UNIVERSITATEA TEHNICĂ "GHEORGHE ASACHI" DIN IAŞI

R E C T O R A T U L

Către

_____

_____

Vă facem cunoscut că, în ziua de **23.10.2020** la ora **10:00** link-ul **https://meet.google.com/rch-sxtm-csd**, va avea loc **susţinerea publică online** a tezei de doctorat intitulată:

"INTEROPERABILITATEA SISTEMELOR DE MANAGEMENT AL CERCETĂRII (ADVANCES IN DIGITAL RESEARCH REPOSITORIES)"

elaborată de domnul **PĂNESCU ADRIAN-TUDOR** în vederea conferirii titlului ştiinţific de doctor.

Comisia de doctorat este alcătuită din:

1. Prof. univ. dr. ing, Matcovschi Mihaela, Universitatea Tehnică „Gheorge Asachi" din Iaşi	preşedinte
2. Prof. univ. dr. ing, Manta Vasile Ion, Universitatea Tehnică „Gheorge Asachi" din Iaşi	conducător de doctorat
3. Prof. univ. dr. ing, Mocanu Mariana Ionela, Universitatea „Politehnica" din Bucureşti	referent oficial
4. Prof. univ. dr. ing, Ciocârlie Horia, Universitatea din Craiova	referent oficial
5. Prof. univ. dr. ing, Păstrăvanu Octavian, Universitatea Tehnică „Gheorge Asachi" din Iaşi	referent oficial

Cu această ocazie vă invităm să participaţi la susţinerea publică a tezei de doctorat.

RECTOR,

Prof.univ.dr.ing. DAN CAŞCAVAL

Secretar universitate,

Ing. Cristina Nagîţ

Rezumat teză de doctorat

DEZVOLTAREA SISTEMELOR DE STOCARE A PUBLICAȚIILOR ȘTIINȚIFICE

Candidat:

ADRIAN-TUDOR PĂNESCU

Îndrumător:

VASILE ION MANTA

Facultatea de Automatică și Calculatoare

Universitatea Tehnică "Gheorghe Asachi" din Iaşi

România, 2020

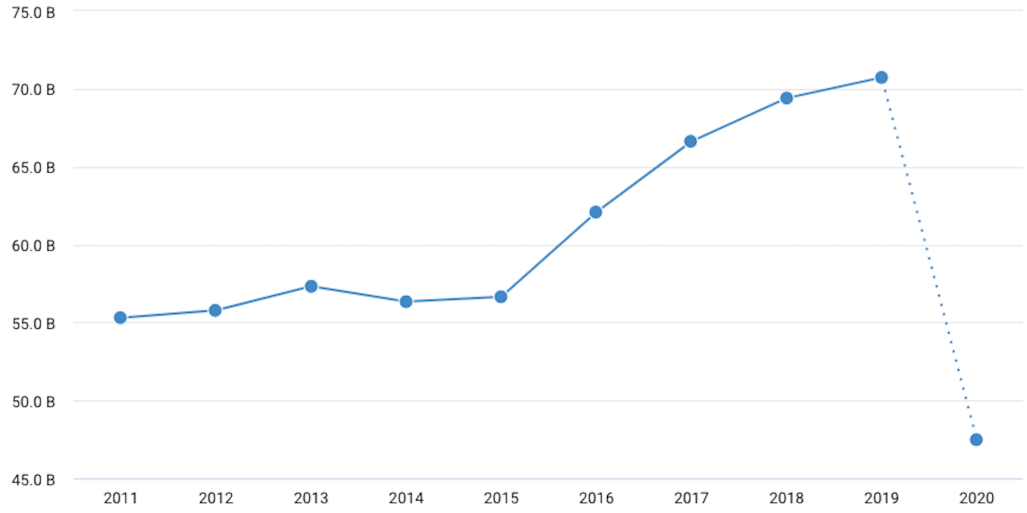
Bibliotecile, colecții structurate de resurse informaționale, joacă un rol esențial în conservarea memoriei colective a civilizației umane. Resursele pe care acestea le administrează iau diferite forme, atât la nivel structural (manuscrise, materiale audiovizuale, articole), cât și la nivel semantic (cronici istorice, publicații științifice, beletristică), și sunt puse fie la dispoziția publicului larg sau a unei comunități restrânse, cum este personalul unei universități.

Odată cu apariția *World Wide Web (WWW)* la sfârșitul secolului XX, bibliotecile au implementat noile tehnologii *online*, dând naștere astfel bibliotecilor *digitale*. Progresia rapida a acestora a fost la rândul său influențată, printre altele, de evoluția proceselor specifice cercetării științifice, evoluție marcată de o serie de schimbări și evenimente.

