Dosarul „Astra Data Mining” (paginile 1-43) este coordonat de
Ștefan Baghiu și Vlad Pojoga

Arhivele romanului românesc și posibilități de digitizare

Andreea COROIAN-GOLDIȘ, Daiana GÂRDAN, Emanuel MODOC, Teodora SUSARENCO, David MORARIU, Cosmin BORZA

Universitatea „Babeș-Bolyai” din Cluj-Napoca, Facultatea de Litere;
Universitatea „Lucian Blaga” din Sibiu, Facultatea de Litere și Arte
Babeș-Bolyai University of Cluj-Napoca, Faculty of Letters;
Lucian Blaga University of Sibiu, Faculty of Letters and Arts
Personal e-mail: andreeacoroian@gmail.com, alexandra.gardan@gmail.com,
ermodoc@gmail.com, susarenco98@gmail.com, david.morariu@ulbsibiu.ro, cosmi_borza@yahoo.com

The Archives of the Romanian Novel and Digitization Possibilities

The present study assesses the existing archives that gather and preserve in digital versions the Romanian novel published in the 19 th century. Digitalization projects, both in Romania and in Europe, are the main subjects of this research. Adding to this, we share the particular and common experiences of the *Astra Data Mining: The Digital Museum of the 19th Century Romanian Novel* in terms of the process of digitizing the 157 novels in our corpus. Other details such as the specific editions digitized, the process of selection, as well as any difficulties met along the way are also provided. At the same time, the activities of our project are also corroborated with similar European projects. In order to address our digitization practice, we also explore a theoretical and methodological framework borrowed by Stanford Literary Lab in terms of conceptualizing what the American researchers call ”the published”, ”the archive” and ”the corpus”.

Keywords: Romanian literature, 19th century novel, canon/archive, digital humanities

Digitizarea la nivel național și european: definiție, proiecte, provocări

Definiția unanim acceptată, la nivel internațional, a digitizării este enunțată astfel: „Digitizarea presupune captura digitală, transformarea din formă analogă în formă digitală, descrierea și reprezentarea obiectelor de patrimoniu și a documentației referitoare la acestea, procesarea, asigurarea accesului la conținutul digitizat și prezervarea pe termen lung”[1]. Digitizarea constă, așadar, în transpunerea unui document din format tradițional în format digital, prin diverse mijloace.

Referindu-ne la spațiul românesc, Biblioteca Națională Română definește digitizarea astfel:

Digitizare reprezintă procedeul prin care informația este capturată în format digital (imagine, document text, fișier audio, etc.) cu ajutorul unui echipament tehnic digital (cameră digitală, scanner, etc.). Când vorbim despre digitizarea documentelor, de cele mai multe ori ne referim la imaginea paginii capturată de un astfel de echipament – pur și simplu o poză a documentului – sau o versiune full-text, în care documentul este stocat folosind caractere text/ scrise. Forma neprocesată a documentului (*plain-text*),

reprezintă varianta integrală a documentului, folosind caractere ASCII sau Unicode, pentru acestea existând posibilitatea efectuării unei căutări în text (cuvinte sau fraze), însă se pierde structura și aspectul original al documentului.

O versiune „codificată" (*encoded*) a documentului va include informații suplimentare sau *markup* de diferite feluri, pentru a exprima structura documentului, formatarea sau alte informații pe care creatorul a dorit să le evidențieze și să-i ofere acestuia funcții speciale. Acest tip de codificare *markup* este folosit frecvent în asociere cu limbajele SGML sau XML și acest gen de informație este aplicată documentelor cu text integral[2].

Bibliotecile universitare din consorțiu dețin, la ora actuală, arhive relativ bogate de texte digitizate, accesibile (cu excepția bibliotecii timișorene) în regim *open-source.* În cazul Bibliotecilor Central Universitare din Iași și Cluj, „bibliotecă digitală" este ușor de accesat de pe site-urile aferente. Selecția de volume și periodice digitizate conține, în special, exemplare din colecții speciale de carte/revistă veche sau rară, criteriile de selecție pentru digitizare constând în 1. gradul de degradare (cărți care trebuite protejate), respectiv 2. exemplarele clasate la categoria *high demand*. Biblioteca digitală a BCU Cluj permite consultarea, ajutată de mai multe filtre (interesant de observat este că la categoria „Literatură" sunt indexate 2495 de unități, iar cf. filtrului privind data apariției, cele mai multe unități – 76826 – sunt apărute între 1900 și 1999, urmate de 42816, între 1800 și 1899; cele mai vechi documente digitizate (4) se încadrează cu data de apariție între 1475 și 1499; Total: 122839[3]). La BCU Timișoara secțiunea „Resurse electronice" poate fi consultată doar pe baza contului de utilizator și pe bază de parolă din interiorul rețelelor Universității de Vest. BCU București are o platformă digitală – „Restitutio" – adică un depozit de tip *open acces*, disponibil oricui; cuprinde un număr de 655 de documente digitizate. La București pe subiectul literatură sunt 164; cele mai multe sunt din perioada 1900-1999 (473), iar cele mai vechi (7) datează din perioada 1483-1499[4].

