CUPRINS

OBIECTIVE GENERALE ALE PROIECTULUI

Prin cercetările ce se vor întreprinde pe parcursul derulării proiectului TEMACOD, acesta îşi propune realizarea următoarelor **obiective generale**:

a) Îmbunătățirea, coordonarea şi eficientizarea proceselor de digitizare a resurselor informaționale din domeniul tehnic din universități şi alte tipuri de organizații;
b) Sporirea numărului de resurse informaționale tehnice reprezentative digitizate, diversificarea şi conservarea acestora;
c) Creşterea gradului de accesibilitate a publicului utilizator de Internet la resursele informaționale tehnice.

Pentru atingerea acestor obiective proiectul **TEMACOD** are în vedere şi realizarea următoarelor **obiective specifice** în cadrul fiecărui obiectiv general:

Obiectivul general (a)

1. Identificarea sistemelor de informare şi documentare care vehiculează documente tehnice si coordonarea modalităților de digitizare a resurselor informaționale în organizații;
2. Soluționarea şi coordonarea aspectelor logistice şi documentare pe care le presupune digitizarea pentru înlăturarea paralelismelor si suprapunerilor, pe fiecare arie tematică în parte;
3. Crearea unor mecanisme de coordonare, gestionare şi întreținere a procesului de digitizare, care să permită o arhitectură coerentă a întregului sistem, la nivel de proiect.

Obiectivul general (b)

1. Identificarea unui segment documentar reprezentativ de resurse informaționale, pe fiecare arie tematică;
2. Dezvoltarea unui sistem performant şi eficient în domeniul tehnologiei informației, care să permită crearea unei platforme unice şi unitare de acces la resursele digitizate, precum şi conservarea materialului digitizat.

Obiectiv general (c)

1. Aplicarea reglementărilor legale referitoare la drepturile de autor şi la proprietatea intelectuală a resurselor informaționale;
2. Instituirea unui sistem unic si unitar de catalogare şi/sau clasificare şi descriere a resurselor informaționale, pe fiecare arie tematică în parte.

Programul național de informatizare a societății româneşti, cu un sens mai larg de integrare a culturii şi ştiinței naționale în circuitele mondiale, generează în mod natural şi logic un nou tip de acces la valorile tehnice şi ştiințifice din țară şi din lume, un nou tip de relații între marea categorie a celor dornici de a şti şi grupul restrâns ca dimensiuni al celor care oferă informații: instituțiile educaționale, de cercetare, de cultură şi artă.

În cadrul acestui program, un rol important revine procesului de informatizare a bibliotecilor, care a trecut de la faza de înregistrare a datelor în cataloage, la cea de dezvoltarea de sisteme integrate de bibliotecă. În prezent ne aflăm în etapa de constituire şi dezvoltare a colecțiilor de date digitale în biblioteci, care va permite ulterior interconectarea bibliotecilor digitale în rețele, bazate pe infrastructuri de comunicații.

Proiectul propus în cadrul Programului 4, „Parteneriate în domeniile prioritare": *Tehnici pentru managementul continutului digital-TEMACOD* are ca obiective principale: dezvoltarea de biblioteci digitale, bazate-Web, la nivelul instituțiilor de învățământ superior implicate în contract, interconectarea lor şi accesarea rapidă via Internet, printr-o interfață prietenoasă şi eficientă.

Tema abordată este de mare actualitate în întreaga lume, se pune problema construirii unei Biblioteci digitale la nivel mondial, accesibilă prin Internet, gratuită şi în format multilingv, care să conțină materiale semnificative.

În constituirea acestei biblioteci digitale mondiale, un rol important va reveni mediului academic, bogat în resurse educaționale, de cercetare şi care are deja o mare parte din fondul infodocumentar digitizat.

Spre deosebire de bibliotecile tradiționale, o bibliotecă digitală are colecțiile stocate în format digital şi accesibile prin calculatoare conectate local sau la distanță, prin rețele de calculatoare. De asemenea, ea oferă un sistem performant de regăsire a informației.

Ca rezultat al dezvoltării Internetului şi a potențialului său de căutare, bibliotecile digitale cum ar fi Biblioteca Europeană şi Biblioteca Congresului sunt acum dezvoltate într-un mediu bazat-Web. Bibliotecile publice, cele şcolare şi academice sunt de asemenea capabile să dezvolte website-uri care să permită preluarea de informație ştiințifică digitizată, din categoria cărților scrise, a celor audio sau video, schimbând fundamental noțiunea de resursă educațională.

Multe biblioteci academice sunt activ implicate în constituirea depozitelor de cărți, buletine ştiințifice, teze de doctorat şi alte lucrări care pot fi digitizate sau au fost de la început preluate în format digital. Aceste depozite sunt accesibile în rețeaua universității, dar şi în afara ei, cu anumite restricții, ce se referă în special la protejarea rezultatelor cercetării şi care sunt în concordanță cu politica adoptată de instituție privind accesul liber.

OBIECTIVELE ETAPEI DE EXECUȚIE

ACTIVITATE: Analiza formatelor documentelor tehnice pe suport electronic

Obiectiv: Identificarea tipurilor de formate în care se regăsesc documentele pe suport electronic

ACTIVITATE: Analiza normelor de conversie a conținutului cuprins în documentele tehnice de la formatul tradițional la formatul electronic prin tehnici de digitizare

Obiectiv: Identificarea celor mai noi tehnici de trecere a documentelor din format tradițional în format electronic. Definirea instrumentelor pentru realizarea depozitelor de conținut digital.

ACTIVITATE: Analiza conținutului web; analiza formatelor documentelor tehnice regăsite pe web

Obiectiv: Realizarea unui studiu asupra tipurilor de conținut web şi a formatelor documentelor tehnice

ACTIVITATE: Elaborarea modelului conceptual al sistemelor de informare şi documentare cu conținut tehnic.

Obiectiv: Elaborarea principiilor de bază şi a metodologiei identificării şi creării sistemelor infodocumentare pentru documente cu conținut tehnic.

REZUMATUL ETAPEI

Proiectul de cercetare TEHNICI PENTRU MANAGEMENTUL CONTINUTULUI DIGITAL şi-a propus în etapa 2, conform Planului de activităţi, atingerea obiectivelor prin finalizarea a patru activităţi de bază şi anume: analiza formatelor documentelor tehnice pe suport electronic, analiza normelor de conversie a conţinutului cuprins în documentele tehnice de la formatul tradiţional la formatul electronic prin tehnici de digitizare, analiza conţinutului web- analiza formatelor documentelor tehnice regăsite pe web, elaborarea modelului conceptual al sistemelor de informare şi documentare cu conţinut tehnic.

Activităţile de cercetare efectuate au avut ca scop identificarea şi analiza tipurilor de formate regăsite pe suport electronic fie online, fie pe unităţi fixe sau mobile de memorie, care pot avea în componenţa lor diferite obiecte informaţionale. Deasememea s-a analizat conţinutul web în care informaţia se găseşte sub formă nestructurată, dar şi structura şi importanţa sistemelor de gestiune de baze de date în care regăsim informaţie structurată.

Partea aplicativă a activităţii de cercetare a constat în definirea unui model conceptual pentru managementul informaţiei tehnice în sisteme infodocumentare, implementat la Universitatea Transilvania din Braşov. Acesta va constitui un depozit digital care va cuprinde documente specifice activităţii de cercetare şi educaţionale din universitate.

Metoda de cercetare a avut la bază un studiu teoretic asupra tipurilor şi analizei tipurilor de formate a documentelor cel mai des utilizate în mediul informatizat. Deasemenea a avut loc şi o cercetare practică desfăşurată cu preponderenţă la partenerul Universitaea Transilvania din Braşov în vederea definirii şi inmplementării unuim depozit digital instituţional.

Pe perioada desfăşurării etapei 2 au avut loc întâlniri cu partenerii, workshop-uri, dezbateri ale aspectelor identificate spre rezolvare.

Raportul de cercetare pentru această etapă poate fi considerat ca fiind o sinteză din punct de vedere teoretic a principalelor activităţi ale etapei, cât şi iniţierea punerii în aplicaţie a concluziilor teoretice.

Analiza formatelor documentelor tehnice pe suport electronic. Munca de alegere a unui anumit format este îngreunată de multitudinea de opţiuni disponibile. S-au studiat mai multe posibilităţi, luând în considerare avantajele şi dezavantajele fiecăruia şi având mereu în vedere scopul utilizării fişierului. S-au identificat formate pentru principalele tipuri de obiecte informaţionale conţinute: text, imagini, audio, video, multimedia şi s-au evidenţiat principalele caracteristici pentru care se optează pentru un format sau altul.

Analiza normelor de conversie a conţinutului cuprins în documentele tehnice de la formatul tradiţional la formatul electronic prin tehnici de digitizare. În acest capitol s-a făcut o

scurtă prezentare a diferitelor modalităţi în care se poate realiza achiziţia imaginilor în formă digitală, plecând de la o carte. Deasemenea s-a definit o arhitectură a unui sistem de digitizare şi s-au dezvoltat tehnicile de conversie prin Recunoaşterea Optică a Caracterelor (Optical Character Recognition-OCR). Pe lângă un scurt istoric al conversiei documentelor din format tradiţional prin scanare şi OCR-izare, am prezentat tehnologiile curente şi produse software pentru conversia în format digital. S-a evidenţiat şi suportul lingvistic al produselor OCR cât şi alte particularităţi.

Analiza conţinutului web; analiza formatelor documentelor tehnice regăsite pe web. S-a analizat tipul de conţinut web şi creatorii de conţinut online. Cele mai frecvent utilizate formate de realizare a paginilor web au constituit baza de plecare în analiza tipurilor de documente care circulă în reţeaua Internet sau altor tipuri de reţele.

Elaborarea modelului conceptual al sistemelor de informare şi documentare cu conţinut tehnic. Cercetare din etapa 2 se încheie cu elaborarea unui model de management al informaţiei în structuri infodocumentare cuconţinut preponderent tehnic. S-a pornit dela definirea depozitelor digitale şi a organizaţiilor care dezvoltă crearea şi prezervarea pe termen lung a arhivelor digitale.

S-au evidenţiat şi preocupările actuale pentru implementarea depozitelor digitale în universităţile din România, activitate ce este la început de drum. Pentru realizarea depozitului digital la Universitatea Transilvania din Braşov s-a folosit platforma de cercetare interdisciplinară ASPECKT.

S-au evidenţiat tehnologii pentru alegerea platformelor şi a software-urilor pentru realizarea Depozitului pilot. Pagina web a depozitului digital ASPECKT-Dspace se poate accesa la adresa http://aspeckt.unitbv.ro/dspace.

Aparatul critic al raportului de cercetare este constituit din bibliografie selectivă şi webgrafie selectivă.

1. Reprezentarea digitală a informației

Teoria informației răspunde la două întrebări fundamentale din teoria comunicației: care este cea mai bună compresie a datelor (răspuns: entropia H) şi care este cea mai bună rată de comunicație pentru transmiterea datelor (răspuns: capacitatea canalului *C*). Din acest motiv unii consideră teoria informației ca fiind o submulțime a teoriei comunicației. Într-adevăr, ea are contribuții fundamentale în fizica statistică (termodinamică), ştiința calculatoarelor (complexitatea Kolmogorov sau complexitatea algoritmilor), inferențe statistice şi la probabilitate şi statistică.

La începutul anilor 1940 se credea că creşterea ratei de transmisie a informației peste un canal de comunicație creşte probabilitatea erorii. Shannon a surprins comunitatea ştiințifică dovedind că această afirmație nu este adevărată atunci când rata de comunicație este sub capacitatea canalului. Capacitatea canalului poate fi calculată simplu din caracteristicile de zgomot ale acestuia. Shannon argumenta că procese aleatoare ca muzica şi vocea au o complexitate ireductibilă sub care semnalul nu poate fi comprimat. Acest lucru l-a numit el Entropie, prin asociere cu utilizarea cuvântului în termodinamică, şi justifică faptul că dacă entropia sursei este sub capacitatea canalului comunicația asimptotic fără erori poate fi atinsă.

Astăzi teoria informației reprezintă punctul extrem al mulțimii tuturor schemelor de comunicație posibile. Limita compresiei de date este o extremă a setului de idei de comunicație. Toate metodele de compresie necesită descrierea datelor cel puțin la nivelul acestui minim. La cealaltă extremă este maximul transmisiei date cunoscut ca şi capacitatea canalului. Astfel toate metodele de modulație şi compresie a datelor se situează între aceste limite.

Teoria informației deasemenea sugerează mijloace de a obține aceste limite extreme ale comunicației. Totuşi aceste scheme teoretice optime de comunicație, se dovedesc a fi nefezabile computațional. Aceasta este mai curând datorită fezabilității computaționale a schemelor de modulare şi demodulare simple pe care le folosim şi nu datorită codificării aleatoare şi regulii de decodificare „nearest neighbor” propuse de demonstrația lui Shannon a teoremei capacității canalului. Progresele din domeniul circuitelor integrate şi proiectării codului ne-au permis obținem câteva din câştigurile sugerate de teoria lui Shannon. Un bun exemplu de aplicare a acestor idei a teoriei informației este utilizarea codurilor corectoare de erori pe compact discuri.

Cercetările moderne privind aspecte ale teoriei informației în comunicație sau concentrat pe teoria informației în rețele: teoria ratelor de comunicație simultană de la mai mulți emițători la mai mulți receptori într-o rețea. Unele câştiguri de rată de comunicație între emițător şi receptor sunt neaşteptate şi toate au o simplitate matematică certă. O teorie unificatoare rămâne totuşi a fi găsită.

Ştiința calculatoarelor (complexitatea Kolmogorov). Kolmogorov, Chaitin şi Solomonoff au avansat ideea că, complexitatea unui şir de date poate fi definită ca lungimea celui mai scurt program binar care generează acel şir de caractere. Deci complexitatea este dată de lungimea descrierii minime. Această definiție a complexității este universală, independentă de calculator şi de o importanță fundamentală. Complexitatea Kolmogorov pune bazele teoriei descriptive a complexității. Din fericire, complexitatea Kormogolov este aproximativ egală cu entropia Shannon H dacă secvența este aleasă aleator dintr-o distribuție care are entropia H. Aşadar legătura între teoria informației şi complexitatea Kolmogorov este perfectă. Într-adevăr considerăm complexitatea Kolmogorov a fi mai fundamentală decât entropia Shannon. Ea este limita compresiei datelor şi ne duce spre o procedură consistentă logic pentru inferență.

Există aici o plăcută relație de complementaritate între complexitatea algoritmică şi complexitatea computațională. Putem privi complexitatea computațională (complexitatea în timp) şi complexitatea Kolmogorov (lungimea programului sau complexitatea descriptivă) ca două axe corespunzând timpului de rulare a programului şi lungimii programului. Complexitatea Kolmogorov se concentrează pe minimizări de-a lungul celei de-a doua axe iar complexitatea computațională se focalizează pe minimizare de-a lungul primei axe. S-au făcut puține încercări pentru minimizarea simultană pe ambele axe.

Fizică (Termodinamică). Mecanica statistică este locul de naştere al entropiei şi a celei de-a doua legi a termodinamici. Entropia întotdeauna creşte. Pe lângă alte lucruri a doua lege ne permite să rejectăm orice pretenție de „perpetuum mobile".

Matematică (Teoria probabilității şi statistică). Mărimile fundamentale ale teoriei informației – entropie relativă şi informație mutuală – sunt definite ca funcționale ale distribuției probabilitice. În schimb, ele caracterizează comportamentul unor secvențe lungi de variabile aleatoare şi ne permit să estimăm probabilitatea unor evenimente rare şi să găsim cel mai bun exponent al erorii în testarea ipotezelor.

Filozofia ştiinței (Briciul lui Occam). William of Occam a spus „Causes shall not be multiplied beyond necessity" sau parafrazându-l „cea mai simplă explicație este cea mai bună". Solomonoff şi ulterior Chaitin, au argumentat că se obține o procedură de predicție universal bună dacă se ia o combinație ponderată a tuturor programelor care explică datele şi se observă ce tipăresc ele în continuare. Mai mult, această inferență va funcționa în multe probleme care nu sunt rezolvate prin statistică. De exemplu această procedură va prezice în final următorii digiți ai numărului π. Când această procedură este aplicată la aruncarea unei monede care generează cap cu probabilitate 0.7, acest lucru va fi de asemenea inferat. Când este aplicată la bursă procedura trebuie să găsească legile bursei şi să le extrapoleze optimal. În principiu, aşa o procedură ar fi găsit legea lui Newton din fizică. Bineînțeles, asemenea inferență este impracticabilă, deoarece a

elimina toate programele care nu reuşesc să genereze datele existente durează exagerat de mult. Am putea prezice ce se întâmplă mâine pentru sute de ani de acum înainte.

Economie (Investiții). Investiții repetate într-o piață staționară duce la creşteri exponențiale ale bunăstării. Rata de creştere a bunăstării este dual ă ratei entropiei pieței. Paralela între teoria investiției optimale în burse şi teoria informației este evidentă. Se poate dezvolta o teorie a investițiilor pentru a explora această dualitate.

Calcul versus Comunicație. Pe măsură ce construim calculatoare mai mari din componente mai mici atingem atât o limită de calcul cât şi o limită a comunicării. Calculul este limitat de comunicare şi comunicarea limitată de calcul. Acestea devin interdependente şi deci toate dezvoltările din teoria comunicației prin teoria informației ar trebui să aibă un impact direct asupra teoriei computaționale.

1.1.Reducerea datelor

Selecția unei submulțimi de trăsături caracteristice este definită ca procesul de selecție a acelei submulțimi de trăsături *d* dintr-un set de dimensiune mult mai mare *D* care maximizează performanțele de clasificare peste toate submulțimile posibile. Căutarea după o astfel de submulțime este o problemă foarte dificilă. Spațiul de căutare poate fi foarte mare.

1. 2. Entropia şi câştigul informațional

Câştigul informațional şi entropia sunt funcții ale distribuției probabilistice care susțin procesul de comunicare. Entropia este o măsură a incertitudini unei variabile aleatoare. Dându-se o colecție S de *n* eşantioane grupate în *c* concepte țintă (clase), entropia lui S relativă la clasificare este (1):

$$Entropy(S) \equiv -\sum_{i=1}^{c} p_i \log_2(p_i) \quad (0)$$

unde p_i este procentajul din *S* care aparține clasei *i*.

Pe baza entropiei este definită o măsură a gradului de eficiență în selecția trăsăturilor. Această măsură este numită *Câştigul informațional* şi reprezintă de fapt reducerea în entropie, cauzată de gruparea eşantioanelor în acord cu un atribut. Mai precis, câştigul informațional pentru un atribut relativ la o mulțime de eşantioane S este definit ca:

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (0)$$

unde *Values(A)* este mulțimea de valori posibile pentru atributul A şi S_v este submulțimea din S pentru care atributul A are valoarea v.

Utilizând ecuațiile, pentru fiecare trăsătură se calculează câştigul informațional obținut dacă mulțimea este împărțită utilizând acea trăsătură. Se obțin valori între 0 şi 1 fiind apropiate de 1 dacă trăsătura împarte mulțimea originală în două submulțimi cu dimensiuni apropiate. Pentru selectarea trăsăturilor relevante am utilizat diferite praguri. Dacă câştigul informațional obținut pentru o trăsătură depăşeşte pragul acea trăsătură va fi selectată, altfel nu va fi selectată.

Forman a arătat că câştigul informațional eşuează în cazul problemelor reale de clasificare când se lucrează cu multe trăsături.

2. Identificarea şi analiza tipurilor formatelor de documente pe suport electronic

Un format de fişier reprezintă o metodă de organizare a informației în secvențe de biți pentru ca aceasta să poată fi păstrată într-un calculator. Există diferite formate de fişiere pentru documente, imagini, sunete, videoclipuri, toate acestea fiind asociate cu anumite extensii. Câteva exemple de asociații: documente cu .doc, imagini cu .JPEG, .html pentru pagini.

Munca de alegere a unui anumit format este îngreunată de multitudinea de opțiuni disponibile. Înainte de a trece la fapte, trebuiesc studiate mai multe posibilități, luând în considerare avantajele şi dezavantajele fiecăruia şi având mereu în vedere scopul utilizării fişierului.

Când considerăm toate aceste aspecte putem avea în minte următoarea listă:

-Este formatul patentat sau controlat de organizații publice de standardizare?

-Au fost făcute publice specificațiile formatului?

-Este vechimea formatului un risc?

-Este posibil un upgrade al software-ului? Dacă da, poate versiunea nouă să deschidă fişiere mai vechi?

-Este posibil ca software-ul să fie retras de pe piață?

-Este posibil ca formatul să nu mai fie folosit şi, ca urmare, să dispară software-ul prin care fişierele puteau fi editate?

-Este permisă extragerea datelor dintr-un anume format pentru a putea fi folosite mai departe sau indexate, sau este permisă doar vizualizarea acestora?

-Formatul salvează datele la calitatea dorită? Vor fi datele degradate prin salvarea în acest format?

-Sunt compresate datele în acest format? (În general compresia face ca fişierele să fie mai puțin tolerante la erori în transmiterea de date şi pot fi degradate mai uşor când sunt păstrate pe diferite medii de stocare a datelor)

-Este formatul ales un standard acceptat?

2.1. Tipuri de formate

2.1.1. Formate text

DOC - fişier document

Formatul de fişier DOC este cel mai frecvent utilizat de Microsoft Word. Formatul a evoluat de-a lungul timpului astfel încât există defapt mai multe formate ce se regăsesc sub aceeaşi titulatură : doc.

DOC (abreviere de la *document*) este extensia numelui de fişier al documentelor produse de diverse procesoare de text în baza formatului impus de Microsoft Word versiunile 1.0-2003. Este un tip de fişier binar, care pe lângă textul propriu-zis stochează şi datele legate de formatarea acestuia, de asemenea şi elemente non-textuale precum imagini, grafică sau diagrame.

Este un format închis, standardul oficial pentru acest format fiind deţinut exclusiv de către compania Microsoft şi nefiind făcut public niciodată. Cu toate acestea, prin metode de inginerie inversă, s-a reuşit reconstituirea formatului de către alţi producători de software. Astăzi un număr mare de programe non-Microsoft (în special Open Source) pot să deschidă şi să editeze fişiere de tip .DOC fără probleme. Printre acestea se numără Open Office şi AbiWord.

Formatul Doc diferă de la o verisune la alta a Microsoft Office Word, astfel versiunile Word până la 97 au folosit un alt format DOC faţă de versiunile Word dintre 97 şi 2003. Versiunile mai vechi sunt incapabile să citească, să primească sau să printeze corect fişierele .doc generate de noile versiuni.

Principalele operaţiuni întreprinse în fişierele document:

- editarea şi formatarea textelor (folosind fonturi diferite, stiluri şi dimensiuni diferite de font, tipuri de subliniere a caracterelor, culoarea caracterelor etc.);
- inserare imagini, tabele, grafice, diagrame;
- aşezare în pagină sau formatare (dimensiunile paginii, orientarea ei, alinierea textului în pagină etc.);
- printare;
- căutare elemente specifice (cuvinte, fraze, etc.) în cadrul textului.

Dezavantajul formatului este acela că, fiind un format închis, există încă incompatibilităţi de manevrare a documentelor la transferul lor între diferite procesoare de text, ceea ce poate duce la pierderea de informaţii de formatare şi astfel, poate altera calitatea fişierului document.

Fişierele .doc sunt utilizate ca şi formate-sursă pentru unele fişiere eBook. Cu ajutorul unor programe convertoare se pot converti chiar fişierele .doc direct într-un anume format eBook. Totuşi ele nu sunt folosite foarte mult ca şi formate eBook deoarece folosind Word

pentru a citi documentul este foarte uşor de a-l modifica din greşeală, iar „scrolling”(derularea paginii) nu este cea mai bună metodă de a citi o carte. Unele versiuni ale Word-ului au, totuşi, un „review mode” ce poate asigura o citire mai uşoară a acestor documente.

Deşi este considerat învechit şi a fost înlocuit inclusiv de către creatorii săi, fişierul document DOC rămâne, pentru moment, cel mai utilizat format pentru procesoare de text[2)] cu atât mai mult cu cât acesta „oferă utilizatorului flexibilitatea necesară de a converti aproape orice tip de fişier într-un document MS-Word”.

Programele care deschid fişiere document în format DOC sunt

- în Mac OS: Microsoft Word, Apple Pages, Apple AppleWorks, OpenOffice.org Writer, Nuance OmniPage Pro X;
- în Windows: Microsoft Word, Microsoft Word Viewer, OpenOffice.org Writer, Nuance OmniPage Professional 17;
- în Linux: OpenOffice.org Writer.

Formatul de fişier DOC mai este folosit şi de către WordPad Document, însă acesta nu este utilizat pe o scară atât de largă ca şi Microsoft Word Document.

Începând cu Office 2007, Microsoft a înlocuit standardul DOC cu un nou standard, de data aceasta deschis (şi publicat), bazat pe XML şi deci folosind fişiere text în locul celor binare : .DOCX.

DOCX - fişier document în Office Word 2007

DOCX este o variantă îmbunătățită de format de fişier document introdus de MS Word 2007 (parte din MS Office 2007) ce „se bazează pe tehnologiile standard XML şi ZIP”.

Diferența dintre cele două formate (DOC şi DOCX) constă în felul cum a fost creat fiecare şi nu în conținut, astfel că, spre ex., un proces verbal, o scrisoare comercială, o invitație la o aniversare sau orice alt document ce poate fi editat în Word poate fi salvat în oricare format de fişier document.

Spre deosebire de formatul doc, DOCX este un format deschis, adică „specificațiile formatului sunt disponibile prin intermediul unei licențe care nu prevede plata unor drepturi de autor”. Prin urmare, formatul permite utilizatorilor să facă schimb de documente chiar dacă nu folosesc acelaşi procesor de text, fără riscul de a altera în vreun fel conținutul lor, ei putând deschide, edita şi salva fişierele .docx.

De asemenea, „.docx nu este un fişier binar, deci nu necesită un program compatibil pentru a-l deschide, şi este de dimensiuni mult mai mici decât echivalentul lui în format doc.

Un fişier .docx deşi pare a fi un singur fişier, este de fapt un pachet de fişiere arhivate împreună.” Pentru a-l deschide este nevoie de o aplicație software capabilă să dezarhiveze fişierele .docx (WinZip, PKWare Unzip, şi Stuffit Expander), „capabilă să înțeleagă şi modul de

lucru folosit de .docx pentru a gestiona diferitele componente ale fişierului : documentul, schema, proprietăţile, obiectele incluse în document etc."

Datorită dimensiunilor sale reduse şi a compatibilităţii cu alte procesoare de text non-Microsoft, acest tip de format de document se pretează foarte bine la trimiterea fişierelor prin poşta electronică, prin reţele sau prin Web, el putând fi deschis şi citit de aproape oricine îl primeşte.

Totuşi, chiar dacă formatul de fişier .docx este un format de fişier text ce poate fi citit cu uşurinţă prin dezarhivarea pachetului şi deschiderea folder-ului conţinând document.xml, deschiderea fişierului în acest fel alterează formatările de font şi de paragraf, precum şi alte obiecte integrate în fişier, astfel că se preferă folosirea unui convertor docx, care uşurează salvarea documentului dintr-un format în altul cu o configurare minimă, păstrând caracateristicile iniţiale complete ale documentului. Procesul de conversie de la un format la altul implică, în esenţă, extragerea de text şi orice obiecte înglobate, precum şi stilul de formatare, apoi re-salvarea acestora în alt format.

Cele mai folosite convertoare pentru fişiere .docx sunt Zamzar, Docx Converter sau Docx2Doc, aplicaţii web ce fac online conversia dintr-un format în altul, nefiind nevoie de instalarea vreunui soft.

Programele care deschid fişiere document în format DOCX sunt:

- în Mac OS: Microsoft Word 2008, Apple Pages, Penergy docXConverter;
- în Windows: Microsoft Word 2007, Microsoft Word with Compatibility Pack, Panergy docXConverter, OxygenOffice Professional, OpenOffice.org with Odf-Converter, NativeWinds Docx2Rtf;
- în Linux: OxygenOffice Professional, OpenOffice.org with Odf-Converter.

DOT- fişier document Word templates

DOT este un fişier document predefinit, folosit ca şablon pentru realizarea de mai multe documente cu formate similare, cum ar fi rapoarte de afaceri, plicuri de scrisori, cereri, formulare etc.[14)]

Corporaţia Microsoft este proprietarul fişierului cu extensia .dot folosit la editare de texte în Word. „Aceste fişiere cuprind caracteristici de formatare standard ca: setările paginilor (mărime, orientare, margini, număr de coloane ş.a.m.d.), stiluri de fonturi, culori de text, dar şi bare de instrumente (toolbars) particularizate sau macrocomenzi pentru a completa textele predefinite, ca de ex.: un nume de companie sau o adresă într-un document realizat cu ajutorul fişierului template (fişierul-şablon).

Şabloanele DOT sunt utile în special la realizarea de documente de acelaşi tip, cu aceleaşi formatări ori caracteristici de editare, sau la crearea personalizată a template-urilor pentru o varietate de documente ca: scrisori, faxuri, plicuri."

De exemplu, un template pentru o scrisoare de afaceri va avea un antet (letterhead) pe prima pagină, numerotarea paginilor şi eventual un câmp de generare automată a datei. În plus, pentru a personaliza stilurile pentru diferitele părți ale scrisorii (zona adresei, salutul, cuprinsul scrisorii, semnătura etc.) template-urile pot cuprinde câmpuri de introducere a textului indicându-se unde merg puse aceste elemente.

Macrocomenzile încorporate pot fi folosite pentru a completa informații în mod automat de fiecare dată când este creat un document nou folosind şablonul DOT salvat. Acestea grupează o succesiune de acțiuni ce se doresc executate direct, prin activarea unui buton, a unei combinații de taste sau eventual a unei opțiuni de meniu. Cea mai simplă modalitate de creare a unui macro este înregistrarea acțiunilor realizate, după care se stabileşte modul său de execuție, optând pentru una din asocierile anterior enumerate. Astfel, grupul de acțiuni înregistrat se va repeta ori de câte ori e nevoie, cu un efort minim.

Şablonul DOT poate fi realizat prin crearea unui document în Microsoft Word şi salvarea lui ca template ce poate fi deschis pe viitor folosind caseta de dialog a noului document. În plus, Microsoft Word este prevăzut cu un număr din cele mai folosite şabloane pentru scrisori, rapoarte, site-uri web şi publicații.

Şablonul normal DOT este cadrul uzual pe care se construieşte un document, el setează marginile şi fontul Times New Roman de dimensiune 10. Pentru creşterea eficienței în crearea documentelor, se pot folosi şi alte şabloane, predefinite sau definite de către utilizator.

Crearea unui şablon se face ca un document normal, acordând atenție deosebită atributelor de formatare şi introducând informațiile comune tuturor documentelor ce vor fi create pe baza acelui şablon. Fişierul creat se va salva cu tipul Template şi i se va atribui extensia .dot

Deşi fişierele .dot însele nu reprezintă un risc asupra securității informatice, unele pot cuprinde macrocomenzi încorporate, acestea fiind o potențială sursă de viruşi informatici, astfel încât ar trebui evitate fişierele .dot conținând macro-uri din surse necunoscute.

Programele care deschid fişiere document în format DOT sunt:

- în Mac OS: Microsoft Word şi NeoOffice Writer
- în Windows: Microsoft Word [14)]

Dezavantajul formatului este acela că fiind un fişier binar şi închis (ca şi .dot) are dimensiuni mari şi întâmpină greutăți la transferul documentelor între diverse aplicații software.

DOTX : Word 2007 Document template

DOTX este formatul noului editor de text Microsoft Word 2007 (parte din Microsoft Office 2007); el este un fişier template şi „o combinaţie între arhitectura XML şi arhivarea ZIP pentru reducerea dimensiunii."

Prin adoptarea formatului de fişier bazat pe XML noul format template asigură o transmitere uşoară a datelor între diferite aplicaţii, platforme şi Internet browser-i fără ca documentele astfel transmise să sufere modificări de orice fel.

Alte avantaje ale formatului Office XML:

- **Fişiere compacte.** Fişierele sunt comprimate în mod automat şi pot fi până la 75 la sută mai mici, în unele cazuri. Formatele Office XML utilizează tehnologia de comprimare ZIP pentru a stoca documente, oferind economii de cost prin reducerea spaţiului de disc necesar pentru stocarea fişierelor şi reducând lăţimea de bază necesară pentru a trimite fişierele prin poşta electronică, prin reţele şi Internet. Când se deschide un fişier, acesta este dezarhivat automat, când se salvează un fişier, acesta este automat arhivat la loc;
- **Recuperarea îmbunătăţită a fişierelor deteriorate.** Fişierele sunt structurate modular, menţinând diferitele componente de date din fişier separate între ele. Astfel, se permite deschiderea fişierelor, chiar dacă o componentă din fişier (de ex. o diagramă sau un tabel) este deteriorat;
- **Detectare mai uşoară a documentelor ce conţin macrocomenzi.** Fişierele salvate utilizând sufixul implicit „x" (.docx, .dotx, .pptx etc.) nu pot conţine macrocomenzi VBA (Visual Basic for Applications) sau controale ActiveX şi astfel nu implică riscurile de securitate asociate cu aceste tipuri de cod încorporat. Doar fişierele care au extensia numelui terminată în „m" (.docm, .dotm, .xlsm etc.) pot conţine macrocomenzi VBA şi controale ActiveX, care sunt stocate într-o secţiune distinctă din fişier. Astfel, extensiile de nume de fişier facilitează distingerea fişierelor care conţin macrocomenzi de cele care nu conţin, uşurând identificarea de către software-ul antivirus a fişierelor ce conţin cod cu potenţial dăunător;

Faţă de un fişier document obişnuit (.docx) la deschiderea unui fisier .dotx se dechide defapt o copie a template-ului, ceea ce protejează şablonul de modificări accidentale.

Fişierul cu extensia .dotx nu are macrocomenzi şi este folosit la depozitarea documentelor electronice, cum sunt: memo-uri, rapoarte, cărţi, foi de calcul, diagrame etc.

Acest format aduce îmbunătăţiri privind gestionarea informaţiei şi a fişierelor, recuperarea datelor şi interoperabilitatea sistemelor din reţea :

- Orice aplicație software bazată pe XML poate avea acces şi lucra cu fişierele document în acest nou format, chiar dacă nu face parte din sistemul Microsoft Office sau, mai mult, chiar dacă nu este un produs Microsoft.
- Utilizatorii pot, de asemenea, folosi transformările standard pentru a elimina sau adăuga date în aceste fişiere.
- În plus, problemele legate de securitatea informatică sunt drastic reduse deoarece informațiile sunt depozitate pe suport XML, care este în esență un simplu text. Astfel, datele pot trece de programele antivirus fără nici un impediment.

Programele care deschid fişiere .dotx sunt:

- în Mac OS: Microsoft Word 2007 sau versiunile precedente cu suport Open XML
- în Windows: Microsoft Word 2007 sau versiunile precedente cu suport Open XML); OpenOffice.org Writer; Adobe LiveCycle Designer

RTF- Rich Text Format

Creat de Microsoft în 1987, ca parte a Microsoft Word 3.0 pentru Macintosh, RTF reprezintă formatul de fişier folosit pentru transferul de documente; este un format universal, adică „poate fi citit şi scris de aproape toate procesoarele de text şi chiar de cele mai multe editoare de text". Acest tip de format este mult mai uşor de generat decât PDF sau PostScript, şi mai prietenos pentru procesoarele de text decât HTML. Astfel că multe programe de procesoare de text permit salvarea unui document în RTF selectând opțiunea „Save As"

RTF este o aplicație ce permite utilizatorului să editeze conținutul unui document. El este standardizat pentru crearea de fişiere de text formatat. Practic „este un format de fişier ce codifică graficele şi textele formatate pentru a permite transferul uşor între diferite aplicații şi sisteme de operare."

Față de fişierele de text simple, de bază, un fişier .rtf conține informații referitoare la stilul textului, mărime şi culoare. RTF foloseşte setul de caractere ANSI (American National Standards Institute), PC-8, Macintosh sau pentru calculatoare IBM pentru a controla descrierea şi formatarea documentelor. Formatul suportă diferite fonturi (italic, îngroşat, subliniat), precum şi imagini, grafice sau diagrame, iar orice document salvat în RTF îşi păstrează atributele de formatare intacte la transferul între diferite procesoare de text.

Spre deosebire de majoritatea formatelor de procesare de text, codurile RTF pot fi făcute citibile de către om (human-readable). Când se deschide un fişier RTF într-un editor de text, textul alfa-numeric este lizibil si elementele limbajului de formatare (markup language) nu sunt greu de descifrat, lucru ce a constituit o raritate la vremea cînd acesta a fost lansat. Însă,

fişierele RTF produse de majoritatea programelor, ca MS Word, conţin şiruri atât de mari de numere ale codurilor de control necesare pentru a asigura compatibilitatea cu programele mai vechi, încât cele mai multe din ele au un ordin de mărime mai mare decât textul în sine şi sunt foarte greu de citit. Din contră, formatele binare ca şi formatul DOC de la MS Word au numai câteva frânturi de text lizibil.

De asemenea, în comparaţie cu formatul DOC de la Microsof Word sau cu formatele Open Document (DOT), RTF nu suportă macrocomenzile. De aceea el este recomandat în locul acestor formate pentru că „este protejat şi are şanse minime de a fi infectat de viruşii informatici. Acest format nu răspândeşte, deci, viruşi când este transmis dintr-un sistem computaţional în altul prin poştă electronică. Salvând şi trimiţând date în format RTF se conferă o siguranţă asupra securităţii fişierului celui care primeşte documentul." Totuşi extensia RTF nu garantează întotdeauna că un fişier primit este sigur, deoarece Microsoft Word deschide fişierele standard DOC redenumite cu extensia RTF şi rulează, astfel, orice macrocomandă conţinută în el. De aceea, este necesară examinarea manuală a fişierului într-un simplu editor de text ca Notepad sau folosirea de „file command" în sistemele de tip Unix, pentru a determina dacă un fişier suspect este sau nu de tip RTF.

Fiind recunoscut de mai toate sistemele de operare indiferent de versiunile lor, RTF este un „format util pentru documentele text cu o formatare de bază, cum sunt: diverse manuale de utilizare, rezumate, scrisori şi alte documente de informare simple, ce necesită cel puţin o formatare de text (italic, bold sau subliniere), alinierea textului la stânga, dreapta sau central precum şi atributele fonturilor şi marginilor documentului."

Aplicaţia SIL International's Toolbox (http://www.sil.org/computing/toolbox), ce dezvoltă şi editează dicţionare, foloseşte RTF ca cea mai comună formă a sa de fişier document . Fişierele RTF produse de Toolbox sunt folosite atât în Microsoft Word cât şi în celelalte procesoare de text ce recunosc formatul RTF.

Datorită interoperabilităţii sale, a simplităţii şi a cerinţelor mici de procesare a unităţilor de prelucare a calculatoarelor (UPC), RTF este un format important pentru instrumentele de citire a cărţilor electronice (ebook reader) şi unele dispozitive, cum este BeBook, lucrează cel mai bine cu acest format.

Programele care deschid fişiere .rtf sunt:

- în Mac OS: Apple TextEditor, Apple Pages, Microsoft Word, Nuance OmniPage ProX sau orice editor de text ce suportă text formatat;
- în Windows: Microsoft WordPad, Microsoft Word, Corel WordPerfect Office X4, Nuance OmniPage Professional 17 sau orice editor de text ce suportă text formatat.

XLS: Microsoft Excel File Format

Microsoft Excel este cel mai utilizat program de calcul tabelar. Un document în format Excel este salvat cu extensia *.xls.*

Excel aparține grupei de programe Microsoft Office şi este disponibil pentru Microsoft Windows cât şi pentru Mac OS. Actuala versiune disponibilă este pentru *Microsoft Excel 2007* (din 30 noiembrie 2006 pentru firme, respectiv 30 ianuarie 2007 pentru utilizatori) ca şi pentru Mac OS *Microsoft Excel 2008* (din ianuarie 2008).

Excel-ul se poate folosi la organizarea datelor, efectuarea de calcule matematice, generarea de rapoarte bazate pe cifre sau pentru a crea diverse grafice.

