society, requiring the engagement of a much wider range of stakeholders: digital and research infrastructures, policy decision makers, funders, industry, and the public itself.

4. Towards better Understanding Researcher Strategies in Cross-lingual Event Analytics

With an increasing amount of information on globally important events, there is a growing demand for efficient analytics of multilingual event-centric information. Such analytics is particularly challenging due to the large amount of content, the event dynamics and the language barrier. Although memory institutions increasingly collect event-centric Web content in different languages, very little is known about the strategies of researchers who conduct analytics of such content. In this paper we present researchers' strategies for the content, method and feature selection in the context of cross-lingual event-centric analytics observed in two case studies on multilingual Wikipedia. We discuss the influence factors for these strategies, the findings enabled by the adopted methods along with the current limitations and provide recommendations for services supporting researchers in cross-lingual event-centric analytics.

Zpráva je pracovníkem do mezinárodního oddělení předložena nejpozději při vyúčtování cesty do 2 týdnů po jejím ukončení. Bez cestovní zprávy nebude provedeno vyúčtování. Při výjezdu více pracovníků na tutéž služební cestu s týmž programem lze odevzdat společnou cestovní zprávu.



5. Adding words to Manuscripts: from PagesXML to TEITOK

Library digitalization projects almost always use a page-driven file format for the description of manuscript transcriptions. But for a searchable corpus, a text-driven file format such as TEI/XML is much more appropriate. This article shows how the TEITOK corpus framework provides a two-stage approach, dealing first with transcription in a page-driven manner, and afterwards converting losslessly to a text-driven format, leading to a fully searchable corpus closely linked to the manuscript images.

6. A Web-Centric Pipeline for Archiving Scholarly Artifacts

Scholars are increasingly using a wide variety of online portals to conduct aspects of their research and to convey research results. These portals exist outside of the established scholarly publishing system and can be dedicated to scholarly use, such as myexperiment.org, or general purpose, such as GitHub and SlideShare. The combination of productivity features and global exposure offered by these portals is attractive to researchers and they happily deposit scholarly artifacts there. Most often, institutions are not even aware of the existence of these artifacts created by their researchers. More importantly, no infrastructure exists to systematically and comprehensively archive them, and the platforms that host them rarely provide archival guarantees; many times quite the opposite.

7. Finding Person Relations in Image Data of News Collections in the Internet Archive

The multimedia content in the World Wide Web is rapidly growing and contains valuable information for many applications in different domains. For this reason, the Internet Archive initiative has been gathering billions of time-versioned web pages since the mid-nineties. However, the huge amount of data is rarely labeled with appropriate metadata and automatic approaches are required to enable semantic search. Normally, the textual content of the Internet Archive is used to extract entities and their possible relations across domains such as politics and entertainment, whereas image and video content is usually neglected. In this paper, we introduce a system for person recognition in image content of the Internet Archive. Thus, the system complements entity recognition in text and allows researchers and analysts to track media coverage and relations of persons more precisely. Based on a deep learning face recognition approach, we suggest a system that automatically detects persons of interest and gathers sample material, which is subsequently used to identify them in the image data of the Internet Archive. We evaluate the performance of the face recognition system on an appropriate standard benchmark dataset and demonstrate the feasibility of the approach with some use cases. Scientific Claims Characterization for Claim-Based Analysis in Digital Libraries

8. Scientific Claims Characterization for Claim-Based Analysis in Digital Libraries

In this paper, we promote the idea of automatic semantic characterization of scientific claims to explore entity-entity relationships in Digital collections; our proposed approach aims at alleviating time-consuming analysis of query results when the information need is not just one document but an overview over a set of documents. With the semantic characterization, we

Zpráva je pracovníkem do mezinárodního oddělení předložena nejpozději při vyúčtování cesty do 2 týdnů po jejím ukončení. Bez cestovní zprávy nebude provedeno vyúčtování. Při výjezdu více pracovníků na tutéž služební cestu s týmž programem lze odevzdat společnou cestovní zprávu.



propose to find what we called “dominant” claims and rely on two core properties: the consensual support of a claim in the light of the collection’s previous knowledge, and the authors’ assertiveness of the language used when expressing it. We will discuss useful features to efficiently capture these two core properties and formalize the idea of finding “dominant” claims by relying on Pareto dominance. We demonstrate the effectiveness of our method regarding quality by a practical evaluation using a real-world document collection from the medical domain to show the potential of our approach.

9. Automatic Segmentation and Semantic Annotation of Verbose Queries in Digital Library

In this paper, we propose a system for automatic segmentation and semantic annotation of verbose queries with predefined metadata fields. The problem of generating optimal segmentation has been modeled as a simulated annealing problem with proposed solution cost function and neighborhood function. The annotation problem has been modeled as a sequence labeling problem and has been implemented with Hidden Markov Model (HMM). Component-wise and holistic evaluation of the system have been performed using gold standard annotation developed over query log collected from National Digital Library (NDLI). In component-wise evaluation, the segmentation module yields 82% F1 and the annotation module performs with 56% accuracy. In holistic evaluation, the F1 of the system has been obtained to be 33%.

Zpráva je pracovníkem do mezinárodního oddělení předložena nejpozději při vyúčtování cesty do 2 týdnů po jejím ukončení. Bez cestovní zprávy nebude provedeno vyúčtování. Při výjezdu více pracovníků na tutéž služební cestu s týměž programem lze odevzdat společnou cestovní zprávu.