Proiectul Biblioteca Digitală Națională al BNR este în regim de open acces din 2007, cu motor de căutare pe site-ul digitool.bibnat.ro; (conține 694 de resurse Carte românească veche și bibliofilă; 617 periodice românești vechi și 1077 de carte străină veche)[5].

Există, la nivel național, o politică națională pentru digitizare. Sub coordonarea Bibliotecii Naționale a României a fost numită (aprobată de Guvernul României pe 18 ianuarie 2008) o Comisie de specialitate pentru coordonarea activităților prevăzute de Politica publică pentru digitizarea resurselor culturale și realizarea bibliotecii digitale a României, a Ministerului Culturii și Cultelor, (OM nr. 2244/15.04.2008). Misiunea sa e de a coordona activitatea de digitizare a patrimoniului cultural scris. Această comisie a elaborat, în 30.10.2009, un ghid de digitizare pe modelul *Europeana.eu*, iar ultimul raport de activitate a comisiei, disponibil pe site-ul BNR, datează din 2009, dar lista de inventar a documentelor digitizare în bibliotecile din România este actualizată în trimestrul IV 2018 și cuprinde 14951 de intrări; lista de inventar a documentelor digitizate în BNR, actualizată în oct. 2019, cuprinde 13176 de intrări. Într-un răspuns oferit de BCU București, instituția notează proiectul *Lib2life – Revitalizarea bibliotecilor și a patrimoniului cultural prin tehnologii avansate*, ce se desfășoară în perioada 2018-2020. Finanțat de UEFISCDI P1 – Dezvoltarea sistemului național de CD, Proiecte complexe realizate în consorții CDI (PCCDI). Proiectul este coordonat de BCU „Carol I" București, fiind dezvoltat de un consorțiu alcătuit din 6 parteneri: cele 4 biblioteci central universitare din București, Iași, Timișoara, Cluj-Napoca, Universitatea Politehnică București și Institutul Național de Cercetare Dezvoltare în Informatică. Biblioteca Universitară „Lucian Blaga" din Sibiu are o platformă de open-acces la resursele electronice (aproximativ 2000 de unități digitizate) și e colaborator la *Europeana.eu*.

Biblioteca Universitară „Transilvania" din Brașov desfășoară proiecte locale de digitizare, fără finanțare externă, în baza planului anual de acțiuni. Se digitizează documente de importanță pentru specificul interesului local (158 de reviste și 5 monografii). Biblioteca Județeană „Panait Istrati" din Brăila care colaborează, și ea, la Europeana.eu cu 2424 unități constând în documente literare[6]. În fine, Biblioteca Metropolitană București desfășoară activități specifice digitizării începând cu anul 2007.

La ora actuală în colecțiile BDD, numărul obiectelor digitale existente se ridică la 101.352, însumând: publicații periodice – 87.520/200 titluri, Carte Veche Românească – 1.376, Patrimoniul Cultural Armean – 670, Cărți poștale – 820, Bibliografie școlară – 2.324, Mica Alexandru/ Documente audio – 778, carte – 7.864. Conform ultimul raport de activitate de pe site-ul instituției, în 2018, au fost scanate 210 documente. De subliniat că biblioteca desfășoară și activități de prelucrare: în 2018, 44.147 pagini au fost prelucrate în BookRestorer, 1.562 documente au fost prelucrate digital, 350.952 de pagini au fost prelucrate în CVision PDFCompressor și 4.384 documente au fost verificate și arhivate (4.165 fotografii BAR și 183 titluri CSIER).

În 30 septembrie 2005 Comisia Europeană publică *I2010: Communication on Digital Libraries*, unde stabilește ca obiectiv strategic și realizarea bibliotecii digitale europene. *Europeana.eu* este proiectul Comisiei Europene din august 2006 prin care se dorește constituirea unei biblioteci europene prin unificarea corpusului de documente digitizate ale celor 25 de state membre la data respectivă („Our aim is to arrive at a real

European digital library, a multilingual access point to Europe's digital cultural resources", Viviane Reding, Commissioner for Information Society and Media, în comunicatul de presă de la lansarea proiectului 25 august 2006[7]). Europeana.eu a fost dublat, între 2007 și 2008, de un alt proiect, *European Digital Library Project*, finanțat tot de CE prin proiectul *eContentplus*, coordonat de Biblioteca Națională a Germaniei, integrează cataloagele bibliografice și colecțiile digitale ale bibliotecilor naționale din Belgia, Grecia, Islanda, Irlanda, Lichtenstein, Luxemburg, Norvegia, Spania și Suedia.

