Závěrečná zpráva OR 2019, Hamburk, Německo

Vlastimil Krejčíř, Ústav výpočetní techniky, Masarykova univerzita

Zpráva je rozdělena na část stručnou (1 strana), která jen shrnuje v bodech program konference Open Repositories 2019 (dále jen OR), a podrobnou, která se věnuje jednotlivým přednáškám do větších detailů. Konference byla 4denní od 10. - 13. 6. 2019, přičemž první den byl věnován workshopům, samotná konference pak trvala zbylé 3 dny. Celý program běžel v 5 paralelních sekcích, budu tedy reportovat pouze ty, kterých jsem se zúčastnil, případně doplním informace z jiných sekcí, které jsem získal od jiných účastníků konference.

V dokumentu uvádím bodový přehled důležitých trendů, výběr zajímavých přednášek, obecnou zprávu (použita do cestovního příkazu pro mateřskou instituci, tedy ÚVT MU) a samotnou podrobnou zprávu z jednotlivých přednáškových bloků a na konci i jednotlivých přednášek.

Obecné trendy a o čem se mluvilo na konferenci:

- přechod do cloudu (Amazon AWS)
- virtualizace/kontejnery (Docker)
- řešení škálovatelnosti (růst objemu dat)
- přívětivost uživatelských rozhraní služeb
- DSpace 7
- Oxford Common File Layout standard

Zajímavé přednášky:

- A Multi-Tenancy Cloud-Native Digital Library Platform (Yinlin Chen)
- DSpace Clustering via Puppet, HAProxy and CephFS (Bernd Nicklas, Paul Münch)
- Flexible metadata: the key to a single repository for all types of output (Ben Summers, Tom Renner) – poster
- DSpace 7 – The Power of Configurable Entities (Lieven Droogmans)
- The Challenges and Charms of a Cloud-Based Repository Infrastructure Transition: A Case study from the University of Illinois (Seth Robbins)

Zpráva pro mateřskou instituci

OR je konference primárně zaměřená na oblast digitálních knihoven, repozitářů, open access a publikování.

Letošní zaměření konference se týkalo uživatelských rozhraní a user experience (UX) - jakým způsobem budovat UI a služby, aby byly atraktivní pro koncového uživatele. V tomto směru jsme měli možnost slyšet převážně motivační přednášky a zkušenosti z jiných institucí. Kromě obecných doporučení typu "sledujte jak to dělá Facebook/Twitter" byly i zajímavé přednášky u zkušenostech s průzkumy u koncových uživatelů - zajímavé bylo, že se našly i instituce (i v USA), které tyto průzkumy z důvodu finanční a logistické náročnosti nedělaly (nepovažovaly výsledný přínos za tak velký).

Obecným trendem zejména v USA (v Evropě se tento trend nezdá tak silný) je přechod na externí komerční cloudové služby, nejvíce se skloňoval Amazon (AWS, primárně úložiště S3, ale také využívání celé infrastruktury, včetně integrace s vývojem software např. na GitHubu apod.). Uvedeno bylo několik přednášek, které popisovali technické i administrativní překážky při přecházení do cloudového prostředí Amazonu. Primární motivace pro přechod byla škálovatelnost, zejména s výhledem do budoucna (současný i předpokládaný nárůst objemu dat by se mohl výrazně prodražoval, pokud by se musel řešit v rámci inhouse HW infrastruktury). Nemalou motivací byla i snaha ušetřit, samotný přechod však vždy levnější nebyl (zejména kvůli technickým komplikacím).

Další obecným trendem je kontejnerizace, zejména používání Dockeru a nástrojů pro správu (Ansible, Puppets aj.).

Z mého pohledu byla klíčová série přednášek o DSpace nové verze 7. Po dvou a půl letech vývoje nás čeká řada větších změn, pakliže mohu hodnotit, tak k lepšímu. Řada funkcí, které jsme doposud zajišťovali vlastními silami, by měla být nativně k dispozici.