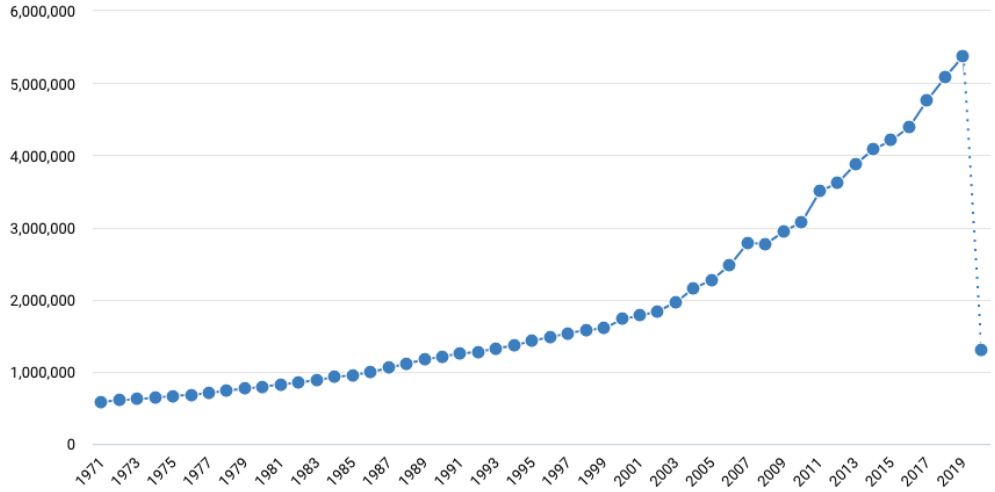
Un prim factor este reprezentat de sporirea finanțării, ce a dus la o creștere direct proporțională a volumului de publicății științifice (Fig. 1); pentru biblioteci aceasta înseamna o creștere a fluxului de materiale noi ce trebuie administrate și diseminate.

Un alt factor important este reprezentat de mișcarea *Open Access (OA)*. Liberalizarea diseminării conținutului, generată de apariția WWW și a bibliotecilor digitale, a pus semne de întrebare privind modelul de afaceri al marilor edituri din zona științifică, precum Elsevier sau Springer Nature, care nu mai puteau justifica integral costurile percepute pentru editare și tipărire. OA reprezintă una din variantele propuse pentru schimbarea acestui model, fiind urmărită suportarea costurilor nu de către cei care doresc să acceseze publicațiile, ci de autori sau, mai realist, de către finanțatorii acestora. OA poate fi implementat în diferite moduri, cum este tipul *gold*, în care se plătește o taxa pentru publicare, sau tipul *green*, în care autorii sunt liberi să facă disponibilă online o variantă proprie a publicației utilizând, de exemplu, biblioteca digitală a instituției căreia sunt afiliați. Aceasta a generat atât creșterea volumului de materiale administrate de bibliotecile digitale (2 milioane de publicații OA *green* au fost stocate până acum conform [HHC19]), cât și necesitata implementării de funcționalități care să suporte noul tip de conținut (de exemplu, facilitățile de *embargo*).

OA este unul din fenomenele ce compun un curent mai amplu, *Open Science*, a cărui țintă este să facă publice toate materialele provenite din activitățile de cercetare; în afara de publicațiile tradiționale, aici sunt incluse, printre altele, și seturile de date, *software*-ul folosit pentru analiză sau experimente, sau figuri și modele. Din



(a) Evoluția totalului granturilor de cercetare, în lire sterline.



(b) Evoluția numărului total de publicații științifice.

Figure 1: Evoluția activităților de cercetare științifică, conform `https://app.dimensions.ai`.

nou, bibliotecile digitale trebuie să fie adaptate pentru a suporta aceste noi tipuri de conținut.

Curentul Open Science a fost generat de așa numita *criză de replicare* a rezultatelor științifice, care a scos la suprafața o serie de publicații ce prezintă rezultate ce nu pot fi replicate independent (exemple în [BL17; ENK16]), fie datorită unor erori de analiză, fie datorită falsificării datelor suport sau a rezultatelor analitice.

Noile politici impuse ca urmare a acestei crize stipulează necesitatea publicării nu doar a rezultatelor, cât și a materialelor ce au fost folosite pentru generarea acestora[Dir17; Nat], aceasta adăugând presiune suplimentară asupra bibliotecilor ce trebuie să prezinte toate artefactele într-un mod consolidat și elocvent.

