UNIVERZA V LJUBLJANI

FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Matjaž Kragelj

PROBLEMATIKA TRAJNEGA OHRANJANJA DIGITALNIH VIROV

DIPLOMSKO DELO NA UNIVERZITETNEM ŠTUDIJU

Ljubljana, 2010

Univerza
v Ljubljani

Fakulteta za računalništvo
in informatiko

Tržaška 25
1000 Ljubljana, Slovenija
telefon: 01 476 84 11
faks: 01 426 46 47
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Št. naloge: 01716/2010
Datum: 01.12.2010

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kandidat: **MATJAŽ KRAGELJ**

Naslov: **PROBLEMATIKA TRAJNEGA OHRANJANJA DIGITALNIH VIROV**

THE PROBLEM OF PERMANENT PRESERVATION OF DIGITAL SOURCES

Vrsta naloge: Diplomsko delo univerzitetnega študija

Tematika naloge:

Zaradi vse več elektronskega gradiva se pojavlja vprašanje, kako zagotoviti njegovo trajno ohranjanje. S to problematiko se ukvarjajo mnoge institucije, razviti so modeli in rešitve, obstaja tudi zakonodaja, npr. Zakon o varstvu dokumentarnega in arhivskega gradiva. Manjkajo pa procesi, ki bi natančneje določili, kako izvajati zajem, digitalizacijo, izgradnjo, vzdrževanje ter varovanje elektronskih publikacij. V okviru diplomske naloge analizirajte in opišite probleme s področja trajnega ohranjanja digitalnih vsebin, s katerimi se soočajo predvsem knjižnice ter tudi muzeji in raziskovalne organizacije. Preglejte in opišite rešitve, ki se uporabljajo v svetu.

Mentor:

prof. dr. Marko Bajec

Dekan:

prof. dr. Nikolaj Zimic

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

UNIVERZA V LJUBLJANI

FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Matjaž Kragelj

PROBLEMATIKA TRAJNEGA OHRANJANJA DIGITALNIH VIROV

DIPLOMSKO DELO NA UNIVERZITETNEM ŠTUDIJU

Mentor:

izr. prof. dr. Marko Bajec

Ljubljana, 2010

IZJAVA O AVTORSTVU

diplomskega dela

Spodaj podpisani Matjaž Kragelj

z vpisno številko 24940569

sem avtor diplomskega dela z naslovom:

Problematika trajnega ohranjanja digitalnih virov

S svojim podpisom zagotavljam da:

- sem diplomsko delo izdelal samostojno pod mentorstvom izr. prof. dr. Marka Bajca;
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki »Dela FRI«.

V Ljubljani, dne 16.12.2010 Podpis avtorja:

ZAHVALA

Na prvem mestu gre zahvala mojemu mentorju – Prof. dr. Bajcu. Čeprav v času študija ni bil nikdar moj profesor, me je kot mentor vodil pri eni od seminarskih nalog, redno pa se srečujeva na posvetovanjih "Dnevi slovenske informatike". Njegov pristop do dela, način vodenja in usmerjanja študentov, ter obvladovanje širokega spektra informatike, je botrovalo k temu, da je bila izbira mentorja enostavna naloga. Profesor Bajec – najlepša hvala!

Zahvala gre tudi staršema, ki sta mi omogočila študij in vsa ta leta čakala na njegov zaključek.

Pozabiti ne smem Zorana in vodstva Narodne in univerzitetne knjižnice. Njim gre zasluga, da je diplomsko delo ugledalo luč sveta še v tem letu…

Hvala Gorazdu in Aniti za prevod ter lekturo in ne nazadnje Tinetu, za razvoj vseh potrebnih aplikacij in modulov.

Na koncu se zahvaljujem moji Sabini, ki ni niti za trenutek podvomila v končni uspeh.

Don't give up!

KAZALO

SEZNAM UPORABLJENIH KRATIC IN SIMBOLOV

OAIS	*(angl) Open Archival Information System*
URI	*(angl) Uniform Resource Identifier*
FEDORA	*(angl) Flexible Extensible Digital Object Repository Architecture*
WCT	*(angl) Web Curator Tool*
IIPC	*(angl) International Internet Preservatium Consortium*
CCSDS	*(angl) Consultative Committee for Space Data Systems*
SIP	*(angl) Submission Information Package*
AIP	*(angl) Archival Information Package*
DIP	*(angl) Dissemination Information Package*
XPS	*(angl) XML Paper Specification*
ODF	*(angl) Open Document Format for Office Applications*
OOXML	*(angl) Office open XML*
ODS	*(angl) Open Office Spreadsheet*
DC	*(angl) Dublin Core*
MARC	*(angl) Machine-Readable Cataloging*
UNIMARC	*(angl) Universal MARC*
COMARC	*(angl) Cooperative MARC*
OAI-PHM	*(angl) Open Archives Initiative Protocol for Metadata Harvesting*
TEL	*(angl) The European Library*
FOXML	*(angl) Fedora Object XML*
API	*(angl) Application Programming Interface*
SOAP	*(angl) Simple Object Access Protocol*
REST	*(angl) Representational State Transfer*

POVZETEK

S problematiko trajnega ohranjanja dokumentnega gradiva se ukvarja precej institucij, razvitih je precej modelov in rešitev zanje. Rešitve na ključ omogočajo tako uporabo programskih rešitev na eni strani, kot najem podatkovnega prostora in arhiviranja na drugi. Zakonodaja je na tem področju precej jasna, npr. Zakon o varstvu dokumentarnega in arhivskega gradiva. Precej manj določeni so procesi zajema, digitalizacije, izgradnje, vzdrževanja ter varovanja elektronskih publikacij ustanov kot so izobraževalne institucije, knjižnice, muzeji, galerije. Zaradi ohranjanja publikacij naših prednikov – kulturne dediščine in ohranjanja del, nastalih danes (kulturna, nacionalna, raziskovalna dejavnost) se kaže potreba po večji sistematizaciji pri procesu trajnega ohranjanja elektronskih publikacij. Publikacijo želimo trajno hraniti predvsem zaradi uničujočega vpliva okolja, kateremu je izpostavljena in zaradi zagotavljanja dostopa širšemu krogu uporabnikov preko interneta, ne glede na čas in lokacijo. V nalogi so opisane nekatere težave v povezavi z zajemom spletnih vsebin, formati bibliografskih podatkov, možnostjo shranjevanja objektov v različne formate, nakazani so repozitoriji za hrambo vsebin. Ponujene so tudi praktične rešitve, ki jih uporabljajo nekatere institucije pri obvladovanju procesa trajnega ohranjanja e-vsebin.

KLJUČNE BESEDE

OAIS, Fedora, Web Curator Tool, optična razpoznava besedil, spletne publikacije, PDF/A

ABSTRACT

There are many institutions dedicated to the permanent preservation of documentary materials; many models have been developed and also solutions which apply to them. Turnkey solutions enable us to use the software solutions on the one hand and to provide the rental space and archiving data on the other. Legislation in this area is fairly clear, for example The Law about Protection of Documents and Archives. Much less determined are ways of digitization, construction, maintenance and protection of electronic publications of institutions such as educational institutions, libraries, museums and galleries. In order to maintain the publications of our ancestors, namely our cultural heritage, and to preserve the works created today (various cultural, national, research endeavours) there is a need for greater sistematization during the process of permanent preservation of electronic publications. Each and every publication should be preserved permanently in spite of the devastating impact of the environment to which it is exposed. In this way we could ensure access to a wider community of users via the Internet, regardless of the time and location. The present study describes some problems concerning the capture of Web content, formats of bibliographic data and the possibility of storage facilities in various formats. It describes the repositories for the storage of various contents. Finally, it offers the practical solutions enacted by some institutions in managing the process of permanent preservation of e-content.

KEYWORDS

OAIS, Fedora, Web Curator Tool, optical recognition, web publications, PDF/A

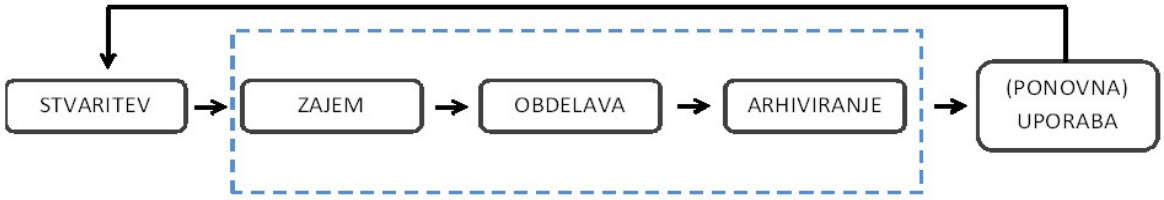
1 UVOD

Z nenehnim in hitrim razvojem informacijske družbe in z njo informacijske tehnologije so postale zahteve in pričakovanja uporabnikov e-storitev, ki vključujejo tudi elektronske vire sorazmerno visoke. Vlogo čitalnic in knjižnic dopolnjujejo baze znanja in e-vsebine dostopne preko spleta, delež uporabnikov te vrste storitev še vedno strmo narašča. Dostop do kulturnih zakladov, znanstvenih publikacij in vsebin, ki tudi gradijo nacionalno identiteto, zahteva opravilo več aktivnosti, ki skozi faze zajema, obdelave in arhiviranja vsebin lahko ponudijo končnemu uporabniku vsebino, ki mu omogoča

- lažji in hitrejši dostop do virov znanja,
- nove možnosti in boljši dostop do vsebin,

poleg tega pa je z digitalizacijo vsebin moč fizične nosilce informacij zaščiti pred še vedno prehitrim propadanjem. Ugotovitev, da raziskovalca starih tiskov tak način dostopa do informacij prav posebej ne bo ogrel in bo vedno raje znova in znova prelistaval stare kodekse in rokopise, je najbrž res na mestu. Poleg informacije – samega besedila ga pogostokrat pritegne fizična struktura in format publikacije, sestavni materiali, vezava, itd., a za veliko večino uporabnikov, ki v delu preučujejo zgolj vrednost zapisane informacije, predstavlja možnost oddaljenega dostopa do vsebin dodano vrednost informacijske tehnologije.

Funkcija "nudenje elektronskega vira" javnosti, zahteva izvajanje določenih postopkov oz. procesov pri delu s publikacijo, ki jo želimo nuditi uporabniku. Življenjsko pot ali tok publikacije, ki jo za potrebe trajnega ohranjanja in nudenja na voljo javnosti digitaliziramo, lahko na preprost način opišemo tako, kot prikazuje slika 1.1.



Slika 1.1) Življenjska pot publikacije

1.1 Motivacija za diplomsko delo

V Narodni in univerzitetni knjižnici, kjer sem že več let zaposlen, se nenehno srečujem s knjižničnim gradivom, bodisi da gre za gradivo na fizičnih nosilcih, bodisi za digitalizirano ali elektronske vire. Čeprav bi nas lahko dejstvo, da (nam dostopni) prvi rokopisi ali tiski še vedno precej uspešno kljubujejo zobu časa in so se ohranili čez več stoletij, na prvo žogo odvrnili od ideje po digitalizaciji gradiva z namenom trajnega ohranjanja, poglobljen razmislek kaže nasprotno. Drži, da je življenjska doba CD ali DVD ploščka krajša od "starih tiskov", drži tudi, da bo potrebnega še precej časa, da bomo lahko z gotovostjo trdili, da so zapisi na digitalnih nosilcih dovolj obstojni in še vedno uporabni, a se vseeno in v vedno večji meri odločamo za digitalizacijo publikacij. Diplomskega dela sem se lotil z željo po predstavitvi težav in preprek, s katerimi se soočajo knjižnice, muzeji, raziskovalne organizacije pri poskusu trajnega ohranjanja digitalnih vsebin. Hkrati bi rad ponudil praktične rešitve in pristope, ki jih je pri tem početju moč uporabiti in jih uporabljajo tudi sorodne organizacije po svetu, predvsem knjižnice.

1.2 Namen in cilj diplomskega dela

Namen diplomskega dela je prikazati "življenjski" tok (spletne) publikacije, ki jo želimo za potrebe trajnega ohranjanja in nudenja javnosti danes in v prihodnjih letih izbrati, po potrebi digitalizirati (lahko upravljamo z digitalno rojenimi vsebinami), obdelati - tako publikacijo kot njen bibliografski opis, arhivirati in s tem zagotoviti dostop do nje in ponovno uporabo.

Cilj diplomskega dela je prikazati primer uporabe referenčnega modela OAIS (Open archival information system) v praksi, natančneje – na primeru Narodne in univerzitetne knjižnice. Model OAIS je bil sprejet kot ISO standard leta 2003 (ISO 14721:2003) in služi kot meta model za določevanje funkcij, procesov in entitet, za zagotavljanje trajnega ohranjanja digitalnih virov. Sam model ne omejuje uporabnika s potrebnimi tehničnimi rešitvami ali implementacijami, služi kot model na logičnem nivoju za pomoč pri celovitem upravljanju z arhivi publikacij.

1.3 Predvideni rezultati in možnost njihove uporabe

V diplomskem delu je govora o publikacijah, ki jih lahko razporedimo na naslednji način:

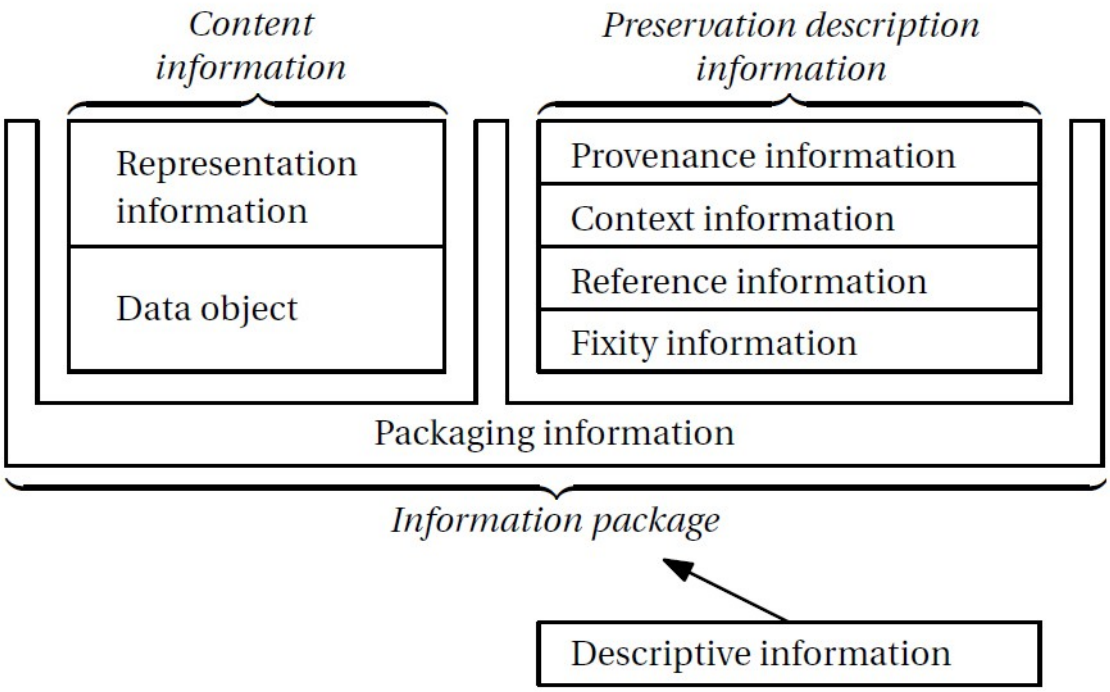
- spletne publikacije,
- digitalno rojene / nastale vsebine ter
- publikacije, ki so nastale kot plod digitalizacije fizičnih nosilcev.

Ker gre v grobem za tri skupine virov informacij, so za zajem publikacij potrebne različne aktivnosti. V diplomskem delu bomo poleg predstavitve težav, ki nastopajo pri zbiranju in urejanju gradiva za trajno ohranjanje, poskušali podati praktična napotila in principe, ki so lahko v pomoč tako posamezniku, ki mu je mar za lastno delo, ki ga je proizvedel v večletnem obdobju profesionalnega raziskovanja ali zgolj zbiranja različnih virov informacij, kot na drugi strani knjižnicam, muzejem, galerijam izobraževalnim zavodom, itd. Zaradi problematike procesa trajnega ohranjanje digitalnih virov je na trgu zaznati vsakodnevne izboljšave obstoječih sistemov in razvoj novih, naprednejših in uporabniku bolj prijaznih izdelkov. V luči tega spoznanja se ne slepimo, da bo pričujoče diplomsko delo v pomoč pri reševanju vseh potreb v procesu trajnega ohranjanja publikacij, upamo le, da bo dovolj nazorno ponazorila pasti in nevšečnosti, ki se pojavljajo pri poskusu gradnje arhiva in uporabniku nudila praktične rešitve pri spopadanju s to težko in nikoli dokončano nalogo.

2 OAIS REFERENČNI MODEL

Referenčni model OAIS - Open Archival Information System [1] je nastal v okviru posvetovalnega odbora za potrebe vesoljskih podatkovnih sistemov, oz. za potrebe analiz podatkov, pridobljenih iz teh sistemov (CCSDS – Consultative Committee for Space Data Systems), z namenom uvajanja standardov za dolgotrajno ohranjanje podatkov raziskav vesoljskih podatkovnih sistemov. Ker model ne predpisuje načrta ali implementacije rešitve, temveč nakazuje koncepte, ki nastopajo v procesu trajnega ohranjanja digitalnih virov, je svoje mesto kmalu našel v svetu knjižnic in raznih ustanov, ki obvladujejo množično število publikacij.