Osobně jsem v kuloárech diskutoval problém multi-tenancy v prostředí digitálních knihoven - tedy variantu, kdy provoz několika z hlediska uživatele samostatných digitálních knihoven zajišťuje pouze jeden systém. Z projektu DARIAH (spolupráce s Filozofickou fakultou) se podařilo zaplatit i výjezd jedné ze studentek, které v rámci svých bakalářských prací tuto problematiku zkoumají. Měli jsme tak možnost rozdělit síly a zaměřit se na více alternativních řešení, zejména pak na nadstavby systému Fedora: Islandora, Samvera, Hyrax, Hyku. Nelze říct, že by šla většina institucí cestou multi-tenancy systému, například i komerční poskytovatel systému DSpace (SaaS) instaluje pro každou digitální knihovnu separátní server. Zdá se, že právě pro multi-tenancy bude vhodnější některá z nadstaveb systému Fedora.

Řada institucí (i konsorcií institucí) stále vyvíjí na zelené louce vlastní systémy pro digitální knihovny - DSpace, Fedora aj. jsou zde již dlouho a historicky si s sebou vezou i řadu problémových vlastností, které nejsou schopni všichni akceptovat. Byť řada přednášejících i v kuloárech připustila, že vlastní řešení není optimální a prodražuje se. Z tohoto pohledu mě zaujal systém HAPLO, který aktuálně řeší malá komunita ve Velké Británii - obecný systém pro budování digitálních knihoven, který si (údajně) bere z existujících systémů to dobré - vypadá zajímavě, nicméně komunita je zatím malá, pro ostré nasazení příliš riskantní (ale je vhodné jej sledovat).

Poslední dvě zajímavé věci:

* Oxford Common File Layout standard: formát pro dlouhodobé ukládání dat, zatím v beta verzi, nicméně tlačen Oxford University, hodně se o tom mluvilo.

* Plan S - dostat od roku 2021 do Open Access všechno financované státem (koalice řady zejména evropských zemí), proběhla panelová diskuse, měl jsem z toho pocit, že se neví kudy kam...

Závěrečná zpráva z konference s podrobnější sumarizací jednotlivých přednášek, kterou jsem vypracoval pro AKVŠ, by měla být brzy k dispozici na https://akvs.cz/zahranicni-cesty/zpravy-ze-zc/

Workshopy

Workshopy byly zaměřeny na uživatele různých systémů (DSpace, Fedora, Samvera, Islandora, Invenio), na dílčí technologie, projekty a standardy (GLAMpipe, Counter, DataCite DOI, OpenAIRE, Wikibase, CRIS aj.). Navštívil jsem workshopy ***Islandora for All: ISLE Workshop***, ***Getting started with DSpace 7 (advanced)***, ***DataCite DOI Services for Repositories – make all research outputs persistent***.
U Islandory jsme přímo zkoušeli na vlastních počítačích její instalaci, u DSpace jsme zkoušeli nové REST API a použití Angularu. Workshop DataCite DOI byl spíše prezentační a představil firmu DataCite a služby, které poskytuje: přidělování DOI, reporty, napojení na PID Graph (projekt EU Freya, EventData, GraphQL).

Konference

Hlavní téma konference bylo User Experience (UX) – jak představit službu koncovému uživateli tak, aby ji chtěl opravdu využívat a jak udělat uživatelské rozhraní, aby bylo efektivní.

Uvedla to již úvodní motivační přednáška *(Jeff Gothelf) o důležitosti uživatelských rozhraní, které službu/technologii prodávají a jsou nezbytná pro úspěch.*

Track P1A: UX in practice: *Série přednášek typu „my to děláme takto a takto, vypadá to u nás hezky“. Na řadě institucí poměrně nemalé úsilí věnovali práci s uživatelem s cílem mu maximálně usnadnit interakci s nabízenými službami (výběr různých skupin uživatelů, několik kol testování) – všichni se vesměs shodli, že je to náročný proces – nedělali jej všude z důvodu nedostatku lidí i financí.*