Cercetarea prezentată în această teză de doctorat s-a concentrat pe analizarea noilor provocări din zona bibliotecilor digitale și pe formalizarea și implementarea de soluții pentru rezolvarea acestora. Toate rezultatele obținute au fost recenzate și publicate după cum urmează:

1. Adrian-Tudor Pănescu, Tibor Šimko și Christine Vanoirbeek. ,,Targeted Annotation of Scientific Literature and Data Resources in Invenio Digital Libraries". În: *Proceedings of Open Repositories* (2014). URL: http://hdl.handle.net/10024/97585.

2. Adrian-Tudor Pănescu și Vasile Manta. ,,Current Issues In Research Output Management". În: *Buletinul Institutului Politehnic din Iaşi, secțiunea Automatică și Calculatoare* (Dec. 2016).

3. Adrian-Tudor Pănescu și Vasile Manta. ,,RDF-based workflows for the figshare research data repository". În: *Proceedings of the 21st International Conference on System Theory, Control and Computing (ICSTCC)* (2017), pag. 860–865. DOI: 10.1109/ICSTCC.2017.8107145.

4. Adrian-Tudor Pănescu și Vasile Manta. ,,Smart Contracts for Research Data Rights Management over the Ethereum Blockchain Network". În: *Science & Technology Libraries* 37.3 (Jul. 2018), pag. 235–245. DOI: 10.1080/0194262X.2018.1474838.

5. Christopher Frederick Isambard Blumzon și Adrian-Tudor Pănescu. ,,Data Storage". In *Good Research Practice in Non-Clinical Pharmacology and Biomedicine*. Ed. Anton Bespalov, Martin C. Michel, și Thomas Steckler. 2020. pag. 277–297. DOI: 10.1007/164_2019_288.

6. Adrian-Tudor Pănescu, Teodora-Elena Grosu și Vasile Manta. ,,SLAM: An ETL System for Performing Digital Library Migrations". Acceptată pentru publicare în *Information Technology and Libraries*. URL: https://ejournals.bc.edu/index.php/ital.

Bibliotecile digitale de tip *repository*

Prima parte a cercetării s-a concentrat pe explorarea ecosistemului curent al bibliotecilor digitale. În primul rând este analizat un tip special de platformă, și anume bibliotecile de tip *repository*; sunt explorate două tipuri, și anume cele *instituționale* și cele destinate seturilor de date.

Repository-urile instituționale sunt definite în literatură ca *„seturi de servicii pe care universitățile le oferă personalului pentru administrarea și diseminarea materialelor digitale”*[Lyn03]; funcțiile pe care acestea le îndeplinesc pot fi împărțite în trei categorii:

1. Selecție, clasare și catalogare
2. Conservare
3. Diseminare

Un aspect notabil al definiției precedente este faptul că aceasta nu prescrie tipurile de artefacte pe care un repository trebuie să le stocheze; deși în mod tradițional acest tip de platformă este utilizat pentru publicații (articole, cărți, dizertații), în ultimii ani, ca urmare a fenomenelor descrise anterior, s-a recurs și la stocarea altor tipuri de informații provenite din cercetare, precum seturi de date sau software științific.

La nivel practic, aceasta tranziție poate fi materializata fie prin implementarea de noi funcționalități în soluții deja existente sau prin dezvoltarea de noi platforme de tip repository. Pentru a aprofunda înțelegerea cerințelor curente, cercetarea a considerat ca și studiu de caz repository-urile dedicate administrării seturilor de date. Spre deosebire de un repository instituțional, acest tip de platformă prezintă o serie de cerințe particulare.

În primul rând, un repository dedicat datelor științifice trebuie să fie capabil să manipuleze fișiere complexe, atât ca și dimensiuni (care pot fi de ordinul *petabytes*) cât și ca structură (ierarhii de directoare cu mii de fișiere). Sistemele de stocare trebuie să asigure durabilitate și redundanță, pentru a asigura că datele din spatele studiilor științifice, care de cele mai mult ori sunt stocate într-o singura locație, nu pot fi pierdute ireversibil. În plus, spre deosebire de publicațiile tradiționale, care nu sunt strâns legate de un anumit format electronic, pentru date trebuie să existe

procese ce asigură ca ele pot fi utilizate și după ce tehnologia folosită inițial pentru generarea lor a fost scoasă din uz.

