

Zpráva z cesty na konferenci Open Repositories Conference 2024

Švédsko, Göteborg, 3. - 6. 6. 2024

Účastníci: Vlastimil Krejčíř, Petra Kubálková

Web: <https://or2024.openrepositories.org/>

Abstrakt

Ve dnech 3. až 7. června se konala 19. mezinárodní konference o repositářích - Open Repositories (OR). Open repositories 2024 nabídl paralelně 11 workshopů, 130 prezentací a panelových diskuzí. V řadě paralelních tracků nabídla konference prostor ke sdílení aktuálních trendů ve výše zmiňovaných oblastech pro široké spektrum posluchačů spojených primárně se světem akademických knihoven, od softwarových vývojářů až po manažery. Konference je známá tím, že je zaměřena na sdílení praxe "how to" více než na "how come".

Kromě prezentací nabízí i široký prostor pro osobní kuloárová jednání. Z pohledu projektu CARDS je zajímavé zejména sledování aktuálních trendů v oblasti repositářů, které by měli být napojeny na vyhledávací službu Discovery - zajímavý je zde jak technický, tak i manažersky logistický pohled na danou problematiku, a to vzhledem i ke komplementárnímu projektu EOSC-CZ. Oba zástupci projektu navštěvovali odlišné tracky, každý dle svého zaměření (Vlastimil Krejčíř technické, Petra Kubálková manažerské).

Časový průběh

Vlastimil Krejčíř (dále VK)

- 2. 6. přilet v pozdně odpoledních hodinách
- 3. 6. workshopy, účast na InvenioRDM workshop
- 4. až 6. 6. konference OR 2024
- 6. 6. návrat do ČR

Petra Kubálková (dále PK)

- 2. 6. přilet v ranních hodinách
- 3. 6. workshopy, účast na Unlocking Accessibility: PDF Annotation with PDF/UA Standard and WCAG Guidelines a Getting running with ARK persistent identifiers
- 4. až 6. 6. konference OR 2024
- 6. 6. návrat do ČR

Program



**Spolufinancováno
Evropskou unií**

3. června – Pre-conference Workshop Day

08:30-09:30 Registrace

09:30-12:30 Workshopy

12:30-14:00 Oběd

14:00-17:00 Workshopy

Úterý 4.června – Konference

08:00-09:00 Registrace

09:00-12:30 Konference

12:30-13:30 Oběd

13:30-17:15 Konference

17:15-18:45 Poster session & Welcome reception

Středa 5.června – Konference

09:00-12:30 Konference

12:30-13:30 Oběd

13:30-17:00 Konference

19:00-23:00 Networkingová večeře v Kajskjul8

Čtvrtek 6.června – Konference a doprovodný program

09:00-12:30 Konference

12:30-13:30 Oběd

13:30-17:00 Fedora User Group Meeting (open for all) – bez účasti NTK

13:30-17:00 COAR Meeting Day 1 (open for all) – bez účasti NTK

Pátek 7.června

09:00-16:00 COAR Meeting Day 2 (COAR members only) – bez účasti NTK

Shrnutí průběhu konference a nejzajímavějších přednášek a získaných informací

Pre-conference Workshop day

InvenioRDM workshop (VK)

Co je InvenioRDM? Jedná se o platformu určenou převážně pro ukládání vědeckých dat. V podstatě obsahuje základní funkce (uložení dat, metadat, vyhledávání, autentizace, autorizace, možnost tvořit skupiny uživatelů, kteří si mohou ukládaná data "recenzovat", integrace s ORCID, řízené slovníky apod.). Technicky je to postaveno hodně na JSON a Python, veškerá konfigurace se tedy děje na bázi těchto jazyků. Obecně se mi InvenioRDM jeví jako systém vhodný pro ukládání zejména velkých objemů vědeckých dat (stovky TB) pro oblast přírodních věd, systém