Informacijski paket, ki vstopa v sistem, je po OAIS sestavljen iz vsebine, ki jo želimo ohraniti in jo delimo na objekt in opis objekta na eni in potrebne informacije, ki nam bodo v pomoč pri umeščanju in iskanju objekta v sistemu (vir, njegovo povezanost in odnos z drugimi viri, referenčne informacije in informacije, ki nas obveščajo o pristnosti in nespremenljivosti vira) na drugi strani.

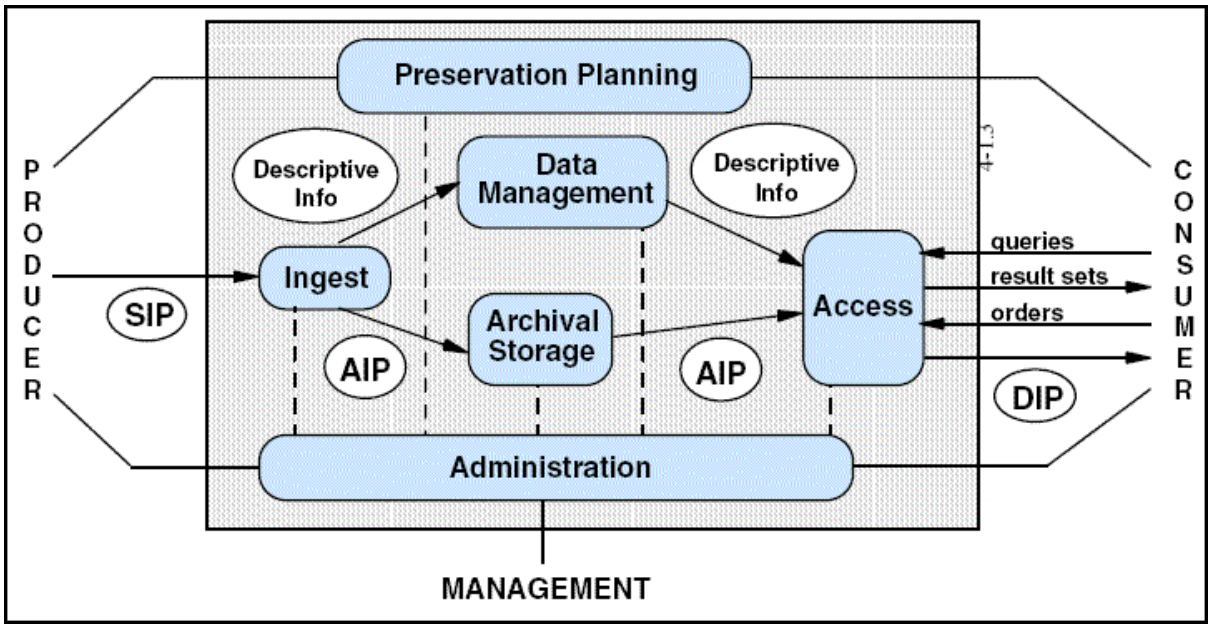


Slika 2.1) Informacijski paket, kot ga razume model OAIS

Glavne funkcije, ki jih podpira model OAIS, so naslednje:

- zajem (ingest),
- arhiviranje in skladiščenje (Archival storage),
- upravljanje z (meta) podatki (Data Management),
- upravljanje sistema (Administration),
- vzdrževanje arhiva (Preservation Planning),
- zagotavljanje dostopa (Access),

poleg teh pa še podporne funkcije, ki skrbijo za nemoteno delovanje sistema (upravljanje s platformo sistema, mreže, varnosti, itd). Na sliki 2.2 je podana shema referenčnega modela OAIS.



Slika 2.2) Model OAIS

Ob omenjenih funkcijah nastopajo v modelu tudi entitete, ki predstavljajo informacijske pakete, oz. vsebine, ki jih vlagamo v sistem.

- SIP (submission information package), paket, ki ga vlagamo v sistem,
- AIP (archival information package), arhivski paket (razširjen z dodatnimi informacijami o objektu, ki ga vlagamo (npr. z opisnimi podatki),
- DIP (dissemination information package), paket, ki je prilagojen končnemu uporabniku.

Ne bomo se spuščali v natančnejši opis OAIS modela, za razumevanje principa trajnega ohranjanja zadostuje grobi oris, ki smo ga podali. Omenimo še, da za trajno ohranjanje elektronskih publikacij oz. elektronskih virov iz vidika knjižnic ne velja Zakon o varstvu dokumentarnega in arhivskega gradiva ter arhivih (ZVDAGA) in termina arhiviranje ne interpretiramo tako, kot ga navaja ta zakon.

3 ZAJEM PUBLIKACIJ PREKO SPLETA

3.1 Kaj so spletne publikacije?

Zakon o obveznem izvodu publikacij [2], v 2. členu definira spletne publikacije kot »elektronske knjige, elektronski časopisi in časniki, dostopni po spletu ter spletne strani in podobno« in jih v 4. členu opredeli kot predmet obveznega izvoda. »Publikacija« je po tem zakonu vsak zapis informacije na kateremkoli nosilcu, ki je izdan, založen, izdelan, prirejen ali fizično ali elektronsko distribuiran za uporabo v javnosti;

[3] med spletne publikacije poleg knjig, časnikov, časopisov in člankov, ki so objavljeni na spletu, uvršča tudi »spletna mesta (strani) organizacij, oseb in dogodkov, portale, spletno dostopne storitve, podatkovne zbirke, spletne novice, spletne konference (forumi), spletne biltene (newsletters) ter različne druge elektronske vsebine kot so video in zvočni zapisi, interaktivne zemljevide in mestne načrte, računalniške programe, računalniške igre, spletno umetnost (web art), mrežne dnevnike (blogi), hitre strani (wiki), učenje na daljavo (e-learning) in podobno«.

Iz definicije Zakona o obveznem izvodu in Pravilnika, objavljenega in sprejetega s strani Narodne in univerzitetne knjižnice se sama po sebi odpira globel problematike trajnega hranjenja digitalnega gradiva. Internet kot medij predstavlja prostor najhitrejšega spreminjanja (dodajanje, odvzemanje, urejanje) informacij in je že zgolj zaradi tega dejstva nemogoče zagotoviti natančen posnetek stanja tega medija v katerikoli izbrani časovni točki. Prva težava, na katero naletimo pri poskusu ohranjanja kulturne dediščine spleta je, kot ugotavljamo (pre)hitra sprememba informacij na mediju, glede na minimalni čas, potreben za izvedbo posnetka le-tega. Čeprav je težava teoretično nepremostljive narave, jo lahko elegantno premostimo na način, da zagotovimo konsistenco za nek časovni okvir. Kot

analogni primer lahko ponudimo zaplet pri zagotavljanju konsistence razpršene podatkovne baze. Če bi želeli ohranjati konsistenco razpršene podatkovne baze v vsakem trenutku, bi prej ko slej prišli do točke, kjer ne bi bilo možno zagotavljati opravljanja transakcij za uporabnika, saj bi sistem vse napore vložil v samo zagotavljanje konsistence podatkov med razpršenimi lokacijami. Pri zajemanju spleta (web harvesting) lahko tako ponudimo posnetek spleta v nekem dovolj širokem časovnem intervalu. Problematika okrog omenjene težave ima smisel le v dovolj ozkem časovnem pogledu, če namreč neko spletno mesto zajemamo skozi celo leto (npr. vsak teden enkrat), bomo zanamcem lahko ponudili precej dobro sliko o tem spletnem mestu ter vsebinah na njem, saj bomo v desetih letih vsebino zajeli več kot petsto krat. Težava, ki je sprva morda ne opazimo, je pa precej bolj pereča kot prva, je (ne)sposobnost robota za zajemanje spletnih mest, kjer je vsebina dosegljiva zgolj preko iskalnih form. Če so spletni roboti – roboti za preiskovanje spleta učinkoviti pri zajemanju vsebin statične narave (html, txt, xml, txt, PDF, itd.), kjer je vsebina nekega spletnega mesta shranjena na datotečnem sistemu spletnega strežnika, pa trčimo ob (zaenkrat) nepremostljivo oviro ob poskusu zajema spletnih portalov – tistih, kjer se vsebina prikaže šele po klicu spletne forme z našim iskalnim nizom kot parametrom funkcije klica. Če bi v praksi ta nemoč robota obsegala le iskalne portale kot so Google, Najdi.si, Yahoo itd, to ne bi predstavljalo posebne težave. Srž problema tiči v dejstvu, da predstavlja količina informacij dinamičnih strani precej večji delež kot je delež statičnih strani. Ker uporablja spletni robot URI naslove za potovanje po straneh in njihovo zajemanje, pri take vrste spletnih mestih ostane brez moči.
Za primer vzemimo digitalni repozitorij znanstvenih publikacij ali digitalno knjižnico, kjer je na voljo na sto tisoče objektov. Zbrana količina informacij je za robota nedosegljiva, saj brez parametra (iskalnega niza) ne uspe pridobiti nabora informacij. Rešilna bilka, ki jo spletne strani običajno vsebujejo, je bodisi indeks strani – kazalo (po vsebini, seznam avtorjev, seznam del, itd.) ki ga robot uporabi za preiskovanje in nabiranje informacij – objektov za zajem. Alternativna možnost je npr. dinamična gradnja imenika objektov na spletnem mestu.
Iz podatkovne baze objektov zgradimo dokument tipa XML, html, txt ali podobno, v katerem navajamo pot (relativno ali eksplicitno) do objektov v repozitoriju. Dinamično sestavljen dokument je robotu dobro napotilo za dostop do vsebin, saj vsebuje pot do vseh objektov.

3.2 Oblika zapisa zajetih objektov

International internet preservation consortium – IIPC [4] oz. mednarodni konzorcij za ohranjanje spletnih vsebin je mednarodna institucija, ki jo sestavljajo v veliki večini nacionalne in digitalne knjižnice ter arhivi. Glavno poslanstvo konzorcija je ohranjanje spletnih vsebin oz. znanja na spletu za prihodnje rodove. Lahko bi rekli, da gre pri tej nalogi za tri faze – **zajem, dostop in ohranjanje gradiva,** kar nas spomni na sliko 2.2 in model OAIS. Leta 1996 so pripadniki konzorcija razvili format datoteke .arc, v katerega spletni roboti pišejo na spletu zajeto gradivo. Datoteko sestavlja več tekstovnih blokov – arc zapisov (glava + telo), kjer glava opisuje vrsto objekta, datum zajema, dolžino objekta (telesa), telo pa predstavlja del spletne strani, binarno kodo objekta (npr. slike ali fotografije na strani). Posamezna .arc datoteka je po privzetih nastavitvah dolžine velikosti max. 100mb in vsebuje množico spletnih strani domene, ki jo zajemamo. Iz same .arc datoteke je moč ročno ali programsko rekonstruirati zajeto spletno mesto. Razširitev .arc datoteke predstavlja datoteka tipa .warc (Web archive file format). Nastala je v okviru istega konzorcija kot nadaljevanje razvoja arhivskega formata. V primerjavi s starejšo arc datoteko vsebuje warc datoteka precej večji nabor opisnih – meta informacij tako o zajemu, kot o objektih samih, omogoča stiskanje podatkov zajema itd. Maja 2009 je warc pridobil ISO standard (ISO 28500:2009).

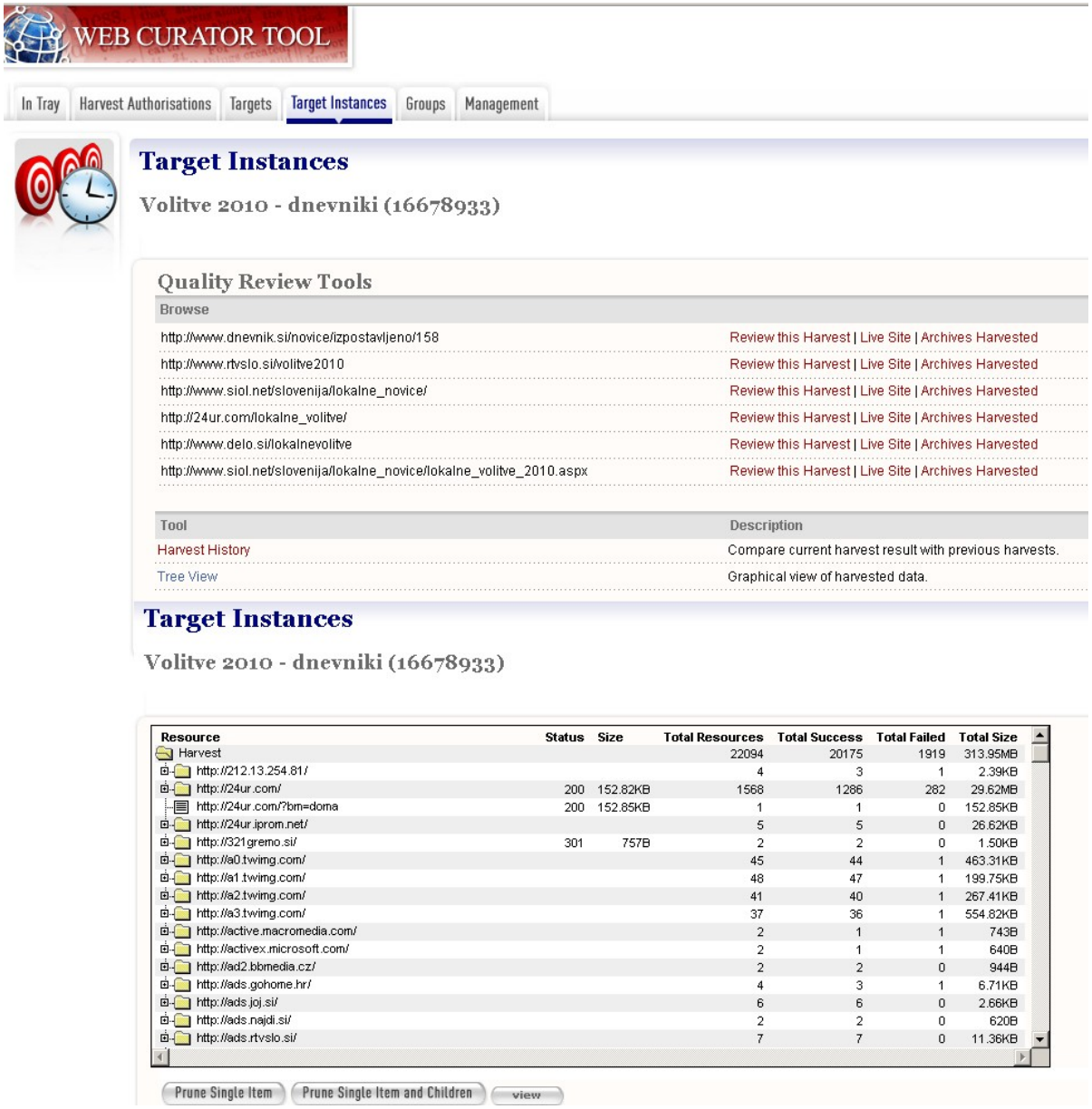
Med programskimi orodji, ki uporabljajo warc format za shranjevanje zajetih spletnih strani, omenimo Heritrix (v razvoju je sodeloval IIPC) in Httrack website copier. Oba programska produkta sta brezplačna.

Nacionalna knjižnica Nove Zelandije in British Library sta v sodelovanju z IIPC izdala orodje Web Curator Tool – WCT [5]. Gre za spletni vmesnik, ki uporablja Heritrix za potrebe zajema spleta, samo orodje pa omogoča precej več, kot le zajem posameznih domen.

Programski paket podpira relacijsko podatkovno bazo Oracle, PostgreSQL, MySQL, najnovejša verzija pa tudi MSSQL. V tabelah podatkovne baze so zbrani podatki o posameznem zajemu, avtorju, času, relacijah do objektov na datotečnem sistemu, pravicah, uporabnikih itd. Spletni vmesnik je namenjen predvsem knjižničnim delavcem in za razliko od modula, ki ga uporablja Heritrix, ne zahteva poglobljenega tehničnega znanja. Omogoča popoln nadzor nad procesom zajemanja spletnih vsebin. Upravljavcu sistema so poleg določanja spletnih domen za zajem ponujene možnosti kot so določevanje globine zajema posamezne domene (bodisi glede na velikost objektov ali število skokov po strukturi drevesa),

možno je omejevanje zajema spletnih strani, ki so del zunanje domene (kot zunanji vir domene), možno je ločevanje vrste objektov, ki naj se zajemajo, časovni interval oz. frekvenco zajemanja, pasovno širino, namenjeno posameznemu robotu, maksimalno število niti, ki lahko vzporedno tečejo in zajemajo spletne vsebine, format zapisa (arc, .warc), itd. V več nacionalnih knjižnic so se spletnega zajema lotili različno. V nekaterih zajemajo splet celotne hrbtenice nacionalnega omrežja in to popolno. Robot tako preiskuje in zajema vse URI naslove, ki jih "vidi" in njihovo vsebino. Drugod so si izbrali napornejšo pot, a le-ta poskrbi za manjšo količino zajetih podatkov. Kot primer lahko navedemo Narodno in univerzitetno knjižnico v Ljubljani (NUK), ki nabor spletnih domen, primernih za zajem določi s pomočjo Pravilnika o vrstah in izboru elektronskih publikacij za obvezni izvod. Zaradi potencialno ogromnih količin podatkov dosegljivih na spletu se posamezno domeno vključi v nabor zajetih vsebin, če poleg dejstva, da pripada domeni .si, ali je rezultat dela slovenskega avtorja oz. govori o Sloveniji, zadosti filtru, da vsebuje »kulturno ali znanstveno vsebino« posebnega pomena.