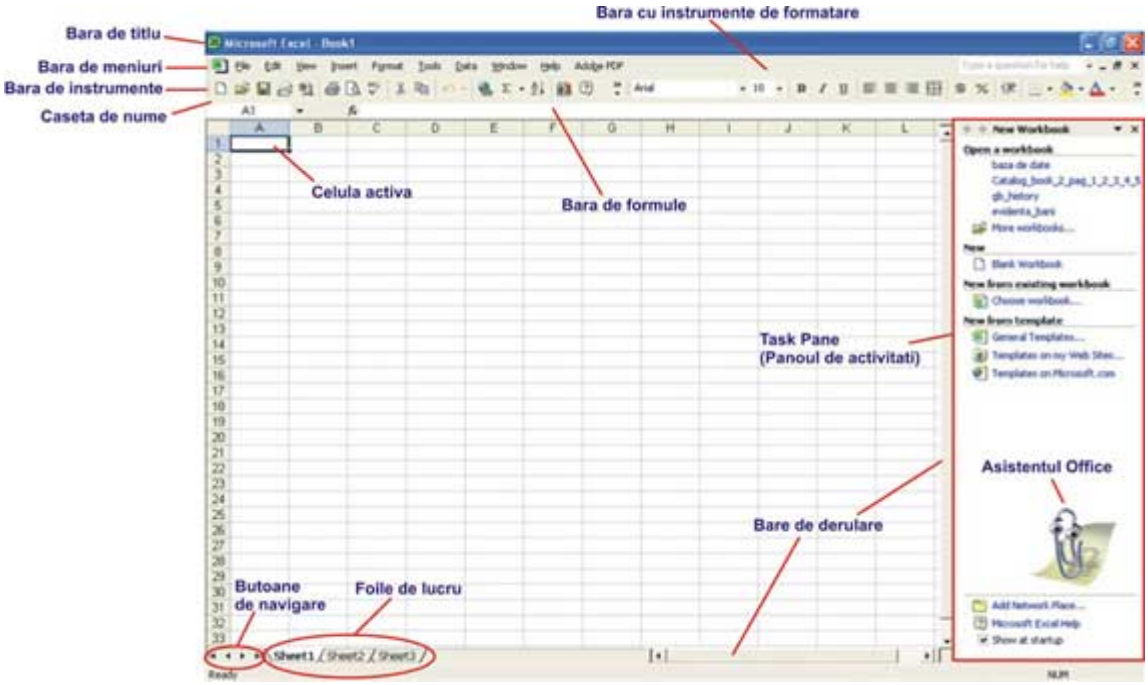


Fig. 2.1. Prezentarea aplicației Microsoft Excel

Tabel 1. Lista de programe care poate deschide fişiere .xls

Nume EXE	*Nume produs*
Excel.exe	Sistemul Microsoft Office 2007
Excel.exe	Microsoft Office 2000
Excel.exe	Microsoft Excel
Excel.exe	Microsoft Office XP
Excel.exe	Microsoft Office 2003
Moc.exe	Microsoft Open XML Converter
pdfcreator.exe	PDFCreator
scalc.exe	
xlview.exe	Sistemul Microsoft Office 2007

În Microsoft Office 2007, Microsoft prezintă formate noi de fişier pentru Word, Excel şi PowerPoint cunoscute ca formate Office Open XML. Aceste noi formate de fişier facilitează integrarea cu surse de date externe şi oferă dimensiuni reduse pentru fişiere şi operațiuni

îmbunătățite de recuperare de date. În Office Excel 2007, formatul implicit pentru un registru de lucru Excel este formatul de fişier XML Office Excel 2007 *(.xlsx)*. Alte formate XML disponibile sunt formatul XML şi cu macrocomenzi activate Office Excel 2007 *(.xlsm)*, formatul de fişier pentru un şablon Excel Office Excel 2007 *(.xltx)* şi formatul de fişier pentru un şablon Excel cu macrocomenzi activate Office Excel 2007 *(.xltm)*.

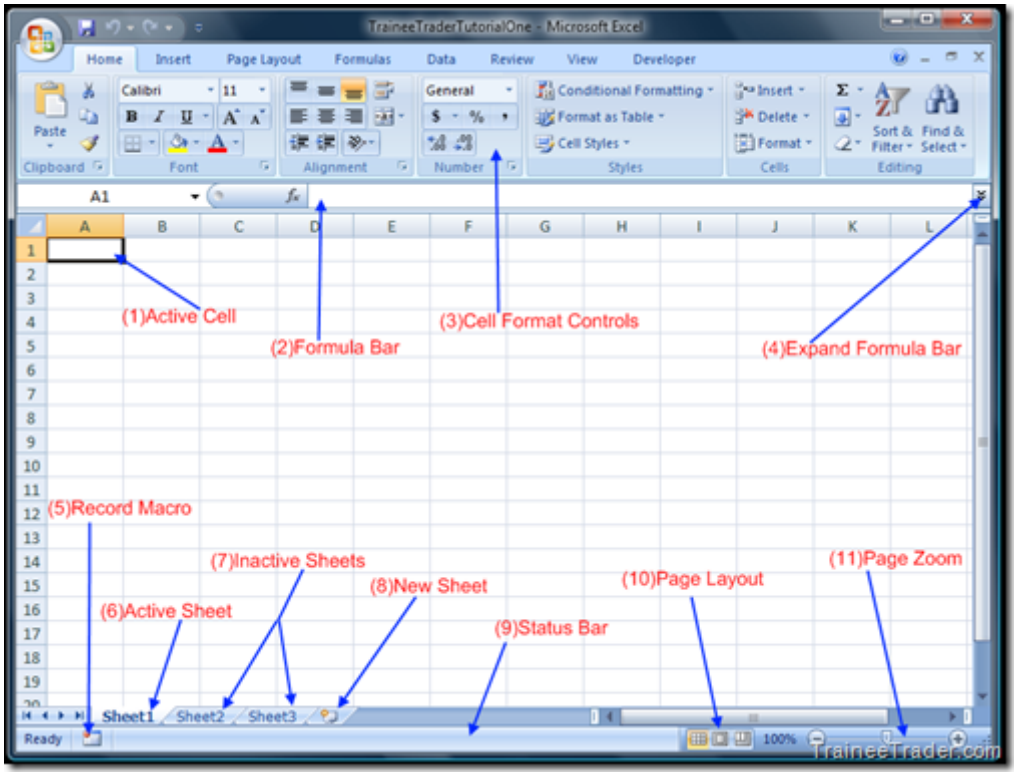


Fig. 2.2. Formatul de fişier pentru un şablon Excel cu macrocomenzi activate Office Excel 2007

Tabel 2. Formate Excel

Format	Extensie	Descriere
Registru de lucru Excel	.xls	Format de fişier binar
Registru de lucru Excel	.xlsx	Formatul de fişier Office Excel. Nu poate stoca cod de macrocomenzi Microsoft Visual Basic pentru aplicaţii VBA sau foi de macrocomenzi Microsoft Office Excel
Registru de lucru Excel (cod)	.xlsm	Formatul de fişier Office Excel cu macrocomenzi activate. Stochează cod VBA sau foi de macrocomenzi Excel
Registru de lucru binar Excel	.xlsb	Formatul de fişier binar Office Excel (BIFF).
Şablon	.xltx	Formatul de fişier Office Excel pentru un şablon. Nu poate stoca cod de macrocomenzi VBA sau foi de macrocomenzi Office Excel
Şablon (cod)	.xltm	Formatul de fişier Office Excel pentru un

		şablon, cu macrocomenzi activate. Stochează cod de macrocomenzi VBA sau foi de macrocomenzi Excel
Şablon Excel	.xlt	Formatul de fişier binar Excel pentru un şablon Excel.
Foaie de calcul	.xml	Format de fişier foaie de calcul
Program de completare Excel	.xlam	Programul de completare Office Excel cu macrocomenzi activate, un program suplimentar, proiectat pentru a executa un cod suplimentar. Acceptă utilizarea VBA şi a foilor de macrocomenzi Excel
Program de completare Excel	.xla	Programul de completare Excel, un program suplimentar proiectat pentru a executa un cod suplimentar; acceptă utilizarea VBA.
Registru de lucru Excel	.xlw	Un format de fişier Excel care salvează numai foi de lucru, foi diagramă şi foi de macrocomenzi; există posibilitatea deschiderii unui registru de lucru în acest format de fişier în Office Excel, dar nu şi salvarea unui fişier Excel în acest format.

XLSB: Excel binary workbook

XLSB: Registru de lucru binar non-XML

Descrierea extensiei de fişier *.xlsb*:

Este un registru de lucru cu stocare de informaţii în formă binară şi este folosit pentru deschiderea şi salvarea documentelor mai rapid şi mai eficient, fiind destinat în special pentru documente foarte mari, o cantitate mare de date, cu zeci de mii de rânduri, şi/sau de mai multe sute de coloane.

Fişierele cu extensie *.xlsb* pot include tabele de numere, text sau ambele numere şi text, formule, conexiuni de date, diagrame şi imagini. Conţinutul registrului de lucru este de obicei organizat într-o grilă şi include date numerice, date structurate şi formule.

Cel mai uşor mod de a deschide un fişier *.xlsb* este de a da dublu click pe el, iar computerul decide singur. Dacă fişierul nu se deschide înseamnă că nu este instalată aplicaţia care poate vizualiza şi edita fişiere *.xlsb*.

De asemenea se poate utiliza Notepad sau alt editor de text pentru a deschide fişierul *.xlsb*.

Atunci când un registru de lucru este salvat în formatul *.xlsb*, datele din foaia de calcul sunt memorate în Biff (Binary Interchange Format de fişier).Fiecare parte Biff este salvată ca o

serie de înregistrări Biff şi chiar dacă diferitele tipuri de înregistrări Biff conţin informaţii diferite, fiecare înregistrare are acelaşi format de bază. Fiecare înregistrare începe cu un număr de înregistrare urmat de o dimensiune de înregistrare. Aceste două elemente sunt apoi urmate de câmpurile înregistrării.Pentru a converti un fişier *.xlsb* în alt tip de fişier se deschide fişierul *.xlsb* implicit în programul său şi se salvează fişierul deschis ca alt tip de fişier sau se foloseşte un fişier de conversie.

XLSM: Microsoft office excel 2007 document

.XLSM: Registru de lucru care permite macrocomenzi

Descriere fişier

Fişierele cu extensie *.xlsm* sunt foi de calcul create cu Microsoft Excel care conţin macrocomenzi activate VBA (Visual Basic for Applications – limbaj de programare derivat din Visual Basic Standard). Doar fişierele care au extensia numelui de fişier terminată în *m* (cum ar fi *.docm* şi *.xlsm*) pot conţine macrocomenzi VBA şi controale ActiveX, care sunt stocate într-o secţiune distinctă din fişier. Extensiile de nume de fişier distincte facilitează distingerea fişierelor care conţin macrocomenzi de cele care nu conţin, facilitând identificarea de către software-ul antivirus a fişierelor care conţin cod cu potenţial dăunător, prin urmare fişierele cu extensia *.xlsm* nu pot fi considerate sigure în cazul în care nu provin de la surse de încredere (pot fi virusate).
(O macrocomandă poate fi un mic program sau script care automatizează sarcinile comune. Aceste script-uri rulează în cadrul programelor şi sunt create de utilizator.)

XLT: Excel templat

Dacă avem de creat mai multe documente cu o structură asemănătoare şi cu informaţii comune, în loc să le introducem sau copiem de mai multe ori, vom crea un document şablon, pe care-l vom putea folosi la crearea celorlalte documente.

Şablonul se creează ca un document normal dar se salvează cu tipul „Template”; fişierul corespunzător va avea extensia *.xlt*. Pentru a putea fi deschise uşor cu File - New - General, se recomandă salvarea şabloanelor în directorul implicit "Microsoft Office\Templates".
În vederea obţinerii documentului creat pe baza unui şablon, se deschide şablonul, se efectuează modificările necesare şi se salvează conţinutul (ca fişier *.xls*).

XLTM: Excel 2007 template

Un fişier cu extensia *.xltm* este un fişier şablon activat pentru macrocomenzi (macrocomenzile pot să conţină viruşi), care poate stoca cod de macrocomenzi VBA (Visual Basic for Applications) sau foi de macrocomenzi Excel 4.0 *(.xlm*).
Este un format nou şi se aplică la Word, Excel şi Power Point.

ASCII (TXT)

Un fişier **text** este un tip de fişier în care datele sunt stocate ca o secvență de caractere, într-o codificare predefinită (de obicei ASCII, dar mai recent şi Unicode). Este unul din cele două tipuri canonice de fişiere, celălalt tip fiind cel de fişier **binar**, în care datele sunt stocate ca o secvență de biți. Diferența dintre cele două tipuri de fişiere este semi - arbitrară, orice fişier text este, până la urmă, un fişier binar, în schimb, nu orice fişier binar este un fişier text. Prin definiție, un fişier text este codificat unitar, conținutul său fiind lizibil şi editabil direct, prin intermediul unui editor simplu de text (Notepad, vi, emacs, gedit etc.). Fişierul text este considerat independent de platformă.Într-un fişier text, tipul de date este considerat explicit ca fiind textul neformatat. Organizarea unui astfel de fişier este pe rânduri (care se generează la apăsarea tastei Enter / Return), fiecare rând de text fiind delimitat de unul sau mai multe caractere de sfârşit de rând (caractere EOL). Acestea diferă în funcție de sistemul de operare folosit pentru crearea şi editarea fişierului. Astfel, Windows foloseşte două caractere de control (ASCII 13 urmat de ASCII 10, sau, mai scurt CR+LF) pentru a semnaliza sfârşitul unei linii de text, pe când sistemele tip UNIX (incluzând aici şi Linux şi Mac OS X) folosesc numai caracterul LF, iar sistemele Mac OS pre-Unix (versiunile 9 sau mai vechi), folosesc doar caracterul CR. Astăzi, majoritatea sistemelor şi editoarelor de text sunt perfect capabile să se folosească şi să convertească automat fişierele text între diverse arhitecturi. O excepție notabilă în acest sens este editorul Notepad care în cazul în care este pus să deschidă un fişier text generat pe o altă platformă, va afişa caracterul de control LF sau CR după fiecare rând, neîmpărțind textul în rânduri.

Cele mai întâlnite tipuri de codificări sunt:

-ANSI: folosit în mod implicit în Notepad, dar este limitat la 256 de caractere; se recomandă pentru caracterele limbii engleze, deoarece o astfel de codificare ocupă mai puțin spațiu şi necesită un timp redus de procesare;

-Unicode: standard folosit pentru a reprezenta simboluri text din toate limbile, putând recunoaşte mai mult de un milion de caractere; include diferite metode;

-UTF-8: numărul de octeți pe care este ținut minte un caracter variază de la 1, pentru literele latine, şi până la 6 pentru limbile cu accente şi diferite simboluri; caracterele nelatine ce erau ținute minte şi prin codificarea de tip ANSI sunt procesate mai eficient în acest caz.

Pentru utilizator, un astfel de fişier, vizualizat ca atare, apare ca text neformatat şi nearanjat în pagină. Fişierele text nu pot stoca decât caractere printabile (litere şi cifre) şi un număr foarte limitat de caractere de control.

Astfel, în structura fişierului nu se pot stoca elemente de formatare precum culori sau imagini. Fişierele text pot fi folosite pentru activități simple, precum luatul de notițe. Cu toate acestea, un fişier text poate stoca date de intrare pentru un program terț. Astfel de utilizări, datorită portabilității uriaşe a formatului (neținând seama de problema sfârşitului de linie) şi datorită faptului că conținutul fişierului este uşor lizibil şi editabil ca text a dus la apariția multor standarde care codifică informații de diverse feluri ca text lizibil stocat în fişiere text şi care procesate cu ajutorul unor alte programe iau altă formă. Exemple în acest sens sunt standardul XML (şi aplicațiile sale - HTML - care stochează pagini formatate într-un fişier text prin intermediul unor descriptori textuali sau SVG - care stochează imagini ca text lizibil în fişiere text). Mai mult, codul sursă al oricărui program, în orice limbaj de programare, este salvat pe disc sub formă de fişier text.

Tradițional, fişierele text stochează datele în format ASCII, pe 7 biți. Multe protocoale asociate rețelelor informatice permit transmiterea corectă a informațiilor în acest format, dar nu permit transmisia corectă de fişiere binare (în care codificarea are loc pe 8 biți), ducând, în cazul transmisiei, la coruperea fişierului rezultat (din cauza conversiei ce are loc între cele două sisteme la apariția caracterelor de sfârşit de rând). Astfel, există diverşi algoritmi şi soluții software pentru codificarea fişierelor binare în format compatibil ASCII pentru transmitere (cel mai comun este Base64). Cu toate acestea, sistemele informatice stochează toate datele într-un mod identic - ca şiruri de biți. Mai mult, unele sisteme de operare nu oferă rutine şi proceduri specifice fişierelor text, tratând toate fişierele ca şiruri binare.

Deşi acest tip de format nu se mai foloseşte des, el are avantajele lui:

-dimensiuni reduse datorate lipsei formatării;

-risc redus de a împrăştia viruşi, conținând doar caractere;

-compatibilitate pe toate platformele;

-nu necesită cumpărarea unui software, deoarece majoritatea editoarelor cu care se deschid sunt încorporate în sistemul de operare.

2.1.2. Formate imagine

Reprezentarea unei imagini pe calculator a constituit un pas firesc, odată cu creşterea puterii de calcul a PC-urilor. O imagine este constituită din pixeli (puncte), denumirea de pixel provenind de la picture elements, el fiind elementul fundamental ce sta la baza unei imagini. Fiecare pixel capată o anumita culoare dintr-un spectru variabil şi este memorat ca un numar, ce va fi înregistrat intern într-o serie de biți. Cu cât numarul de biți alocați este mai mare, cu atât creşte posibilitatea de a afişa nuanțe coloristice mai apropiat de realitate.

Pentru o imagine monocrom, fiecare pixel va deține culoarea alb sau negru, deci el va fi memorat ca 0 sau 1. Pentru o imagine în 4 culori, pixelii vor putea lua valori de la 0 la 3 (în format binar 00, 01, 10 şi 11 - deci 2 biți), iar pentru 256 de culori (2 la puterea a opta) vor fi necesar câte 8 biți pentru fiecare pixel. Imaginea astfel creată va fi totuşi săracă, un numar de culori rezonabil fiind 65536 (16 biti), iar pentru perfecțiune fiind necesare 16 milioane de culori (24 de biti).

Cunoscând faptul ca un byte este format din 8 biți, pentru o imagine pe 24 de biți, un pixel este memorat pe 24/8=3 bytes. Astfel, pentru o imagine la rezoluția de 1024x768 vor fi necesari 1024x768x3=2.359.296 bytes pe disc. Acesta este spatiul ocupat de o imagine necomprimata, ce poartă de obicei extensia **BMP**, fiind formatul standard, cel mai simplu, dar şi cel mai nepractic. În timp, s-a constatat că majoritatea imaginilor nu necesită o fidelitate perfectă şi au fost concepuți numeroşi algoritmi de compresie, mulți fiind asemănători cu cei folosiți la compresia datelor, dar care duc în multe cazuri la pierderi de calitate. În acest mod, o imagine de 2 MB poate ocupa doar 100 KB în format comprimat, pierderea de calitate fiind în general insesizabila unui ochi amator. În timp, s-a constatat ca majoritatea imaginilor nu necesita o fidelitate perfecta şi au fost concepuți numeroşi algoritmi de compresie, mulți fiind asemănători cu cei folosiți la compresia datelor, dar care duc în multe cazuri la pierderi de calitate. În acest mod, o imagine de 2 MB poate ocupa doar 100 KB în format comprimat, pierderea de calitate fiind în general nesemnificativă unui ochi amator.

Este important de cunoscut ce format de fotografie trebuie folosit în anumite situații, deoarece unele formate de imagine sunt bune pentru a obține o balanță optimă dintre calitate si mărimea imaginii, în timp ce alte formate sunt perfecte pentru a recupera sau a modifica mult mai uşor caracteristicile unei fotografii (contrast, luminozitate, culoare). Cele mai relevante trei formate în fotografia digitala sunt: Raw, Tiff si JPEG. In continuare vor fi prezentate caracteristicile fiecarui format.

TIFF- Tagged Image File Format

TIFF este un acronim al Tag(ged) Image File Format şi totodată unul din cele mai populare şi flexibile tip de fişier în format raster. Deşi este foarte cunoscut datorită flexibilității şi

puterii de compresie, el este considerat în egală măsură destul de complicat şi misterios în acelaşi timp. Aceasta deoarece formatul TIFF este extensibil şi cu multe caracteristici de care un programator ar putea avea nevoie atunci cînd lucrează cu imagini grafice. Din această cauză el este considerat şi ca cel mai confuz format grafic. Tagged Image File Format are mai multe formate interne. Multe programe suportă formatele de bază, însă nu suportă toate formatele existente.

Fişierul TIFF este organizat pe trei mari secţiuni:

1. *Antetul* - În ciuda complexităţii fişierului TIFF, acesta are un antet foarte simplu. El conţine 3 câmpuri şi are lungimea de 8 octeţi.
2. *Directorul Imaginii* - Este o colecţie de informaţii folosite pentru a descrie imaginea bitmap care urmează. Poate părea puţin cam confuz deoarece Directorul Imaginii poate apărea oriunde în fişier, dar acest lucru nu trebuie să neliniştească deoarece el este pus într-o listă simplu înlănţuită, împreună cu ceilalti Directori, listă spre care pointează Primul ID din Antet. Spre deosebire de structurile întâlnite până acum, structura Directorului Imaginii nu este fixă, ea depinzând de tipul imaginii căruia îi serveşte ca şi antet. În această structură se pot adăuga sau scoate noi câmpuri, exact ca si paginile dintr-un dosar.
3. *Zona de Date* în care sunt stocate imaginile bitmap - Fisierele TIFF contin numai imagini de tip bitmap, deşi cu ajutorul câmpurilor nu este exclusă nici stocarea unor imagini vectoriale, sau a unor texte (vezi si formatul GIF).

Din aceste trei sectiuni este nevoie doar de primele două, existând posibilitatea ca Zona de Date să lipsească. TIFF are reputatia de a fi un format complicat datorită faptului că în interiorul fişierului numai antetul poate fi găsit la o locatie fixă, restul sectiunilor variind de la fişier la fişier. Fiecare Director al Imaginii împreună cu bitmap-ul asociat formează un subfişier. Nu există o limită a numărului de subfişiere pe care le poate conţine un fişier TIFF.

Formatul TIFF suportă cei mai mulţi algoritmi de compresie. Versiunea 4.0, are incluşi numai algoritmii Run Length Encoding, versiunea 5.0 adaugă la acestea şi algoritmul LZW, întâlnit si la formatul GIF, iar versiunea 6.0 adoptă si algoritmul JPEG, algoritm care se descurcă de minune cu împărţirea fisierului în cărămizi (o bucată dreptunghiulară din imaginea bitmap, care a fost introdusă datorită imaginilor TIFF foarte mari).

Totodată TIFF este un format comun utilizat pentru diverse aplicaţii pentru imagini, inclusiv cele pentru scanare şi fax. Microsoft Office Document Imaging utilizează formatul

TIFF, folosind posibilitatea acestuia de a conține text acceptat de recunoaşterea optică a caracterelor (OCR) (OCR: traduce imaginile de text, cum sunt documentele scanate, în caractere de text efective. Este cunoscut şi ca recunoaştere de text.). Când se scanează documente noi, acestea sunt salvate în format TIFF (cu extensia tif) şi orice text OCR este stocat în fişierul TIFF împreună cu imaginea.

Se pot deschide şi edita fişiere TIFF create cu Office Document Imaging în multe alte aplicații grafice. În acest caz, orice text OCR conținut în fişier, va fi pierdut. Dacă se doreşte să se acceseze din nou textul din fişierul TIFF, în Office Document Imaging, va trebui să se reexecute OCR.

Detalii tehnice:

Office Document Imaging creează fişiere în următoarele formate:

Monocrom Un bit per pixel, compresie G4

Tonuri de gri 8 biți per pixel, compresie JPEG

Color 24 biți RGB, compresie JPEG

Office Document Imaging acceptă:

- Toate tipurile de compresii listate în specificația TIFF 6.0.
- Diferite tipuri de compresii pentru fiecare pagină a unui document multipagină.
- Imagini TIFF cu adâncimea culorii de 1-bit, 4-biți, 8-biți sau 24-biți (atât paletă cât şi non-paletă).
- Spațiere culoare RGB şi CMYK.
- Imagini rearanjate şi redimensionate.

Office Document Imaging nu acceptă:

- Spațiere culoare YCbCr, cu excepția imaginilor JPEG.
- Spațiere culoare CIE Lab.
- Imagini cu mai mult de cinci eşantioane per pixel sau cu o dimensiune a eşantionului mai mare de 32 biți.
- Imagini în format Planar.

TIFF este un standard care se foloseşte în mod special în industria tipografică. Fisierele TIFF sunt semnificativ mai mari decat cele JPEG si pot fi şi ele necompresate sau compresate folosind compresia lossless. Spre deosebire de JPEG, fisierele TIFF pot fi de 16-biți/channel sau 8-biți/channel si pot fi folosite layere pentru a stoca mai multe imagini intr-un singur fisier TIFF.

Fisierele TIFF sunt o opțiune excelentă de backup şi de a fi folosite ca fişiere intermediare, care pot fi editate mai târziu, din moment ce nu sunt folosite artifacte de compresie. Multe camere de fotografiat au optiunea de a folosi fisiere TIFF, dar acestea pot consuma un spatiu destul de mare si destul de important de pe cardul de memorie a aparatului. În cazul in care camera suportă formatul RAW este indicat folosirea acestuia deoarece este o alternativă superioara fată de TIFF.

Avantajele formatului:

- Format universal, recunoscut de majoritatea programelor.
- Calitate foarte bună a fotografiilor.
- Compresie (opțională) prin protocol LZW.
- Format ideal pentru stocarea imaginilor în curs de prelucrare sau prelucrate.
- Codificare până la 64 biti.
- Permite gestionarea transparenței (strat alfa).
- Conversie în orice alt format fără pierdere de calitate.

Dezavantajele formatului:

- Fişierele de imagine au dimensiuni foarte mari.
- Nu este adecvat pentru utilizarea în camerele foto digitale.
- Nepotrivit pentru utilizare pe Internet.

GIF- Graphics Interchange Format

Formatul Gif a fost creat pentru elemente grafice de dimensiuni mici. Modul de lucru se bazează pe reducerea paletei de culori a graficii la maximum 256 de culori. Adesea, ceea ce se percepe ca fiind o culoare sunt de fapt mai multe: de exemplu, galbenul este compus din mai multe nuanțe sau, în cazul culorilor digitale, din pixeli de culoari individuale.

Tehnologia GIF reduce numărul de culori folosind *intercalarea* pentru cele care nu se găsesc în paleta GIF standard de 256 de culori.

GIF sau *Graphics Interchange Format* (Format de schimb de grafică), este formatul de fişier folosit de cele mai multe persoane pentru a face schimb de grafică. Devenit popular la origine pe CompuServe, GIF s-a răspândit către alte servicii on-line şi apoi către Internet şi către Web. Orice browser ce suportă grafică, suportă GIF.

Pentru imaginile cu o mulțime de blocuri de culoare uniformă, dimensiunile fişierelor GIF tind să fie mici. Astfel, se preferă GIF-urile pentru bannere (steaguri) sau pentru imaginile cu zone întinse de culoare uniformă, cum ar fi graficele cu bare sau pictogramele. Deci desenele simple pe care le creează cei mai mulți dintre noi funcționează cel mai bine cu GIF. Grafica artistică densă şi fotografiile funcționează mai bine cu JPEG.

Fişierele GIF dau, de asemenea, anumite opțiuni de afişare a paginii Web care nu se obțin întotdeauna cu fişierele JPEG. Se pot face culorile din imaginile GIF transparente față de orice se află în fundalul imaginii şi se pot salva imaginile GIF în format *intercaled (întrețesut).* Imaginile salvate în acest mod şi apoi descărcate (download-ate) de către un browser, apar mai întâi într-o rezoluție foarte scăzută şi apoi într-o rezoluție progresiv mai clară, până apare întreaga imagine. Această facilitate face imaginile GIF preferabile pentru afişarea rapidă a unei grafici brute care se îmbunătățeşte pe măsura trecerii timpului şi pentru crearea efectelor speciale extravagante.

Este desigur unul dintre cele mai cunoscute formate pentru fişiere de imagini: GIF (*Graphic Interchange Format*) a fost dezvoltat de CompuServe în 1987 şi este foarte des utilizată pe WWW grație portabilității sale. Fortmatul utilizează o paletă pâna în 256 de culori din cei 24 biți de culoare din spațiul RGB. Extensia acestor fişiere este .GIF.

Suportă animații şi alocă câte o paletă de 256 de culori pentru fiecare frame. Poate fi folosit pentru reproducerea de filmulețe cu rezoluții mici. Limitările de culoare fac ca formatul Gif să fie nepotrivit pentru reproducerea pozelor sau altor imagini cu culori continui, dar este foarte potrivit pentru imagini simple ca exemplu grafice sau logo-uri cu suprafețe de culori unanime.

Imaginile gif utilizează un algoritm de compresie LZW - denumit dupa cei care l-au creat, Lempel, Zif şi Welch în 1985- această tehnică permite reducerea dimensiunii imaginii fără degradarea vizuală a imagnii. La imagini color este posibilă o comprimare de 3:1, uneori se poate obține chiar şi o rată de comprimare de 5:1. Codificarea LZW divizează informațiile de imagine. Fiecărei succesiuni îi este atribuit un index, care este salvat într-un tabel. În consecință, informația propriu-zisă nu este formată din puncte individuale de imagine, ci din indexurile din tabel.

La prima vedere, comprimarea LZW lucrează fără pierderi în cazul fişierelor GIF, deoarece fişierul GIF decomprimat este identic cu cel dinainte de comprimare. Problema este însă urmatoarea: fişierele GIF salveaza numai 8 biți pe punct de imagine, fapt ce micşorează

adâncimea de culoare la 256 de culori. Dar, fiind vorba despre plăci grafice care pot reprezentă pâna la 16 milioane de culori, această adâncime este prea mică.

Cu ceva timp în urmă, GIF a fost cel mai important format de fişier în Usenet pentru transferul de imagini color prin intermediul liniei telefonice. Însă, pentru ca Unisys a protejat compresia LZW prin patent, au existat tot mereu probleme de licență la aplicațiile comerciale. Acest fapt şi metodele moderne de comprimare care au aparut între timp au dus la înlocuirea lui GIF cu JPEG.

Mai vechi decât JPEG, pentru unele tipuri de imagini, GIF este superior în privința calității imaginii, a ratei de comprimare sau a ambelor, dar are un defect: o paletă redusă de culori, deoarece, prin convenție, GIF are maximum 256 culori (codificare pe 8 biti). De exemplu, arii largi cu exact aceeaşi culoare, sunt comprimate foarte eficient de algoritmul GIF. Dar, pentru o imagine fotografica, transpusă digital prin scanare sau înregistrată cu aparate fotografice digitale, adâncimea de culoare este de cel putin 24 biți (16 milioane de culori), ceea ce determină o alterare severă a pozei.

Imaginile optime pentru comprimare prin metoda GIF sunt cele cu linii şi cu suprafețe uniform colorate, cum sunt cele desenate cu programe de grafică (de exemplu Microsoft Paint). Cu cât imaginea conține mai multe nuanțe, cu atât JPEG se descurcă mai bine. GIF comprimă bine marginile netede, cum ar fi chenarele fotografiilor sau conturul literelor aplicate peste o imagine, pe care JPEG le redă mai difuz, datorită rotunjirilor inerente în calculele matematice pe care le efectuează algoritmul de compresie.

În plus, formatul GIF permite animații simple. Formatul GIF este foarte folosit pentru realizarea site-urilor Internet (butoane, bare de defilare, logo-uri, etc).

Avantaje

- Format universal, fară copyright, foarte folosit pentru Internet, recunoscut de orice nagivator, fără să aibă nevoie de plugin-uri;
- Comprimă imaginea fără pierderi de calitate, până la dimensiuni foarte mici;
- Posibilitate de animare;
- Poate gestiona transparența.

Dezavantaje

-Codificare doar pe 8 biți;

- Reproducere de calitate redusă a imaginilor fotografice.

JPEG- Joint Photographic Experts Group

Formatul de fişier JPEG (*Joint Photographic Experts Group*) a fost dezvoltat de C-Cube Microsystems în 1992 şi aprobat în 1994 ca standartul ISO 10918-1. În ciuda adâncimii mari de culoare (16 milioane de culori) imaginile scanate cu JPEG pot fi stocate pe harddisk, ocupând puțin spațiu. Acest lucru este posibil datorită unei comprimări a datelor bine pusă la punct, care micşorează dimensiunile fişierului la minim. Fişierele JPEG pot fi recunoscute dupa extensia: .jpeg, .jpg, .jpe, .jfif, .jfi, .jif (containers).

Formatul JPEG foloseşte pentru comprimare transformarea cosinus discretă (DCT). În spatele acestui sistem se ascunde o transformată Fourier, care schimbă dispersia pixelilor într-o dispersie de frecvență de amplitudine. Suprafețelor mari şi uniforme de imagine li se atribuie participații de frecvență mici, în schimb detaliile fine primesc participații de frecvență mai mari. Comprimarea apare ca urmare a faptului că părți de imagine cu participații de frecvență mai mari capată o pondere mai mică, iar amplitudinile lor vor fi egale cu zero.

La comprimare se pierd unele date de imagine, însă algoritmul de compresie utilizat verifică în permanență că la decomprimare imaginea să nu sufere o pierdere a calității. La comprimare, fiecare imagine este divizată în blocuri de 8x8 puncte de imagine. Pentru început, fiecare bloc trebuie să suporte o transformare cosinus. Scopul acesteia este de a cuprinde modificarea de culoare de la un bloc la altul. În etapa a doua, algoritmul verifică dacă diferentele de culoare dintre blocuri sunt vizibile cu ochiul liber. Diferentele dintre blocuri vecine care nu sunt observabile cu ochiul liber sunt ignorate.

În ultima etapa vor fi comprimate celelalte blocuri în mod obişnuit şi vor fi stocate pe harddisk într-un fişier JPEG. În funcție de conținutul imaginii, această metodă economiseşte între 50 şi 70 de procente din spațiu de stocare, fără să fie observată o pierdere a calității.

O consecință a comprimarii puternice cu JPEG este faptul că aceste fişiere pot fi transferate prin modem prin intermediul liniei telefonice. Astfel, majoritatea imaginilor din Internet sunt codificate JPEG. JPEG aduce cu sine, în schimb, şi o serie întreaga de dezavantaje. Dacă în cazul fotografiilor comprimarea decurge fără probleme, în schimb la desenele tehnice, de exemplu, apar probleme serioase.

O proprietate foarte utila a algoritmului JPEG este capacitatea acestuia de a avea un grad variabil de comprimare, ales de utilizator. Aceasta înseamnă că dacă se doreşte obținerea unui fişier de imagine cât mai mic, se poate alege o rată mare de comprimare, în dauna calității; invers, pentru a menține calitatea la o cotă ridicată, se alege un grad redus de comprimare. La prima comprimare a imaginilor, chiar si pentru un grad mediu, pierderea de calitate este minoră. Aşa se face că toate aparatele fotografice digitale, pentru a realiza economie de spațiu pe cartela de memorie, prezintă posibilitatea de a comprima imaginea descărcată de pe CCD în imagine comprimată JPEG în diverse rapoarte (1:5 - 1:20), fără o alterare semnificativă. Dacă însă fişierul JPEG este deschis, prelucrat, salvat, apoi din nou deschis, prelucrat ... de mai multe ori, pierderile de calitate devin tot mai evidente. Cât de mult se poate comprima o imagine? Dacă se pleacă de la un fişier imagine de tip TIFF de 1 MB (să presupunem), putem obține un fisier jpg de 100 K (rata de compresie 1:10), fără alterări importante ale calității vizibile a imaginii. O comprimare la o rata de 1:40 - 1:50 aduce fişierul-imagine la 20 - 25 K cu pierderi perceptibile dar moderate. Este posibilă comprimarea de până la 1:1000, imaginea obținută având doar 1 K si p3strează mai multe detalii decât oferă multe programe de manipulare a imaginilor în modul "thumbnail".

Avantajele formatului

-Format universal, fără copyright, recunoscut de orice program de vizionare şi prelucrare a imaginilor;

-Foarte bun raport calitate+dimensiuni;

-Cel mai folosit format pentru arhivarea fotografiilor;

-Utilizat aproape exclusiv pentru afişarea fotografiilor pe internet;

-Codificare pe 24 biți (până la 16 milioane de culori).

Dezavantalele formatului

-Format cu pierderi,datorită algoritmului de compresie;

-Fişierele prea mari pentru grafice;

-Nu gestionează straturi Alfa (transparență).

Fişiere gif întrețesute şi fişiere jpg progresive

Randarea progresivă reprezintă apariția gradată a graficii, fiind o metodă utilă pentru a păstra atenția pe durata înărcării. Imaginea apare mai întâi "înceţoşată" şi apoi, în mod progresiv, devine mai clară.

Există două metode de randare progresivă în Web. Cea mai populară şi mai eficientă este folosirea fişierelor GIF întreţesute. Procedeul se poate aplica oricărui fişier GIF şi, de obicei, în programele grafice sau instrumente plug-in există opțiuni care să ajute.

Fişierele JPG *progresive* reprezintă răspunsul JPEG la întreţesere. Nu se poate face întreţeserea unui fişier JPG, dar există programe care creează ceea ce este cunoscut sub numele de fişier JPG progresiv. Exemplu: Photoshop 4.0 are opțiunea de randare progresivă pentru JPG.

BMP- bitmap

De când există PC-urile, informațiile de imagine se salvează în aşa-numitele fişiere bitmap. Sub sistemul de operare Windows, fişierele de format BMP se folosesc pentru a stoca imagini, icon-uri, cursoare, pointeri, etc.

Acestea pot fi recunoscute datorită extensiei *.BMP. Faptul că acest format, care ocupă mult spațiu de stocare, a supravieţuit atâta timp este de înțeles: în fişierele BMP sunt stocate informațiile aproape aşa cum Windows reprezintă intern imaginile. Aici, informația de culoare a fiecărui punct de imagine este stocată în 1, 4, 8, 16 sau 24 biți. Cu cât sunt rezervați mai mulți biți de culoare pentru fiecare punct de imagine, cu atât este mai mare adâncimea culorii şi implicit calitatea imaginii. Acest format prezintă însă şi un dezavantaj vizavi de celelalte: în masura în care este mai mare adâncimea de culoare, creste şi spațiul ocupat de fişier pe harddisk. Astfel se adună rapid câțiva MB, chiar şi în cazul unor imagini color mici, la o rezoluție de 640x480 pixeli. Fiecare fişier BMP este compus din mai multe părți:

-un header;

-un tabel de culoare (opțional);

-datele pentru punctele de imagine.

În header se află, de exemplu, informații despre câte puncte are înalțimea şi lățimea imaginii şi unde sunt stocate în fişier informațiile de imagine. În tabelul de culoare este stabilit modul în care sunt reprezentate culorile pe paleta de culori a plăcii grafice. Acest lucru este important mai ales atunci când imaginea este salvată cu 16 milioane de culori, însă placa grafică

poate reprezenta numai 256 de culori. La o adâncime de culoare de 16 şi 24 de biţi nu există un tabel de culoare, deoarece în acest caz informaţia de culoare este stocată direct în bitmap.

Deci informaţiile de imagine sunt stocate într-un fişier BMP în moduri diferite. Utilizatorul nu trebuie să-şi bată capul cu aceste amănunte, deoarece filtrele de import ale tuturor programelor de prelucrare de imagini afişeaza corect informaţiile de imagine.

Fişierele BMP pot fi transferate fără probleme între calculatoare. Deoarece datele nu sunt comprimate, la reprezentarea pe monitor sau la imprimare nu are de suferit calitatea imaginii. Dar cei care vor dori să colecţioneze pe harddisk mai multe imagini, este bine să evite fişierele BMP. În caz contrar se vor trezi rapid cu un harddisk plin.

Pixelii dintr-o imagine de tip bitmap sunt comprimaţi folosind o combinaţie a trei moduri de lucru:

- *Modul RLE* (run length encoded), transformă extenisia din BMP în RLE, este de 2 octeţi şi reprezintă de la 1 la 255 de pixeli, toţi de aceeaşi culoare. În secvenţa 04 07, de exemplu, urmează patru pixeli având culoarea 07;
- *Modul escape*, primul octet este 0 iar următorii octeţi semnifică una dintre următoarele trei posibilităţi: 0 reprezintă sfârşitul unei linii de imagine, 1 reprezintă sfârşitul fişierului bitmap, 2 indică o comandă delta. O comandă delta mai este urmată de încă doi octeţi care reprezintă deplasarea pe orizontală şi pe verticală privind locul unde va apare următorul pixel, în raport cu pixelul curent;
- *Absolute mode*. Prin acest mod primul octet are valoarea zero iar următorul octet are o valoare de 3 sau mai mare reprezentând numărul de pixeli necomprimaţi care urmează. Un exemplu de folosire a modului absolut este 00 03 09 08 06 care interpretat în mod adecvat ne spune că urmează trei pixeli având valorile 09, 08 şi 06.

PNG- Portable Network Graphics

Înca din 1994 a aparut ideea de format grafic nou, fără pierderi de calitate, care să suporte 24 de biţi de culoare şi mai ales, să fie gratuit. Astfel, în octombrie 1996 s-a nascut **PNG** (Portable Network Graphics), un format excepţional pentru acea vreme, dar care s-a impus destul de greu. Avantajul sau cel mai important erau pierderile de calitate nule, iar ca dezavantaj putem enunţa viteza destul de scazută a algoritmului de compresie. PNG este folosit în special pentru compresia imaginilor schematice, dar şi a celor de tip fotografie unde nu se acceptă pierderi de calitate. Dacă pentru mult timp au fost preferate alte formate, cum ar fi GIF sau JPG, din simplul motiv al compatibilitatii şi - de ce nu - al comodităţii, în prezent, orice aplicaţie din domeniu

poate lucra cu acest format. Odată cu introducerea suportului în browsere (fie direct, fie prin diverse plugin-uri), succesul i-a fost asigurat, el fiind în prezent cel mai performant format grafic de pe piață din sfera Internet-ului, însă totuşi nu este atât de răspândit ca GIF sau JPG. „Norocul" său a fost faptul că prin definiție, el este destinat web-ului, chiar dezvoltatorii lui numindu-se The World Wide Web Consortium (W3C). PNG este unul din puținele cazuri din istorie în care un format mai performant dar initial necunoscut, a putut cuceri piata. Să ne amintim de muzica şi de formatul MP3, care este în prezent este destul de slab ca performanțe, dar extrem de popular, alternativele chiar dacă sunt mai bune din toate punctele de vedere, sunt ignorate de piață,.