by měl i při těchto objemech škálovat. Předpokladem je, kromě primárního cíle bezpečného uložení, že uložená data budou poskytována dalším aplikacím pro zpracování (vědecké výzkumné prostředí) - což byl obecný trend na celé konferenci (v oblasti přírodních věd). Software InvenioRDM používá projekt EOSC-CZ a měl by být napojen i na vyhledávací službu (centrální index) v PNG. Workshop umožnil velmi rychle získat technický i obecný vhled do systému InvenioRDM a možnostech jeho užití. V dopolední sekci byl systém představen a formou řady krátkých přednášek vybraných uživatelů systému InvenioRDM jsme se seznámili s možnými případy užití, s výhodami i nevýhodami systému. Poměrně zajímavé bylo povídání, jak převáděli jejich staré verze Invenia, které mimochodem používaly jako metadatový formát MARC21, na novější verze. Prý už by to za žádnou cenu nechtěli dělat znovu 😊. Odpolední sekce se rozdělila na dva tracky, zúčastnil jsem tracku vývojářského, ve kterém jsme systém InvenioRDM instalovali. Instalace probíhala na vlastních počítačích s pomocí vývojářů a zkušených uživatelů. Bylo vidět, že systém InvenioRDM je v intenzivním vývoji, instalace nebyla jednoduchá, troufám si tvrdit, že bez pomoci vývojářů (konkrétně Maxe Mösera z TU Wien) bych s tím strávil několikanásobně více času. Systém je postavený na jazyce Python a vyžaduje dobře sladěné prostředí a vzájemnou kompatibilitu konkrétních knihoven a komponent. Většina účastníků měla vůbec problém dát dohromady dané prostředí. Mně se nakonec s vydatnou pomocí systém rozjet podařilo, ale dalo nám to pořádně zabrat 😞.

Archival Resource Key (ARK) (PK)

ARK slouží jako trvalé identifikátory nebo stabilní důvěryhodná reference pro informační objekty a identifikují cokoliv digitálního, fyzického nebo abstraktního. ARKs je podobná identifikátorům DOIs, URNs, and Handles. Aliance ARK je otevřená globální komunita podporující infrastrukturu ARK. Od roku 2001 bylo vytvořeno přibližně 8,2 miliardy ARK více než 1250 organizacemi – knihovnami, datovými centry, archivy, muzei, vydavateli, vládními agenturami a prodejci.

Konference (výběr nejzajímavějších příspěvků)

Den první

Keynote příspěvek (Gustav Nilsson, neurolog, Karolinska Institutet, Švédsko) se dotýkal problematiky Open Access, které máme podle něj zvládnuté perfektně technicky, ale bohužel už méně politicky (Gold OA a stálé navyšování poplatků). Kritika padla i na současný systém peer-review (mělo by se platit recenzentům) a na to, že se nezveřejňují statisticky nezajímavé studie, což zkresluje výsledky celého systému vědy. Na dotaz z publika "Co jsme my, knihovníci, udělali stran OA špatně?" odpověděl "Vy nic, to my vědci jsme se dost nesnažili."

Developers track

Byl představen standard OpenCost (<https://github.com/opencost-de/opencost/tree/main/doc>), spolu se systémem pro zpracování, správu a vyhodnocování APC poplatků. Toto by mohlo být potenciálně zajímavé i pro CARDS, jakožto služba nabídnutá členským institucím (byť to sám neumím posoudit, je to tedy spíš na úrovni tipu). Viz <https://www.opencost.de/en/project/>

V poměrně hodně přednáškách se objevila problematika AI, přesněji řečeno se mluvilo o machine learning (ML). Využívá se na různé operace nad daty pro jejich analýzy, hledání podobností apod. Například systém Archipelago na New York University pracuje na podobnostech obrázků (mají katalog miliónů obrázků a neví, co s tím).

Zajímavý byl systém/framework OSF na podporu kolaborativního vědeckého prostředí. Vědci takto mohou sdílet svá data (např. přes OneDrive nebo GoogleDrive), diskutovat, anotovat apod. Je možné institucionální i individuální členství. Viz <https://osf.io/>

Systém INFRAFINDER pro porovnávání infrastruktur. Nakrmí se informacemi a následně provede srovnání vybraných infrastruktur (discovery, publikační systém, repozitář), vypadalo to velmi užitečně <https://infrafinder.investinopen.org/solutions>

Zajímavá byla i přednáška o efektivitě přesunu repozitářů a jiného sw do cloudu, ale neřekl bych, že tam zaznělo něco vyloženě převratného. Z hlediska poměru nákladů a přínosů je klíčové, jestli je konkrétní software na běh v cloudu již vytvořen (např. běží v kontejnerech), tj. u starších softwarů nepřipravených cloud je přesun do cloudu akcí, která se nevyplatí.

