Výzkumné projekty jako hnací síla pro vývoj open source a rychlá cesta na trh

RODA, SCAPE a E-ARK – případová studie

Hélder Silva, Miguel Ferreira, Luís Faria
KEEP SOLUTIONS, LDA
R. Rosalvo de Almeida, no5
Braga, Portugal
{hsilva, mferreira, lfaria} @keep.pt

Abstrakt

Výzkumné projekty, zvláště v oblasti počítačové vědy, trvale poskytovaly výstupy jako open source produkty nebo updaty k dlouholetým open source projektům. K tomu dochází díky sdílné a otevřené povaze vědeckého výzkumu i open source hnutí, které umožňuje opětovné využití komunitami, což pozitivně ovlivňuje rozvoj jak výzkumu, tak open source produktů. Ale slouží-li open source projekty společnosti a řeší-li problémy skutečného světa, může se rychlost výzkumu dostat do střetu se setrvačností aplikace ve skutečném světě. Přesto mohou výzkumné projekty přinést velice potřebnou inovaci open source projektů a ty mohou otevřít potřebnou cestu na trh, kde investoři hledají výstupy výzkumů, jež financují, a ujišťují se, že rozpočet vynaložený na výzkum skutečně pomůže komunitě a zlepší svět.

Toto pojednání představuje analýzu této dynamiky v případové studii o systému RODA, open source digitálním repozitáři, využívaném v paměťových institucích jako jsou archivy, a dvou výzkumných projektech: SCAPE, který je zaměřen na škálovatelné služby digitální ochrany, a E-ARK, zaměřený na standardizaci informačních balíků, integraci s reálnými aplikacemi a uchování databází.

Článek se dále pokouší identifikovat osvědčené postupy pro použití stávajících open source projektů ve výzkumu a ujistit se, že výstupy výzkumů jsou přenášeny do hlavních verzí open source projektů a najdou si cestu k uživateli.

Kategorie a deskriptory subjektu
H.3.7 [**Informační Systémy**]: Úchovávání a získávání informací - Digitální knihovny

Klíčová slova
Ochrana, Repozitář, Výzkum, Open Source, Integrace

1. PŘEDMLUVA

RODA[1] je open source digitální repozitář speciálně navržený pro archivy, jehož prvořadým cílem je dlouhodobá ochrana a autentičnost. Byl vytvořen v roce 2006 během dvouletého projektu, který vedl portugalský Národní archiv ve spojení s Univerzitou Minho, a který později vedl k vytvoření společnosti KEEP SOLUTIONS[2]. Ta doteď pokračuje ve vývoji RODA, zaštituje její open source komunitu a na vyžádání poskytuje komerční služby pro údržbu, podporu a vývoj dalších funkcí.

1 http://www.roda-community.org

2 http://www.keep.pt

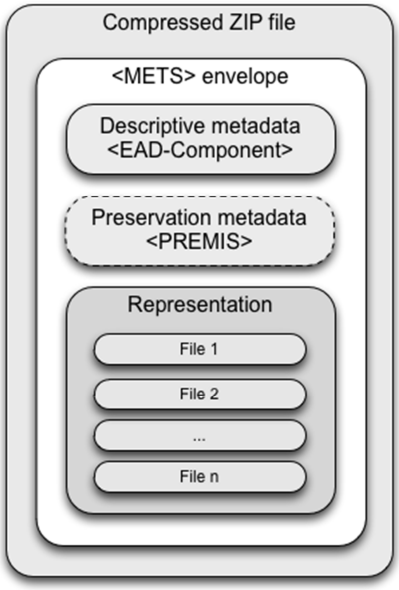
SCAPE[3] byl projekt spolufinancován Evropskou komisí ze Sedmého rámcového programu. Běžel od roku 2011 do roku 2014 a mířil na rozvoj škálovatelných služeb pro plánování a výkon institucionálních prezervačních strategií na open source platformě, která řídí poloautomatizované pracovní toky pro rozsáhlé, heterogenní kolekce komplexních digitálních objektů.