Track P2E: Developers track: *se zabýval převážně technickými tématy, konkrétním řešením různých aspektů. Například přenosu dat z Archivematicy do DSpace, integrací DSpace s jiným systémy přes prostředí GLAMpipe, archivací dat, provozem systémů v AWS (Amazon cloud), multi-tenancy (provoz více digitálních knihoven nad jedním jádrem), DSpace clustering (přes HAProxy a na CephFS).*

Track P3B: *Research data and the Oxford Common File Layout standard: nový standard pocházející z Oxfordu, určený pro transparentní ukládání digitálních objektů. Zatím ve vývoji.*

Track P3E: *Developers track: věnován praktickým vlastnostem DSpace 7 (REST API, Docker), zajímavá byla přednáška o Internet Archive a problémy s ukládáním a správou velkého objemu dat, dále budování repozitáře nad systémem Fedora (ukázka zpracování amerického denního tisku).*

Track P4C: *Large-scale repozitories: ukázka repozitáře na Stanford University, 500 TB dat, používají vlastní řešení, uvažují o přechodu na Fedoru + Samveru + Hyrax/Hyku.*

Track P4A a P5A: *Introducing and Extending DSpace 7: úvod do DSpace 7 – Angular UI, theming, submission workflow, REST API, Entities, DSpace-CRIS, Dspace-GLAM. Přednášky se podrobně*

věnovaly novým vlastnostem v DSpace 7, submision workflow, user interface, nová koncepce entit (typovaných Items) atp.

Track P7D: *Design principles for Cloud-based repozitories: představen systém Archipelago, který má být jako cloudová služba pro muzea, archivy a podobné instituce v oblasti New Yorku. V další přednášce popsán proces přenosu univerzitní repozitáře (University of Illinois) do Amazon cloudu, jak z technické, tak administrativní a finanční stránky. Poslední přednáška tracku byla o vývoji software v Amazon cloudu v projektu OSF financovaném Mellonovou nadací.*

Podrobnější shrnutí přednášek, které jsem shlédl

10. 6. 2019

Workshopy

Islandora for all: ISLE Workshop
David Keiser-Clark, Bethany Seeger (ISLE, USA)

Islandora je je systém pro budování repozitářů a digitálních knihoven, který je postaven na systému Fedora a jako nadstavbu nad Fedorou používá systém Drupal. Islandora je rozšířena především ve Spojených Státech, odhadem má komunita asi cca 50 produkčních instalací. ISLE je komunita, která pomáhá s instalací a údržbou Islandory, především pomocí vydávání Islandory pro Docker. Workshop se saměřil na seznámení uživatelů s Islandorou formou hands-on: instalovali jsme s pomocí Dockeru Islandoru na vlastních noteboocích (bohužel vinou pomalého připojení k internetu se to zdaleka všem nezdařilo). Islandora je bezesporu jedním z možných řešení univerzálního repozitáře – jeden repozitář dokáže spravovat více nezávislých digitálních knihoven (každá s jinou strukturou dat, jinými metadaty apod.) - například <URL>.

Getting Started with Dspace 7 (advanced)
Tim Donohue (USA), Art Lowel (@mire, Belgie), Andrea Bollini (4science, Itálie)

Odpolední workshop byl určen vývojářům a podrobněji nastínil možnosti Dspace 7 z hlediska programování systému, customizací a rozšíření systému. Workshop navazoval na dopolední část, která byla cílena na správu systému a kam jsem se pro velký zájem bohužel nedostal. Postupně jsme si formou hands-on vyzkoušeli řadu nových funkcí Dspace 7. Nejdůležitější změnou ve srovnání s předchozími verzemi je podpora kompletního REST API – DSpace 7 bude možné kompletně ovládat pomocí standardních metod na programovacím jazyku nezávislého vzdáleného volání (do verze 6 včetně to bylo možné pouze částečně, nebo jen prostřednictvím Java API). Měli jsme možnost si toto API na místě vyzkoušet, podrobně jsme se seznámili s jeho funkcemi a formátem (JSON). Druhá část workshopu se týkala nového uživatelského rozhraní, které je na REST API navázáno, vytvořeného pomocí technologie Angular. Seznámili jsme se se strukturou kód, užitím Angularu a možnými customizaci a rozšířeními, které lze pomocí Angularu provádět.

