# 5. *SUCCEED* PREPORUKE*

Ovaj odeljak donosi skup preporuka za poboljšanje intereoperabilnosti i zaštitu tekstualne/štampane građe. Njihov cilj je da pomognu ustanovama (istraživačkim grupama, kompanijama i organizacijama za očuvanje kulturnog nasleđa) u odabiru posebnog formata ili standarda za svoje aktivnosti u vezi sa digitalizacijom. Preporuke su podeljene u tri dela, a svaki od njih se odnosi na specifičan aspekt digitalizacije:

- *Dugoročno očuvanje* – ovaj deo se bavi formatima i standardima koji su u vezi sa glavnim datotekama, metapodacima i OCR[1] rezultatima;
- *Onlajn isporuka* – ovaj deo se bavi formatima i standardima koji su u vezi sa datotekama za isporuku, opisnim metapodacima, OCR rezultatima i identifikatorima;
- *Napredne tehnologije i tehnologije za podršku* – ovaj deo se bavi uputstvima za semantičke tehnologije, lingvističke izvore i pakovanje alata.

Ova podela je izvršena iz praktičnih razloga: ukoliko neka institucija sprovodi digitalizaciju radi zašitite građe, pažnja treba da bude usmerena na deo o dugoročnoj zaštiti. Ukoliko institucija sprovodi digitalizaciju radi pristupa, onda će je interesovati deo o onlajn isporukama. Ukoliko se digitalizacija sprovodi iz oba gore navedena razloga (radi očuvanja i radi pristupa), trebalo bi potražiti delove o dugoročnoj zaštiti i onlajn isporuci. Na kraju, ukoliko postoji vizija o korišćenju novih i naprednih tehnologija radi poboljšanja procesa digitalizacije, treba uzeti u obzir deo koji se odnosi na napredne tehnologije i tehnologije za podršku.

Svaki pojedinačni aspekt o kojem se govori u ovom odeljku može da ima nekoliko preporučenih i alternativnih naziva (kao što su *format* ili *standard*). Na primer, odeljak *Format glavne datoteke – tekstualni dokumenti* ima TEI[2] i PDF/A[3] kao preporučene formate, a kodirani standardni UTF-8 tekst kao alternativni. To znači da su TEI i PDF/A podjednako primenljivi; odabir formata zavisi od specifičnog izbora ili iskustva pojedine institucije. To takođe znači da UTF-8 ima izvesna ograničenja zbog kojih i jeste alternativan, a ne osnovni izbor. Ipak, ukoliko institucija nema odgovarajuće izvore za kreiranje PDF/A ili TEI dokumenata (na primer, odgovarajući softver ili broj zaposlenih) ili postoji neki drugi razlog zbog kojeg se preporučeni format ne koristi (na primer, politika institucije), predložen je alternativni format i on se smatra odgovarajućim. U pomenutom primeru, UTF-8 je alternativni format i u većini slučajeva neće

* Tekst je deo Preporuka za metapodatke i formate podataka za onlajn dostupnost i dugoročno očuvanje (*Recommendations for metadata and data formats for online availabilitz and long-term preservation*) do kojih je došlo u okviru projekta *Succeed*, pokrenutog sa ciljem da se olakšaju procesi digitalizacije u svim vrstama institucija u Evropi.

[1] OCR (*Optical Character Recognition*) – Optičko prepoznavanje znakova.

[2] TEI (*Text Encoding Initiative*) – Inicijativa za obeležavanje teksta.

[3] PDF/A je verzija PDF formata, namenjena dugoročnom očuvanju digitalnih dokumenata.

biti toliko složeno napraviti ga. Čak i ako institucija odluči da koristi alternativni format, trebalo bi da teži prelasku na onaj preporučeni, jer je to najpogodniji način postupanja u pojedinim aspektima digitalizacije.