E-ARK[4] je pokračující projekt spolufinancovaný Evropskou komisí v Rámcovém programu pro konkurenceschopnost a inovace. Bude běžet od roku 2014 do roku 2017 a míří na vývoj celoevropské metodologie pro elektronickou archivaci dokumentů, syntézu nejlepších existujících národních a mezinárodních postupů pro uchování záznamů a databází, které zůstanou i postupem doby autentické a použitelné.

Tento článek poskytuje celkový pohled na výzkumné projekty, jmenovitě SCAPE a E-ARK zahrnuté do projektu RODA, na to, jak byly jejich výsledky začleněny do hlavních funkcí, jak jsou zaměřeny na budoucí vývoj a jaký dosah má výzkum na uživatele.

2. RODA

RODA je komplexní systém digitálního repozitáře, který poskytuje funkčnost pro všechny základní jednotky referenčního modelu OAIS. RODA plně implementuje pracovní postupy Ingest, který ověřuje SIPy a migruje digitální objekty do formátů vhodných pro dlouhodobou ochranu, poskytuje přístup skrze různé způsoby vyhledávání a navigaci díky dostupným datům, stejně jako vizualizaci a přístup k uchovaných digitálním materiálům. Funkce data managementu dovoluje archivářům vytváření a modifikování deskriptivních metadat a definuje pravidla pro ochranné akce, například plánování kontroly integrity dat na uložených digitálních objektech nebo zahájení migračních procesů. Administrativní procesy umožňují nastavení přístupových práv k datům a oprávnění pro každého uživatele nebo skupinu.



Obrázek 1: Struktura RODA SIP

Dříve než je RODA schopna přijmout Informační balíčky, které se v OAIS nazývají Submission Information Packages (SIP), musí mezi producentem dat a repozitářem

3 http://www.scape-project.eu

4 http://eark-project.com

proběhnout formální nebo neformální dohoda, aby se mohl specifikovat jejich obsah (například specifikace požadovaných sad informací nebo v jakém standardu by měli být zakódované) a platný časový rámec. Jelikož má RODA svůj vlastní formát SIP (viz Obrázek 1), kdokoli, kdo chce ukládat do repozitáře, musí vytvořit SIPy, které jsou vyhovující pro RODA. Aby toho bylo dosaženo, může se použít desktopový nástroj (RODA-in), který dovoluje vytvoření a upload SIPů pro RODA, anebo je možné přímo použít rozhraní pro webového uživatele RODA (RODA-WUI).

Oba tyto přístupy mají svá omezení, protože neškálují a užívají vlastní specializovaný formát SIPů. To může být problematické, když je nutné vytvořit masivní SIPy a existující systémy nevytváří vhodné SIPy pro RODA. To lze obejít pomocí vývoje programů, které vhodně propojí existující systémy s RODA tím, že vytvoří SIPy ve formátu vhodném pro RODA, ale existuje mnoho systémů a ne všechny instituce mají prostředky k vývoji vlastních integrací jen pro tento účel. Pokud nejsou zavedeny směrnice pro tvorbu a sdílení informačních balíčků jednotným způsobem, ať už jako doporučení nebo zákonem dané požadavky, integrace a sdílení informací mezi systémy nebo entitami ze stejné země je těžké a stává se ještě těžší v širším mezinárodním kontextu.

Po ingestu referenční model OAIS doporučuje, aby byly vykonány akce, které zajistí, že informace dostupné v repozitáři zůstanou nadále přístupné a srozumitelné pro uživatele. Tyto akce jsou definovány a monitorovány procesem plánované ochrany, kterým může být i jednoduchý proces, jako je detekce poškozených souborů pomocí kontrolních součtů, nebo komplexnější jako je migrace formátů souborů a kontrola kvality uložených dat. Z technického pohledu RODA a její konzervační aktivity (sada plug-inů, které mohou být prováděny manuálně nebo může být naplánováno pozdější spuštění) ulehčují vykonání těchto konzervačních úkolů. Ale z pohledu managementu musí být opodstatněné rozhodnutí pro výběr optimálních ochranných úkolů, aby se zajistil nepřetržitý přístup k informacím dostupným v repozitáři.