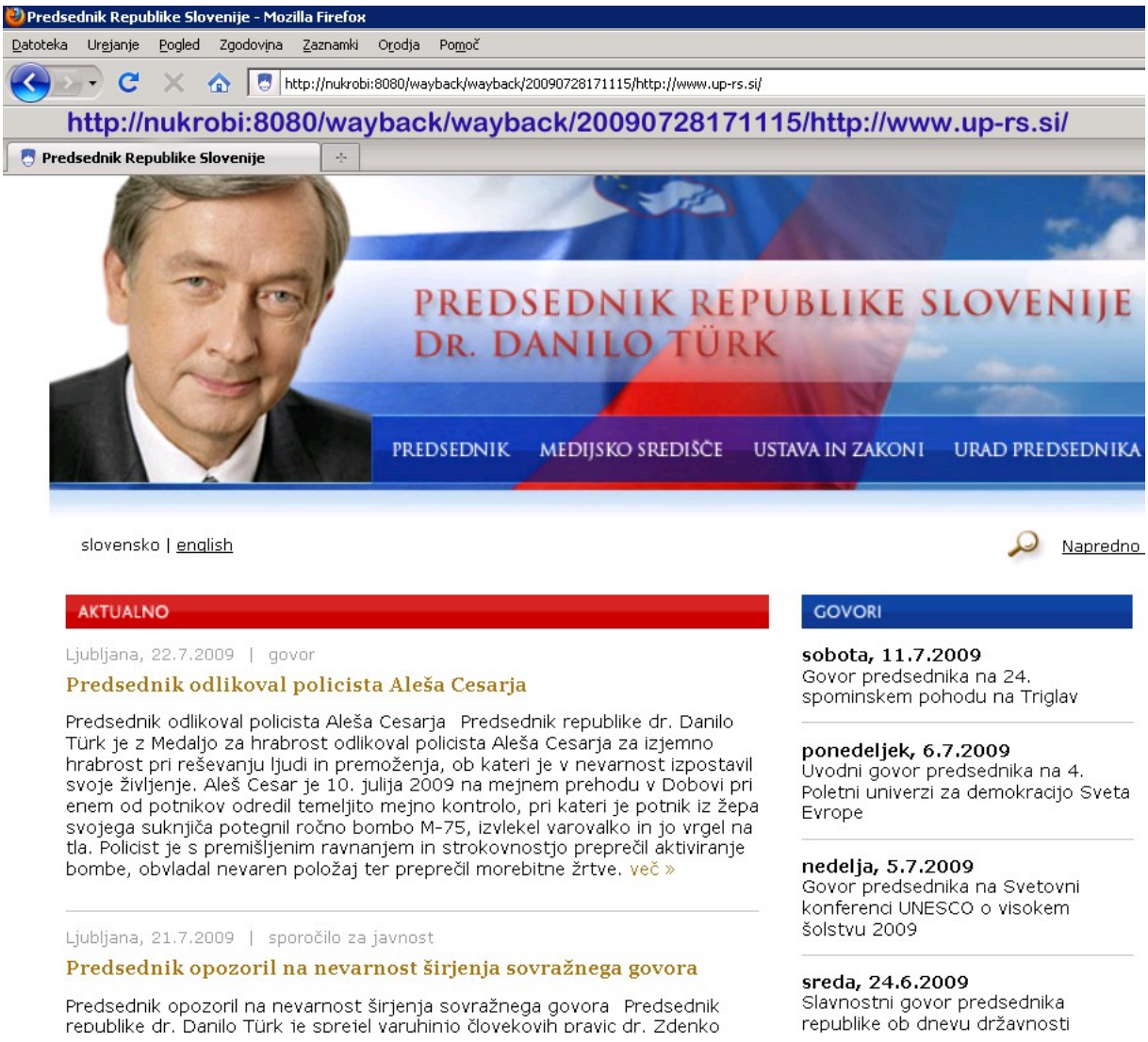
Rezultat zajemanja je, kot smo že omenili, datoteka/datoteke tipa warc ali arc, shranjene na datotečnem sistemu strežnika. Orodje WCT omogoča sprehod po drevesu vsebin posameznega zajetja in rezanje / odstranjevanje vsebin, kot nam kaže slika 3.1.



Slika 3.1) Modul za urejanje zajetih vsebin

Pri tem početju pregledujemo drevesno strukturo zajetih spletnih domen, iz arhiva umikamo oz. brišemo dele spletnih vsebin, ki predstavljajo šum, oz. informacijo, ki sicer predstavlja del spletne domene, ni pa vitalnega pomena za ohranitev spletišča (reklame, video ali zvočne vsebine, forume, spletne dnevnike, formularje, spletne prijavnice, itd). Tako je omogočeno arhiviranje zgolj tistih vsebin, ki jih želimo ohraniti (določeno PDF datoteko, besedilno datoteko, le točno določeno video vsebino, itd).

Aktivnost zajemanja, urejanja in arhiviranja spletnih vsebin s pomočjo za to pripravljenega prosto dostopnega in brezplačnega orodja WCT dopolnjuje aktivnost nudenja vsebin končnemu uporabniku. V okviru istega konzorcija – IIPC je nastal produkt **Wayback machine**. Prosto dostopen program, pisan v Javi je moč namestiti in uporabljati na različnih platformah (Unix, Linux, MAC OS,Windows okolje). Wayback machine je spletna aplikacija in omogoča pregledovanje zajetih vsebin preko spletnega brskalnika v realnem času. Za vsako zajeto spletno domeno je uporabniku na voljo kronološki seznam url povezav do zajetih in arhiviranih spletnih mest. Uporabnik ima možnost preiskovanja arhiva spletnih strani in dostopa do vsebin, ki jih je spletnem robotu uspelo zajeti, upravljavec pa jih je shranil v arhiv. V sklopu konzorcija so bile razvite dodatne spletne aplikacije, ki omogočajo indeksiranje in iskanje po polnem besedilu zajetih spletnih domen. Na ta način postane arhiv tudi spletni iskalnik. Kadar zajemamo spletno domeno z namenom ohranjanja popolne vsebine (oblika strani s css elementi, ves nabor objektov na strani, ki ga lahko zajamemo, javascript funkcije, spletne obrazce, itd) postane spletni portal za uporabnika enak originalnemu. Poleg dejstva, ki nakazuje, da gre za arhiv - URI spletnega mesta (slika 3.2), je edina omejitev za uporabnika dejstvo, da je spletni portal, ki ga preiskuje, vedno starejšega izvora. V trenutku preiskovanja vsebin je lahko na pravem spletnem mestu kopica novih vsebin, lahko pa v realnosti spletnega mesta sploh ni več.



Slika 3.2) Zaslonska slika zajetega spletnega mesta

4 (ARHIVSKI) FORMATI IN DATOTEČNI STANDARDI

Z uporabo spletnih robotov in aplikacij za upravljanje in prikazovanje zajetih vsebin postanejo funkcije **zajema, arhiviranja** in **uporabe** razmeroma preprosto opravilo. Precej večja nejasnost je z arhiviranjem publikacij, ki niso javno dostopne na spletu oz. niso sestavni del spletnih mest. Med publikacije, poleg tekstovnih datotek na spletu, katere smo v prejšnjem poglavju že arhivirali, v knjižničnem svetu uvrščamo predvsem tiskano gradivo, rokopise, notno, slikovno ter avdio in video gradivo, zemljevide, v zadnjem času pa prihaja tudi do poskusa arhiviranja 3d objektov. Že bežni pogled na različne tipe in formate gradiva nam daje slutiti, da je težko najti skupni imenovalec za arhiviranje omenjenega gradiva. Na naslednjih straneh bomo poskusili orisati težave pri izboru (arhivskih) standardov za shranjevanje publikacij in na praktičnih primerih prikazali delo z njimi.

4.1 PDF/A

PDF (Portable Document Format) format, ki ga je družba Adobe Systems lansirala leta 1993, je precej poenostavil izmenjavo različnih vrst podatkov med uporabniki (osebnih) računalnikov. Uporablja se za predstavitev (dvodimenzionalnih) dokumentov, neodvisno od programske in strojne opreme, ter operacijskega sistema, v katerih so bili ustvarjeni. PDF datoteka torej obdaja dokument vključno z njegovimi sestavnimi deli kot so teksti, pisave, slike, vektorska grafika. Z uporabo pretvornikov je moč predstaviti v PDF obliki tudi 3d objekte. Ker gre za odprt format, je tako izgradnja kot pregledovanje PDF dokumentov mogoča z razvojem programske opreme drugih podjetij, ne le podjetja Adobe. Pri podjetju Adobe zatrjujejo, da bo datoteke PDF, bodisi verzije 1 (iz leta 1993) ali naslednic, možno brez večjih težav in omejitev pregledovati tudi v prihodnosti. Ker format PDF [6] ni primeren za trajno ohranjanje digitalnih virov, so pri podjetju Adobe šli korak dlje in razvili format PDF/A.

Kakšna je razlika med PDF in PDF/A standardom in zakaj bi ga uporabili?

Format PDF/A je namenjen arhiviranju oz. trajnemu ohranjanju publikacij. Pridobil je ISO standard za arhiviranje in trajno ohranjanje leta 2005 (ISO 19005-1:2005)

Na prvem mestu velja omeniti zahtevo po vključenih pisavah. Format PDF/A zahteva vključevanje uporabljenih pisav (fontov) v sam dokument, saj je to potreben pogoj za zagotavljanje popolne reprodukcije dokumenta. Nemalokrat se nam namreč pripeti, da v posameznih PDF datotekah, ki jih prebiramo, na zaslon ali tiskalnik ne dobimo vseh znakov, ki jih je uporabil avtor. Če gre za izgubo šumnikov, lahko manjkajoče znake prepoznamo in dokument preberemo in uporabimo, vsekakor pa tak dokument ni primeren za trajni arhiv. Standard PDF/A hkrati ne dovoljuje uporabe slojev. Kot že rečeno, je datoteke PDF moč ustvarjati iz različnih aplikacij in mnoge od njih dopuščajo oz. podpirajo uporabo slojev - layerjev (AutoCad, Microsoft Visio, Adobe Illustrator, Adobe InDesign itd). Z njimi lahko poskrbimo za večjezičnost, za notico o avtorstvu (lahko je na zaslonu nevidna, pri tiskanju pa vidna), v sloje postavimo grafične objekte, vidne ali nevidne itd. Format PDF, verzija Acrobat 7 (PDF 1.6) pri izvozu podpira ohranjanje slojev. V PDF pregledovalniku uporabnik dostopa do datoteke na način, kot jo je videl avtor. S takim načinom shranjevanja izvornih datotek postavimo bodočega uporabnika npr. čez petdeset let v kočljivo situacijo. Težko se bo namreč odločil, kateri sloji so bili v končnem izdelku uporabljeni, kateri ne, kateri sodijo skupaj, kateri bi morali ostati nevidni itd. Izvoz v obliko PDF/A nam to izbiro prepoveduje – vsi elementi PDF datoteke morajo biti dostopni in vidni, zaradi zmede, ki bi lahko nastala ob uporabi slojev, pa je uporaba le-teh pri shranjevanju v PDF/A prepovedana.

4.1.1 Primerjava PDF/A-1a in PDF/A-1b

ISO (International Organization of Standardization) deli PDF/A standard v dva nivoja ustreznosti. Zahteve pri formatu **PDF/A-1a (Level A)** določajo možnost opisa teksta v Unicode obliki, zahtevo po reprodukciji dokumenta, pri čemer ne sme priti do optične dvoumnosti (primer: težava s sloji) in zahtevo po pravilni strukturi dokumenta. Zahteva pri zapisu dokumenta v format **PDF/A-1b (Level B)** kot zadosten pogoj določa zgolj zagotavljanje reprodukcije brez možnosti optične dvoumnosti. V tabeli 1 so prikazane glavne razlike v formatih PDF/A-1a in PDF/A-1b.

	ISO 19005-1:2005: PDF/A 1a (Level A)	ISO 19005-1:2005: PDF/A 1b (Level B)
Cilj	Ustvariti arhivske PDF dokumente in zagotoviti polni dostop do celotne vsebine.	Ustvariti arhivske PDF dokumente in zagotoviti

		reprodukcijo vidnih elementov (slike).
Metapodatki	Podatki, kot so avtor, naslov, datum nastanka, vir, itd., morajo biti v zapisu, združljivim z XMP (Extensible Metadata Platform. http://www.adobe.com/devnet/xmp/)	
Zaščita in varnost	Nastavitve za varnost niso dovoljene, dokument mora biti brez zaščite. Dovoljen mora biti vpogled in uporaba dokumenta brez omejitev.	
Barve	Barvne sheme morajo biti definirane.	
Kompresija	LZW kompresija ni dovoljena, JPEG2000 kompresija ni dovoljena.	
Transparentnost	Ni dovoljena.	
Uporaba PDF slojev	Niso dovoljeni.	
Pisave	Pisave, uporabljene v dokumentu, morajo biti vključene v dokument.	
	Za vsak uporabljen znak v dokumentu mora obstajati Unicode preslikava	
Opombe	Opombe, predstavljene kot govor ali video, niso dovoljene.	
Reference (sklici / povezave)	Sklici na zunanje slikovno gradivo ali vsebino dokumenta niso dovoljeni.	
Programski jeziki	Vključevanje jezika JavaScript ni dovoljeno.	
Akcije	Določene akcije, kot npr. zagon filmskih ali zvočnih datotek, pošiljanje ali brisanje formularjev (forms) niso dovoljene.	

Tabela 1) Primerjava PDF/A 1a (Level A) in PDF/A 1b (Level B)

4.2 Open XML paper specification (Open XPS)

V primerjavi s standardom PDF, ki obstaja od leta 1993 in je postal standard ISO leta 2001 za verzijo PDF/A, je format XPS [7] na voljo od konca leta 2006 kot del operacijskega sistema Microsoft Vista. XPS zapis opisuje vsebino, ki naj bi jo prikazal zaslon oz. natisnil tiskalnik. V tem smislu najbolj spominja na sodoben »spool« zapis, primerljiv z PCL (Printer Command Language), AFP (IBM Advanced Function Printing) in PostScript. Vsaj v operacijskem sistemu Vista in Windows7 si je te vrste datoteke moč ogledati tudi v Internet Explorer brskalniku. Glavni namen zapisa XPS je doseči večuporabnost dokumenta, ki ni vezan na programsko ali strojno opremo. Format Open XPS je postal leta 2009 ECMA standard.

Ali predstavlja XPS alternativo PDF/A?

Struktura formata XPS se v primerjavi z PDF/A izkaže za pomanjkljivo. Konverzija iz PDF v XPS je vedno neizgubna, obratno pa takšne vrste preslikava ni vedno mogoča. Glavne pomanjkljivosti formata XPS v primerjavi z PDF/A so naslednje:

- XPS ne omogoča funkcije pretiska (overprint),
- XPS ne dovoljuje uporabe uporabniško določenih grafičnih mask v dokumentih,
- XPS ne podpira JBIG2 procedure kompresije,
- Uporaba JPEG2000, ki prihaja s formatom PDF/A-2, ni podprta v XPS.

4.3 Open Document (Open Document Format for Office Applications - ODF)

Open Document Format [8] je odprt komprimirani datotečni format, ki je namenjen shranjevanju in izmenjavi pisarniških dokumentov, ki jih je moč urejati. Med take dokumente spadajo besedilni dokumenti (zabeležke, poročila, knjige in podobno), preglednice, predstavitve, podatkovne zbirke in grafikoni. Standard so razvili pri industrijskem združenju OASIS (organizacija za napredek standardov za strukturirane informacije) in temelji na datotečnem formatu na osnovi XML, ki so ga ustvarili pri OpenOffice.org. ODF je bil pri OASIS potrjen kot standard 1. maja 2005. Osnutek za standard ISO ISO/IEC 26300 je bil potrjen 3. maja 2006.

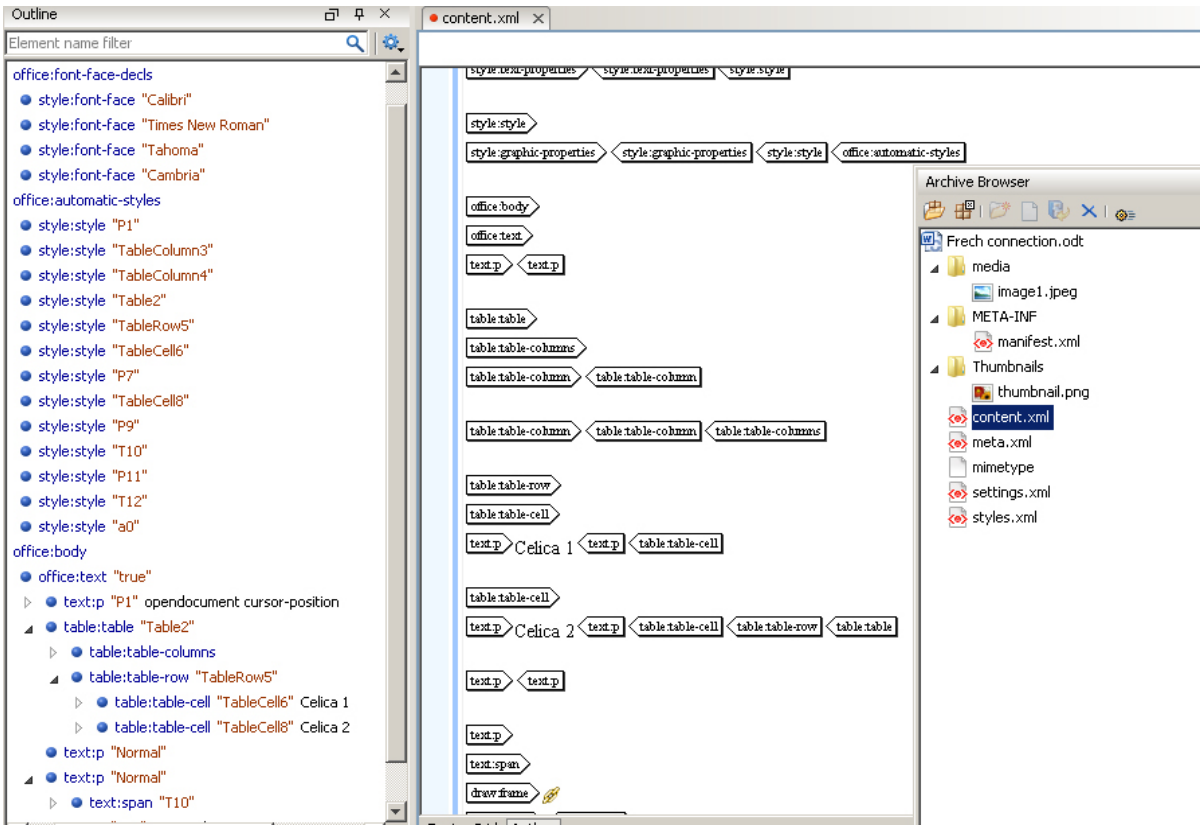
Standard je razvijalo več organizacij in je prosto dostopen, lahko ga uporabi in implementira kdorkoli, brez kakršnih koli omejitev. Format Open Document predstavlja odprto alternativo zaprtim, lastniškim formatom za dokumente.

