DOI: https://doi.org/10.53486/9789975155649.52
CZU: 025.31:004.65

BIBLIOGRAPHY DATABASE METADATA – CHALLENGES AND SOLUTIONS

METADATELE DIN CADRUL BAZELOR DE DATE BIBLIOGRAFICE – PROVOCĂRI ȘI SOLUȚII

Irina COJOCARU[152]**, PhD Student**

Abstract: *Any function of bibliographic databases cannot be performed without metadata. Creating and managing metadata is a key activity and a commune responsibility of content producers, distributors and users, as important as the content itself. Open and relevant metadata contribute to increasing the visibility of digital scientific content.*
Key words*: bibliographic databases, metadata, standard, format, scientific content, visibility.*

JEL CLASSIFICATION: I23

1. Introducere

Bazele de date bibliografice furnizează informații despre publicații științifice - cel mai răspândit rezultat al activității de cercetare. Aceste baze sunt utilizate atât pentru regăsirea informației în scopul documentării bibliografice, care stă la baza noilor cercetări, cât și pentru determinarea și calcularea mai multor indicatori, în special celor bibliometrici. Astfel, aceste instrumente tind să satisfacă așteptările diverselor părți interesate, precum cercetători, organizații de cercetare, organisme de finanțare a cercetării, factori de decizie, mediu de afaceri și publicul larg. Indicatorii și informațiile furnizate de bazele de date bibliografice sunt utilizați pentru evaluarea cercetării, asigurarea calității, promovarea și sporirea vizibilității publicațiilor științifice, și cel mai important - pentru căutarea și regăsirea conținutului.

Niciuna dintre aceste funcții ale bazelor de date bibliografice nu poate fi realizată fără metadate. Metadatele – date despre date – în cea mai largă accepțiune, se referă atât la informații, cât și la fluxul acestora, iar crearea și administrarea metadatelor reprezintă o activitate-cheie, la fel de importantă, ca și conținutul în sine. Anume metadatele din domeniul cercetării formează o rețea informațională în cadrul comunității științifice, care interconectează cercetătorii, organizațiile de cercetare, organismele de finanțare, editurile și bibliotecile științifice. O circulație liberă a informației în cadrul acestei comunități poate facilita un suport adecvat al activităților științifice și eventual determina eficiența cercetărilor în sine (Science, 2016). Totuși, metadatele din domeniul științific în particular și din orice domeniu, în general, sunt guvernate de o multitudine de standarde, formate, vocabulare și reguli. Asigurarea metadatelor deschise, cartografierea acestora pentru definirea unor cerințe de interoperabilitate, definirea unui vocabular consistent pentru metadatele științifice și analiza necesităților cercetătorilor în materie de metadate, sunt doar unele dintre provocările și problemele privind metadatele din cadrul comunicării științifice.

Prezenta lucrare își propune să elucideze rolul și importanța metadatelor în cadrul comunicării științifice digitale, în special în cadrul bazelor de date bibliografice, să sintetizeze provocările asociate utilizării acestora și să propună un șir de recomandări pentru depășirea provocărilor și problemelor respective.

2. Metadatele din cadrul bazelor de date bibliografice

Motoarele de căutare academice și bazele de date bibliografice (ASEBD) reprezintă la moment destinația standard, unde sunt accesate publicațiile științifice actualizate. Astfel de resurse

[152] E-mail: irisha.cojocaru@gmail.com, Moldova State University

precum Web of Science (WoS), Scopus, Google Scholar, Dimensions, BASE, EbscoHost, ProQuest și Crossref sporesc accesibilitatea și respectiv vizibilitatea producției științifice în creștere continuă, prin filtrarea celor mai relevante informații. Studenții și cercetătorii își încep căutările pe web anume cu ASEBD, care au ajuns să fie obiectivul prin care este privită știința (Haines et al, 2010). În cadrul prezentei lucrări, vom utiliza o definiție mai largă a bazelor de date bibliografice, ca set structurat de metadate bibliografice (de ex. titlu, tip de publicație, an, autor, identificatori etc.) necesare pentru calcularea numărului de publicații (și a citărilor dintre acestea).

Un număr mare de țări au dezvoltat baze de date bibliografice naționale, pentru a asigura acoperirea cât mai completă a conținutului științific elaborat în limbile naționale sau din domeniile, care nu sunt tradițional acoperite de revistele indexate în ASEBD internaționale. Căutarea publicațiilor științifice prin intermediul acestor resurse furnizează rezultate semnificativ diferite, cauzate atât de diferența numărului de înregistrări stocate în bazele de date respective, dar și, mai important, de calitatea metadatelor bibliogafice utilizate.