Proces plánované ochrany je definován jako úkol odpovědný za *monitorování prostředí OAIS archivu a poskytující doporučení a ochranné plány za účelem zajištění, že informace uschované v OAIS archivu zůstanou dlouhodobě přístupné, srozumitelné a dostatečně použitelné pro určenou komunitu, i kdyby se stalo původní počítačové prostředí zastaralé (1).* Tento proces může být periodickým, manuálním nebo polo-automatizovaným způsobem vykonáván správcem repozitáře. Jak ale informace v repozitáři přibývají a stávají se rozmanitějšími, ruční monitorování všech rizik postihujících formáty souborů a plánování správných kroků pro jejich provedení může být neuskutečnitelné. Automatizace některých kroků v plánování ochrany se stává rozhodující pro udržení důvěryhodného repozitáře a autentičnosti chráněných digitálních objektů.

Toto jsou pouze některá z omezení zjištěných při implementaci RODA před spuštěním tohoto výzkumného projektu a zároveň se stala předmětem dalšího výzkumu. V dalších částech představíme a popíšeme výsledky výzkumného projektu a proces integrace výsledků výzkumu do open source projektu.

3. VÝZKUMNÉ INICIATIVY A VÝSLEDKY

V projektu SCAPE byla RODA využita jako referenční implementace díky své integraci se Scout[5], monitorovacím nástrojem pro ochranu, s Plato[6], nástrojem pro plánování ochrany, a s Taverna[7], na doméně nezávislým workflow management systémem, který slouží ke

5 http://openplanets.github.io/scout

6 http://www.ifs.tuwien.ac.at/dp/plato/intro

7 http://www.taverna.org.uk

spouštění úloh pro uchovávání. Tyto integrace umožňují RODA, aby přijala životní cyklus uchovávání, který nepřetržitě monitoruje existenci rizik, vymýšlí plány ochrany k jejich zmírnění, vykonává plány transferu obsahu a opětovně kontroluje, jestli jsou všechny problémy vyřešeny.

3.1 Scout

Scout, monitorovací nástroj ochrany, podporuje škálovatelný proces plánované ochrany pomocí implementace automatizované služby pro sběr a analýzu informací v chráněném prostředí (6). Scout pracuje tak, že konfiguruje zdrojové adaptéry (source adapters), které získávají a normalizují informace z různých zdrojů pro záchranu informací ve znalostní bázi. Tyto zdroje, jak je ilustrováno na Obrázku č. 2, mohou být obsahové (například využití systémových souborů), organizační politiky, formátové a nástrojové registry, web (například zpracování přirozeného jazyka k extrakci znalostí z webových stránek), nebo dokonce lidské znalosti. Ve spojení se souborem informací Scout dovoluje tvorbu dotazů, mechanizmus umožňující vyhledávání ve shromážděných informacích, aby byly zjištěny změny. Navíc mimo systém dotazů mohou být vytvořeny spouštěče (které hlídají podmínky), které je budou pravidelně vyhodnocovat. Pokud se té podmínce nevyhoví, Scout umožňuje například e-mailové hlášení. Po hlášení mohou být zahájené další akce z plánu ochrany. Scout je momentálně integrován s RODA, k vytvoření této integrace byly udělány určité změny, aby bylo možné konfigurovat Scout pro monitorování více aspektů repozitáře. RODA má dostupné API, které umožňuje integraci s ostatními systémy, ale ne vždy API dodává všechny potřebné informace. Na instalaci RODA jsme chtěli monitorovat jak obsah (tedy soubory), tak události v repozitáři (například ukončení *ingestu*). Aby se dala tato integrace provést, byly potřeba následující funkce:

- Reportovací API[8]: OAI-PMH (7) rozhraní, které bude hlásit v repozitáři události jako je zahájení a ukončení *ingestu*
- FITS plug-in: RODA plug-in zodpovědný za charakterizaci každého souboru v RODA za pomoci nástroje FITS[9].