DataCite DOI Services fo Repozitories – making all research outputs persistent
Robin Dasler (DataCite,Německo)

Toto nebyl klasický workshop, ale spíše přednáška – přednášející představila firmu DataCite a její služby, ve druhé půli jsme pak formou hlasování online „ladili“ ideální repozitář – například

výběrem toho, co je pro nás v repozitáři nejdůležitější, co je nejdůležitější pro uživatele apod. DataCite je firma, která se podobně jako například Crossref zabývá přidělováním DOI, nicméně necílí primárně na vydavatele časopisů. Kromě přidělování DOI se snaží zároveň přidat do své nabídky další služby (například statistiky, reporty o DOI, kontrolu odkazů, kontrola provenance – kdo, kdy a jak aktualizoval u konkrétního článku apod.). DataCite spolupracuje úzce na projektu FREYA (financovaný EU), který má za cíl vyrobit jakousi chytrou nadstavbu na současnými systémy perzistatních identifikátorů (Handle.net, DOI, …). Zavádí pojem PIDGraph – systém na strojové propojení identifikátorů (například mohu spojit identifikátor článku s identifikátorem vědeckých dat, ze kterých výzkum v článku vychází) a nad tímto grafem dotazovací jazyk GraphQL. Technicky to realizuje pomocí konceptu EventData, který již DataCite používají – tento koncept slouží ke sledování změn v objektu spojeném s nějakým DOI, například událostí je změna metadata nějakého článku. Na takovou událost lze reagovat, například propojením článku s vědeckými daty.

Open Repositories (konference)

11. 6. 2019

Outcomes over output: a user-centric approch to building successful system
Jeff Gothelf

Úvodní obecnější přednáška o důležitosti marketingu a zejména budování uživatelských rozhraní služeb a jejich designu (tedy jak daná aplikace vypadá – je to mnohdy důležitější, než co daná aplikace umí). Ukazoval řadu příkladů, jak by to mělo a nemělo vypadat (brát si příklady z úspěšných produktů), jednoduchost a použitelnost – stále se skloňoval pojem UX (user experience – jak danou službu vnímá uživatel).

Track P1A: UX in practice

Jisc Open Research Repository: Delivering a compelling User Experience
Tom Davey (Jisc, UK)

Přednášející představil Open research hub (JISC) – ukázal uživatelské rozhraní, administrátorské rozhraní i rozhraní pro vkládání dat, které krok po kroku rozebral (proč to udělali právě takto). Závěr: není snadné to pro uživatele udělat tak, aby to měl pohodlné (bylo to „UX challenging“). V rámci průzkumu posbírali požadavky cca 70 institucí a následně předváděl slepé uličky uživatelského designu, shrnul jak by se mělo správně postupovat (3 kola testování s uživateli). Desposit se jim údajně po redesignu zlepšil.

Uncomplicating the bussiness of repositories
Emily G. Morton, Katherine Lynch (Univ. Of Pennsylvania Libraries, USA)

Přednáška ukazovala, jak dělali upgrade UI na Univ. Of Pennsylvania. Používají software Colenda (repozitář) a Omeka (publikační platforma). Automatizují, kde to je trochu jde (metadata z OPACu, hledání výrazu pouze v rámci geograficky vyznačené lokace – například v okolí nějakého města). Repozitáře „integrují“ na webu v rámci systému OPenn. Mluvili i o výběru materiálů k digitalizaci a dlouhodobému uchovávání.

