

METADATA

Premise

Dezvoltarea explozivă a rețelei Internet înregistrată în ultima perioadă a condus la o extindere considerabilă a documentelor în formă electronică.

De asemenea, bibliotecile parcurg din ce în ce mai evident și mai rapid drumul de la biblioteca tradițională la biblioteca digitală care manipulează colecții mari de tip electronic în cadrul unui sistem de partajare atât a resurselor (achiziție, catalogare partajată, schimb de date, etc.), cât și a serviciilor oferite utilizatorilor (navigare Internet, împrumut interbibliotecar, furnizare la distanță a documentelor electronice, etc.).

Potențialurile de localizare a resurselor sunt încă reduse și ineficiente. Modul de organizare a informației este neuniform și, în consecință, greu manevrabil și laborios de administrat. De aceea, procesul de regăsire a informației relevante pentru un anumit domeniu de interes la un moment dat este dificil. Cunoscutele motoare de căutare precum Altavista, Lycos, Yahoo, Metacrawler, Ecila, etc. permit indexarea automată a resurselor disponibile pe Web și actualizează con-

tinuu bazele de date ale locațiilor. Acest tip de indexare este util în cazul colecțiilor relativ mici pe un anumit domeniu. În momentul în care numărul resurselor depășește o anumită limită, atunci apar probleme cu seturile mari de referințe și probleme legate de intersecția și suprapunerea interdisciplinară din punct de vedere semantic a termenilor folosiți. Înregistrările complexe necesită metode de căutare și regăsire complexe. Standardele formale ca TEI sau MARC ar putea satisface cerințele dar necesită un consum de timp mare pentru crearea și întreținerea înregistrărilor.

O soluție pentru rezolvarea acestei probleme, care să medieze cele două extreme, implică crearea unei înregistrări care să fie mai informativă decât o intrare de index, dar mai puțin completă și complexă decât o înregistrare de catalogare.

Pe baza acestor considerente s-a ajuns la introducerea conceptului de metadata.

Ce este metadata?

Metadata reprezintă date despre date.

Metadata este data care descrie atributele unei resur-

Voichița Dragomir
B.C.U. București
Oficiul de Calcul

■ METADATA ■

se. Poate fi deci o dată de tip bibliografic sau poate include descrieri cu referire la conținutul, termenii și condițiile de utilizare, acoperire și caracteristici tehnice sau de acces pentru o anumită resursă de tip electronic.

Metadata este instrumentul prin care utilizatorii, fie ei oameni sau programe, regăsesc o resursă (structurată pe nivele) sau alta fără a-i cunoaște a priori caracteristicile.

În contextul bibliotecii digitale, constituită din multiple componente denumite obiecte digitale, metadata este o metodă de structurare a informației în formă digitală.

Așadar:

- **metadata** este termenul care definește orice tip de date utile pentru identificarea, descrierea, gestiunea, localizarea resurselor electronice și furnizarea accesului la obiecte. În acest context există o varietate de formate specifice metadatelor, începând cu descrierile bibliografice, descriptorii, cuprinsul, indeșii simpli nestructurați, apoi diferite formate adecvate mediului Web ca formatele PICS - Platform for Internet Context Selection și Dublin Core, ajungând la formate structurate de tip Marc, TEI (Text Encoding Initiative) și la formatele complexe ca

EAD - Encoded Archival Description sau CIMI - Computer Interchange of Museum Information utilizate pentru materialele de tipul arhivelor;

- **metadata structurală** definește tipurile, versiunile, relațiile și alte caracteristici ale materialelor digitale;
- **metaobiectul** este obiectul ce furnizează referințe la un set de obiecte digitale (un metaobiect este de fapt o listă de identificatori ai obiectelor digitale - aşa numitele „handle”) și asupra căruia se efectuează procesarea pentru identificare, selectare, evaluare și accesare;
- **mediul meta informațional** reprezintă multitudinea serviciilor de informare puse la dispoziția utilizatorilor, local și la distanță, împreună cu toate mijloacele de coordonare în folosirea acestor servicii care permit accesul, evaluarea și utilizarea oricărei resurse de informații în conformitate cu standardele și protocolele în vigoare.

