Zpráva ze zahraniční služební cesty

Jméno účastníka cesty	**Jan Hutař**
Pracoviště – instituce, adresa	ODO
Pracoviště – zařazení	1.5
Důvod cesty	**návštěva Královské knihovny v Haagu, seznámení s procesem digitalizace a LTP systémem DIAS**
Místo – město	**Haag**
Místo – země	**Nizozemí**
Datum (od-do)	**10-12.6.2009**
Podrobný časový harmonogram	10.6. – odpoledne příjezd do Haagu 11.6. – návštěva KB 12.6. – návštěva KB, odjezd do Prahy
Spolucestující z NK	Jiří Polišenský, Tomáš Foltýn
Finanční zajištění	IOP-NDK
Cíle cesty	seznámení s procesem digitalizace a LTP systémem DIAS
Plnění cílů cesty	splněno
Program a další podrobnější informace	viz níže
Přivezené materiály	
Tištěné přílohy a elektronické dokumenty	
Datum předložení zprávy	20.6.2009
Podpis předkladatele zprávy	

11.-12.6.2009

2003 edepot začal fungovat
edepot byl součástí nového oddělení R&D, kam spadá DP

- oddělení začínalo se 7 lidmi, dnes okolo 100 (různé projekty)
- 1/3 lidí jsou staff KB aby si udržela knihovna ty znalosti (zbytek je najmutý na projekty)

provoz edepotu je dnes v rámci odd. akvizice a provozu

Národní vs. mezinárodní edepot

- mezinárodní je původní, teď se začínají soustředit na domácí dokumenty (vláda, univerzity apod.)
- webarchiv – nedělají zatím kompletní sklizně, začínají celkově, sklízejí 500 webů – kriteria výběru podle normálních pravidel akvizice pro vědecké publikace
- edepot bude archivem pro národní program digitalizace
- 8mil Euro pro projekt Metamorfosis – pro všechy instituce ALM (mají na webu kriteria pro digitalizaci)
- přístup do edepotu je jen onsite v případě licencovaných publikací

METS MODS MIX ;-)

archivace ekonomických publikací evrospské společnosti Nereus

Národní knihovna Číny momentálně kopíruje všechny postupy KB a bude je využívat u sebe

Barbara Dorren, Jos Pijpers (techničtí analytici edepotu)

new edepot – pro digitalizaci, do starého edepotu (DIASu) digitalizované dokumenty dávat nebudou, počkají si na nový systém 2011/2012

je jedna lokace, data centrum 2 zálohy v budově (disky) – v různých částech a patrech KB

- migrace ještě žádné neprovedli
- 2 lokace Hilversum –národní data center (TV, rozhlas apod.)– záloha na páskách
- k managementu záloh používají TSM

NCDP- national coalition for DP

najít projekt Peprs

- mají upsky, diesel generátory jako záloha
- dvojitá kontrola klimatu (1 se rozbije, druhá jede) 21°C
- IBM HW je méně tolerantní na změny teplot v provozu směrem nahoru
- FM200 plyn jako zhášedlo

síť

- gigabit ethernet připojen k edepotu
- stejné prostředí jako všechny ostatní systémy KB
- FC na SAN + online backup
- Surfnet – poskytovatel sítě (přes lightpath)

servery

- P570, 520, IBM servery
- používají virtualizaci ibm (LPAR) i vmware

storage

- mají DS6800 – 2TB jen na provoz systému
- obsah edepotu není uložen na discích ale na optických mediích (plasmon G638 – už překonaná technologie) – optická knihovna – nelze z toho vymazávat
- optická media kodak, vydrží až 50 let reálně, což nic neznamená, protože okolí SW se stejně mění
- LT media nic neřeší
- od roku 2012 uvažují o fast LTA technologii Silent Cube (12 disků – lze nastavit rychlost spinu > šetří energii, když dokument nikdo nechce, nehýbe se to ;-) + se prodlužuje životnost , umí kontrolovat checksumy

BackUp

- TSM – součást edepotu (DIASU) – používají to pro backup všech systémů
  - obsah DIASu
  - LTP data
  - ostatní systémy
  - obyč. klient + tivoli data protector
- DB2

přístup

- Edepot byl původně dark archiv- pak změna, chtěli přístup > nasazeny webseal servery /windows + TSM access servery
- to vysvětluje proč dias vypadá jak vypadá, access nebyl na začátku v plánu
- pro články v pdf nemají user copy
- pro digitalizované dokumenty pak ano
- user copies jsou mimo DIAS v jiném úložišti
- v DIASu je cash vrstva – občas se musí promazat