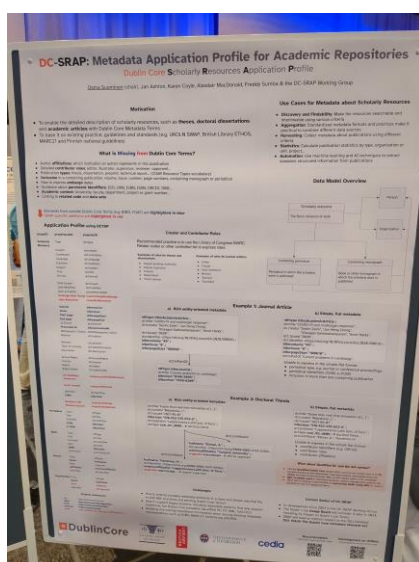
Panel

Panelová diskuse nebyla až tak diskuse, byli tam zástupci různých repozitářových systémů: DSpace, Fedora, ePrints, Dataverse, InvenioRDM, Islandora, Samvera. Tito zástupci měli každý cca 7 minut na to, aby daný systém v kostce představili. Byla to spíš rychlá propagace, i když místy i spíš propaganda, ale jako přehled pro neznalého uživatele se to dá považovat jako za dobrý úvod.

Postery (výběr)



Poster autorek z amerických univerzit (Arkansas a Houston) se zabýval problematikou objektivního určení "veřejného blaha". Máme obrovské množství historického materiálu (v případě posteru fotografií) a jak máme určit, co se vyplatí nebo nevyplatí skenovat. Cílem je mít nějakou metodiku, která by na základě řady parametru objektivně určila míru smysluplnosti pro digitalizaci a archivaci konkrétních sbírek a materiálů. Jednou z motivací je i boj s úředním šimlem, protože úředníci mají rádi čísla (to na posteru není, to jsem se od autorek dozvěděl mimochodem).



Poster se dotýkal problematiky metadatového formátu Dublin Core, resp. DCTERMS. Pomocí těchto formátů nelze optimálně popsat závěrečné kvalifikační práce, autoři posteru se tedy pokusili o rozšíření DC. V návrhu přidávají další pole a je otázka zda-li uspějí.

Presentations: Regional Collaborations

Blok byl složen ze tří prezentací: Building Community to Coordinate Ireland's Fragmented Open Repository Network, Collaboration Across Borders, PHAIDRA's Journey Towards Sustainable and Open Repositories. Prezентující sdíleli důležitá zjištění a identifikovali potřeby vedoucí ke spolupráci při budování repozitářů. Nejzajímavějším byl první příspěvek, který sdílel zkušenosti z budováním National Open Access Repositories - **Building Community to Coordinate Ireland's Fragmented Open Repository Network** - Cillian Joy, Stephanie Ronan, Christopher Loughnane, iniciativu vede Universita v Galway. V rámci prezentace byly nasdíleny (zatím neoficiální) výsledky průzkumu jehož tématem bylo mapování repozitářů.

Průzkum potvrdil podobnou situaci v ČR, že prostředí repozitářů je velice různorodé a instituce čelí výzvám při jejich správě, ať na straně financování, technologického vývoje nebo využití. Nutností pro zmírnění všech překážek je udržitelné financování pro zajištění odborníků, aktualizací a vývoj prostředí. Zároveň je nutné nastavit standard pro kvalitu metadat. Dobrou praxí pro obě zmíněné oblasti je právě budování konsorcií a neustále vysvětlování důležitosti otevřeného přístupu k vědeckým informacím.

Navrhovaným postupem pak je community-driven solutions, které nezapomíná na silnou infrastrukturu, posilování kvality metadat, implementace nových technologií, budování kapacit a vzdělávání komunity, spolupráce v rámci konsorcií a propagace otevřeného přístupu. V rámci prezentace byla sdílena ukázka vzdělávací centra [NORF](#).

Den druhý

Developers track

Stručně lze zmínit informace o standardu RO-Crate pro popis vědeckých dat (<https://www.researchobject.org/ro-crate/>).