## 5. 1 DUGOROČNO OČUVANJE

Ovaj deo preporuka bavi se formatima za glavne datoteke, zatim za opisne, strukturalne i administrativne metapodate, kao i za OCR rezultate. Razlog za odabir posebnog formata je u tesnoj vezi sa njegovim faktorima održivosti,[4] što se pre svega odnosi na otvorenost i rasprostranjenost.

### Format glavne datoteke – fotografije

Preporučeni format: TIFF[5].

Alternativni format: JPEG2000(JP2).

Preporučeni format za očuvanje fotografija je TIFF. To je najpopularniji format, bilo da je reč o postojećim preporukama (94% njih upućuje na TIFF), bilo da je reč o rezultatima *Succeed* ispitivanja (87% ispitanika navodi TIFF). Format je dobro dokumentovan i ima veoma dobru podršku u softverima za skeniranje, OCR, manipulaciju i konverziju. Preporučene karakteristike TIFF formata prikazene su u tabeli 19.

| KARSKTERISTIKE | PREPORUKA |
|---|---|
| Prostorna rezolucija | Najmanje 300dpi. Finalna rezolucija zavisi od tipa dokumenta. Cilj je da sve bitne karakteristike dokumenta budu jasno vidljive. Indeks kvaliteta[6] može biti od pomoći prilikom izračunavanja finalne rezolucije. |
| Dubina boje | 24 bita za fotografije u boji, 8 bitova za crno-bele. |

[4] Dostupno na: http://digitalpreservation.gov/formats/sustain/sustain.shtml.

[5] TIFF – Tagged Image File Format.

[6] Dostupno na: http://www.clir.org/pubs/abstract//reports/ub53.

| | |
|---|---|
| Kompresija | Bez kompresija ili LZW[7] kompresija. |
| Verzija | 6.0. |
| Redosled bitova | *Little endian.* |
| Profil boje | ICC profil[8]. |
| Broj strana | 1 strana po dokumentu (jednostrani TIFF). |

Tabela 19. *Kratak pregled preporučenih karakteristika TIFF formata*

Alternativni format za glavnu datoteku je JPEG2000 Deo 1 (Osnovni model) – JP2. Prilično je popularan u postojećim preporukama (53%), ali nije naročito zastupljen u trenutnim procesima digitalizacije (14% učesnika *Succeed* ispitivanja koristi ga kao glavnu datoteku). U pogledu korišćenja formata, izgleda da je JPEG2000 pre format u razvoju, nego što je opšteprihvaćen. Dobro je dokumentovan, ali i prilično komplikovan. Može da se ponaša i kao glavna datoteka i kao izlazni dokument, te je stoga posebno interesantno imati ga u vidu u izradi glavne datoteke. Nažalost, JPEG2000 nema široku softversku podršku, iako je u toku razvijanje alata koji ga podržavaju na različite načine (na primer, *OpenJPEG*[9], *Jpylyzer*[10], IIIF[11]). Zbog trenutnih ograničenja, ovaj format je i označen kao alternativni.

## Format glavne datoteke – tekstualni dokumenti

Preporučeni format: TEI, PDF/A.

Alternativni format: standardni kodirani UTF-8 tekst.

Za zaštitu dokumenata koji su dostupni u tekstualnoj formi, preporučujemo formate TEI ili PDF/A.

TEI format je orijentisan na prezentaciju teksta i obuhvata njegove različite karakteristike – strukturalne ili konceptualne. Veoma je prilagodljiv, što istovremeno može biti i prednost i mana. Srećom, postoje višestruki korisnički orijentisani TEI formati, kao što je *TEI Lite*, za elemente koji su dovoljni za jednostavne dokumente. *TEI Lite* je najčešće upotrebljavan korisnički

[7] LZW – Lempel-Ziv-Welch.

[8] ICC – International Color Consortium.

[9] Dostupno na: http://www.openjpeg.org/.

[10] Dostupno na: https://github.com/openplanets/jpylyer.

[11] Dostupno na: http://iiif.io/.

orijentisan format. TEI je popularan u digitalnim humanističkim naukama, što takođe ukazuje da je dobar izbor za očuvanje teksta. Više informacija o TEI možete naći u odeljku 0.