Intrată în UE în 2007, România, prin Biblioteca Națională a României, realizează un *Studiu de fezabilitate privind digitizarea, prezervarea digitală și accesibilitatea online a resurselor bibliotecilor* pornind tocmai de la aceste inițiative și direcții strategice europene. Obiectivele care au stat la baza acestui demers se referă la: transpunerea în format electronic a patrimoniului cultural scris; promovarea patrimoniului cultural scris la nivel european; protejarea valorilor de carte bibliofilă și manuscrise; protejarea documentelor aflate într-o stare avansată de deteriorare; îmbunătățirea posibilităților de acces la documente, local sau la distanță, cu impact asupra creșterii numărului de utilizatori și a categoriilor acestora; posibilitatea consultării simultane de către mai mulți utilizatori a aceluiași document; oferirea unui mod de consultare a documentelor modern, în acord cu noile tehnologii, independent de spațiul și programul de funcționare al bibliotecii (cu respectarea restricțiilor de copyright); îmbunătățirea calității procesului de consultare a documentelor; creșterea numărului de resurse electronice realizate direct în format electronic, fără echivalent tradițional (tipărit).

La acel moment, s-au aplicat chestionare în interiorul sistemului național de biblioteci; cele mai importante observații la momentul respectiv au fost că toate bibliotecile dețineau cataloage electronice, că unele au derulat proiecte locale de digitizare cu caracter individual și de o mică amploare, iar cel mai important minus pentru realizarea proiectului de digitizare era citat numărul mic de specialiști IT din biblioteci; de asemenea, se sesiza la acel moment necesitatea achiziționării unui nou soft de bibliotecă unitar la nivel național pentru a exista compatibilitate între procesele de digitizare derulate la nivel local. În procesul de identificare a proiectelor de digitizare la acel moment sunt citate proiecte de digitizare la BCU București (aprox. 20000 de pagini), la Biblioteca Academiei Române (peste 36000 de facsimile digitale din manuscrisele Eminescu), Biblioteca Metropolitană din București – peste 5500 de pagini de bibliografie școlară, Biblioteca Județeană „Panait Istrati" (Brăila) – 30000 de pagini literatură școlară română și Biblioteca Județeană „Gh. Asachi" din Iași – peste 2594 de pagini și 950 de imagini (alte colecții specifice la Biblioteca județeană din Brașov și la Ministerul Culturii și Cultelor). *Europeana.eu* oferă acces la peste 50 de milioane de articole digitalizate. Dintre acestea, din România provin doar 154.000, adică doar 0,3% din total: „În 2017, România nu a trimis nimic la Europeana"[8], a precizat Dan Matei, care este și membru în grupul de experți din statele membre pentru proiectul bibliotecii digitale europene.

Contributorii români la Biblioteca digitală *Europeana.eu* sunt următorii: BCU Cluj 58.374; Biblioteci/instituții de cultură locale din România (6840), dintre care Biblioteca Județeană „Panait Istrati" Brăila (2,424), Timis County Library (1,442), Images of Old Cluj (686), Biblioteca Județeană „V.A. Urechia" Galați (501), Dolj County Library (372), „Octavian Goga" Cluj County Library (282), Cluj County Library (281), Cluj County Center for the Preservation of Traditional Culture (199), Arhivele Naționale, Serviciul Județean Cluj (194), EuropeanaLocal Romania (152), Hunedoara County Library (147), Cluj County Centre for the Preservation and Promotion of Traditional Culture (110), Aman Library (30), Biblioteca Județeană „G. T. Kirileanu" Neamț (8), Books about Cluj County (8), Museum of Dacic and Roman Civilisation in Deva (4); Biblioteca Naționala a Romaniei – 4324; Institutul National al Patrimoniului Bucuresti – 4114; Biblioteca Academiei Romane – 1703; Bcu Sibiu – 549.

Reunirea totală a corpusurilor digitizate de la nivelul României ar fi făcut, de asemenea, obiectul unui proiect al Ministerului Tehnologiei, Informației și Comunicațiilor inclus în „Agenda Digitală pentru România" – 2020; intitulat *Culturalia*. Proiectul își propune digitizarea a circa 550000 de bunuri culturale; la 27 mai 2018, pe blogul dedicat proiectului (blog.culturalia.ro), inițiatorii anunță că „(în sfârșit) proiectul este în faza de contractare între Ministerul Fondurilor Europene și Ministerul Culturii". Principala diferență între proiectul *Culturalia* și *Europeana* ar fi aceea că în vreme ce *Europeana* este un catalog colectiv alimentat cu date offline doar de către instituții culturale, *Culturalia* își propune să fie un catalog partajat care să poată fi, deci, alimentat cu date online nu doar de către instituții culturale, ci și de către publicul larg. Un articol semnat de Răzvan Chiruță, publicat în 12 martie 2018, în *Suplimentul de cultură* arată cum biblioteca digitală a României (*Culturalia*) este un proiect rămas încă „virtual". Pe site-ul dedicat, Culturalia.ro, apare mesajul datat 07.07.2017 „Aici veți găsi catalogul partajat național și indexul bibliotecii digitale a României". Întrebându-se de ce acest proiect care trebuia să existe din 2014 a fost propus spre finanțare abia la finele lui 2017, Chiruță a cerut clarificări Ministerului Fondurilor Europene; răspunsul îi aparține lui Dan Matei, fost director al Institutului de Memorie Culturală și initiator al

proiectului. Se pare că proiectul s-a lovit de reticența managerilor bibliotecilor din țară care au solicitat un număr prea mare de norme pentru fișarea obiectelor digitizate; astfel, doar 29 de parteneri (un număr extrem de mic) au răspuns proiectului. Două dintre instituțiile care dețin la ora actuală cele mai mari corpusuri de documente digitizate nu au răspuns invitației de a fi parte a proiectului – este vorba de Biblioteca Națională a României și BCU Cluj[9].

