Zpráva ze zahraniční služební cesty

Jméno účastníka cesty	**Mgr. Jan Hutař**
Pracoviště – instituce, adresa	Národní knihovna ČR, Klementinum 190, Praha 1, 110 00
Pracoviště – zařazení	Odbor digitální ochrany 2.3
Důvod cesty	- účast na konferenci iPRES 2010 - konzultace k LTP systému s odborníky ze světových knihoven - navnímání nových trendů
Místo – město	**Vídeň**
Místo – země	**Rakousko**
Datum (od-do)	**18.-22. září 2010**
Podrobný časový harmonogram	– 18.9. cesta vlakem do Vídně – 19.9. účast na workshopu – 20.-22.9. účast na konferenci – 22.9 návrat vlakem do Prahy
Spolucestující z NK	Mgr. Marek Melichar (19.-22.9)
Finanční zajištění	- VaV.INST.0131
Cíle cesty	- viz důvod cesty
Plnění cílů cesty	- účast na konferenci - jednotlivé konzultace s odborníky
Program a další podrobnější informace	viz níže
Přivezené materiály	
Datum předložení zprávy	22.10. 2010
Podpis předkladatele zprávy	

19.9

tutorial T3: Logical and bit-stream preservation integrated digital preservation using Plato and EPrints

Adam Field/ David Tarrant – eprints southampton http://www.eprints.org/

Francouzský LTP systém SPAR nemá zatím funkční preservation planning, zatím jen v plánu
PLATO – na vytvoření preservation plánu
eprints- dobject management sw

preservation plan- jak vypadá – plato udělali vzor – 10 částí
trigger for re-evaluation > sledovat nové metody, sledovat nové nástroje

TRAC je v procesu ISO, takže bude brzo normou!

- eprints – hybrid storage
- use the best features of each storage type
- local delivery from the cloud
- je tam storage controler – plug-iny na různé typy HW – řídí kam dokumenty jdou – location of files na disky, pásky apod. (sun cloud, amazon S3, local discs) – tj. to samé jako tessella (nabízí to i rosetta? v budoucnosti by to mohlo být důležité!)

- každý cloud posílá zpátky po uložení kontrolní součet a pokud nesouhlasí s tím, co se tam poslalo, není to uloženo dobře
- v případě vypnutí storage, lze to seamlessly přehazovat

droid – validace formátů
pronom – risk analysis – ve formě score v eprints preserv2.googlecode.com, klikne se na riskantní files a lze vidět možnost uploadu preser. plan, preserv. action atd., ke stažení možnosti 5 filů bez risku toho původního filu
otázky typu -je tady lepší formát k dispozici? plato

- eprints dělá validace formátů, risk analysis
- drží object revisions
- po 6 měsících se to může z lokálního disku odlévat na cloud, droid vidí i do zazipovaných files, takže no problem
- do eprints lze nahrát preservation plan pro soubory, kt. mají risky (preservation plan musí vzniknout v nástroji PLATO – viz níže)

preservation plan vytváření- PLATO
- vše ve slidech[1]
- je potřeba držet kontext, proč to vzniklo, insitutional settings, constraints, user community

triggers- proč vzniká nový PP?
- nová kolekce
- změna profilu kolekce
- změna prostředí
- periodic review apod.

file format identification –pronom, droid
http://p2-registry.ecs.soton.ac.uk/

choose sample object – zatím ručně v plato, eprints má pravidla, plato vyvíjí nástroj, kt. proběhne celou kolekci a vybere vhodné dokumenty do vzorku pro vytvoření PP

1. krok – identify requirements, to do stromové struktury, high level goals > drobnosti, nebo brainstorming stakeholderů
    a. object charakt.
    b. record charakteristics
    c. process char.
    d. costs

requirements musí být měřitelné!

v platu nejsou placené nástroje, lze je ale zapojit, pokud si stáhneme plato na lokální PC
minimee – tool na migraci, validaci a porovnání výsledků
- CRIB stále funguje a je v rámci PLATO
- plato dělá porovnání toolů na konci

2. evaluate experiment

[1] eprints – hledat preservation – jsou tam ty materiály, eprints wiki, prezentace z tutorialu viz http://files.eprints.org/581/

- posouzení výsledků vzhledem k nárokům
- porovnání nástrojů a přehled všech nastavení a našeho ohodnocení je pak v analysis části (analyse results)
- části o cost se musí spočítat manuálně, vychází z LIFE projektu
- na konci plánu je vidět vše, včetně doporučených nástrojů apod. , plán lze vytisknout, uložit, exportovat

- **uploadne se tam ten plán z plato a pokud je tam nějaká akce, tak se to na pozadí udělá samo v eprints > tj. proběhne migrace, originální soubor se zachová!**
- uživatel pak vidí aktuální verzi i odkaz na originál