PDF/A je ISO standard koji služi za arhiviranje različitih vrsta dokumenata u digitalnom obliku. To je relativno nov format i zbog toga nije označen kao format glavne datoteke ni u postojećim preporukama ni u trenutnoj praksi, do koje se došlo u okviru *Succeed* ispitivanja. Ipak, on se zasniva na veoma popularnom, PDF formatu, koji 23% ispitanika koristi za glavnu datoteku. Stoga je razumljivo, barem za one koji ga već koriste, da sa običnog PDF formata pređu na PDF/A. Veoma je važno razlikovati PDF/A od PDF formata. PDF/A je arhivski format koji jeste zasnovan na PDF formatu, ali uvodi specifična ograničenja/zahteve koji obezbeđuju odgovarajuću vizuelnu prezentaciju dokumenta i ostalih karakteristika. Na primer, on zahteva da fontovi budu ugrađeni u dokument, ICC profile boja i ne dozvoljava kodiranje.

Postoje tri uzastopne verzije PDF/A formata, a svaka od njih ima nekoliko nivoa usaglašenosti. Sadržaj im je sledeći:[12]

- Nivo B – obezbeđuje odgovarajući izgled dokumenta; predstavljen je u verziji PDF/A-1;
- Nivo A – predstavlja nadogradnju nivoa B, ali dodatno zahteva strukturne informacije o dokumentu; predstavljen je u verziji PDF/A-1;
- Nivo U – omogućava da tekst dokumenta bude izdvojen i adekvatno protumačen; predstavljen je u verziji PDF/A-2.

Takođe, uzastopne verzije formata obezbeđuju nove mogućnosti. Najvažniji aspekti svake verzije su sledeći:

- PDF/A-1 – uvodi ograničenja koja se odnose na fontove, boje itd.;
- PDF/A-2 – uvodi mogućnosti različitih slojeva u dokumentu, dopušta JPEG2000 kompresiju i priloge u dokumentu;
- PDF/A-3 – nudi fleksibiniji sistem priloga.

Nijedna od verzija nije zastarela, stoga se svaka može koristiti za arhiviranje.

Svaki od ovih formata nudi različite tehnike koje mogu da se upotrebe, kao i različite nivoe usaglašenosti.

Alternativni format za prezentaciju teksta je standardna *Unicode* tekstualna datoteka (kodirana sa UTF-8). Ovaj format je alternativni zbog nedostatka podrške za strukturalne informacije, jer jednostavno prezentuje skup znakova. Preporučujemo upotrebu kodiranog UTF-8, koji je

[12] Dostupno na: http://www.pdflib.com/knowledge-base/pdfa/.

kompatibilan sa formatom ASCII[13] i može da kodira različite dijakritičke znake. Za skladištenje tekstualnih datoteka pogodno je koristiti standardne forme[14] UTF-8.

U slučaju dokumenata pisanih starom grafijom, posebno onih sa specijalnim znacima kojih trenutno nema u *Unicode* standardu, preporučujemo upotrebu MUFI[15] specifikacije (kodnih tačaka). Time bi se smanjio rizik od kolizija kodnih tačaka između tekstualnih izvora digitalizovanih u okviru različitih projekata ili pomoću različitih softverskim alata.

Takođe, moguće je ugrađivanje MUFI znaka u sâm *Unicode* (na primer, 152 MUFI znaka dodata su standardu Unicode 5.1). Detaljnije o MUFI specifikaciji možete videti u odeljku 0.

## Format za opisne metapodatke

Preporučeni formati: DCMES (*Dablinsko jezgro*), [16] MODS[17].

Alternativni format: MARC 21.