În loc de concluzii, este necesar să insistăm asupra faptului că toate proiectele de digitizare din România au fost gândite nu ca necesități ale cercetării active, ci mai degrabă ca modalități de prezervare patrimonială și/sau ca modalități de acces ale publicului la volume foarte căutate. S-a sesizat prea puțin utilitatea acestor materiale pentru cercetarea propriu-zisă, rezultând de aici specificul corpusului digitizat, pentru care lipsește cu desăvârșire infrastructura teoretică și metadatele necesare cercetărilor presupuse de formalismul digital.

Edițiile, arhiva, corpusul

Într-unul dintre pamfletele[10] care urmau să fie reunite în volumul *Canon/Archive*[11], cercetătorii de la Stanford Literary Lab făceau câteva distincții esențiale între ceea ce numeau „edițiile", „arhiva" și „corpusul". Astfel, aplicat proiectului nostru, *ediții* cuprinde totalitatea producției de roman românesc care reprezenta realitatea editorială a secolului al XIX-lea. Acest segment literar, „orizontul fundamental al tuturor cercetărilor cantitative"[12], poate fi survolat (dar nu și consultat) cu ajutorul unui instrument lexicografic de tipul DCRR-ului[13]. În plus, *arhiva* desemnează tot ceea ce s-a păstrat – în biblioteci, colecții personale, fonduri și arhive – și care poate fi accesat și consultat în format fizic. Nu în ultimul rând, *corpusul* reprezintă o secțiune a arhivei care poate fi supusă unor cercetări literare (cantitative sau nu). Deoarece, ținând cont de logica interrelațională a celor trei noțiuni, avem de a face cu o ierarhizare determinată cantitativ (edițiile sunt mai numeroase decât arhiva, iar arhiva mai voluminoasă decât corpusul), cea mai mică unitate de măsură a ceea ce poate fi studiat în câmpul formalismului digital este corpusul.

Dificultățile ce țin de nevoia de egalizare cantitativă între cele trei tipuri de corp textual sunt foarte pertinent explorate de către cercetătorii de la Stanford Literary Lab, care aveau la dispoziție un corpus de 4000 de romane englezești publicate între 1750 și 1880, deosebit de inegal în raport cu producția totală de roman. Mai mult, ceea ce desemnează un corpus pentru formalismul cantitativ propus de cercetătorii americani constituie – sau ar trebui să constituie – un eșantion reprezentativ selectat *aleatoriu*. În schimb, arhiva celor de la Stanford Literary Lab suferă – și o anunță chiar colaboratorii proiectului – de o anumită „prejudecată de selecție", mai exact: din corpusul lor total de romane, existau discrepanțe de ordin genologic (aveau mai multe romane gotice decât istorice, de pildă) și temporal (segmentele de timp vizate erau reprezentate, în corpus, inegal). Astfel, pentru a compensa aceste prejudecăți de selecție (și care țin de disponibilitatea și de accesul cercetătorilor la arhiva de roman englezesc), Stanford Literary Lab a optat, în realizarea experimentul citat, pentru un eșantion de 674 de romane. Nu vom insista asupra dificultăților de accesare a metadatelor din acest corpus[14]; concluzia extrasă din experiența americană a elaborării unei *arhive* românești este, în schimb, pertinentă pentru ceea ce urmează în prezentarea arhivei românești propuse în proiectul nostru: că orice demers colectiv care dezvoltă o cercetare cantitativă asupra unei producții literare se poate realiza doar interstițial, anume în spațiul dintre multiple instituții care iau parte la un asemenea proiect.

Particularizând conceptualizările anterioare la situația romanului românesc în secolul al XIX-lea și la proiectul de digitalizare *Astra Data Mining*, rezultă că două dintre cele trei noțiuni, arhiva și corpusul, se suprapun aproape fără rest, întrucât – și acest aspect e important de accentuat – *arhiva* romanului românesc de secol al XIX-lea își are, în sfârșit, un *corpus* în proporție de 92,3% complet. Din start, toate dificultățile întâmpinate de cei de la Stanford (relevanța corpusului raportat la arhivă, gradul de acoperire al arhivei relaționat cu edițiile „reale" apărute etc.) sunt în mare măsură anulate în cazul nostru. Sigur că – și asta e de domeniul simplei evidențe – arhiva romanului românesc de secol XIX este „privilegiată" numai și datorită diferențelor cantitative enorme dintre o cultură precum cea engleză și cultura noastră. Însă faptul că s-a reușit o suprapunere aproape totală între arhivă și corpus, respectiv că arhiva în sine acoperă aproape 90% din producția totală de roman documentată în cel mai important instrument lexicografic dedicat romanului românesc, *Dicționarul cronologic al romanului românesc de la origini până în 1989*, trebuie privite ca pe niște câștiguri indeniabile: mai mult decât o selecție aleatorie, un eșantion reprezentativ, corpusul digital realizat de Astra Data Mining constituie, din toate punctele de vedere, *o arhivă totală* a romanului românesc publicat în ediție în secolul al XIX-lea. Ținem la acest detaliu – publicat *în ediție* – și pentru că trebuie să avem în vedere o altă realitate publicistică a secolului al XIX-lea: romanul în foileton, care constituie – și aici avem am luat în considerare doar foiletoanele care au fost finalizate – un segment relevant cantitativ și cultural, aproximativ 32% din producția reală de roman românesc în perioada vizată. Felul în care am tratat gestionarea romanelor în foileton în relație cu arhiva pe care am digitalizat-o ține de limitările tehnologice cu care ne-am confruntat, dar și cu condițiile de păstrare