Structura informației digitale

Structura informației digitale poate fi descrisă prin câteva aspecte ce au fost relevante în cadrul sistemului experimental inițiat în 1995 la Library of Congress prin proiectul NDLP (National Digital Library Project) de convertire a documentelor istorice într-o

formă digitală accesibilă prin Internet:

- obiectele digitale sunt adesea în directă legătură sau în relație cu alte obiecte din perspectiva parte/intreg (exemplu: textul digitizat care poate fi constituit din pagini, capitole, index, ilustrații, etc. sau un program software compus dintr-un anumit număr de fișiere);
- un obiect digital poate fi stocat în diferite formate cu posibilitatea de convertire dintr-un format într-altul (exemplul unei imagini stocate sub format gif și convertită într-un format jpg sau tiff sau o pagină text reprezentată în format SGML sau PostScript);
- versiuni diferite, actualizate continuu (exemplul paginii WEB a unei instituții);
- drepturi și restricții asupra informației (dreptul de copyright, număr licențe de folosire);
- posibilități tehnice de accesare a informației în funcție de caracteristicile și performanțele sistemului informatic constituit din rețele de calculatoare LAN sau WAN, a calculatoarelor integrate în rețea (viteză de transmisie, dimensiunea materialului digital);

Indiferent de caracteristicile și structura informației digitale, aceasta trebuie să aibă o identitate aparte prin: calitate, siguranță, continuitate, acuratețe, valoare, pertinență, accesibilitate, u-

■ METADATA ■

nicitate și diversitate.

Arhitectura sistemului informațional bazat pe informația digitală se sprijină pe câteva principii:

- serviciile oferite utilizatorilor trebuie să fie flexibile, adaptabile pentru orice nivel de pregătire și competență profesională;
- colecțiile trebuie să fie ușor de gestionat și manevrat;
- arhitectura informației trebuie să ofere posibilitatea de a selecta informația valoroasă, de a furniza termenii și condițiile de utilizare, de a asigura securitatea în momentul transmitemii ei prin rețelele de informare.

Arhitectura informației utilizează, la rândul ei, trei concepte de bază:

- tipul de date (descrie proprietățile tehnice ale datei: format sau metodă de procesare);
- metadata structurală (descrie tipurile, versiunile, relațiile, etc);
- metaobiectul (obiectul ce furnizează referințe către un set de obiecte digitale).

Scopul

Scopul principal al demersului de definire a metadatei este de a identifica un set cât mai simplu de elemente pentru descrierea resurselor electronice disponibile în rețea, mai

precis acele elemente care ar duce la localizarea cu precizie a resursei, independent de subiectul sau complexitatea acesteia.

Astfel s-a constituit setul **Dublin Core Metadata Element Set** sau pe scurt

Dublin Core care propune un număr minimal de elemente necesare pentru a facilita descoperirea obiectelor de tip document (Document like objects) într-un mediu de rețea precum Internet.

Setul a fost elaborat de directoratul Dublin Core găzduit de Departamentul de Cercetare și Proiecte speciale a corporației OCLC. Lucrarea a fost finalizată în decembrie 1996. Toate elementele au caracter descriptiv, unora dintre ele fiind preferabil să li se asociază un vocabular controlat.

Caracteristici

- simplitate, astfel încât să poată fi utilizat atât de specialiștii în descrierea resurselor cât și de necatalogatori;
- interoperabilitate semantică;
- consens internațional, prin participarea activă la promovarea și experimentarea setului a peste 20 țări din America de Nord, Europa, Australia și Asia;
- extensibilitate și flexibilitate prin faptul că oferă o alternativă pentru modelele de descrie-

re mai elaborate cum ar fi formatul Marc;

- modularitate pe Web, consorțiul W3C inițiind deja implementarea formatului RDF specific utilizării pe WWW.