První z nich je přímo integrovatelná se Scoutem (využití zdrojového adapteru pro reportovací API repozitáře), zatímco druhá potřebuje dodatečný nástroj: C3PO - Clever, Crafty, Content

8 https://github.com/openplanets/scape-apis

9 http://projects.iq.harvard.edu/ts

Profiling of Objects[10], který analyzuje technické vlastnosti velkých sad objektů podle metadat generovaných nástroji k charakterizaci jako jsou FITS a Apache Tika[11] a poskytuje souhrnné informace k daným technickým vlastnostem (například velikost souboru, MIME type, komprimační schéma). Výstup FITS plug-inu je vložen do C3PO, který ho vzápětí ve Scoutu konfiguruje na zdrojové informace (za použití zdrojového adaptéru pro C3PO).

Ve Scoutu je zformulována sada podmínek, které musí být splněny, aby bylo prokázáno, že neexistují žádná rizika pro dlouhodobou ochranu a které musí splňovat OAIS kompatibilní repozitář: "Repozitář se musí držet zdokumentovaných zásad a procedur zajišťujících ochranu informace před všemi možnými nepředvídanými událostmi (1)".

3.2 PLATO

Plato je nástroj pro dlouhodobé plánování ochrany, implementuje dobře zdokumentovanou a ověřenou metodologii plánování ochrany a integruje registry a služby pro ochranné akce a charakteristiky (2, 10).
Plato poskytuje webové rozhraní, které umožňuje vybudování interaktivních plánů ochrany, přičemž provází plánovače skrz dobře definovaný rozhodovací proces:

1. Definice vysokoúrovňových požadavků a jejich rozložení na měřitelná kritéria,
2. Vyhodnocení potenciálních plánů dlouhodobé ochrany použitím vybraných nástrojů na zvládnutelné podskupiny objektů, což by mělo pokrýt základní znaky analyzované kolekce;
3. Analýza výsledků a rozhodování, zda by měla být daná strategie aplikována, a pokud ano, může být vytvořen proveditelný plán ochrany.

Na úplném konci procesu vytváření plánu ochrany dostaneme výstupní plán ve formátu XML. Pro umístění plánu do RODA byla vytvořena služba nazvaná Plan Management API[12], která dovoluje tvorbu, vyhledání, aktualizaci a smazání plánů ochrany repozitáře. Vedle operací CRUD[13] také umožňuje vyhledávání pomocí SRU (14) jako vyhledávacího protokolu a CQL (11) jako syntaxe pro reprezentování dotazů. Pro správcovské účely API povoluje monitorování plánů ochrany v repozitáři (například jestli jsou aktivní, jestli jsou vykonávány v určitém čase, zda jsou provedeny úspěšně atd.).

Jakmile je vložen plán do RODA, je k němu přiřazen jedinečný identifikátor. Tímto způsobem, pokud je plán proveden a nastanou změny v intelektuální entitě, je RODA schopna vše propojit. Dosáhne toho pomocí vytvoření PREMIS události (9) (pro každou intelektuální entitu zvlášť), která spojí plán ochrany a reprezentační soubory dané intelektuální entity. Při procházení časového přehledu ochrany specifické intelektuální entity v RODA, pokud byla vytvořena ochranná událost na základě ochranné akce provedené v kontextu s plánem ochrany, může být tento plán okamžitě konzultován, aby bylo popsáno, proč byla akce provedena. Dojde k zobrazení všech rozhodovacích procesů na detailní úrovni, které byly učiněny k výběru provedených specifických akcí, včetně testovaných alternativ, použitých vzorků, výsledků experimentu, konečných rozhodnutí a detailů provedení.
Schopnost RODY provádět úkoly plánované ochrany (skrz akce plánované ochrany), spolu s možností přidání externího nástroje pro budování plánu dlouhodobé ochrany, který formálně popisuje požadavky, analyzuje alternativní řešení k omezení bezpečnostních rizik a umožňuje dobře zdůvodněná rozhodnutí, umožňuje RODA ještě větší podporu digitálních

10 https://github.com/peshkira/c3po

11 http://tika.apache.org

12 https://github.com/openplanets/scape-apis

13 CRUD je zkratka pro Vytvoř, Získej, Obnov a Smaž (Create, Retrieve, Update and Delete)

procesů plánované ochrany, které jsou definovány jako nutné pro důvěryhodnost digitální ochrany a repozitářů v normě OAIS a ISO 16363 [5].