a revistelor care găzduiau fragmente foiletonistice și cu gradul lor de truvabilitate. În final, am optat pentru omiterea cvasi-totală a foiletoanelor din arhiva propusă, optând pentru păstrarea doar a acelor foiletoane care au ajuns să fie publicate și în ediție.

Între arhivare și digitizare

Câteva date legate de premisele de la care am pornit se impun. Selecția corpusului a fost dirijată exclusiv după un criteriu elementar: tendința totalizatoare. Dat fiind faptul că miza centrală a acestui proiect a fost livrarea unei arhive digitale *exhaustive* a romanului românesc din secolul al XIX-lea, discuțiile premergătoare, expuse la începutul acestui studiu, referitoare la spații cu o producție și o tradiție literară imense, au fost evitate. Unul dintre avantajele pe care implementarea formalismului digital în culturi minore (i.e. tinere, cu o producție literară modestă din punct de vedere cantitativ) este tocmai acesta: posibilitatea de a livra arhive exhaustive ale unui secol întreg de literatură. Mai departe, cel mai mare impediment a fost dat de accesibilitatea corpusului. Două probleme – de așteptat – au fost întâmpinate: 1. dispariția totală din orice bibliotecă/fond de carte a unora dintre titlurile indexate în DCRR-1, imposibilitatea, deci, de consultare și digitizare a unui exemplar, fie el și în formulă reeditată, și 2. dificultatea cu care exemplarele existente au fost făcute accesibile pentru digitizare. Ne referim aici atât la condițiile de păstrare și accesare a acestor volume în fondurile bibliotecilor (de găsit doar la căutarea manuală, pe fișele tradiționale, scrise de mână, exemplare recondiționate manual – „accesorii" care îngreunau așadar procesul de scanare și de prelucrare), cât și la gradul de reticență cu care unele dintre bibliotecile care dețineau bune părți din corpusul necesar au întâmpinat inițiativele acestui proiect.

Cea mai mare parte a corpusului vizat se află la Biblioteca Central Universitară „Lucian Blaga" din Cluj-Napoca, cu fondul conex al Facultății de Litere din cadrul Universității „Babeș-Bolyai" (fondul bibliotecii de literatură română). Vorbim aici despre un total de 82 de romane scanate din 157, așadar peste 50% din corpusul total, restul de circa 48% fiind accesibil în alte 4 biblioteci din țară, care urmează să fie menționate, de asemenea, în segmentul de față. Obținerea unui acord inter-instituțional s-a dovedit foarte utilă. Odată corpusul făcut accesibil, echipa responsabilă de digitizarea materialului aflat în centrul universitar clujean (formată din Andreea Coroian-Goldiș, Daiana Gârdan, Emanuel Modoc și Cosmin Borza), a demarat procesul de digitizare, care, din nou, cum era de așteptat, a presupus o serie de dificultăți.

În primul rând, timpul alocat fiecărei scanări s-a dovedit mai mare decât preconizat inițial, din cauze care țin atât de capriciile tehnice și proprietățile aparaturii utilizate, cât și de formatul edițiilor. Insistența – auto-impusă – de a selecta, acolo unde erau de găsit (cu mențiunea că acest lucru s-a reușit în cazul majorității romanelor), edițiile princeps a condus la dificultatea digitizării lor în măsura în care, gândit și formalizat pentru standardele edițiilor actuale de carte, instrumentul principal utilizat pentru acest proces s-a dovedit dificil de manevrat acolo unde aveam de-a face cu formate de ediții care au dispărut ulterior de pe piață și care au ridicat necesitatea unor foarte fine reglaje, costisitoare atât în termeni de timp, cât mai ales în termeni de calitate (ne referim la imposibilitatea generării unei ediții digitizate perfect curate și simetrice).

În al doilea rând, recondiționarea edițiilor princeps de către fondul bibliotecar în care au fost depuse și păstrate a condus adesea la pierderea unor informații prețioase pentru orice demers de *data mining* și de analiză cantitativă bazată pe metadate. Un asemenea caz este, de pildă, lipsa colofonului, acolo unde ar fi trebuit să existe, și imposibilitatea accesării, așadar, a informațiilor legate de tiraj. Lipsa unor coperte sau deteriorarea aproape completă a acestora se înscrie în aceeași categorie. În același timp, alte elemente de dificultate pentru desfășurarea proceselor de scanare, procesare, curățare și „minare" ale romanelor au fost reprezentate de cazurile în care recondiționarea unor pagini rupte s-a făcut cu metode mai mult sau mai puțin improvizate (lipire cu bandă adezivă sau de hârtie). Au existat și situații în care lipseau total unele bucăți de pagini.