Elemente

1. Titlu: numele obiectului atribuit de autor sau editor;
2. Autor: persoana sau organizația responsabilă pentru conținutul intelectual al obiectului;
3. Alt agent: persoanele precum editorii, traducătorii, redactorii sau ilustratorii care au realizat contribuții intelectuale semnificative pentru lucrare;
4. Editor: agentul sau agenția care asigură disponibilitatea obiectului în rețea;
5. Data: data publicării;
6. Descriere: descrierea conținutului resursei, inclusiv rezumatele;
7. Subiect și cuvinte cheie: domeniul tratat în lucrare (de preferat să se utilizeze vocabulare controlate);
8. Tip obiect: de exemplu dicționar, nuvelă, curs, raport, teză;
9. Forma: de exemplu fișier postscript sau fișier executabil Windows;
10. Identificator: sir de caractere sau număr utilizat pentru identificarea unică a obiectului;
11. Relație: relații cu alte obiecte cum ar fi capitolul dintr-o carte;
12. Sursa: obiecte, fie tipărite sau electronice, din care obiectul descris derivă dacă este cazul;

■ METADATA ■

13. Limba: limba conținutului;
14. Acoperire: localizarea spațială și temporală a obiectului;
15. Drepturi de copyright.

Elementele au fost definite ținând cont de proprietățile intrinseci ale obiectului descris și de posibilitățile de extindere a informațiilor. Elementele sunt optionale, repetabile, modificabile, în funcție de necesități, și independente sintactic.

Obiectivul inițial propus de diferitele organisme care au demarat proiecte referitoare la Metadata, organisme ca OCLC (Online Computer Library Center), NCSA(National Center for SuperComputing Applications) sau Library of Congress, acela de a specifica elementele necesare pentru descoperirea obiectelor de tip document - așa numitele DLO-uri (Document Like Objects), va fi dezvoltat în diferite direcții:

- extinderea setului Dublin Core pentru includerea altor tipuri de obiecte ca servicii sau colecții;
- extinderea setului Dublin Core în vederea abordării unor aspecte funcționale noi, altele decât cele de descoperire a resurselor, cum ar fi: controlul arhivării, autentificarea utilizatorilor, mecanisme de plată;
- stabilirea unor metode standard pentru extensibilitate.

În prezent există câteva formate deja definite ca răspuns la mobilitatea ne-predictibilă a resurselor în rețea Internet și anume formatele URN și PURL.

Proiecte

Pentru punerea în practică a conceptelor menționate anterior s-au dezvoltat o serie de proiecte interesante. Printre acestea:

- **Proiectul de catalogare a Resurselor Internet** al corporației OCLC, care are ca scop crearea, implementarea, testarea și evaluarea unei baze de date bibliografice pentru materialele accesibile prin Internet în format USMARC completat prin câmpul 856, definit pentru publicațiile globale, cu informații referitoare la posibilitățile de acces și la localizarea electronică.

- **Proiectul American Memory** al Bibliotecii Congresului în cadrul inițiativei de constituire a **Colecțiilor digitale** - National Digital Library Project (NDLP)

- **Proiectul eLIB** - electronic Library care reunește o serie de biblioteci academice din Marea Britanie

- **Proiectul DLI** - Digital Library Initiative susținut și sponsorizat de NSF - National Science Foundation, DARPA - Department of Defense Advanced Research Projects Agency și NASA - National Aeronautics and Space Administration, a fost inițiat în 1995

și include 6 proiecte susținute de 6 importante universități: University of California, Berkley; University of California, Santa Barbara; University of Illinois, Urbana - Champaign; University of Michigan; Stanford University și Carnegie Mellon University având ca scop cercetarea mijloacelor avansate de colectare, stocare, organizare și utilizare pe scară largă a resurselor distribuite de cunoștințe ce înmagazinează informație pe diferite suporuri electronice.

Programul își propune să anticipateze evoluțiile tehnologice ale unei biblioteci digitale și reunește efortul unor echipe multidisciplinare de cercetători pentru a dezvolta și implementa serviciile, performanțele și capabilitățile bibliotecii digitale la nivel global.

Printre partenerii principali care susțin din punct de vedere tehnologic proiectele se numără: Xerox, IBM, Hewlett-Packard, Digital Equipment, Sun, Illustra, Oracle, etc.

Partenerii care furnizează conținutul informațional includ: Elsevier, John Wiley, MIT Press și alții.

Proiectele studiază o varitate mare de probleme care includ interoperabilitatea semantică, interoperaabilitatea și compatibilitatea între diversele formate și protocoale de lucru, căutarea prin imagini a conținutului, noi modele pentru

■ METADATA ■

documentele distribuite, strategii de management, aspecte sociale și economice ale bibliotecilor digitale, noi forme de colaborare, noi reprezentări ale cunoștințelor, instrumente de cercetare și navigare multilingve.