3.3 Taverna

Taverna je systém řízení workflow nezávislý na doméně, jde tedy o sadu nástrojů používaných k návrhům a realizacím vědeckých pracovních postupů (16). Zahrnuje i Taverna Engine (užívaný pro přijímání workflow), která pohání jak Taverna Workbench (desktopová aplikace klienta) tak i Taverna Server (který provádí vzdálené pracovní toky). Taverna je také dostupná jako nástroj příkazového řádku pro rychlejší provedení workflow z terminálu bez dohledu GUI.
Užívání Taverna Workbench umožňuje interaktivní vytváření workflow. Jako workflow je definována "automatizace (obchodního) procesu, úplná nebo částečná, během které jsou dokumenty, informace nebo úkoly předávány od jednoho účastníka k druhému k provedení akce, podle dané sady procesních předpisů"[14]. V praxi, a obzvlášť v případě Taverna, to může být způsob, jak popsat, spravovat a sdílet komplexní vědecké analýzy. Toho může být dosaženo kombinací několika složek tzv. Services, buď sekvenčně, nebo paralelně, které mají několik typů:

- Webové služby (místní nebo vzdálené; ve formátu REST nebo WSDL);
- Místní skripty (Bash skripty, R skripty);
- Beanshell (útržky kódu Java)
- Místní služby (předdefinovány Beanshell pro specifické úkoly, jako manipulace se soubory/XML/textem, propojitelnost databáz přes JDBC, atd.);
- Dílčí workflow

Po dokončení návrhu workflow je možné jej okamžitě spustit v Taverna Workbench, Taverna Server nebo v příkazovém řádku Taverna.
Jelikož RODA původně neposkytovala funkce pro správu a provoz plánů dlouhodobé ochrany (dostupné v repozitáři), které v tomto případě obsahuje workflow Taverna, musely být provedeny změny. Proto byly vytvořeny dvě nové API:

- Data connector - REST API pro manipulaci s intelektuálními entitami a jejich souvisejících reprezentací v repozitáři;
- Plan management - REST API pro znovuzískání dostupných plánů ochrany z repozitáře, ke správě jejich stavu (povolit/zakázat) a spouštění (spuštění probíhá, spuštění úspěšné nebo spuštění selhalo).

Nestačí mít mechanismus k manipulaci s daty v repozitáři a správu plánu ochrany, potřebný je také mechanismus pro zpracování plánu ochrany a k provozu Taverna Suite. Kvůli tomu byl vytvořen nástroj zvaný Plan Management WebApp. Kromě správy plánů dlouhodobé ochrany dostupných v repozitáři (tvorba, editování a smazání), umožňuje také vykonání ochranného plánu.

Při spuštění plánu ochrany, Plan Management WebApp vytáhne z repozitáře celkový plán (původně šlo vytáhnout pouze metadata pro účely výpisu), nastaví stav na "spuštění probíhá", identifikuje objekty, které musí být změněny a vytáhne je z repozitáře. Poté izoluje proveditelný plán (například workflow Taverna), vykoná ho v Taverna Suite a poskytne vyčleněné objekty jako výstup a shromáždí výsledky. Dále, pokud všechno probíhá podle očekávání a pokud je potřeba odeslat tyto výsledky zpátky do repozitáře, učiní tak pomocí Datového konektoru API. Pro ukončení nastaví status plánu na "vykonáno" nebo "selhání".