În al treilea rând, calitatea edițiilor a reprezentat o problemă în cele mai multe cazuri. Lucrând cu un segment de literatură prevalent minor, lipsa de valoare estetică s-a tradus, de multe ori, și în lipsa de calitate a editurii care a scos romanul (acesta este, de pildă, cazul autorilor cu edituri proprii, vezi Teochar Alexi). Precaritatea și transparența hârtiei, pe lângă problema dimensiunii edițiilor mai sus menționate, a ridicat o cu totul altă serie de probleme la scanare. Aparatura nu reușea, de cele mai multe ori, decât cu eforturi repetate, să înregistreze foi care aveau o anumită culoare și să citească textul aflat pe ele. În același sens, acolo unde foile erau atât de subțiri încă textul se putea vedea de pe cealaltă parte, era în aceeași măsură dificil ca aparatura să înregistreze textul.

Exemplarele care nu se regăseau la Biblioteca Central Universitară clujeană au fost de găsit, în cea mai mare parte, la Biblioteca Astra din Sibiu și la Biblioteca Academiei Române din București. Dificultățile procesului de digitizare au fost, în mare parte aceleași, legate de trăsăturile fizice ale edițiilor folosite.

O descriere particularizată a activităților desfășurate în cadrul proiectului de față la Sibiu redă,

în fond, aceeași schemă de evoluție a etapelor dinainte planificate și urmărite în vederea obținerii principalelor obiective propuse. În ceea ce o privește pe prima, aceasta s-a subsumat operației de căutare a romanelor publicate în secolul al XIX-lea în cataloagele clasice ale Bibliotecii Județene „Astra" Sibiu.

De fapt, se impun până în acest moment trei precizări care să clarifice tot atâtea aspecte, precizări pe care le vom dezvolta apoi în mod individual, întrucât îndeplinesc un rol major în schematizarea planului de lucru din etapa incipientă.

În primul rând, această operație de căutare a implicat mai degrabă o operație de verificare, întrucât romanele erau în acel moment deja selectate, urmând ca această verificare să genereze o listă de treizeci și șapte de romane. În al doilea rând, Biblioteca „Astra" a fost, fapt ce reiese din informațiile anterioare, principalul partener al proiectului la Sibiu. Îl numim principalul și nu singurul, deoarece am colaborat și cu una dintre filialele Bibliotecii Universității „Lucian Blaga" din Sibiu. Este vorba despre Biblioteca Facultății de Teologie „Andrei Șaguna", din fondul căreia am scanat unul dintre romanele lui N. D. Popescu – *Radu al III-lea cel frumos. Novelă originală*, publicat în 1864. Cea de-a treia precizare vizează cataloagele clasice ale Bibliotecii „Astra", cataloage alcătuite din fișe ordonate alfabetic, și timpul alocat verificării și notării cotelor care corespund titlurilor căutate. Biblioteca este dotată cu un catalog electronic ce permite accesarea fără niciun fel de dificultate a marii majorități de cărți care îi compun fondul, însă – având în vedere anul publicării celor mai multe dintre romanele digitizate în proiect – acestea se află în depozitul clădirii vechi a instituției și în fondul „colecțiilor speciale", fond care se găsește în aceeași clădire (Corpul A). Consecința evidentă rezidă în faptul că romanele nu sunt indexate în baza de date online, ceea ce reprezintă în mod evident un prim dezavantaj tocmai din cauza faptului că operația de verificare a necesitat un timp mai îndelungat. În aceeași categorie a dezavantajelor specifice cataloagelor tradiționale și cărților nesolicitate frecvent poate fi inclusă și situația în care cota trecută pe fișa din catalog să nu (mai) corespundă cu titlul romanului.

Revenind la tematica articolului, trebuie să menționăm faptul că Biblioteca „Astra" nu dispune de o arhivă destinată romanului românesc, informație care nu transformă nicidecum această bibliotecă într-un caz izolat din România, ba dimpotrivă, efectul este cel opus operației de particularizare. Importante de punctat în acest context sunt, pe de o parte, faptul că absența unei arhive nu afectează vizibil posibilitățile de digitizare, *Astra Data Mining. Muzeul Digital al Romanului Românesc din secolul al XIX-lea* fiind unul dintre exemplele care certifică această convingere, iar pe de altă parte, faptul că beneficiile obținute din implementarea unui asemenea proiect pot fi identificate mai ales în acest domeniu de activitate. Arhiva romanelor digitizate constituie nu doar un mijloc de prezervare a romanelor din secolul al XIX-lea, dar și un nou corpus de texte care pot fi supuse analizei. Acesta este, de fapt, avantajul proiectelor care își propun vizitarea unui număr foarte mare de texte dintr-o anumită perioadă, întrucât acoperirea în întregime a unui interval de timp reprezintă obiectivul de bază.