Najpopularniji format za opisne metapodatke je *Dablinsko jezgro* (puno ime je *Dublin Core Metadata Element*, skraćeno DCMES), svetski priznat ISO standard. Ovaj format, kao glavni za opisne metapodatke, u pogledu dugoročnog očuvanja, navodi 71% postojećih preporuka i 59% ispitanika. To je XML format, jednostavan i praktičan za korišćenje. Jednostavnost DCMES formata istovremeno je i prednost i mana. Dobro je to što, zahvaljujući njegovoj jednostavnosti, veliki broj institucija može da ga koristi. Nedostatak je to što značenje posebnih elemenata standarda nije precizno, pa može dovesti do različitih nesporazuma. Za detaljniji opis može se koristiti terminologija *Dablinskog jezgra* (DCTerms)[18], jer sadrži sve elemente DCMES-a, kao i dodatne elemente koji omogućavaju precizniji opis.

MODS format je prilično popularan, sa relativno visokim stepenom primenljivosti među korisnicima (16% ispitanika ga koristi za čuvanje, u 47% postejećih preporuka navodi se kao dobra opcija). Zasniva se na XML formatu, može da sadrži potpuniji opis nego *Dablinsko jezgro*; zasniva se i na MARC 21 formatu (iako ne može da podrži kompletne MARC 21 zapise), te se lako može napraviti od postojećih MARC 21 zapisa.

MARC 21 se, takođe, navodi u postojećim preporukama i ispitivanju. Međutim, nije naročito preporučljiv, zbog izvesnih problema sa interoperabilnošću. On ima specifičnu šemu kodiranja za prenos podataka (MARC 21 komunikacioni format), ali ona nije jednostavna, nije

[13] ASCII – American Standard Code for Information Interchange.

[14] Dostupno na: https://en.wikipedia.org/wiki/Unicode_equivalence.

[15] MUFI – Medieval Unicode Font Initiative.

[16] DC – Dublin Core.

[17] MODS – Metadata Object Description Schema.

[18] DCTerms – Dublin Core Metadata Initiative Terms.

samoopisujuća i definitivno nije čitljiva ljudskim okom. Dodatni problem predstavlja mogućnost različitog kodiranja MARC 21 zapisa. To može prouzrokovati dodatne probleme, kao na primer da početni elementi, navedeni u zaglavlju MARC 21 zapisa, zavise od znakova, a ne od bitova (neki znaci mogu da zauzmu više od jednog bita, u zavisnosti od kodiranja). Ovo znači da se kodiranje mora znati unapred (pre obrade) i da nije prisutno u samoj datoteci. Zbog svega toga, MARC 21 format je predložen kao alternativni.

## Format za strukturalne metapodatke

Preporučeni format: METS[19].

Jedina opcija za strukturalne metapodatke je METS format. Za njega u praksi ne postoji alternativa. Već ga koristi 36% ispitanika, dok se u postojećim preporukama pojavljuje u 59% slučajeva. Zasnovan je na XML formatu, jednostavan za primenu i podržava raznovrsne specifične formate, kao što su MODS, ALTO,[20] TextMD,[21] MIX[22] i PREMIS[23] (sve ih preporučuje *Succeed* projekat). Stoga je to najbolja opcija za dugoročno očuvanje strukturalnih metapodataka (u praksi, i jedina).

## Format za administrativne metapodatke

Preporučeni formati: PREMIS, MIX, TextMD.

Što se tiče administrativnih metapodataka, postojeće preporuke, kao i ispitanici, navode format PREMIS za očuvanje, a MIX ili NISO Z39-87 za dobijanje tehničkih metapodataka o fotografijama.

MIX format je zasnovan na XML-u i najpopularnija je implementacija NISO Z39-87[24] standarda. Lako se može integrisati u METS. Stoga se preporučuje za skladištenje tehničkih metapodataka o fotografijama. PREMIS je, zapravo, jedini format koji se u praksi koristi za skladištenje metapodataka za očuvanje. Uključen je u 41% postojećih preporuka, a 22% ispitanika navodi da ga koristi. PREMIS takođe lako može da se integriše u METS, jer je zasnovan na XML formatu. Aktivno se razvija (Redakcioni odbor trenutno radi na verziji 3.0) i ima sopstvenu PREMIS ontologiju za prezentovanje informacija putem semantičkih tehnologija. Institucije koje su učestvovale u ispitivanju nisu mnogo koristile PREMIS. Ni postojeće

[19] METS – Metadata Encoding and Transmission Standard.