În linii generale, metadatele sunt date care furnizează informații despre alte date. Metadatele au diverse scopuri: ajută utilizatorii să găsească informații relevante și să descopere resurse, sprijnă organizarea resurselor electronice și furnizarea identității digitale, facilitează arhivarea și conservarea resurselor (Hare, 2016). Metadatele adecvate, exacte și furnizate în timp util contribuie la sporirea detectabilității conținutului digitital, concept devenit extrem de important în lumea interconectată de astăzi. Una dintre caracteristice funadamentale ale metadatelor este faptul că, deși pot fi lizibile de către oameni, scopul lor de bază este să fie procesate de calculator pentru căutare, sortare sau afișare. În ultimii ani, această caracteristică de bază a fost suplinită de emergența web-ului semantic, care facilitează integrarea și reutilizarea datelor între aplicații și domenii.

Metadatele se încadrează în trei mari categorii, conform NISO (Riley, 2017):

- metadate descriptive: elemente ce facilitează identificarea ori descoperirea, precum titlu, rezumat, autor, cuvinte cheie;
- metadate structurale: denotă modul în care sunt structurați anumiți itemi; de exemplu, cum sunt ordonate paginile pentru a forma capitole;
- metadate administrative: informații care permit gestionarea unor resurse, precum momentul în care este creat un item, tipul de fișier, cine l-a accesat etc.

Conform lui Riley (2017), metadatele sunt puternic integrate în comunicarea științifică digitală, inclusiv domeniul editorial, de la crearea și producerea informațiilor la marketing, descoperire, diseminare și impact, cât și cel al bibliotecilor și bazelor de date bibliografice. Această omniprezență a metadelor este determinată de un aspect crucial al lumii interconectate – interoperabilitatea, fiind definită drept abilitatea datelor și metadatelor eterogene să fie partajate între diferite sisteme (Day, 2005). Este esențial de reținut că orice tip de conținut digital, în special cel online, are două tipuri de public-țintă: pe de o parte utilizatorii în căutarea informațiilor relevante, și pe de altă parte – motoarele de căutare. Pentru exemplificare, studiul lui Walker (2016) privind corelarea dintre metadate și volumul de vânzări a demonstrat clar că, cărțile cu o descriere bibliografică, inclusiv imaginea coperții, deci metadate complete și corecte, generează un volum mediu de vînzări per ISBN cu 75% mai mare decât titlurile cu descriere incompletă.

Datorită semnificației majore a metadatelor, crearea, colectarea și partajarea acestora sunt guvernate de principii, tehnici, formate și standarde menținute și revizuite de organisme internaționale, cum ar fi Biblioteca Congresului, Comitetul de coordonare al Research Data Alliance, IFLA (International Federation of Library Associations and Institutions) și NISO (National Information Standards Organization). MARC, BIBFRAME, Dublin Core, ONIX, XML sau OAI-PMH sunt doar o mică parte dintre standardele și resursele utilizate pentru managementul medatelor (Bascones & Staniforth, 2018). Spre exemplu, Dublin Core Metadata Initiative (DCMI, http://dublincore.org) menține un set de 15 metadate de bază, împreună cu definiții ale altor termeni privind metadatele, care pot facilitatea interoperabilitatea din și între domenii.

Interoperabilitatea la nivel tehnic depinde de existența unor sintaxe (limbaje) standard, precum XML (Extensible Markup Language) și utilizarea unor protocoale de comunicare, precum standardul Z39.50, utilizat pentru căutarea în colecții distribuite ale datelor bibliografice, de tipul cataloagelor bibliotecare sau protocolul OAI-PMH.

Dintre acestea, XML, limbajul extensibil de marcare, este omniprezent pentru managementul metadatelor practic din orice domeniu de activitate. XML transformă documentele dintr-un amestec de text și obiecte într-o colecție sortabilă, ajustabilă, ierarhică de unități de conținut, iar conținutul este descris în termeni structurali și folosește etichete lizibile de calculator pentru a comunica definiții structurale, semnificații și relații. XML este prin excelență simplu, extensibil și universal compatibil, întrucât poate fi înțeles de orice program, sistem de operare, browser sau bază de date (EBSCO, 2021).