Struktura dokumenta ODF je sledeča:

- koren dokumenta,
- metapodatki dokumenta (uporabljeni elementi Dublin Core),
- telo dokumenta (telo besedila, preglednice, prezentacije, slikovni objekti,..),
- nastavitve aplikacije,
- skripte,
- deklaracije pisav,
- stili,
- definicija oblike strani dokumenta.

Več o strukturi odf dokumenta je na voljo tu
http://www.oasis-open.org/committees/download.php/12572/OpenDocument-v1.0-os.PDF

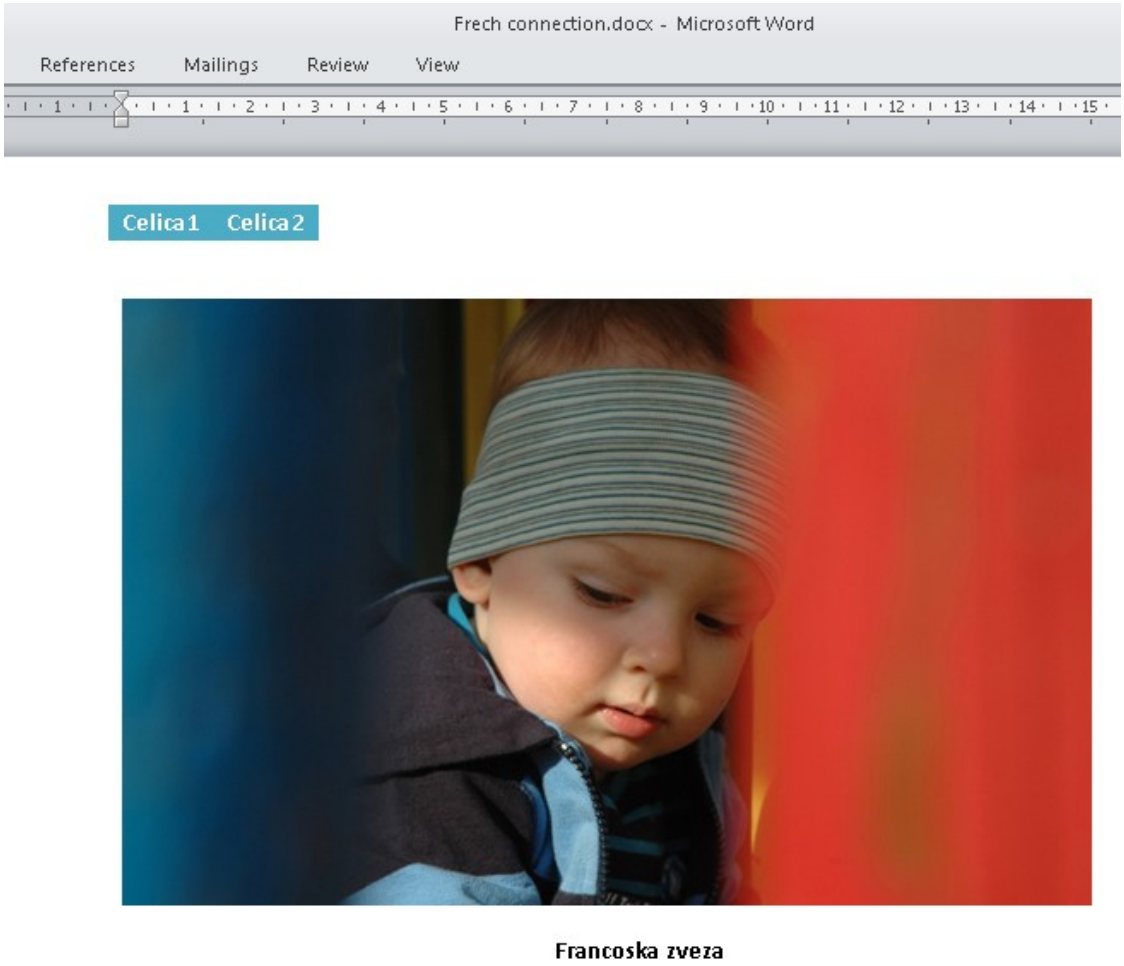
Na sliki 4.1 je prikazan pogled na odf datoteko, kot jo vidimo v XML urejevalniku (Oxygen XML editor). Kot vidimo, predstavlja odf datoteka komprimiran XML zabojnik, v katerem je vsebina.



Slika 4.1) Vsebina zabojnika odf

4.4 Office open XML (OOXML ali OpenXML)

Office Open XML [9] je datotečni format, strukturiran v XML in stisnjen (uporabljen ZIP algoritem). Po strukturi objektov v dokumentu je soroden formatu ODF. Razvit je bil s strani Microsofta in je prav tako kot ODF namenjen opisovanju preglednic, tabel, tekstov, prezentacij itd. Standardiziran je bil s strani ECME (European Computer Manufacturers Association), organizacije, ki deluje na področju standardov informacijsko komunikacijskih sistemov. Format OOXML je nadaljevanje razvoja XML zapisovanja podatkov, ki so ga uporabili že leta 2003 v zbirki Microsoft Office 2003. Verzija iz leta 2003 je vsebino – dokument zapisala v en XML dokument, kjer so si prostor delili tekstovni in binarni objekti – predstavljeni kot BASE64 nizi, vse brez uporabe kompresije. OOXML je v tem pogledu precej drugačen, saj datoteko s končnico .docx sestavlja komprimirana shramba objektov.



Slika 4.2) Pisarniška datoteka iz paketa Microsoft Office 2007

Za primer si poglejmo sliko (slika 4.2), ki predstavlja datoteko, katere vsebina je izjemno preprosta – tabela z dvema celicama, slika in napis pod njo. V prejšnjih verzijah Microsoft Office dokumentov vpogled v strukturo dokumenta (binarna) ne bi bil možen oz. ne bi bil uporaben. Z vpeljavo XML podatkovnega skladišča, kjer shranjujemo objekte, ki sestavljajo datoteko (publikacijo), pa postane vpogled v samo jedro dokumenta mogoč.

Strukturo dokumenta sestavljajo tri mape (_rels, docProps in word) ter datoteka, ki opisuje vsebino - [Content_Types].xml

V mapi **_rels** nastopajo seznam objektov, ki sestavljajo dokument. Dokument sestavljajo objekti, ki ga opisujejo in samo telo dokumenta.

docProps – mapa, ki jo sestavljajo datoteke, ki opisujejo dokument.

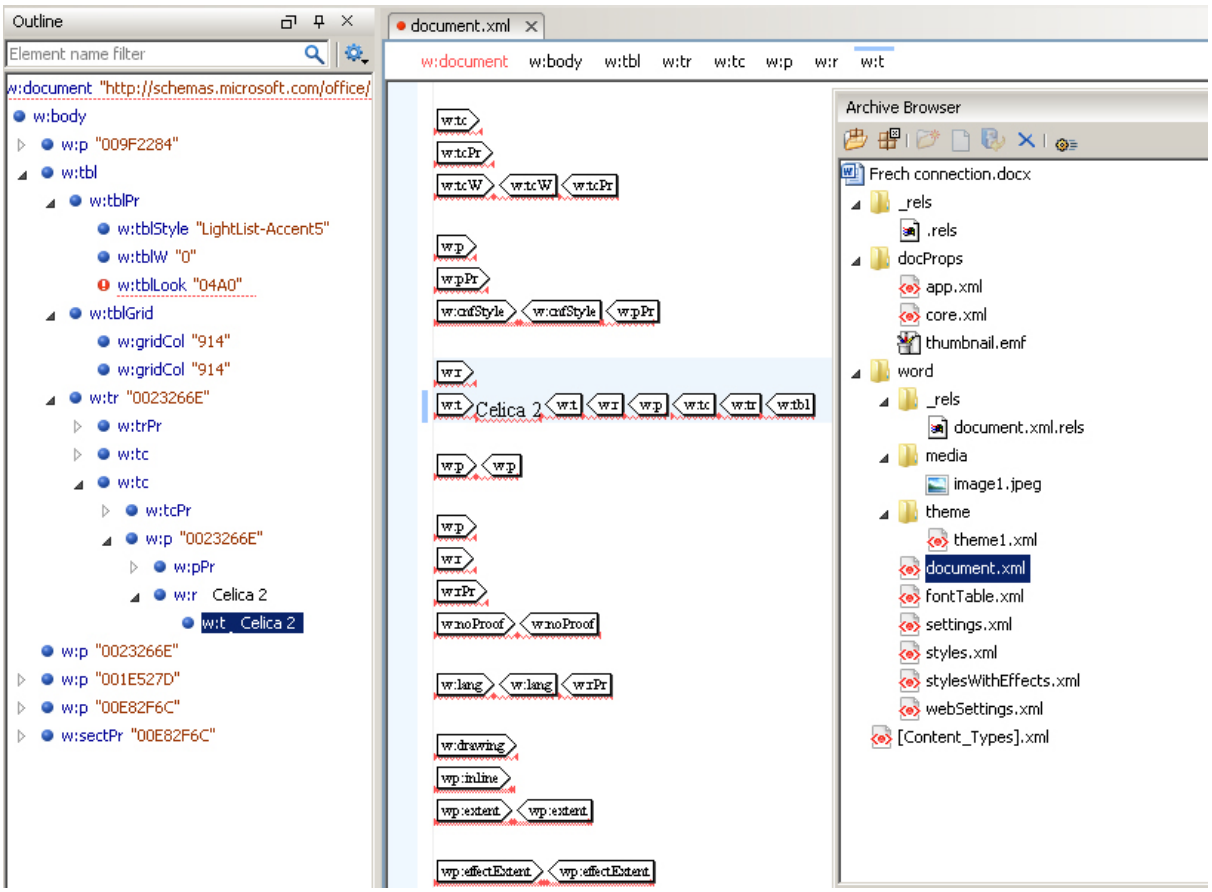
- opis aplikacije (verzija, podatki o zaščiti dokumenta, število znakov, strani v dokumentu, itd.).

- opis sistemskih podatkov (lastnik, zadnji urednik, datum posega v dokument, itd).

Za opisovanje je uporabljen Dublin Core z dodanimi lastnimi polji.

V mapi **word** se skriva jedro dokumenta z definicijami pisav, stilov, povezav na objekte, ki so del dokumenta npr.slike.

Dokument vsebuje npr. informacijo o uporabljenih pisavah v dokumentu. Z uporabo te informacije lahko določi, katere pisave uporabiti v primeru prikaza in uporabe pri uporabniku, ki podpore za v dokumentu uporabljeno pisavo nima. Slika 4.3 prikazuje sestavne dele Microsoft Office 2007 Wordove datoteke.



Slika 4.3) Vsebina zabojnika OOXML

Za razliko od formata ODF OpenXML ni ISO standard in po napovedi Gartnerja, to tudi ne bo postal (http://www.gartner.com/resources/140100/140101/iso_approval_of_oasis_opendo_140101.PDF). Gartner se namreč sklicuje na dejstvo, da je konzorcij OASIS sprejel ODF format kot uradni XML dokumentni format. V slabost formatu in standardu priča dejstvo, da gre za zaprt standard, razvit v enem podjetju, brez možnosti spreminjanja nadgradnje s strani zunanjih organizacij ali posameznikov.

Iz sveta besedilnih datotek za hip preidimo med tabele in opazujmo pristop zapisovanja le-teh skozi prizmo odprtokodnega ODF formata in Microsoft Office OOXML formata. Pri uporabi tabel (npr. v aplikaciji Excel) je moč nazorno prikazati različen pristop v shranjevanju in slabost uporabe lastnosti XML strukture, saj naj bi bila le-ta (že po definiciji) opisljiva - Extensible Markup Language, oz. označevalni jezik. Primer nakazuje na neprimerno označevanje zastavic - <c r="A1" t="s">, ki uporabniku ne povedo kaj dosti.

Primer zapisa v Excelu:

v celico A1 vpišemo »Črt« (brez narekovajev)
v celico A2 »Kragelj« (brez narekovajev)
v celico A3 vpišemo »Črt« (brez narekovajev)

ODS (Open Office Spreadsheet):

…

```
<table:table-row table:style-name="ro1">
    <table:table-cell office:value-type="string" table:style-name="ce1">
      <text:p>
        Črt
      </text:p>
    </table:table-cell>
</table:table-row>
```

```
<table:table-row table:style-name="ro1">
    <table:table-cell office:value-type="string" table:style-name="ce1">
      <text:p>
        Kragelj
      </text:p>
    </table:table-cell>
</table:table-row>
<table:table-row table:style-name="ro1">
    <table:table-cell office:value-type="string" table:style-name="ce1">
      <text:p>
        Črt
      </text:p>
    </table:table-cell>
</table:table-row>
```

OOXML:

Kot shrambo podatkov uporablja OOXML več datotek. V datoteko sheet1.xml zapiše med drugim naslednje:

```
…
<row r="1" spans="1:1" x14ac:dyDescent="0.25">
    <c r="A1" t="s">
        <v>
            0
        </v>
    </c>
</row>
<row r="2" spans="1:1" x14ac:dyDescent="0.25">
    <c r="A2" t="s">
```

```
        <v>
            1
        </v>
    </c>
</row>
<row r="3" spans="1:1" x14ac:dyDescent="0.25">
    <c r="A3" t="s">
        <v>
            0
        </v>
    </c>
</row>
```

V datoteko sharedStrings.xml pa

```
<sst xmlns="http://schemas.openxmlformats.org/spreadsheetml/2006/main" count="3" uniqueCount="2">
    <si>
        <t>
            Črt
        </t>
    </si>
    <si>
        <t>
            Kragelj
        </t>
    </si>
</sst>
```

OOXML predvideva večkratno uporabo **istih** podatkov v celicah. S tem, ko vrednosti ne zapisuje neposredno v datoteko, privarčuje na prostoru, saj namesto vrednosti ponuja referenco na objekt. Kaj pa velikost datoteke? Dolžini datotek zgoraj opisanega primera sta sledeči: 2.85kb za ODS in 8.59kb za OOXML. Glede na to, da je velikost preglednice zanemarljive velikosti (vsebina v le dveh celicah), na osnovi rezultatov tega primera ne

smemo sklepati na splošno kvaliteto algoritmov za kompresijo podatkov. Iz načina shranjevanja datotek v shrambo je videti, da bi moral biti OOXML format boljši, ko se vrednosti polj ponavljajo.

Primer 2:

Tokrat preglednico v Excelu napolnimo v večji meri - 26 x 10.000 polj, torej 26.000 celic. V vsaki je naključna vrednost RAND()*100, ki vrne v celico število v intervalu 0-100.

Evidentno je, da se bo veliko število vrednost celic ponovilo. Proti vsem pričakovanjem so velikosti datotek sledeče:

ODS - 4.302kb, OOXML - 5.029kb.

OOXML je navkljub nestandardnim XML poimenovanjem (krajši opisi v zastavicah) in uporabo sharedstrings.xml datoteke porabil več diskovnega prostora. Sharedstrings.xml datoteko OOXML uporablja le za pisanje **nizov**. V kolikor v celici-polju ne nastopa niz, nam koncept Microsofta v produktu Office 2007, 2010 ne koristi.

4.5 Kateri format uporabiti?

V prejšnjem poglavju smo na kratko predstavili nekatere izmed obstoječih formatov za shranjevanje publikacij. Glede na dejstvo, da je format PDF/A pridobil ISO standard kot arhivski format, bi morala biti izbira preprosta, a seveda ni.

Zapis v omenjenem formatu nam vsekakor zagotavlja dostop do dokumenta v prihodnosti – podjetje Adobe je programsko kodo za pregledovalnik javno objavilo in s tem vzpodbudilo razvoj različnih programskih rešitev za dostop do informacij v zapisu PDF. Prav tako nam omejitve pri izgradnji ali konverziji PDF ali drugih datotek v PDF/A format zagotavlja vedno enako informacijo – vedno berljivo (zaradi vključenih pisav) in nenazadnje zagotavlja nam reprodukcijo dokumenta – publikacije. Slabe strani takšne vrste zapisa pa so predvsem naslednje:

- odpovedati se moramo skriptnim dodatkom v dokumentu (javascript ni dovoljen);
- vključevanje postscripta ni dovoljeno;
- 3d objekti v dokumentu niso dovoljeni;

- omejena uporaba spletnih formularjev;
- akcije kot npr. zagon video vsebine, pošiljanje spletnih obrazcev (form) ni dovoljeno;
- prepovedana je lzw in jpeg2000 kompresija;
- prepovedana je zaščita dokumenta z geslom ipd.