Building interfaces for all users
William Hicks (University of North Texas Libraries, USA)

Ukázka systému Aubrey pro budování digitálních knihoven – 3 různé digitální knihovny v jednom repozitáři: 2,6 milionu položek v 1125 kolekcích, jednotka je objekt = více reprezentací stejných dat. Primárně mají skonované stránky (jedna položka průměrně 15 stran). Zvažovali udělat mezi uživateli průzkum použitelnosti, vytipovali 3 skupiny uživatelů (dle pracovního zaměření), ale následně po dalším rozboru se jim to rozpadlo na skupin 16 (například speciální skupina nevidomí apod.). Navíc by se na řadu uživatelů nemuselo dostat, protože jsou často mimo univerzitu (uváděl např. „grant writers“). Finálně se tedy kvůli nedostatku lidí a času rozhodli průzkum nedělat. Vyhodnotili si nakonec jen data z Google Analytics (věk uživatelů, použitý hardware, …). Výsledky byly zajímavé, příkladem na podrobná metadata se prokliklo jen 4,8 % uživatelů, nejvíce lidí přišlo z Googlu a byli to noví uživatelé (nikdy předtím tam nebyli). Rozhodli se radikálně předělat homepage, ukazoval jak a proč udělali dané změny a jak to teď vypadá.

Track P2E: Developer track

Série kratších přednášek

Automating OAIS compliant digital preservation using Archivematica and DSpace
Hrafn Malmquist (University of Edinburg, UK)

Ukázka, jak dostávali výstup (DIP i AIP) z LTP systému Archivematica do Dspace (současná Archivematica má v modulu pro ukládání dat i podporu pro uložení do Dspace).

Using Dspace as a backend service – Workflow-centric repository development in practice
Ari Hayrinen (University of Jyvaskyla, Finsko)

Popisoval situaci u nich na univerzitě – mají řadu systémů, jejichž propojení se děje na ruční bázi a bylo by vhodné to automatizovat. Představil systém GLAMpipe – prostředník, který umí mluvit s různými systémy prostřednictvím REST API (s DSpace, KOHA aj.). Například pro odevzdávání závěrečných prací mají vlastní aplikaci, z které přes REST API tečou data do Dspace.

Longleaf: a repozitory-independent utility for applying digital preservation processes to files
Benjamin Pennell, Jason Casden (Univ. Of North Carolina at Chapel Hill University Libraries, USA)
Longleaf je open source aplikace pro příkazovou řádku na správu a ukládání dat – aplikace, která hlídá fixitu dat a stará se o jejich replikaci (mohu si definovat kam a co se bude replikovat), to vše bez dodatečných mezivrstev pouze na souborovém systému (dle Oxford Common File Layout). Po přednášce byla poměrně živá diskuse ohledně tohoto řešení, které je sice jednoduché, ale padly otázky ohledně škálovatelnosti a řada dalších námitek (podobných software je na světě celá řada).

A Multi-Tenancy Cloud-Native Digital Library Platform
Yinlin Chen (Virginia Tech, USA)
Ukázka mixu repozitářů v Amazon cloudu (AWS), vše postaveno na bázi microslužeb (AWS lambda: speciální „aplikace“, například pro přístup k objektu, přístup k metadatům apod.), které spolu komunikují. využitím Apache Airflow (doplnit dle slajdů). Kvůli výkonu používají velmi silně cachování, z hlediska správy velká automatizace. Pro dlouhodobé ukládání BagIt přes speciální software na automatizaci a správu procesů Apache Airflow. Příklad automatizace: nahrají soubor do Amazon S3 úložiště, to spustí službu, která ze souboru extrahuje metadata a ta uloží do metadatového záznamu. Zároveň umožňují souběh mnoha různých digitálních knihoven (tzv. multitenancy) prostřednictvím speciální vrstvy nad službami AWS. Jako hlavní výhodu Amazon cloudu uváděl spolehlivost a rychlou dostupnost (geograficky rozprostřená data a služby po světě).