Această observație poate fi ilustrată printr-un exemplu concret care relevă utilitatea proiectului de față. Biblioteca „Astra" dispune, aspect ușor de remarcat în momentul accesării site-ului propriu[15], și de o secțiune destinată „Bibliotecii digitale". Dacă ne-am opri la această informație, beneficiile digitizării romanelor din cadrul proiectului ar scădea considerabil, deoarece biblioteca este dotată cu mijloacele necesare întreprinderii unei astfel de activități. Diferența intervine la nivelul „criteriilor de selecție" pe care biblioteca le impune: „publicații unicat; publicații relevante pentru Sibiu și Transilvania; publicații foarte solicitate; documente scanate la solicitarea cititorilor". Dintre cele patru criterii, numărul cel mai mare de romane digitizate din secolul al XIX-lea s-ar fi datorat cel mai probabil celui dintâi, fapt evident numai la o verificare a celorlalte trei. Ultimele două dintre acestea ies din discuție de la bun început din cauza simplului fapt că nu există solicitări însemnate pentru romane atât de vechi. Absența cererii nu trebuie pusă, însă, doar pe seama unei incongruențe între orizontul de așteptare al lectorului de astăzi și universul romanesc din secolul al XIX-lea (din cauza caracteristicilor limbii vorbite/scrise în perioada respectivă și a absenței calităților literare ale anumitor texte), ci și pe seama faptului că multe dintre aceste romane aparțin unor autori deloc/puțin cunoscuți. În ceea ce privește cel de-al doilea criteriu, dacă ar fi să ne oprim la „publicațiile relevante pentru Sibiu" există în totalul celor 157 de romane digitizate în cadrul proiectului doar două romane publicate pe plan locl. Este vorba despre *Petru Rareș, principele Moldaviei. Nuvelă istorică originală*, scris de At. M. Marienescu și publicat în 1862 la tipografia *Filtsch*, respectiv unul dintre romanele lui Alexi Theochar – *Ai carte, ai parte. Roman umoristic* – publicat în 1878 la aceeași tipografie, numită în acea perioadă *W. Krafft*. Bineînțeles că dacă am lua în considerare acea relevanță „pentru Transilvania", datele ar suferi modificări considerabile datorită numărului de romane publicate la Cluj și, mai ales, la Brașov. Și totuși cele mai multe dintre romanele digitizate au fost publicate la București. Scopul exemplului a fost, însă, de a releva încă o dată rolul proiectului, date fiind câteva remarci valabile pentru desfășurarea acestuia la Sibiu, remarci care fac evidentă o posibilă concluzie: digitizarea unui asemenea corpus de romane ar fi fost greu realizabilă în afara proiectului.

Descrierea celei de-a doua etape, cea în care a avut loc procesul de digitizare propriu-zis, pornește de la câteva detalii tehnice. Echipamentul de scanare și softul utilizat au prezentat un real avantaj în cadrul acestui proces. Luând în considerare caracterul fragil al *documentelor* scanate – majoritatea cărților făcând parte fie din colecțiile speciale, fie din cadrul fondului vechi de carte – scanerul CZUR ET16 a permis prezervarea documentelor originale datorită senzorului foto încorporat, ba mai mult, a făcut posibilă transpunerea în format digital a cărților prin intermediul software-ului.

Dificultățile apărute în timpul scanării au fost implicit legate de trei aspecte: formatul paginilor, calitatea hârtiei și defectele filelor. Astfel, marginile foarte înguste și inflexibilitatea cotorului, impedimente accentuate, totodată, și de fragilitatea hârtiei, au pus reale probleme în timpul scanării. Romane precum: *Meșterul Manole sau Fundarea monastirei Curții de Argeș* de N. D. Popescu, exemplar din anul 1882, sau *Iubita* lui Traian Demetrescu, exemplar din anul 1896, sunt doar câteva exemple care, din cauza defectelor precum pete, aranjarea în pagină a textului, însemnări făcute cu diverse instrumente de scris, au prelungit timpul alocat scanării și pre-procesării. De asemenea, obișnuit cu un spațiu alb pe care apar blocuri de text bine delimitate și la distanțe (relativ egale de cotor și margini), scanerul nu a recunoscut uneori textul din romanele a căror ediție digitizată corespundea cu anul apariției.

Stadiul de procesare care a urmat s-a desfășurat în aplicația CZUR și a fost împărțită în două etape: pre-procesarea romanelor – etapă în care varianta jpg. a fiecărei pagini a fost editată și verificată și post-procesarea – fază în care, prin intermediul funcției OCR (proces ce permite recunoașterea caracterelor dintr-o imagine scanată a unui text și transpunerea acesteia într-un fișier text) au fost create variantele PDF și Word ale romanului editat. Trebuie menționată aici ineficiența acestui soft în cazul romanelor din secolul al XIX-lea al căror alfabet de tranziție (alfabet mixt chirilic-latin, prezent între anii 1830-1862 în România) a permis doar parțiala recunoaștere a caracterelor, întrucât alfabetul chirilic vechi nu se regăsește între cele 186 de limbi cu care operează versiunea acestui program. Iată, în linii mari, modul în care s-a desfășurat proiectul la Sibiu și descrierea primelor două etape ale acestuia.