[20] ALTO – Analyzed Layout and Text Object.

[21] TextMD – Technical Metadata for Text.

[22] MIX – Microsoft Image Exchange.

[23] PREMIS – Preservation Metadata: Implementation Strategies.

[24] Standard koji nije ISO standard.

preporuke nisu često upućivale na njega. U stvari, nije bilo nikakvih nagoveštaja tehničkih metapodataka za tekstualne dokumente. Zbog toga se čini da je opravdano koristiti format koji je već dobro integrisan sa preporukama za strukturalne metapodatke ili sa preporukama za očuvanje. Takav je TextMD – XML format koji se lako može koristiti kako u METS formatu tako i u PREMIS-u. Osim toga, podržavaju ga alati za karakterizaciju (na primer, JHOVE)[25].

## Format za dobijanje OCR rezultata

Preporučeni formati: ALTO, PAGE.

Alternativni format: kodirani standardni UTF-8 tekst.

ALTO format je naveden u 29% postojećih preporuka. Razvijen je radi proširenja METS formata, u cilju obezbeđivanja informacija o koordinatama (ALTO format), kao i strukturalnih informacija (METS). U odeljku 0 navedene su prednosti i mane ALTO formata. Osnovna prednost je interoperabilnost, čitljivost (zasnovana na XML-u) i jednostavnost. Glavni nedostatak je što podržava ograničen broj tipova regija i što nema podršku za opis logičkih struktura (za ovo je potreban kontejner format, kao što je METS). Neke od komercijalnih mašina za OCR podržavaju ispis u ALTO formatu, a ALTO format se upotrebljava u pojedinim tekućim aktivnostima (na primer, u okviru projekta *Europeana Newspapers*[26])

Jedan od glavnih dizajnerskih ciljeva PAGE formata bio je da se omogući detaljan i precizan opis bilo koje informacije koja se može dobiti iz date slike, prevazilazeći sva ograničenja postojećih formata (kao što je ALTO) i dozvoljavajući njegovu upotrebu u aplikacijama koje zahtevaju veoma precizno predstavljanje sadržaja (kao što je evaluacija uspešnosti). PAGE format nema veliki broj korisnika, ali privlači sve veću pažnju, jer je korišćen u inicijativama i projektima kakvi su *IMPACT Centre of Competence,*[27] *eMOP*[28] ili *Transcriptorium*[29].

Alternativni format je standardna kodirana UTF-8 tekstualna datotetaka. Alternativan je zbog nedostatka podrške za strukturalne informacije, pošto je datoteka samo skup znakova. Preporučujemo upotrebu kodiranog UTF-8 formata, jer je kompatibilan sa ASCII i može da kodira različite dijakritičke znake. U takvim tekstualnim datotekama, dobro je koristiti normalizovane forme UTF-8 za skladištenje rezultata OCR-a.

U slučaju istorijskih dokumenata, posebno onih sa specijalnim znacima kojih trenutno nema u *Unicode* standardu, preporučujemo upotrebu MUFI specifikacije (kodnih tačaka) pri osposobljavanju OCR mašine (što će rezultirati MUFI karakterima u izlaznom OCR rezultatu).

[25] Dostupno na: http://sourceforge.net/projects/jhove/.

[26] Projekat „Evropeana novine“.

[27] Dostupno na: http://www.digitisation.eu/.

[28] Dostupno na: http://www.europeana-newspapers.eu/.

[29] Dosupno na: http://transcriptorium.eu/.

Na ovaj način, smanjuje se rizik od kolizija kodnih tačaka između tekstualnih izvora digitalizovanih u okviru različitih projekata ili pomoću različitih softverskim alata. Takođe, moguće je ugrađivanje MUFI znaka u sâm *Unicode* (na primer, 152 MUFI znaka dodata su standardu Unicode 5.1). Detaljnije o MUFI specifikaciji možete videti u odeljku 0.