[14] Citováno z http://www.taverna.org.uk/introduction/why-use-workows

Na jedné straně skutečnost, že máme dvě nové API, činí integraci RODA s nástroji/systémy třetí strany snadnější. Na druhou stranu mít možnost spouštění plánů ochrany s workflow Taverna (které jsou úzce spjaté s plánem dlouhodobé ochrany) zlepšuje plnění ISO 16363, protože lépe splňuje požadavky plánu dlouhodobé ochrany.

Zpráva o shodě systému prezentována výše, nazvaná SCAPE Preservation and Watch Suite nebo SCAPE Preservation Environment, která spojuje RODA, Scout, Plato a Taverna s ISO 16363, je hodnocena pouze z pohledu softwarové technologie (tudíž ignoruje organizační, finanční nebo fyzické infrastrukturální požadavky) ukazuje, že 69 z požadavků je plně podporováno, 2 jsou podporovány částečně, 6 není podporováno a 31 je mimo rozsah (je ignorováno) (4). Téměř všechny požadavky jsou podporovány pouze tímto softwarovým balíkem, zbývající mohou být podpořeny manuálními procedurami, což je velké zlepšení proti předešlým verzím.

4. BUDOUCÍ VÝZKUM

RODA je použita jako jeden z pilotů v projektu E-ARK, který bude rozvíjet celkem 6 různých pilotů. Tento pilot bude řešený ve spolupráci s Portugalskou agenturou pro reformu veřejných služeb (AMA)[15] a Instituto Superior Técnico[16], protože RODA bude řešením pro dlouhodobou archivaci a tyto dvě organizace budou poskytovat data.

Jedním z cílů je podpora celoevropského formátu SIP, který usnadní vytváření informačních balíků, jejich transfer a import do archivů způsobem, který je účinný, spolehlivý a použitelný ve všech evropských zemích. Dalším cílem je zlepšit proces ingestu do RODA, aby byl více flexibilní a upravitelný, tedy aby se snadněji integroval do systému třetích stran bez potřeby lidského zásahu, což by činilo systém více škálovatelným.

Pilot bude demonstrovat adekvátnost celoevropské struktury SIP navržené v E-ARKu k podpoře obsahových typů v současnosti podporovaných RODA (například relační databáze, textové dokumenty, video, audio a CD obrazy) a poskytne rozhraní pro automatickou tvorbu SIP pomocí systému správce dokumentů.

Tento projekt se také soustředí na přístup k obsahu, speciálně komplexního obsahu, jakým jsou relační databáze, nalezení škálovatelných metod k poskytnutí přístupu do archivačních databází a také poskytnutí metod k umožnění pokročilých analýz a opětovné použití databázového obsahu kupříkladu pomocí OLAP technologií (3).

5. CESTA NA TRH

Jako kterýkoli open source projekt má RODA svůj zdrojový kód volně přístupný. Jak je také dobrým zvykem ve vývoji softwaru, má zdrojový kód RODA více verzí. RODA užívá Git (15) jako systém pro správu svých verzí a publikuje zdrojový kód v hlavním repozitáři[17] GitHub.

Když začne nový projekt, je vytvořena větev (8) hlavního zdrojového kódu, což umožňuje oddělení vývojových trendů. Na konci specifického projektu jsou všechny vývojové trendy zanalyzovány a je rozhodnuto, které z nich budou integrovány do další oficiální verze. Tato analýza musí proběhnout, protože ne všechny vývojové trendy mají širokou použitelnost a tudíž nemusí být vhodné pro širší publikum nebo mohou mít vážný dopad na další funkce, které byly předtím vyvinuté komunitou.

15 http://www.ama.pt

16 http://tecnico.ulisboa.pt

17 https://github.com/keeps/roda

Všechny vývojové trendy, hlavní i alternativní, vytvořené pro výzkumné projekty, jsou zveřejněny v GitHubu a jsou volně přístupné, aby je komunita mohla dále vyvíjet a stavět na nich. Ovšem pouze hlavní verze je průběžně udržována hlavními vývojáři a používaná jako základ pro nové výzkumné projekty. To zajišťuje, že vývojové práce jsou soustředěné a projekt není příliš roztříštěn.