Toate proiectele construiesc platforme de testare, să numitele „testbeds” pentru activități de cercetare și dezvoltare în scopul creării fundamentului pentru o bibliotecă digitală distribuită operațională (distribuită însemnând că părțile componente ale bibliotecii pot fi deopotrivă, răspândite pe Internet sau pot coexista în cadrul aceluiași sistem).

Sistemele de testare sunt construite pentru a asigura verificarea ideilor teoretice pe un suport practic și pentru a da ulterior tuturor posibilitatea de accesare și participare la efortul intelectual al comunității DLI.

Cele 6 proiecte dezvoltă abordări diferite.

Proiectele de la Illinois și Berckley dezvoltă sisteme complexe cu un număr mare de utilizatori, Illinois concentrându-și atenția asupra documentelor text structurate manual, iar Berckley asupra recunoașterii automate a documentelor imagini. Proiectele Carnegie Mellon și Santa Barbara studiază posibilitățile de manipulare a celor mai noi medii care până nu demult

erau imposibil de indexat și cercetat: Carnegie Mellon investighează segmentarea și indexarea documentelor video prin utilizarea recunoașterii de voce, Santa Barbara indexează hărți prin utilizarea procesării automate a imaginilor (Proiectul ADL - Alexandria Digital Library). Stanford și Michigan studiază elementele de legătură intermediare necesare pentru operarea proceselor la nivelul bibliotecilor digitale globale, încercând să găsească reprezentările convenabile pentru interoperarea între formatele folosite de diferite servicii de căutare și să identifice sursele cele mai potrivite pentru o cercetare în conformitate cu o anumită cerere de căutare.

Proiectul DLI va continua cu faza II.

Alte proiecte de dezvoltare a bibliotecii digitale și experimentare a setului Dublin Core ce se derulează în Statele Unite ale Americii, Australia și Europa sunt:

SUA

CORC (Cooperative Online Resources Cataloguing)
<http://www.oclc.org/research/projects/corc/index.htm>

Proiectul de cercetare CORC va explora posibilitățile de creare, în cooperare, a unui catalog al resurselor Internet. Printre obiectivele proiectului se numără:

- integrarea într-un singur sistem a formatelor Marc și Dublin Core;
- import-export Marc și Dublin Core;
- import-export în format RDF/XML;
- integrarea metadatelor în regim de partajare la nivel local și global.

Digital Library Catalog

<http://sunsite.berkeley.edu/Catalog>

Proiectul include cărți, discursuri și alte materiale de tip text în format HTML, rapoarte tehnice (în diferite formate), fotografii, gravuri și alte materiale vizuale, clipuri video și audio.

Everglade Information Network & Digital Library

<http://everglades.fiu.edu/>

Colecția va include rapoarte tehnice, comunicări științifice, programe ale conferințelor, hărți, diapoziitive, fotografii, documente legale și de arhivă acoperind, în principal, subiecte legate de cercetări ecologice.

Florida International University Digital Library

<http://www.fiu.edu/diglib/>

Proiectul de bibliotecă digitală se va concentra asupra documentelor de tip imagine, audio și video inclusiv prezentări multimedie și module curriculare, acoperind orice subiect abordat în procesul de predare și cercetare al universității.

■ METADATA ■

Gateway to Educational Materials

<http://gem.syr.edu>

GEM este o inițiativă a Departamentului american pentru Educație și Biblioteca Națională pentru Educație. Obiectivul este de a îmbunătăți organizarea și accesibilitatea colecțiilor substantiale, dar necatalogate, de materiale educaționale existente disponibile pe diferite site-uri universitare, federale și comerciale pe Internet.

Mantis

<http://purl.oclc.org/mantis>

Mantis este un proiect de cercetare al corporației OCLC ce testează utilizarea și integrarea formatelor: Dublin Core, RDF, XML, Scorpion și Kilroy.

Monticello Electronic Library

<http://www.solinet.net/monticello/>

Biblioteca digitală Monticello are ca obiectiv principal realizarea legăturii între resursele distribuite regionale indiferent de sursa sau tipul informației. Este utilizat setul Dublin Core pentru asigurarea interoperabilității semantice între bazele de date ce gestionează informație pe suport electronic și pe diferite formate ca SGML, EAD, MARC, etc.