În încercarea de a corecta deficiențele fişierelor de tip GIF, a aparut formatul PNG (**P**rotable **N**etwork **G**raphic) care prezintă codificare pe 8 biți (PNG-8) sau pe 24 biți (PNG-24), inclusiv gestionarea transparenței pe 256 nivele. Formatul PNG este recunoscut de practic toate programele moderne de prelucrare şi navigare pe Internet. Pentru aceeaşi imagine, fişierele PNG sunt mai mici decât cele în varianta GIF. Formatul PNG-24 suportă şi imagini cu tonuri continue, dar dimensiunile fişierelor sunt mai mari decât cele JPEG.

Avantaje:

- Format universal, folosit pentru Internet, recunoscut de orice nagivator modern, fără să aibă nevoie de plugin-uri.
- Comprimă imaginea fără pierderi de calitate, până la dimensiuni foarte mici. Poate gestiona transparența mai bine decat formatul GIF.

Dezavantaje:

- Incompatibil cu versiunile vechi de navigare Internet.
- Fotografiile sunt comprimate mai puțin decât formatul JPEG.
- Fara posibilitate de animare.

DIB- device independent bitmap

DIB nu este un format de fişier imagine, ci este formatul în care pot fi păstrate imaginile în memorie de aplicațiile Windows. De obicei programele care doresc afişarea diferitelor formate de fişiere imagine recurg la acest format ca şi format intermediar.
DIB reprezintă o hartă de pixeli, independentă de dispozitiv.

Tabloul de pixeli poate fi memorat necomprimat sau folosind o metodă de compresie de tip RLE (Run-Length Encoded) pe 4 sau pe 8 biți. Metoda de stocare a valorii pixelilor este specificată de valoarea câmpului biCompression din headerul formatului.

În formulare şi rapoarte există posibilitatea să se afişeze imagini, precum şi ilustrații, sigle sau fotografii. Pentru aceasta, imaginile trebuie mai întâi stocate. Access furnizează câteva metode de stocare a imaginilor.

O posibilitate este următoarea: încorporarea imaginilor într-un câmp de tip Obiect OLE dintr-un tabel al bazei de date.

OLE este tehnologia utilizată pentru a partaja fişiere între diversele programe Office. De exemplu, când se inserează o foaie de calcul Excel într-un document Word sau se inserează un diapozitiv Microsoft PowerPoint într-un desen Microsoft Visio, se utilizează OLE. Un câmp Obiect OLE se utilizează când este necesară stocarea imaginilor (sau legături la ele) şi a altor fişiere din alte programe Office direct în baza de date.

Această metodă este cel mai simplu de implementat deoarece se utilizează ecranele şi instrumentele furnizate de Access. De asemenea, imaginile devin parte a bazei de date şi se deplasează împreună cu ea. Nu va fi nevoie niciodată să se actualizeze legăturile la fişierele imagine.

Încorporarea imaginilor, însă, poate duce rapid la creşterea dimensiunii bazei de date şi la o execuție lentă. Aceasta se întâmplă în special când se stochează fişiere GIF şi JPEG, deoarece OLE creează fişiere bitmap suplimentare care conțin informații despre afişare pentru fiecare fişier imagine, iar aceste fişiere suplimentare pot fi mai mari decât imaginile inițiale. În plus, această metodă acceptă numai formatele de fişier grafic Windows Bitmap (.bmp) şi Device Independent Bitmap (.dib). Dacă se doreşte să se afişeze alte tipuri obişnuite de fişiere imagine, cum ar fi imaginile GIF şi JPEG, trebuie să se instaleze software suplimentar.

GPX- GPX exchange format

GPX este un format de prezentare a datelor GPS pentru aplicații software derivat din XML (Extensible Markup Language). Poate fi folosit pentru a descrie coordonate şi rute. Formatul nu este supus vreunei licențe şi poate fi folosit fără a plăti vreo taxă.

Marcajele pe care le foloseşte stochează locații, altitudini sau date temporale şi astfel GPX poate fi folosit pentru schimbul de date între echipamentele GPS şi aplicațiile software.

Acest aplicații pot fi folosite împreună cu programe de imagistică prin satelit, cum ar fi Google Earth.

În GPX, un grup de puncte fără vreo legătură secvențială între ele este tratat ca un grup de coordonate. Un grup ordonat de puncte poate fi tratat ca o rută sau ca un traseu. Un traseu reprezintă înregistrarea locurilor pe unde a trecut o persoană, pe când rutele reprezintă destinații posibile în viitor. Aşadar, în cazul traseelor pot exista marcaje temporale pentru fiecare punct din traseu (pentru că cineva a notat când a trecut pe acolo), dar în cazul rutelor este puțin probabil ca acestea să fie furnizate. Datele minimale care trebuie incluse într-un document GPX sunt latitudinea şi longitudinea pentru o coordonată, restul variabilelor fiind opționale.

EPS- Encapsulated postscript

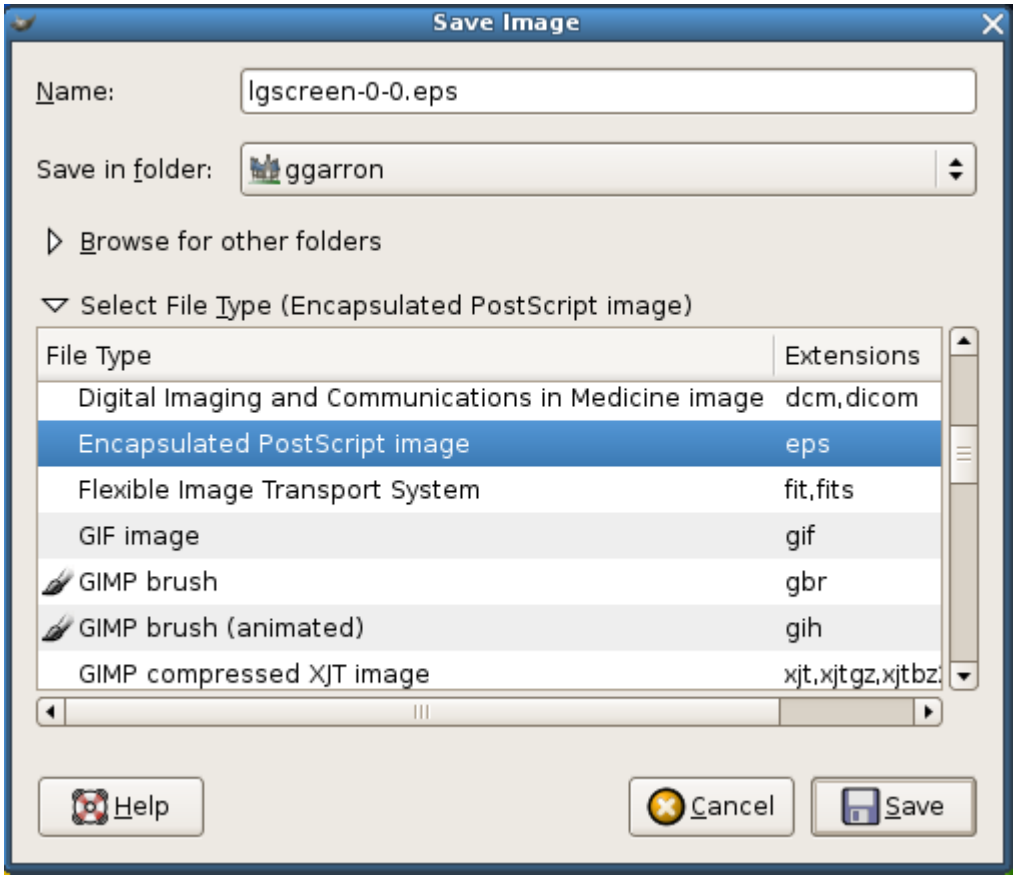


Fig. 2.3. Encapsulated postscript

Formatul de fişier Encapsulated PostScript (*EPS*) poate conține atât grafică vectorială cât şi bitmap şi este acceptat teoretic de toate programele pentru grafică, ilustrații şi machetare a paginilor. Formatul *EPS* este utilizat pentru transferul ilustrațiilor PostScript între aplicații. Când se deschide un fişier *EPS* care conține grafică vectorială, Photoshop rasterizează imaginea, convertind grafica vectorială în pixeli._(Rasterizarea = convertirea imaginilor si textului unui document in imagini raster - un model electronic constând in pixeli - pentru a fi tiparite sau

printate. Aplicația RIP transformă informațiile vectoriale în instrucțiuni executabile de echipamentele de tipărire.)

Formatul *EPS* acceptă modurile de culoare Lab, CMYK, RGB, Culoare indexată, Bicromie, tonuri de gri şi Bitmap, şi nu acceptă canalele alfa. *EPS* nu acceptă traseele de tăiere. Formatul Desktop Color Separations (DCS) - o versiune a formatului *EPS* standard care permite salvarea separațiilor de culoare ale imaginilor CMYK. Pentru a tipări fişiere *EPS*, trebuie utilizată o imprimantă PostScript.

Photoshop utilizează formatele EPS TIFF şi EPS PICT pentru a permite deschiderea de imagini salvate în formate de fişiere care creează previzualizări, dar care nu sunt acceptate de Photoshop (cum ar fi QuarkXPress).

Encapsulated PostScript poate avea extensiile: *.eps* , *.epsf* (grafice şi imagini porționate ale documentului), *.epsi (bitmap).*

CDR: Corel draw format

CorelDRAW este un program de grafică vectorială, creat pentru a rula pe platforme Microsoft Windows (cel mai important pachet de grafică sub Windows). El permite crearea unor ilustrații mai deosebite, publicitate la nivel profesional, rapoarte anuale. Este un program de desenare orientat pe obiecte. Din multe puncte de vedere este asemanător cu programele de proiectare asistată pe calculator (AutoCad). Acest program se aseamană cu programele de pictură care folosesc grafica realizată prin puncte şi dau iluzia picturii pe pânză. Programele bazate pe grafică vectorială, cum este CorelDRAW sunt de preferat celor ce realizează grafica prin puncte. Avantajele graficii vectoriale sunt precizia şi flexibilitatea.

Tabel 3. Fişiere recunoscute de CorelDRAW

Adobe Illustrator	(AI)
Adobe Photoshop	(PSD)
Adobe Type 1 Font	(PFB)
ANSI Text	(TXT)
AutoCAD Database	(DWG)
AutoCAD Interchange Format	(DXF)
CALS Compressed Bitmap	(CAL)
Computer Graphics Metafile	(CGM)
Corel ArtShow 5	(CPX)
Corel DESIGNER	(DSF or DES)
Corel DESIGNER	(CDT)
Corel Paint Shop Pro	(PSP)
Corel Painter	(RIF)
Corel PHOTO-PAINT	(CPT)
Corel Presentation Exchange	(CMX)
Corel Presentations	(SHW)
Corel R.A.V.E.	(CLK)
CorelDRAW	(CDR)

CorelDRAW Compressed	(CDX)
Encapsulated PostScript	(EPS)
FPX	
Frame Vector Metafile	(FMV)
GEM File	(GEM)
GEM Paint	(IMG)
GIF	
GIMP	(XCF)
Hewlett-Packard Plotter	(PLT)
HTML	
JPEG	(JPG)
JPEG 2000	(JP2)
Lotus PIC	(PIC)
MacPaint Bitmap	(MAC)
Macromedia Flash	(SWF)
Macromedia FreeHand	(FH)
MET Metafile	(MET)
Micrografx 2.x, 3.x	(DRW)
Micrografx Picture Publisher 4	(PP4)
Micrografx Picture Publisher 5	(PP5)
Microsoft PowerPoint	(PPT)
Microsoft Word Document	(DOC sau RTF)
NAP Metafile	(NAP)
OS/2 Bitmap	(BMP)
Pattern File	(PAT)
Macintosh PICT	(PCT)
Picture Publisher File	(PPF)
Portable Document Format	(PDF)
Portable Network Graphics	(PNG)
PostScript	(PS sau PRN)
PostScript Interpreted	(PRN)
PostScript Interpreted	(PS)
Rich Text Format	(RTF)
Scalable Vector Graphics	(SVG)
SCITEX CT Bitmap	(SCT)
TIFF Bitmap	(TIF)
TrueType Font	(TTF)
Visio	(VSD)
Windows Bitmap	(BMP)
Windows Metafile Format	(WMF)
WordPerfect Document	(WPD)
WordPerfect Graphic	(WPG)
XPixMap Image	(XPM)

CMX: Vector image file format

CMX: Corel Presentation Exchange Image

Cmx este un format vectorial grafic, un fişier *.cmx* conţinând o descriere a modului în care se trasează imaginea şi nu doar o descriere a modului de afişare. Apoi imaginea poate fi afişată la orice dimensiune, fiind interpretată, scalată şi afişată de către browser-ul plug in.

Fiecare obiect într-o imagine vectorială este stocat ca un element distinct, cu informații despre poziția sa din imagine, dimensiune, culoare etc. O imagine în format vectorial este de rezoluție-independentă. Ea poate fi redimensionată, fără a pierde nici un detaliu, pentru că este stocată ca un set de instrucțiuni, nu o colecție de pixeli.

Acest format nu este potrivit pentru fotografii obişnuite, din "viața reală", el funcționând foarte bine pentru planuri, hărți, miniaturi (clipart) şi grafice de afaceri. Imaginile *.cmx* pot fi vizualizate numai de Netscape Navigator TM 2.0 pentru Windows 95 şi Windows NT.

RAW: Raw image format

RAW (Raw image format) este un format de imagine în care sunt stocate datele brute, neprocesate, capturate de senzorii unei camere foto digitale sau ale unui scanner.

Atunci când au apărut scannerele şi camerele foto digitale, senzorii nu aveau calitatea necesară pentru a captura prea multe nuanțe de culoare, aşa că spațiul de culoare sRGB era suficient pentru stocarea imaginii capturate. Pe măsură ce calitatea senzorilor s-a îmbunătățit, dispozitivele de captură au ajuns să poată capta o plajă mai mare de culori şi nivele de luminozitate decât cea care poate fi stocată într-un fişier JPEG, de exemplu. Ca atare, cineva trebuia să decidă în ce fel să fie expusă imaginea: oricât de bun ar fi fost senzorul, imaginea finală din spațiul sRGB sau Adobe RGB trebuia în mod necesar să reprezinte o gamă dinamică mai mică decât cea capturată de senzori. Cele mai multe dispozitive moderne de captură de calitate rezonabilă, atât scannere cât şi camere de fotografiat, permit utilizatorului să aleagă între două moduri majore de stocare a imaginii: fie imaginea este stocată în spațiul sRGB lăsând dispozitivul să aleagă metoda de compresie a gamei dinamice, fie este stocată sub formă brută, urmând ca toți paşii asociați cu compresia gamei dinamice să fie efectuați ulterior de către fotograf. Această a doua variantă de stocare produce imagini în format brut, aşa-numitul ***RAW format.***.

Pe lângă informația brută provenită de la senzori, formatul RAW conține informații conexe despre imagine (informații care sunt stocate de altfel şi în imagini JPEG), plus informații legate de senzor şi o reprezentare redusă a imaginii pentru identificare ei vizuală, fără a fi necesară procesare suplimentară. RAW este un format relevant în fotografia digitală, alături de Tiff sau JPEG.

RAW este o alternativa superioara fata de TIFF. Fişierele RAW sunt mai reduse şi pot reține chiar mai multe informații despre imagine .

Fisierele RAW pot oferi o plajă mult mai extinsă de ajustări decât fişierele JPEG: peste 50% mai multe detalii la expunere (acest lucru depinde însă de capacitatea senzorului). Doarece fişierele RAW conţin o gamă mai variată de culori, balansul pe alb poate fi corectat fără piederi mari de detalii. Folosint convertoarele RAW cum ar fi plugin-ul Photoshop Camera, RAW este mai uşor şi produce rezultate mult mai bune.

Marele avantaj al lucrului în formatul RAW este dat de faptul că asigură o mare flexibilitate, fiind formatul care oferă şi cel mai mult spaţiu pentru modificarea imaginii, păstrând în acelaşi timp calitatea.

Formatul RAW oferă :

- Imagini cu cel mai mare interval dinamic - putând înregistra mai bine zonele luminoase şi umbrele
- Imagini al căror balans de alb poate fi reglat în orice moment
- Un mai bun control al clarităţii în comparaţie cu alte formate de fişiere
- O mai mare flexibilitate în corectarea erorilor de expunere

RAW-urile conţin toate informaţiile înregistrate de cameră în momentul captării imaginii, însă au două mari defecte: ocupă mult spaţiu şi necesită un editor specializat pentru a le deschide.

2.1.3. Formate audio

AIFF Audio Interchange File Format

Audio Interchange File Format Audio File Interchange Format (AIFF) este un format audio standard utilizat pentru stocarea datelor de sunet pentru calculatoare personale şi alte dispozitive electronice audio. Formatul a fost co-dezvoltat de Apple Computer, în 1988 pe baza Interchange Electronic Arts - Format de fişier (IFF, utilizate pe scară largă la sistemele Amiga) şi este cel mai frecvent utilizat pe sisteme de computere Apple Macintosh.

Datele audio într-un fişier AIFF standard sunt impulsuri-modulare necomprimate cod (PCM). Există, de asemenea, o variantă comprimată de AIFF cunoscut sub numele de AIFF-C sau aifc, cu diverse codec-uri de compresie definite.

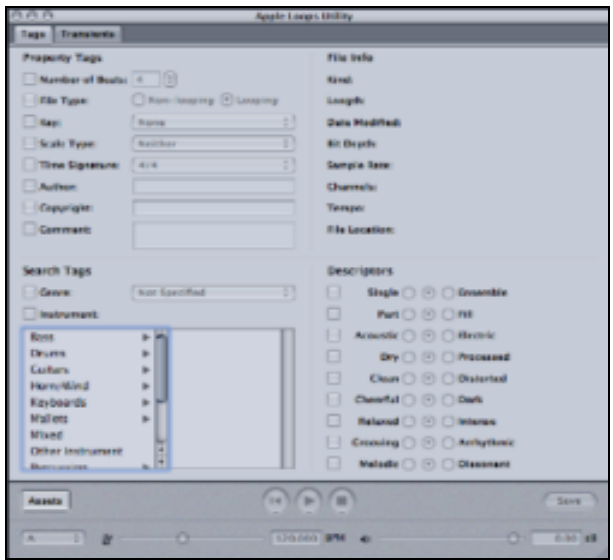


Fig. 2.4. Apple Loops Utility

Standard AIFF este un format de conducere (împreună cu SDII şi WAV), utilizat la nivel profesional în aplicații audio şi video, iar spre deosebire de formatul mai bine cunoscut lossy MP3, acesta este non-comprimat (care SIDA rapidă de streaming de mai multe fişiere audio de la disc pentru a cererii), şi fără pierderi. Ca orice format non-comprimat fără pierderi, foloseşte spațiu pe disc mult mai mult decât MP3-aproximativ 10 MB pentru un minut de audio stereo la o rată eşantion de 44,1 kHz şi un eşantion de 16 biți. În plus față de datele audio, AIFF poate include date de tip bucla punct, iar nota muzicala a unei probe, pentru a fi utilizate de către prelevatorii hardware şi aplicații muzicale.

Extensiile pentru formatul standard AIFF sunt .aiff sau .aif. Pentru variantele comprimate se presupun a fi .aifc, dar si .aiff sau .aif sunt acceptate de aplicațiile audio.

Apple a mai creat recent o extensie la formatul AIFF sub forma Apple Loops (buclă) folosit mai ales de GarageBand şi Logic Audio, care permit includerea de date pentru trecerea ritmului şi a tempo-ului în aplicații mai comune.

Un fişier AIFF este împărțit într-un număr de părți fiecare identificată de un ID .

Tipuri de părți în fişierele AIFF:

- Common Chunk (required)
- Sound Data Chunk (required)
- Marker Chunk
- Instrument Chunk
- Comment Chunk
- Name Chunk
- Author Chunk
- Copyright Chunk
- Annotation Chunk
- Audio Recording Chunk
- MIDI Data Chunk
- Application Chunk

WAV - Waveform audio format

WAV este o prescurtare a „Waveform Audio Format", adică „format de undă audio", standard dezvoltat de IBM si Microsoft pentru pastrarea fisierelor audio pe PC. Este un format fără compresie folosit de obicei pentru secvențe scurte de sunet și înregistrări vocale. Raspindirea sa a scazut in ultimul timp, datorita mai ales dimensiunii mari a fisierelor si a aparitiei formatului mp3.

Acest format de fișier audio stochează sunetele sub formă de unde. În funcție de mai mulți factori, un minut de sunet poate stoca o dimensiune foarte mică, precum 644 de kiloocteți sau una foarte mare, cum ar fi 27 de megaocteți.

WAV este compatibil cu Windows, Macintosh, şi sisteme de operare Linux.

Deşi un fişier WAV poate conține fişiere audio comprimate, forma cea mai comună WAV conține fişiere necomprimate audio în module liniare cod format (LPCM). Formatul standard de fişiere audio pentru CD-uri, de exemplu, este LPCM-codificate, care conține două canale de 44100 de eşantioane pe secundă, 16 biți per eşantion. Deoarece LPCM utilizează o metodă de stocare necomprimată, care pastreaza toate mostrele de pista audio, utilizatorii profesionişti sau experți audio poate folosi formatul WAV pentru calitate maxima audio. WAV audio poate fi, de asemenea, editat şi manipulat cu o relativă uşurință utilizând software-ul. Format WAV sprijină comprimatul audio, folosind, pe Windows, Audio Compression Manager.

Datorită structurii simple, bazate pe formatul IFF, este folosit cu o varietate de aplicații software.

Datorită dimensiunilor mici ale fişierelor comprimate, dar formate lossy cum ar fi MP3, ATRAC, AAC, Ogg Vorbis, WMA acestea sunt folosite pentru a stocarea şi transferul datelor audio. Dimensiunile mici ale acestor fişiere permit o transmitere rapidă pe Internet precum şi un consum mai mic de spațiu pe mass-media de memorie. Cu toate acestea, formatele lossy comprimate pierd din calitatea audio.

Formatul WAV este limitat la fişiere de cel mult 4GB, ceea ce ar însemna circa 6,8 ore de calitate CD-audio, motiv pentru care a fost necesară depăşirea acestor limite prin crearea formatului W64.

De când frecvența transferului de date a unui fişier WAV a variat de la 1Hz la 4,3 GHz şi numărul de canale este mai mare de 65536, fişierele .wav au fost folosite pentru date nonaudio.

WMA- Windows Media Audio

Windows Media Audio este o tehnologie audio de compresie de date dezvoltată de Microsoft. Numele poate fi utilizat pentru a se referi la formatul de fişier audio sau codecurile sale audio. Ea face parte din cadrul Windows media.

WMA este un format „proprietate” de compresie audio ce oferă posibilitatea protejării fişierelor audio împotriva copierii ilegale printr-o tehnică numită „gestiune numerică de drepturi”.

WMA există în patru forme (codecuri):

- WMA standard- codec original un concurent pentru MP3 şi codec-uri Real Audio, cel mai răspândit pe Internet fiind „lizibil” pentru mai multe browser-e;

- WMA Pro- codec mai nou şi mai avansat, susține mai multe canale şi are o înaltă rezoluție audio, dar mai puțin răspândit;

- WMA Lossless- comprimă datele audio, fără pierderi de fidelitate audio, oferind o calitate sonoră identică cu a originalului,

- WMA Voice – în special folosit pentru codificarea vocii, comprimarea se face cu o viteză mai mică sau egală cu 20kbit/s.

Primul codec WMA s-a bazat pe munca lui Henrique Malvar şi a echipei sale care au lucrat la proiectul MSAudio. Primul codec finalizat a fost prevăzut ca o continuare MSAudio 4.0. Mai târziu a fost lansat oficial ca Windows Media Audio, ca partener a Windows Media Technologies 4.0.

Windows Media Audio Standard (WMA) este cel mai utilizat codec din cele enumerate mai sus.

Versiuni disponibile WMA: Windows media Audio 2 îm 1999, Windows Media Audio 7 în 2000, Windows Media Audio 8 în 2001 şi Windows Media Audio 9 în 2003 (cu codecurile: Windows Media Audio 9 Professional, Windows Media Audio 9 Lossless şi Windows Media Audio 9 Voice) , Windows Media Audio Professional 10 în 2007.

Windows Media Audio 9 are o frecvență de 44,10 sau 48 kHz, ocupă 16 B, similar CD-ului, sunetul având aceeaşi calitate şi o rată cuprinsă între 64 şi 192 kB/s. În această variantă calitatea sunetului este cu 20% mai bună decât a versiunii din 2003. Comparativ cu MP3 calitatea sunetului este mai mare la aceeaşi rată de biți.

AVI Audio Video Interleave

AVI este un format multimedia de tip „container” conceput ca parte a Video pentru Windows de către Microsoft în 1992 pentru a stoca date audio-video, permițând sincronizarea audio-video.

Cele mai multe formate AVI folosesc extensiile dezvoltatede Matrox Open DML Group, care neoficial se numesc AVI 2.0.

Într-un fişier AVI, fiecare element audio sau video poate fi comprimat de orice codec. Formatul DivX este codec-ul folosit pentru datele video şi formatul Mp3 pentru datele audio dar

pot fi folosite şi alte codec-uri, cum ar fi: XviD, MPEG pentru video şi Mp2 şi WAV pentru audio.

În acest format pot fi reunite într-un singur fişier a unei piste video şi până la 99 canale audio, ceea ce permite de exemplu subtitrarea unui film în mai multe limbi.

Multe fişiere AVI folosesc acum extensii ale formatelor dezvoltate de Matrox Open DML în februarie 1996. Aceste fişiere sunt acceptate de Microsoft şi sunt numite neoficial AVI 2.0.

DAT- Digital Audio Tape

Digital Audio Tape (DAT sau R-DAT) este un suport de înregistrare şi mijloc de redare audio numeric pe bandă magnetică de 3,14 mm dezvoltat de Sony în anii 80. Ca aspect este similară casetei audio dar este mai mică 73x54x10,5mm şi face parte din categoria mijloacelor de înregistrare digitale asigurând copierea exactă a unui alt material digital, spre deosebire de copierea pe CD.

Tehnologia DAT este stâns legată de cea a magnetoscoapelor, folosind un cap rotativ pentru înregistrare diferența constând în cantitatea mult mai mare de date ce poate fi stocată.

DAT nu a fost prima casetă audio- digitală, a mai existat o tentativă în 1972 în Japonia, dezvoltată de Denon pentru înlocuirea discurilor de vinilin, dar nu s-a dezvoltat într-un produs de consum.

În 1976 primul succes comercial al DAT-ului a fost dezvoltat de Sony folosind un aparat de înregistrare fabricat de Honeywell conectat la un harware ce folosea programe de codare- decodare create de Sony.

La scurt timp după Sony, 3M a introdus propria linie de înregistratoare şi formate ale acestor casete pentru a putea fi folosite în studiourile de înregistrări din Minneapolis, Minnesota.

La sfârşitul anilor 80, Recording Industry Association of America a făcut lpbby împotriva introducerii dispozitivelor DAT în SUA, pentru a preveni folosirea acestora pentru copierea discurilor sau casetelor.

Acest gen de acțiuni au dus la încetarea fabricării Replayer-ului DAT pe scară largă.

Acest gen de format de stocare a fost utilizat pe scară largă în industria de înregistrări audio profesionale şi în radiodifuziune datorită faptului că permitea codificarea fără pierderi şi cantitatea de date stocate varia între 1,3 şi 80 GB pe o bandă de 60 – 180 metri.

Transferul de date de pe bandă pe hard disc se face prin intermediul unei conexiuni digitale de tip AES/EBU.

WMV- Windows Media Video

WMA este numele unei familii de codec-uri video dezvoltate de Microsoft. Pe Internet se întâlneşte frecvent acest tip de fişier. Ca cea mai mare parte a codec-urilor „proprietate" acestă tehnologie nu corespunde cu cea a normelor internaționale (ISO/IEC MPEG), Microsoft dorind intrarea acestor codec-uri pe cele mai multe obiecte de larg consum (telefoane, DVD playere, decodoare TV) altele decât cele obişnuite, a încercat introducerea versiunii 9, WMV9, în consorțiile industriale (3GPP, DVB, ATSC, DVD-Forum, Blu-ray).

Fiind în parte în neconcordanță cu normele MPEG-4 AVC/H 264, ce au fost inițiate şi dezvoltate de crecetători din companii internaționale şi universități din toată lumea, pe de altă parte fiind obligați de consorțiile industriale de a prezenta un codec standardizat, Microsoft a înscris codec-ul WMV9 la SMPTE (society of Motion Picture and Television Engineers) pentru a-l standardiza sub numele VC-1.

În 2005, codec-ul VC-1 devine stabil şi are statut de standard internațional.

VC-1 are versiune high definition, a fost implementat pe HD DVD de la NEC şi pe discurile Blu-ray de la sony, care probabbil vor fi înlocuitorii DVD- playerelor de astăzi.

Codecuri Windows Media Video :

- Windows Media Video v.7
- Windows Media Video Screen v.7
- Windows Media Video v.8
- Windows Media Video v.9
- Windows Media Video v.9 Screen
- Windows Media Video 9 Advanced Profile (FourCC: WVC1)
- Windows Media Video 9 Image
- Windows Media Video 9.1 Image
- VC-1

2.1.4. Formate video

Standardele MPEG. MPEG este acronimul pentru Moving Picture Experts Group (grupul experților în filme) creat de organizația internațională pentru standardizare (ISO).

Grupul a fost format pentru a stabili standarde de compresie şi transmisie audio şi video. Prima întâlnire a grupului a avut loc în oraşul Ottawa din Canada în mai 1988.

Metodologia de compresie MPEG este considerată asimetrică deoarece codorul este mai complex decât decodorul. Codorul trebuie să fie algoritmic sau adaptabil pe când decodorul execută acțiuni fixe. Acest fapt este considerat un avantaj în aplicațiile de difuzare în masă unde numărul de codoare scumpe este mic dar numărul decodoarelor ieftine este mare.

Standardele MPEG sunt compuse din mai multe părți. Fiecare parte acoperă un anumit aspect al întregii specificații. Standardele specifică de asemenea şi profilele şi nivelele. Profilele sunt destinate să definească un set de instrumente care sunt valabile, iar nivelele definesc gama de valori potrivite pentru proprietățile acestora. MPEG a standardizat următoarele formate de compresie şi standarde auxiliare:

•MPEG-1: Acest standard a apărut în anul 1993. Utilizat la codarea ilmului şi a sunetului asociat pentru medii digitale de stocare cu viteze până la 1,5 Mbit/s. Este primul standard pentru compresie audio şi video. Principiul de proiectare a fost permiterea codării filmului şi a sunetului asociat la rata de biți a unui disc compact (CD). Pentru aceasta MPEG-1 reduce rata semnalului de transmisie pentru imagini, dar utilizează şi o frecvență a pozelor de doar 24-30 Hz, rezultând o calitate moderată. Include popularul format de compresie audio MP3.

•MPEG-2: Standardul a fost lansat în anul 1995. Era utilizat pentru codarea generică a filmelor şi a informațiilor audio asociate. MPEG-2 suportă întrepătrundere şi o înaltă definiție. Standardul este considerat important deoarece a fost ales ca şi schemă de compresie pentru televiziunile digitale ATSC, DVB şi ISDB, servicii digitale prin satelit precum Dish Network, SVCD, DVD şi Blu-ray.

•MPEG-3: Se ocupă de standardizarea compresării scalabile şi de rezoluții multiple şi a fost creat pentru compresie HDTV însă a fost descoperit că este redundant şi a fost contopit cu MPEG-2, astfel că nu există standardul MPEG-3. A nu se confunda MPEG-3 cu MP3 care este stratul audio 3 al standardului MPEG-1.

•MPEG-4: A apărut în 1998. A fost creat pentru codarea obiectelor audio-vizuale. MPEG-4 foloseşte unelte de codare cu complexitate sporită pentru a atinge factori de compresie mai mari decât MPEG-2. Pe lângă codarea video mai eficientă, MPEG-4 se apropie mai mult de aplicațiile grafice. În profile mai complexe, decodorul MPEG-4

devine un eficient procesor de interpretare grafică, iar fluxul de date compresate descrie forme tri-dimensionale şi texturi ale suprafețelor.

Formatul MOV

Formatul Quicktime (.mov) funcționează ca un recipient care conține una sau mai multe piste, fiecare conținând un singur tip de date: audio, video, efecte sau text. Fiecare pistă conține un flux de date codate digital (folosind un codec specific) sau o referință către fluxul de date aflat într-un alt fişier. Pistele sunt menținute într-o structură de date ierarhizată care este formată din obiecte numite atomi. Un atom poate fi părintele altor atomi sau poate conține date media, dar nu poate fi ambele.

Abilitatea de a conține referințe abstracte de date pentru datele media şi separarea datelor media de listele de editare a pistelor duc la faptul că formatul Quicktime este printre cele mai potrivite pentru editare, fiind capabil să importe şi să modifice datele pe loc fără necesitarea rescrierii întregului fişier.

Formatul OGG

OGG este un format standard gratuit, deschis developării publice, întreținut de fundația Xiph.Org.

Creatorii formatului susțin că acesta nu este restricționat de brevete de software şi este proiectat să producă un eficient flux de date şi o manipulare eficientă a datelor multimedia digitale de calitate înaltă.

Numele „Ogg” se referă la formatul care poate multiplexa mai multe codecuri gratuite, deschise developării publice, pentru date audio, video, text şi meta.

În cadrul arhitecturii multimedia Ogg, Theora furnizează stratul video, în timp ce codecul orientat spre muzică Vorbis acționează ca strat audio. Ca şi strat audio mai pot fi utilizate şi Speex, FLAC sau OggPCM.

Termenul „Ogg” este deseori folosit ca referință la formatul audio Ogg Vorbis care este de fapt audio codat folosind Vorbis în formatul de stocare Ogg. Anterior, extensia .ogg era folosită pentru orice conținut distribuit prin Ogg, însă începând cu anul 2007, fundația Xiph.Org a cerut ca extensia .ogg să fie utilizată doar pentru codecul Vorbis pe baza îngrijorărilor legate de compatibilitate inversă. Fundația a decis să creeze un nou set de extensii ale fişierelor şi tipurilor media pentru a descrie diferitele tipuri de

conținut precum .oga pentru fișiere cu conținut exclusiv audio, .ogv pentru video cu sau fără audio (inclusiv Theora), şi .ogx pentru aplicații.

Formatul AVI

AVI, acronimul de la „Audio Video Interleave" (Interclasare audio video), este un format de stocare multimedia digital introdus de Microsoft în 1992 ca parte a tehnologiei „Video pentru Windows". Fişierele AVI pot conține atât date audio cât şi date video într-un fişier care permite redare sincronizată audio-video, însă nu e necesar ca acestea să apară împreună (exemplul unui clip fără sunet).

AVI este o derivare a formatului intitulat RIFF care împarte datele unui fişier în bucăți. Fiecare bucată este identificată de o etichetă FourCC(cod de patru caractere) . Un fişier AVI ia forma unui fişier de tip RIFF format dintr-o singură bucată care este divizată în două părți mai mici obligatorii şi o parte opțională.

Prima parte obligatorie este identificată de eticheta „hdrl" şi reprezintă antetul fişierului care conține informații legate de video precum înălțimea, lățimea şi rata cadrelor. A doua parte obligatorie este identificată de eticheta „movi" şi conține toate datele audio-video care alcătuiesc filmul. Partea opțională este identificată de eticheta „idx1" şi indexează adresele relative ale bucăților de date din fişier.

Asemănător formatului RIFF, datele audio-video conținute în partea „movi" pot fi codate sau decodate de un software numit codec. La crearea fişierului, codecul translatează între datele neprelucrate şi tipul de date compresate folosit în fişier. Un fişier AVI poate stoca datele audiovideo în aproape toate formele de compresie, inclusiv : Full Frame (necompresat), Intel Real Time (Indeo), Cinepak, Motion JPEG, Editable MPEG, VDOWave, ClearVideo/RealVideo, QPEG şi MPEG-4 video.

Secțiunea ce urmează nu descrie formatul OpenDML AVI (aşa numit-ul "AVI 2.0"), dezvoltat de compania Matrox, ci doar specificațiile Microsoft.

Fişierele AVI sunt identificate după FourCC-ul „AVI " (inclusiv spațiul, al patrulea caracter) din antetul RIFF. Toate AVI-urile au cel puțin două liste esențiale de părticele, care conțin formatele fluxurilor, respectiv fluxurile de date propriu-zise. Un AVI poate avea, opțional, o porție ce conține locațiile celorlalte porții de date din fişier. Atunci AVI-ul ar avea următoarea structură:

RIFF ('AVI '

LIST ('hdrl' ...)

LIST ('movi' ...)

['idx1' (<AVI Index>)]

Aici, „hdrl” şi „movi” sunt ID-urile celor două liste esențiale despre care vorbeam. „hdrl” defineşte formatul de date şi este prima listă necesară. „movi” conține datele pentru secvența audiovideo şi este cea de-a doua listă necesară. Porția cu ID-ul „idx1” conține indecşii. Fişierele TREBUIE să aibă aceste trei componente în ordine.

Listele „hdrl” şi „movi” au ca date subporții (subchunks). Următorul exemplu prezintă într-un mod ceva mai extins structura unui fişier RIFF AVI:

RIFF ('AVI '

LIST ('hdrl'

'avih'(<Main AVI Header>)

LIST ('strl'

'strh'(<Stream header>)

'strf'(<Stream format>)

['strd'(<Additional header data>)]

['strn'(<Stream name>)]

LIST ('movi'

{SubChunk | LIST ('rec '

SubChunk1

SubChunk2

['idx1' (<AVI Index>)]

Antetul AVI principal

Lista „hdrl” începe cu sub-chunkul „avih” (AVI header), ce conține antetul AVI principal, informații valabile pentru întregul fişier AVI, cum ar fi numărul de fluxuri din fişier (en. streams –spre exemplu, un fişier şi video şi audio are două fluxuri), lățimea şi înălțimea secvenței AVI (în pixeli), numărul de microsecunde dintre cadre, numărul total de cadre şi alte informații. Pentru mai multe detalii, rmăreşte linkul http://msdn.microsoft.com/en-us/library/ms779632(VS.85).aspx, unde este definită structura AVIMAINHEADER.

Antetele fluxurilor AVI

După „avih” (AVI main header) urmează una sau mai multe liste „strl”. Pentru fiecare flux de date (audio, video) este necesară o listă „strl”. O astfel de listă conține informații despre fluxul asociat ei şi trebuie să aibă o porție pentru antetul fluxului („strh” – stream header chunk) şi o porțiune pentru formatul fluxului („strf” – stream format chunk). Opțional, o listă „strl” poate conține porțiunile „strd” (stream header data chunk) şi „strn” (stream name chunk).

Porțiunea cu ID-ul „strh” alcătuieşte o structură AVISTREAMHEADER. Pentru mai multe informații, urmează linkul http://msdn.microsoft.com/en-us/library/ms779638(VS.85).aspx.

Porțiunea cu ID-ul „strf” trebuie să urmeze imediat după „strh”. Acesta descrie formatul datelor din flux, deci depinde de tipul acestuia. Pentru fluxuri video, informația regăsită în porția „strf” alcătuieşte o structură BITMAPINFO (vezi http://msdn.microsoft.com/en-us/library/ dd183375(VS.85).aspx), iar pentru fluxuri audio, o structură WAVEFORMATEX (http://msdn. microsoft.com/en-us/library/ms788112(VS.85).aspx).

Dacă porțiunea „strd” este prezentă, atunci ea trebuie să urmeze porțiunii „strf”. Formatul şi conținutul ei sunt definite de driver-ul codec. În general, driverele folosesc informația din această porțiune pentru configurări. Aplicațiile care redau, creează sau modifică fişiere AVI nu au nevoie să interpreteze informația. Ele doar o transferă la sau de la driver ca un simplu bloc de memorie.

Porțiunea opțională cu ID-ul „strn” conține un şir de caractere (cu caracterul nul la sfârşit) descriind fluxul.

Anteturile fluxurilor din lista „strl” sunt asociate cu fluxurile propriu-zise (datele audio sau video) din lista „movi” în funcție de ordinea în care apar listele „strl”. Astfel, prima lista „strl” se aplică fluxului 0, a doua fluxului 1, şi aşa mai departe.

Datele fluxurilor (lista „movi”)

După toate informațiile din antet, urmează lista „movi”, ce conține datele propriu-zise, altfel spus, cadrele pentru fluxurile video şi sample-urile pentru fluxurile audio. Porțiunile de date pot fi aşezate direct în conținutul listei „movi” sau grupate în liste „rec ” ce sunt aşezate, bineînțeles în interiorul lui „movi”. Listele-grup cu ID-ul „rec” sunt

folosite pentru a determina aplicația de redare (orice AVI Player) să citească datele din fişier dintr-o singură parcurgere. Acest lucru este util, spre exemplu, pentru vizionarea filmelor în format AVI de pe un CD.

În „rec" se află, aşadar, o mulțime de subporțiuni, care conțin date audio, sau vizuale. ID-ul fiecărei sub-porții este un FourCC format din doi octeți ce eprezintă index-ul fluxului de date de care aparține (se ține cont de ordinea în care apar listele „strl" în fişier) la care se adaugă două caractere ASCII ce definesc tipul de date din porțiune(deci în total 2+2=4 octeți).