Un al doilea aspect important este reprezentat de *metadate*, informații ce descriu seturile de date. Dacă pentru publicații acestea tind să fie standardizate, indiferent de domeniu de cercetare sau zona geografică, pentru date poate fi observată o eterogenitate generată de particularitățile fiecărei arii și tip de experiment științific. Metadatele au însa un rol crucial în cazul datelor științifice, reprezentând prima metodă prin care un observator extern poate înțelege sintaxa și semantica setului descris.

În acest sens, comunitatea științifica a conceput principiile *FAIR*, ce descriu proprietățile pe care seturile de date trebuie să le îndeplinească, prin intermediul metadatelor și a platformei pe care sunt stocate:

1. *Findability*: un set de date trebuie să poată fi ușor găsit utilizând facilitățile platformelor online. Aceasta presupune ca metadatele să fie cât mai descriptive, să utilizeze un format structurat, ușor de citit atât de oameni cât și de sisteme automate, și să includă identificatori unici și durabili, precum cei de tipul *Digital Object Identifier (DOI)*.

2. *Accesibility*: seturile de date trebuie să fie ușor de accesat și descărcat; în plus, metadatele trebuie să rămână accesibile și în eventualitatea în care resursa descrisă nu mai este publică.

3. *Interoperability*: datele trebuie să poată fi ușor migrate între sisteme, acest lucru fiind asigurat de caracterul structurat și complet al metadatelor cât și de facilitățile de transfer oferite de repository.

4. *Reusable*: după cum a fost demonstrat de către criza de replicare, datele trebuie să poată fi reutilizate pentru a desfășura experimente similare. În acest sens metadatele trebuie să prezinte o imagine cât ma detaliată, incluzând detalii tehnice privind instrumentele și procesele ce trebuie utilizate pentru întrebuințarea datelor.

Al treilea aspect considerat vizează cerințele de ordin legal ce dictează stocarea și utilizarea datelor. Pe de o parte, politici precum Regulamentul General privind Protecția Datelor (GDPR), sau legislația ce formalizează protejarea subiecților din

studiile științifice prevăd reguli clare în legătura cu modul în care datele trebuie anonimizate sau pseudoanonimizate, stocate și diseminate. Pe de altă parte, spre deosebire de publicații, pentru care drepturile de autor și reutilizarea sunt aspecte clar stipulate, pentru date există un număr suplimentar de dimensiuni ce trebuie considerate; de exemplu, suita de licențe *Creative Commons (CC)*[1] include o serie de clauze ce pot fi utilizate în combinații cu diferite grade de restricționare a reutilizării. O aplicație de tip repository nu trebuie doar să ofere utilizatorilor facilitățile necesare conformării cu aceste aspecte, ci în plus trebuie să le implementeze și prezinte într-un mod ușor de înțeles, chiar și în lipsa cunoștințelor de ordin legal.

Înțelegând cerințele pe care trebuie să le îndeplinească atât repository-urile instituționale cât și cele destinate noilor tipuri de artefacte provenite din cercetare, pot fi deduse și anumite cerințe pe care trebuie să le îndeplinească o platformă capabilă să administreze orice tip de produs științific, indiferent de dimensiuni, format sau domeniu. În continuarea tezei sunt considerate trei astfel de funcționalități, fiind prezentate soluții practice ce utilizează tehnologii din diverse arii de cercetare ale informaticii, cu scopul general de a schița imaginea noii generații de platforme de tip repository.

Administrarea drepturilor de autor folosind *smart contracts*

După cum a fost menționat anterior, administrarea drepturilor de autor și reutilizare a artefactelor științifice neconvenționale aduce o serie de provocări noi; de exemplu, în [Eyn+16] este identificat faptul ca autorii pot fi rezervați în a publica datele suport ale unui studiu științific cu scopul evitării situațiilor în care acestea sunt folosite pentru a publica rezultate noi, fiind pierdute astfel oportunități de publicare, importante în cariera academică.

O soluție pentru aceste probleme este reprezentată de *smart contracts*, o metodă de a defini contracte tradiționale utilizând un limbaj de programare. Deși formalizata în 1997[Sza97], aceasta tehnica a fost popularizată recent, odată cu apariția tehnologiilor de tip *blockchain*. Pe de o parte, aplicațiile de tip blockchain au nevoie de o metoda de a descrie tranzacții, mai ales în domeniul financiar, iar pe de altă parte, oferă atât o metoda de a stoca în mod trasabil și verificabil contractele, cât și infrastructura necesara executării acestora.