DSpace Clustering via Puppet, HAProxy and CephFS
Bernd Nicklas, Paul Münch (Philipps-Universität Marburg, Německo)
Přednáška o optimalizaci výkonu DSpace s pomocí clusterování (tedy replikace DSpace na více serverů a následně rozložení zátěže – přístupové požadavky se střídavě rozdělují na jednotlivé servery s DSpace). Replikace instancí DSpace se děje pomocí Puppets (instalace, konfigurace, správa, včetně firewallu) – některé věci však všechny instance sdílí: databázový stroj, SOLR index. Použití open source HAproxy pro rozložení zátěže. Uváděli instalaci DSpace o velikosti 120 TB, kde má taková forma clusterování smysl.

Poster Reception

Posterová sekce byla velice obsáhlá (cca 60 posterů), vybírám proto jen ty, které mě zaujali nejvíce. Obecně ukazovaly postery řadu řešení na různých institucích – jednotlivé technologie, repozitáře, jak na cloud, jak na větší data.

Flexible metadata: the key to a single repository for all types of output
Ben Summers, Tom Renner (HAPLO, Velká Británie)
Poster, který mne zaujal asi nejvíce – představil nový systém pro budování digitálních knihoven s názvem HAPLO. Současné známé systémy (DSpace, Fedora, Invenio aj.) si s sebou historicky táhnou řadu problémů, systém HAPLO je budován v podstatě na zelené louce a snaží se z každého z těchto systémů si vzít to dobré a zahodit to špatné. Podle prezentujících (zde Tom Renner) se zdá, že systém zvládne řadu formátů metadata včetně metadat strukturovaných, nemá problém s uložením i struktur souborů, provozem logicky oddělených digitálních knihoven nad jedním jádrem a další zajímavé vlastnosti. Nevýhodou je to, že má zatím téměř nulovou komunitu, takže je otázka, jestli do budoucna přežije.

A native iPad app for the DSpace 7 REST API
Keith Gilbertson (Virginia Tech, USA)
Ukázka aplikace pro iPad prostřednictvím které lze přistupovat k DSpace verze 7 (aplikace využívá REST API).

Automated metadata generation using machine learning
Harish Maringanti (University of Utah, USA)
Možnosti využití strojového učení pro generování a extrakci metadat.

CDS Videos: The new platform for CERN videos
CERN
Ukázka ukládání videí v CERNu, celkem pěkné – bylo vidět, že s velkými objemy dat umí pracovat.

12. 6. 2019

Track P3E: Developer track

Dspace 7 – Creating High-Quality Software: Update to Development Practices
Andrea Bollini (4Science, Itálie)
Přednáška stručně shrnovala změnu filozofie vývoje Dspace 7 ve srovnání s dřívějšími verzemi. Zmiňovali možnost využití Dockeru pro testování Dspace (větve 4x až 7x), včetně možnosti stáhnout si testovací data. Shrnuli současný stav REST API (pro access pokrytí cca 79 %, pro write režim cca 76 %, zatím to nevypadá, že by ve verzi 7 bylo to pokrytí 100%).

Web Data Engineering: A Technical Perspectives on Web Archives
Helge Holzman (Internet Archive, USA)
Shrnutí aktuálních problémů práce s velkými daty ve web archivu, jaké používají techniky (například analýza jazyka), jak data ukládají. V současnosti spravují přes 40 PB dat, sklízejí asi 5000 stránek za sekundu.

Tools and Techniques for a Repository-Centric Architecture with Fedora

Joshua A. Westgard (University of Maryland, USA)
Ukazoval postupy budování repozitáře na systémem Fedora – první polovina přednášky byla dosti technická a primárně určena pro uživatele Fedory. Ve druhé části ukazoval archiv denního tisku, jakým způsobem se vypořádali s formátem klasických novin (rozbití na články, ale umí zobrazit i po stranách apod.).

Track P4A: Large-scale repositories

Sustaining a Large-Scale Repository Architecture: Behind the Scenes of the Stanford Digital Repository
Michael Giarlo, Justin Coyne (Stanford University, USA)
Ukázka systému SDR, provozovaného na Stanfordu – v současnosti objem dat cca 500 TB. Celkově mají v digitalizaci, repozitářích a souvisejících věcech asi 400 zaměstnanců, z toho je 10 vývojářů. Vše mají v clusteru (desítky počítačů). Vše jim běží na jejich vlastnoručně vyvíjeném systému, nicméně přednášející říkal, že by bylo vhodnější více využívat open source software třetích stran. Do budoucna uvažují o Fedora + Samvera + Hyrax | Hyku.