Australia DSTC (Distributed Systems Technology Centre)

<http://www.dstc.edu.au/RDU/>

Centrul DSTC participă în grupul de lucru RDF al consorțiului W3C pentru dezvoltarea unei specificații a setului **Dublin Core** în RDF, a instrumentelor necesare utilizatorului pentru crearea RDF compatibil Dublin Core și a unor interfețe de căutare.

EdNA (Education Network Australia)

<http://www.otfe.vic.gov.au/edna/dc5edna.html>

EdNA este un proiect cooperativ între toate statele australiene și toate sectoarele de educație: școli, licee, universități. EdNA folosește standardul Dublin Core.

The PANDORA Project

<http://www.nla.gov.au/policy/pandje97.html>

Biblioteca Națională a Australiei (NLA) dezvoltă o arhivă electronică numită PANDORA care să furnizeze accesul pe termen lung la publicațiile australiene online. Pentru a eficientiza căutarea online, proiectul va implementa un sistem de descriere a documentelor arhivate bazat pe atributele setului Dublin Core.

Europa

DONOR (Directory of Netherlands Online Resources)

<http://www.konbib.nl/donor/index-en.html>

Obiectivul este de a crea o infrastructură validă pentru managementul și regăsi-

rea informației pe rețea națională de cercetare a Olandei-SURFnet. DONOR va permite gestiunea coordonată a documentelor și metadatelor.

Electronic Library Image Service for Europe (ELISE II)

<http://severn.dmu.ac.uk/elise/>

Serviciul ELISE va opera pe baza modelului client/server, utilizând standardele Z39.50 și Dublin Core.

EULER (European Libraries and Electronic Resources in Mathematical Sciences)

<http://www.emis.de/projects/EULER/>

Obiectivul proiectului EULER finanțat de Comunitatea Europeană este de a furniza accesul orientat utilizator, integrat în rețea, la publicațiile matematice. Prin utilizarea descrierilor bazate pe Metadata (setul Dublin Core), EULER va integra: baze de date bibliografice, cataloage publice online de bibliotecă, publicații electronice de nivel academic, arhive online de literatură gri, îndepărtând resurselor Internet din domeniul matematicii.

The German Educational resources Server (Deutscher Bildungs-Server)

<http://dbs.schule.de/indexe.html>

Site-ul cuprinde aproximativ 2000 documente Web reprezentând materiale educaționale de la studenți, profesori, editori și autorități de stat de învățământ.

■ METADATA ■

INDOREG: Internet Document REGistration

<http://www.purl.dk/rapport/html.uk/>

Proiectul Centrului de Biblioteci Daneze (DBC) are ca scop înregistrarea tuturor publicațiilor internet care se încadrează în criteriile de selecție și să ofere accesul la aceste documente prin sistemul DanBib.

Metadata Project

<http://www2.sub.uni-goettingen.de>

Proiectul explorează utilizarea metadatei din perspectiva bibliotecii și analizează impactul dezvoltărilor înregistrate în domeniul regăsirii resurselor de informație în mediul de rețea asupra regulilor tradiționale de catalogare. Proiectul este o parte dintr-un proiect mai mare ce include câteva biblioteci și instituții germane.

The Nordic Metadata Project

<http://linnea.helsinki.fi/meta/>

Tările nordice dezvoltă un proiect de creare și implementare a unui sistem de partajare a metadatelor care, bazat pe setul Dublin Core, să furnizeze servicii utilizator mult îmbunătățite.

Proiectul BIBLINK

<http://hosted.ukoln.ac.uk/biblink/>

Este un proiect bazat pe cooperarea a câteva biblioteci naționale și Uniunea Europeană, având ca scop stabilirea unei legături electronice între editori și Agențiile Naționale Bibliografice în vederea schimbului de înregistrări de tip metadata între aceștia.

Proiectul DESIRE

<http://www.nic.surfnet.nl/surfnet/projects/desire/desire.html>

DESIRE abordează două căi de regăsire a informației: servicii de regăsire după subiect, bazate pe

selecție manuală și descrierea resurselor de calitate pe de o parte, servicii de regăsire la nivel regional bazate pe metadate generate automat de motoarele de căutare web.