[1] `https://creativecommons.org/`

Pentru administrarea drepturilor de autor a fost definit un contract care la nivel conceptual are trei pași:

1. Publicare: în acest pas autorii unui, spre exemplu, set de date științifice face cunoscut faptul că acesta a devenit public și poate fi descărcat și refolosit. La acest pas pot fi stipulați și termenii sub care setul de date poate fi folosit, autorul având libertatea să aleagă orice condiții (de exemplu, concluzii noi obținute din analiza setului de date nu pot fi publicate în următorul an). Acești termeni, alături de alte metadate, vor fi înregistrate pe blockchain, în mod durabil și verificabil.

2. Cedare: în acest pas setul de date este transferat de la autor la un alt utilizator; acest fapt este din nou înregistrat pe blockchain, fiind consemnați și termenii sub care se pot reutiliza datele (definiți fie la pasul anterior fie personalizați în acesta, oferind și mai mult control autorilor).

3. Reutilizare: în acest pas este înregistrat faptul că pornind de la setul de date inițial au fost generate și trimise spre publicare noi rezultate. Înregistrarea acestui pas pe blockchain este crucială în execuția smart contractului, întrucât astfel poate fi verificat faptul că termenii de reutilizare au fost respectați.

La nivel practic, acest contract a fost implementat folosind *Solidity*[2], limbajul de programare standard al platformei de blockchain *Ethereum*[3]. Implementarea a permis analizarea fezabilității integrării unei astfel de funcționalități pe o platforma de tip repository, demonstrând că oferă o mai mare flexibilitate în alegerea termenilor de reutilizare și un control mai strict în ceea ce privește respectarea acestora. În același timp, succesul unui astfel de sistem este strâns legat de participarea unui număr cât mai mare de cercetători și edituri la blockchain, pentru a garanta executarea ultimului pas descris anterior, în care se validează reutilizarea corecta a rezultatelor științifice. Cu toate acestea, tehnologia smart contract reprezintă o soluție viabilă pentru rezolvarea unei probleme lipsite momentan de alte soluții formale.

[2] `https://solidity.readthedocs.io/`
[3] `https://ethereum.org/`

Modelarea metadatelor folosind RDF

Odată ce conținutul platformelor de tip repository a evoluat de la publicații tradiționale la neconvenționale, acestea s-au lovit de o nouă provocare, și anume modul în care stochează metadatele. Dacă pentru publicații, de cele mai multe ori, sunt folosite vocabulare statice, existente în industrie de câteva decenii (de exemplu, *Dublin Core Metadata Terms(DCMT)*[4] a fost definit inițial în 2002), pentru seturi de date sau software științific acest lucru nu este întotdeauna posibil. Acest fapt se datorează fie variației mari în ceea ce privește câmpurile necesare descrierii publicațiilor[5], fie evoluției câmpurilor bibliografice în sine ca urmare a noilor realități din domeniul cercetării (de exemplu, în taxonomia CRediT[6] definiția câmpului *autor* este extinsă pentru a ilustra detaliat rolul fiecărei persoane ce a contribuit la un proiect de cercetare).

La nivel înalt, această problemă poate fi rezolvată prin utilizarea unui model bibliografic flexibil, care să permită atât combinarea mai multor vocabulare în descrierea unei publicații, cât și definirea de câmpuri noi, aplicabile fie și pentru o singura intrare dintr-un repository. Un model standardizat ce oferă aceasta flexibilitate este *Resource Description Framework (RDF)*[7], în care orice informație este transpusă ca și o *propoziție*, sau *tripletă*, cu subiect, predicat și adverb; propozițiile pot fi vizualizate ca și grafuri orientate, similare cu cel din Fig. 2. După cum se poate observa în aceasta figură, termenii folosiți pot fi definiți în ontologii consacrate; astfel, metadatele ce descriu o intrare bibliografică pot fi modelate ca și grafuri complexe de propoziții ce utilizează termeni din ontologii disjuncte.

Ca și experiment practic, aceasta parte a cercetării a considerat soluția repository Figshare[8], ce folosește un model relațional pentru stocarea metadatelor, și a conceput un nou sistem bazat pe RDF. În afara de dezavantajele prezentate anterior, modelul relațional al acestui repository îngreuna diseminarea metadatelor în diversele formate folosite de utilizatorii săi, majoritatea bazate pe *Extensible Markup*

[4] `https://www.dublincore.org/specifications/dublin-core/dcmi-terms/`

[5] De exemplu, un model al unui artefact arheologic poate avea ca și metadate informații despre locația geografică unde a fost descoperit sau era și societatea cărora le aparține, în timp ce un set de date generat în cadrul unui studiu clinic nu va conține astfel de informații