OPM – open provenance model – to má v sobě – bude standardem za chvíli

20.9. pondělí

keynote – project SCAPE?
project trident – research microsoft
http://www.microsoft.com/windowsazure/

Markus Enders - METS based information package for preserving web archives
http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/enders-70.pdf

- heritrix, web curator vytvářejí SIPy a DIPy v různých formátech (single vs. container),
- popis v dc, warc pro jedno nebo více url

common model for infomation package format chtějí

- základem data modelu je website (z jedné nebo více webpages), one AIP per website,
- webpage má asociované objekty – obrázky např.
- vypadá to, že dělají migrace aipů, např. webpage do pdf
-

BL – ukládají data a metadate na repozitáři – metadata database (pro operační metadata) – pro long-term preservation metadata, v METSu
AIP – zip nebo warc+ METS descriptor
ukládají preservation metadata – události ještě před ingestem, vztahy mezi bytestreamy v případě migrací

metadata
website, webpage a asoc. objekty jsou definované v structMAP as nested elementy
descr. metadata on website level – DC, MODS vytvořený z DC
rights metadata – proprietary schema

file definitions – 3 file groups – digital manifestation (all content files), logfile (logy z crawleru), viralfiels (soubory s virama, v aip, ale zvlášť, nelze zpřístupnit)

- mets potom může mít třeba 120mb pro 800MB warc!! 2MB warc odpovídá 2MB mets > to se jim zdálo moc > make it simple, nedefinujte objekty, o kt. nemame co říct, pokud lze extrahovat nějakou vlastnost později, nedělejme to dnes
- virus check na container level> není potřeba <file> element v premisu
- vyjímečně na urovni souboru

struktura warcu – lze udělat později
necpat logy do mets, pokud jsou metadata ve warcu, není nutné je extrahovat nyní

BL má web s mets profily, bude to tam vyvěšené ten profil pro WA

Maurizio Lunghi, Angela Di Iorio - Relevant Metadata to Preserve "Alien" AIP

archives ready to the aips transmission ARTAT
jak uložit cizí aip – musí být standard na přenos metadat a poskytnout tak možnost uložení pro třetí strany v jakémkoliv repozitáři
PML – preservation metadata layer
mpeg21 part 2 – metadata container, používají v KB NL

- jak vyměňovat informace mezi repozitářem kt. používá mets a druhým, kt. používá MPEG21

k aip se pripoji pml, které ho popisuje (pml core a pml redundant – překlad semantiky metadat aip do premisu)

Session 2 (Panel) Preserving Web Archives: One Size Fits All?

NLAustrálie- ochrana jen bit level, spíš počítají s emulací
LC – preserve the bits and do some preservation action

- dělat validace, charakterizace je časově náročné
- ke každé websitě dělají katalog. záznam
-

HUL – připrava na ochranu, hned po sklizení přípravné kroky – planning

- mají k tomu základní metadata v premisu
- zabaleno v metsu v repozitáři> získat od nich popisy co vlastně dělají?

BNF- zatím nic – dávají to do LTP

- technology watch
- emulace?
- budou ukládat arcy, bit stream, pak až budou potřebovat, tak je vytáhnou, udělají z nich WARCy,
- vyrábějí plug-in do JHOVE2 aby viděl do arku a mohl udělat charakterizaci a validaci formátů
- zatím neví jak s migracemi, ale počítají s nimi
- PRONOM bude v RDF

CZ -
NZ – neřeší, zatím do LTP rozhodně dávat nebudou? jak to mají s tím web curatorem?dělají charakterizaci filů v arcu pro výběrové sklizně z curatoru, ale trvá to dlouho plus nevidí žádný přínos

FIN – bit level zatím, žádná migrace, žádná emulace

Amy Kirchhoff, Eileen Fenton, Stephanie Orphan, Sheila Morrissey
Becoming a certified trustworthy digital repository: The Portico experience

mají první TRAC audit provedený třetí stranou!
timescale

- march 2009 se o auditu dozvěděli
- april – poslali self report
- june – dostali materiály
- aug- poslali první balík – organizační a dalších 4 packety, pak je navštívila CRL – jeden den
- oct- additional documents
- dec –CRL poslalo draft zprávy

komentáře
january 2010 oznámení o certifikaci a finální report

documentation je jako důkaz o auditu
vzniká dokument o nárocích na auditory – musí být certifikovaní – CCSDS a MOIMS-RAC working group na tom dělají
Requirements for bodies ….