Datorită complexității și diversității formatelor, standardelor și vocabularelor care guvernează metadatele din domeniul comunicării științifice și gestionării activității de cercetare, sunt necesare eforturi colaborative pentru definirea unui vocabular consistent în domeniu, cât și pentru cartografierea metadatelor în scopul definirii cerințelor de interoperabilitate. Aceste provocări comune au determinat crearea Metadata2020 (https://metadata2020.org/) - un efort de colaborare care pledează pentru metadate mai bogate, deschise, conectate și reutilizabile pentru toate rezultatele științifice, care vor contribui la promovarea activității științifice în beneficiul societății. Reunind cei mai importanți actori cu atribuții la metadate în sectorul de cercetare (100 organizații și peste 170 de persoane), Metadata2020 vine cu un set amplu de proiecte, recomandări, publicații, bune practici și instrumente pentru furnizarea și partajarea unor metadate mai bogate pentru comunicarea științifică. Conform acestei inițiative (Metadata2020, 2020), metadatele din sectorul de cercetare trebuie să fie:

- COMPATIBILE: deschise, interoperabile, parsabile, acționabile pentru diverse sisteme, lizibile pentru oameni, pe cât este posibil;
- COMPLETE: să reflecte conținutul, componentele și relațiile, să fie comprehensive, pe cât este posibil;
- CREDIBILE: cu proveniență clară, de încredere și exacte;
- PRELUCRATE (CURATED): menținute în timp, pentru a reflecta actualizările și elementele noi.

Respectarea acestor principii pentru metadate va avea un impact semnificativ, deoarece metadatele complete facilitează detectabilitatea, descoperirile și inovația, metadatele conectate elimină decalajele dintre sisteme și comunități, iar metadatele reutilizabile elimină duplicarea efortului.

3. Identificatorii persistenți în comunicarea științifică

Ecosistemul comunicării științifice digitale în general și bazele de date bibliografice în particular au devenit asociate în timp cu o multitudine de identificatori digitali, obiectivul principal al cărora este soluționarea provocărilor privind dezambiguizarea și identificarea entităților. Peisajul identificatorilor persistenți este unul complex, interconectat și se referă nu doar la identificarea persoanelor, ci la o varietate tot mai mare de resurse științifice: publicații, seturi de date, software, metadate bibliografice, organizații. Un identificator este un număr sau o etichetă alfanumerică, care este lizibilă pentru calculator sau om și identifică în mod persistent un obiect, un document, o persoană, un loc, o organizație sau orice entitate, atât în lumea reală, cât și pe Internet (Ouvrir la science, 2020). Identificatorii persistenți reprezintă blocuri componente fundamentale ale științei deschise, asigurând relaționarea stabilă între diverse tipuri de resurse, cât și prezervarea acestora pe termen lung.

Cei mai importanți identificatori ai persoanelor includ ORCID, Scopus Author ID și ResearcherID, dar desigur nu se limitează la acestea, acoperind de asemenea diverși identificatori utilizați la nivel național. ORCID (Open Researcher and Contributor ID, https://orcid.org/) este un

registru de identificatori unici pentru cercetători, gratuit, deschis și mobil. A fost dezvoltat pentru a soluționa problema atribuirii corecte a rezultatelor științifice cercetătorilor individuali. Sistemul se bazează pe colaborarea între edituri, universități, organisme de finanțare, cercetători și alte părți interesate din domeniul comunicării științifice și este centrat pe cercetător. Mai mult, prin intermediul identificatorului persistent ORCID, cercetătorii preiau controlul asupra prezenței sale digitale, asumându-și rezultatele cercetării și conectându-le la sistemele cu care interacționează. Deoarece ORCID nu este dependent de instituția căreia îi este afiliat cercetătorul, această prezență digitală într-un format structurat urmează cercetătorul pe tot parcursul carierei, în diverse organizații. Stabilind conexiuni cu orice sistem local de management al informațiilor științifice, ORCID devine un canal de comunicare între diverse sisteme prin care circulă înregistrări privind listele de publicații, seturile de date sau informații privind finanțarea obținută.