## 5.2 ONLAJN ISPORUKA

### Format datoteke za isporuku podataka

Preporučeni formati: JPEG, PDF, JPEG2000 (JP2), *Epub*, MOBI dobijen iz *ePUB*.

Datoteke za isporuku namenjene su krajnjem korisniku – lake su za korišćenje i jednostavno ih je prikazati. Takođe vredi razmotriti upotrebu nekoliko formata za isporuku za specifične digitalne objekte, pošto korisnici mogu da imaju drugačije izbore.

JPEG format navodi većina postojećih preporuka (82%), kao i učesnika u *Succeed* ispitivanju (71%). U osnovi, to je format slike koji koristi *lossy* kompresiju[30] za smanjivanje veličine slike. Praktično svi veb-pretraživači podržavaju JPEG format, uključujući i pretraživače za mobilne telefone.

PDF format je najpopularniji među učesnicima *Succeed* ispitivanja (77%), veoma je popularan i u postojećim preporukama (53%), ali zahteva prikaz specijalnih softverskih alata na računaru. Neki veb-pretraživači su, u poslednje vreme, ugradili podršku za PDF (na primer, *Firefox* ili *Chrome*), tako da, u pojedinim slučajevima, to više ne predstavlja prepreku. Osim toga, PDF podržava progresivno preuzimanje podataka, kao i višestruke slojeve, te se može koristiti za slike ili tekstualne sadržaje ili i za jedne i za druge.

JPEG2000 je opcija koju takođe treba uzeti u obzir pri onlajn isporuci, naročito kada se želi visoka rezolucija slike. JPEG2000 podržava popločavanje[31] i različite nivoe rezolucije; stoga je savršen format za ovakvu aplikaciju. Zahteva da se odgovarajući softverski alati prikažu u veb-pretraživaču korisnika, s tim što alati koji ovo podržavaju već postoje (na primer, IIIF,[32] *OpenSeadragon*[33]). Zahvaljujući tome, produkcione glavne datoteke moguće je koristiti kao direktan izvor za onlajn isporuku digitalnog sadržaja.

Za čitače elektronskih knjiga obavezan je tekstualni format. U ovom pogledu, najpopularniji su *ePub* i MOBI, zbog čega se i preporučuju; mogu direktno da se konvertuju iz formata ALTO, PAGE ili iz standardnog kodiranog UTF-8 formata. Što se MOBI formata tiče, važno je napomenuti da je to vlasnički format (zaštićen autorskim pravima). Razlog za preporuku je to što ga podržavaju *Kindle* uređaji, a oni su trenutno veoma popularni među čitačima elektronskih

[30] Tehnika kompresije sa „gubitkom kvaliteta“.

[31] Tiles.

[32] Dostupno na: http://iiif.io/.

[33] Dostupno na: http://openseadragon.github.io/.

knjiga. Najbolji način upotrebe MOBI formata za pružanje podrške širem krugu korisnika i njihovih uređaja jeste da se *ePub* sačuva kao osnovni format za isporuku i da se posle konvertuje u MOBI (dostupni su besplatni alati).

OCR rezultati mogu se spakovati u jednu datoteku prezentacionog formata, na primer, u PDF, ili u pojedinačnu datoteku koja je u formatu koji se koristi za zaštitu OCR rezultata.

## Format opisnih metapodataka za onlajn isporuku

Preporučeni: DCMES (*Dablinsko jezgro*), EDM.

*Skup elemenata podataka Dablinskog jezgra* (DCMES)[34] obavezan je za sve institucije koje žele da obezbede onlajn opisne metapodatke. To je osnovni set od 15 elemenata, koji pruža opšte informacije o izvoru.

*Dablinsko jezgro* je među učesnicima *Succeed* ispitivanja najpopularniji onlajn format za metapodatke (69%). To je jedini format koji mora da bude podržan pri implementaciji OAI-PMH komunikacije (OAI-PMH je široko prihvaćen protokol za prikupljanje metapodataka, koji koriste *Evropeana* i Digitalna javna biblioteka Amerike). Iako je jednostavan i veoma popularan, osnovna mana mu je nedostatak preciznog tumačenja svakog elementa. To može dovesti do nedoslednosti, na primer, na nivou agregatora za metapodatke.