Poleg omenjenih omejitev, ki neposredno vplivajo na izgradnjo PDF/A dokumenta, moramo v zakup vzeti še posredne težave. Ena glavnih težav je prav gotovo »interoperabilnost« vsebine. Kot primer vzemimo digitalno zbirko (repozitorij) dokumentov s področja znanosti. Čeprav lahko vsakega od dokumentov pregledujemo in natisnemo, pa ni moč (ponovno) uporabiti objektov, ki sestavljajo neko publikacijo. Med znanstveniki in raziskovalci na področju kemije, matematike, računalniške arhitekture, fizike in podobnih, v dokumentih (raziskavah, poročilih, rezultatih, nalogah,..) kar mrgoli matematičnih, kemijskih, fizikalnih in ostalih simbolov, ki sestavljajo različne formule, izjave, sklepe, teoreme, zakone ali pravila. Uporaba teh objektov – formul postane težavna iz vsaj dveh vidikov. Optično razpoznavanje besedila, ki ga nad publikacijo sicer lahko opravimo in ga shranimo v podatkovno zbirko za potrebe iskanja po polnem besedilu, nam seveda ne bo preveč koristilo, saj lahko upravičeno pričakujemo napake pri optični razpoznavi. Tudi po morebitnem uspešnem iskanju in identifikaciji dokumenta trčimo ob še večjo težavo – objekt je za nadaljnjo uporabo izgubljen, saj ni shranjen kot objekt znotraj PDF datoteke. Le stežka ga bomo brez težav izvozili v urejevalnik, kjer ga želimo uporabiti, mu morebiti spremeniti tip pisave, velikost, itd.

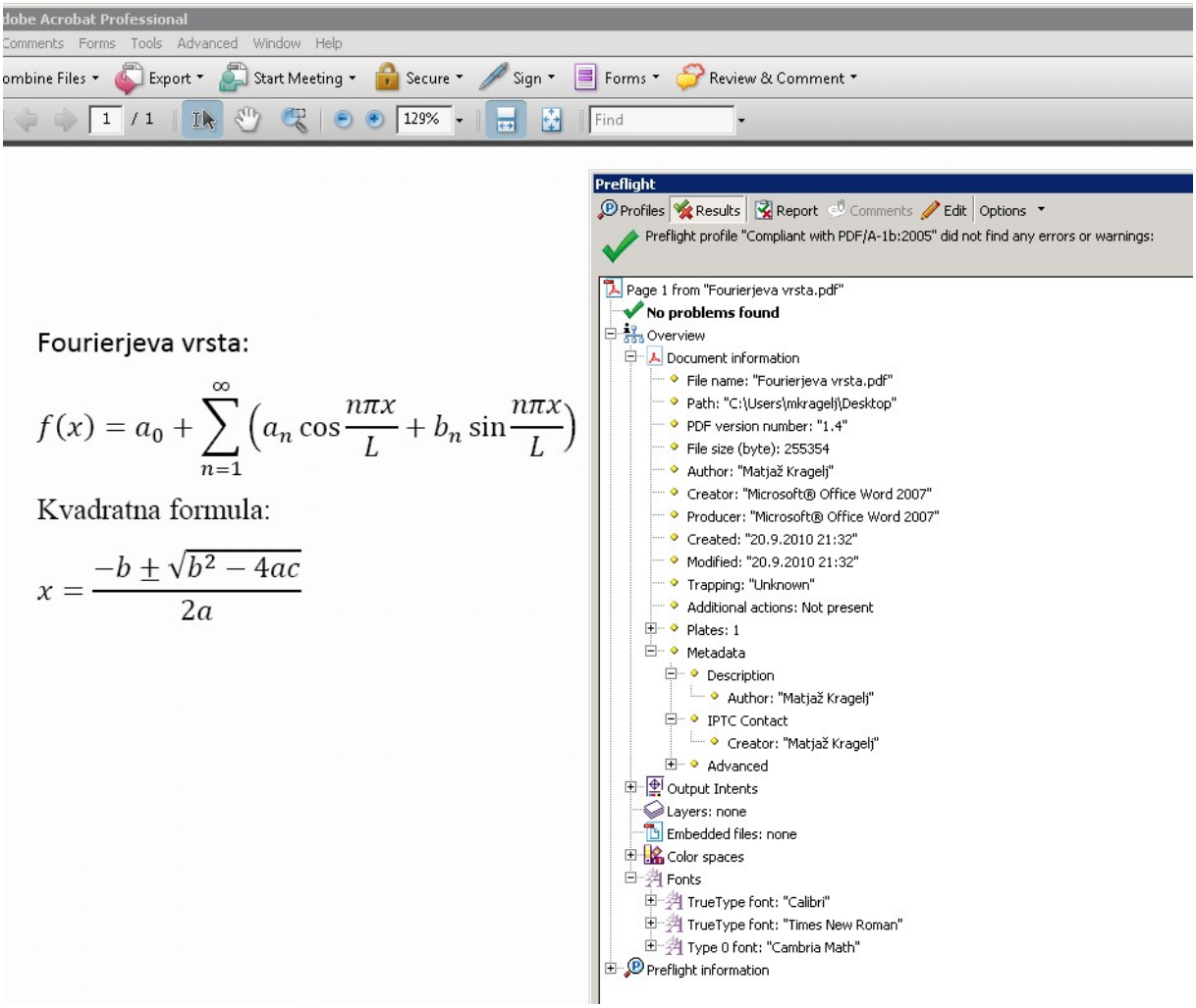
Primer:

Ustvarili smo dokument in vanj vnesli dve matematični formuli, dokument pa izvozili v PDF/A arhivski format - slika 4.4. Poleg videza datoteke v Adobe Acrobatu smo dodali rezultat testiranja z orodjem Preflight. Le-to preveri strukturo dokumenta in nas obvesti, ali dokument (v našem primeru PDF/A) zadosti potrebnim pogojem za takšno poimenovanje, samo orodje pa (če je mogoče) doda potrebne parametre in metapodatke dokumentu, da je skladen z definicijami formata (v našem primeru PDF/A). Iz slike je razvidno, da Preflight ni našel težav v datoteki tipa PDF/A, kar pomeni, da je naša datoteka pripravljena za arhiv.

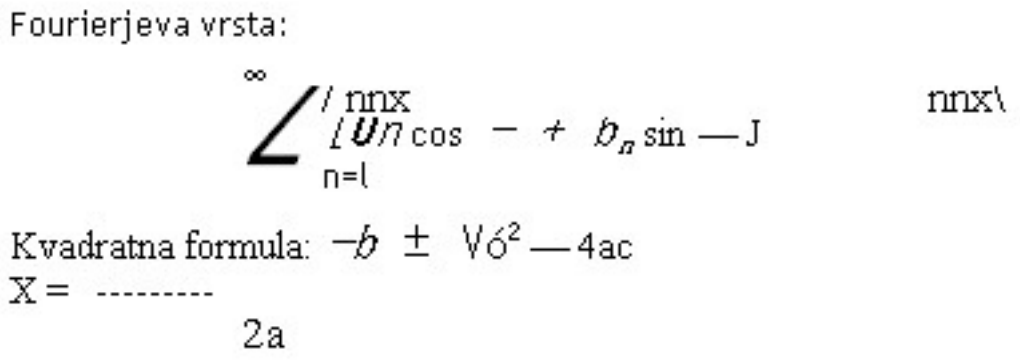
Kaj pa se zgodi, če želimo nad njo opraviti optično razpoznavo (OCR)?

S programom Abby Fine Reader, (ki sodi med najboljše in najbolj uporabljane na področju optične razpoznave besedila) opravimo optično razpoznavo nad datoteko, vsebino datoteke

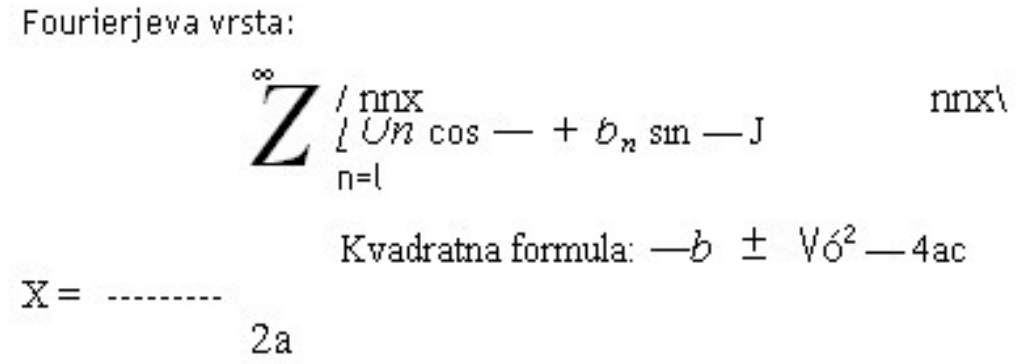
PDF (ki sedaj vsebuje tudi besedilo za sliko – text under image) pa preko odložišča prenesemo v urejevalnik – npr. Word 2010. Tako elementi besedila (matematična formula) kot umestitev razpoznanega besedila v odložišče in prezentacija le-tega končnemu uporabniku predstavljajo izredno veliko težavo za aplikacijo Abby Fine Reader. Rezultat je viden na sliki 4.5. Težavo poskusimo odpraviti na drug način. Po postopku optične razpoznave rezultat v aplikaciji Abby Fine Reader shranimo neposredno v Word 2010 format (.docx). Rezultat je viden na sliki 4.6. Poskusimo še tretji prijem. Po konverziji datoteke iz Word urejevalnika v PDF/A (kot vemo, datoteka te vrste ni zaščitena pred kopiranjem) poskusimo vsebino neposredno naložiti v odložišče (ctrl+a, ctrl+c, ctrl+v). Rezultat je prikazan na sliki 4.7.



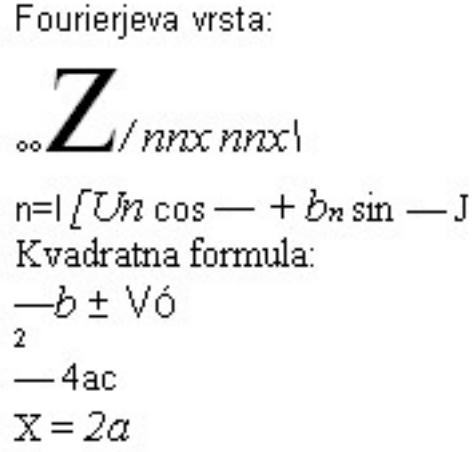
Slika 4.4) PDF/A Dokument, v katerem sta dve matematični formuli



Slika 4.5) Vsebina optično razpoznane PDF/A datoteke, zajem preko odložišča



Slika 4.6) Izvoz iz aplikacije Abby Fine Reader v Microsoft Word 2007-2010



Slika 4.7) Izvoz vsebine iz aplikacije Abby Fine Reader v PDF prek odložišča

Čeprav je format PDF/A ISO standard za področje arhiviranja, nam prikazani primer ilustrira, da zgolj arhivska kopija publikacije **ne zadošča** za njeno uporabo v **raziskovalne** namene.

Dostop do objektov, ki niso zgolj besedilo, je za aplikacije, ki razpoznavajo besedilo, sila trd oreh. Ena od možnih rešitev, ki omogoča uporabo sestavnih delov dokumenta, je nedvomno zagotavljanje arhiviranja originalnih datotek – datotek, v katerih je avtor oddal publikacijo. Omenjena rešitev je zagotovo vedno izhod v sili, saj bo na tak način arhiviranja do datoteke vedno moč dostopati in jo uporabiti in urejati. Slabost takega načina razmišljanja predstavlja zagotavljanje programske opreme, ki omogoča to početje. Kot primer naj navedemo avtomatično **nezdružljivost nazaj** pri Microsoft Office izdelkih. Brez posebnih vmesnikov ni moč dostopati do vsebine datotek pisanih v starejših verzijah tega produkta. Upoštevajoč dejstvo, da gre za produkte istega podjetja, je skrb vzbujajoče razmišljanje, kako dostopati do informacij ustvarjenih iz drugih aplikacij, ki jih (morda) sploh ni več na tržišču.

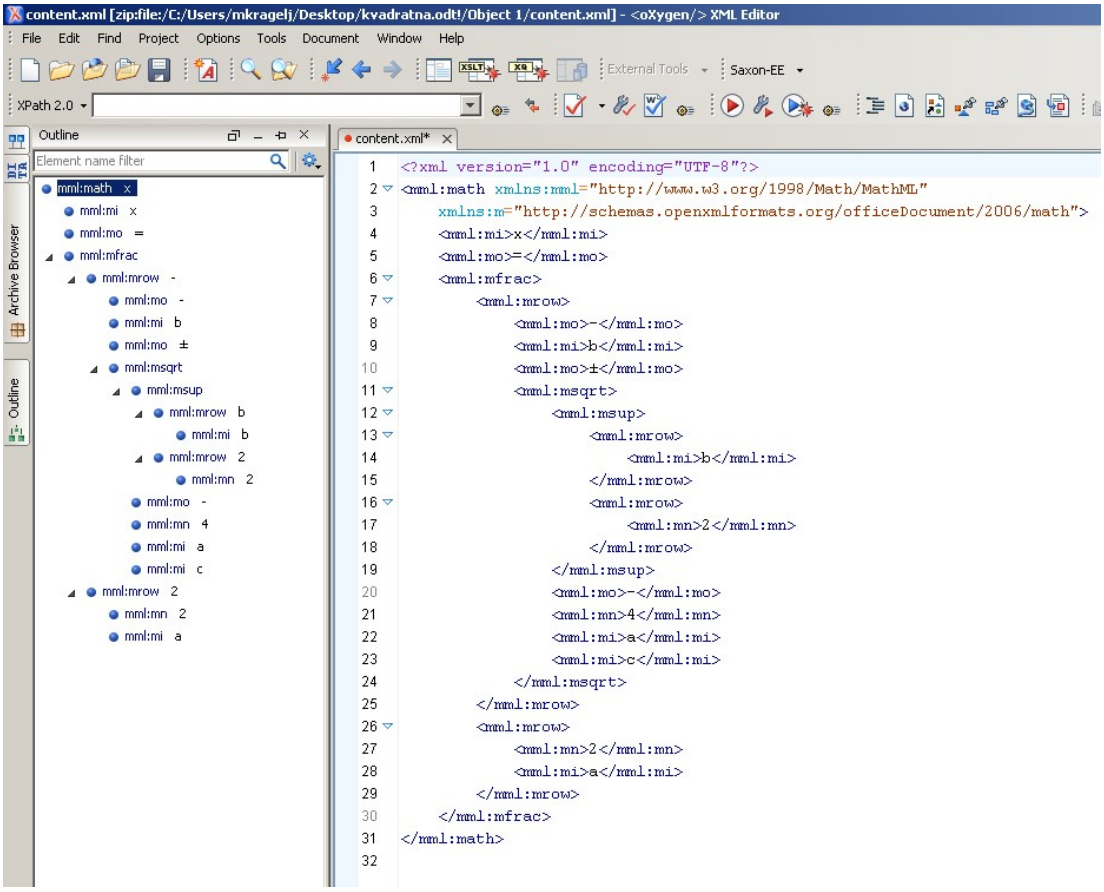
Z možnostjo shranjevanja publikacij v ODF zapis z dnem, ko je postal format ISO standard (november 2006), ali v Microsoftovo OOXML obliko (januar 2007), se je uporabnikom ponudila možnost, ki je z Adobe PDF/A ni moč doseči. V strukturiranem XML zapisu je shranjena informacija, ki uporabniku omogoča izvoz in ponovno uporabo delov publikacije – objektov. Kot objekt publikacije mislimo v tem poglavju npr. sliko, (matematične) formule, simbole itd. V spodnjem primeru prikazujemo način shranjevanja takih objektov in možnost dostopa do njih za uporabo.

Primer:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

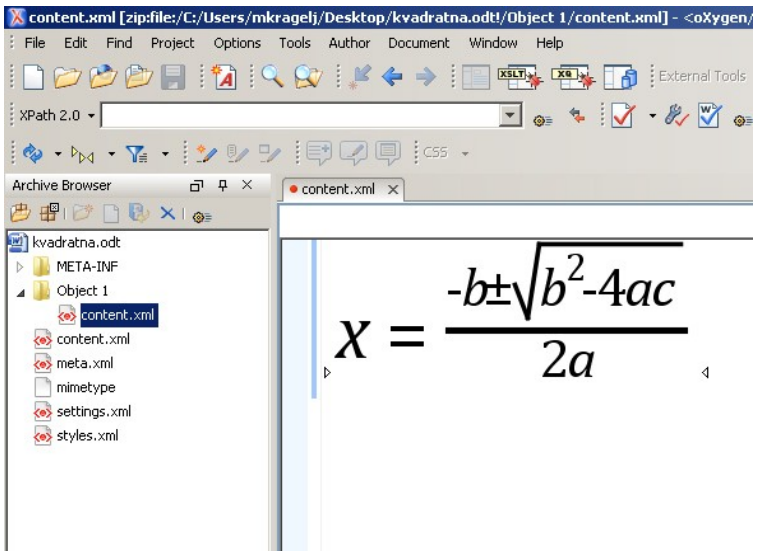
Slika 4.8) Formula kvadratne enačbe napisana v MS Word 2010

Če dokument z vsebino iz slike 4.8 shranimo v Open Document Format in si ga ogledamo v XML editorju (uporabljen je <oXygen/> XML Editor 12.0), dobimo naslednjo XML vsebino objekta – slika 4.9.



Slika 4.9) XML zapis matematične formule iz slike 4.8

XML urejevalnik nam dobljeno XML datoteko interpretira tako, kot prikazuje slika 4.10.