Porția AVI Index („idx1")

Această porțiune este opțională, putând urma după lista „movi". Conține o listă cu ID-urile porțiunilor de date (cadrele, sample-urile audio), mărimea, locațiile lor în fişier. În listă sunt incluse, de asemenea, şi grupurile de porțiuni de date, porțiunile „rec ". Porțiunea „idx1" alcătuieşte o structură AVIOLDINDEX (http://msdn.microsoft.com/en-us/library/ms779634(VS.85).aspx). Dacă fişierul AVI conține această porțiune finală, atunci câmpul dwFlags din structura AVIMAINHEADER (adică porția „avih") trebuie să aibă setată caracteristica AVIF_HASINDEX.

Alte părți pot fi cele cu ID-ul „JUNK". Acestea sunt folosite pentru alinierea datelor în fişierul AVI.

O aplicație ar trebui să ignore conținutul acestei părți.

Digital Video Data (DV DATA) în fişierele AVI

În cele ce urmează va fi prezentată specificația Microsoft pentru stocarea digitală a secvențelor video. Conformându-se la această specificație, fişierele AVI sunt şi vor fi compatibile cu versiuni curente şi viitoare ale arhitecturii video pentru platforma Windows.

În cele ce urmează, se descrie structura AVI-urilor ce conțin DV data. Se definesc FourCC-uri specifice pentru fluxurile DV data şi obiecte de comandă (en. handlers) pentru compresia / decompresia fluxurilor. De asemenea, se defineşte structura unui flux de DV data şi se specifică cele două metode de stocare a DV data într-un fişier AVI.

2.1.5. Formate multimedia

Multimedia (MM) înseamnă abilitatea de a achiziționa, manipula, combina şi reda informații de la o mare varietate de medii, ce includ text, grafică, sunet, imagine fixă sau video. Multimedia nu este o tehnologie, ci mai degrabă un termen ce descrie un număr de tehnologii care lucrează împreună. Multimedia a transformat interacțiunea om-calculator şi a permis crearea unei noi familii de produse în diverse domenii precum:

- *acces la cunoaştere* – multimedia este modul cel mai rapid, eficient şi ieftin de a permite indivizilor accesul la informare, punându-le la dispoziție adevărate enciclopedii electronice;
- *administrarea documentelor şi a înregistrărilor* – datorată apariției unor documente tot mai complexe ca şi conținut care trebuiau gestionate de întreprinderi şi instituții comerciale;
- *educație şi instruire* – furnizând materialul potrivit pentru şcolarizarea şi instruirea unor categorii variate de subiecți;
- *reclame* – în mod practic nu există nici o limită în folosirea informației multimedia în astfel de aplicații;
- *controlul şi monitorizarea proceselor în timp real* – împreună cu bazele de date, prezentările multimedia de informații au un rol efectiv în operațiile de monitorizare şi control a sistemelor precum sistemele de transport, sistemele de supraveghere a pacienților, etc.

Această nouă generație de produse permite atât integrarea elementelor multimedia în mediile existente cât şi o nouă abordare în ceea ce priveşte procesul muncii. Aşadar, multimedia poate extinde aplicațiile existente, dar poate conduce şi la regândirea în mod revoluționar a procesării informației în diverse domenii cum ar fi economia, ştiința, arta, educația, ingineria. Folosirea multimedia generează beneficii pentru toți utilizatorii sistemelor informatice. Ea îmbunătățeşte calitatea şi cantitatea informației prezentate utilizatorului, precum şi interacțiunea om-calculator.

Organizarea documentelor multimedia

Organizarea eficientă a informației de tip multimedia cade în sarcina sistemelor de gestiune a bazelor de date multimedia (SGBDMM). Un astfel de sistem este un sistem de gestiune a bazelor de date performant care suportă tipuri de date multimedia pe lângă cele clasice, alfanumerice, şi care are capacitatea de a manipula cantități foarte mari de informație multimedia integrând strâns trei tehnologii fundamentale, şi anume:

- bazele de date;
- sistemele de regăsire a informației;
- sistemele de stocare ierarhică a informației.

Bazele de date multimedia pot fi definite ca sisteme de baze de date care pot stoca, manipula şi interoga informație prezentată sub formă de text, audio, video, obiecte grafice şi imagini statice alb-negru sau color.

Bazele de date multimedia sunt tot mai de actualitate în lumea de azi a calculatoarelor datorită faptului că oferă posibilitatea de a prelucra într-o manieră convenabilă şi flexibilă diferite tipuri de obiecte cu care interacționăm în viața noastră de zi cu zi. Ca atare, într-o bază de date multimedia putem întâlni pe lângă tipurile de date clasice dintr-o bază de date şi tipuri noi precum:

- *Date imagini* – acestea sunt foarte comune în bazele de date multimedia şi ele pot fi simple figuri, icoane, imagini medicale precum razele X, etc.;
- *Date video* – acestea sunt date similare fişierelor video (film) care sunt stocate în baza de date;
- *Date audio* – acestea sunt date asemănătoare celor stocate în fişierele audio şi de cele mai multe ori sunt de fapt melodii, voce, etc.;
- *Date document* – acestea sunt date de forma fişierelor text tradiționale în care informația este stocată sub formă de text, numai că mărimea acestor texte este foarte mare.

Obiectele multimedia sunt diferite de textul tradițional sau de documentele numerice în sensul că acestea necesită de regulă o mai mare cantitate de memorie internă şi externă pentru memorare. De asemenea, operațiile aplicate obiectelor multimedia sunt

diferite (de ex. afişarea unei figuri sau vizionarea unui video clip sunt diferite de afişarea unui paragraf de text).

Un sistem de gestiune a bazelor de date multimedia (SGBDMM) trebuie să ofere un mediu adecvat pentru utilizarea şi gestiunea datelor multimedia. În completarea funcțiilor tradiționale ale sistemelor de gestiune a bazelor de date, un sistem de gestiune a bazelor de date multimedia trebuie să includă funcții specifice precum:

- gestiunea unor date de tip multimedia precum imagini, video, grafică, audio, etc.;
- gestiunea unor volume mari de date multimedia;
- furnizarea unor scheme de gestiune a spațiilor de stocare performante şi rentabile ca şi cost;
- stocarea eficientă şi gestiunea livrării datelor multimedia;
- soluții eficiente pentru indexarea şi regăsirea datelor, folosind date multimedia drept criteriu de căutare;
- să suporte diferite formate pentru datele multimedia;
- să suporte diferite funcții specifice bazelor de date precum inserările, ştergerile, căutările şi modificările;
- să optimizeze realizarea interogărilor şi prelucrărilor.

Obiectele multimedia sunt, de regulă, obiecte binare mari BLOB (binary large objects). Este obişnuit ca un video clip să ocupe mai mult de 100 MB spațiu de stocare. Pe un sever video, este posibil ca sute de video clipuri să fie stocate. Datorită cantității uriaşe de spațiu de stocare necesar, un sistem de gestiune a bazelor de date multimedia necesită un mecanism sofisticat de gestiune a spațiului de stocare, care de asemenea trebuie să aibă un preț rentabil. Schemele de gestiune a spațiului de stocare trebuie să suporte, de asemenea, şi operațiile fundamentale necesare la nivelul bazei de date.

Adesea, un SGBDMM trebuie să ia în considerare şi următoarele aspecte:

- compunerea şi descompunerea obiectelor multimedia;
- operații ale obiectelor multimedia care implică sincronizarea acestora;

- persistența obiectelor;
- regăsirea informațiilor multimedia pe baza conținutului;
- acces concurent şi mecanisme de blocare pentru procesarea distribuită;
- securitatea datelor;
- consistența datelor, integritatea referențială şi refacerea datelor în caz de accident;
- tranzacții lungi şi tranzacții imbricate;
- indexarea şi clusterarea datelor.

Spre deosebire de SGBD-urile tradiționale, în SGBDMM replicarea datelor nu este încurajată din cauza volumului mare de date vehiculat.
Pentru aplicațiile multimedia simple, utilizarea modelului client-server pentru accesarea bazei de date multimedia, este considerată adecvată.

Aplicațiile multimedia complexe pot impune existența unui server video şi utilizarea unui SGBDMM cu o arhitectură dinamică.

Arhitectura unui SGBDMM

O bază de date multimedia are arhitectura structurată pe trei nivele:

- nivelul extern al interfeței cu utilizatorii;
- nivelul conceptual de compunere a obiectelor;
- nivelul intern al mediului de stocare a datelor.

Sarcinile specifice nivelului extern al interfeței includ în principal afişarea obiectelor, procesarea interogărilor şi compunerea şi descompunerea obiectelor. Afişarea obiectelor permite utilizatorilor să găsească entități multimedia existente care apoi pot fi refolosite. Prin intermediul interogărilor, fie bazate pe text fie pe vizualizare, utilizatorii specifică un număr de condiții pentru entitățile multimedia dorite şi obțin o listă de obiecte posibile a satisface aceste condiții. Obiectele potrivite sunt apoi reutilizate. Resursele multimedia, spre deosebire de cele de tip text sau numerice, nu pot fi localizate efectiv utilizând un limbaj de interogare bazat doar pe text. Chiar şi folosind limbajul

natural pentru formularea interogării este greu de regăsit cu precizie o fotografie sau o secvență video cu un anumit conținut. Cercetările privind regăsirea informațiilor pe baza conținutului se concentrează asupra mecanismelor care permit utilizatorilor să găsească efectiv obiecte multimedia reutilizabile, incluzând poze, sunete, video şi alte forme. După găsirea cu succes a obiectelor dorite, interfața bazei de date trebuie să-l ajute pe utilizator să compună/descompună documentele multimedia.

Cel de-al doilea nivel lucrează în legătură cu nivelul interfeței pentru a gestiona obiectele. In mod tipic, compunerea obiectelor necesită o serie de legături, precum legăturile de asociere, legăturile de similaritate şi legăturile de moştenire pentru a specifica diferitele relații dintre obiecte. Aceste legături sunt specificate fie prin intermediul interfeței grafice utilizator a bazei de date, fie prin intermediul unor funcții ale API.

Ultimul nivel, nivelul intern al gestiunii mediului de stocare a datelor, include două aspecte legate de performanță: clusterarea şi indexarea datelor. Clusterarea se referă la organizarea fizică a informației multimedia pe hard disc sau pe alt mediu optic de stocare, astfel încât atunci când se încearcă regăsirea acestor date, sistemul să fie capabil să acceseze datele binare în mod eficient. De obicei, performanța regăsirii datelor trebuie să garanteze un anumit nivel de calitate a serviciului oferit (QoS) şi să realizeze sincronizarea obiectelor multimedia. Indexarea înseamnă că un mecanism de localizare mai rapid este esențial pentru a găsi adresa fizică a unui obiect multimedia. Câteodată, schemele implică date complexe sau structuri de fişiere.

Modele de date

Datele multimedia de tip continuu, precum video şi audio, implică folosirea conceptelor specifice precum fluxul datelor, compoziție temporală, eşalonare în timp şi sincronizare. Aceste concepte sunt diferite de cele utilizate în cazul modelelor convenționale de date şi în consecință modelele de date convenționale nu sunt compatibile cu sistemele de baze de date multimedia. O problemă cheie cu care se confruntă un SGBDMM este descrierea structurii mediilor dependente de timp într-o formă adecvată interogării, actualizării, regăsirii şi prezentării datelor.

În acest sens, printre primele realizări se găseşte *modelul VideoSTAR* care reprezintă un standard pentru baze de date video şi este proiectat ca suport pentru aplicații video care partajează şi reutilizează date video şi metadate. Metadatele sunt implementate ca o bază de date complexă. Sistemul conține un model principal de indexare valabil pentru multe tipuri de aplicații cu câteva entități predefinite precum: persoane, evenimente etc. Fiecare aplicație poate îmbunătăți puterea descriptivă a modelului de bază cu tipuri specifice de indecşi, idee folosită în multe alte modele.

Un alt model de date descris de Zhong prezintă o schemă generală pentru modelarea obiectelor video, care încorporează caracteristici vizuale de nivel scăzut şi grupări ierarhice. Modelul oferă o structură generală pentru extragerea obiectelor video, pentru indexare, împărțirea pe categorii şi prezentarea noilor segmente video şi algoritmi de urmărire a obiectelor pe baza culorii relevante şi caracteristicilor traiectoriei. Prin obiecte video, autorii se referă la obiectele de interes, incluzând cele mai importante regiuni ale imaginilor de nivel scăzut, obiectele în mişcare şi grupe de obiecte care satisfac constrângeri spațio-temporale. Aceste obiecte video pot fi extrase automat prin segmentarea obiectelor şi utilizarea unor mecanisme de urmărire a obiectelor. După aceea obiectele sunt stocare într-o bibliotecă digitală. Obiectele pot fi grupate şi fiecărui grup îi pot fi asociate adnotări semantice de nivel ridicat. Se propune un model ierarhic de reprezentare a obiectelor în care obiecte din niveluri diferite pot fi indexate, căutate şi grupate în concepte de nivel ridicat. Interogările bazate pe conținut pot doar regăsi segmente video şi pot seta punctele de intrare în video în funcție de caracteristicile spațio- temporale ale obiectelor video interogate şi nu se referă la regăsirea unităților video dedicate din întregul flux video.

Cercetările mai recente se concentrează pe sisteme informaționale multimedia, în care SGBDMM manipulează datele multimedia într-o manieră structurată şi care folosesc sisteme de recunoaştere a datelor multimedia.

Multe din modelele conceptuale ale datelor multimedia sunt implementate ca scheme de baze de date sau ca sisteme obiectuale sau relaționale şi folosesc serviciile standard ale SGBD- urilor precum independența datelor (abstractizarea datelor), neutralitatea aplicației (deschisă), controlarea accesului multiutilizator (control concurent), toleranța la erori (tranzacții, recuperare) şi controlul accesului.

Aplicațiile multimedia sunt, în general, distribuite folosind mai multe servere pentru a satisface cerințele de stocare şi prelucrare. Pentru gestiunea distribuției eficiente a fluxului datelor acestea pot fi gestionate de SGBD-uri sub forma BLOB-urilor (*Binary Large Object*) şi pentru gestiunea accesului la distanță, folosind diferite protocoale de comunicare, diverse tehnologii de acces la date: ODBC (*Open Database Connectivity*), ADO (*ActiveX Data Object*), OLEDB (*Object Linking and Embedding DataBase*), JDBC (*Java DataBase Connectivity*).

Proiectarea de modele pentru schimbul de date, adică descrierea datelor împreună cu descrierea fluxului datelor sau ca un flux de sine stătător este un punct important într-un sistem multimedia distribuit. Datele multimedia descriptive pot oferi, de exemplu, componente de rețea active cu date utile în administrarea şi îmbunătățirea politicilor de scalare a mediilor şi în construirea de buffere cu conținut. Datele descriptive pot asista clientul în selectarea modului de livrare a datelor multimedia, de exemplu vizualizare imediată sau nu a datelor.

Standardul SQL/MM şi MPEG-7

Organizația internațională *International Organization for Standardization and International Electrotechnical Commission* (ISO-IEC), grupul de lucru SQL a dezvoltat o extensie a limbajului SQL numită SQL/MM (Multimedia).

Pe de altă parte, standardul MPEG-7 conține o modalitate de descriere a conținutului diferitelor date multimedia. Elementele MPEG-7 sunt codificate ca documente XML pe baza regulilor limbajului pentru descrierea definițiilor datelor (DDL). DDL oferă mijloacele pentru definirea structurii, conținutului şi semanticii documentelor XML. Atât MPEG-7 cât şi SQL/MM introduc câte un model conceptual pentru datele multimedia, în vederea utilizării în sistemele de baze de date multimedia.

De exemplu, definirea tipului folosit pentru imagini statice în cele două limbaje se face cu ajutorul lui *SI_StillImageType* în SQL/MM şi a tipului *StillRegionType* în MPEG-7.

Pe lângă definirea unei imagini sunt propuse tipuri pentru definirea caracteristicilor folosite la compararea imaginilor. De exemplu, *SI_FeatureList* oferă o listă a tuturor caracteristicilor de nivel scăzut disponibile, listă care conține în principal

mijloace de indexare a culorii şi texturii. Pentru culoare este propusă o valoare-histogramă (*SI_ColorHistogram*) şi două caracteristici nonhistogramă: o valoare medie (*SI_AverageColor*) şi un vector al culorilor dominante (*SI_PositionalColor*). Caracteristica textură (*SI_Texture*) conține valori reprezentând caracteristicile texturii imaginii.

Culoarea medie a imaginii e definită ca media artimetică a valorilor culorii tuturor pixelilor imaginii şi este similară - dar nu e identică - cu culoarea predominantă a imaginii. Histograma culorii unei imagini este specificată ca distribuția culorilor unei imagini măsurată de-a lungul spectrului celor N culori (64 sau 128 sau 256). Textura unei imagini e măsurată folosind trei factori: asprimea, contrastul şi direcționalitatea. Asprimea indică mărimea elementelor repetitive ale imaginii. Contrastul este variația luminii față de întuneric. Direcționalitatea indică existența sau nu a unei direcții predominante. Textura este folosită pentru căutarea unei imagini cu un format anume.
Comparativ cu SQL/MM, modelul conceptual al MPEG-7 este mai bogat. La nivelul secvenței DDL ce defineşte o imagine statică sunt înglobate atât elementele descriptive ale unei imagini statice, cât şi descrierile semantice şi elementele de descompunere spațială ale imaginii, care nu sunt acoperite de SQL/MM.

În descrierea altor tipuri de date multimedia se păstrează aceeaşi abordare.

În concluzie, SQL/MM acoperă partea sintactică a descrierilor multimedia, dar nu are nici o modalitate de a descompune o imagine şi nu există nici o soluție pentru utilizarea descrierilor semantice.

Majoritatea modelelor pentru datele multimedia utilizate de SGBDMM folosesc variante ale SQL/MM. De exemplu, Oracle 10g foloseşte în componenta *InterMedia* tipul *ORDImage* care are elemente similare cu tipul *SI_StillImage* prezentat, îmbunătățit cu informații detaliate despre formatul şi compresia imaginii.
SQL/MM se integrează bine cu SGBD şi de aceea oferă soluții pentru indicarea drepturilor de acces, soluții de recuperare a datelor şi este operațional, în sensul că metodele pentru regăsire şi prelucrarea imaginilor sunt asociate cu ierarhia tipurilor de date.

Pentru ca MPEG-7 să ofere funcționalitatea SQL/MM, acesta trebuie folosit împreună cu un SGBD care să stocheze şi să permită indexarea documentelor MPEG-

7. Dezavantajul major al variantei MPEG-7 este că nu oferă mecanisme de regăsire şi prelucrare proprii.

Cu toate acestea, formatul datelor folosit de SQL/MM nu permite interschimbul datelor multimedia şi a metadatelor cu alte sisteme software.

Produse pentru gestiunea bazelor de date multimedia

SGBD-urile relaționale şi cele obiectuale sunt, în acest moment, cele mai promițătoare platforme tehnice pentru implementarea unui model de date multimedia, pentru prelucrarea şi optimizarea interogărilor, deoarece oferă următoarele facilități:

- permit definirea de noi tipuri de date plecând de la cele existente şi definirea de funcții şi metode de acces la datele asociate;
- oferă soluții extensibile de indexare, prelucrare şi optimizare;
- gestionează datele multimedia extern şi intern. Gestiunea externă a datelor multimedia presupune stocarea acestora ca fişiere de sine stătătoare, iar în baza de date sunt păstrate doar numele acestor fişiere. Gestiunea internă presupune stocarea datelor în baza de date sub forma obiectelor distincte.

SGBDMM se bazează, în principal, pe serviciile sistemului de operare pentru stocarea şi regăsirea fişierelor.

Primul val al SGBDMM a fost, la mijlocul anilor ’90, reprezentat de *MediaDB* (numit acum *MediaWay*), JASMINE şi ITASCA.

Aceste sisteme puteau manipula diverse tipuri de date şi ofereau mecanisme pentru inserarea, interogarea, regăsirea şi actualizarea datelor. Majoritatea sistemelor au dispărut de pe piață după câțiva ani.

Al doilea val de SGBDMM comerciale manipulează conținutul multimedia prin intermediul tipurilor de obiecte complexe definite pentru diferite tipuri de medii. Cele mai avansate soluții sunt oferite de Oracle 10g, IBM DB2 şi IBM Informix.

ORACLE interMedia

*Inter*Media este o componentă care extinde funcționalitățile sistemului de gestiune a bazelor de date Oracle permițând stocarea, gestiunea şi regăsirea datelor

multimedia: a imaginilor, a secvenţelor video, a datelor audio şi a altor tipuri media eterogene, într-o manieră integrată cu tipuri de date tradiţionale.
Oracle *inter*Media permite accese SQL standard utilizând servicii, operatori si metadate pentru gestiunea tipurilor de date multimedia.

Oracle *inter*Media nu controlează dispozitivele de captură multimedia şi nu are funcţii pentru redarea datelor multimedia ci facilitează gestiunea datele multimedia stocate în baza de date.

Conţinutul multimedia poate implica rapid cantităţi uriaşe de informaţie. Aplicaţii precum publicarea pe Web, e-comerţ, gestiunea activelor media au condus la creşterea dramatică a generării şi consumării de date de tip media precum poze, sunete, muzică, voce, video, etc. Pentru prima dată, securitatea, controlul administrativ, performanţa, scalabilitatea şi accesul gestionat profesional la sistemele de informaţii ale întreprinderilor este disponibil şi pentru conţinutul multimedia utilizat în Web site-uri corporative sau în aplicaţiile multimedia.

Oracle *inter*Media utilizează obiecte de tip dată, similare cu cele din clasele JAVA sau C++ pentru a descrie date de tip imagine, audio, video. *Inter*Media permite stocarea, gestiunea şi regăsirea datelor multimedia provenind din diferite surse de date.
Astfel, *inter*Media gestionează datele multimedia stocate în baza de date sub forma BLOB-urilor (*binary large objects*) caz în care aceste date se află sub control tranzacţional, sau a fişierelor multimedia gestionate direct de sistemul de operare, formatul utilizat în acest caz fiind BFILE (*file-based large objects*) şi a datelor multimedia stocate pe un server web. In aceste ultime două cazuri, datele multimedia sunt stocate în afara bazei de date fără a beneficia de contolul tranzacţional al acestora. In acest caz, un pointer către aceste date este stocat în baza de date sub control tranzacţional, iar datele multimedia sunt stocate în fişierul extern de tip BFILE. Datele multimedia stocate în afara bazei de date pot furniza un mecanism adecvat pentru gestiunea unor depozite media de mărimi foarte mari, pre-existente sau noi, care sunt stocate sub forma unor fişiere obişnuite pe medii de memorare read-only sau reincriptibile. Aceste date pot fi importate ulterior în BLOB-uri în orice moment pentru a le putea supune controlului tranzacţional.

Metadatele, atributele şi metodele obiectelor sunt întotdeauna stocate în baza de date sub controlul lui Oracle *inter*Media. Dacă datele multimedia sunt stocate în

interiorul sau în exteriorul bazei de date, *inter*Media gestionează metadatele acestora pentru toate tipurile media şi automat le extrage din acestea. Aceasta face ca dezvoltatorii de aplicații să fie scutiți de povara cunoaşterii caracteristicilor formatelor datelor multimedia.

*Inter*Media foloseşte tipuri obiectuale pentru descrierea datelor multimedia, astfel:

- tipul *ORDAudio* pentru datele audio,
- *ORDImage* pentru imagini statice,
- *ORDVideo* pentru secvențe video şi
- *ORDDoc* pentru date eterogene.

Toate aceste tipuri conțin şi informații despre sursa datelor, tipul de dată folosit fiind *ORDSource.*

Instanțele acestor tipuri conțin atribute, inclusiv metadate, metode şi datele propriu-zise multimedia. Metadatele sunt foloste pentru stocarea informațiilor despre date, ca de exemplu mărimea fişierului multimedia, formatul de compresie, etc.

Definițiile tipurilor multimedia se găsesc în schema *ORDSYS.*

La inserarea datelor multimedia în baza de date Oracle sub controlul *inter*Media, indiferent de modalitatea de stocare a datelor multimedia, în interiorul sau exteriorul bazei de date, sunt automat extrase metadate pentru datele audio, imaginile statice şi secvențele video. Metadatele extrase automat sunt:

- informații legate de stocarea datelor: locația şi numele sursei datelor şi locul unde sunt stocate datele, în baza de date sau sub forma unui fişier extern;
- data şi ora curentă;
- formatul datelor;
- tipul în format MIME al datelor;
- pentru datele audio: formatul de codare, numărul canalelor audio, rata de eşantionare, tipul compresiei, durata secvenței;
- pentru imaginile statice: formatul imaginii, dimensiunile unui cadru, formatul de compresie al imaginii;

- pentru datele video: dimensiunile unui cadru, rezoluția unui cadru, rata de derulare a secvenței, durata secvenței, numărul total al cadrelor, formatul compresiei.

Funcționalitatea cheie într-o bază de date multimedia este modalitatea de regăsire eficientă a datelor multimedia continue şi non-continue. O metodă general folosită pentru regăsirea bazată pe conținut a obiectelor multimedia se bazează pe simpla extracție a proprietăților obiectelor multimedia. În recunoaşterea bazată pe conținut, se adaugă interpretarea semantică a obiectelor. Interpretarea semantică poate fi adăugată la indexare sau poate fi obținută printr-un proces de indexare semiautomat.

*Inter*Media suportă imagini digitale, statice, în două dimensiuni stocate ca şi reprezentări binare ale unor obiecte sau scene ale lumii reale, utilizând cele mai populare formate de fişiere şi scheme de compresie. Utilizatorii pot uşor să stocheze imagini în baza de date, fără să fie experți în caracteristicile diferitelor formate de fişiere pentru imagini, deoarece nu este necesar să se facă explicit conversii ale fişierelor la/de la un format intern folosit de Oracle pentru memorarea acestora. Oracle *inter*Media suportă următoarele tipuri de fişiere pentru imagini: TIFF, JFIF (JPG), BMP, TARGA, EXIF, PCX, PICT, GIF, CALS, SUN RASTER, FPIX, PNGF, PPMF, PGMF, PBMF, WBMP, RPIX. Utilizatorii sunt scutiți de cunoşterea complexității tehnologiilor de compresie/decompresie, ei trebuind doar să solicite ca imaginile să fie comprimate/decomprimate. Oracle *inter*Media suportă cele mai populare şi eficiente scheme de compresie, incluzând CCITT G3/G4, ISO/CCITT JPEG, etc.

Recunoaşterea imaginilor bazată pe conținut este o problemă importantă asociată sistemelor de gestiune a bazelor de date. Odată cu creşterea volumului colecțiilor imaginilor digitale care pot fi eventual stocate în baza de date, creşte şi dificultatea regăsirii imaginilor relevante. Pentru rezolvarea acestei probleme există două metode şi ambele utilizează metadate pentru regăsirea imaginilor:

- folosind informații introduse manual în tabele, ca de exemplu: titluri, cuvinte cheie descriptive preluate dintr-un vocabular limitat şi scheme de clasificare predefinite,
- folosind caracteristicile imaginilor, caracteristici extrase automat şi recunoaşterea obiectelor pentru clasificarea conținutul imaginii.

InterMedia permite combinarea celor două alternative prin proiectarea unei tabele ce conține imagini. Pentru aceasta se folosesc date de tip text pentru descrierea semnificației semantice a imaginii şi tipul de dată *ORDImageSignature* pentru interogări bazate pe conținut ce folosesc atributele esențiale ale imaginii.

InterMedia este o extensie a SGBD Oracle introdusă începând cu versiunea 9i şi oferă facilități pentru stocarea imaginilor, funcționalități de recunoaştere a imaginilor şi facilități de conversie a formatului, prin introducerea unui nou tip de obiect *ORDImage* şi a metodelor şi funcțiilor asociate. Recunoaşterea formelor este realizată prin posibilitatea de extragere a unui vector de caracteristici ale imaginii din diferite atribute vizuale.

Un sistem de recunoaştere bazat pe conținut prelucrează datele din imagine şi creează o abstractizare a conținutului pentru atributele vizuale. Interogările lucrează cu abstractizarea imaginii şi nu cu imaginea propriu-zisă.

Criteriile de căutare folosite de *Inter*Media sunt culoarea, textura şi conturul şi poziția. Pozițiile acestor atribute vizuale în cadrul imaginii sunt reprezentate prin coordonate. Aceste coordonate nu sunt folosite în mod independent pentru recunoaşterea formelor ci doar împreună cu unul din cele trei atribute vizuale.

Imaginea odată inserată în baza de date este analizată şi este stocată câte o reprezentare compactă a conținutului sub forma unui vector de caracteristici numit semnătura imaginii. Semnătura imaginii este extrasă prin segmentarea acesteia în regiuni, pe baza petelor de culoare care compun imaginea. Fiecare regiune are asociate informații despre culoare, textură şi contur. Semnătura conține aceste informații pentru fiecare regiune care formează imaginea şi informații despre culoare, textură şi contur pentru reprezentarea acestor atribute în întreaga imagine.

Atributul *culoare* memorează informații despre distribuția culorilor în întreaga imagine. Această distribuție conține date despre intensitatea fiecărei culori.

Atributul *textură* reprezintă şabloanele din cadrul imaginii, precum granularitatea şi netezimea. Spre deosebire de atributul contur, textura este foarte sensibilă la caracteristicile care apar cu mare frecvență în imagine.

Conturul este determinat de tehnicile bazate pe segmentare. Conturul este caracteristica unei regiuni de culoare uniformă.

Locația reprezintă poziția componentelor culoare, textură şi contur. Recunoaşterea imaginilor stocate într-o bază de date se face prin compararea lor cu o imagine model care poate fi o imagine stocată în baza de date, din afara bazei de date sau o imagine vectorială.

În procesul de căutare, se atribuie o *pondere* fiecărui atribut vizual în funcție de importanța lui. Valoarea fiecărei ponderi reflectă cât de sensibil trebuie să fie procesul de căutare față de un anumit atribut. Valorile ponderilor trebuie să fie între 0 (atribut nesemnificativ) şi 1 (atribut extrem de important în procesul de căutare).
Asemănarea dintre două imagini pentru fiecare atribut vizual este calculată ca *scorul* sau *distanța* dintre imagini, respectând atributul. Scorul ia valori în intervalul 0 (nu există diferență) – 100 (diferență maxim posibilă). Scorul întregii imagini se calculează ca sumă a scorurilor atributelor ponderată cu importanța fiecărui atribut.

În procesul de căutare se foloseşte o valoare – *prag de semnificație*. Dacă suma ponderată este mai mică sau egală cu valoarea pragului atunci imaginile se potrivesc, altfel nu.

Pentru creşterea vitezei de căutare în bazele de date de mari dimensiuni care conțin date multimedia este utilă crearea şi utilizarea unui index folosit pentru căutarea printre semnăturile imaginilor. Pentru aceasta se foloseşte un *index de domeniu* sau *index extensibil* deoarece acesta suportă obiecte complexe. Baza de date Oracle şi *inter*Media cooperează pentru definirea, construirea şi întreținerea unui index pentru datele de tip imagine, index de tip *ORDImageIndex*. Odată creat, indexul este automat actualizat ori de câte ori imaginile sunt inserate, modificate sau şterse din baza de date. Datele indexului sunt stocate în două *tablespace*-uri care trebuie create în prealabil: unul care conține datele indexului curent şi celălalt este un index intern creat pe aceste date.

Recunoaşterea formelor este o procedură complexă. Algoritmul de recunoaştere a formelor implementat în Oracle poate fi folosit numai în anumite condiții, şi anume:

- dacă obiectul sau obiectele căutat(e) ocupă o parte semnificativă a imaginii;
- dacă nu există elemente irelevante suprapuse peste o parte a obiectului căutat;
- dacă obiectul căutat se află în aceeaşi parte a imaginii;
- dacă dimensiunile relative ale obiectului în cele două imagini, imaginea de referință şi cea în care se face căutarea sunt apropiate;

- dacă obiectul căutat este fotografiat din acelaşi unghi în ambele imagini;
- dacă obiectele adiacente din imagine au culori distincte;
- dacă imaginea este formată doar din câteva forme simple.

Pentru a îndeplini aceste condiţii, se pot decupa succesiv zone din imagine, zone în care se realizează căutarea şi se pot utiliza diferite combinaţii de ponderi ale atributelor folosite în procesul de căutare.

Structura pentru indexare multidimensională *ORDImageIndex* permite accelerarea accesului la vectorul de caracteristici stocate.

De asemenea, *inter*Media poate gestiona date audio având următoarele formate: AIFF, AIFF-C, AUFF, WAV, MPEG I, MPEG II, MPEG III şi formate audio Real Networks (prin intermediul serverului de streaming Real Networks) sau date video cu formatele: QuickTime, AVI, MPEG şi formate video Real Networks (prin intermediul serverului de streaming Real Networks). El extrage în mod automat informaţiile de tip metadate din aceste formate şi le memorează în atributele obiectelor audio sau video din Oracle *inter*Media. Datele audio sau video propriu-zise pot fi stocate fie local în baza de date Oracle sau pot fi referite din surse externe menţionate anterior. Sunt recunoscute schemele de compresie ADPCM şi MU-LAW pentru datele audio şi AVI Indeo pentru datele video.

Oracle *inter*Media combinat cu Oracle şi cu unelte de dezvoltare a aplicaţiilor furnizate de parteneri ai firmei Oracle, constituie o platformă puternică pentru dezvoltarea şi întreţinerea aplicaţiilor Web de la nivelul întreprinderilor.

IBM DB2 Universal Database Extenders

DB2 UDB, serverul de baze de date al firmei IBM, dispune şi el de tipuri de date noi care permit stocarea şi gestionarea datelor multimedia. Pe lângă tipurile de date, aşa numite clasice: numerice, şir de caractere, date calendaristice, etc., există şi tipuri de date care permit stocarea şi gestionarea datelor de tip audio, video, voce, figuri, etc. Acestea pot fi memorate în baza de date în câmpuri de date de tip BLOB (*binary-large-object*), şiruri binare mari de lungime variabilă care pot păstra date nestructurate cum sunt figurile, vocea, imaginile video, etc. De asemenea, există şi tipuri de date care pot stoca date de tip şir de caractere de lungimi foarte mari numite CLOB (*character large object*

string) care pot stoca documente text, sau tipuri de date care permit stocarea unor şiruri grafice de lungime fixă sau variabilă care pot fi chiar de dimensiuni foarte mari precum GRAPHIC, VARGRAPHIC, LONG VARGRAPHIC, DBCLOB.

DB2 UDB este îmbogăţit şi cu componente numite *Extenders* pentru a permite gestionarea unor tipuri noi şi complexe de date precum: imaginile, secvenţele video, secvenţele audio, obiectele spaţiale, documentele text, etc.

IBM DB2 Universal Database Extenders extinde gestiunea datelor, incluzând suport pentru gestiunea imaginilor, a secvenţelor video, audio şi a obiectelor spaţiale. Toate aceste tipuri de date sunt modelate, accesate şi manipulate prin intermediul unui suport comun. Extensiile multimedia permit importul şi exportul obiectelor multimedia şi a atributelor acestora în interiorul şi în afara bazei de date, controlând accesul la tipurile de date neconvenţionale, cu acelaşi nivel de protecţie ca în cazul datelor tradiţionale şi navigând prin sau extrăgând obiectele găsite din baza de date.

De exemplu, *DB2 Image Extender* defineşte un tip de dată distinct, *DB2IMAGE* ce are asociate funcţii definite de utilizator pentru stocarea şi manipularea fişierelor de tip imagine. Conţinutul fişierului imagine pe care-l descrie *DB2Image* poate fi stocat ca *BLOB* sau în afara bazei de date, în sistemul de fişiere.

DB2 Image Extender oferă funcţionalităţi de căutare similare bazate pe tehnologia QBIC (*Query by Image Content*) pentru imaginile stocate în tipul *DB2IMAGE*. Tehnologia QBIC oferă posibilitatea de interogare sau căutare a imaginilor pe baza conţinutului lor.

DB2 NetSearch Extender oferă utilizatorilor şi dezvoltatorilor de aplicaţii o metodă inteligentă de căutare a documentelor text care sunt stocate în baza de date sau în fişiere sistem externe. De asemenea, el permite gestiunea documentelor structurate de tip XML, HTML, GPP.

DB2 Spatial Extender permite memorarea şi gestiunea datelor spaţiale ale unor caracteristici geografice cum ar fi: obiecte geografice (de tipul râurilor, pădurilor, munţilor, etc.), o zonă bine definită în spaţiu (de ex. o zonă de siguranţă aflată în jurul unui obiectiv, o zonă în care operează un anumit business), un eveniment care apare la o locaţie ce poate fi definită (de ex. un accident auto care se produce într-o anumită intersecţie). *DB2 Spatial Extenders* permite gestionarea acestor date spaţiale asemănător celor tradiţionale şi folosirea lor în cadrul diferitelor tipuri de aplicaţii.

IBM Informix DataBlades

IBM Informix DataBlades este o extensie a serverului de baze de date Informix al firmei IBM care aduce puterea Internetului direct în baza de date. Astfel el permite:

- extinderea funcționalităților serverului dinamic IBM Informix;
- facilități specifice pentru gestiunea conținutului;
- facilități pentru customizare;
- facilități pentru gestiunea de finețe a conținutului multimedia.

Modulele lui *IBM Informix DataBlades* sunt extensii ale serverului care sunt integrate în interiorul motorului bazei de date, furnizând funcționalități noi pentru aplicații şi crescând performanța sistemului. Cu ajutorul acestor module imaginile, documentele structurate, imaginile video complexe, paginile de informații HTML pot fi memorate, indexate şi accesate în mod direct în baza de date, furnizând astfel un mai mare grad de performanță, securitate şi uşurință în utilizare.
Există mai multe module ale *IBM Informix DataBlades*:

- *Spatial DataBlade* – modul care permite gestiunea inteligentă a informațiilor complexe, geospațiale, împreună cu date tradiționale, păstrând mecanismele eficiente de la nivelul modelului relațional al datelor;
- *Geodetic DataBlade* – modul care suportă interogări bazate pe date spațiale globale şi pe date de timp fără limitările inerente în cazul proiecțiilor pe hărți şi care asigură o precizie ridicată în legătură cu localizarea globală;
- *TimeSeries DataBlade* – modul care furnizează suport sofisticat pentru gestiunea datelor temporale şi de tip serii-de-timp.
- *TimeSeries Real-Time Loader* – modul care este un loader de date care lucrează împreună cu modulele *TimeSeries DataBlade* şi cu *NAG (Numerical Algorithms Group) DataBlade* pentru a atinge o performanță şi o capacitate analitică superioară celei care este posibil să fie atinsă utilizând bazele de date relaționale tradiționale sau soft de sine stătător dedicat analizei de timp-real.

- *Web DataBlade* – modul care este o colecție de unelte, funcții şi exemple care uşurează dezvoltarea aplicațiilor "inteligente" interactive de baze de date în medii Web.

Un punct important, dar dificil de realizat, este formularea şi prelucrarea interogărilor complexe. Nici unul din sistemele menționate nu suportă căutări complexe, ca de exemplu imaginile care conțin persoane stând în fața unui autoturism de o anumită culoare. În acest caz, forma şi culoarea autoturismului nu mai pot fi folosite pentru stabilirea asemănărilor dintre imagini deoarece în acest caz, o parte a autoturismului nu este vizibilă.

2.2. Comparație între formate

Trebuie luat în considerare faptul că scopul final este o prezentare a documentelor unor utilizatori interesați astfel că ieşirile din acest sistem ar trebui sa fie documente ce să permită căutarea în textul acestora pe cat posibil fără a deteriora informația vizuală din documentele originale.

Acestea fiind spuse, devine destul de clar că documentele create trebuie să aibă următoarele caracteristici:

- trebuie sa conțină atât imagini cât şi text;
- să existe posibilitate de căutare;
- să fie citibile de către clienți web uzuali;
- să aibă conceptul de pagină;
- să fie de dimensiuni reduse;
- să fie suportate de către sistemele de recunoaştere - OCR existente (ca documente de ieşire).

Prin analiză criterială am stabilit importanța caracteristicilor şi am luat în considerare mai multe tipuri de documente existente inclusiv un tip de document proprietar care ar putea fi dezvoltat special pentru acest proiect.

În urma analizei s-a ajuns la concluzia că dezvoltarea unui format proprietar nu este cea mai bună soluție, documentul PDF părând a fi cea mai bună soluție (tabel 4):

Tabel 4. Comparație între formate

Caracteristică	Importanța caracter.	TIFF	DOC	PDF	Format proprietar
Image + text	3	0	1	1	1
Cautare	4	0	0.5	1	1
Compatibilitate cu clienții existenți	5	0.75	1	1	0
Paginabile	2	1	0.75	1	1
Dimensiune	1	0.75	0.6	0.8	1
Suportabilitate	6	1	1	1	0
Punctaj acumulat		**12.5**	**18.1**	**20.8**	**10**

În plus față de celelalte documente, formatul PDF permite aşezarea textului recunoscut sub imaginea inițială (text under the image) astfel că utilizatorul va vedea imaginile dar va căuta în text. Acest lucru asigura o calitate ridicată a vizualizării chiar şi în cazul în care recunoaşterea a fost de o calitate mai scăzută.

Documentele PDF permit şi înmagazinarea de metainformații în acestea. Mecanismele de protecție ale documentelor PDF nu sunt de neglijat fiind posibil să protejam un document la salvare, copiere, tiparire, copy-paste, etc.