Track P4A: Introducing DSpace 7

Dspace 7 – The Angular UI from a user's perspective
Ignace Deroost, Art Lowel (Atmire, Belgie)
Přednáška ukazoval změny v uživatelském rozhraní Dspace 7, obhajovalo užití systému Angular. Dochází ke změně designu úvodní obrazovky kvůli optimalizaci zobrazení na mobilních zařízeních (na prvním místě bude vyhledávací formulář, facety se odsunují až níže), výpis položek v komunitě nebo kolekci bude upravitelný, například bude možné vypsat položky s náhledy ve formě tabulky (vhodné například pro obrázky). Admin panel bude připínací a schovatelný, především však dostupný odkudkoli v daném kontextu. Dojde ke změně lokalizace, překlad by měl jít vytvořit snadněji. Opět bude možnost vytvářet vlastní vzhledová témata. Předpokládá se, že ve verzi 7 zatím nebude speciální modul na statistiky. Velmi pravděpodobně bude nutné většinu customizací ze starších verzí Dspace předělávat ručně (nicméně by to mělo v Angularu jít snadněji).

Dspace 7 - Enhanced Submission & Workflow
Giuseppe Digilio (4Science, Italy)
Přednáška se podrobně věnovala novému systému vkládání dat do Dspace. Měl by být snadněji upravitelný. Bude nový MyDSpace, v závislosti na roli uživatele bude možné využívat filtry (zobraz tasky jen moje, všechny, všechny ve workspace, archivované, dle typu jako článek, prezentace apod.). Přibude možnost importu metadat z externích zdrojů (BibTeX, EndNote), metadata bude možné plně customizovat (definovat si jak bude vypadat formulář), stejně tak bude možné k jednotlivým bitstreamům připojit plný metadatový a customizovatelný záznam. Snadno konfigurovatelné by měly být i všechny kroky vkládání, změny dozná systém validací vložených dat, přibude autocomplete na formulářích (například napojený na autoritní bázi). V souvislosti s tím dojde ke změně formátu souboru input_forms.xml, bude k dispozici transformace na nový formát, ale minimálně ruční kontrola po převodu bude následně nutná!

Track P5A: Extending DSpace 7

DSpace 7: open for integrations
Andrea Bollini (4Science, Itálie)
Ukázka, jak lze v novém DSpace 7 využít REST API: napojení na virtuálního asistenta Amazon Alexa, streamování videa aj.

DSpace 7 – The Power of Configurable Entities
Lieven Droogmans (Atmire, Belgie)
DSpace 7 zavádí novou kategorii objektů, kterými jsou entity. V podstatě se jedná o typované Items spolu s vazbami mezi nimi, jejichž cílem je umožnit zavádět specializované typy dat, například profil autora (spolu s vazbou na jeho publikace), projekt, časopis, ročník časopisu apod. Vztahy i entity lze snadno definovat pomocí XML. Zároveň jsou zavedena tzv. virtuální metadata, která je možno snadno přenášet mezi jednotlivými entitami, například časopis má ISSN, které se následně automaticky ukládá i u všech článků časopisu – vše opět konfigurovatelné přes XML. Pro každý typ entity je možné definovat vlastní vzhled (landing page), vlastní facety pro různé typy entit.

Extending DSpace 7: Dspace-CRIS and Dspace-GLAM for empowered repositories and digital libraries
Giuseppe Digilio (4Science, Itálie)
Ukázka nadstaveb Dspace-CRIS and Dspace-GLAM – jedná se o rozšíření, která vyvíjí 4Science, implementující některé funkce a standardy, které samotný DSpace neposkytuje: SignPosting, ResourceSync, OpenAIRE, FAIR apod. Umožňuje budování autorit, propojování na projekty, bibliometriky apod.