Un alt bloc component important al ecosistemului de cercetare, care de asemenea contribuie la sporirea vizibilității, atât a autorilor, cât și a grupurilor de cercetare și organizațiilor sunt identificatorii organizaționali. Relevanța acestora rezidă în posibilitatea de identificare exactă a apartenenței instituționale, afilierii, transparenței finanțării activităților de cercetare și rezultatelor acesteia. GRID, baza de date a identificatorilor organizaționali, a fost dezvoltată de compania comercială Digital Science, însă va fuziona din punct de vedere al conținutului cu un sistem similar de identificatori dezvoltat pe bază comunitară – ROR (https://ror.org/), care furnizează identificatori unici, deschiși și sustenabili despre peste 98000 organizații de cercetare.

Identificarea publicațiilor științifice de asemenea este facilitată de un șir de instrumente, precum DOI, ISNI, ISSN, ISBN, Handle etc. Datorită exploziei fără precedent a publicațiilor electronice, cel mai utilizat în prezent este identificatorul obiectului digital DOI, care din 2012 a devenit standard internaţional ISO 26324. Ca identificator, DOI este unic prin faptul că poate identifica un produs, o lucrare sau o parte componentă, precum capitol, tabel sau diagramă; poate fi alocat atât versiunior tipărite ale lucrărilor, cât și celor electronice; este acționabil, conectându-se la orice resursă specificată de editor. DOI în sine este un identificator permanent, dar metadatele asociate acestuia, precum și conexiunea pe care o asigură, pot fi modificate de către proprietar în orice moment (Warren, 2015). Datorită DOI, sistemele de informații științifice instituționale pot face tranziția de la sisteme orientate în interior spre rețele de informații capabile să facă conexiunea cu numărul actualizat de citări și indicatori altmetrici ai publicațiilor.

Identificatorii permit motoarelor de căutare și resurselor din ecosistemul de comunicare științifică să găsească, să consume, să facă schimb și să traseze mai ușor legăturile între publicații și alte rezultate digitale asociate acestora. Toți acești identificatori se bazează pe metadate și evoluează continuu pentru a facilita interoperabilitatea și interacțiunea dintre sisteme. Astfel, dacă la inregistrarea DOI pentru publicații, editorul înregistrează metadatele privind ORCID, atunci ulterior, profilul ORCID al cercetătorului este automat completat cu publicația care deține identificatorul DOI respectiv. Iar sistemul de managament al revistelor științifice recenzate OJS (Open Journal System), dezvoltă continuu interfețe pentru schimb automat de metadate cu unele dintre cele mai relevante resurse și baze de date bibliografice, precum DOAJ (Directory of Open Access Journals) sau Index Copernicus.

4. Recomandări privind metadatele

În baza celor enunțate poate fi conturat un set de recomandări, relevante pentru părțile interesate din sistemul comunicării științifice digitale, inclusiv componentele acestuia precum bazele de date bibliografice:

- Utilizarea identificatorilor persistenți devine obligatorie la toate etapele ciclului de viață al unei publicații științifice și la toate nivelurile procesului editorial. Facilitând dezambiguizarea și identificarea unică a entităților din cadrul comunicării științifice, și respetiv a sistemului de cercetare-inovare, identificatorii și în special interconectarea acestora (ex. OpenAIRE

ResearchGraph) permit generarea unei imagini cât mai veridice privind producția științifică la nivel global.

- Informațiile privind drepturile de autor trebuie incluse în metadatele la nivel de publicație individuală, acestea devenind astfel mai detectabile, în special pentru revistele cu acces deschis și totodată reprezintă una dintre cerințe de conformare cu Planul S (https://www.coalition-s.org/about/), o inițiativă globală menită să consolideze și să accelereze tranziția la acces deschis. Multe motoare de căutare și servicii de descoperire (inclusiv Google Scholar) sprijină căutarea și filtrarea după licența drepturilor de autor, pentru a ajuta utilizatorii să găsească conținut în acces deschis, iar Creative Commons – CC (Licențele Creative) a lansat propriul său motor de căutare pentru a ajuta utilizatorii să găsească conținutul, care poate fi reutilizat sub o licență CC.
- Rezumatele tuturor rezultatelor științifice ar trebui să fie disponibile în mod deschis, prin intermediul depozitelor de încredere, în formate lizibile de calculator. Acesta este dezideratul promovat de „Inițiativa pentru rezumate deschise" (I4OA, https://i4oa.org/), care încurajează editorii să înregistreze rezumatele deschise în calitate de metadate pentru înregistrarea DOI la Crossref, pe cât este posibil. Adoptarea acestei prevederi va spori semnificativ detectabilitatea, vizibilitatea și impactul publicațiilor, dar și va facilita mineritul datelor și a textului din rezumate, contribuind la dezvoltarea instrumentelor de cercetare din domeniul Învățării Automate și Inteligenței Artificiale.
- Citările publicațiilor ar trebui să fie deschise și incluse în metadatele publicațiilor, întrucât acestea reprezintă un bun public, de importanță atât pentru comunitatea academică, cât și pentru publicul larg. Inițiativa pentru citări deschise (I4OC, https://i4oc.org) își propune să promoveze disponibilitatea datelor despre citări, astfel încât acestea să fie structurate (exprimate în formate comune, care pot fi accesate și citite automat), separabile (citările pot fi accesate și analizate fără a accesa articolele sau cărțile în care sunt indicate) și deshise (accesibile și reutilizabile). Aceasta sporește detectabilitatea, potențialul pentru analiză și interconectare, inclusiv facilitează dezvoltarea instrumentelor care sprijină cercetătorii în identificarea rapidă, analiza și urmărirea conexiunilor din cadrul activității științfiice.