Sistemul trebuie sa fie gândit astfel ca toate documentele intermediare să fie păstrate, în cazul unor erori nemaifiind necesară reluarea întregului proces (de la scanare la salvare PDF) ci doar părți ale acestuia.

3. Conținutul web

3.1. Identificarea formatelor de documente pentru conținutul web

PDF - Portable Document Format

Portable document format PDF (Format portabil de documente) – este un format de fişier creat de firma Adobe în 1993 pentru schimbul de documente. PDF este utilizat pentru reprezentarea documentelor într-un spațiu bidimensional într-o manieră independenta de aplicațiile software, hardware şi sistemele de operare.

Fiecare fişier PDF cuprinde o descriere completă despre un plan general pentru documentele cu două dimensiuni (sau cu trei dimensiuni) care cuprinde textul, fontul, imaginile şi un vector grafic 2D cu conținutul documentului. Standardul PDF a fost oficial publicat doar în anul 2008.

Fundamente tehnice

PDF combina 3 tehnologii:

- un subset a limbajului de programare PostScript pentru descrierea paginii, pentru generarea graficelor şi structurii de bază;
- un sistem pentru includerea / eliminarea fonturilor pentru a permite ca fonturile să există în acelaşi document cu datele;
- un sistem structurat de memorare care să permită ca aceste trei elemente precum şi orice conținut asociat să poată exista împreună într-un singur fişier, de asemenea să permită compresia datelor.

PostScript

Este un limbaj de descriere a paginii care rulează într-un interpretor pentru a genera imaginea. Acest proces necesită multe resurse. PDF este un format de fişier, nu un limbaj de programare (comenzile pentru controlul fluxului cum ar fi *if* şi *loop* sunt eliminate, în timp ce comenzile grafice cum ar fi *lineto* sunt păstrate). Adesea PDF este generat dintr-o sursă PostScript. Comenzile grafice care sunt generate de codul PostScript sunt colectate şi însemnate; orice fişier, grafic sau font la care se referă documentul sunt preluate şi memorate de asemenea în fişier. Apoi totul este comprimat într-un singur

fişier. Aşadar, toate informaţiile generate de PostScript (font, structura de bază, dimensiunile) rămân nemodificate.

Ca format de document PDF-ul are câteva avantaje comparativ cu PostScript:

- PDF conţin tokenizarea şi interpretarea rezultatelor codului sursă PostScript, pentru o corespondenţă directă între modificările item-urilor din descrierea pagini PDF şi modificările asupra modului de prezentare a paginii;
- PDF suportă transparenţă grafică în timp ce PostScript nu;
- PostScript este un limbaj de programare interpretiv cu o stare globală implicită, unele instrucţiuni care însoţesc descrierea unei pagini pot afecta modul de reprezentare a paginilor următoare. Aşadar, toate paginile de procesat în documentul PostScript trebuiesc procesate în ordine pentru a determina modul de reprezentare corect pentru pagina dată, în timp ce fiecare pagină din documentul PDF este neafectată de caracteristicile celorlalte pagini. Ca rezultat, view-erele PDF permit utilizatorului să sară rapid la paginile finale dintr-un document lung, în timp ce un viewer PostScript necesită procesarea secvenţială a tuturor paginilor înainte de a fi capabil să afişeze pagina cerută.

Structura fişierului

Fişierul PDF constă din primitive de *obiecte*, de 8 tipuri diferite:

- *Boolean* - valori booleene, reprezentând *adevărat* sau *fals*
- *Numbers* (numere)
- *Strings* (şiruri de caractere)
- *Names*
- *Arrays* – colecţii ordonate de obiecte
- *Dicţionare* – colecţii de obiecte indexate prin Nume
- *Streams* – de obicei conţin mari cantităţi de date
- *NULL* – obiectul mull

Obiectele pot fi fie *direct* (incluse într-un alt obiect) fie *indirect*. Obiectele indirecte sunt numărate printr-un *număr obiect* şi printr-un *număr generat*. O tabelă de indecşi numită *xref table* păstrează biţii de ofset, pentru fiecare obiect indirect, faţă de

începutul fişierului. Această organizare permite un acces aleator eficient la obiectele din fişier şi de asemenea permite ca mici modificări să poată fi efectuate fără rescrierea întregului fişier (actualizare incrementală). Începând cu versiune PDF 1.5, obiectele indirecte pot de asemenea să fie localizate într-un stream special cunoscut ca *object streams*. Această tehnică reduce dimensiunea fişierelor care au un număr mare de obiecte indirecte mici şi este de ajutor pentru Tagged PDF.

Există două plane generale pentru fişierele PDF – neliniar (ne optimizat) şi liniar (optimizat). Fişierele PDF neliniare consumă mai puțin spațiu pe disc decât partea de numărare liniară, aşadar sunt mai lenți din punct de vedere a accesului deoarece porțiuni din datele necesare pentru a asambla pagina documentului sunt împrăştiate în tot documentul. Fişierele PDF liniare (de asemenea mai sunt numite „optimizate" sau „optimizare web") sunt construite într-o manieră care asigură ca ele să fie citite într-un browser web (utilizând un plug-in), ele sunt scrise pe disc într-un mod liniar (cum ar fi în ordinea paginii). Fişierele PDF pot fi optimizate utilizând software Adobe Acrobat sau pdfopt care face parte din Ghostscript.

Modelul imagini

Idea de bază despre reprezentarea graficelor în PDF este foarte asemănătoare cu cea din PostScript, excepție este doar utilizarea unui modul de transparență care a fost adăugată în PDF 1.4

Graficele PDF utilizează un sistem de coordonate carteziene independente de dispozitiv pentru a descrie suprafața unei pagini. O descriere a paginii PDF poate utiliza matrici pentru a scala sau roti elementele grafice. Conceptul cheie din PDF este acela că specificarea graficelor, care este o colecție de parametrii grafici care pot fi modificați, salvați şi restaurați de descrierea paginii. PDF a avut 24 proprietății grafice dintre care cele mai importante sunt:

- matricea transformării curente (CTM), care determină sistemul de coordonate;
- clipping path;
- spațiul culorilor;
- constanta alfa - care este o componentă cheie pentru transparență.

Graficele vectoriale

Graficele vectoriale în PDF, la fel ca în PostScript, sunt construite cu căi. Căile sunt de obicei compuse din linii şi curbe cubice Bezier, dar pot fi de asemenea construite din liniile externe din text.

Similar cu PostScript, PDF-ul nu permite o singură cale pentru a grupa corect curbele şi liniile. Căile pot fi o trăsătură, filtre sau utilizate pentru clipping. Trăsăturile şi câmpurile pot utiliza orice set din grafic-ul selectat, incluzând *patter-nuri.*

PDF-ul suportă câteva tipuri de patternuri. Cel mai simplu este *paternul tiling* în care fiecare parte din şirul de descriere este specificată să desenată repetat. Aceasta poate fi un pattern de culoare tliling, cu culoarea specificată în obiect, sau un patern tiling necolorat în care diferă specificarea culorii în timpul în care tarenul este desenat.

Începând cu PDF 1.3 există de asemenea un patern *shading*, care desenează continuu variind culorile. Există 7 tipuri de patternuri *shading* dintre care cele mai simple sunt *radial shade*(Type 2) şi *axial shade*(Type 3).

Imagini raster

Imaginile raster în PDF (numite imagini XObjects) sunt reprezentate prin dicționare cu stream-uri asociate. Dicționarul descrie proprietățile imaginii şi stream-ul conține imaginea. (De obicei, o imagine raster poate fi inclusă direct în descrierea paginii ca şi o imagine *inline*). Imaginile sunt de obicei *filtrate* în scopul comprimării. Filtrele de imagini suportate în PDF sunt următoarele:

- **ASCII85Decode** – un filtru vechi utilizat pentru a pune stream-urile pe 7-bit ASCII;
- **ASCIIHexDecode** – similar cu cel de sus dar mai puțin compact;
- **FlateDecode** – un filtru uzual bazat pe DEFLATE sau algoritmul Zip;
- **LZWDecode** – un filtru vechi bazat pe compresia LZW;
- **RunLengthDecode** – o metodă simplă de compresie pentru stream-uri cu date care se repetă utilizând algoritmul Run-length encoding.

Filtrele specifice pentru imagini sunt:

- **DCTDecode** – un filtru cu pierderi bazat pe standardul JPEG;

- **CCITTFaxDecode** – un filtru fără pierderi bazat pe metoda de compresie CCITT fax;
- **JBIG2Decode** – un filtru cu sau fără pierderi bazat pe standardul JBIG2 introdus în PDF 1.4;
- **JPXDecode** – un filtru cu sau fără pierderi bazat pe standardul JPEG2000 introdus în PDF1.5;

În mod normal toate imaginile conținute în PDF sunt incluse în fișier. Dar PDF permite ca fisierele cu maginii să fie memorate în afara fișierului prin utilizarea *streamu-urilor externe* sau *imagini alternante*. Subseturile standardizate ale PDF, PDF/A şi PDF/X interzic aceste tehnici.

Text

Textul în PDF este reprezentat prin *elemente text* în paginile care conțin stream-uri. Un element text specifică ce *caracter* trebuie desenat la poziția respectivă. Caracterele sunt specificate utilizând *codificare* în funcție de fontul selectat. Un obiect font în PDF este descris ca un set de una sau mai multe litere, în funcție de dimensiunea fontului. Fişiere cu fonturi care pot fi incluse se bazează foarte mult pe formatul standard al fonturilor digitale: Type 1, TrueType şi OpenType. În plus PDF suportă şi Type 3 varianta în care componentele din fişier sunt descrise de operatorii grafici din PDF.

În timp ce caracterele text sunt afişate utilizând codurile caracterelor (numere întregi) pentru a mapa heroglifele în fontul curent se utilizează o codificare. Există un număr de codificări, incluzând *WinAnsi, MaRoman* şi un număr mare de codificări pentru limbajele din Asia. Mecanismul de codificare din PDF este proievtat pentru fontul Type 1 şi regulile utilizează reguli complexe pentru aplicarea lui în fonturile TrueType.
Pentru fonturi mari sau fonturi cu heroglife nestandard, este utilizată codificarea specială *Identity-H* şi *Identity-V*. În această codificare fiecare font trebuie să furnizeze o tabelă *ToUnicode* de informații semantice despre caracterele care vor fi prezente.

Transparența

Modelul de imagine original prezentat în PDF a fost similar cu cel din PostScrip *opaque*: fiecare obiect este desenat pe pagină complet eliminând orice a fost desenat

înainte în aceeaşi locaţie. În PDF 1.4 modelul imaginii a fost extins pentru a permite transparenţa. Când este utilizată transparenţa, noile obiecte interacţionează cu obiectele deja marcate pentru a crea mai multe efecte.

Extensia transparenţei are la bază concepte cheie de *grupuri de transparenţă, modele amestecate, shap* şi *alpha* modelul este apropiat de trăsăturile modelului Adobe Illustrator din versiunea 9. modelul amestecat are la bază modelul utilizat de AdobePhotoshop în acel moment. Când a fost publicată specificaţia PDF 1.4 formularele pentru calcularea modelelor amestecate au fost ţinute secrete de Adobe. Fişierul PDF poate conţine elemente interactive cum ar fi adnotaţiile şi câmpurile formei.

Securitate şi semnătură

Fişierele PDF pot fi criptate pentru securitate, sau pot fi semnate digital pentru autentificare. Standardul de securitate furnizat de Adobe constă din două metode şi două parole diferite „user password" şi „owner password". Documentul PDF poate fi protejat de parolă pentru deschidere (user password) şi documentul poate fi de asemenea restricţionat pentru anumite operaţii: listare, copiere text şi grafic în afara documentului, modificarea documentului şi adăugarea sau modificarea notelor text. Oricum, toate operaţiile (excepţie deschiderea documentului prin parolă de protecţie, daca există) pot fi restricţionate prin parolă de „owner" sau „user", sunt triviale datorită existenţei softurilor „pdf crack".

Variante de PDF

O mare parte din subseturile PDF au fost sau sunt standardizate pentru diferite componete:

- PDF/X – pentru imprimare şi grafică ISO 15930
- PDF/A – pentru arhivare în medii ca întreprinderi / guvern]/ biblioteci]/etc ISO 19005
- PDF/E – pentru portabilitate şi motorul de desenare (utilizabil din ISO TC171)
- PDF/UA – pentru fişiere PDF universal accesibile

HTM, HTML- Hypertext Markup Language

Unul din primele elemente fundamentale ale WWW (World Wide Web) este **HTML** (Hypertext Markup Language); acesta descrie formatul primar în care documentele sînt distribuite şi văzute pe Web. Multe din trăsăturile lui, cum ar fi independența față de platformă, structurarea formatării şi legăturile hipertext, fac din el un foarte bun format pentru documentele Internet si Web. Redus la esență, HTML (HiperText Markup Language - HTML) este un set de coduri speciale care se inserează într-un text, pentru a adăuga informații despre formatare şi despre legături. HTML se bazează pe SGML (Standard Generalized Markup Language).

HTML a fost dezvoltat inițial de Tim Berners-Lee la CERN in 1989. Tim Berners-Lee a utilizat ca model SGML (un standard internațional în plină dezvoltare). SGML avea avantajul unei structurări avansate şi al independenței de platformă dar proiectarea lui a avut în vedere mai mult structura semantică a documentului decît modul de formatare. Flexibil, SGML putea fi caracterizat ca o specificare pentru descrierea altor formate. Standardul oficial HTML este World Wide Web Consortium (W3C), care este afiliat la Internet Engineering Task Force (IETF). W3C a enunțat cateva versiuni ale specificatiei HTML, printre care si HTML 2.0, HTML 3.0, HTMLl 3.2, HTML 4.0 şi, cel mai recent, HTML 4.01. In acelaşi timp, autorii de browsere, cum ar fi Netscape si Microsoft, au dezvoltat adesea în browserele lor. În unele cazuri, cum ar fi tagul Netscape, aceste extensii au devenit standarde *de facto* adoptate de autorii de browsere. Documentele HTML sînt documente în format ASCII şi prin urmare pot fi create cu orice editor de texte. Au fost însa dezvoltate editoare specializate care permit editarea într-un fel de WYSIWYG deşi nu se poate vorbi de WYSIWYG aîta vreme căt navigatoarele afişeaza acelaşi document oarecum diferit, în funcție de platforma pe care rulează. Au fost de asemenea dezvoltate convertoare care permit formatarea HTML a documentelor generate (şi formatate) cu alte editoare. Evident conversiile nu pot păstra decât parțial formatările anterioare deoarece limbajul HTML este înca incomplet.

Un fisier HTML este de fapt un fişier text. Este sursa paginii de Web. Acest fişier este scris în limbajul HTML sau HTM (ambele însemnând practic acelasi lucru) pe care, la afişare, browserul îl interpreteaza. Orice fisier începe cu eticheta <html> si se terminã cu eticheta </html>. Fişierul este compus din:

- **header**, inclus între etichetele <head> si </head> , conține informația introductivă de formatare a paginii
- **corp**, inclus între etichetele <body> si </body>

Structura generală a unui document HTML este următoarea:

<HTML>

<HEAD>

<TITLE> </TITLE>

</HEAD>

<BODY>

</BODY>

</HTML>

Între aceste etichete este inclus întregul document HTML. Ele comunică browserului unde începe şi unde se termină documentul.

<HEAD> </HEAD>

Între aceste etichete sunt incluse titlul, deja menționat, precum şi alte informații despre documentul HTML. Aceste elemente, numite *metatag*-uri sunt deosebit de importante pentru ca pagina să fie cât mai bine cotată de către motoarele de căutare. Ele vor face obiectul unui capitol separat.

Metatag-urile şi antetul în ansamblul său nu sunt vizibile pentru vizitator în pagina Web. Totuşi, la fel ca întreg codul HTML al paginii, antetul poate fi vizualizat selectând din meniul browserului opțiunea *View > Source.*

<TITLE> </TITLE>

Stabileşte **titlul** documentului HTML.

Titlul documentului este deosebit de important deoarece este unul din criteriile folosite de motoarele de căutare pentru indexarea paginii.

<BODY> </BODY>

Conține descrierea textului şi elementelor paginii. În corpul documentului se stabilesc, deci, aspectul şi conținutul paginii Web. Elementele conținute în această secțiune sunt vizibile în pagină.

Limbajul HTML se bazează pe *tag-uri* . Acestea sunt comenzile, sau instrucțiunile pe care un browser le interpretează, ca să poată afişa corect pagina. Exemplu de tag: **<body>** . Semnele „<" (mai mic) si „>" se numesc *paranteze unghiulare*. Majoritatea tagurilor se folosesc în perechi; fiecare pereche are un tag de deschidere, (sau de început), şi unul de închidere (de sfârsit). Tag-urile şi textul dintre doua taguri se numeşte *element HTML*, iar textul dintre tag-ul de început şi cel de sfârsit formeaza *coținutul elementului*. Exemplul de mai sus („<body>") este tag-ul de deschidere pentru elementul „body". Tag-urile de închidere au caracterul „/" în plus fata de cele de deschidere. Scopul celor doua taguri „<body>" si „</body>" este să specifice browserului că elementul dintre ele este corpul paginii.

Sintaxa oricarui tag este:

<TAG atribut1="val" atribut2="val">Text</TAG>

Regulile de folosire a etichetelor, atributelor şi valorilor acestora reprezintă **sintaxa** limbajului HTML. Similar limbajelor de programare, în HTML respectarea sintaxei limbajului este determinantă pentru o bună funcționare a documentului. Documentele HTML sunt interpretate de browser exact aşa cum au fost ele scrise. Prin urmare, orice greşeală de sintaxă se va reflecta direct în aspectul paginii Web, conducând, de cele mai multe ori, la o funcționare defectuoasă a acesteia. O pagina web poate conține:

- text
- imagini
- fişiere audio
- fişiere video

XHTML- Extensible Hypertext Markup Language

XHTML (eXtensible HyperText Markup Language) este un limbaj de marcare ce are aceleaşi capabilităţi expresive ca şi HTML, dar cu o sintaxă mai strictă. XHTML poate fi considerat ca încrucişarea dintre HTML şi XML, în multe privinţe fiind o reformulare a HTML în XML. La data de 26 ianuarie 2000, Consortiul Web a facut publică o nouă specificaţie-recomandare intitulata **XHTML**, reprezentând o familie, bazată pe XML , de tipuri de documente şi module care extinde arhicunoscutul standard HTML. XHTML 1.0 este primul "descendent" al acestei familii, fiind în fapt reformularea tipurilor de documente HTML4 in termenii meta-limbajului XML. Dezvoltatorii de pagini si de aplicatii Web vor migra astfel de la HTML 4, bazat pe relativ vechiul şi complexul SGML (Standard Generalized Markup Language), la XML (eXtensible Markup Language) - extensibil şi mai usor de utilizat, cu un viitor întrevazut a fi strălucit.

Arborele genealogic al XHTML

Avantajele unei aplicaţii XHTML sunt multiple, important este că pot fi citite de toate dispozitivele XML în timp ce păstreaza compatibilitatea cu toate browserele de Internet mai vechi sau mai noi, fără a necesita specificaţii suplimentare.

Un document XHTML se compune din trei părţi principale:

- declaraţia de conformitate sau DOCTYPE, defineşte tipul documentului creat. În cazul unui document XHTML, aceasta este <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">. Editorul HTML va şti să o introducă singur.
- secţiunea **<head>** conţine meta-marcajele necesare clasificării şi indexării corecte a paginilor pe internet. Urmează imediat după declaraţia de conformitate şi este

delimitat în interiorul marcajului <head></head>. Cele mai importante meta-marcaje conținute în antet sunt <title></title> - titlul paginii, <meta name="description" content="" /> - descrierea paginii şi <meta name="keywords" content="" /> - cuvintele-cheie care descriu cel mai bine conținutul paginii.

- secțiunea **<body>** , reprezintă pagina propriu-zisă, delimitată de marcajul <body></body>. Aici apar toate celelalte marcaje folosite.

Documentele XHTML pot fi etichetate atit "text/html" cât şi "text/xml", elementul rădacină fiind <html>. Declarea tipului de documente XHTML se va face prin intermediul construcției DOCTYPE, ca în SGML, existând trei tipuri de definții de documente conform specificației HTML 4: tipul *strict* , *tranziţional* si *pentru cadre (frames)*, ca în exemplul următor:

<!DOCTYPE html

PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"

"DTD/xhtml1-strict.dtd">

Structura de baza a unui document este următoarea (fig.1):

<!DOCTYPE ...>
<html>
<head>
<title>... </title>
</head>
<body> ... </body>
</html>

XHTML foloseşte *tag-uri*pentru a structura documentul. Deosebirile față de HTML (deosebiri care derivă din faptul că XHTML este bazat pe XML) sunt următoarele:

- Tag-urile (şi atributele lor) se scriu cu litere mici (XML este case-sensitive)
- Toate tag-urile trebuie închise
- Valorile atributelor trebuie închise în ghilimele

XHTML este un limbaj care foloseste markere şi are aceeaşi putere de expresie ca şi HTML dar mai are în plus un avantaj, acela de a respecta sintaxa XML. Pe cand HTML este o aplicație de tip SGML (Standard Generalized Markup Language), care este un

limbaj foarte flexibil, XHTML este o aplicație a XML care este un su bset mult mai restrictiv al SGML Documentele XHTML trebuie să fie bine formate, de aceea ele permit ca procesarea automata sa fie făcută cu unelte standard XML, spre deosebire de HTML care necesită un parser destul de complex, tolerant şi personalizat.

Scopul folosirii XHTML-ului este acela de a înlocui limbajul clasic HTML, de a schimba tehnologia folosită cu una mai performantă, realizând un pas important in dezvoltarea web-ului; în plus, este susținut de către organizația World Wide Web Consortium (W3C).

XML: Extensible Markup Language

XML (e**X**tensible **M**arkup **L**anguage), descendent al SGML (Standard Generalized Markup Language), este un meta-limbaj utilizat în activitatea de marcare structurală a documentelor, a cărui specificație a fost dezvoltată începand cu 1996 in cadrul Consortiului World Wide Web (W3C), de un grup de cercetare condus de Jon Bosak de la Sun Microsystems, la care au aderat ulterior o serie de grupuri de experți din comunitătile academice (Text Encoding Initiative, NCSA, James Clark) şi industriale (SUN, IBM, Netscape, Oracle, Adobe ,etc.).

Acest limbaj nu procesează în nici un fel datele sau informația. Menirea lui principală este doar de a formata şi structura informația. Principalul avantaj al XML este compatibilitatea. Informația structurată cu ajutorul XML va fi citită şi interpretată ulterior în acelasi fel, indiferent de dispozitiv, fie el mobil, palmtop, PC sau Mac. XML a fost elaborat pentru:

- separarea **sintaxei** de **semantică** pentru a furniza un cadru comun de structurare a informației;
- **construirea de limbaje de mark-up** pentru aplicații din orice domeniu;
- **structurarea informației** în viitor;
- **asigurarea independenței** de platformă şi suport pentru **internaționalizare.**

XML este un format de fişier asemănător cu Microsoft Word sau un fişierAdobe Acrobat, o foaie de calcul tabelar sau o pagină online HTML. Oricum, are proprietăți diferite de acestea şi anume:

• standardul XML nu este controlat de o singura organizație; este un `open' standard la care oricine poate contribui sau folosi;
• fisierle XML nu sunt salvate ca date `binare', ci mai degrabă ca plain text. Asta înseamnă că sunt indepente de platformă şi pot fi citite de oameni;
• fisierele XML nu au o structură specifică, însa folosesc un set de reguli de bază (dar stricte) . Astfel, se pot reprezenta prin fişiere XML multe tipuri de date şi informații, de la documente până la fişiere cu imagini şi tranzacții financiare;
• regulile XML pot fi folosite să restricționeze structura tipurilor de date - deci pot fi create noi standarde. Structura este de tipul self- descriptive şi fiecărei date îi este asociat un `tag' pentru a o descrie intr-un fel. Acest lucru permite ca fişierele să poată fi validate de un calculator , dar chiar şi de oameni pâna la un anumit nivel; sunt doar secvențe de text create pentru a structura, depozita şi transporta informație. Important este că definind reguli stricte dar neimpunând restricții asupra structurii, se pot dezvolta formate `standard XML' (numite Schemas) care să reprezinte un tip particular de date. Există "Standard XML schemas" pentru multe tipuri de informații, de la tranzacții de vânzări în business, până la formatarea ştirilor.

Documentele XML sunt realizate din unități de stocare numite entități, ce conțin date parsate sau neparsate. Datele parsate sunt realizate din caractere, unele dintre ele formând date caracter iar altele ca marcaje. Marcajele codifică o descriere a schemei de stocare a documentului şi structura logică. XML furnizeaza un mecanism pentru a impune constrângeri asupra schemei de stocare şi a structurii logice.

Cuvantul *marcaj (markup)* a fost folosit inițial pentru a descrie anumite adnotări, note marginale în cadrul unui text cu intenția de a indica tehnoredactorului cum trebuie listat un anumit pasaj. Generalizând, putem defini marcajul drept orice acțiune de a interpreta explicit o porțiune de text.

Un *marcaj (tag)* este un şir de caractere delimitat de caracterele "<" si ">". *Datele caracter* reprezintă conținutul marcajelor.

Coduri fisier ". XML"

Un document XML este un **arbore ordonat etichetat**:

- **date caracter** - noduri frunză ce conţin datele
- noduri **elemente** etichetate cu
    - un nume (adesea numit şi tipul elementului)
    - o mulţime de atribute, fiecare din ele având un nume si o valoare

XML-ul este un limbaj cu o sintaxa simplă şi care permite doar structurarea datelor într-o manieră proprie prin definirea propriilor taguri. Această facilitate de structurare a datelor a permis folosirea sa pentru a dezvolta limbaje noi precum si pentru a fi folosit in noi standarde de stocare a datelor. Nu in ultimul rând, XML-ul poate fi folosit pentru a schimba date intre aplicţiile care au nevoie de a comunica într-un limbaj comun.

ASP- Application Server Provider

Există câteva definiţii date termenului de **ASP (Application Service Provider -** furnizor de aplicaţii livrate ca servicii), printre care şi următoarele două, date organizaţii de specialitate:

• **Asociaţia Americană pentru IT** oferă următoarea definiţie:

"Un ASP este orice companie care distribuie şi întreţine automat aplicaţii şi servicii web pentru clienţi/utilizatori, direct prin Internet sau printr-o reţea privată."

• **Aberdeen**, o binecunoscută companie de analiză de marketing şi poziţionare pe piaţa IT, defineşte ASP-ul astfel:

"ASP-ul este un tip de afacere care oferă consumatorilor/clienţilor acces la anumite software-uri găzduite într-o locaţie administrată centralizat."

• Conform dicţionarului online **Webopedia** ASPurile sunt:

"...terţe părţi care întreţin şi distribuie clienţilor soluţii şi servicii de tip software dintr-o locaţie administrată centralizat."

Toate aceste definiții pun accent pe faptul că ASP-ul **distribuie şi întreține aplicații software prin Internet**. Cu alte cuvinte, un ASP "închiriază" aplicații software pe Internet clienților care nu pot sau nu doresc să investească în licențe pentru software, hardware, personal IT specializat şi alte resurse, oferindu-le acestora posibilitatea de a prelua controlul total asupra software-ului. Conform sitului www.ASPnews.com, ASPurile se împart în 4 subcategorii:

• ASPuri pentru corporații - oferă aplicații high-end pentru afaceri;

• ASPuri locale/regionale - oferă numeroase tipuri de aplicații pentru afaceri mai mici (locale sau regionale);

• ASPuri specializate - oferă aplicații pentru anumite nevoi, de genul serviciilor pentru website-uri sau de resurse umane;

• ASPuri pentru piețe verticale - programe utilitare pentru un anumit tip de industrie, cum ar fi cea imobiliară.

Un ASP trebuie să combine cunoştințele legate atât de servicii, networking, tehnologie şi management, cât şi de administrarea, integrarea, managementul şi suportul aplicațiilor. Din punct de vedere al utilizatorului, el cumpără un singur serviciu de la un singur distribuitor: ASP-ul. În realitate, există un sistem complex care susține ASP-ul şi care oferă diferite calificări în diferite domenii de tipul networking-ului sau al aplicațiilor. În marea majoritate a cazurilor, componentele şi serviciile ASP-urilor sunt distribuite printr-o serie de parteneriate între diferiți furnizori. Cel mai simplu mod de distribuție poate include un singur creator de programe, susținând propria aplicație, pe care o distribuie pe Internet sub termenii unui contract.

Cel mai la îndemână exemplu de serviciu ASP este webmailul. Chiar dacă mailurile pot fi accesate şi printr-un program specializat (ex: Outlook, Firebird, etc.), marea majoritate a utilizatorilor persoane fizice preferă să se folosească de serviciul web-based (Yahoo, Gmail, Hotmail, etc.). Avantajele unui astfel de serviciu: posibilitatea de a-l accesa de pe orice computer, lipsa unor cerinte software speciale, faptul ca nu necesită mentenanță. Aceste avantaje se mențin in cazul oricarui tip de serviciu ASP. Pentru acele

servicii profesionale oferite contra cost, un alt mare avantaj este partajarea costurilor de dezvoltare între toți utilizatorii sistemului. Serviciile web distribuite in regim ASP (application service provider) reprezintă una dintre laturile cele mai avansate ale Internetului. In esență, o astfel de aplicație oferă servicii online cu nivel de complexitate ridicat, servicii care se adresează unui target larg (ex: utilizatorii de mail, persoanele tinere, companii mici si mijlocii indiferent de domeniu) sau restrâns (avocati, agenții imobiliare, agenții turistice, contabili, etc.).

ASPurile folosesc ca platformă de distribuție a serviciilor un sistem bazat pe soluții web. Fig. 2 arată cum utilizatori aflați online se pot conecta prin Internet la serverele web ale ASP-urilor localizate în spatele unui firewall. Serverul de web accesează informațiile din baza de date printr-o retea private.

ASP: Active Server Pages

ASP (Active Server Pages) este una dintre metodele de construire dinamică a paginilor de Web. Începuturile WWW- ului au constat în livrarea de paginii HTML prin protocolul HTTP. Acestea constau în text formatat, imagini şi link- uri care permiteau navigarea de la o pagină la alta. În curând s-a simțit nevoia interacțiunii, posibilitătii de a prelua răspunsuri de la client, de a furniza pagini personalizate, de a afişa dare prelucrate din diverse surse care nu sunt statice. Astfel, a apărut partea dinamica a Web-ului, parte ce a dus la o dezvoltare substanțială a Internetului prin e-commerce, aplicații online de tot felul, de la simpla completare a unor formulare electronice până la site-uri care simulează burse, bănci. În prezent, prin pagini dinamice de web se pot crea aplicații complexe care înlocuiesc din ce în ce mai mult aplicațiile clasice client- server pentru că sunt simplu de întreținut, nu necesită instalare pe calculatoarele client şi pot fi accesate de oriunde. În acest cadru, ASP este una din opțiunile principale pentru crearea de pagini dinamice. ASP-urile se utilizeaza pentru:

- Editarea dinamică, schimbarea sau adăugarea de conținut la o pagină Web;

- Generarea de răspunsuri cererilor utilizatorilor sau informațiilor trimise prin formularele HTML;

- Accesarea oricărei date sau baze de date şi returnarea de rezultate către navigator;

- Personalizarea paginilor Web, făcându-le mai utile pentru utilizatorii particulari;

- Simplitatea şi viteza de execuție a ASP față de CGI şi Perl;

- Asigurarea securității, deoarece codul ASP nu este vizibil prin navigator;

- Posibilitatea de a fi afişate de orice navigator, fişierele ASP sunt returnate în format HTML;

- Minimizarea traficului prin rețea.

La dezvoltarea unui site care utilizeaza ASP-uri pot interacționa aproape toate tehnologiile Web dinamice existente. În fig.1 este prezentat modul în care se integreaza diferitele componente de sistem cu aplicațiile.

Caracteristica majoră a ASP constă în capacitatea sa de includere a scripturilor direct într-un fişier pe care-l accesează navigatorul, generând astfel *pagini dinamice*.

Dar, față de celelalte scripturi de pe server, fişierele ASP pot să conțină cod HTML sau XML, incluzând chiar scripturi pentru client şi referiri la componentele COM (Component Object Model) prin care realizează o serie de sarcini, cum ar fi conectarea la o bază de date.

- text;
- tag-uri HTML sau XML;
- comenzi ASP.

CGI-Computer generated imagery

Computer-generated imagery (CGI) **CGI (Imagini generate pe computer)** este aplicarea domeniului graficii pe calculator (mai exact 3D Computer Graphics) şi a efectelor seciale în filme, programe de televiziune, reclame, jocuri video. CGI a fost folosit prima dată în filme, în anul 1973. Progresele în sfera CGY sunt raportate în fiecare

an la Siggraph, o conferință anuală pe tema graficii pe calculator si tehnicilor interactive, care adună la Los Angeles mii de profesionişti din domeniu.

CGI: Common Gateway Interface

CGI (Common Gateway Interface) este un protocol standard de comunicare între documentele Web şi aplicațiile localizate pe serverul Web. Este metoda prin care un server web poate să prezinte vizitatorilor informația prin intermediul unui browser web, în funcție de cererile acestora, interactiv, accesând baze de date sau documentație, şi extrăgând datele cerute.

CGI (Common Gateway Interfaces) s-a impus ca cea mai eficientă, stabilă şi uşor de înțeles modalitate de manipulare a informației generate în mod dinamic pe Web. Este de fapt acea parte a server-ului Web care poate comunica cu alte programe care rulează pe sistem. Cu ajutorul acestei interfețe, serverul Web poate apela un program. Comunicarea prin Internet se bazeaza pe protocoale. Protocolul utilizat pentru serverele Web, pentru comunicațiile dintre servere si programele care le accesează (clienți)este **HTTP (HyperText Transfer Protocol).** Ca orice protocol pe Internet, el este în acelaşi timp un mecanism extrem de complex şi un dialog simplu bazat pe principiul cerere-răspuns: programul care cere (clientul) emite o cerere spre server (dupa o secvență inițială de conectare). Cererea pe WWW porneşte în general începând de la un document în format HTML şi poate să ia o mare varietate de forme: de la o cerere simplă, de exemplu aducerea unui alt document, pâna la completarea form-urilor HTML. Datele transferate de către protocol pot veni în formate diferite, de la text simplu pâna la imagine sau videoclip. Mai mult, este posibil ca datele ce compun un document să se afle fizic pe calculatoare diferite, putând fi accesate doar prin alte protocoale de comunicație, nu direct via HTTP. Comunicația se derulează între clientul care a inițiat cererea şi serverul Web accesat. Serverul are la dispozitie mecanismul specificat de standardul CGI. CGI specifică exact sub ce formă trebuie date mai departe informațiile sosite sub forma unei cereri de la client. Programele care recepționează cererile se numesc programe CGI: sunt practic programe activate pe server, indirect, de către utilizatori externi, datorita caracterului interactiv (dinamic) al documentelor HTML. Aceste programe pot fi scrise în principiu în orice limbaj de programare de uz general. CGI specifică modul în care

trebuie să se facă transferul cererilor sosite la serverul Web(prin HTTP) către programele CGI. CGI este deci o interfață independentă de limbaj, ce permite realizatorilor aplicațiilor Web să genereze *documente dinamice*. Numite uneori impropriu *scripturi CGI*, acestea pot fi scrise în aproape orice limbaj ce are posibilități de acces la variabilele de mediu şi poate produce o ieşire.

Scripturile CGI trebuie sa aibă o anumită extensie, de obicei extensia implicită este .cg dar poate fi si .pl, .js , sau oricare alta extensie, in funcție de limbajul de programare folosit la realizarea acestor scripturi. Pentru ca un programator HTML să poată folosi programe cgi acesta trebuie sa ceară drepturi de acces administratorului serverului Web la directorul cgi-bin. Aşa cum am mai specificat, programele CGI pot fi scrise în orice limbaj de programare care foloseşte intrări şi ieşiri standard (STDIN, STDOUT). Limbajele cu ajutorul cărora se pot realiza programe CGI se împart în două categorii, şi anume: programe compilate C, C++, Turbo Pascal, Fortran, Ada, etc. sau programe interpretate PERL, AppleScript, shell-urile UNIX, ect.

Scopul CGI este furnizarea unui mecanism flexibil şi convenabil pentru extinderea funcțiilor unui server HTTP peste limitele unor simple preluări si afişări de fişiere.

Cele mai răspândite aplicații ale CGI sunt:

- prelucrarea datelor inserate în formulare (care necesită un răspuns);
- interogarea unor baze de date pentru o anumită informație (se realizează cu motoarele de căutare: Altavista, Lycos, Yahoo, Infoseek prin interogări SQL);
- documente virtuale (documente HTML complexe, care conțin text, imagini, fişiere de sunet sau video).

PHP- Hypertext Preprocessor

PHP (Hypertext Preprocessor) este unul din cele mai folosite limbaje de programare server-side. Limbajul PHP s-a născut în 1994 din nevoia lui Rasmus Lerdorf de a afla câte persoane îi viziteaza CV-ul online. El a denumit setul de scripturi create PHP, acronimul pentru Personal home page. Pe parcursul următorilor trei ani limbajul a evoluat, adevăratul succes cunoscându-l când Zeev Suraski si Andi Gutmans au rescris motorul PHP de la cap la coadă, motor care poartă din versiunea 4 a PHP numele Zend, o

combinație de litere din prenumele creatorilor săi: Zeev si Andi. Fiind open-source, PHP beneficiază de suport activ din partea comunității online, acesta fiind şi motivul creşterii explozive a numărului site-urilor bazate pe PHP.

Se foloseşte în principal înglobat în codul HTML dar începând de la versiunea 4.3.0 se poate folosi şi în mod „linie de comandă" (CLI), permițând crearea de aplicații independente. Este unul din cele mai importante limbaje de programare web open source şi server-side existând versiuni disponibile pentru majoritatea web serverelor şi pentru toate sistemele de operare. Conform statisticilor este instalat pe 20 de milioane de situri web şi pe 1 milion de servere web.

Caracteristici:

- Familiaritatea : sintaxa limbajului este foarte uşoară combinând sintaxele unora din cele mai populare limbaje Perl sau C;

- Simplitatea : sintaxa limbajului este destul de liberă. Nu este nevoie de includere de biblioteci sau de directive de compilare, codul PHP inclus într-un document executându-se între marcajele speciale;

- Eficiența : PHP-ul se foloseşte de mecanisme de alocare a resurselor, foarte necesare unui mediu multiutilizator, aşa cum este web-ul;

- Securitate : PHP-ul pune la dispoziția programatorului un set flexibil şi eficient de măsuri de siguranță;

- Flexibilitate : fiind apărut din necesitatea dezvoltării web-ului, PHP a fost modularizat pentru a ține pasul cu dezvoltarea diferitelor tehnologii. Nefiind legat de un anumit server web, PHP-ul a fost integrat pentru numeroasele servere web existente;

- Gratuitate : este probabil cea mai importantă caracteristică a PHP-ului. Dezvoltarea PHP-ului sub licența open source a determinat adaptarea rapidă a PHP-ului la nevoile web-ului, eficientizarea şi securizarea codului.

PHP foloseste extensii specifice pentru fişierele sale: .php, .php3, .ph3, .php4, .inc, .phtml. Aceste fişiere sunt interpretate de către serverul web, iar rezultatul este trimis în formă de text sau cod HTML către browser-ul clientului. Cand accesăm o pagină HTML serverul care o gazduieşte trimite pagina HTML către browser spre afişare. In cazul unei pagini PHP serverul citeşte codul PHP, il interpretează şi generează dinamic pagina HTML care este trimisă browserului spre afişare. Acesta este motivul pentru care utilizatorii folosesc PHP pentru construirea unor pagini cu conținut dinamic.

Când PHP-ul parcurge un fişier, de fapt "citeste" textul până când întâlneşte una din etichetele speciale care îi spun să înceapă să interpreteze textul ca pe cod PHP. Se execută codul până când este întâlnită eticheta de inchidere. Apoi se "citeşte" din nou textul mai departe. Acesta este motivul pentru care se poate adăuga cod PHP în interiorul HTML-ului. Codul PHP este delimitat de unul din urmatoarele seturi de etichete de deschidere şi inchidere:

<?php	?>	etichete recomandate
<script language="php"?>	</script>	
<?	?>	folosirea lor necesită anumite setări pe server
<%	%>	etichete tip ASP, folosirea lor necesită anumite setări pe server

RSS: Formate de fluxuri web

RSS este o familie de fluxuri web, specificate în XML, şi folosite pentru a distribui ştiri, noutăți sau sumaruri de către portalurile care oferă asemenea servicii.

Un flux web (sau „feed") este un format de date, folosit pentru a oferi utilizatorilor conținut (în formă integrală sau în rezumat) împreună cu link-ul către sursă şi o serie de elemente descriptive. Se mai numeşte şi flux RSS, feed RSS, webfee, stream RSS sau canal RSS. Fluxurile web oferă acest conținut cel mai adesea sub forma unui fişier XML**.**

Apărut în 1997 şi utilizat pentru prima dată de Netscape, feed-ul RSS a devenit un mod foate popular de distribuie a informațiilor. Practic, cam toate site-urile cu adevărat importante ce oferă material updatat zilnic folosesc serviciul RSS.

Fluxul de informații de tip RSS este marcat pe site-urile ce oferă acest serviciu într-un mod uşor de identiifcat. Abonarea la fluxurile de tip RSS implică fie existența unei aplicații desktop destinată acestui lucru (feed reader, de genul client email - MS Outlook, Mozilla Thunderbird, Apple Mail - sau browser web - Internet Explorer, Mozilla Firefox,Apple Safari), fie o aplicație sau site web (Google Reader, MyYahoo, Bloglines, etc).