Dle výzkumných iniciativ a výsledků prezentovaných v oddíle 3, které se převážně vztahují k projektu SCAPE, jsou Scout, Plato a Taverna externí služby RODA, které se s ní integrují pomocí tří API a jednoho plug-inu. Tyto API byly vyvinuty jako nezávislé na systému repozitáře, a některé z nich jsou implementovány pro jiné systémy repozitářů[18]. Samotné API jsou vyvinuty pro dodržování standardů (jako OAI-PMH, PREMIS, Dublin Core, METS), aby byly flexibilní (minimální povinné informace) a aby měly co možná nejmenší dopad na základní datové modely. Všechny tyto parametry umožňují API, aby byly sloučené s hlavním zdrojovým kódem a poslány do další verze RODA. Plug-in byl implementován jako modulární a uzavřený softwarový komponent přidávající specifickou funkci a umožňující snadné spojení a dodání s další verzí.

Budoucí výzkum prezentován v oddílu 4, který se převážně věnuje projektu E-ARK, popisuje budoucí vývoj, který má mnohem hlubší dopad na datový a obchodní model RODY. Změna formátu SIP může představovat informační omezení nebo flexibilitu, která může mít dopad na provádění ingestu a toho, jaké informace mohou nebo musí být uchovány v systému. Toto je akceptované riziko pro schopnost sloučení změn s hlavním zdrojovým kódem při přinášení výsledků projektu uživatelům. Riziko je sníženo srovnáním cílů výzkumného projektu s plánem samotného projektu open source, což zajišťuje, že zásadní změny jsou výhodné pro celou komunitu. Testování změn v reálných případech udržuje neustálý kontakt s cílovou komunitou. To je dáno povahou projektu E-ARK, který je financován rámcovým programem pro konkurenceschopnost a inovace a který nestanovuje cíle projektu nadneseně a nesplnitelně, ale stanovuje je k vytvoření příznivého ekosystému a růstu trhu. To se zhmotnilo v projektu E-ARK zaměřeném na piloty, kde se řídí a testují vývojové trendy v reálných podmínkách, integrují systémy s referenčními institucemi v evropském kontextu, zajišťuje se směřování na komunitu a testování v reálných podmínkách.

6. ZÁVĚR

RODA je kompletní digitální repozitář zajišťující funkčnost pro všechny hlavní jednotky referenčního modelu OAIS. Přesto, stejně jako kterýkoli software, který chce být úspěšný, musí být přístupný dalšímu zlepšení. Tato zlepšení mohou být vyvolána potřebami vlastní komunity stejně jako změnami v komunitě dlouhodobé ochrany dat, která musí být nepřetržitě monitorována pro udržení aktuálnosti RODA pomocí zkušeností z oboru. To, že má RODA základ v open source technologiích a dobře zavedených standardech jako METS (13), EAD (12) a PREMIS, usnadňuje její zlepšování. To ukazují zlepšení vytvořená v projektu SCAPE stejně jako vylepšení, na které se přijde v průběhu projektu E-ARK. A další skvělá výhoda open source je skutečnost, že tyto zlepšení jsou volně a okamžitě k dispozici.

Je však potřeba určité plánování a design za účelem zmírnění úsilí nutného ke sloučení a publikování výstupů výzkumu open source produktů. Výstupy z výzkumu často nejsou připravené k produkci, jsou specifické pro doménu a mohou mít dopad na platformu, která narušuje existující funkčnost. Snadná rozšiřitelnost open source aplikací, např. použití pluginů, je důležitá vlastnost umožňující rychlé zahrnutí nových výzkumných výstupů do hlavních verzí, zejména pokud jsou funkce specifické pro doménu. Také použití rozhraní API

[18] http://wiki.opf-labs.org/display/SP/Repository+APIs

pro integraci při užití standardů je flexibilní a vytvořené pro co nejmenší dopad na datový model a může mít zásadní význam pro pro koncového uživatele.