Slika 4.10) Interpretacija XML zapisa matematične formule

Do podobnega rezultata pridemo z uporabo Office Open XML formata.

Objekti - deli dokumentov so torej shranjeni v XML datotekah (vsak objekt zase). Z objekti lahko poljubno manipuliramo v urejevalniku, ki podpira delo z XML dokumenti. Rezultat XML datoteke v spletnem brskalniku je zaradi sheme, ki jo ima definirano in jo uporablja za interpretacijo –

<mml:math xmlns:mml="http://www.w3.org/1998/Math/MathML"
xmlns:m="http://schemas.openxmlformats.org/officeDocument/2006/math">

kar formula sama – izpis v matematični obliki.

Opisani primer nazorno kaže težave in pasti arhivskega formata in ponuja možno rešitev za dostop in uporabo do objektov dokumenta. Preprost razmislek nam pove, da je pravo razmišljanje v smeri shranjevanja tako arhivskega formata za shranjevanje publikacij na eni strani kot zapisovanje le-teh v obliko, ki uporabniku ponuja manipulacijo z njenimi sestavnimi deli na drugi strani. Z arhivskim formatom dosežemo možnost reprodukcije publikacije, njeno trajno ohranjanje, videz dokumenta, kakršen je bil ob njegovem nastanku in uporabo za prihodnje rodove, ter zapis v obliko, ki nam omogoča zajem in manipulacijo z njeno vsebino – uporabo v raziskovalne namene, oz. ponovno uporabo, kar je eden glavnih ciljev referenčnega modela OAIS.

V prihodnjih poglavjih se bomo najprej dotaknili bibliografskih podatkov in vlogi, ki jo imajo kot sestavni del publikacije, kasneje pa še načinu shranjevanja publikacije v repozitorij – shrambo digitalnih vsebin in prednosti, ki nam jih tak način dela z dokumenti prinaša.

5 BIBLIOGRAFSKI PODATKI IN NJIHOVA IZMENJAVA

Elektronski dokument, ki je lahko predstavljen na spletu - spletni dokument, digitalno rojen dokument, dokument, ki je nastal na računalniku in digitalizirana kopija fizičnega objekta lahko sestavlja **n** objektov. Pod terminom objekt na tem mestu razumemo digitalni zapis, ki sestavlja publikacijo, je torej njen sestavni del. Publikacija je lahko predstavljena v tiff, jpeg, jpeg2000 formatu, lahko je PDF ali PDF/A dokument, lahko je wav ali mp3 datoteka, v kolikor imamo opravka z zemljevidi, imamo na voljo med drugimi format mrSid itd. Objekte lahko shranjujemo v visoki ločljivosti (vsaj 300dpi), predvsem z namenom arhiviranja, kot primarni vir za potrebe trajnega hranjenja, ponovne uporabe, tiska. S tem želimo zavarovati publikacijo pred ponovno digitalizacijo, izpostavitvi neugodnim pogojem (npr. stare rokopise). Poleg visoko ločljivih zapisov publikacije shranjujemo objekte publikacije v nižji ločljivosti (72-120) za javno uporabo – preko spletnih portalov, potreb pošiljanja preko e-pošte ali za druge načine uporabe.

Pri shranjevanju dokumentov oz. njegovih sestavnih delov – objektov ne smemo pozabiti na morda najpomembnejši del publikacije – bibliografske podatke. V knjižničnem katalogu COBISS se za opisovanje bibliografskih podatkov uporablja format COMARC (izvedenka formata UNIMARC). Podatkovna baza vsebuje več kot štiristo polj. Na srečo lahko bibliografske podatke enoumno in dovolj jasno opišemo s strukturo Dublin Core [10] formata. Format zapisa je ISO standard (ISO 15836:2003). V osnovi predstavlja Dublin Core besednjak petnajstih ključnih elementov (contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, type), katerih vrednosti enolično opisujejo publikacijo. Ime »Dublin« nosi po mestu Dublin, Ohio, kjer je nastal, »core« pa zaradi dejstva, da taka struktura metapodatkovnega opisa (v knjižničnem prostoru – bibliografskih podatkov) predstavlja širok, splošen in generičen zapis, ki je lahko uporaben za opisovanje obširnega nabora različnih tipov entitet.

Ta način opisovanja se je do sedaj v izmenjavi podatkov med partnerji izkazal za zelo uporabnega, saj je razmeroma preprost, tako za kreiranje in urejanje zapisov, kot tudi za uvoz oz. izvoz podatkov. Pomanjkljivost oz. slaba lastnost je natanko to – njegova preprostost. Pri uporabi tega formata za strukturiranje metapodatkov se zavestno odločimo zavreči del bogatega nabora bibliografskih informacij, ki lahko opisujejo neko publikacijo (če je le-ta

zavedena v kataložnem sistemu COBISS, ali pač v kakem drugem knjižničnem sistemu, ki ga organizacija uporablja za opisovanje elektronskih vsebin).

Do nedavnega so bili metapodatki v obliki Dublin Core (DC) zadosten vir informacij za potrebe posredovanja dostopa do gradiva preko neke partnerske organizacije (npr. Evropske knjižnice – The European Library). Sčasoma se je izkazalo, da opisovanje publikacij s takšnim naborom informacij ne zadošča več.

Razlogi:

- Metapodatki pridobljeni iz različnih virov, so uporabljeni za različne namene (različne tipe objektov).
- DC shema pogostokrat ne zadošča za nedvomni opis publikacije.
- Različni ponudniki vsebin publikacij uporabljajo različne meta podatkovne sheme.

Možna rešitev:

- Ponudniki vsebin publikacij (npr. Europeana) ponujajo razširljivo DC shemo (npr. DC Extended).

Čeprav se je format DC oz. DC extenended izkazal za trenutno najprimernejšega pri opisovanju publikacij za potrebe izmenjave bibliografskih vsebin, je potrebno zaradi nepredvidljive prihodnosti na področju standardizacije in zahtev ponudnikov vsebin zagotavljati čim bolj polne / bogate zapise. Poleg formata Dublin Core je torej potrebno shraniti dodane opise dokumenta – publikacije, oz. shraniti celotni-popoln originalni zapis opisa publikacije – npr. zapis v knjižničnem katalogu COBISS. Iz zapisa je vedno mogoče opraviti preslikavo v format DC, DC extended, ISO 690, ISBD, ESE (http://www.europeana.eu/schemas/ese/ESE-V3.2.xsd), itd.

6 PROCES IZMENJAVE BIBLIOGRAFSKIH VSEBIN

S petim okvirnim programom, ki je trajal med leti 1998 in 2002 (http://cordis.europa.eu/fp5/)- so Evropske nacionalne knjižnice s sredstvi Evropske unije znotraj petega okvirnega programa pričele organizirano izvajati digitalizacijo in urejati vprašanja o trajnem ohranjanju digitalnih virov. Portal Evropske knjižnice (TEL) (http://www.theeuropeanlibrary.org) je bil prvi korak v smeri izmenjave bibliografskih vsebin in trajnega ohranjanja, hkrati pa predstavlja predvsem mejnik v sistematizaciji in obvladovanju e-vsebin, s katerimi se knjižnice resneje ukvarjajo šele zadnja leta.

Evropska knjižnica – TEL je spletni servis, ki ponuja brezplačen dostop do virov oseminštirideset nacionalnih evropskih knjižnic v petintridesetih jezikih. Osnovna ideja spletnega portala evropske knjižnice je na enem mestu ponuditi dostop do digitalnih objektov in knjižničnih katalogov vseh evropskih nacionalnih knjižnic. Evropska knjižnica tako omogoča dostop do dveh različnih tipov vsebin:

- dostop do vsebine – bibliografskih zapisov – knjižničnih katalogov (z uporabo protokola Z39.50 in SRU/SRW);
- bibliografske podatke digitalnih objektov ter povezave na te objekte – če je mogoče, preko protokola OAI-PMH.

6.1 Stanje v Sloveniji

V letu 2005 je IZUM knjižnicam ponudil dostop do strežnika z nameščenim Z39.50 servisom in s tem ponudi možnost dostopa do lokalnega knjižničnega kataloga. S tem je bil dosežen eden od temeljev za izmenjavo in ponovno uporabo bibliografskih podatkov ter njihova interoperabilnost. Standard Z39.50 je bil sprejet leta 1988 in doživel nekaj sprememb (leta 1992, 1995 in 2003) in čeprav počasen in okoren, kot ga nekateri radi cinično označujejo, je postala prav izmenjava, bolje rečeno enosmerni tok bibliografskih podatkov (iz kataloga knjižnic je preko tega protokola dovoljeno le brati) ključna za gradnjo lokalnih spletnih servisov z bibliografsko vsebino. Narodna in univerzitetna knjižnica je s spletnim mestom Digitalne knjižnice Slovenije (http://www.dlib.si) in uporabo protokola Z39.50 na relaciji IZUM – NUK uresničila to interoperabilnost v praksi že novembra leta 2005.

Open Archives Metadata Harvesting Protokol (OAI-PMH) je protokol za izmenjavo metapodatkov. Trenutno je v uporabi verzija 2.0. Evropska komisija, ki medtem gradi nov spletni servis – Europeana (http://www.europeana.eu), kjer poleg gradiva nacionalnih knjižnic Evrope vključuje tudi zbirke muzejev, arhivov idr., pričakuje od članice (npr. knjižnice), da preko spletnega portala deluje kot OAI repozitorij. Vsaka entiteta (knjižnica, muzej, arhiv, itd.) služi kot strežnik, s katerega Evropska knjižnica črpa informacije (metapodatke) in gradi centralni indeks kot jedro centralnega spletnega portala.

Postopek lahko strnjeno opišemo na naslednji način: knjižnica generira bibliografske podatke v obliki, primerni za uporabo z OAI-PMH protokolom, torej sintaktično pravilne, ter shranjene v XML zapisu, roboti Evropske knjižnice oz. Europeane pa v določenih časovnih intervalih zajemajo ("harvestirajo") tako pripravljene podatke. Meta podatkovnega standarda za uporabo pri OAI-PMH protokolu ni. Tudi TEL ne daje strogih navodil, v kakšni obliki naj partnerji ponujajo bibliografske podatke svojih publikacij. Predlagajo svojo različico DC, ki se imenuje TELAP ali pa enega od MARC formatov, pri čemer dajejo prednost MARCXML obliki (to je zgolj standardiziran MARC zapis v XML obliki s predpisano XML shemo).

TEL je razvil rešitev za vse partnerje, in sicer OAI-PMH servis - programsko arhitekturo, ki omogoča posredovanje metapodatkov centralnemu strežniku. Prednost tega strežnika ob nastanku je bila v tem, da se instituciji-partnerju ni bilo potrebno ukvarjati s pretvorbo metapodatkov, saj strežnik omogoča uvoz metapodatkov iz skoraj vseh metapodatkovnih shem (DC, DCq, TELAP, MARC, MARCXML) in jih pretvarjal v Dublin Core format (DC). S povečevanjem števila partnerskih organizacij (predvsem nacionalnih knjižnic), ki so zapise o gradivu na ta način posredovale Evropski knjižnici, mnogovrstnosti gradiva v fondih le-teh, dodatnih storitvah, ki jih je Evropska knjižnica preko spletnega portala pričela ponujati uporabnikom, je postala potreba po razširjeni shemi potrebnih atributov za opisovanje publikacij (kot smo že omenili) nujna. Nabor potrebnih informacij, ki sestavljajo opis vsake publikacije, se je sicer povečal, a za prikaz ponazoritve principa delovanja bistveno ne vpliva.

Vsak zapis, strukturiran v Dublin Core obliki, predstavljeni v XML datoteki, je sestavljen iz treh delov:

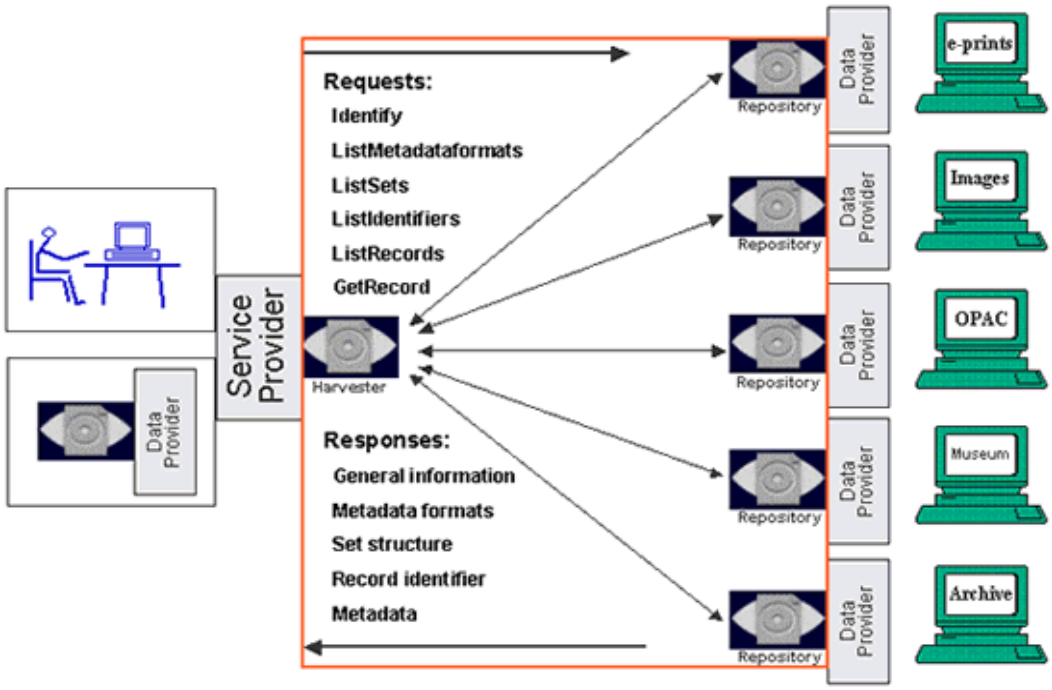
- **glave** – vsebuje vsaj identifikator OAI elementa, lahko pa še dodatne elemente, ki pospešijo zajem zapisov – npr. "časovni žig", ki omogoča selektivni zajem, ter druge statusne atribute;

- **metapodatkov** – vsebuje metapodatkovni zapis v Dublin Core formatu – lahko tudi v drugih ali zgolj razširjen DC;
- **("about")** – opcijsko. Lahko vsebuje dodatne informacije o zapisu ali dodatno razlago vrednosti metapodatkov.

Lahko bi rekli (v primeru, ko se uporablja za opis knjižničnega gradiva), da predstavlja Dublin Core podmnožico obsežnejšega in natančnejše strukturiranega knjižničnega formata COMARC. Kot tak je namenjen predvsem predstavitvi metapodatkov (digitalnih) objektov na spletnih portalih.

POTEK "POGOVORA" OAI-PMH STREŽNIKA IN KLIENTA IN PRIMER ZAPISA OBJEKTA V DC FORMATU.

Potek zajema podatkov se prične s podano zahtevo na OAI-PMH strežnik. Ta vrne rezultat v eni od vnaprej določenih oblik v XML formatu. Trenutno TEL uporablja svojo različico DC meta podatkovne sheme TELAP, ki vsebuje razširjen nabor polj iz DC. Spletni portal Europeana uporablja zapis oz. shemo ESE. Ker so podatkovne zbirke lahko z vidika prenosa preko spleta precejšnje in bi izgradnja in prenašanje velikih XML datotek lahko predolgo trajala, se prenos zbirke razbije v poljubno število nizov, vsak niz pa se konča s poljem (resumption token), s katerim strežniku pošljemo zahtevo za naslednji niz zapisov. Če polja ni, aplikacija za zajem predvideva, da je prišla do konca podatkovne zbirke in preneha z zbiranjem podatkov. Zajem poteka prek protokola http in potreben čas za prenos podatkov je sorazmeren z velikostjo zbirke, ki jo prenašamo. Na sliki 6.1 se nahaja ponazoritev servisa OAI-PMH, ki zajema vsebine različnih ponudnikov, in jih na enem mestu posreduje uporabniku.



Slika 6.1) OAI-PMH servis

Tako pripravljeni viri predstavljajo interoperabilnost med ponudnikom informacije in odjemalcem - v našem primeru evropskim knjižničnim spletnim portalom. Na ta način se zgradi centralni spletni indeks, ki na enem mestu omogoča iskanje po naboru vseh digitalnih zbirk vseh evropskih knjižnic, muzejev, arhivov in sorodnih inštitucij, ki so gradivo posredovale.

Hitrost iskanja po virih zbranih na ta način in katalogih evropskih knjižnic je neprimerljiva. Pri digitalnih zbirkah je na spletnem portalu Evropske knjižnice centralni indeks v vlogi podatkovne zbirke, zato so iskalni časi nizki, dostop do metapodatkov pa hiter in zanesljiv, kar ne moremo trditi za iskanje po katalogih. Kot smo že omenili, je pri iskanju po katalogih potrebna vzpostavitev povezave z oddaljenimi strežniki, iskanje, prenos informacij s pomočjo protokola Z39.50, SRU ali SRW in predstavitev oziroma posredovanje uporabniku.

Partnerske organizacije pri gradnji in razvoju Evropske knjižnice poskušajo časovno pregrado premostiti tudi z gradnjo centralnega indeksa katalogov. V časovno manj uporabljenih terminih uporabe knjižničnih katalogov (npr. ponoči) se sproži zajem kataloga (vseh novih oziroma spremenjenih zapisov) in gradi predpomnjen indeks. Tako imajo uporabniki dostop do nabora bibliografskih zapisov publikacij obdelanih do prejšnjega dne, kar je cena za pospešitev iskanja.