- data z masové digitalizace tam dávat nebudou, počkají na nový systém, zatím je budou držet vedle

DIAS v. 1.3
existuje už v.2 (DNB) ale v KB ji nechtějí, protože to stejně nesplňuje jejich požadavky

- musí jet na AIX (operační systém IBM) > tj. HW musí být také IBM
- testy na jiném HW – powerPC
- na dias core není ovšem na jiném HW support!!
- ibm je velká společnost, dias není její prioritou!
- navíc dias jako takový je poskládán z komponentů IBM, z nichž některé už IBM dále nevyužívá a není k nim tedy podpora ani interně v IBM

KB si vytvořilo vlastní ingest modul – pre processing (PEARL)

- digitalizace – data a metadata mimo edepot (jsou v TSM archivu)
- KB nevyužívá kolibri
- mají spec. systém na zpřístupnění CD (od IBM)
- IBM preservation manager (nová součást Diasu) zatím nevyužívají a neví jestli budou
- mají tam zatím jen pdfka
- rozdíly mezi 1.3 a verzí 2
    - nejsou markantní
    - 1.3 má support pouze v KB
- závislost na hw ibm je jeden z důvodů, proč musí dias z kb pryč + to že nevyhovuje, už je to starý sw a kb má dnes jiné nároky než před lety

security

- TAM WebSeal frontend
- RTM. real time monitoring
- ITIL – change management

zajímavost

- v r. 2006 kb vyhodila všechen HW na kterém dobíhala maintenance, spočítali si, že provozování starého hw (vyžaduje dodatečné smlouvy o maintenance) je v konečném důsledku dražší než pořízení nového včetně maintenance
- všechy pásky byly z důvodu bezpečnosti sešrotovány /rozdrceny

---

Marcel Ras – general workflow

odd. akvizice/ odd. katalogizace a metadata managementu/ odd. edepotu

harvestují nově data (publikace) z univerzit

- v katalogu mají 13mil .článků
- manuálně jen tituly periodik jsou vytvářeny v katalogu, články se vytvářejí automaticky vedle v katalogu článků
- load kapacita edepotu je 400.000 článků/den

DP je o organizaci (nejen technice)

na počátku byla 2 nové oddělení (provoz edepotu a R&D odd (digitalizace a DP))
digitalizace je pod R&D odd (80 lidí)

Edepot oddělení

- na ingestu berou všechno (mají ale seznam preferovaných formátů)

- zatím digitaliz. věci mimo edepot- ani to tam dávat nebudou
- mají to včetně bibliografických, technických a strukturálních metadat

DIAS neumí uložit webové stránky
- důvodem je velmi omezená práce s metadaty (v.1.3 má snad jen 15 polí!)- tj. řeší to tak, že mají metadata v databázi mimo dias
- omezení velikosti AIP- max. 16GB
- problém s velkými AIP je i na straně přístupu – nelze v diasu rozdělit aip na více kousků
- ingest je dělán pomocí tzv. batch builderu – vytvořeno pro KB na přání od IBM

- ve smlouvě, kt. se uzavírá s publishery je psáno, že tito nemají vliv na typ migrací v archivu (pozor na AiP Beroun a manuskriptorium)
- v případě migtrací budou ukládat první originální verzi DO, pak tu poslední logicky a ty, kde byla změna DO markantní

Ingest/pre ingest

DIP pro end usera – vidí link na pdfko a metadata- viz web KB, link vede do diasu

dias pošle dip, ale access jako takový nemá
- vydavatel pošle DO jako .zip (pdf + metadata + special files), pošle to do tzv. post office (ftp server)
- zkontroluje se to
- během ingestu spolupráce s katalogem, lze stáhnout záznam z katalogu a zpětně holding DO se objeví v katalogu
- batch builder (ibm sw) dělá virus check, md5 a vytváří SIPy +konverze metadat
- neumí to JHOVE ani charakterizaci (PRONOM apod.)
- v rámci projektu DARE by pronom měl umět
- normalizaci to taky neumí (jen word.doc do pdf), v tomto případě, pokud někdo pošle .doc a oni to změní na pdf, tak uchovají oba soubory (DO)