Následovala série přednášek od implementaci vědeckých výzkumných prostředí. V principu většinově využívali nástroj Jupyter notebook společně s Docker kontejnery a spuštěním v cloudových řešeních jako jsou Kubernetes. S větší nebo menší účastí konkrétního vědce připraví image kontejneru společně s příslušnými vědeckými daty, po spuštění kontejneru lze pak nad těmi daty provádět analýzy (obvykle ukazovali grafy). Zmiňované projekty: Binder, resp. BinderHub (<https://mybinder.org/>, <https://binderhub.readthedocs.io/en/latest/>), SciLifeLab (<https://serve.scilifelab.se/>).

Lessons learned track

V rámci tracku proběhla představení některý repozitářových řešení, vesměs nic zásadně převratného. Zajímavý byl příspěvek [Tyng-Ruey Chuanga](#) z Taiwanu, kde používají software CKAN (<https://www.ckan.org/>) a jako perzistentní identifikátory ARK, jako jeden důvod byly uvedeny menší náklady.

Problems solving track

Zajímavá přednáška poukazující na odvrácenou stránku AI: vysoká spotřeba elektrické energie a zejména velmi agresivní chování AI serverů při harvestování dat pro učení, které často zahrnuje repozitáře a knihovní systémy (v Metropolitan New York Library Council došlo v době spuštění

ChatGPT k přehlčení všech jejich 28 repozitářů). Byly představeny nějaké pomocné techniky, jak se podobnému scénáři bránit (zmíněny například HTML meta tagy *noAI* a *noImageAI*).

Repository Trust and Certification track

Sekce se zabývala zkušenostmi různých institucí s úspěšným i neúspěšným získáním certifikátu CoreTrustSeal. Velmi zajímavá byla i následná diskuse k problematice. Získání CoreTrustSeal je poměrně nepříjemný, náročný, frustrující a dlouhotrvající proces (například samotná příprava klíčového dokumentu může vyžadovat 2 lidi na plný úvazek na dva měsíce).

Den třetí

Developers track

Jedna přednáška o virtuálním vědeckém prostředí na TU Wien (v podstatě stejné principy jaké jsou uvedeny ve zprávě výše), opět postaveno na Jupyter technologii a kontejnerech.

Další příspěvky

Nejzajímavější byla přednáška **Herberta van de Sompela (FAIRiCat: Supporting Discovery of a Repository's Interoperability Affordances)** o návrhu nového standardu FAIRiCat. Cílem standardu je umožnit repozitářům, digitálním knihovnám, knihovním systémům a v podstatě jakýmkoli webům vystavit strojově čitelnou informaci o tom, jaké služby poskytují pro získání obsahu (kromě standardního HTML). Je tedy možné takto například "inzerovat" přítomnost OAI-PMH serveru, speciálního API aj.

Zajímavá byla také přednáška **Looking up from the weeds: seeing what's next for OA by learning from the past**, kterou přednesl Richard David Jones z Cottage Labs, United Kingdom. Motivem přednášky bylo zamyšlení nad minulostí a budoucností OA. Pár postřehů z příspěvku, které jsou volně parafrázované. Na začátku knihovny věřily, že OA stlačí ceny vydavatelů odborných zdrojů dolů, ale mýlily se. Vydavatelé vždy najdou způsob, jak se přizpůsobit OA a přitom vydělat peníze, prostě je nelze obejít. Pro udržení a rozvoj OA je rozhodující dobrá komunita a veřejná správa, která OA podporuje, stejně tak potřebujeme robustní infrastrukturu a prostor pro malé experimentování. Konsorcium nebo řešitel, který má dostatek financí pak bude udávat směr v budoucím vývoji OA. Důležitá je podle Jonese také změna uvažování o repozitářích. Repozitář nesmí být pouze úložištěm pro data, ale naopak musí být místem pro začátek dalších inovací.



Celou konferenci ukončil příspěvek od Monica Granados, která představila iniciativu “**Open Climate Campaign**” společnosti Creative Commons propagující otevřený přístup k výzkumu klimatu a biologické rozmanitosti. Již během přednášky se některé univerzity, skrze své zástupce na konferenci, přihlásily ke kampani. Více o iniciativě: <https://openclimatecampaign.org/>