Bibliografske zapise digitalnih vsebin bi prav tako lahko s pomočjo protokola OAI-PMH prenašali na slovenske spletne iskalnike, npr. Najdi.si in gradili oziroma dodajali v centralni indeks portala.

7 SHRANJEVANJE PUBLIKACIJ – E-VSEBIN

V začetku diplomskega dela smo govorili o zajemanju spleta, veliki količini dokumentov, ki jih na ta način zajemamo v shrambe (.arc, .warc datoteke), nato smo se posvetili najprej formatom za zapisovanje publikacij za trajni dostop in uporabo, nazadnje še bibliografskim podatkom, njihovem izvozu in posredovanju centralnim indeksom, ki ponujajo centralno iskanje, vzdržujejo podatkovno zbirko redundančnih podatkov (vsaka izmed partnerskih institucij ima vsebine še vedno pri sebi in jih preko svojih vmesnikov ponuja javnosti), ostala nam je še najbolj zahtevna naloga – shranjevanje publikacij.

Publikacije je potrebno shraniti na način, da zadostijo vsaj naslednjim zahtevam:

- informacija ostane celotna – popolna,
- uporabniku nudi uporabno vrednost – ponovno uporabnost,
- publikacije ni potrebno ponovno digitalizirati,
- publikacija je najdljiva in enolično določena,
- uporabniku je dostop do nje zagotovljen in omogočen,
- zagotovljeno je njeno trajno ohranjanje.

Na tržišču je brez večjih težav v vsakem trenutku moč najti podjetja, ki se ukvarjajo z razvojem in implementacijo informacijskih sistemov, ki to omogočajo. Še lažje je poiskati podjetje, ki bi nam takšne vrste aplikacijo razvilo, povsem po našem okusu in željah.

Na trgu je ta čas moč zaslediti več "velikih" ponudnikov – več platform, ki so razvite in podprte do te mere, da se uporabljajo za potrebe trajnega ohranjanja digitalnih vsebin v produkcijske namene.

Pri raziskavi trga smo se tako omejili na "največje tri", to so Fedora Commons, dSpace in Eprints. Izbira pravega produkta za gradnjo arhiva vsebin ni preprosta naloga, saj ni zaslediti enotnih ocen glede posameznega produkta izmed teh.

Vse tri rešitve so v praksi močno zastopane in uporabljene. Vse tri so prosto dostopne in odprtokodne in imajo širok krog razvijalcev sistema. Nekatere večje razlike med njimi so:

Dspace: [11] trenutno najbolj uporabljana in razširjena platforma. V tem trenutku ima največ aktivnih inštalacij, sam produkt pa je deležen neprestanih posodobitev. Malo za njim je po

implementacijah zastopan **EPrints** [12]. V prid Dspace govori med drugim dejstvo, da je realizacija platforme podprta na več odprtokodnih podatkovnih bazah. (EPrints – MySQL, Dspace tudi PostgreSQL). Medtem ko Eprints sloni na programskem jeziku Perl, sta Dspace in Fedora javanska produkta. Vsi trije produkti so testirani in inštalirani tudi na Windows platformi. Če lahko produkta EPrints in Dspace štejemo kot paketa »out of the box« - v paketu je inštalacija platforme in vmesnikov, je šel razvoj **Fedore** [13] v drugo smer. Produkt vsebuje zgolj inštalacijo platforme in podatkovno bazo (Derby), uporabnik pa si lahko izbere tudi katero izmed drugih prosto dostopnih, odprtokodnih podatkovnih baz (MySQL, PostgreSQL), poleg tega pa tudi MSSQL in Oracle. Produkt vsebuje le administratorski modul, ki omogoča dostop do repozitorija, saj pričakuje od uporabnika razvoj ali uporabo katerega od že obstoječih »front-end« vmesnikov. Tak vmesnik preko ponujenih Fedora API-jev dosega z uporabo HTTP, SOAP in REST protokolov repozitorij, iz katerega črpa in prikazuje podatke na portalu ali jih ponuja naprej, ter polni podatkovno zbirko. Za Fedoro je na spletu dostopnih nekaj prostodostopnih aplikacij tipa »Front-end«, ki so pri ostalih dveh (Eprints, Dspace) že vključeni v produkt.

Izmed treh produktov smo se odločili za opis Fedore, ki ponuja največ, čeprav je proces njene prikrojitve najtežji in najbolj zamuden.

Fedora predstavlja kratico za *Flexible Extensible Digital Object Repository Architecture.* Je platforma, ki omogoča shranjevanje digitalnih objektov, vsebin z namenom trajnega hranjenja, dostopanja in urejanja le-teh. Je ena od treh najbolj znanih, priljubljenih in uporabljanih sistemov za takšne potrebe, odlikuje jo visoka skalabilnost, saj zmore hraniti in upravljati več kot deset milijonov objektov. Ker je sistem Fedora načrtovan predvsem za trajno ohranjanje digitalnih virov, je seveda skladen z referenčnim modelom OAIS.

V tem poglavju bomo definicijo objekta razumeli v naslednjem kontekstu. Objekt predstavlja zapis dokumenta ali zgolj njen del v digitalni obliki, ki je računalniško berljiv.

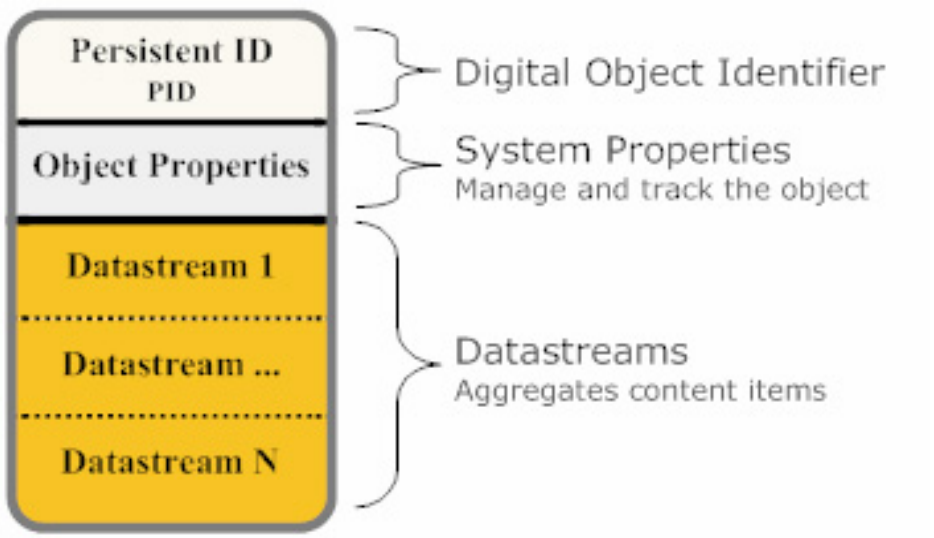
Kot primer uporabe dela s publikacijo vzemimo monografijo "Ivan Cankar: Hlapci"

Za potrebe trajnega hranjenja in zagotavljanja trajne dostopnosti smo se odločili shraniti naslednje zapise publikacije oz. njenih objektov:

- **arhivski PDF/A objekt** z opravljeno razpoznavo besedila
  (za uporabo uporabnikom ter za trajno hranjenje);

- **txt datoteko** (za izgradnjo celotnega besedila - uporabo iskanja po celotnem besedilu ter za predogled publikacije);
- **skenograme** v visoki ločljivosti (za morebitno potrebo po rekonstrukciji publikacije);
- **popoln bibliografski zapis**, ki ga dobimo v katalogu COBISS (shranimo popolno informacijo o publikaciji – vse atribute, ki jo opisujejo);
- **zapis v DC formatu** (za skrajšan opis publikacije, za nudenje informacije o publikaciji drugim spletnim servisom, digitalnim knjižnicam).

Fedora si podatke o publikaciji in njenih sestavnih delih shranjuje v podatkovni model (poenostavljen prikaz) kot ga prikazuje slika 7.1



Slika 7.1) Pogled na podatkovni model za publikacijo

Publikacija shranjena v Fedori je praviloma sestavljena iz treh delov in sicer iz:

- enoličnega identifikatorja (Persistent ID PID),
- sistemskih lastnosti, potrebnih za sledenje in upravljanje publikacije v repozitoriju (Object Properties) ter
- nizov podatkov oz. objektov publikacije (Datastreams)

Podatki so shranjeni v shrambi, ki jo pri Fedori poimenujejo FOXML (Fedora Object XML). Gre za format, ki neposredno predstavlja Fedorin podatkovni model. Preprosta ponazoritev zapisa FOXML bi bila takšna:

```
<digitalObject PID="uniqueID">

        <!-- there are a set of core object properties -->
        <objectProperties>

                <property/>

                <property/>

                ...

        </objectProperties>
        <!-- there can be zero or more datastreams -->
        <datastream>

                <datastreamVersion/>

                <datastreamVersion/>

                 ...

        </datastream>
        <!-- there can be zero or more disseminators -->

        <disseminator>

                <disseminatorVersion/>

                <disseminatorVersion/>

                 ...

        </disseminator>

</digitalObject>
```

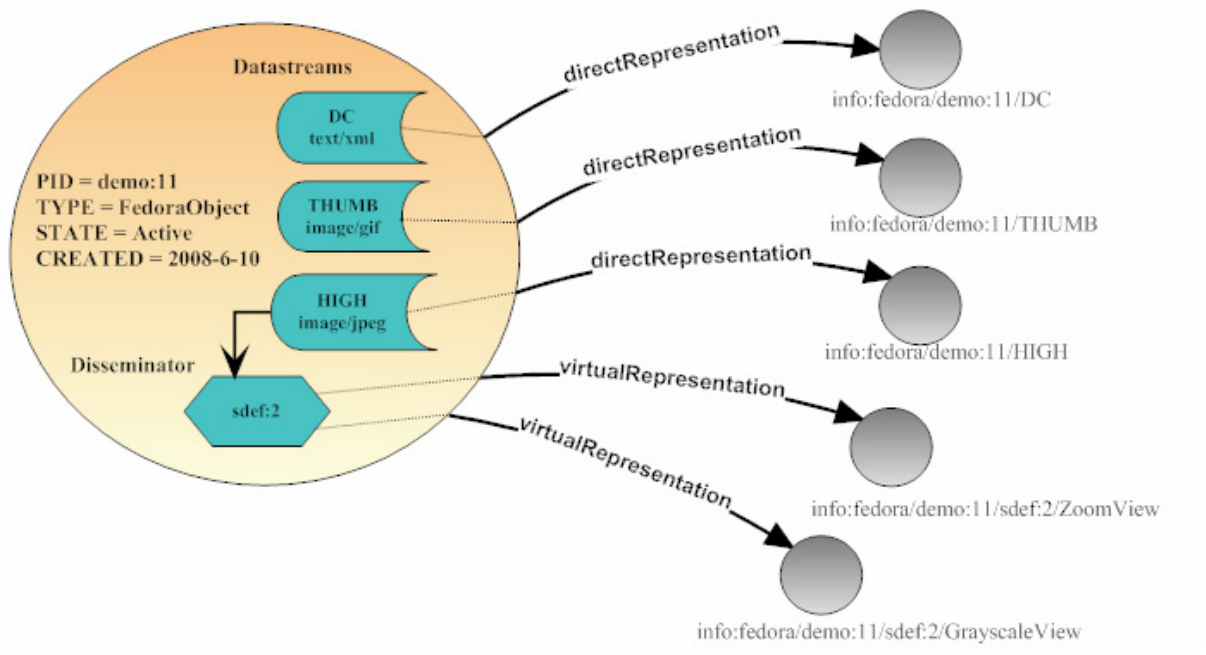
V lastnosti <objectProperties> Fedora zapiše informacije o publikaciji kot so identifikator publikacije, ki smo jo vložili, lastnika, datum vnosa, datum zadnje spremembe, stanje publikacije (npr. aktivna), itd.

V lastnosti <datastreams> se zapisujejo "objekti" publikacije. Tu so shranjeni tako tekstovni zapisi o publikaciji – npr. vrednosti polj Dublin Core, kot tudi binarni objekti (slikovno gradivo, PDF, video zapisi, itd).

FOXML shema je dosegljiva na (http://www.fedora.info/definitions/1/0/foxml1-0.xsd)

Zaradi preprostega načina zapisovanja informacij lahko shranjuje tako tekstovne kot slikovne dokumente, izvajalno kodo, multimedijske vsebine (avdio, video, slikovno gradivo). FOXML je torej XML zapis shrambe, v katerem nastopa opis publikacije in njeni sestavni deli – objekti. Razločevanje med dvo- in tridimenzionalnimi objekti ni potrebno niti pomembno, prav tako ne med videozapisom ali rokopisom, časopisnim člankom ali zvočnim posnetkom. Z opisom publikacije v Dublin Core (DC) formatu poskrbimo za opis publikacije in za njeno dostopnost, najdljivost, enoličnost. V Fedori je moč shranjevati objekte ali zgolj povezave – relacije na objekte, ki so lahko del druge publikacije znotraj repozitorija, v drugem repozitoriju ali na popolnoma drugi lokaciji. Format objektov (sestavnih delov) postane irelevanten, pomemben je le identifikator, ki nas pripelje do njega ter zmožnost uporabnika, da ta tip datoteke lahko odpre.

Slika 7.2 prikazuje publikacijo in njene sestavne dele na način, kot so shranjeni v repozitoriju vsebin. Poleg FOXML zapisa lahko Fedora sprejema zapise tudi v obliki METS (Metadata Encoding and Transmission Standard) in MPEG21/DIDL (Moving Pictures Expert Group Multimedia Framework Digital Item Declaration Language)



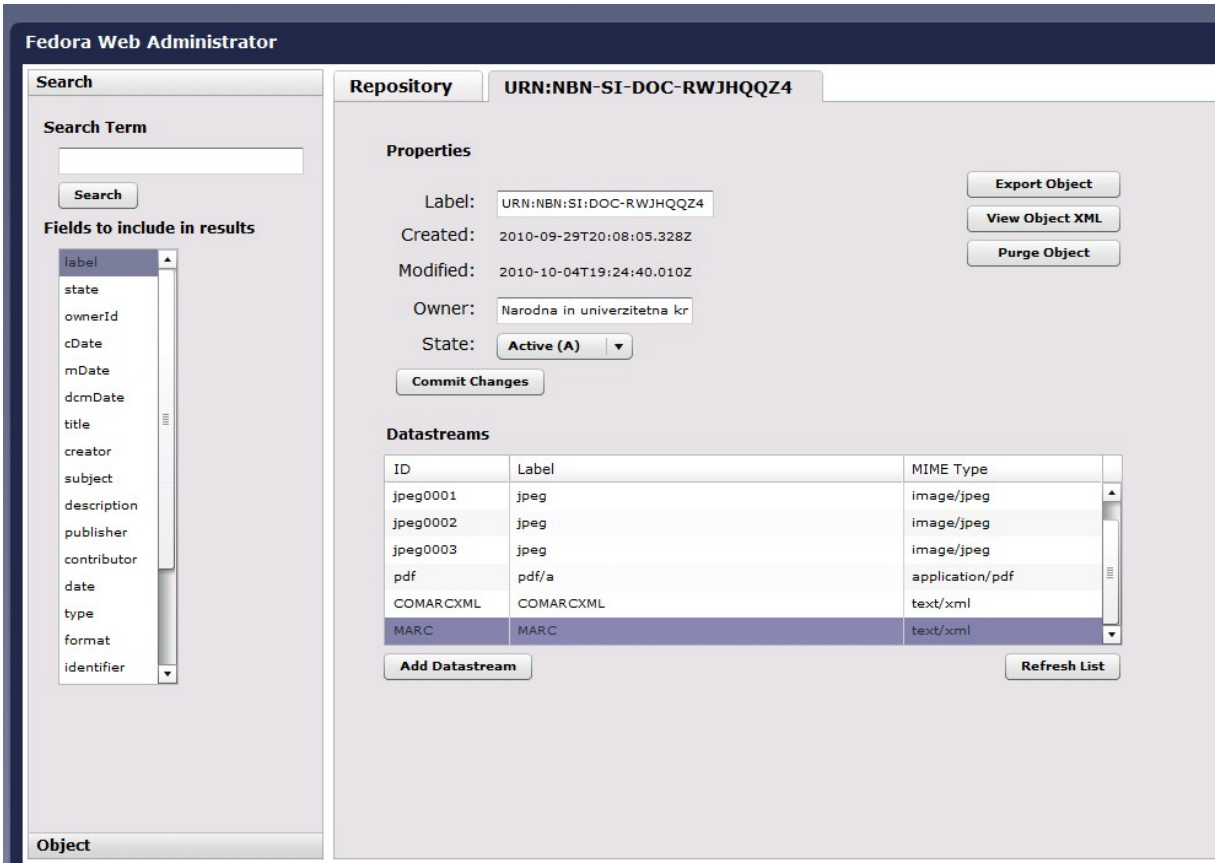
Slika 7.2) Pogled na perspektivo dostopa do objektov

7.1 VLAGANJE VSEBIN

Proces vlaganja vsebin (v našem primeru publikacij) je preprost. Na voljo imamo več možnosti, ki se konceptualno razlikujejo in sicer:

7.1.1 Vlaganje posamezne publikacije

Orodje, ki je uporabniku na voljo ob namestitvi produkta, ponuja preprost uporabniški vmesnik za vlaganje vsebin. Zaslonska slika vlaganja objektov Cankarjevih Hlapcev preko spletne aplikacije izgleda tako, kot prikazuje slika 7.3. V aplikacijo smo vnesli nekaj objektov, ki sestavljajo publikacijo (PDF datoteko, jpeg skenograme v visoki ločljivosti, meta podatkovne opise v Dublin Core, MARC in COMARC formatu). Vnos vsebin na ta način je preprost, a zamuden.