uživatelské rozhranní
- nemá- jen příkazová řádka, otevřená pro různé části diasu v několika oknech !!
- SIP je .zip, AIP ukládáno také jako .zip/.tar a DIP to samé
- metadata jsou vedle v DB2

v rámci SIPu je kontrola v ingestu
- zda jsou tam data a metadata od vydavatele (ne jako mají formu a zda jsou komplet, jen zda tam vůbec jsou)
- producent se nedozví, zda to proběhlo korektně (žádný feedback)
- kontrola md5 ne, vydavatelé to nedodávají

publisher profiles
- není v diasu ale v lokální bázi

projekt DARE
- digital academic repositories> 100.000 dokumentů bude uloženo do edepotu
- partneři (univerzity) musí respektovat pravidla předávání, metadat a formátů

Barbara – spec. odd. DP bylo původně, dnes je to rozhozené

DIAS neumí PREMIS (bude snad ve verzi 2.3 pro Dassault – příští rok)
- všechny projekty, na kt. se KB podílí musí být ku prospěchu edepotu
- KB právě aktualizuje svou DP policy

- jsou v podobné situaci jako NKP- ohledně digitalizace – nemají DP finální policy, ani formáty ustanovné apod. pro masovou digitalizaci

DARE – 42 file formats přijímají, což ale znamená udělat průzkum ohledně migrací apod. pro DP pro všechny z nich…

Dias nepodporuje verzování DO a velmi špatná metadata

preservation manager – zatím to nemají, ještě uvidí jestli to k něčemu bude (view paths)

- je to mimo dias – ve verzi 2 je to dohromady, není to ale integrované
- vzniklo to na návrh KB v r. 2008- zatím nedodáno

v diasu není lokální knihovna formátů (Němci mají svoji vlastní a to jen proto, že to nepracuje s Pronomem)

v KB vzniká analýza PREMISu

- na základě toho, co chtějí těmi metadaty podporovat z procesy v systému (funkcionalitu)
- důležité hledisko je „jak ten který element vytvořit“ (půjde to lehce nebo ne?) a nač jej použít

DIAS 1.3 má interní metadatový formát s 15 elementy (naprosto nedostačující), verze 2 (DNB) má nový data model

verze 2 pro DNB vznikla bez jakýchkoliv kontaktů a konzultací s KB – tj. evidentně to, že stejný systém používají 2 zákazníci neznamená pro IBM, že jeho vývoj bude konzultovat s oběma ;-) KB byli docela v šoku

- verze 2 má navíc vylepšené migrace tools (i přesto nelze změnit např. část metadat v aip aniž by se musel migrovat i datový objekt a reingestovat)

KB chce nový systém outsorcovat jako službu – nadiktuje podmínky a hotovo, toto poskytne velkou flexibilitu a ušetří lidi

ovládání a monitoring diasu je nedostatečné (nemá to GUI)

KB se chce IBM v budoucnu vyhnout právě pro jejich způsob práce

---

Trudie – National Edepot
webform – dias nemá, vytvořili si sami
customized ingest workflow pro určité producenty- dias neumí, rosetta ano

---

Hilde
dias se rozjel v 2002

- doprovázeno rozsáhlou interní reogranizací (edepot odd. v rámci akvizice vzniklo) + R&D oddělení vzniklo
- o systému začali přemýšlet už v r. 1992
- 1997 proof of concept s IBM

proč se rozhodli pro IBM – nabídla vývoj, jako jediný tehdy z uchazečů přiznali, že o tom nic neví, ale že to zkusí

stavěli DIAS na články z Elsevieru a na CDčka (batch builder 2003)

- 2002 skupina pro DP (KB, IBM + odborníci)
- od počátku bylo jasné, že v diasu není DP funkcionalita a IBM to ani neslibovala

- KB chtěla DP funkcionalitu později (Files format registry zapojení, UVC apod.) – dodnes to nemají
- preservation manager se testuje
- nelze jednoduše (bez migrace DO) měnit metadata v AIP, to je důvod proč mají v diasu minimum metadat v aip
- je blbost kvůli změně metadat migrovat a re-ingestovat DO – v případě statisíce změn…
- v posledních letech je patrné, že IBM totálně ztratila ponětí o novinkách v DP, KB jim to sama musela vysvětlovat (sun, exlibris neustále na konferencích, ibm nikdy)
- DIAS nelze outsourcovat hw
- dias není vhodný na nic jiného než jsou ečlánky (rozhodně ne na masovou digitalizaci – má špatný ingest)
- chtěli po ibm lepší monitoring a ingest – nikdy k tomu nedošlo
- problémem je, že nelze udělat vlastní moduly, protože to celé visí na AIX