Proiectul SAFARI

<http://safari.hsv.se/index.html.en>

SAFARI (the Spreading of research information to the general public over the Internet) este un sistem pentru diseminarea pe Internet a informației științifice în Suedia.

Proiectul SCRAN

<http://www.scran.ac.uk>

SCRAN (Scottish Cultural Resources Access Network) este un proiect dedicat constituirii unei rețea care să furnizeze accesul la resurse multimedia pentru activitățile educationale, de cercetare și studiu a istoriei și culturii scoțiene.

GLOSAR

Ecila	Motor de căutare pe Internet	http://www.ecila.com/
ANSI	American National Standard Institute	
Altavista	Motor de căutare pe Internet	http://www.altavista.digital.com
CIMI	Computer Interchange of Museum Information	
	Înregistrările CIMI sunt construite pe baza formatului SGML pentru informația specifică muzeelor.	
DLO	Document Like Objects - obiecte tip document	
DTD	Document Type Definition - descrierea unui tip anumit de limbaj de marcate, descriere alcătuită prin folosirea SGML	
EAD	Encoding Archival Description - format bazat pe SGML dezvoltat pentru descrierile utilizate în arhive.	

■ METADATA ■

EDI	Electronic Data Interchange - Format pentru schimbul mesajelor structurate în vederea validării tranzacțiilor între diferitele sisteme
HTML	Hyper Text Markup Language - limbaj de tip DTD SGML care permite marcarea textului astfel încât să fie publicat online cu fonturile și legăturile corespunzătoare
HTTP	HyperText Transfer Protocol - protocol aflat deasupra protocolului TCP/IP, utilizat de servere și clienți pentru schimbul documentelor hypertext
Infoseek	Motor de căutare pe Internet http://infoseek.go.com
LAN	Local Area Network - rețea locală de calculatoare
Lycos	Motor de căutare pe Internet http://www.lycos.com
MARC	MAchine Readable Cataloguing - format bazat pe ISO 2709 pentru schimbul informației bibliografice în formă citibilă de calculator
Metacrawler	Motor de căutare pe Internet http://www.metacrawler.com/
Metadata	Date care furnizează informație despre o resursă
NISO	American Information Standards Organisation
PICS	Platform for Internet Context Selection format adecvat mediului Web
Postscript	Limbaj de descriere pagina (PDL - Page Description Language) folosit pentru descrierea textelor și imaginilor ce urmează a fi tipărite.
PURL	Persistent Uniform Resource Locator - din punct de vedere funcțional este un URL care însă, în loc să pointeze direct către locația unei resurse Internet, face trimiterea către un serviciu intermedier care găsește URL-ul și îl retrasmite către client. În limbaj Web procedura se numește redirectare standard HTTP.
RDF	Resource Description Framework Este un format în dezvoltare în cadrul consorțiului W3C și va furniza suportul pentru utilizarea metadatei în diferite domenii (regăsirea resurselor, proprietate intelectuală, etc.)
SGML	ISO 8879 Standard Generalized Mark-up Language Standard internațional pentru descrierea textului electronic.
TEI	Text Encoding Initiative - format pentru codarea și schimbul textelor citibile de calculator
URL	Uniform Resource Locator - format standard folosit pentru specificarea localizării unui obiect în mediul WWW
URN	Uniform Resource Name Este o resursă internet care permite utilizatorului să regăsească o anumită informație prin intermediul unei agenții sau a unui director care este capabil să ofere informații multiple despre localizarea elementului căutat. Așadar nu este necesară cunășterea adresei exacte a resursei ca în cazul URL.
Z39.50	Standard ANSI/NISO pentru regăsirea informației prin căutarea în baze de date eterogene distribuite pe diferite servere. Se bazează pe modelul Client/Server.
XML	Extensible Markup Language - versiune prescurtată a formatului SGML, proiectată pentru utilizarea în Internet.
Yahoo	Motor de căutare pe Internet http://www.yahoo.com
WAN	Wide Area Network - rețea de arie largă
W3C	World Wide Web Consortium - Organizație responsabilă cu dezvoltările Web