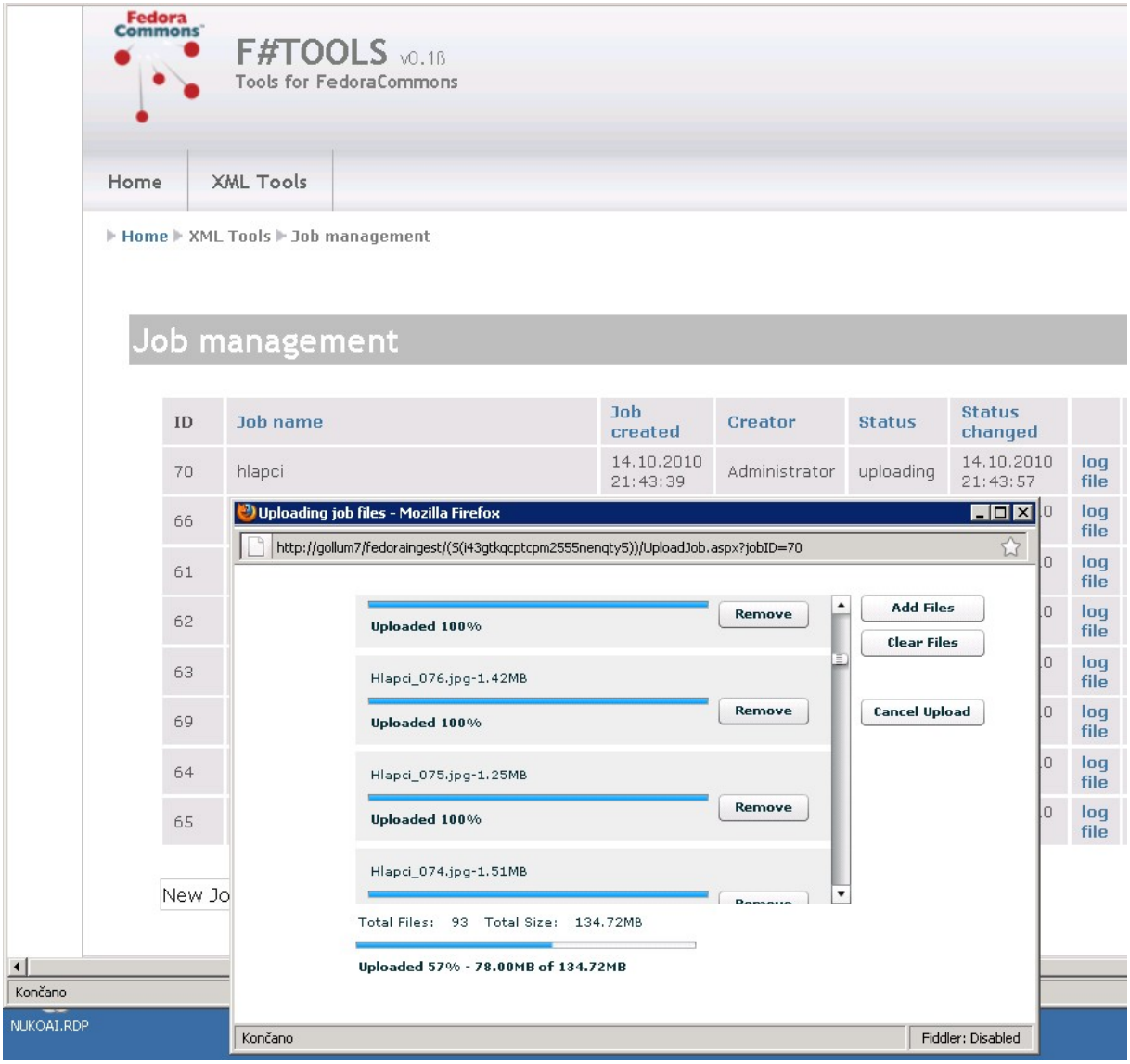
Slika 7.3) Spletni vmesnik za dodajanje vsebin

Ker je Fedora odprtokodni in prosto dostopen repozitorij, kateremu podpora in število uporabnikov strmo narašča, je preko skupnosti uporabnikov **mogoče dobiti izdelane spletne vmesnike** za lažje delo z repozitorijem. Prosto dostopni spletni vmesniki ponujajo bodisi zgolj podporo za shranjevanje objektov v celoto, ki jo oddamo v repozitorij (da se izognemo zamudnemu vlaganju posameznih objektov), do celotnih sistemov za upravljanje z vsebinami, podprtih s Fedora repozitorijem.

7.1.2 Paketno vlaganje vsebin

Ob ugotovitvi, da ima povprečna publikacija (tiskano gradivo) vsaj nekaj deset strani, bi zaradi potrebne časovne komponente polnjenje arhiva s prikazanim postopkom sistem ob resni uporabi postal praktično neuporaben, saj je aktivnost vlaganja objektov publikaciji zamuden posel. Na srečo sistem Fedora podpira vnašanje vsebin z uporabo API-jev – programskih vmesnikov, ki omogočajo implementacijo modulov za paketni vnos vsebin.

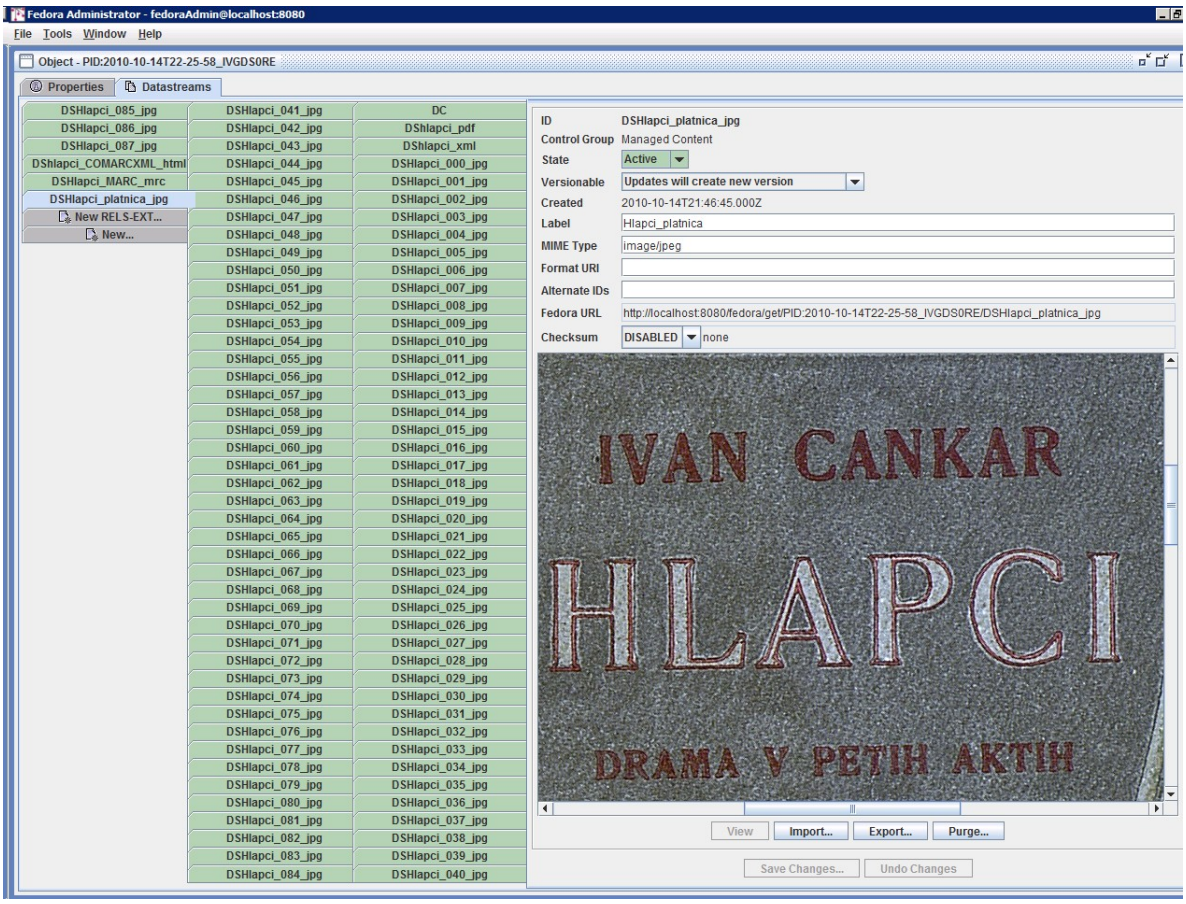
Uporabnik - vlagatelj v poljubnem programskem jeziku sestavi XML datoteko v formatu FOXML, kjer poskrbi za pravilno strukturo datoteke (glave, kombinacija tekstovnih – opisljivih elementov in binarne kode z "mime-type" opisi formatov objektov, ki jih vstavlja. Datoteka vsebuje publikacijo, z vsemi njenimi sestavnimi deli - objekti. Tako pripravljeno datoteko ali množico datotek (več pripravljenih publikacij) v enem koraku oz. klicu funkcije za vlaganje vloži preko klica prožilca, časovnega razporejevalnika, direktno, itd. v repozitorij vsebin – slika 7.4.



Slika 7.4) Implementacija vmesnika za paketno dodajanje objektov publikacij

7.2 Dostop do vsebin

Na sliki 7.5 je predstavljen pogled na dostop do publikacije, ki smo jo ravnokar vložili (Cankarjevi Hlapci) in njenih sestavnih delov. Vloženi publikaciji je moč dodajati nove objekte, brisati obstoječe, dodajati nove verzije objektov – Fedora sama skrbi za delo z verzijami.



Slika 7.5) Pogled na publikacijo v repozitoriju Fedora

S stališča strežnika oz. platforme Fedora delimo uporabnike na več nivojev in sicer na upravljavce sistema, ponudnike vsebin (content providers) in uporabnike vsebin. Fedora omogoča komuniciranje in s tem omogoča dostop do vsebin na več načinov:

- preko aplikacije (vgrajene javanske aplikacije ali lastnim potrebam razvite),
- preko spletnega vmesnika,
- preko implementacij prikrojenih spletnih vmesnikov,
- preko implementacij vmesnikov/aplikacij – klicev API funkcij.

Aplikacije dosegajo podatke shranjene v Fedori z uporabo API klicev in sicer

- upravljanje (management) – preko http ali SOAP,
- dostop (access) – preko http ali SOAP,
- iskanje (search) – preko http ali SOAP,
- dostop do OAI API vmesnika, dostopnega preko http.

Dokumenti, shranjeni v arhivu Fedora, so po procesu vnosa in nastavitvi parametrov na voljo javnosti. Do publikacije lahko dostopamo z uporabo http protokola kot npr:

http://nukfedora:8080/fedora/objects/PID%3A2010-10-14T22-25-58_IVGDS0RE/

ali zgolj do nekega objekta (datastreama) publikacije kot npr.

http://nukfedora:8080/fedora/objects/PID%3A2010-10-14T22-25-58_IVGDS0RE/datastreams/DSHlapci_000_jpg

Zaradi uporabe trajnega enoličnega identifikatorja je omogočen dostop in uporaba arhiviranih e-vsebin različnim vrstam uporabnikov (fizičnim osebam, organizacijam, servisom, spletnim portalom, itd). Platforma Fedora poleg dostopa te vrste omogoča tudi zajem podatkov z uporabo protokola OAI-PMH, ki smo ga omenili v prejšnjem poglavju. Organizacije kot so Evropska knjižnica in spletni servis Europeana z zajemanjem vsebin po tem postopku gradita in vzdržujeta centralni indeks partnerskih organizacij (nacionalnih knjižnic, muzejev, arhivov idr.) in s tem omogočata dostop do vsebin z enega spletnega mesta.

8 TRAJNO OHRANJANJE E-VSEBIN

Z dokončanjem zadnje faze – shranjevanjem publikacije v shrambo digitalnih vsebin še zdaleč nismo na koncu procesa ohranjanja digitalnih virov. Če želimo slediti cilju, ki smo si ga zadali (ohranitev kulturne, nacionalne in znanstvene dediščine zanamcem), moramo zagotoviti trajno varovanje, dostop, identiteto in sledljivost dokumentov. Elektronsko rojene dokumente (predvsem znanstvene publikacije) želimo ohranjati z namenom ponovne uporabe zaradi same vsebinske vrednosti dokumentov, starejše (npr. rokopisno gradivo) pa tudi z namenom varovanja originalov pred izpostavljanjem okolju (kemični in fizični vplivi, ki škodijo gradivu).

Količina podatkov z masovno digitalizacijo strmo narašča. V današnjem času je zaradi eksponentne rasti količine digitaliziranega gradiva fokus usmerjen predvsem v shranjevanje informacij. Razen pri publikacijah, ki predstavljajo biser kulturne dediščine (npr. rokopisno gradivo in prvi tiski) tiff formata za zapis barvne ali sivinske palete praktično ne uporabljamo, saj kljub nizki ceni za bit shranjene informacije ne bi bilo mogoče zagotoviti diskovnega prostora za trajno shranjevanje tolikšne količine informacij. Kljub ugotovitvi, da identične kopije iz visoko ločljivih jpeg datotek ni moč izdelati, je približek, ki ga na ta način lahko zagotavljamo dovolj dober, da se nacionalne in univerzitetne knjižnice (po Evropi) zatekajo k tej rešitvi. Na ta način je moč na enakem diskovnem prostoru ohranjati precej večjo količino publikacij.

Čeprav je publikacija shranjena v repozitoriju – digitalni shrambi, je potrebno poskrbeti za vzdrževanje arhiva. Z uporabo orodij, ki smo jih opisali in so večinoma prosto dostopna in brezplačna, se levji delež potrebnih sredstev seli iz programske na strojno opremo. Poskrbeti je potrebno za nikoli dovolj velika **diskovna polja**, RAID tehnologijo in **redundanco**, diskovno **varnostno kopiranje**, **potresno in požarno varnost** strežniških prostorov in **alternativno lokacijo** (ali več njih), ki mora biti prav tako **izolirana pred zunanjimi vplivi**.

Narodna in univerzitetna knjižnica, Ljubljana je v letu 2009 pridobila nov računalniški center in komunikacijsko vozlišče, ki je prvi od pogojev za trajno in varno ohranjevanje digitalnega gradiva. Bibliografske podatke lahko, kot smo že večkrat omenili, brez težav in razmeroma poceni pretakamo po internetu k različnim ponudnikom iskalnikov in spletnih portalov, vsebine, ki predstavlja 99% količine podatkov pa žal ne. Prvi razlog je hitrost, drugi pa enormne količine informacij. Digitalni arhiv, njegovo vzdrževanje in varovanje je v domeni

institucije, ki razpolaga z (digitalnim) gradivom, oz. gre za nacionalno vprašanje, ki ga je potrebno rešiti na nivoju države, saj je odnos do kulturne, nacionalne in znanstvene dediščine tudi ogledalo samega naroda.

9 ZAKLJUČEK

V diplomskem delu smo ponazorili možen pristop procesa trajnega ohranjanja e-vsebin. Proces zajema fazo **izbora, zajema, manipulacije, shranjevanja in vzdrževanja** arhiva digitalnih publikacij. Večinoma smo aktivnosti zgolj podali in jih grobo opisali, saj je bil cilj diplomskega dela ponazoritev problematike in narave dela z e-vsebinami in trajnim ohranjanjem le teh, ter prikaz dobre prakse, ki jo uporabljajo v Narodni in univerzitetni knjižnici in sorodnih inštitucijah po svetu. Zaradi vedno cenejših kapacitet za shranjevanje vsebin in hitrih podatkovnih poti (interneta) na eni, ter pridobljenega znanja in sodelovanja organizacij, vpletenih v ta proces na drugi strani prehajamo iz zgolj potreb in želja po trajnem ohranjanju e-vsebin na masovno digitalizacijo in procesiranje milijonov strani besedil, rokopisov, slikovnega gradiva, ne le v svetu, ampak tudi pri nas. Kljub vedno naprednejši tehnologiji pa ostaja cilj – digitalizacija vse kulturne in nacionalne dediščine trenutno nič več kot zgolj - utopija.

10 VIRI IN LITERATURA

[1] (2002) Consultative, Committee for Space Data Systems,
"Reference Model for anOpen Archival Information System (OAIS)"
http://public.ccsds.org/publications/archive/650x0b1.PDF

[2] (2006) "Zakon o obveznem izvodu publikacij (ZOIPub)", Ur. list RS št. 69/06 – ZOIPub
http://zakonodaja.gov.si/rpsi/r06/predpis_ZAKO3606.html

[3] (2007) "PRAVILNIK o vrstah in izboru elektronskih publikacij za obvezni izvod"
(Ur. list RS št. 90/07)
http://www.uradni-list.si/1/objava.jsp?urlid=200790&stevilka=4422

[4] (2010) "International Internet Preservation Consortium"
http://netpreserve.org/

[5] (2010) "Web Curator Tool Project"
http://webcurator.sourceforge.net/

[6] (2007) *Olaf Drümmer, Alexandra Oettler und Dietrich von Seggern*
"PDF/A in a Nutshell – Long Term Archiving with PDF"

[7] (2010) "Open XML Paper Specification"
http://en.wikipedia.org/wiki/Open_XML_Paper_Specification

[8] (2010) "Open Document Format for Office Applications"
http://en.wikipedia.org/wiki/OpenDocument

[9] (2010) "Office Open XML"
http://en.wikipedia.org/wiki/Office_Open_XML

[10] (2010) The Dublin Core Metadata Initiative
http://www.dublincore.org/

[11] (2010) Dspace
http://www.dspace.org

[12] (2010) EPrints
http://www.eprints.org

[13] (2010) Fedora Commons
http://fedora-commons.org/

11 KAZALO SLIK