5. Concluzii

Metadatele și identificatorii persistenți facilitează fluxurile de lucru și tranzițiile între sisteme, susțin analiza și detectabilitatea și oferă tuturor părților interesate o perspectivă mai largă asupra activității de cercetare. Anume metadatele facilitează procesarea interogărilor și fac distincția între sursele de informații legitime și cele dubioase într-o lume în care volumul informațiilor și complexitatea căutărilor cresc exponențial. Astfel, metadatele țes tapiseria complexă a informațiilor din era digitală, prin interconectarea producătorilor, distribuitorilor și utilizatorilor de conținut (Warren, 2015).

Crearea și gestionarea metadatelor reprezintă o activitate-cheie și o responsabilitate comună a producătorilor, distribuitorilor și utilizatorilor de conținut, la fel de importantă ca și conținutul în sine, iar principiul GIGO „Garbage In, Garbage Out" este perfect aplicabil pentru metadate. Prin urmare, este important ca toți producătorii, distribuitorii și utilizatorii de conținut științific să adopte și să utilizeze standarde și vocabulare relevante, să reducă pe cât posibil factorul erorii umane în procesul de generare și procesare a metadatelor, prin implementarea unor sisteme și instrumente software pentru generarea și gestionarea metadatelor, în scopul asigurării interoperabilității acestora. Metadatele deschise, relevante, complete și actuale determină o vizibilitate sporită pentru conținutul științific, contribuind la o detectabilitate mai exactă și asigurând o prezență digitală de încrede la nivel de autor, organizație, domeniu științific sau țară.

Referințe

1. Bascones, M. & Staniforth, A. (2018). What Is All This Fuss About? Is Wrong Metadata really Bad for Libraries and Their End-users? *Insights*, 31: 41. http://doi.org/10.1629/uksg.441

2. Day, M. (2005). Metadata. In: S.Ross, M.Day (Eds), *DCC Digital Curation Manual*. Digital Curation Center.
3. Haines, L. L., Light, J., O'Malley, D., & Delwiche, F. A. (2010). Information-seeking behavior of basic science researchers: implications for library services. *Journal of the Medical Library Association : JMLA*, 98(1), 73–81. https://doi.org/10.3163/1536-5050.98.1.019
4. Hare, J. (2016). *What is metadata and why is it as important as the data itself?* OpenDataSoft. https://www.opendatasoft.com/blog/2016/08/25/what-is-metadata-and-why-is-it-important-data
5. *Open identifiers for open science*. (2020). Ouvrir la science, France. https://www.ouvrirlascience.fr/open-identifiers-for-open-science/
6. Riley, J. (2017). *Understanding Metadata: What is Metadata, and What is it For?*: A Primer. NISO.
7. Science, D., & Porter, S.. (2016). *Digital Science White Paper: A New 'Research Data Mechanics'* (Version 1). Digital Science. https://doi.org/10.6084/m9.figshare.3514859.v1
8. *The XML – First Competitive Advatage*. (2021). EBSCO. https://www.ebsco.com/apps/assets-publishers-materials/Advantages_of_XML.pdf
9. Walker, D. (2016). *Nielsen Book US Study: The Importance of Metadata for Discoverability and Sales*. The Nielsen Company US, LLC.
10. Warren, J. (2015). Zen and the Art of Metadata Maintenance. *The Journal of Electronic Publishing*, 18(3). https://doi.org/10.3998/3336451.0018.305