[6] `https://casrai.org/credit`

[7] `https://www.w3.org/RDF/`

[8] `https://fighsare.com`

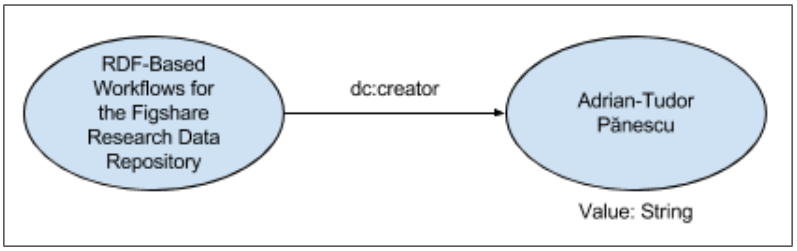


Figure 2: O propoziție RDF vizualizată ca și un graf orientat. Predicatul folosește definiția din ontologia Dublin Core, după cum este marcat de prefixul *dc:*.

Language (XML)[9].

RDF are ca și format de serializare standard RDF-XML, în care toate tripletele sunt exportate într-un document XML. Pornind de la aceasta serializare și folosind Extensible Stylesheet Language Transformations (XSLT), o metodă de transformare a documentelor XML, poate fi obținut oricare dintre formatele utilizate de repository. Un avantaj important al aceste abordări, subliniat de implementarea experimentală, este ca transformările XSLT pot fi mult mai ușor modificate decât modulele curente ale aplicației, ce sunt strâns cuplate în arhitectura generală a platformei.

Deși implementarea experimentală a dovedit avantajele acestei tehnici de modelare a metadatelor, integrarea în sistemul de producție implică o serie de provocări. În primul rând, o parte din logica aplicației nu poate folosi modelul RDF (de exemplu, logica legată de documente confidențiale sau sub embargo) și din acest motiv baza de date relațională trebuie să continue să funcționeze în paralel cu depozitul RDF. În plus, toate intrările bibliografice curente trebuie migrate către modelul RDF, această operațiune prezentând o complexitate ridicată, mai ales în cazul unui sistem ca Figshare ce administrează peste patru milioane de documente și care nu poate fi oprit pentru operațiuni conform politicilor sale de funcționare. Acest ultim aspect este studiat în următoarea parte a cercetării. În ciuda acestor dificultăți, implementarea experimentală e dovedit avantajele principale ale modelului RDF, și anume flexibilitatea în definirea metadatelor și facilitarea dezvoltării de noi funcționalități pentru un repository.

[9] `https://www.w3.org/XML/`

Migrarea datelor între biblioteci digitale

Migrarea conținutului între biblioteci digitale este o operațiune de complexitate ridicată întrucât trebuie să se asigure faptul că nu vor exista pierderi sau coruperi de informații. În plus, fiind dată natura mediului online și deservirea de către o aceeași platformă a mai multe comunități dispersate geografic, de cele mai multe ori migrarea trebuie să aibă loc fără întreruperea sau perturbarea funcționarii bibliotecii.

Necesitatea unei astfel de migrări poate apărea din mai multe motive; unul dintre acestea, după cum a fost menționat anterior, este reprezentat de scoaterea din uz a unei platforme, atunci când aceasta nu mai poate asigura ficționalitățile necesare utilizatorilor. Alte motive pot fi reprezentate de schimbări administrative, financiare sau operaționale la nivelul comunității care deține biblioteca digitală, sau dorința de a inventaria și cura corpusul bibliografice.

În aceasta parte a cercetării a fost considerată dezvoltarea unui sistem de migrare a datelor intre repository-uri; la nivel înalt, au fost considerate trei principii pe care acesta trebuie să le îndeplinească:

1. Posibilitatea de reutilizare: deși în literatura sunt prezentate soluții pentru astfel de migrări, nu exista o varianta ce poate fi refolosită pentru a executa multiple astfel de operațiuni. Așadar, noul sistem trebuie să poată fi reutilizat cu modificări minimale indiferent de tipul de repository din care se extrag sau în care se depun datele.

2. Idempotența: sistemul trebuie sa permită rularea procedurilor de migrare ori de cate ori este nevoie pentru, de exemplu, corecturi sau îmbogățiri ale informațiilor, fără a fi corupte sau duplicate intrările bibliografice in repository-ul final.

3. Capacitate de mitigare a erorilor: procedurile de migrare trebuie sa ruleze fără supravegherea unui operator, având implementate mecanisme de raportare și tratare a excepțiilor ce pot fi cauzate, printre altele, de eterogenitatea datelor bibliografice, erori de programare sau probleme ale canalelor de comunicare online.