Modalitatea de abonare la un flux RSS al unui site depinde foarte mult de clientul de RSS folosit; de pildă, MS Outlook 2007 sau Mozilla Firefox:

Abrevierera RSS provine din referirea la următoarele standarde:

*** Really Simple Syndication;**

*** Rich Site Summary;**

*** RDF Site Summary.**

RSS este un fişier XML; el descrie conținutul unui site, fiind actualizat odată cu acesta. Inițial, utilizat de Netcape pentru a crea pagini *My Netscape*, RSS a fost adoptat de serviciile de informare (news syndication services), weblog-uri. RSS este important pentru că permite informarea despre actualizări ale resurselor online fara a face apel la e-mailuri sau Newslettere, implicit reducând posibilitatea de a primi spam şi viruşi. Un fişier RSS include un logo, linkul la site, un câmp de citire şi itemii de noutate; fiecare item constă din URL-ul articolului respectiv, titlul, o descriere . Sisteme numite agregatoare - *aggregators* sau *harvesters*, citesc RSS-urile şi informează abonații lor asupra modificărilor de pe site, informația fiind astfel transmisă unei audiențe largi - *syndicated*. Informația din fişierele RSS este stocată in baze de date, care pot fi accesate cu multiple criterii de căutare.

Utilizatorii care vor să afle noutățile apărute pe site vor folosi un cititor de RSS ("RSS reader"), cititor de feed ("feed reader") sau un agregator ("aggregator"), care

poate fi instalat pe calculator sau poate fi integrat în browser (precum Feedster sau Blogdigger). Există agregatoare care permit gruparea mai multor surse de ştiri într-una singură. Utilizatorul se înscrie la un RSS prin introducerea URL-ului şi astfel este început procesul de subscriere. Cititorul de RSS verifică în mod regulat feed-urile utilizatorului, downloadează update-urile pe care le gaseşte şi oferă o interfață pentru a citi feed-urile.

SVG: Scalable vector graphics

Lista de programe care poate deschide fişiere cu extensia *.SVG.*:

- firefox.exe;
- gimp-2.6.exe;
- ImgBurn.exe;
- Inkscape.exe;
- Paint Shop Pro 9.exe;
- sbrowser.exe.

SVG (Scalable Vector Graphics) este un limbaj pentru descrierea de imagini 2D folosind XML (Limbajul XML - *eXtensible Markup Language* - este limbajul ce oferă un format pentru stocarea şi transmiterea de date). Acest format de fişier este folosit pentru salvarea de fişiere vectoriale grafice. Fişierele SVG pot fi vizualizate pe web şi sunt susținute de browsere moderne care permit ca imaginile vizalizate să aibă o înaltă calitate grafică.

Este un standard al organizației W3C a cărui proiectare a început în anul 1999. Permite definirea imaginilor prin 3 metode: text, grafică vectorială şi bitmap-uri.

Deşi există aplicații specializate pentru crearea şi editarea de SVG-uri, în acest scop poate fi folosit orice editor text. Vizualizarea unei imagini *SVG* poate fi realizată cu orice browser modern.

În momentul de față *SVG*-ul are mai multe profile pentru a se adapta mai bine la diferite constrângeri. Astfel, profilele SVG Tiny şi SVG Basic au fost create special pentru dispozitivele mobile cu resurse limitate, iar profilul SVG Print este destinat mediilor de printare a documentelor. Pentru animarea unei imagini SVG organizația W3C recomandă standardul SMIL. Pe lângă recomandarea oficială mai există şi alte soluții, precum ar fi ECMAScript.

Principalele elemente din componența unui fişier *SVG* sunt:

- Paths – „Căi” ce pot fi folosite pentru descrierea conturului unei forme. Conturul poate rămâne gol sau poate fi umplut. Căile pot fi folosite şi pentru a specifica zonele de decupare;

- Forme de bază – Specificația *SVG* oferă posibilitatea folosirii formelor de bază: dreptunghi, cerc, elipsă, linii şi poligoane. Acestea pot fi construite şi cu ajutorul căilor care au acelaşi contur;

- Text – Pentru a specifica textul ce apare într-o imagine trebuie folosite elemente de tipul text;

- Painting – Se referă la posibilitatea de a umple formele specificate în *SVG*. Pentru aceasta poate fi folosită o singură culoare, o culoare cu transparență, un gradient sau un model;

- Culoare - Propietatea culoare este folosită pentru specificare culorii;

- Gradient şi model – Folosite pentru colorarea formelor specificate;

- Decupare, mascare - Pot fi folosite zone de decupare sau de mascare;

- Filtre – Descriu diferite efecte aplicate imaginilor;

- Interactivitate – O imagine SVG are posibilitaea de a interacționa cu utilizatorul. Astfel, la apăsarea unui buton sau la folosirea mouse-lui pot fi pornite diferite scripturi;

- Linkuri – Un document poate conține legături către alte pagini sau elemente web;

- Scripting – Într-un SVG pot fi definite scripturi cu diverse funcții;

- Animații – Pentru un SVG pot fi specificate diverse tipuri de animații;

- Font - Nu este necesar ca utilizatorul final să aibă instalate diferitele seturi de caractere folosite; fonturile pot fi incluse în imagine;

- Metadata – Pentru intregrarea mai bună se oferă şi opțiunea specificării de metadate (acestea sunt datele care descriu datele propriu-zise).

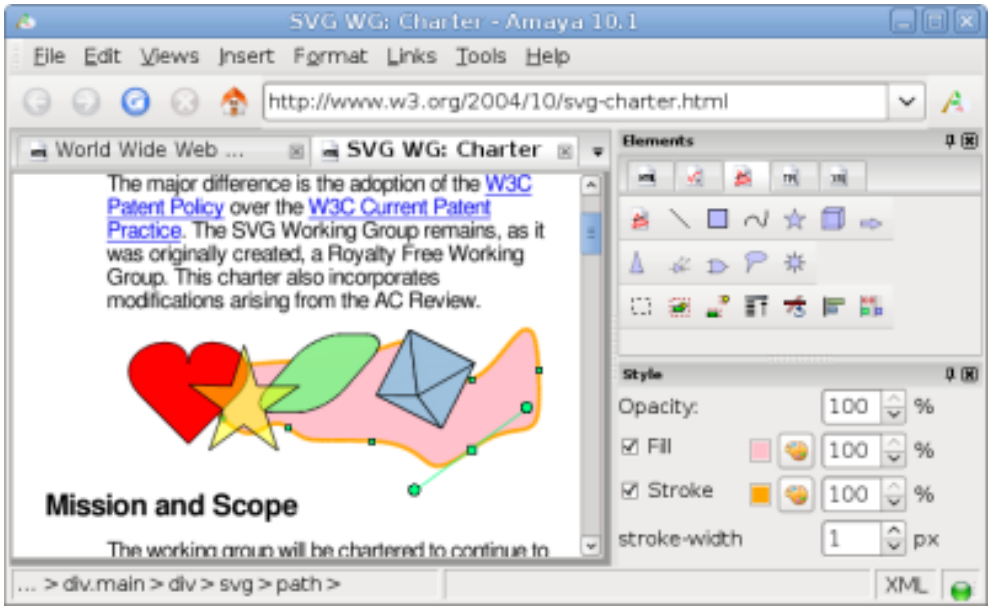


Fig.3.1. Formatul SVG

SVG este util pentru prezentare şi comunicare, reprezentarea datelor statistice, ştiințifice şi medicale, pentru acces la informație pentru persoane cu dizabilități etc.

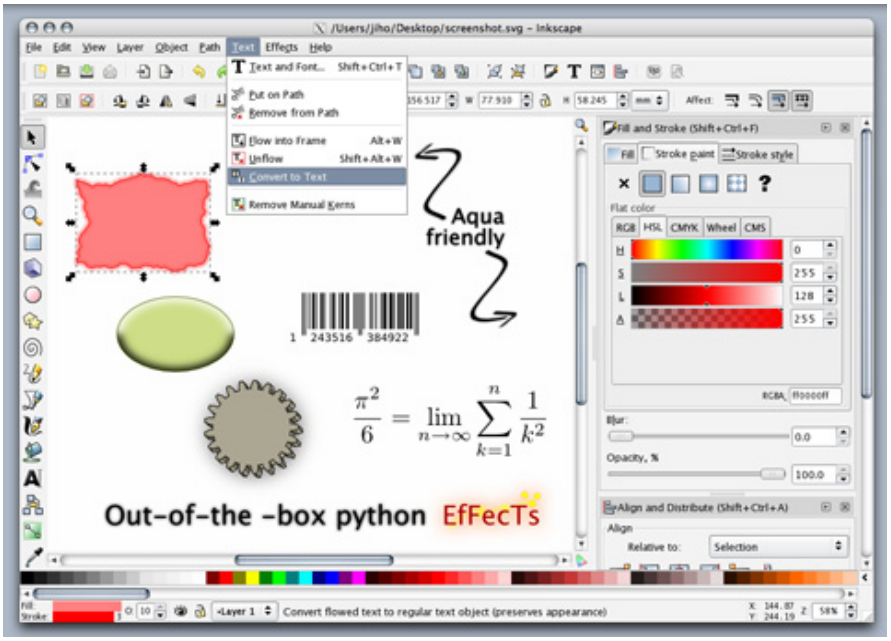


Fig.3.2. Formatul SVG grafice, formule

VRML: Virtual Reality Modeling Language

VRML (Virtual Reality Modeling Language)

este un format standard de fişier care prezintă imagini 3D a diferitelor forme într-o mediu interactiv.

Fişierele *VRML* sunt de obicei numite "lumi" şi au extensia WRL (de exemplu, island.wrl - *http://yallara.cs.rmit.edu.au/~rotaylor/web3d/journal.shtml*).

Prima versiune a *VRML* a fost specificată în noiembrie 1994. Versiunea actuală şi funcţională este VRML97 (ISO / IEC 14772-1:1997); VRML a fost acum înlocuit de X3D (ISO / IEC 19775-1). Termenul *VRML* a fost inventat de Dave Raggett într-un document depus la Prima Conferinţă Internaţională de World-Wide-Web în anul 1994, dar el a fost dezvoltat de Tony Parisi şi Peter Kennard.

În 1997 a foat finalizată o nouă versiune a formatului şi VRML97 (cunoscut ca VRML2 sau VRML 2.0) şi a devenit un standard ISO. VRML97 a fost folosit pe internet pe unele homepages personale şi site-uri, cum ar fi "CyberTown", care a oferit chat 3D utilizând Blaxxun Software-ul. Formatul a fost promovat de către Cosmo SGI Software-ului; atunci când SGI a fost restructurat în 1998, divizia a fost apoi preluată de Computer Associates care a decis să nu dezvolte sau să distribuie software-ul.
Consorţiul VRML a schimbat numele în Web3D şi a început să lucreze la succesorul VRML –X3D.

VRML a provocat mult interes, dar nu a fost niciodată utilizat pe scară largă. Un motiv pentru acest lucru poate fi fost lipsa de lăţime de bandă disponibilă. Experimentarea *VRML* a fost în primul rând în domeniul educaţiei şi cercetării.

VRML este util pentru o varietate de aplicaţii, inclusiv pentru vizualizarea datelor, analize financiare, Entertainment, educaţie, proiectare asistată de calculator, mall-uri virtuale, interfeţe la informaţii, vizualizare ştiinţifică.

Scopul limbajului *VRML* - *Virtual Reality Modeling Language* a fost crearea unui standard prin care să se facă tranziţia de la interfaţa bidimensională a Web-ului la una tridimensională şi care să poată fi şi manipulată dinamic.

VRML permite descrierea de obiecte 3D şi combinarea lor în scene şi lumi virtuale. El poate fi folosit la crearea de lumi interactive, care să conţină: imagini, animaţii, clipuri audio sau video. Într-o lume VRML utilizatorul şi nu computerul are controlul. Obiectele dintr-o scenă pot interacţiona unele cu altele prin evenimente sau pot interacţiona cu evenimente utilizator. Lumile pot fi legate între ele sau cu alte documente

HTML. O lume creată poate fi distribuită pe Internet, văzută şi explorată interactiv de mai mulţi vizitatori în acelaşi timp.

Pentru a putea vizualiza o lume VRML, browser-ul de Internet (Internet Explorer, Netscape Navigator, HotJava, Opera etc.) are nevoie de un *VRML browser* (sub forma unui plugin de cele mai multe ori) ce permite utilizatorului (numit *vizitator* în VRML) să navigheze în lumea virtuală, să obţină diferite vederi ale scenei în funcţie de punctul de vedere ales, să interacţioneze cu elementele scenei etc.

4. Conversia documentelor din format tradiţional în format electronic

Pe plan mondial, proiectele de digitizare atrag tot mai mulţi cercetători, ingineri, specialişti în calculatoare, în tehnologia informaţiei şi a comunicării. Interesul mare în această direcţie se explică prin fondurile substanţiale care au început să fie investite în acest tip de cercetare:

- Colorado Digitization Project a fost iniţiat în 1998 şi a reunit eforturile bibliotecilor, muzeelor, societăţilor istorice şi arhivelor din Colorado pentru a creşte accesul utilizatorilor la colecţiile speciale şi resursele unice deţinute de aceste instituţii. Proiectul presupune un catalog comun de metadate şi a dezvoltat instrumente speciale pentru creatorii de baze de metadate.

- Conferința IFLA din anul 2003, ținută în Turcia, a avut ca temă de dezbatere subiectul digitizării: „Introducing Digitization into Turkish Libraries: Current Attitudes and the Way Forward".

- Un alt proiect important şi complex de digitizare a fost lansat în Spania. AGID - Archivo General de Indias - a inițiat un proiect deosebit de ambițios pentru digitizarea colecțiilor sale. Mai mult de un milion de pagini de documente privind istoria Spaniei au fost digitizate.

- Uniunea Europeană a lansat, de asemenea, foarte multe proiecte de standardizare a cataloagelor de publicații, de digitizare a publicațiilor şi de creare a portalurilor de acces multilingve la aceste biblioteci virtuale.

Noile tehnici de digitizare, recent dezvoltate, combină în mod optim robotica, electronica digitală şi tehnica de calcul prin crearea liniei de digitizare care oferă în premieră o soluție complet integrată. Cele două avantaje majore ale liniei de digitizare sunt productivitatea foarte ridicată şi menținerea integrității fizice a cărții supuse procesului de digitizare. Manipularea documentelor, transferul şi digitizarea lor sunt automate, asigurându-se prezervarea fidelă a conținutului informației precum şi a formei originale a documentului indiferent de tipul acestuia (cărți, reviste, colecții de ziare).

A. În octombrie 2002, firma elvețiană 4Digitalbooks ASSY, cea care a conceput DIGITIZING LINE, a instalat un prim echipament de acest tip la Green Library, una dintre bibliotecile Universității Stanford, California. Rezultat direct al unui prototip pentru formatul maxim A4, prima versiune industrială a lui DIGITIZING LINE permite scanarea complet automată a volumelor sau a ziarelor până la formatul A2. Echipamentul realizează în premieră mondială întoarcerea automată a paginii pentru întreaga gamă de formate de pagină.

Bibliotecile Universității Stanford au în total aproximativ opt milioane de cărți digitizate printre care toate titlurile publicate de Stanford University Press, inclusiv conservarea lucrărilor începând cu 1923. În prezent, sistemul DIGITIZING LINE instalat la Stanford funcționează la capacitate maximă pentru a scana câteva mii de cărți anual.

Proiectul prin care s-a dezvoltat linia de digitizare la Biblioteca Universității din Southampton se numeşte BOPCRIS. Una dintre principalele lui activități este scanarea documentelor parlamentare, rare şi istorice, folosind DIGITIZING LINE, pentru prima dată în iunie 2004 în Marea Britanie şi a doua oară în lume. Linia de digitizare scanează 600 de pagini / oră, ceea ce a făcut posibil accesul la documentele istorice, inaccesibile utilizatorilor până în acel moment. (DigiBook, nr. 9, December 2004).

În România, guvernul a început să acorde interes tot mai mare acțiunilor de digitizare, în special a materialelor vechi, pentru a oferi acces mai mare şi pentru a le păstra intacte, în acelaşi timp. Câteva exemple prezentate în cele ce urmează arată stadiul actual al digitizării în țara noastră:

ANBPR (Asociația Națională a Bibliotecilor Publice) - Comisia pentru digitizarea documentelor: propune un proiect de digitizare ca o metodă de păstrare şi conservare a documentelor în original;

CIMEC (Centrul Institutului de Memorie Culturală): lansează ideea digitizării de fotografii şi crearea unei baze de date;

Biblioteca Universității „Transilvania" a inițiat programul „Leonardo da Vinci" în anul 2003. Proiectul „*Abordări inovatoare în cadrul bibliotecii universitare - formarea de bibliotecari şi specialişti specializați în comunicare electronică, web-design, transmisie de informație digitală, capabili să conceapă produse şi servicii utilizând ICT*" din anul 2003 s-a finalizat cu realizarea de produse şi servicii de o calitate deosebită, rezultate diseminate în cadrul universităților şi instituțiilor participante la proiect: S-au obținut documente digitizate diseminate folosind două instrumente importante: catalog Web şi catalog pe CD.

Serviciile oferite se referă în primul rând la punerea la dispoziția publicului cititor a unei varietăți de resurse originale, incluzând cărți, documente, hărți, atlase, toate făcând parte din fondul rar şi vechi şi, în acelaşi timp foarte valoros. Devin accesibile informații complete referitoare la fondul rar: titlu, scanarea coperții şi a paginii de titlu, descriere bibliografică, clasificare zecimală universală (CZU) şi cuvinte cheie.

Un alt proiect „Leonardo da Vinci” a demarat în anul 2005, reunind instituții de renume din Braşov, Prima Şcoala Românească, Arhivele Naționale, filiala Braşov, şi Biblioteca Județeană Braşov şi având ca obiectiv crearea unui *Centru de Pregătire* prin intermediul căruia să se asigure pregătirea profesională continuă în cadrul organizațiilor, universităților, îmbunătățirea serviciilor publice şi relațiilor interuniversitare, interorganizaționale, legăturile dintre universități şi instituții sau alte companii. Un alt pas important a fost făcut prin caştigarea grantului CNCSIS în anul 2003, „*Cercetari privind metodele moderne de conservare a colecțiilor, cu aplicații în managementul bibliotecii digitale*”. În acest proiect, echipa de cercetare a testat şi cercetat o nouă bază teoretică şi experimentală pentru proiectare şi realizare de sisteme tehnice de conservare şi arhivare a documentelor tradiționale şi a celor electronice din biblioteci.

În toate aceste proiecte româneşti soluția de digitizare aleasă a fost ***DIGITIZING LINE***, produsul realizat în Elveția şi care a fost deja utilizat în Universități din întreaga lume. El realizează scanarea documentelor (cărți, reviste, ziare) cu ajutorul unei tehnologii foarte avansate, întoarcerea paginilor şi stocarea informațiilor fără afectarea calității şi integrității documentelor.

Posibilitatea de a folosi o gamă largă de dispozitive scanner este una din caracteristicile importante ale unui sistem de digitizare bun.

4.1. Metode de digitizare

4.1.1 Cerințe în procesul de digitizare

În acest capitol vom face o scurtă prezentare a diferitelor modalități în care se poate realiza achiziția imaginilor în formă digitală, plecând de la o carte. Unele variante au fost încercate în mod concret, rezultatele confirmând aşteptările noastre.

O primă problemă care apare este cea a manipulării cărților în timpul procesului de digitizare, principalele probleme întâlnite fiind gradul ridicat de degradare, respectiv modul de legare al acestor cărți, de obicei destul de strâns, ceea ce împiedică deschiderea perfect dreaptă (fig. 4.1.).

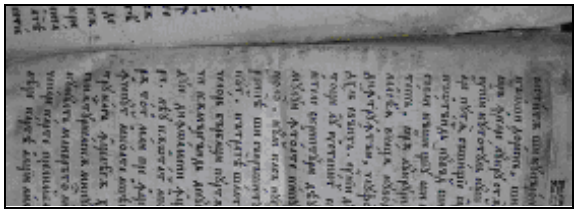


Fig. 4.1. Modul de legare al cărților

Pentru a evita deteriorarea cărților va trebui deci să luăm în considerare faptul că acestea nu vor putea fi manipulate în aceleaşi condiții ca documente nelegate.

În general există mai multe metode de a realiza digitizarea unei document:

1. Folosirea unui scanner ;
2. Folosirea unui aparat foto digital (camere planetare) ;
3. Scanarea filmelor folosite în fotografia clasică.

4.1.2 Scanare

Această variantă are ca avantaj incontestabil costul şi rezoluțiile ridicate de scanare atinse, dar exista o serie de limitări majore:

• Pentru a obține o imagine decentă, cartea trebuie să fie amplasată cu fața în jos pe sticlă, apoi ridicată, trebuie dată pagina şi aşezată din nou. Aceste operații repetate pot deteriora o carte.

• Cartea trebuie presată puternic la cotor pentru a face paginile cât mai paralele cu dispozitivul de scanare.

• Chiar şi prin apăsare puternică, unele cuvinte sau imagini vor fi distorsionate, iar spre cotor va apărea o nuanță mai închisă.

• În cazul unor documente speciale, dimensiunea standard de scanare (undeva pe la A4) poate să nu ajungă.

4.1.3 Fotografiere digitală

O soluție acceptabilă din punct de vedere al costurilor, dar cu pierdere în rezoluție, comparativ cu soluția scanării. Avantajul major, poate chiar hotărâtor, este faptul că operațiile asupra cărții de digitizat sunt mai puțin agresive, fiind nevoie de simpla aşezare şi deschidere a cărții, pentru fiecare pagină fiind necesară doar operația de întoarcere a paginii. Cel mai important lucru este cel că nu trebuie să apăsam cartea pentru a obține rezultate bune.

În scopul prelucrărilor ulterioare (de exemplu OCR), avem nevoie de anumite rezoluții minime. În următorul tabel avem câteva valori utilizate în descrierea aparatelor digitale şi conversia acestora în alte unități de măsură. (tabel 5)

Tabel 5. Caracteristici tipice ale aparatelor digitale

Megapixeli	Dimensiuni (pixeli x pixeli)	Dimensiune pe disk (Mb)	Dimensiune in inch la 300 dpi
0.3	640 x 480	0.9	2.1 x 1.6
2.1	1792 x 1200	6.15	6 x 4
3.2	2160 x 1440	8.9	7.2 x 4.8
4.3	2400 x 1800	12.4	8 x 6
5.1	2608 x 1952	14.6	8.7 x 6.5
8	3264 x 2448	22.9	10.9 x 7.3

Un alt criteriu ce merită luat în calcul în cazul fotografierii digitale îl reprezintă numărul de puncte de focalizare pe care îl are camera digitală. Un număr mai mare de puncte de focalizare va asigura o calitate mai buna a imaginii, chiar dacă pagina nu este perfect întinsă, caz frecvent în cazul cărților cu număr mare de pagini unde manipularea poate fi dificilă.

Varianta cea mai bună ar fi folosirea unei camere planetare, aparat special construit pentru digitizarea cărților, dar cu costuri mult mai mari, având un preț orientativ de 15.000 dolari. Acest aparat combină avantajele camerei digitale şi cele ale scanerului.

4.1.4 Scanarea fotografiilor analogice

O soluție extremă ar fi scanarea filmelor realizate prin fotografiere clasică. Pentru această avem nevoie de un film cu sensibilitate mare, un aparat analogic şi un scaner care permite scanarea negativelor.

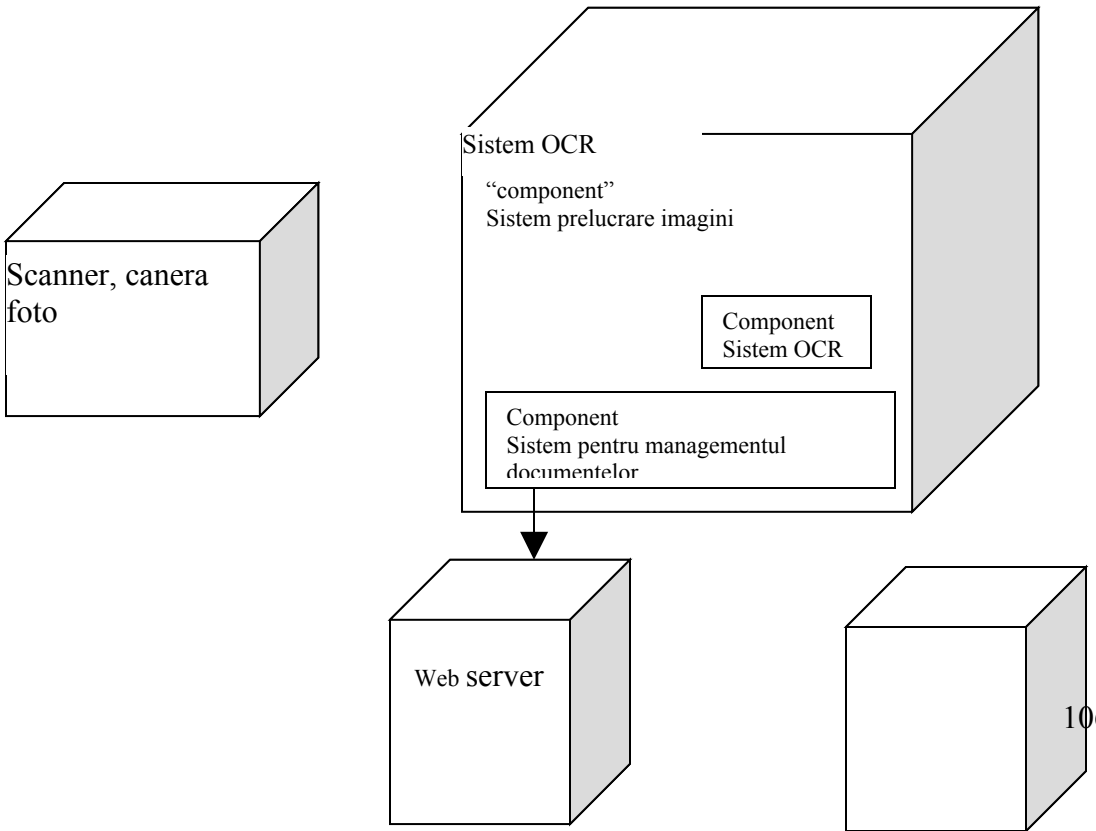
4..2 Arhitectura unui sistem de digitizare

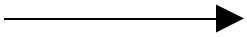
Un asemenea sistem are ca scop îmbunătățirea procesului de recunoaştere de caractere prin utilizarea unui sistem hardware – software adecvat acestei activități şi prin definirea unor procese clare de utilizare a acestuia.

Digitalizarea imaginilor se face cu ajutorul unei camere foto digitale sau a unui scanner special pentru a nu fi necesară deschiderea completă a cărților, astfel evitându-se deteriorarea acestora.

Sistemul este compus dintr-o componentă de achiziție de imagini (scanner sau cameră foto digitală), un sistem software de preprelucrare a imaginilor, un sistem software de recunoaştere a caracterelor şi un sistem de înmagazinare a informațiilor acumulate în diversele faze ale procesului.

Documentele rezultate pot fi apoi prezentate atât utilizatorilor web cât şi utilizatorilor locali prin intermediul unor module de prezentare personalizate în funcție de cerințe. În figura următoare se pot vedea componentele importante ale sistemului şi legăturile dintre acestea. Săgețile prezintă curgerea informațiilor în interiorul sistemului nu dependențe intre componente (fig. 4.2.).





internet

Fig.4.2. Schema bloc a sistemului de digitizare

Trebuie luat în considerare faptul că scopul final este o prezentare a documentelor unor utilizatori interesați astfel că ieşirile din acest sistem ar trebui să fie documente care să permită căutarea în textul acestora pe cât posibil fără a deteriora informația vizuala din documentele originale.

Documentele create trebuie să aibă următoarele caracteristici:

- trebuie să conțină atât imagini cât şi text;
- să existe posibilitatea de căutare;
- să poată fi citite de către clienți web uzuali;
- să aibă conceptul de pagină;
- să fie de dimensiuni reduse ;
- să fie suportate de către sistemele OCR existente (ca documente de ieşire.

4.3 Moduri de livrare a conținutului digital

Pentru a livra conținutul (imaginile pe care s-a făcut OCR) propunem ca soluție documentele Acrobat. Acest tip de documente conțin mai multe nivele informaționale. Nivelul text se poate pune atât deasupra nivelului imagine cât şi sub acesta. Cele mai utile modalități de prezentare a conținutului în acest caz sunt: text sub imagine şi respectiv text peste imagine.

4.3.1 Text sub imagine

În literatură acest mod de prezentare se numeşte "text under the image". Adobe a creat acestă funcționalitate pentru a pune la dispoziție o funcționalitate de căutare în documentele cu conținut "sensibil" în care recunoaşterea nu este perfectă şi în care nu ne permitem ca utilizatorul final să vadă greşelile dar totuşi existând posibilități de selecție / căutare a textului de sub imagini, selecția reflectându-se asupra imaginii de deasupra.

4.3.2 Text peste imagine

În literatură este cunoscut sub numele de “text over the image”. Textul original de pe imaginile pe care s-a făcut OCR se şterge automat şi este înlocuit de textul recunoscut.

4.4. Alegerea unui format

Prin analiza criterială am stabilit importanta caracteristicilor şi am luat în considerare mai multe tipuri de documente existente inclusiv un tip de document proprietar care ar putea fi dezvoltat special pentru acest proiect (tabel 6).

În urma analizei s-a ajuns la concluzia că dezvoltarea unui format proprietar nu este cea mai bună soluție, documentul PDF câştigând:

Tabel 6 : Caracteristici ale diferitelor formate de documente

Caracteristica	Importanță caracteristică	. TIFF	DOC	PDF	Custom format
Image + text	3	1	1	1	1
Căutare	4	0	0.5	1	1
Compatibilitate cu clienții existenți	5	0.75	1	1	0
Paginabile	2	1	0.75	1	1
Dimensiune	1	0.75	0.6	0.8	1
Suportabilitate	6	1	1	1	0
Punctaj acumulat		12.5	18.1	20.8	10

În plus față de celelalte documente, formatul PDF permite aşezarea textului recunoscut sub imaginea inițială (text under the image) astfel că utilizatorul va vedea imaginile dar va căuta în text.

Acest lucru asigură o calitate ridicată a vizualizării chiar şi în cazul în care recunoaşterea a fost de o calitate mai scăzută. Documentele PDF permit şi înmagazinarea de metainformații în acestea.

Mecanismele de protecție ale documentelor Pdf nu sunt de neglijat fiind posibil să protejam un document la salvare, copiere, tipărire, copy-paste, etc .

Sistemul trebuie să fie gândit astfel ca toate documentele intermediare să fie păstrate, în cazul unor erori nemaifiind necesară reluarea întregului proces (de la scanare la salvare PDF) ci doar parți ale cestuia.

4.5. Tehnici de conversie prin Recunoaşterea Optică a Caracterelor (Optical Character Recognition-OCR)

4.5.1. Scurt istoric al conversiei documentelor din format tradiţional prin scanare şi Recunoaşterea Optică a Caracterelor (OCR)

Recunoaşterea textului din imagini a fost un subiect mult discutat de-alungul timpului. „Recunoaşterea optică a caracterelor (OCR) transformă imagini de text, cum ar fi documentele scanate, în caractere de text. Cunoscută şi sub numele de recunoaşterea textului, OCR face posibilă editarea şi reutilizarea textului conţinut de imaginile scanate. OCR utilizează o formă de inteligenţă artificială, cunoscută sub numele de recunoaşterea modelului, pentru identificarea individuală a caracterelor unui text dintr-o pagină, inclusiv semnele de punctuaţie, spaţiile şi sfârşitul de linie”. (http://office.microsoft.com/ro-ro/help/HP030812551048.aspx). De la începuturile din 1950 tehnica a fost în permanenţă îmbunătăţită. Script-urile recognoscibile au fost la început numerele arabe şi apoi alfabetele latine, japoneze, chineze. Multe tipuri diferite de formate pe hârtie pot fi citite astăzi prin OCR-izare. Tehnicile de recunoaştere a caracterelor au cunoscut perioade diferite de dezvoltare. Se identifică două momente în care acestea au avut de suferit ca utilitate şi dezvoltare. Primul moment a fost în anii '80, atunci când au apărut programele de birotică. Acestea facilitau crearea documentelor direct în format digital (documente "born digital"). Al doilea moment a fost implementarea crescândă a noilor tehnologii şi utilizarea Internetului. Se reevaluează importanţa manuscriselor şi trecerea lor în format digital. Dispozitive mobile mobile cu microcamere au acum incorporate unităţi de procesare capabile de recunoaştere în timp real al caracterelor.

OCR a apărut în 1950 în Statele Unite ale Americii, în aceeaşi perioadă în care apărea calculatorul UNIVAC. În anii 1960 IBM crează propriul program de recunoaştere capabil să citească numere tipărite şi scrise.. Tot în acest an, s-au mecanizat operaţiunile poştale, astfel scrisorile erau sortate cu ajutorul unor dispositive mecanice cu OCR. În 1974, Ray Kurzweil a dezvoltat un program capabil de a recunoaşte caracterele tipărite în orice font. În anii 1980 dispozitivele de recunoaştere şi-au redus dimensiunile datorită progreselor din sfera semiconductorilor şi a microprocesoarelor.

4.5.2. Starea actuală a tehnologiei OCR

Recunoaşterea precisă a fontului latin, text scris la maşină nu este considerat o problema rezolvată în aplicațiile unde imaginile clare sunt puse la dispoziție prin scanarea documentelor printate.

Rata preciziei depăşeşte 99%; acuratețea totală poate fi atinsă doar prin reverificare umană. Alte arii cum ar fi cele care includ recunoaşterea scrisului de mână şi a celui printat în alte fonturi (scripts), în special cele cu un număr mare de caractere, sunt în continuare subiectul cercetărilor în domeniu.

Rata preciziei poate fi măsurată în mai multe feluri, precum şi modul în care acestea sunt măsurate pot afecta foarte mult rata raportată de precizie. De exemplu, dacă nu este folosit un dicționar pentru a corecta cuvintele nonexistente găsite de către soft, o marja de eroare de 1% la litere (acurateţe de 99%) poate duce la o marja de eroare de 5% (acurateţe de 95%) sau mai mult în cazul în care fiecare cuvânt cu o litera greşită este luat ca şi greşit.

Recunoaşterea on-line a caracterelor este deseori confundată cu recunoaşterea optică a caracterelor. OCR este un sistem de recunoaştere de caractere **off-line**, unde sistemul recunoaşte formele statice a caracterelor, în timp ce recunoaşterea **on-line** a caracterelor implică recunoaşterea mişcării dinamice a scrisului de mână. De exemplu, recunoaşterea on-line, cum ar fi recunoaşterea gesturilor în SO Penpoint sau în Tablet PC, poate preciza dacă o linie orizontală a fost desenată dinspre dreapta spre stânga sau invers. Recunoaşterea on-line a caracterelor este, de asemenea, echivalentă şi cu alți termeni, cum ar fi: recunoaşterea dinamică a caracterelor, recunoaşterea în timp real a caracterelor şi recunoaşterea inteligentă a caracterelor denumita şi ICR.

Sistemele dinamice de recunoaştere on-line au devenit cunoscute ca produse comerciale în ultimii ani. Printre acestea se numără şi dispozitivele periferice pentru asistență personală digitală asemenea celor care folosesc SO Palm. Corporația Apple a patentat acest produs. Algoritmii folosiți în aceste dispozitive au avantajul faptului că ordinea, viteza şi direcția liniilor de segment individuale la introducere sunt cunoscute.

De asemenea, utilizatorul este obligat să folosească doar forme specifice. Aceste metode nu pot fi folosite în softuri care scanează documente de hârtie, astfel recunoaşterea cu acurateţe a documentelor scrise de mână este încă o problema mare.

Rata acurateței de 80% până la 90% a fişierelor cu scrisul de mână lizibil poate fi atinsă, dar cu o asemenea acurateţe tot mai apar zeci de greşeli pe pagină, astfel aceasta tehnologie este folositoare doar în aplicații limitate.

Recunoaşterea textului cursiv este o zona activă de cercetare, cu rate de recunoaştere chiar mai mici decât cea a recunoaşterii scrisului de mână. Rate mai mari de recunoaştere a textului nu vom putea obține fără ajutorul informațiilor gramaticale. De exemplu, recunoaşterea cuvintelor întregi folosind un dicționar este mai uşoară decât încercarea de analiză individuală a caracterelor.

Cunoaşterea gramaticii limbii textului scanat poate de asemenea să ajute, de exemplu, să determine dacă un cuvânt este verb sau substantiv pentru o mai mare acurateţe. Formele individuale de caractere cursive pur şi simplu nu conțin informații suficiente pentru a recunoaşte cu acurateţe (mai mult de 98%) toate caracterele scrisului de mână. Este necesar a înțelege că tehnologia OCR este o tehnologie de bază de asemenea folosită de aplicațiile de scanarea avansată. Din aceste motive, o soluție de scanare avansată poate fi unică, patentată şi protejată de drepturile de autor, deşi este bazată pe tehnologia OCR de bază.

Recunoaşterea optică a caracterelor se realizează în doi paşi:

-utilizarea unui dispozitiv pentru scanarea informației tipărite într-un calaculator ca imagine bit-map;

-aplicarea algoritmilor de recunoaştere a caracterelor pentru crearea fişierului text.

Acurateţea atinsă în recunoaşterea caracterelor din imagini este cel mai important factor care determină eficacitatea şi fezabilitatea unui produs OCR.

Software-ul OCR este proiectat pentru a asigura conversia documentelor scrise olograf sau tipărite, digitizate pein scanare, într-o formă care se pretează procesării computaționale. Fişierele text în reprezentare ACSII sau Unicode sunt produse ale perogramelor de OCR-izare. Sistemele OCR îşi au originile în recunoaşterea "pattern-urilor" şi inteligența artificială. Recunoaşterea vizuală a caracterelor utilizând tehnici ca oglinzi şi lentile şi recunoaşterea digitală a caracterelor utilizând scanere şi algoritmi specifici se regăsesc în tehnicile OCR.

