Zpráva ze zahraniční služební cesty

Jméno a příjmení účastníka cesty	**Rudolf Kreibich**	
Pracoviště – dle organizační struktury	8.1 ODF	
Pracoviště – zařazení		
Důvod cesty	**Konference LAWA**	
Místo – město	**Paříž**	
Místo – země	**Francie**	
Datum (od-do)	**14-16.11.2011**	
Podrobný časový harmonogram	14.11 let Praha > Paříž 15.11. účast na LAWA konferenci 16.11 odlet Paříž > Praha	
Spolucestující z NK		
Finanční zajištění	VZ136+VZ137+rozpočet NK	
Cíle cesty	Účast na konferenci LAWA	
Plnění cílů cesty (konkrétně)	cíle naplněny	
Program a další podrobnější informace	viz níže	
Přivezené materiály		
Datum předložení zprávy		
Podpis předkladatele zprávy		
Podpis nadřízeného	Datum:	Podpis:
Vloženo na Intranet	Datum:	Podpis:
Přijato v mezinárodním oddělení	Datum:	Podpis:

RDF Ontologie – Pierre Senellart (Telecom ParisTech, Webdam project, aplikace Paris)
Současné ontologie (dbpedia, freebase, yaho, uniprot) mají podobné či překrývající se data. Je snaha ontologie normalizovat na úrovni entit i vztahů mezi nimi a vytvořit tak kombinovanou ontologie z více bází. Mapování se provádí pomocí pravděpodobnostního modelu.

Experiment s propojením YAGO + Dbpedie se stále optimalizuje. Problém s dočasnými vztahy např. prezidenství. Ontologie se neustále vyvíjí, jak pracovat s dynamickým obsahem.

http://webdam.inria.fr/wordpress/
http://webdam.inria.fr/wordpress/?p=893
http://wiki.dbpedia.org/About
http://www.mpi-inf.mpg.de/yago-naga/yago/

Pokud budem někdy provádět seménatickou analýzu textů, bude dobré zkontrolovat současné ontologie např. na bázi wikipedie či NTK PSH.

INTERNET MEMORY FOUNDATION

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.

http://internetmemory.org/en/
Prezentovali hlavně infrastrukturu: z 200TBS na 1 PB = large-scale crawling – HDFS/HBASE
Sociální sítě sklízejí ve spolupráci s Hanzo Archives
řeší vizualizace běhu hadoop.

DATA PRIVACY OF SEARCH QUERIES, Vicenc Torra
Problematika nabízení search-query dat. Zatím u nás jen otázka, jak lidé u nás hledají v archivu, zatím search logy jiných institucí nearchivujeme.
Základ je anonymizovat uživatele. Ovšem otázka jak toto řešit na sociálních sítích. Zpřístupnit už anonymizovaná data, přičemž uchovat shluk hledání, tj. aby se nevědělo kdo hledal, ale že někdo hledat to a to.
Stejně postupovat při vztazích na sociálních sítích.

FOREST, Marilena Oita
a Pierre Senellart
Otázka jak se vypořádat s obrovským množství uživateli generovaným obsahem (blogy, tweety, zprávičky, wiki články, diskuzní příspěvky atd) a jak v něm najít to podstatné.
vytvořit WEB CHANNEL: agregované feedy z různých (i sociálních) sítí.
Standardizovaná struktura: Titul, popis, čas vydání
URL:TITLE:DESCRIPTION
Jednotlivé uzly (nodes) shlukovány pomocí tag paths (Strukturální vlastnost feedu). Feedy analyzovat na základě překvapivosti a statistické sémantické hustoty. Neb co je překvapivé s sémantické husté, tak je relevantní.
Stejné aplikovat i na samostatné stránky, tj. nelimitovat se jen na feedy.
Tj. kurátoři již při sklizni, nikoliv lidsky v post procesu.

TEXT, ENTITY, TIME ANALYSIS – Marc Spaniol
Webový archivy jsou zlatým dolem pro časovou analýzu: veřejená prohášení, výrobky, instituce a společnosti, technologie, vývoj pojmů a frází.
Uživatelé: Sociologové, politologové, market výzkumníci.

Identifikovat relevantní události a intervaly. Získat pojmy a fráze. Identifikovat, odlišit a sledovat entity. Shlukovat informace podle důležitosti. Postupně analyzovat data na větší úrovni, až po celý web.

Problém jak odlišit jména. – se řeší hlavně předchozím zaznamenaným vztahem. Oni řeší pomocí graphu a dotazování ontologie. Stejně tak řeší i časový údaj o události. Byli schopní z věty strojově zjistit že se jedná o hráče, který hraje v určitém klubu a že v nějaký čas odešel. Tj. namapovali událost do ontologie.

http://www.lawa-project.eu/index.php/news/aida_an_online_tool_for_accurate_disambiguation_of_named_entities_in_text_a
aplikace: http://www.mpi-inf.mpg.de/yago-naga/aida/

YAHOO RESEARCH: TIME EXPLORER

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.

Zajímali se o hledání na základě času, tj. hledání textů s prognózami.
Jejich aplikace Time Explorer: http://fbmya01.barcelonamedia.org:8080/future – žel mi to nikdy neběžela..
Diversity aware hledání: zdroj, místo, čas, jazyk, nároz, rank. Pozná entity i vztahy.
Zdroje jako NY times, málo diverzifikovaný na rozdíl od webarchivních kolekcí.
Workflow: výtah entit, TimeML, Sentiment, ontologické mapování na YAGO, anlýza obrázů (sentiment, rozpoznání tváře).
Solr: na úrovni vět a dokumentů.
Jen pro anglické texty.
Objevli např. Vztah mezi pojmy války v Jugoslávii a Husajnem – tehdy se o Mileševičovy mluvilo ve spojitosti s Husajnem.

Degeneracy based community evaluation, Michalis Vazirgiannis

WWW je graph, Společnské sítě a citační indexy jsou vztahové graphy. Takové vztahy mohou být propojené (WWW) či potvrzené (sítě důvery, sociální sítě atd.). Neustále se mění tvarem i velikostí.
Hledání komunit a jejich vyhodnocování v graphech. Různé přístupy (huby, autority, centrálita, mezičlínovost, strukturální podobnost apod.)

Řešili k-core analýzu, rankovali autory v citacích.. Např. knihy co měli 119 spoluautorů.
http://graphdegeneracy.org/

UK Webarchive
Chybí jím kontext k obrázkům
Chtěli by profilovat webové stránky, jestli odpovídá kurátorskému profilu.
Používají geografické entity (PSČ), ale mají jich 15 miliónů.
Vědí co mají, ale nevědí jestli je to užitečné.

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.