Sistemul propus, denumit *Stateful Library Analysis and Migration (SLAM)*, este conceput sub forma unui proces *extract, transform, load (ETL)* și compus dintr-o

colecție de module disjuncte. În primul rand, atât pentru repository-ul din care sunt extrase datele bibliografice cât și pentru cel în care vor fi depozitate sunt implementați doi *conectori*; scopul acestora este de a izola logica specifică diferitelor tipuri de repository, nefiind astfel necesară modificarea nici unei alte componente pentru refolosirea sistemului.

A doua componentă este cea destinată analizării metadatelor; aceasta operațiune, deși opțională în cadrul unei migrări, este utilă în depistarea erorilor și neconcordanțelor ce pot apărea mai ales în cazul colecțiilor vaste, cu materiale publicate de-a lungul mai multor decenii. Aceasta componentă include un *crosswalk*, un tip special de document ce definește modul în care metadatele din repository-ul sursă trebuie să fie procesate înainte de a fi depozitate în repository-ul destinație. În plus, este inclus un motor de căutare și o interfața grafică pentru analiză, implementate folosind Elastic Stack[10], ce facilitează verificarea transformărilor ce urmează a fi operate și identificarea potențialelor probleme.

Cea de-a treia componentă este motorul de execuție ce va transfera fiecare intrare bibliografica și fișierele asociate pe repository-ul destinație. Aceasta este implementata sub forma unui automat secvențial, pentru fiecare intrare bibliografica fiind executați un număr de pași predefiniți (aplicarea transformărilor din *crosswalk*, încărcarea fișierelor etc.), execuție ce este serializată în mod persistent. Această implementare asigura idempotența migrării, întrucât există o evidență a tuturor pașilor deja executați. În plus, deoarece pentru fiecare intrare bibliografica se execută o nouă instanță a automatului, este facilitată și mitigarea erorilor și derularea nesupravegheată a migrării; operatorii vor verifica la finalul întregii operațiuni dacă au exista erori, și după îndepărtarea cauzelor acestora vor executa din nou procedura de migrare, fiind considerați doar pașii ce nu au fost deja parcurși anterior.

SLAM a fost testat prin derularea a cinci migrări pe parcursul unui an, totalizând peste 80000 de intrări bibliografice transferate; evaluarea a fost realizată prin urmărirea celor trei principii enunțate anterior. În ceea ce privește posibilitatea de reutilizare, între migrări au fost modificate, conform așteptărilor, doare componentele de conectare la repository-uri și crosswalk-ul; timpul necesar acestor schimbări a fost în medie de două săptămâni pentru un singur programator. Aceasta durată este o îmbunătățire față de migrările descrise în literatură, care au înregistrat

[10] `https://www.elastic.co/elastic-stack`

durate de ordinul lunilor sau anilor[Tuy+18].

Capacitatea de mitigare a erorilor a fost testată *involuntar*, toate migrările întâmpinând probleme pentru anumite intrări bibliografice. De exemplu, pentru una din instante 300 de astfel de intrări, dintr-un total de 15000, nu au putut fi migrate ca urmare a unei erori în definiția crosswalk-ului. Aceste probleme au fost înregistrate și raportate operatorului la finalul primei execuții; după rezolvarea problemei, procesul de migrare a fost executat din nou, fiind executați doar pașii lipsă pentru cele 300 de întrări bibliografice și finalizând cu succes migrarea. Aceasta a demonstrat și idempotența sistemului, întrucât, de exemplu, nu au fost generate duplicate ale unor intrări deja migrate.

Aceasta evaluare a demonstrat viabilitatea arhitecturii sistemului de migrare și a evidențiat și posibilele direcții de îmbunătățire. În primul rând, sistemul de analiză trebuie extins pentru a fi ușor utilizabil și de către personalul fără abilitați tehnice, dar care deține cunoștințe despre particularitățile colecției migrate sau a proceselor din biblioteca sursă, fiind cel mai în măsura să identifice și să propună soluții pentru eventualele inconsistențe. De asemenea, sistemul poate fi îmbunătățit pentru a extrage metadate și din alte surse externe, cum sunt Elesevier Scopus[11] sau Crossref[12], acestea putând fi folosite pentru corectarea și îmbogățirea informațiilor din repository; acest tip de automatizare este benefic mai ales în cazul colecțiilor vaste, în care nu este viabilă verificarea și corectarea manuală a fiecărei intrări bibliografice.