4.5.3.Tehnologii curente, produse software pentru OCR (tabel 7)

Tabel 7. Tehnologii curente, produse software pentru OCR

Nume	Licența	Sistem de operare	Descriere
ExperVision Typereader OpenRTX	**Comercial**	Windows, Mac OS X, Unix, Linux, OS/2	ExperVision Inc. a fost fondat în 1987, tehnologia şi produsul lor OCR. Au luat cele mai mari note în testarea independentză făcută de UNLV în anii în care a participat
ABBYY FineReader OCR		Windows, Mac OS X	Pentru a lucra cu interfețe locale este necesar suportul lingvistic corespunzător.
AnyDoc Software OCR for AnyDoc	Comercial	Windows	Lucrează cu structuri, semistructuri şi documente nestructurate
OmniPage	Comercial	Windows, Mac OS	Produs de Nuance Communications
Readiris Windows,	Comercial	Mac OS X	Produs de I.R.I.S. Group of Belgium. Ediții Asian şi Middle Eastern
CuneiForm	BSD variant	Windows, Linux, BSD, MacOSX.	Enterprise-class system, multi language, poate salva textul formatat şi tabele de recunoaştere complexe a oricărei structuri
Puma.NET	BSD	Windows	.NET OCR SDK bazat pe tehnologie
CVISION Technologies, Inc. PdfCompressor and Maestro Recognition Server	Comercial	Windows	Rapidă, precisă, volum mare de date.
GOCR	GPL	diverse (open source)	Dezvoltare de început.
Microsoft Office Document Imaging	Comercial	Windows	Foloseşte motorul ScanSoft OCR .
Microsoft Office OneNote 2007 l	Comercia	Windows	
NEOPTEC DATASCAN	Comercial	Windows	Soft pentru procesare automată a cererilor şi

.			chestionarelor.
NovoDynamics VERUS	Comercial		Produs specializat în limbile din orientul mijlociu
Ocrad	GPL	Unix-like, OS/2	
Brainware	Comercial	Windows	Extracţii şi procesări de date din documente în orice sistem backend; printre documentele cunoscute se număra: chitanţe, declaraţii
HOCR	GPL	Linux	Hebrew OCR
OCRopus	Apache	Linux	Poate utiliza Tesseract
PDF OCR X	Comercial	Mac OS X	Utilitate drag and drop care poate converti fişiere PDF în fişiere text folosind OCR. Foloseşte Tesseract
ReadSoft	Comercial	Windows	Scanează şi clasifică documente oficiale cum ar fi: chitanţe, facturi
Alt-N Technologies' RelayFax Network Fax Manager	Comercial	Windows	Utilitar OCR multi-lingvistic care converteşte documente fax în documente editabile (.doc, .pdf,…) în limbi diferite.
Scantron Cognition	Comercial	Windows	Pentru lucrul cu interfeţe locale, suport lingvistic
SimpleOCR	Freeware şi Comercial	Windows	
SmartScore	Comercial	Windows, Mac OS	Pentru note muzicale.
Tesseract	Apache	Windows, Mac OS X, Linux	Creat de către Hewlett-Packard; sub dezvoltare curentă de către Google

4.5.4. Suportul lingvistic al produselor pentru OCR

Tabel:7 Suportul lingvistic al produselor OCR

Nume **Ultima versiune** **Anul lansării**	**Limbi recunoscute**
ExperVisionTypeReader & OpenRTK	English, French, German, Italian, Spanish, Portuguese, Danish, Dutch, Swedish,

7.0 2007	Norwegian, Hungarian, Polish, Simplified Chinese, Traditional Chinese, Russian, Finnish şi Polynesian
ABBYYFineReader OCR 10.0 2009	Abkhaz, Adyghian, Afrikaans, Agul, Albanian, Altai, Armenian (Eastern, Western, Grabar), Avar, Aymara, Azerbaijani (Cyrillic), Azerbaijani (Latin), Bashkir, Basic, Basque, Belarusian, Bemba, Blackfoot, Breton, Bugotu, Bulgarian, Buryat, C/C++, COBOL, Catalan, Cebuano, Chamorro, Chechen, Chinese Simplified, Chinese Traditional, Chukchee, Chuvash, Corsican, Crimean Tatar, Croatian, Crow, Czech, Dakota, Danish, Dargwa, Dungan, Dutch (Netherlands and Belgium), English, Eskimo (Cyrillic), Eskimo (Latin), Esperanto, Estonian, Even, Evenki, Faroese, Fijian, Finnish, Fortran, French, Frisian, Friulian, Gagauz, Galician, Ganda, German (Luxemburg), German, Greek, Guarani, Hani, Hausa, Hawaiian, Hebrew, Hungarian, Icelandic, Ido, Indonesian, Ingush, Interlingua, Irish, Italian, JAVA, Japanese, Jingpo, Kabardian, Kalmyk, Karachay-balkar, Karakalpak, Kasub, Kawa, Kazakh, Khakass, Khanty, Kikuyu, Kirghiz, Kongo, Koryak, Kpelle, Kumyk, Kurdish, Lak, Latin, Latvian, Lezgi, Lithuanian, Luba, Macedonian, Malagasy, Malay, Malinke, Maltese, Mansy, Maori, Mari, Maya, Miao, Minangkabau, Mohawk, Moldavian, Mongol, Mordvin, Nahuatl, Nenets, Nivkh, Nogay, Norwegian (nynorsk and bokmål), Nyanja, Occidental, Ojibway, Ossetian, Papiamento, Pascal, Polish, Portuguese (Portugal and Brazil), Provencal, Quechua, Rhaeto-romanic, Romanian, Romany, Rundi, Russian, Russian (old spelling), Rwanda, Sami (Lappish), Samoan, Scottish Gaelic, Selkup, Serbian (Cyrillic), Serbian (Latin), Shona, Simple chemical formulas, Slovak, Slovenian, Somali, Sorbian, Sotho, Spanish, Sunda, Swahili, Swazi, Swedish, Tabasaran, Tagalog, Tahitian, Tajik, Tatar, Thai, Tok Pisin, Tongan, Tswana, Tun, Turkish, Turkmen, Tuvinian, Udmurt, Uighur (Cyrillic), Uighur (Latin), Ukrainian, Uzbek (Cyrillic), Uzbek (Latin), Welsh, Wolof, Xhosa,

	Yakut, Zapotec, Zulu
OmniPage 17 2009	Afrikaans, Albanian, Aymara, Basque, Bemba, Blackfoot, Breton, Bugotu, Bulgarian, Byelorussian, Catalan, Chamorro, Chechen, Corsican, Croatian, Crow, Czech, Danish, Dutch, English, Esperanto, Estonian, Faroese, Fijian, Finnish, French, Frisian, Friulian, Gaelic (Irish), Gaelic (Scottish), Galician, Ganda/Luganda, German, Greek, Guarani, Hani, Hawaiian, Hungarian, Icelandic, Ido, Indonesian, Interlingua, Italian, Inuit, Kabardian, Kasub, Kawa, Kikuyu, Kongo, Kpelle, Kurdish, Latin, Latvian, Lithuanian, Luba, Luxembourgian, Macedonian, Malagasy, Malay, Malinke, Maltese, Maori, Mayan, Miao, Minankabaw, Mohawk, Moldavian, Nahuatl, Norwegian, Nyanja, Occidental, Ojibway, Papiamento, Pidgin English, Polish, Portuguese (Brazilian), Portuguese, Provencal, Quechua, Rhaetic, Romanian, Romany, Ruanda, Rundi, Russian, Sami Lule, Sami Northern, Sami Southern, Sami, Samoan, Sardinian, Serbian (Cyrillic), Serbian (Latin), Shona, Sioux, Slovak, Slovenian, Somali, Sorbian, Sotho, Spanish, Sundanese, Swahili, Swazi, Swedish, Tagalog, Tahitian, Tinpo, Tongan,
Readiris 12 Pro & Corporate 2009	American English, British English, Afrikaans, Albanian, Aymara, Balinese, Basque, Bemba, Bikol, Bislama, Brazilian, Breton, Bulgarian, Byelorussian, Catalan, Cebuano, Chamorro, Corsican, Croatian, Czech, Danish, Dutch, Esperanto, Estonian, Faroese, Fijian, Finnish, French, Frisian, Friulian, Galician, Ganda, German, Greek, Greenlandic, Haitian (Creole), Hani, Hiligaynon, Hungarian, Icelandic, Ido, Ilocano, Indonesian, Interlingua, Irish (Gaelic), Italian, Javanese, Kapampangan, Kicongo, Kinyarwanda, Kurdish, Latin, Latvian, Lithuanian, Luxemburgh, Macedonian, Madurese, Malagasy, Malay, Maltese, Manx (Gaelic), Maori, Mayan, Minangkabau, Nahuatl, Norwegian, Numeric, Nyanja, Nynorsk, Occitan, Pidgin English, Polish, Portuguese, Quechua, Rhaeto-Roman, Romanian, Rundi, Russian, Samoan, Sardinian, Scottish (Gaelic), Serbian, Serbian (Latin), Shona, Slovak, Slovenian, Somali, Sotho, Spanish,

	Sundanese, Swahili, Swedish, Tagalog, Tahitian, Tok Pisin, Tonga, Tswana, Turkish, Ukrainian, Waray, Wolof, Xhosa, Zapotec, Zulu, Bulgarian - English, Byelorussian - English, Greek - English, Macedonian - English, Russian - English, Serbian - English, Ukrainian - English + Moldovan, Bosnian (Cyrillic
Readiris 12 Pro & Corporate Middle-East 2009	Arabic, Farsi şi Hebrew
Readiris 12 Pro & Corporate Asian 2009	Simplified Chinese, Traditional Chinese, Japanese şi Korean
CuneiForm 12 2007	English, German, Croatian, Polish, Danish, Portuguese, Dutch, Digits, Czech, French, Romanian, Hungarian, Bulgarian, Slovenian, Lettish, Lithuanian, Estonian, Turkish, Russian, Swedish, Spanish, Italian, Russian-English (mixed), Ukrainian, Serbian
Microsoft Office Document Imaging Office 2007 2007	Accesul la diferite limbi este legat de instalarea unor componente MS Office
NEOPTEC DATA-SCAN 5.7 2009	French, Spanish, English.
NovoDynamicsVERUS Middle East Professional 2005	Arabic, Persian (Farsi, Dari), Pashto, Urdu, inclusiv English and French
NovoDynamicsVERUS Asia Professional 2009	Chineza simplificată şi tradițională , limbile Korean şi Russian, incluzând English
HOCR 0.10.13	Hebrew
OCRopus 0.3.1, 08 ReadSoft Caractere Europene,	Toate limbile suportate de Tesseract prin plug-in-uri, şi suportă Latin script şi English în mod nativ Chineza simplificată şi tradițională, caractere Korean şi Japanese

SimpleOCR 3.5 2008	Engleză, franceză
Tesseract 2.03 2008	Poate recunoaşte 6 limbi, compatibil UTF8, are suport de antrenare

4.5.5. Particularități ale sistemelor OCR actuale

- ***ExperVision***

Sisteme de operare: Windows, Mac OS X, Unix, Linux, OS/2

A câştigat cele mai mari note la testele de performanță făcute de UNLV tehnologiilor OCR.

OpenRTK este o tehnologie pentru clienții OEM sau care au nevoie să integreze OCR-ul în aplicațiile lor (exemplu DIM, FPS).

TypeReader 2008 este un program de recunoaştere a caracterelor creat în special pentru segmentul corporate. Este adresat celor ce au nevoie de transformarea a sute de pagini în documente electronice, dar numai dacă ai nevoie de a face acest lucru rapid şi nu te interesează prea mult calitatea. Viteza este de aproximativ 3 ori mai bună decât cea a lui Abbyy Fine Reader Professional Edition şi de 4 ori mai bună decât a lui OmniPage Professional. La fel ca şi OmniPage Professional, TypeReader include caracteristica "watched folder". Aceasta presupune posibilitatea de a aplica automat procesul de OCR-izare tuturor documentelor ce sunt plasat într-un anumit director din rețea. Viteza de recunoaştere a caracterelor este absolut uimitoare, iar prezența unei astfel de caracteristici face şi mai mare viteza de recunoaştere a caracterelor care era mare. Totuşi, unele asocieri de litere nu sunt recunoscute corespunzător (exemplu "rt" sau "tr").

- ***ABBYY***

Avantajele aplicației Fine Reader OCR 8.0 constau în procesarea facilă a documentelor în format pdf, recunoaşterea pozelor realizate cu camera digitală şi recunoaşterea optică a caracterelor printr-o singură apăsare de buton. Asigurând o capacitate semnificativ îmbunătățită de convertire a imaginilor documentelor în format HTML, FineReader 10 simplifică citirea şi publicarea cărților electronice. Are funcții de

analiza şi păstrare a formatului paginii. Oferă o optimizare a dimensiunii documentelor PDF rezultate, reducându-le de până la 10 ori, păstrând însă calitatea vizuală.

Recognition (ADRTTM) analizează documentul ca o entitate singulară, păstrând toate elementele din structura documentului original, inclusiv textul, tabelele, antetul / subsolul şi numerotarea paginilor. Are dicționare în peste 30 de limbi.

- ***AnyDoc***

Dezvoltă tehnologii pentru a procesa documente structurate, semi-structurate şi nestructurate, de clasificare. Documente structurate, cum ar fi o cerere de credit, folosesc un anumit şablon.

- ***OmniPage***

Converteşte imagini, cum ar fi documentele scanate şi fişierele PDF, în formate de fişiere folosite de aplicații cum ar fi Microsoft Word, Excel, Adobe Acrobat, sau de fişiere HTML.

- ***Readiris***

În 2006, a fost selectat pentru a fi utilizat în Adobe Acrobat 8.0. Softul se află în concurență cu Expervision (TypeReader), ABBYY FineReader , OmniPage, GOCR şi Tesseract.

- ***CuneiForm***

Cuneiform poate transforma texte în format RTF, HTML, text simplu, Microsoft Word sau Microsoft Excel. Poate recunoaşte tabele complicate de orice structură. În 2008 Cognitive Technologies a lansat on-line servicii gratuite de recunoaştere pe penOCR.org.

- ***Puma.NET***

Poate încărca o singura pagină, o recunoaşte şi salvează rezultatul într-un fişier. Poate să returneze rezultatul recunoaşterii într-un şir de caractere. Puma.NET nu poate să încarce pagini multiple şi să salveze rezultatul într-un singur fişier. Nu oferă utilizatorului posibilitatea de a vedea în paralel originalul şi varianta recunoscută.

- ***CVISION Technologies***

Oferă o calitate superioară pentru documentele colorate şi procesează un volum mare de documente.

Localizează rapid un cuvânt într-un document de dimensiuni mari. În plus este optimizat pentru WEB şi permite procesarea PDF-urilor.

- ***OCRopus***

OCRopus este un sistem de ultima oră pentru analiza documentelor şi OCR, ce conține un analizator de layout-uri, un dispozitiv de recunoaştere a caracterelor, modelare statistică a limbajului natural şi capacități multilingvistice.

Motorul OCRopus se bazează pe două proiecte de cercetare: un sistem de recunoaştere dezvoltat în anii 90' de biroul US Census şi pe metode moderne, de mare performanță, de analiza a layouturilor.

Dezvoltarea OCRopus este sponsorizată de către Google. Inițial proiectul a fost gândit ca un ajutor pentru eforturile de conversie masivă a documentelor. Se doreşte de asemenea a fi un excellent sistem OCR pentru multe alte aplicații.

- ***PDF OCR X***

PDF OCR este un program simplu care converteşte un document PDF în document text. Acesta utilizează tehnologie OCR avansată prin care poate să extragă textul din PDF chiar daca textul este conținut intr-o imagine.

Avantaje:

1. Lucrează cu orice fel de PDF chiar dacă este un PDF scanat sau un PDF generat dintr-un anumit document;
2. Interfața drag and drop uşor de folosit;
3. Poate fi utilizat mai multe limbi: Engleză, franceză, germană şi italiană;

- ***Alt-N Technologies- RelayFax Network Fax Manager***

RelayFax network fax server management software automatizează trimiterea, primirea şi managementul faxurilor rețelei din desktopul tau, în timp ce integrează total funcționalitățile faxului cu sistemul de email existent - totul în timp ce furnizează trimiteri şi primiri nelimitate de faxuri, fără nici o taxă lunară sau per-transmitere.

RelayFax se conectează la un număr nelimitat de POP mailboxes, la intervale programate şi colectează mesajele de fax aflate în aşteptare pe care apoi software-ul le trimite prin e-mail, fax sau printează după cum este definit în regulile configurabile ale fax-ului.

Avantaje:

• detectare automată a modemului;

• utilizarea mediului Terminal Services;

• programarea fax-urilor;

• **conversia fax-urilor in PDF sau PNG**;

• tehnologie modernă de trimitere a fax-urilor;

- ***Scantron***

Este o soluție de business care elimină în mod eficient procesele manuale care nu fac diferența între informațiile primite prin fax, mail sau internet.

Avantaje:

1. Reduce timpul şi costul cheltuit pe introducerea manuală a datelor cu până la 90%;
2. Trimiterea reală şi precisă a datelor şi a documentelor;
3. Selecție variată de programe folosite (Visual Basic, C++, VB Script, .NET).

- ***SmartScore***

Acest software ajută să aveți muzică printată într-un mod foarte rapid în computer. Asigură recunoaşterea oricărui format PDF.

- **Tesseract**

Avantaje:

1. Poate identifica dacă textul este proporțional sau monospațiu;
2. Poate fi utilizat în mai multe limbi;
3. Dezavantaje:
4. Este utilizat doar pentru imaginile de format TIFF (asta doar dacă libtiff este instalat);
5. Poate fi utilizat pe Windows, Liniux, Mac OS X, dar din cauza resurselor limitate

dezvoltatorii l-au testat doar pentru Windows si Linux.

- ***ReadSoft***

ReadSoft Documents este o aplicație complexă destinată procesării documentelor de tip structurat şi semistructurat. Este o aplicație completă, adaptabilă oricăror cerințe de business sau volum de documente. Este capabilă de a captura toate formularele de procesat, indiferent de format şi / sau design. Se poate integra uşor cu aplicațiile proprii (SAP, Oracle).

- ***Microsoft Office Document Imaging***

Se recomandă a se utiliza Microsoft Office Document Imaging atunci când se doreşte:

-Scanarea documentelor de o singură pagină sau de mai multe pagini. De exemplu, se pot scana documente pe hârtie pentru arhivare şi reciclarea copiilor pe hârtie;

- Citirea unui document scanat sau a unui fax, rapid şi uşor, pe ecranul computerului. De exemplu, se poate citi în mod conectat un fax pe mai multe pagini;

- Completarea în mod conectat a unui formular simplu care a fost scanat sau primit ca fax. De exemplu, scanați un formular pe hârtie sau deschideți un formular primit ca fax, îl completați cu informațiile solicitate, inclusiv casetele de selectare, apoi returnați formularul prin poşta electronică ;

- Efectuarea recunoaşterii optice a caracterelor (OCR) într-un document scanat sau într-un fax ;

De exemplu, după recunoaşterea textului din documentul scanat sau din fax, se poate căuta un text specific sau se poate copia text în alt program.

- Copierea de text şi imagini dintr-un document scanat sau dintr-un fax şi lipirea lor în orice program Office. Exportul de text şi imagini dintr-un document scanat sau dintr-un fax în Microsoft Word. De exemplu, dace se doreşte copierea de date importante dintr-un fax sau dintr-un document scanat într-o foaie de lucru Microsoft Excel ;

- Căutarea de text într-un document scanat sau într-un fax ;

-Reorganizarea ordinii paginilor într-un document scanat sau fax cu mai multe pagini, la fel de uşor ca rearanjarea hârtiilor într-un dosar. De exemplu, eliminați pagina

de însoțire a unui fax sau îi adăugați pagini suplimentare înainte de a-l trimite la altă persoană ;

-Trimiterea de documente scanate către alții, prin poştă electronică sau ca fax, prin Internet ;

- Adnotarea unui document scanat sau a unui fax şi partajarea lor cu altă persoană. De exemplu, adăugarea de comentarii unui fax şi returnarea lui la expeditor.

Microsoft Office Document Imaging are două componente — una de scanare şi una pentru lucrul cu imagini.

Componenta de scanare permite scanarea documentelor pentru a le face disponibile în computer, utilizând orice scaner instalat. Această componentă furnizează setări prestabilite de scanare pentru controlul scanerului utilizând setări optimizate pentru scopuri specifice. De exemplu, setarea prestabilită de scanare Alb-negru dă cele mai bune rezultate la OCR când scanați pagini de text pentru OCR, în timp ce setarea prestabilită de scanare Color este cea mai bună pentru a scana imagini colorate sau lucrări de artă. OCR este efectuat automat pe documente text imediat după scanare şi se pot scana uşor pagini multiple într-un singur fişier.

Componenta de imagine facilitează vizualizarea pe ecran a documentelor scanate, rearanjează documentele multipagină, selectează şi manevrează textul ecunoscut, adnotează documentele scanate şi fax-urile electronice şi trimite documente către alții prin poştă electronică sau fax.

- ***Fine Reader 7.0***

Deoarece există o multitudine de documente în format clasic, pe hârtie, care trebuie şi ele aduse în formă digitală este foarte important procesul de OCR

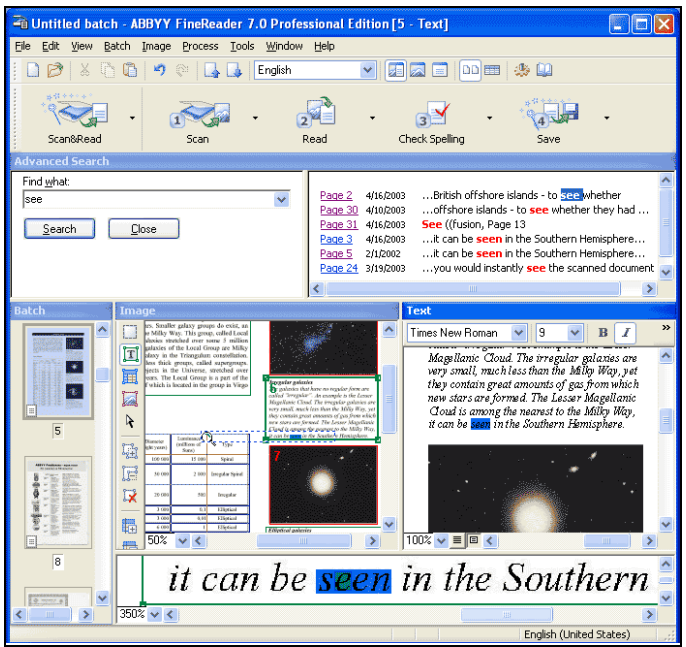


Fig.4.3. Utilitarul FineReader 7.0 de la ABBYY

(Optical Caracter Recognition), în cele ce urmează prezentând unul din cele mai de succes utilitare în acest sens - FineReader.

Datorită faptului că utilizează tehnologia IPA, FineReader permite recunoaşterea foarte bună a conținutului. Renumit deja datorită replicărilor perfecte pentru proiecte complexe şi pentru tabele şi documente de calitate slabă, acum prezintă îmbunătățiri suplimentare. Sunt proiectate proceduri noi de filtrare 'Intelligent background' pentru filtrarea fundalului color, zgomotelor şi altor interferențe şi permite recunoaşterea şi mai corectă a unor documente care până acum constituiau o adevărată provocare pentru utilitarele de OCR-izare, precum: articole cu un fundal foarte colorat, texte tipărite peste imagini, sau urme lăsate de agrafe. Acum nu mai are importanță ce fel de documente se scanează, fie ele poze multi-coloană, texte încadrate, sau prețuri dintr-un ziar, pentru că FineReader asigură acuratețea necesară.

Se pot recunoaşte texte scrise în 177 de limbi, inclusiv engleza, germana, franceza, spaniola, italiana, portugheza, olandeza, suedeza, finlandeza, rusa, ucraineana, bulgara, ceha, maghiara, poloneza, slovaca, malaeziana, indoneziana şi altele. Verificarea ortografiei este disponibilă pentru 34 de limbi. Acest fapt simplifică procesul de introducere a datelor pentru oameni aparținând diferitelor țări şi culturi.

Suportă exportul către aplicațiile uzuale de birou, inclusiv Microsoft Word, Microsoft Excel, Microsoft PowerPoint, Lotus Word Pro, Corel WordPerfect şi Sun StarWriter. Textul recunoscut poate fi salvat intr-o varietate de formate ale fişierelor, inclusiv PDF (cu şase opțiuni de salvare), HTML, Microsoft Word XML, DOC, RTF, XLS, PPT, DBF, CSV si TXT. Se pot converti cu uşurință documentele în format XML, utilizând Microsoft Word XML, un standard în devenire pentru utilizarea fişierelor. Acesta permite utilizatorilor să beneficieze de avantajele fişierelor in format XML,

putând fi manipulat şi căutat prin utilizarea oricărui program care prelucrează standardul XML. Cu ajutorul suportului XML, FineReader 7.0 simplifică procesul de citire şi editare a documentelor word. În momentul exportului documentelor în Microsoft Office Word 2003, FineReader deschide în mod automat imaginea inserată a documentului original, permițând utilizatorilor să editeze şi să vadă în mod simultan documentele originale, eliminându-se astfel necesitatea de comutare între cele două aplicații.

Se pot deschide, citi şi edita fişiere în PDF, care pot fi salvate în orice format sau poate exporta datele către aplicația favorită. Toate fişierele PDF create în FineReader sunt optimizate pentru publicarea pe Web. FineReader suportă exportul în oricare dintre cele patru formate standard ale fişierelor PDF (numai text şi imagini, text suprapus imaginii, text sub imagine, şi numai imagine), cu controale suplimentare. Aceasta oferă utilizatorilor şase opțiuni de salvare în PDF.

Este asigurat suportul pentru periferice multi-funcționale care combină funcționalitățile scanerului, ale imprimantei, fotocopiatorului şi faxului, atât conectate la o stație de lucru, cât şi în rețea. Setările speciale ale programului permit utilizatorilor să deschidă în mod automat imagini scanate, din orice locație a rețelei sau de la un server FTP şi să le recunoască.

5. Elaborarea unui model conceptual pentru un sistem de informare şi documentare

5.1. Depozite digitale

Resursele digitale aflate în permanenţă dezvoltare formează baza capitalului intelectual în cercetare. Acestea trebuie să persiste şi să poată fi căutate, accesate şi înțelese. Reutilizarea datelor (de către utilizatori din domenii diferite), trebuie să fie posibilă imediat ce acestea au fost create cât şi pentru o perioadă mai mare de timp. Aceleaşi tehnici de prezervare a datelor vor susține formarea depozitelor actuale cât şi utilizarea lor ulterioară. Există riscul însă ca multe dintre datele ştiințifice şi tehnice să se piardă pentru generațiile viitoare dacă nu sunt prezervate corespunzător.

Ceretările în domeniu sunt concentrate pe prezervarea informațiilor digitale în ştiință, pornind de la datele primare până la analizarea publicațiilor finale rezultată din cercetare. Este important a dezvolta o hartă (roadmap) şi recomandări pentru dezvoltarea unei e-infrastructuri cu scopul de a menține accesibilitatea informației pe termen lung şi utilitatea informației ştiințifice digitale în cele mai importante depozite.

Pentru a realiza o infrastructură care să ofere acces pe termen lung, de multe ori se aplelează la un sistem deschis de tipul Open Archival Information System (OAIS) Reference Model (ISO 14721). Termenul de acces pe termen lung defineşte o perioadă de timp suficient de lungă care să vizeze impactul produs de schimbarea de tehnologie, inclusiv suportul pentru o nouă media şi noi formate de date.

Paşii parcurşi în e-Ştiință a modificat dramatic procesul de cercetare. Ciclul de scriere şi citire a publicațiilor în secolul trecut care utiliza un singur mediu pentru schimbul de informații a evoluat către o multitudine de resurse digitale. Aceste noi oportunități promovează cercetarea multidisciplinară şi permit reutilizarea rapidă a

informațiilor. Pentru a avea insă acces la cercetări este necesar a întreprinde eforturi coerente şi concrete pentru a prezerva înregistrările digitale.

5.1.1. Organizații implicate în prezervarea conținutului digital

Alianța Europeană pentru Accesul Permanent (European Alliance for Permanent Access) (http://www.alliancepermanentaccess.eu/) Alianța este o coaliție unică trans-sectorială a majorității deținătorilor de informație din știință. Printre membrii alianței se numără: European Science Foundation, Centre National d'Etudes Spatiales, Centre Informatique National de l'Enseignement Supérieur, Joint Information Systems Committee UK, the British Library şi National Archives din Suedia.

Practic, membrii alianței sunt deja activi într-un număr relevant de proiecte de prezervare pe termen lung şi acces cum ar fi CASPAR, DRIVER, PLASNETS, SHAMAN şi DPE. Implicarea partenerilor comerciali în unele dintre aceste proiecte cum ar fi IBM, Microsoft şi SUN este benefică şi vor mări sfera de influență în domeniu a alianței.

Membrii alianței sunt implicați deasemenea şi în dezvoltarea unei noi generații a depozitelor digitale ştiințifice, care iau în considerare eforturile europene de construire a unei e-infrastructuri cum ar fi GENESI-DR, EuroVO-AIDA, acoperind ştiința pământului şi comunităție în astrofizică. Totuşi, prin definiție, proiectele au limite în ceea ce priveşte focalizarea pe un domeniu şi timpul alocat. Alianța Europeană pentru Accesul Permanent îşi propune să furnizeze un cadru conceptual şi strategic care să organizeze şi consolideze eforturile individuale şi să umple eventualele goluri dintre acestea. Ar trebui să se finalizeze o abordare unitară europeană şi în acelasi timp să se încerce o coroborare şi cu alte proiecte din afara Europei.

De aceea alianța a adoptat un plan de lucru strategic pentru a ajuta consolidarea rolului său la nivel național European şi internațional. Planul de lucru va stabili acțiunile care vor ajuta la generarea masei critice de utilizatori şi va accelara progresul pe termen mediu (2008-2010).

Este general recunoscut că un răspuns la provocările prezervării digitale necesită abordări coerente pentru toate tipurile de informație codată digital. Diversitatea comunităților şi a practicilor riscă o muliplicare a soluțiilor parțiale, favorizând aparația unor neconcordanțe.

Accesul permanent trebuie văzut ca fiind ca parte a unei activități de creare a unui depozit digital. Se impune o viziune coordonatoare asupra costurilor, beneficiilor, standardelor şi tehnicilor care pot deriva de aici. Managementul de risk trebuie să se bazeze pe o abordare validă a viitorului.

5.1.2.Preocupări actuale pentru implementarea depozitelor digitale instituționale în universitățile din România

În cadrul unei universități, misiunea asigurării cu material documentar revine bibliotecii. Sistemul național de biblioteci din România cuprinde în cazul bibliotecilor universitare, biblioteci subordonate direct Ministerului Educației, Cercetării şi Inovării, cele patru biblioteci centrale universitare din Bucureşti, Iaşi, Cluj şi Timişoara şi celelalte biblioteci universitare subordonate senatelor universităților.

Dacă în cazul bibliotecilor centrale universitare situația financiară permite achiziții de material documentar şi de baze de date, în cazul bibliotecilor universitare subordonate universităților situația nu este la fel de bună.

În practica achiziției partajate, a celorlalte țări, funcționează consorțiile. Prin aceste consorții se negociază prețuri accesibile la baze de date pentru toată comunitatea academică. Deşi prețurile la publicațiile electronice cresc an de an, „criza jurnalelor” se simte acut în nevoile de documentare ale universităților. Nici până acum nu există în România un consorțiu al bibliotecilor universitare care să funcționeze. Consiliul Național al Cercetării Ştiințifice din Învățământul Superior a achiziționat pentru întreaga comunitate academică românească baze de date importante pentru dezvoltarea cercetării ştiințifice în anii 2007-2008.

În Universitatea Transilvania, biblioteca universitară oferă acces la trei baze de date: Springerlink, Mathscinet şi Forest Science Database. Din anul 2004 biblioteca nu mai posedă abonament pe suport tradițional la nici o publicație străină, iar cele trei baze de date nu acoperă nevoile de informare din toate domeniile. O nemulțumire persistă în faptul ca aceste produse pot fi accesate numai de pe rețeaua universității. Majoritatea doresc ca accesul să sa facă de acasă. Poate de aceea rezultatele cercetării calitative privind atitudinea cadrelor didactice arată că în proporție de 48,1% accesează bazele de

date de câteva ori pe an.

La nivelul Uniunii Europene, sistemul de editare ştiinţifică este văzut ca un element fundamental al sistemului European de cercetare. Referitor la noua paradigmă a comunicării academice şi anume mişcarea „acces deschis" în România există puţine iniţiative.

La Universitatea_Transilvania, publicaţiile proprii: Buletinul Ştiinţific şi alte trei reviste: Jurnal Medical Braşovean, Recent şi PRO Ligno permit accesul la nivel de rezumat, dar nu sunt indexate în Google Scholar. Vizibilitatea şi impactul cercetărilor cadrelor didactice este mic considerând indicatorii scientometrici calitativi.

Statisticile internaţionale reflectă ştiinţa românească într-o lumină nefavorabilă. Producţia ştiinţifică a României raportată anual, numărul de articole la 1 milion de locuitori este un indicator slab. Nici situaţia României în Europa Centrală nu este mai bună.

Universitatea Transilvania a făcut eforturi susţinute şi în anul 2004 este poziţionată pe locul 13 în Carta Albă a Cercetării. Începerea construcţiei **Institutului de Cercetare-Dezvoltare-Inovare: Produse High-Tech pentru Dezvoltare Durabilă,** indică preocupări deosebite de dezvoltare a cercetării ştiinţifice.

Analiza impactului citărilor pentru profesorii universităţii nu indică performanţe remarcabile. Implementarea unui depozit digital ar putea fi soluţia ideală pentru creşterea vizibilităţii producţiei ştiinţifice a universităţii.

Universitatea Transilvania ar putea fi pionierul depozitelor digitale instituţionale din România.

5.2. Elaborarea modelului conceptual al sistemelor de informare şi documentare cu conţinut tehnic. Rezultate obţinute în cadrul Universităţii "Transilvania" din Braşov

Pe baza informaţiilor obţinute în urma documentării asupra stadiului actual de dezvoltare a depozitelor digitale instituţionale în lume, a modului lor de implementare şi a cerinţelor impuse de cercetarea cantitativă efectuată în cadrul comunităţii academice privind dezvoltarea unui depozit digital în Universitatea Transilvania, se va elabora o strategie de marketing al cărei obiectiv va fi creşterea vizibilităţii cercetării ştiinţifice româneşti.

5.3. Implementarea unui depozit instituțional digital la Universitatea TRANSILVANIA din Braşov

Membrii comunității academice din Universitatea Transilvania sunt preocupați de dezvoltarea cercetării ştiințifice şi participă la toate competițiile naționale şi europene lansate pentru accesarea fondurilor destinate cercetării. Numărul de proiecte câştigate este mare, iar producția ştiințifică a universității este remarcabilă. Acest mediu academic este unul în care, cu succes, se poate implementa un depozit digital.

5.3.1. Misiunea şi obiectivele depozitului digital

Depozitul digital este un produs al tehnologiei informației pe care universitatea îl oferă membrilor comunității sale pentru gestionarea şi difuzarea materialelor digitale create de către instituție şi membrii comunității.

Acest produs este în esență un angajament organizatoric privind gestionarea materialelor produse de universitate, organizarea, accesul şi distribuția lor dar şi conservarea pe termen lung.

Prin acest produs se vor colecta şi distribui toate documentele ştiințifice create de membrii universității.

Obiectivul principal al depozitului digital este creşterea impactului cercetării ştiințifice a universității prin promovarea producției ştiințifice a universității prin depozitul digital instituțional şi, în acest mod, vizibilitatea ei pe plan mondial şi european şi valorificarea pentru creşterea aportului românesc la dezvoltarea ştiințifică şi tehnologică.

Echipa de implementare a depozitului digital

Pentru asigurarea implicării persoanelor specializate şi asigurarea unui parteneriat performant se propune ca următoarele departamente şi direcții de cercetare să realizeze pentru început un **DEPOZIT DIGITAL PILOT**.

1. **Departamentul de informatizare**

Departamentul a fost implicat în multe proiecte universitare, dintre care proiectele de dezvoltare tehnologică, proiecte software (Transilvania):

- **AGSIS –** Aplicația de gestiune a studenților şi informațiilor şcolare.
- **Admitere**, Utilizarea aplicației de admitere în întreaga universitate, pentru toate

domeniile ştiinţifice şi pentru toate formele de admitere a candidaţilor la nivel de licenţă, master şi şcoală doctorală: admitere prin lucrări scrise, probe practice, teste grilă, concurs de dosare etc.

- **FRACS**, extindere aplicaţie pentru a corespunde noilor cerinţe de raportare din universitate şi de la Consiliul Naţional al Cercetării Ştiinţifice din Învăţământul Superior a activităţii de cercetare ştiinţifică a tuturor membrilor comunităţii academice
- **Portal.unitbv.ro**, pentru studenţi: registratură electronică – gestiunea electronică online a cererilor şi adeverinţelor cu transferul automat în AGSIS; securitate bazată pe criptare şi coduri de bare
- **Site-ul universităţii**
- Introducerea unui suport de tip **Single Sign On** pentru autentificarea şi autorizarea accesului în reţeaua universităţii pentru cadre didactice şi studenţi, cu acces ulterior la toate aplicaţiile dezvoltate – colaborare cu Departamentul de Infrastructură şi Tehnologie care este în curs de realizare şi a cărui finalizare este prevăzută în 2010.
- Introducerea **semnăturii electronice** pentru toate cadrele didactice din universitate cu finalizare în 2011.

Considerăm că acest departament este cel mai potrivit în a lansa această aplicaţie electronică. Departamentul are expertiza dovedită în implementarea platformei elearning utilizând Moodle (http://moodle.org/), care este un sistem de administrare a cursurilor (CMS – Course Management System), şi are o licenţă de utilizare gratuită („open source"), cu peste 20000 instalări în 171 de ţări, un excelent instrument destinat tutorilor, implicaţi în crearea cursurilor „online" şi administrarea claselor de studenţi.

2. Platforma de cercetare interdisciplinară ASPECKT, Platforma/Laborator de **A**nalize **S**tatistice şi **P**reviziune a fenomenelor **E**conomico-sociale şi **C**ercetări de mar**K**e**T**ing – **ASPECKT.** În cadrul acestei platforme integrate de cercetare ştiinţifică şi dezvoltare aplicativă există condiţiile obiective şi subiective necesare pentru implementarea depozitului digital. Colectivul de cercetare implicat în această platformă va participa la implementarea DEPOZITULUI DIGITAL PILOT.

3. Departamentul de cercetare Produse mecanice de înaltă precizie şi sisteme mecatronice, cu direcția Digitizare şi tratament de imagine

Membrii acestui departament vor participa la implementarea DEPOZITULUI DIGITAL PILOT.

4. Departamentul de cercetare Legislație şi proprietate intelectuală, cu direcția Legislație şi dezvoltare durabilă.

Membrii acestui departament vor asigura consultanță cu privire la dreptul de proprietate a autorilor şi dreptul de a arhiva documente in DEPOZITUL DIGITAL PILOT.

Fig. 5.1. Echipa de implementare a DEPOZITULUI DIGITAL PILOT, ASPECKT-Dspace.

Echipa de implementare (fig. 5.1.) va fi un parteneriat între:

- Facultatea de ştiințe economice;
- Facultatea de inginerie mecanică;
- Facultatea de drept.

Criteriile de selecție a echipei de implementare sunt:

- Departamente prietenoase în raport cu misiunea DEPOZITULUI DIGITAL

PILOT;

- Diversitate în zonele disciplinelor;
- Diversitate de conținut sau formate;
- Exemple de gestiune a diferitelor forme de proprietate intelectuală;
- Arhivarea unor colecții de dimensiuni mici pentru început;
- Echipă cu legături puternice şi încredere.

Membrii echipei au lucrat şi în alte proiecte. Fiecare prezintă expertiză în activitățile desfăşurate.

Planul de dezvoltare al produsului DEPOZIT DIGITAL PILOT

Planul va cuprinde următoarele acțiuni:

- Identificarea echipei de service;
- Identificarea primilor inițiatori care doresc să îşi arhiveze cercetările în depozitul digital;
- Identificarea unor colecții deja existente care să fie arhivate;
- Identificarea conținutului nou pentru arhivare;
- Identificarea conducătorilor;
- Dezvoltarea politicilor;
- Dezvoltarea structurii consultative;
- Identificarea personalului academic implicat în echipa proiectului;
- Identificarea personalului neacademic implicat în echipa proiectului.

Resursele umane pentru service şi asistență vor forma două colective:

- Comitet de asistență;
- Comitet pentru web-design,

care vor avea drept ținte ale activității lor:

- Definirea colecțiilor ce vor fi arhivate;
- Definirea fluxurilor de lucru - în Dspace;
- Arhivarea colecțiilor deja existente;
- Sprijinirea utilizatorilor depozitului prin oferirea de informații telefonice sau (pe o linie) on-line;
- Crearea de pagini cu întrebările cele mai frecvente.

Producția ştiințifică a universității cuprinde: dizertațiile masteranzilor, tezele de

doctorat, lucrările ştiinţifice elaborate de cadrele didactice (Figura 5.2.). Lucrările elaborate de cadrele didactice sunt lucrări publicate în volumele conferinţelor, lucrări publicate în reviste, cărţi (Figura.5.3.).

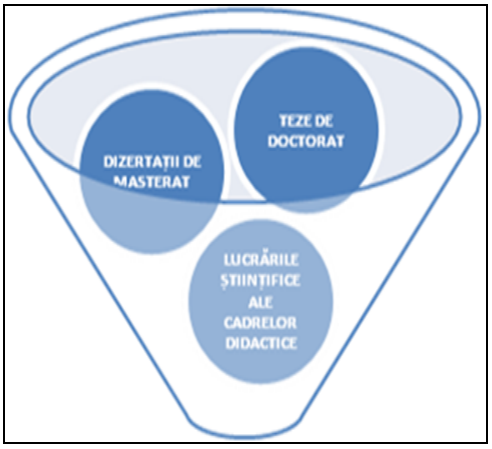


Figura 5.2: Producţia ştiinţifică a universităţii.

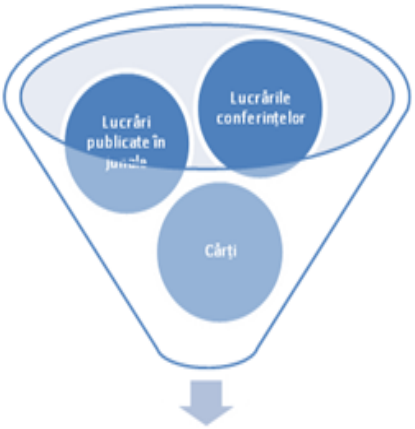


Figura 5.3.: Lucrările publicate de cadrele didactice.

5.3.2. Tehnologii pentru alegerea platformei şi a softului

În urma unei analize tehnice se consideră ca Dspace este soluţia sursei deschise pe care se va implementa DEPOZITUL PILOT în universitatea Transilvania.

Platforma de găzduire va fi platforma ASPECKT a Facultăţii de Ştiinţe Economice:

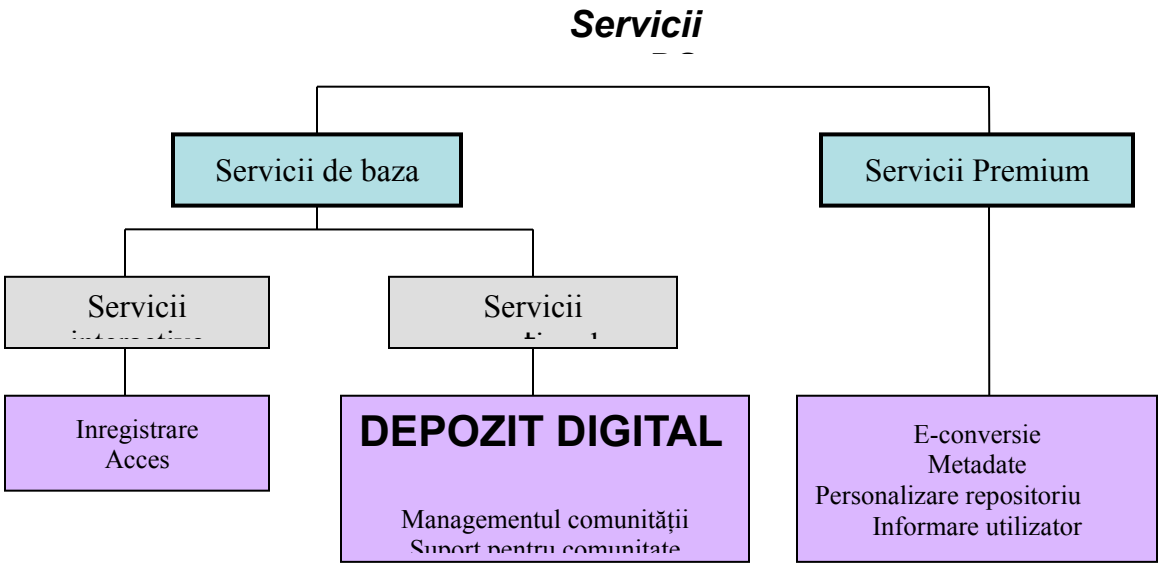


Figura 5.4. Serviciile oferite de platforma Dspace. Sursa (DSpace).

http://aspeckt.unitbv.ro/dspace/ în care DSpace a fost deja instalat.