Pokud je dopad na platformu nevyhnutelný, je důležité identifikování rizik v počátečním stádiu, přijetí jejich existence a sjednocení plánu pro open source projekt s požadavky výzkumu. V těchto případech sleduje zajištění vývoje zájmy komunity a je nutné udržovat blízký kontakt s komunitou, aby se zajistilo, že vývojový trend bude vyhovovat komunitě a reálnému využití.

Ve vývoji softwarových trendů open source a ještě více ve výzkumu, jsou změny nevyhnutelné a dokonce nutné, ale plánování, návrhy a komunikace jsou potřebné k udržení projektu na správné cestě a pro přijetí komunitou.

7. PODĚKOVÁNÍ

Tato práce byla částečně podpořena projektem E-ARK. Projekt E-ARK je ko-financován Evropskou komisí pod CIP-ICT-PSP-2013-7 (Schvalovací číslo grantu 620998)

8. REFERENCE

[1] Reference Model for an Open Archival Information System (OAIS). Technical report, Consultative Committee for Space Data Systems (CCSDS), 2002.

[2] C. Becker, H. Kulovits, A. Rauber, and H. Hofman. Plato: A service oriented decision support system for preservation planning. In *Proceedings of the 8th ACM IEEE Joint Conference on Digital Libraries (JCDL 2008)*, 2008.

[3] S. Chaudhuri and U. Dayal. An overview of data warehousing and olap technology. *SIGMOD Rec.*, 26(1):65–74, Mar. 1997.

[4] M. Ferreira, L. Faria, M. Hahn, and K. Duretec. Report on compliance validation. Technical Report MS63, SCAPE project, 2014.

[5] ISO. Space Data and Information Transfer Systems—Audit and Certification of Trustworthy Digital Repositories. ISO 16363:2012, International Organization for Standardization, Geneva, Switzerland, 2012.

[6] M. Kraxner, M. Plangg, K. Duretec, C. Becker, and L. Faria. The SCAPE planning and watch suite: supporting the preservation lifecycle in repositories. In *iPRES 2013 - 10th International Conference on Preservation of Digital Objects*, 2013.

[7] C. Lagoze and H. V. de Sompel. The open archives initiative: Building a low-barrier interoperability framework. *Digital Libraries, Joint Conference on*, 0:54–62, 2001.

[8] J. Loeliger and M. McCullough. *Version Control with Git: Powerful tools and techniques for collaborative software development.* " O'Reilly Media, Inc.", 2012.

[9] PREMIS Editorial Committee. Data Dictionary for Preservation Metadata: PREMIS version 2.0. Technical report, Mar. 2008.

[10] S. Strodl, C. Becker, R. Neumayer, and A. Rauber. How to choose a digital preservation strategy: Evaluating a preservation planning procedure. In *Proceedings of the 7th ACM IEEE Joint Conference on Digital Libraries (JCDL'07)*, pages 29–38, New York, NY, USA, June 18-23 2007. ACM Press.

[11] The Library of Congress. Common Query Language. http://www.loc.gov/standards/sru/cql/ [Online; accessed 21-October-2014].

[12] The Library of Congress. Encoded Archival Description. http://www.loc.gov/ead/ [Online; accessed 21-October-2014].

[13] The Library of Congress. Metadata Encoding & Transmission Standard. http://www.loc.gov/mets/ [Online; accessed 21-October-2014].

[14] The Library of Congress. Search/retrieve via url. http://www.loc.gov/standards/sru/ [Online; accessed 21-October-2014].

[15] L. Torvalds and J. Hamano. Git: Fast version control system. *URL http://git-scm. com*, 2010.

[16] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, J. Bhagat, K. Belhajjame, F. Bacall, A. Hardisty, A. Nieva de la Hidalga, M. P. Balcazar Vargas, S. Sufi, and C. Goble. The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. *Nucleic Acids Research*, 41(W1):W557–W561, 2013.