[11] `https://www.scopus.com`

[12] `https://www.crossref.org/`

Remarci finale

Cercetarea de fata s-a concentrat pe studiul librăriilor digitale în general și a repository-urilor în particular, examinând provocările pe care acestea le întâmpina în fața schimbărilor ce au loc în domeniul științific. Printre aceste provocări se număra creșterea volumului de publicații și schimbările generate de fenomenul Open Science ca urmare a crizei de replicare.

Aceasta analiză a permis identificarea unui număr de arii pentru care pot fi propuse soluții tehnologice menite să îmbunătățească funcționarea bibliotecilor digitale și implementarea de noi facilitați pentru administrarea și diseminarea noilor tipuri de conținut științific (seturi de date, modele tridimensionale, software științific). În plus, cercetarea oferă o viziune asupra modului în care trebuie concepută arhitectura noii generații de repository-uri, și anume ca și o colecție de module disjuncte, fiecare specializat pe o anumita funcționalitate a bibliotecii și administrat de personalul cel mai în măsura, potrivit experientei și cunoștințelor necesare.

Sumarizând, aceasta cercetare aduce următoarele patru contribuții:

- A fost desfășurata o analiza profundă a bibliotecilor digitale, concentrată pe înțelegerea provocărilor curente; ca și studiu de caz au fost considerate repository-urile destinate stocării seturilor de date suport pentru studii științifice.

- Este propus un sistem de licențiere și administrare a drepturilor de autor pentru publicații științifice, bazat pe smart contracts; acesta se remarca prin flexibilitate în definirea clauzelor și trasabilitate.

- Este prezentată o implementare experimentală a unui nou sistem de administrare a metadatelor, bazat pe RDF; avantajele acestuia sunt reprezentate de flexibilitatea în utilizarea a multiple vocabulare pentru definirea câmpurilor bibliografice și posibilitatea de serializare a metadatelor în diverse formate structurate.

- A fost implementat un sistem de migrare a datelor și informațiilor bibliografice între repository-uri, ce include capabilități de analiză a metadatelor și se remarcă prin capacitățile de execuție nesupervizată și posibilitatea reutilizării componentelor pentru migrări între diferite tipuri de platforme.

Referințe

[BL17] Jennifer A. Byrne and Cyril Labbè. "Striking similarities between publications from China describing single gene knockdown experiments in human cancer cell lines". In: *Scientometrics* 110 (Mar. 2017), pp. 1471–1493. DOI: 10.1007/s11192-016-2209-6 (cit. on p. 2).

[Dir17] Directorate-General for Research & Innovation. *2020 Programme – Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020*. Tech. rep. 3.2. European Commission, 2017. URL: https://web.archive.org/web/20180826235248/http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf (cit. on p. 3).

[ENK16] Anders Eklund, Thomas E. Nichols, and Hans Knutsson. "Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates". In: *Proceedings of the National Academy of Sciences* 113 (28 June 2016), pp. 7900–7905. DOI: 10.1073/pnas.1602413113 (cit. on p. 2).

[Eyn+16] Veerle Van den Eynden et al. *Towards Open Research – Practices, experiences, barriers and opportunities*. Tech. rep. 1. Wellcome Trust, Oct. 2016. DOI: 10.6084/m9.figshare.4055448.v1 (cit. on p. 6).

[HHC19] Daniel Hook, Mark Hahnel, and Ian Calvert. *The Ascent of Open Access*. Tech. rep. 2. Digital Science, Jan. 2019. DOI: 10.6084/m9.figshare.7618751.v2 (cit. on p. 1).

[Lyn03] Clifford A. Lynch. "Institutional Repositories: Essential Infrastructure For Scholarship In The Digital Age". In: *Libraries and the Academy* 3.2 (Apr. 2003), pp. 327–336. DOI: 10.1353/pla.2003.0039 (cit. on p. 4).

[Nat] National Institutes of Health. *NIH Public Access Policy Details*. URL: https://web.archive.org/web/20180421191423/https://publicaccess.nih.gov/policy.htm (cit. on p. 3).

[Sza97] Nick Szabo. "Formalizing and Securing Relationships on Public Networks". In: *First Monday* 2 (9 Sept. 1997). DOI: 10.5210/fm.v2i9.548 (cit. on p. 6).

[Tuy+18] Steve Van Tuyl et al. "Are we still working on this? A meta-retrospective of a digital repository migration in the form of a classic Greek Tragedy (in extreme violation of Aristotelian Unity of Time)". In: *code{4}lib* (41 Aug. 2018). URL: https://journal.code4lib.org/articles/13581 (cit. on p. 12).