Note:

1. Vezi Calimera Guidelines, http://www.calimera.org/lists/guideslines/digitisation.
2. Cf. Comisia de specialitate pentru digitizare – pilonul tematic „Biblioteci", *Ghid de digitizare – pilonul tematic „Biblioteci"*, p. 6, accesibil la: http://www.bibnat.ro/dyn-doc/Ghid%20de%20digitizare_Pilonul%20tematic_Biblioteci_versiunea01_25_11_2009.pdf.
3. Vezi http://dspace.bcucluj.ro/?fbclid=IwAR1WqTfuxnqkSOQ7PniE_SoLwmeS6Jkj-mI3yK69pu5DiWamH178QmQ5tpQ.
4. Vezi http://restitutio.bcub.ro/.
5. Vezi http://www.bibnat.ro/Biblioteca-Digitala-s89-ro.html.
6. Vezi https://www.bjbraila.ro/literatura-romana/.
7. Vezi https://ec.europa.eu/digital-single-market/en/news/european-digital-library-commission-calls-member-states-contribute-european-digital-library.
8. Vezi http://suplimentuldecultura.ro/22100/biblioteca-digitala-a-romaniei-un-proiect-ramas-inca-virtual/?fbclid=IwAR2YLehJIXk99N8zZkQLAf_DuQnbWCGMsjji6StVoSkVQbNCNcEFKP6ymh0.
9. Vezi http://culturalia.ro/
10. Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, Hannah Walser, *Canon/Archive. Large-scale Dynamics in the Literary Field*, January 2016, https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf.
11. Franco Moretti, ed., *Canon/Archive* (New York: n+1 Foundation, 2017).
12. Mark Algee-Hewitt et al., *Canon/Archive. Large-scale Dynamics in the Literary Field*, 2.
13. ****Dicționarul cronologic al romanului românesc de la origini până în 1989* (București, Editura Academiei Române, 2004).
14. Vezi Mark Algee-Hewitt et al., *Canon/Archive. Large-scale Dynamics in the Literary Field*, 2-3.
15. Biblioteca Județeană „Astra" Sibiu, http://bjastrasibiu.ro/biblioteca-digitala/.

Bibliography:

***. *Dicționarul cronologic al romanului românesc de la origini până în 1989*. București: Editura Academiei Române, 2004.

Algee-Hewitt, Mark, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, Hannah Walser. *Canon/Archive. Large-scale Dynamics in the Literary Field*. January 2016. https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf.

Biblioteca Digitală „RESTITUTIO". http://restitutio.bcub.ro/.

Biblioteca Digitală a BCU Cluj. http://dspace.bcucluj.ro/?fbclid=IwAR1WqTfuxnqkSOQ7PniE_SoLwmeS6Jkj-mI3yK69pu5DiWamH178QmQ5tpQ.

Biblioteca Județeană „Astra" Sibiu. http://bjastrasibiu.ro/biblioteca-digitala/.

Biblioteca Județeană „Panait Istrati" Brăila. https://www.bjbraila.ro/literatura-romana/.

Biblioteca Națională a României. http://www.bibnat.ro/Biblioteca-Digitala-s89-ro.html.

Calimera Guidelines. http://www.calimera.org/lists/

guideslines/digitisation.
Chiruță, Răzvan. "Biblioteca Digitală a României, un proiect rămas încă virtual." *Suplimentul de cultură*, nr. 598 (12 martie 2018). http://suplimentuldecultura.ro/22100/biblioteca-digitala-a-romaniei-un-proiect-ramas-inca-virtual/?fbclid=IwAR2YLehJIXk99N8zZkQLAf_DuQnbWCGMsjji6StVoSkVQbNCNcEFKP6ymh0.
"Commission calls on Member States to contribute to the European digital library" Brussels, 25 august 2006. https://ec.europa.eu/digital-single-market/en/news/european-digital-library-commission-calls-member-states-contribute-european-digital-library.
Culturalia. http://culturalia.ro/.
"Ghid de digitizare – pilonul tematic *Biblioteci*". http://www.bibnat.ro/dyn-doc/Ghid%20de%20digitizare_Pilonul%20tematic_Biblioteci_versiunea01_25_11_2009.pdf.
Moretti, Franco (ed.). *Canon/Archive*. New York: n+1 Foundation, 2017.

Prezentul articol a fost realizat în cadrul proiectului *ASTRA Data Mining. Muzeul Digital al Romanului Românesc din Secolul al XIX-lea*, organizat de Complexul Național Muzeal ASTRA și co-finanțat de Administrația Fondului Cultural Național. Proiectul nu reprezintă în mod necesar poziția Administrației Fondului Cultural Național. AFCN nu este responsabilă de conținutul proiectului sau de modul în care rezultatele proiectului pot fi folosite. Acestea sunt în întregime responsabilitatea beneficiarului finanțării.

PROIECT CO-FINANȚAT DE:

Administrația Fondului Cultural Național