- je důležité si uvědomit, že nejpodstatnější částí systému LTP je ingest

- requirements na dias jsou na webu KB (z r. 2000)

Judith – nový LTP systém plány
ingest – od okamžiku doručení PSP od producenta k okamžiku vytvoření AIP
batch builder IBM – jakákoliv malinká změna stojí peníze (KB si dělá mimo to svoje skripty a vytvořila si tak vlastní pre-ingest, čímž vlastně krkolomně obešla IBM)

- BB nemá normalizaci, migraci ani charakterizaci
- nemá role based autorizaci
- není vhodné pro vysoké počty dokumentů

do stávající verze diasu už nechtějí dát ani euro

> nový systém

- starý systém AIP max 2GB
- musí být managable application – stávající dias neumožňuje (It management)
- musí mít roadmap/vývoj
- open standard
- role based authorizace
- protokoly FTP, OAI-PMH, sFTP, FTPS, SWORD
- scheduler na ingest
- nutná funkcionalita vložení z přenosného nosiče (CDčka třeba)
- musí být možné přeskočit některé fáze ingestu
- nakonec ingestu je viruscheck – důvodem je, že se systém během ingestu bude propojovat s externími službami (PRONOM) a tam nikdo nezaručí, že se nedostane do systému (SIP) virus
- roll back scenario musí mít pro různé moduly

kdo vše se na systému podílí a na jeho plánování

- edepot odd.
- DP manager
- IT
- Ingest lidi
- metadata lidi

Judith – analýza Rosetty a Tesselly pro potřeby KB – jen ústně, žádný oficiální dokument nám dát nemůže

Rosetta negativa
- nemá zkušenosti s ejournals (Tessella ano) – pro nás nedůležité
- v Holandsku nemá ExLibris dobrý support – u nás asi naopak
- není možné koupit jen část (buď všechno nebo nic) – což mně ale dává smysl, protože to vše souvisí
- podpora jen Solaris a Linux
- support je jen v Německu
- nejsou si jisti, jestli je R schopná zvládnout hodně dat/den

Rosetta pozitiva
- znají knihovny
- skvělý interface
- umí webarchiving
- znají planets a digital preservation (jsou velmi aktivní)

Tessella
- více flexibilní než Rosetta (asi myšleno v oddělitelnosti modulů)
- nemá tak pěkný interface
- vědí jak na ejournals
- nezávislé na platformě
- znají planets
- mají v tom zabudované dioscuri
- licencing – platí se jen za změny, ovšem to co vytvořili pro BL a TNA je zdarma
- zdrojový kód SW je dostupný (v případě krachu Tesselly apod. je uschován po zakoupení systému u notáře)
- dobrý workflow engine
- není možné vynechat kroky ingestu na základně předchozího kroku
- nejasné, jestli si poradí s velkým množstvím dat/den
- BL rozvázala minulý měsíc spolupráci s Tessellou
- spíše je to pro archivy
-

PTC – windchill

použít pro specifikaci systému
  - systém musí mít vývoj (roadmapu), kt., nebude závislá na zákazníkovi – tj. samostaný vývoj sw
  - systém musí mít základní ingest a access
  - musí mít premis a met
  - firma musí být aktivní na poli DP – skutečně aktivní – mít vlastní vývoj apod., přítomnost na konferencích apod.
  - flexibilita – nezávislost na HW a operačním systému!!! AIX
- musí být managable application – stávající dias neumožňuje (It management)
- musí mít roadmap/vývoj
- open standard
- role based authorizace
- protokoly FTP, OAI-PMH, sFTP, FTPS, SWORD
- scheduler na ingest
- nutná funkcionalita vložení z přenosného nosiče (CDčka třeba)
- musí být možné přeskočit některé fáze ingestu

- nakonec ingestu je viruscheck – důvodem je, že se systém během ingestu bude propojovat s externími službami (PRONOM) a tam nikdo nezaručí, že se nedostane do systému (SIP) virus
- roll back scenario musí mít pro různé moduly
- nutné zvládnout 80tis DO/den na ingestu!!
- nezávislé na platformě
- dostupnost zdrojového kodu
- jak bude probíhat support, přímo z ČR?
- podmínka je běžící systém s digitalizovanými daty v něm kdekoliv na světě