Track P6B: Plan S panel

Diskusní panel k projektu Plan S – koalice různých institucí a národních konsorcií, které financují vědu. Cílem je mít většinu státem financovaného vědeckého výstupu v Open Access v roce 2021 (S = 'shock'). Viz https://en.wikipedia.org/wiki/Plan_S . Panel nebyl extra zajímavý, nic zásadního kromě již známého tam nezaznělo (měl jsem dojem, že nikdo pořádně neví, co s tím).

13. 6. 2019

Track P7D: Design principles for Cloud-based repositories

When the use case tells you to start over
Nate Hill (Metropolitan New York Library Council, USA)
Prezentace systému archipelago.com a konsorcia Metro – jedná se o konsorcium muzeí a knihoven v okolí New Yorku, které chce mít všechny sbírky v jednom systému. Jedná se o poměrně heterogenní prostředí cca 20 institucí, v rámci daného konsorcia se kromě digitalizace a ukládání dělají i granty, book delivery apod., což je nesnadné skloubit. Do nedávna používali software na bázi Islandory verze 7 (dcmny.org), ta však začala nedostačovat. Zvolili tedy cestu vlastního řešení, kdy si ze známých open source projektů (DSpace, Fedora Islandora/Samvera, Eprints, Invenio) vzali (resp. nechali se inspirovat) ty dobré věci. Nový software staví už v cloudu (S3, Azure, …). Zajímavé bylo, že všechna metadata ukládají ve formátu JSON, údajně se v tom velmi dobře rychle vyhledává.

The Challenges and Charms of a Cloud-Based Repository Infrastructure Transition: A Case study from the University of Illinois
Seth Robbins (University of Illinois, USA)
Velmi zajímavá přednáška popsala konkrétní případ přenosu vlastního repozitáře do Amazon cloudu (AWS). Na University of Illinois mají vlastní repozitář zvaný Medusa (Illinois Digital Library), objem dat cca 15 TB. Cílem je převést vše do Amazon Glasure – motivace: zbavit se údržby a práce na vlastním hardware, řešení škálovatelnosti i pro budoucnost, lepší dostupnost a spolehlivost. Představa byla, že prostě vezmou data, tak jak je mají na discích a 1:1 je přenesou do Amazon Elastic File System (EFS). Nakonec se ukázalo, že EFS by bylo velmi drahé, cca 10krát dražší než

klasické S3 úložiště, které nakonec použili. Protože S3 není klasický souborový systém, tak jim to výrazně celý přenos komplikovalo (použili AWS Ruby SDR, museli přepsat aplikace, aby uměli komunikovat s S3. Zároveň chtěli přenést do Amazonu i jejich vlastní image server, což se také ukázalo komplikované. Celkový závěr – nakonec byl celý přenos a vše okolo dražší než provoz univerzitní infrastruktury, ale předpokládá se, že při narůstajícím objemu dat jim Amazon vyřeší problém škálovatelnosti a tím se jim to nakonec vyplatí. Mají i exit strategii pro přechod do Google Azure cloudu. Závěrem si stěžoval, že podpora od Amazonu za moc nestojí.

Forming a CI/CD Pipeline and Cloud-first Culture
Jeremy Friesen (University of Notre Dame, USA)
Projekt financovaný Mellonovou nadací na vybudování komplexního systému pro ukládání a zpřístupnění různorodých dat. V přednášce bylo ukázáno kompletní workflow pro vývoj software v cloudu. Propojení gitHubu s Amazonem – systém automatického deploy workflow (kompilace, testy, schvalování, přenos do cloudu – na vše grafické klikací rozhraní). Začínali na PRIMO od ExLibris, ale nakonec to zahodili, protože měli pocit, že ExLibris se s nimi moc nebavili a nakonec přešli na AWS spolu ElasticSearch, ReactJS a GatsbyJS. Například moc neřešili metadata, protože ta koncového uživatele nezajímají. Detailní informace o projektu jsou na osf.io/cusmx.