Paul Conway
Measuring Content Quality in a Preservation Repository: HathiTrust and Large-Scale Book Digitization

http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/conway-20.pdf

21.9. úterý

dave tarrant
Connecting preservation planning and Plato with digital repository interfaces – viz tutorial
viz odkaz nahoře, je tam web, kde jsou k dispozici materiály k trainingu, který probíhal po UK – 5 oblastí školení pro provozovatele repozitářů – najít a použít v NK?

- evaluation of bit preservation strategies
- jak vybrat vhodnou strategii?
- based on planets a na strategii pro bit repository v dánsku
- jde o výběr řešení
- bit safety, costs, availability, confidentiality
- pillar layer a nad tím něco co konroluje ty bity

měření bit safety je komplexní

- bit repository requirements measuring systém

plato v tom hraje roli, kt. se netýká hw, ale těch dokumentů, kt. se na tom mají ukládat
to pak udělají pro každý pillar – pásky, dvd, cloud, hdd apod.
SLA – service level agreement pro každý pillar a pro každý requirement

Peter McKinney
Preservation Planning: A Comparison between Two Implementations
porovnání plato a plánování na NZ
mají v rosettě 14TB dat, any format any time, 72 know formats, 8945 unknownformats inside
plan

- business view- risk> solution>pplan
- systém view – plan to refine
- NZ záměrně nedělá srovnání nástrojů a výsledků pomocí hodnot a tabulek, váh apod., ale ručně – vizuálně – je tam mnoho rolí lidí, kt. se k tomu musí vyjádřit, zbytek je podobný jako plato
- jeden z důvodů je, že nechtějí vyrábět rozsáhlé plány, ale musí jednat „rychle“

David Pcolar
Chronopolis and MetaArchive: Preservation Cooperation

- chronopolis a metaarchive a university of north texas (16 institucí používajích lockss – dohromady 260TB dat) – jak pracovat dohromady?
- jak sdílet objekty? – různé infrastruktury, operační procesy etc.
- bagit – testy pro základní data transfer, verifikace pomocí kontr. součtů

22.9. středa

Louise Fauduet, Sébastien Peyrard
RDF as a Data Management Strategy in a Preservation Context

- modulární systém založený na oais
- preingest mimo oais
- mají oddělení katalogizace a metadat
- nemají preservation modul
- 3 SLA pro každou kolekci – ingest, ochrana a disemination jsou pomocí toho formalizovány
- jaké formáty jsou dovoleny
- jaká je max. velikost SIPu pro ingest např.

- všechno na papíře, žádné dohadování kdo co! lidi z ltp vs. ostatní
- rdf exportuji z data managementu (je tam mets, sla aj. – struktur. metadata v balíku)
- to co je v rdf zachová strukturu ale zároveň lze různě použít, kombinovat, využít části apod.
- ontologie pro různé entity – např. pro provenance obsahuje digitalizaci, id generation, ocr apod, k tomu přistupují data, agenti, vztahy

SPAR at BnF – rozhovor Sébastien Peyrard

- pustili to v květnu 2010
- zatím nemají preservation planning
- pouze základní OAIS funkce
- zatím jedno workflow pro digitalizované věci, později letos přidají i video (mpeg7 metadata)
- vše na páskách, metadata pak ještě pro případ havárie na discích
- pracovní prostory na discích
- uvnitř to mají v metsu (DC, premis, mix, mpeg7)
- dělá to pro ně někdo externě, v knihovně asi 7 lidí, kteří říkají co to má umět – business analytici
- původně to chtěli na fedoře ale nešlo to, udělali to celé sami – nechali si jen ingest a access částečně
- mají tam i modul pro content management a tato část je pro ně hodně důležitá, takže je to takový hybrid
- možná to bude open source,musí rozhodnout vedení knihovny
- využívá open source nástroje

Hilde van Wijngaarden
Building blocks for the new KB e-Depot
edepot

- KB – nová priorita – digitální knihovna, přístup pro kohokoliv čehokoliv, dlouhodobá ochrana

- kontrakt s googlem z toho vychází samozřejmě
- výstup ltp group – popis digital library services

- **různé úrovně ochrany pro různé kolekce**
- levels for checks, metadata, storage and preservation
- úroven ochrany závisí na riscích, hodnotě dokumentu apod.
- chtějí do nového systému dávat ejournals, ebook, zdigitalizované věci a WA
- flexible modular systém, avoid vendor lock-in
- **start small and expand**
- podzim 2010 procurement for separate components, 2011 development
- 2012 migration

vytvářejí data model a xml schemata na jeho podporu
řeší s každým oddělením co jim digitalizace a dig. knihovna přinese, co chtějí aby jim přinesla, co pro ně bude znamenat apod.

nebudou na google datech provádět dlouhodobou ochranu, jen základní ochranu – kvalita od googlu není vysoká