Serviciile de bază oferite sunt:

- Managementul arhivării pentru a asigura o conservare pe termen lung,
- Stocare persistentă, inclusiv în caz de back-up şi proceduri de recuperare,
- Atribuie fiecărui document un identificator unic, care să fie citat.

Dspace este softul tip sursă deschisă disponibil gratuit utilizat de cea mai mare parte a depozitelor instituționale. Oferă servicii de bază dar şi servicii suplimentare premium, fig. 5.4.

Conținutul depozitului digital

Organizarea conținutului depozitului digital se poate face pe departamente de cercetare sau pe facultăți, la nivelul cărora se va lua şi decizia asupra tipului de materiale ce vor fi depozitate.

Strategii de achiziție a lucrărilor ce vor fi arhivate în depozit sunt:

- inițierea echipei academice în auto-arhivare;
- redactare de instrucțiuni pentru ca autorii să-şi arhiveze singuri articolele;
- redactare de instrucțiuni cu privire la managementul dreptului de autor, documentația de solicitare a permisiunii unui editor, în mod special, permisiunea în numele autorului;
- digitizarea copiilor pe suport tradițional este realizată de o echipă instruită;
- verificarea politicilor editoriale , a dreptului de autoarhivare.

Procedura de achiziție presupune următoarele etape:

- Autorii trimit pe email publicațiile dorite pentru arhivare.
- Echipa proiectului verifică dreptul de proprietate intelectuală şi acordă dreptul de arhivare.
- Autorul îşi arhivează on-line publicațiile.

Dezvoltarea politicilor de achiziție, distribuție şi menținere se face prin:

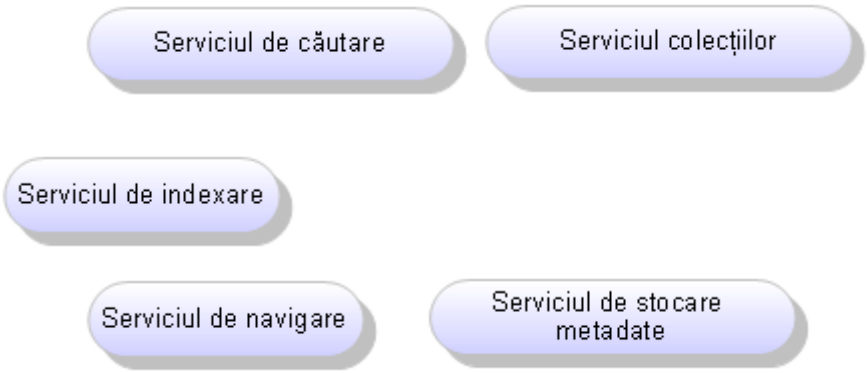
- Politicile de decizie se referă la conținutul depozitului. Trebuie stabilit ce documente se acceptă pentru a fi arhivate, cine poate arhiva documentele, cine va

oferi metadatele. Se vor introduce şi proceduri ce au în vedere calitatea lucrărilor: practic o comisie peer-review care se va activa in principal în cazul proiectelor de diplomă şi care va avea obligația de a verifica şi dacă proiectul nu este plagiat.

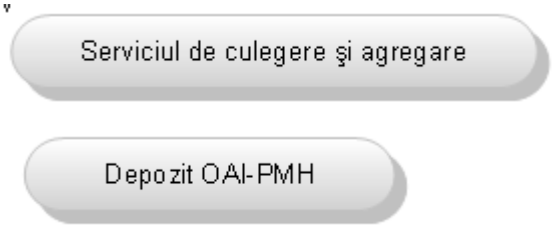
Managementul colecțiilor şi a spațiului informațional se face ținând cont de serviciile oferite. Depozitul digital va oferi servicii de management al colecțiilor, de căutare, de indexare şi stocare a documentelor:

- La nivelul de prezentare avem serviciul interfeței cu utilizatorul şi cel de publicare.

- Managementul spațiului informațional se face ca la orice bibliotecă digitală:

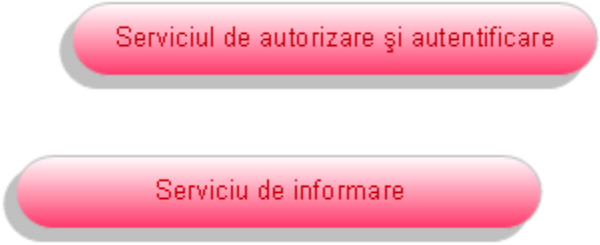


- La nivelul colectiv avem serviciul de culegere şi agregare:



Serviciul de bază este acela de depozitare, stocare a documentelor.

Managementul sistemului funcționează cu un serviciu de autorizare şi autentificare, serviciul de informare şi serviciul de management:



Serviciu de management

Depozitul digital este diferențiat de alte colecții digitale prin următoarele caracteristici:

- conținutul este depozitat într-un depozit digital fie de către creator, deținător sau de terți;
- arhitectura depozitului destionează deopotrivă conținutul şi metadatele;
- depozitul oferă un set minim de servicii de bază, de exemplu introducere, extragere, căutare, controlul accesului;
- depozitul trebuie să fie sustenabil şi de încredere, bine susținut şi bine gestionat.

Serviciile cheie pe care depozitele digitale trebuie să le acopere în câteva zone funcționale sunt (fig. 5.5):

- Acces crescut la resurse;
- Noi modalități de peer-review şi de publicare;
- Managementul informațiilor la nivel corporatist (managementul înregistrărilor şi sistemele de management ale conținutului);
- Partajarea datelor (reutilizarea datelor din cercetare, reutilizarea obiectelor didactice);
- Prezervarea resurselor digitale.

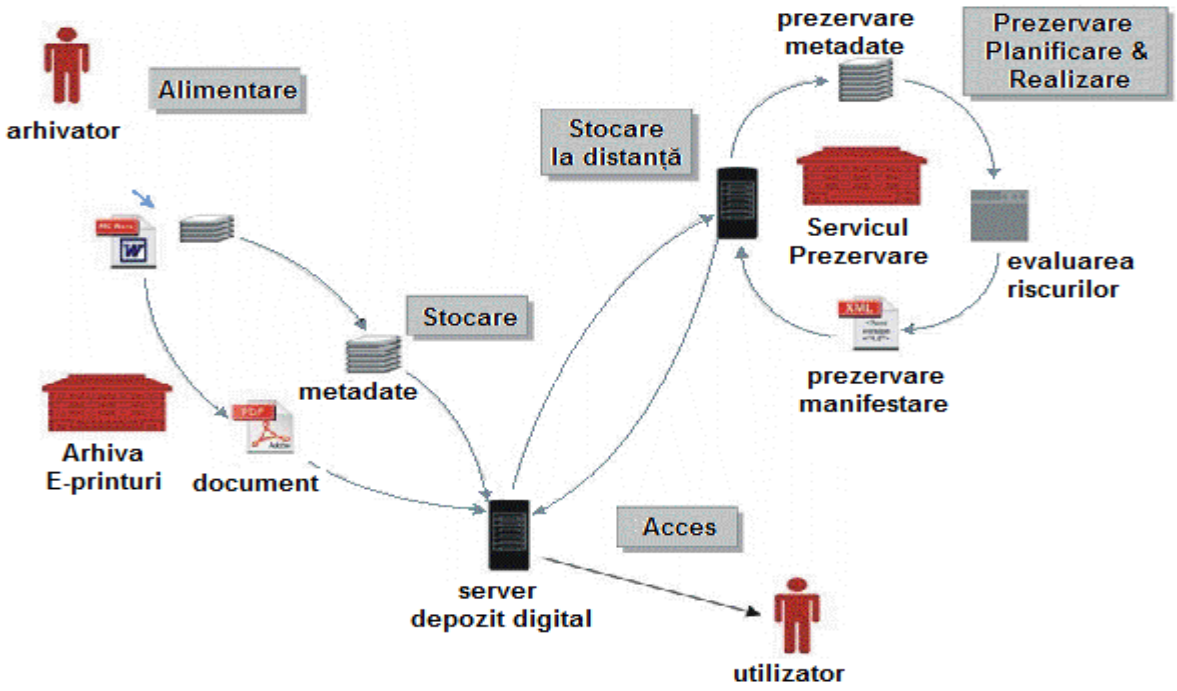


Fig. 5.5: Schema de funcționare a depozitului digital. Sursa (Kosson, 2009).

Depozitele digitale constituie de fapt un complex de servicii gestionate pentru a servi un scop şi un anumit segment de utilizatori. Nu există o definiție care să stabilească distinct granițele a ceea ce este un depozit digital.

Lansarea depozitului digital

Produsul lansat este **DEPOZITUL DIGITAL PILOT** la Universitatea Transilvania **- ASPECKT-Dspace.** Lansarea acestui produs este concepută ca funcționând prin intermediul următoarelor etape:

- Trimiterea de materiale publicitare la toate departamentele de cercetare ;
- **Strategia de promovare a depozitului digital ;**
- **Strategia de profilare**. Această strategie este concepută pentru a convinge cercetătorii de beneficiile depozitului. Comunicarea informațiilor privind avantajele şi beneficiile acestui serviciu sunt transmise utilizatorilor potențiali prin curier electronic al cărui conținut trebuie să evidențieze anumite elemente de bază, prezentate în exemplul de mai jos.

MODEL DE EMAIL TRANSMIS DEPARTAMENTELOR DE CERCETARE

- **Ce este ASPECKT-Dspace** : http://aspeckt.unitbv.ro/dspace/
- Serviciul online dezvoltat arhivează materiale de cercetare publicate de membrii universității Transilvania;
- Aceste documente sunt disponibile de oriunde on-line şi pentru oricine;
- Prin arhivarea cercetărilor în acest depozit, produsele ştiințifice vor putea fi accesate pe o scară mai largă ducând la o mai mare vizibilitate şi impact;
- Depozite similare cu ASPEKT-DSpace sunt dezvoltate în toată lumea;
- Documentele încărcate în Dspace sunt indexate în Google Scholar;
- Depozitul este un instrument care deschide accesul către cercetările din cadrul universității şi le face disponibile pe scară largă şi în întreaga lume;
- Nu este un mecanism de publicare sau susținut pentru reviste ci găzduieşte materiale deja publicate;
- E destinat deschiderii literaturii ştiințifice de cercetare în spiritul accesului liber la informație.

S-au stabilit comunitățile depozitului, adică tipurile de materiale ce se vor încărca în depozit şi anume: Articole ştiinţifice, dizertaţii ale masteranzilor, lucrări studenţeşti, proiecte de diplomă, rapoarte de cercetare ale doctoranzilor, teze de doctorat (Figura.5.6).

In Figura 5.6. se prezintă cazul căutării după autor.

Intr-o primă fază, se încarcă dizertaţiile masteranzilor de la masterul de cercetare din domeniul marketing, şi deci, autorii sunt masteranzii 2009. Valorificând acest exemplu, vor putea fi accesate, după acelaşi criteriu – acela al căutării după autor şi alte categorii de materiale care au fost încărcate în aceeaşi bază de documentare şi care se adresează unui anumit grup ţintă vizat: lucrări studenţeşti, proiecte de diplomă, rapoarte de cercetare.

În Figura 5.7. se prezintă lista completă a autorilor din categoria masteranzi la masterul de cercetare din domeniul marketing şi ale căror lucrări pot fi accesate direct. Autorii sunt clasificaţi în mod ascendent sau descendent şi, de asemenea, se poate opta pentru un anumit număr de rezultate ce sunt afişate pe o pagină.

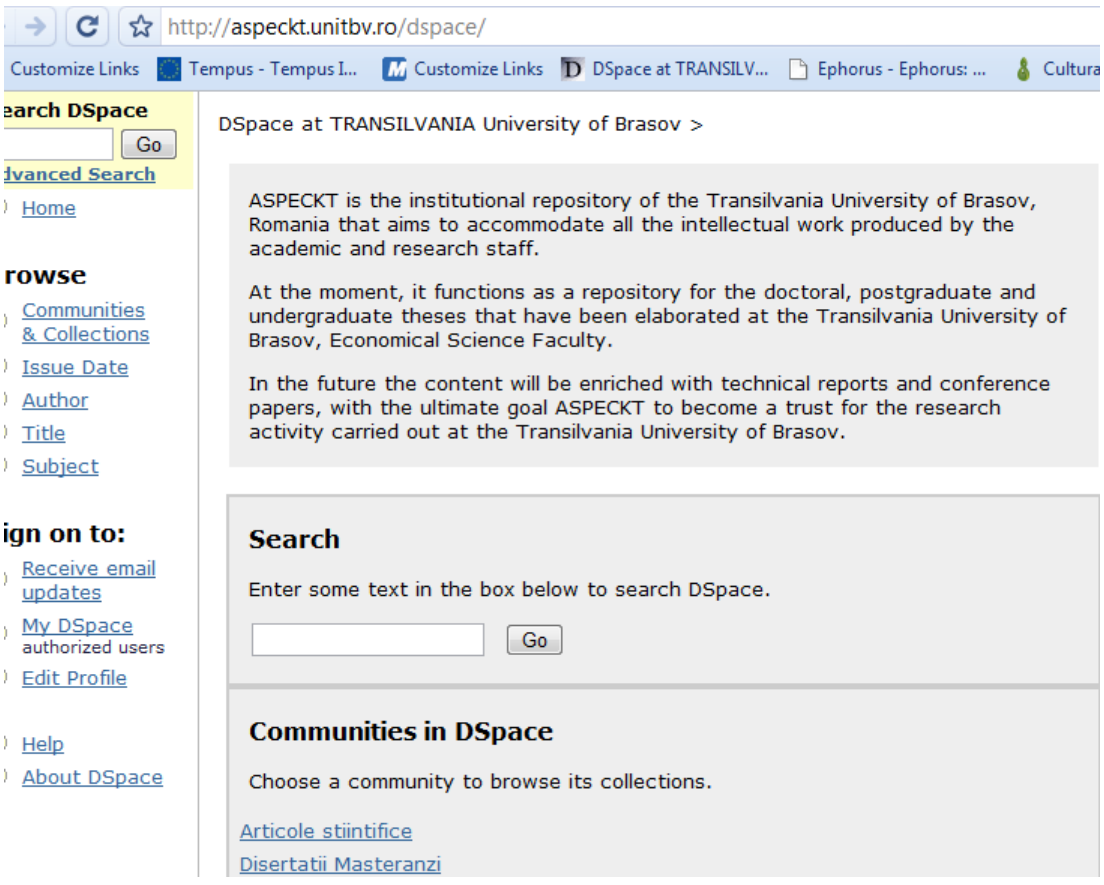


Figura 5.6: Imagine din pagina web a depozitului digital ASPECKT-Dspace.

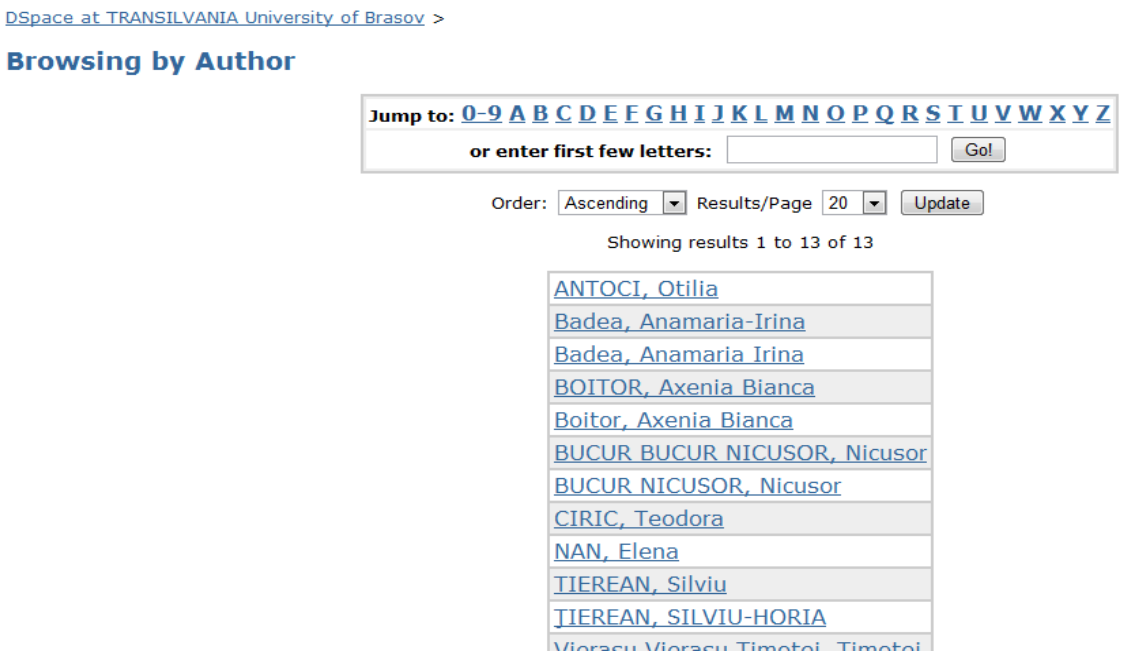
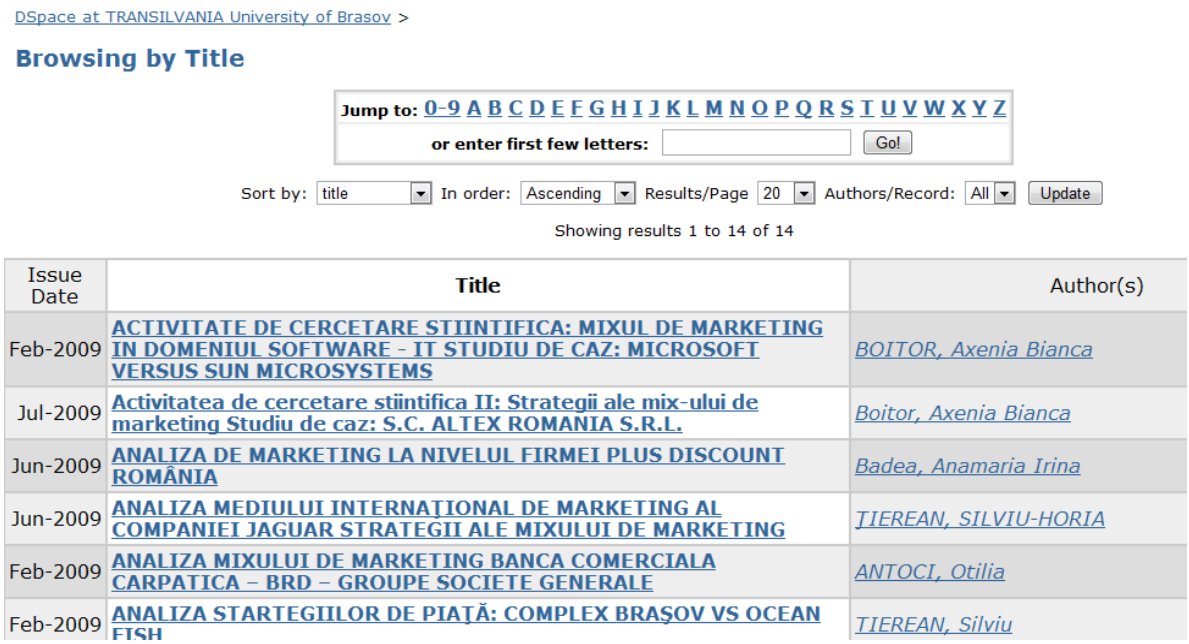


Figura 7: Structura depozitului pentru dizertaţiile de masterat, căutare după autor.

În acelaşi mod se poate efectua căutarea şi după titlul lucrării. În figura 5.8. este prezentat conţinutul depozitului. Avem o imagine completă a documentelor, titlu, autor, data depozitării.



Issue Date	Title	Author(s)
Feb-2009	ACTIVITATE DE CERCETARE STIINTIFICA: MIXUL DE MARKETING IN DOMENIUL SOFTWARE - IT STUDIU DE CAZ: MICROSOFT VERSUS SUN MICROSYSTEMS	*BOITOR, Axenia Bianca*
Jul-2009	Activitatea de cercetare stiintifica II: Strategii ale mix-ului de marketing Studiu de caz: S.C. ALTEX ROMANIA S.R.L.	*Boitor, Axenia Bianca*
Jun-2009	ANALIZA DE MARKETING LA NIVELUL FIRMEI PLUS DISCOUNT ROMÂNIA	*Badea, Anamaria Irina*
Jun-2009	ANALIZA MEDIULUI INTERNAŢIONAL DE MARKETING AL COMPANIEI JAGUAR STRATEGII ALE MIXULUI DE MARKETING	*ŢIEREAN, SILVIU-HORIA*
Feb-2009	ANALIZA MIXULUI DE MARKETING BANCA COMERCIALA CARPATICA – BRD – GROUPE SOCIETE GENERALE	*ANTOCI, Otilia*
Feb-2009	ANALIZA STARTEGIILOR DE PIAŢĂ: COMPLEX BRAŞOV VS OCEAN FISH	*TIEREAN, Silviu*

Figura 5.8.: Structura depozitului după încărcarea dizertaţiilor de masterat, căutare după

titlu.

Formatul prevăzut pentru încărcare este formatul .pdf deoarece el oferă mai multe avantaje, printre care cele mai importante ar fi cele legate de volumul mic al memoriei ocupate şi de facilităţi foarte accesibile de securizare la copiere şi imprimare, acces care ar putea rămâne la nivelul de decizie al autorilor.

DSpace odată instalat este foarte uşor de utilizat. În anexa 8 sunt prezentaţi toţi paşii necesari pentru arhivarea documentelor.

Strategii de promovare a depozitului digital

Strategiile de promovare a depozitului digital se pot încadra în mai multe categorii: de atracţie, de împingere şi de consultare care pot fi implementate fie prin comunicare digitală – e-mail, internet etc., fie prin comunicare directă, bilaterală.

A. Strategia de atracţie. Este strategia de a face depozitul atractiv pentru contribuabili potenţiali. În acest scop un e-mail de antamare ar trebui să cuprindă mai multe elemente de teoria convingerii, ca în modelul de mai jos, ce pun în evidenţă avantajele promovate de depozitul digital.

MODEL DE EMAIL DE TRANSMIS MEMBRILOR COMUNITĂŢII pentru atragerea membrilor comunităţii să arhiveze lucrările ştiinţifice în depozit.

Beneficii pentru cercetătorii care îşi arhivează lucrările ştiinţifice în ASPEKT-Dspace:

- Cercetările sunt puse la dispoziţia umanităţii pe o scară imposibil de realizat cu documentele tradiţionale;
- Face documentele disponibile dispersându-le între mii de publicaţii academice dispersate în multe discipline;
- Documentele sunt indexate de motoarele Google şi Yahoo;
- Cercetarea e stocată într-o centrală, spaţiu de căutare şi sunt distribuite într-o gamă largă de site-uri de cercetare;
- Depozitul reprezintă o alternativă sigură de a face documentele disponibile pe site-uri web personale;
- Accesul la arhive similare în toată lumea;
- Acces uşor pentru studenţi.

B. Strategia de împingere. Această strategie are scopul de a crea o atitudine pozitivă față de depozit şi de a arăta efectul pozitiv, odată ce materialele au fost arhivate. Mesajul trebuie sa fie sugestiv, în acest sens şi să conțină elemente destinate atingerii scopurilor strategiei de împingere, să pună în evidență ce este ASPECKT-Dspace, ca în exemplul următor.

MODEL DE EMAIL TRIMIS MEMBRILOR COMUNITĂȚII pentru a arăta efectul pozitiv al arhivării lucrărilor ştiințifice în depozit:

- Serviciu on-line pentru a găzdui textul materialului de cercetare publicat de membrii universității Transilvania;
- Un produs destinat creşterii vizibilității şi impactului cercetării ştiințifice;
- Oferă condiții pentru cercetare liberă fără bariere de acces;
- Depozit în deplină siguranță pentru eternitate;
- Serviciu ce verifică acordurile dreptului de proprietate.

C. Strategia de consultare. În această etapă scopul este comunicarea bidirecțională. Acest lucru se realizează prin intermediul sondajelor, reuniuni, grupuri de lucru, de tip peer-to-peer de comunicare şi alte forme de solicitare a feedback despre depozit.

Promovarea internă şi marketingul depozitului digital

Promovarea internă şi marketingul depozitului are în vedere atragerea utilizatorilor din cele două sensuri posibile: autori care furnizează materiale spre arhivare şi utilizatori care se documentează pe baza materialelor conținute în depozit.

Această acțiune trebuie, în mod obligatoriu, să creeze anumiți vectori de comunicare şi convingere a țintelor vizate prin oferirea de informații utile, clare, simple şi cu mare impact referitoare la avantajele şi beneficiile utilizării unui astfel de instrument de documentare şi valorificare a cercetării ştiințifice.

Acțiunile de marketing pentru promovarea depozitului digital, fără a fi exhaustive sunt:

1. Lansarea oficială a depozitului instituțional - DEPOZITULUI PILOT

2. Dezvoltarea la nivel de universitate, prin:

- Relaționare online: crearea unui blog pentru promovarea depozitului,

- Diseminarea experienței prin articole ştiințifice, organizarea de workshopuri pentru promovarea experienței implementării depozitului,
- Organizare de cursuri de instruire.

3. Promovarea depozitului în lume

Pentru a creşte vizibilitatea depozitului, acesta va fi înregistrat în cât mai multe liste de depozite şi servicii existente:

- Repertoarul de Depozite Open Access (Directory of Open Access Repositories - DOAR);
- Registrul Depozitelor Open Access (Registry of Open Access Repositories - ROAR);
- OAIster;
- DRIVER.

Un lucru este sigur: depozitele instituționale vor reprezenta memoria unei societăți şi un factor de progres în cercetare.

Sperăm că drumul deschis de Universitatea Transilvania să reprezinte un exemplu de bună practică pe viitor.

Concluzii privind implementarea unui depozit instituțional digital la Universitatea TRANSILVANIA din Braşov

- Membrii comunității academice din universitatea Transilvania sunt preocupați de dezvoltarea cercetării ştiințifice şi participă la toate competițiile naționale şi europene lansate pentru accesarea fondurilor destinate cercetării.
- Numărul de proiecte câştigate este mare iar producția ştiințifică a universității este remarcabilă. Acest mediu academic este unul în care cu succes se poate implementa un depozit digital.
- Depozitul digital este un produs al tehnologiei informației pe care universitatea îl oferă membrilor comunității sale pentru gestionarea şi difuzarea materialelor digitale create de către instituție şi membrii comunității.
- Acest produs este în esență un angajament organizatoric privind gestionarea materialelor produse de universitate, organizarea, accesul şi distribuția lor dar şi conservarea pe termen lung.

- Prin acest produs se vor colecta şi distribui toate documentele ştiinţifice create de membrii universităţii.
- Obiectivul principal al depozitului digital este creşterea impactului cercetării ştiinţifice a universităţii prin promovarea producţiei ştiinţifice a universităţii prin depozitul digital instituţional.
- Realizarea unor depozite care să aparţină instituţiilor de cercetare, astfel ca munca cercetătorilor să fie vizibilă şi crearea unui depozit central la nivel naţional nu solicită alocarea unor surse financiare considerabile, dacă ţinem cont de costul unui sistem operaţional, implementarea lui şi echipa de dezvoltare.
- Fiind o tendinţă nouă în procesul de furnizare a rezultatelor cercetărilor academice, depozitele instituţionale ar trebui să fie implementate în cadrul câtor mai multe instituţii, pentru ca astfel toţi cei implicaţi în munca de cercetare să beneficieze de mijloace moderne de diseminare a cunoaşterii.
- Includerea in OpenDOAR a primului deposit instituţional din România (fig. 5.9.)

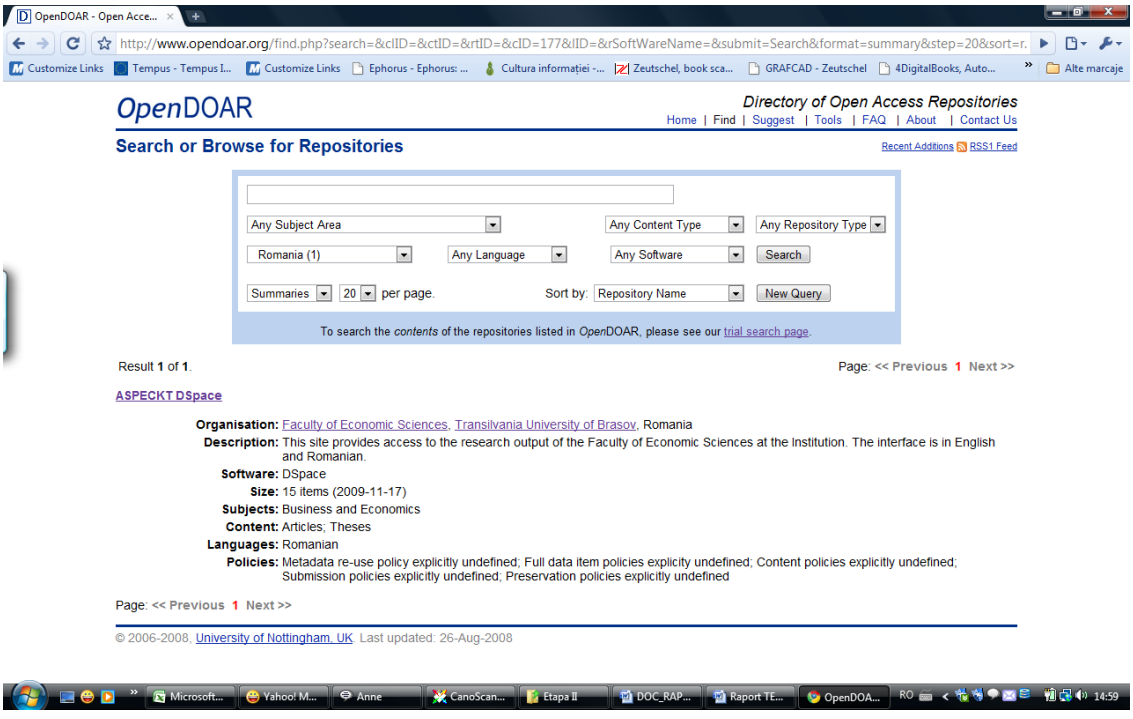


Fig. 5.9. OpenDOAR

6. Concluzii

Trăim într-o eră plină neprevăzut care aduce modele noi de dezvoltare a cercetării ştiinţifice, dar şi dezavantaje pentru instituţiile academice care nu s-au implicat în proiecte de promovare a cercetărilor proprii, care riscă să intre într-un con de umbră în cazul în care nu îşi asigură o oarecare notorietate.

Datorită faptului că totul în prezent se schimbă cu „viteza luminii", trebuie avute în vedere şi nevoile de informare tot mai diverse ale comunităţilor academice. Acestea implică, printre altele şi următoarele:

- folosirea surselor oferite prin intermediu Internetului (motoarele de căutare), care sunt mai uşor de utilizat şi mai accesibile decât depozitele tradiţionale ale bibliotecilor;
- realizarea unor depozite care să aparţină instituţiilor de cercetare, astfel ca munca cercetătorilor să fie vizibilă şi crearea unui depozit central la nivel naţional
- sursele financiare, dacă ţinem cont de costul unui sistem operaţional, implementarea lui şi echipa de dezvoltare.

Fiind o tendinţă nouă în procesul de furnizare a rezultatelor cercetărilor academice, depozitele instituţionale ar trebui să fie implementate în cadrul câtor mai multe instituţii, pentru ca astfel toţi cei implicaţi în munca de cercetare să beneficieze de mijloace moderne de diseminare a cunoaşterii.

Un lucru este sigur: depozitele instituţionale vor reprezenta memoria unei societăţi şi un factor de progres în cercetare.

Sperăm că drumul deschis de Universitatea Transilvania să reprezinte un exemplu de bună practică pe viitor.

7. Note bibliografice si webografice

1. Banciu, D. coordinator, *Contributions to ICT progress - Digital Libraries from Concept to Practice*, Editura Tehnica, 2007, pp.: 77-94
2. Banciu, D., *Informatizarea structurilor infodocumentare*, Editura Ars Docendi, 2007
3. Bănică, L., *Metode şi standarde la nivel de aplicaţie*, Ed.Universităţii din Piteşti, 2005
4. Braşov, U. T. (2008-2012). *Plan strategic.* Braşov.
5. Coravu, R. (2008). Bibliotecile universitare şi accesul deschis la informaţia ştiinţifică. (A. B. România, Ed.) *Revista română de biblioteconomie şi stiinţa informării , 4* (3-4).
6. Campbell, J., *The Academic Library as a Virtual Destination*, EDUCAUSE Review, Volume 41, no.1, 2006, pp.: 16-31
7. Casarosa, V., *The European Digital Library - Discussion guidelines*, Workshop WS1, 2006, www.cimec.ro/ELAG/ELAG-2006-WS1.htm
8. Cristea, L., & Repanovici, A. (2007). Tehnologia informaţiei-motorul trecerii la biblioteca digitală. În R. Angela (Ed.), *BIBLIO Braşov 2007. 2007*, pg. 317-318. Braşov: Editura universităţii Transilvania.
9. DSpace. (2009). *Dspace home federation*. Preluat pe septembrie 1, 2009, de pe DSpace Federation home page, <http://www.dspace.org>.
10. Gadd, E., Oppenheim, C., & Probetts, S. (Acceptat spre publicare). How academics expect to use open-access research papers. *Journal of librarianship and Information Science*
11. Gibbons, S. (2005). Faculty staff do not always understand the jargon used by librarians. The assurance of their work being found on the Web is better understood than the need for consistent metadata, and an unbreakable link better than a persistent URL. Preluat pe Iulie 29, 2009, de pe http://docushare.lib.rochester.edu/docushare/dsweb/Get/Document-19725/amsterdam.ppt

12. Hajjem, C. H. (2006). The Self-Archiving Impact Advantage: Quality Advantage or Quality Bias? University of Southampton,Department of Electronics and Computer Science.
13. Ignat, T. (2008). Biblioteca digitală şi rolul său în activităţile de cercetare ale instituţiei pe care le deserveşte. (A. B. România, Ed.) *Revista română de biblioteconomie şi ştiinţa informării , 4* (3-4).
14. Ignat, T., & Repanovici, A. (2009). Institutional repository marketing: from research to knowledge transfer. Advances in Marketing, Management and Finances, Proceedings of the 3rd International Conference in Management, Marketing and Finances, (MMF'09) (pg. 107-110). Houston,USA: WSEAS.
15. Initiative, B. O. (2002). Preluat pe Mai 23, 2009, de pe http://www.soros.org/openaccess/read.shtml
16. Kevin Yank, *Build your own Database Driven Website using PHP & MySQL*, a guide published by SitePoint, SitePoint Tech Times newsletter, 2009
17. Kosson. (2009). *Depozite digitale*. Preluat pe septembrie 20, 2009, de pe Kosson comunitate virtuală: www.kosson.ro
18. OpenDOAR. (2008). *The Directory of Open Access Repositorie*. Preluat pe Iulie 18, 2009, de pe http://www.opendoar.org/
19. Repanovici, A. (2009). Marketing Research about Attitudes, Difficulties and interest of academic Community about Institutional Repository, PLENARY LECTURE. *Advances in Marketing, Management and Finances, Proceedings of the 3rd International Conference in Management, Marketing and Finances, (MMF'09), Houston,USA,April30-May 2, 2009, ISSN 1790-2769, ISBN 978-960-474-073-4,pag.88-95* (pg. 88-95). Houston,USA: WSEAS.
20. Repanovici, A., & Turcanu, D. (2009). Qualitative and quantitative measures in marketing research for university library resourse assesment. *Books of abstracts, Qualitative and quantitative methods in libraries QQML 2009*, (p. 73). Chania, Greece.
21. RoMEO, S. (2006). *RoMEO News*. Preluat pe Iulie 18, 2009, de pe http://www.sherpa.ac.uk/romeo/
22. Sherpa. (2006). *University of Nottingen*. Preluat pe Iulie 20, 2009, de pe http://www.sherpa.ac.uk/

23. SOROS. (2009). *Open access software*. Preluat pe Noiembrie 14, 2008, de pe http://www.soros.org/openaccess/software
24. Suber, P. (2005). Overview iin Open access- Unrestricted access to published research. *93rd Indian Science Congress* , 7-13.
25. Transilvania, U. (2009). *Institut de cercetare-dezvoltare-inovare:produse high-tech pentru dezvoltare durabilă*. Preluat pe septembrie 3, 2009, de pe http://www.unitbv.ro/institut_prodd/index.php?id=8
26. Transilvania, U. (2009). *Pagina web a universităţii Transilvania*. Preluat pe Septembrie 1, 2009, de pe Departamentul Informatizare: http://www.unitbv.ro/Default.aspx?tabid=1222&language=en-US
27. Ulman, L., *PHP for the World Wide Web*, Editura Teora, 2004
28. Welling, L., Thompson, L., *PHP and MySQL Web development*, 3rd Edition, Editura Teora, 2007
29. Witten, H., I., Bainbridge, D., *How to build a digital library,* Morgan Kaufmann Publishers, 2003
30. http://allformp3.com/extension/what-is-the-dotx-files.html
31. http://www.altn.com/
32. http://answers.com/topic/doc-1
33. http://code.google.com/p/ocropus/
34. http://code.google.com/p/tesseract-ocr/
35. http://dot.extensionfile.net/
36. http://www.anydocsoftware.com/
37. http://en.wikipedia.org
38. http://en.wikipedia.org/wiki/Rich_Text_Format
39. http://euro.ubbcluj.ro/~alina/cursuri/birotica-practic/word/5-8.htm
40. http://euro.ubbcluj.ro/~alina/cursuri/birotica-practic/word/5-3.htm
41. http://ezinearticles.com/?Rich-Text-Format---Easy-to-Use-ormat&id=3204897
42. http://fileinfo.com/extension/doc
43. http://filext.com/file-extension/DOTX
44. http://interglacial.com/rtf/
45. http://kb.iu.edu/data/adnl.html

46. http://office.microsoft.com
47. http://office.microsoft.com/ro-ro/excel/HP100141031048.aspx
48. http://office.microsoft.com/ro-ro/help/HA100069351048.aspx
49. http://office.microsoft.com/ro-ro/products/HA101723691048.aspx
50. http://msdn.microsoft.com/
51. http://pumanet.codeplex.com/
52. http://ro.wikipedia.org/wiki/.doc
53. http://ro.wikipedia.org/wiki/Microsoft_Excel
54. http://z.about.com/d/spreadsheets/1/0/a/2/-/-/excel_2007_screen_parts.gif
55. http://wiki.mobileread.com/wiki/DOC
56. http://www.adobe.com/devnet/flv/pdf/video_file_format_spec_v10.pdf
57. http://www.apple.com/downloads/macosx/productivity_tools/pdfocrx.html
58. http://www.cvisiontech.com/
59. http://www.digitalsquad.ro/
60. http://www.driver-repository.eu/ DRIVER. (2009). *Networking European Scientific Repositories*. Preluat pe iulie 24, 2009
61. http://www.docx.net/what-is-a-docx-file
62. http://www.docx.net/how-to-open-docx-files
63. http://www.docx.net/what-is-a-docx-converter
64. http://www.ecursuri.ro/cursuri/excel-prin-exemple.php
65. www.edlproject.eu/conference/downloads/EDLconf_Gradmann.pdf
66. http://www.encyclopedia.com/doc/1O11-RTF.html
67. http://www.eprintdriver.com/help/v5.0/COM/dllaux/DOCFmt.htm www.europeana.eu/ EUROPEANA project
68. http://www.explaintechstuff.com/how-to-view-docx-files.html
69. http://www.file-extensions.org/dotx-file-extension
70. http://www.fileinfo.com/extension/docx
71. http://www.fileinfo.com/extension/dot
72. http://www.jarte.com/help_new/document_file_formats.html
73. http://www.htmlgoodies.com
74. http://www.microsoft.com/downloads/details.aspx?familyid=ac57de32-17f0-4b46-9e4e-467ef9bc5540&displaylang=en

75. http://www.musitek.com/X/default.html
76. http://www.scantron.com/cognition/
77. http://www.scanstore.com/Document_Management_Solutions/
78. http://www.scribd.com/doc/12831694/manual-microsoft-excel-complet
79. http://www.simpleocr.com/
80. http://www.techterms.com/definition/rtf
81. http://www.traineetrader.com/microsoft-excel-tutorial-series-the-basics-data-entry-and-navigation/
82. www.theeuropeanlibrary.org/telplus -TELplus project
83. http://www.trichview.com/features/files.html
84. http://www.w3.org
85. http://www.word.mvps.org/FAQs/Customization/CreateATemplatePart1.htm
86. http://www.word-tips.com/dotx-and-dotm.html
87. http://www.xml.com