*Evropeanin model podataka* (EDM)[35] uveden je kako bi se omogućilo dostavljanje bogatijih informacija *Evropeaninom* portalu informacija nego u slučaju *Dablinskog jezgra* ili *Evropeana* semantičkih elemenata. Spreman je da podrži sve značajne zahteve institucija za zaštitu kulturnog nasleđa. Ideja je bila da se poveća interoperabilnost metapodataka, da se iskoriste semantičke tehnologije i obezbedi finija granularnost i više semantike. EDM se zasniva na postojećim formatima i standardima, kao što su *Dablinsko jezgro*, SKOS[36] i OAI-ORE[37]. Već ga koristi 22% učesnika *Succeed* ispitivanja. Evropskim institucijama se preporučuje korišćenje EDM za predstavljanje metapodataka o prikazanom sadržaju, jer u potpunosti omogućava integraciju sa *Evropeanom*.

## Identifikacija objekata

Preporučeni: OAI[38] identifikator, DOI[39].

[34] DCMES – Dublin Core Metadata Element Set.

[35] EDM – Europeana Data Model.

[36] SKOS – Simple Knowledge Organization System.

[37] OAI-ORE – Open Archives Initiative Object Reuse and Exchange.

[38] OAI – Open Archive Initiative.

Kada je reč o identifikatorima, postoje dve glavne opcije – OAI identifikator i DOI.

OAI identifikator je besplatno rešenje; zasniva se na imenima domena i pruža mogućnost implementiranja trajnih identifikatora u repozitorijume koji podržavaju OAI-PMH protokol. Nije sagrađen na opštoj infrastrukturi, već na digitalnom repozitorijumu koji implementira OAI-PMH protokol i omogućava OAI identifikatore u OAI-PMH komunikaciji. Oslanja se na imena domena, što znači da jedan deo OAI identifikatora sadrži ime domena servisa koji obezbeđuje OAI-PMH funkcionalnost. To može da stvori konfuziju, na primer, u slučaju da se ime domena promeni.

DOI je usluga koja se plaća, a služi očuvanju trajnog identifikatora digitalnog sadržaja. Zasniva se na *Handle* sistemu[40] i koristi ga više institucija (15% ispitanika). *Handle* sistem je odabrao DOI (koji takođe koristi 15% ispitanika), jer sadrži dodatna svojstva, kao što su postojanost, doslednost i snažna tehnička infrastruktura. Prednost ovakvog pristupa je pouzdana i postojana infrastruktura (koju obezbeđuju *Handle* sistem i DOI), kao i nezavistnost specifične tehnologije (nasuprot OAI-PMH, koji se zasniva na imenima domena).

## 5.3 NAPREDNE TEHNOLOGIJE I TEHNOLOGIJE ZA PODRŠKU

Napredne tehnologije i tehnologije za podršku koje se primenjuju u digitalizaciji imaju za cilj da poboljšaju interoperabilnost, vreme procesuiranja i kvalitet čitavog procesa digitalizacije. Razmatrali smo tri aspekta: semantičke tehnologije, lingvističke izvore i pakovanje alata.

### Povezani otvoreni podaci (*Linked Open Data*)

Preporučeni: RDFa[41], SPARQL[42].

*Povezani otvoreni podaci* (*Linked Open Data* – LOD) uvode nov način razmišljanja o izvorima dostupnim na internetu. Osnovna ideja LOD jeste posedovanje izvora koji su povezani sa drugim izvorima, te je lako pronaći nove izvore i otkriti njihove veze. Izraz *otvoreni* sugeriše posedovanje raspoloživih podataka korišćenjem otvorenih licenci, kao što je *Creative Commons 0 Public Domain Dedication* (koju koristi *Evropeana*). Postoji više standarda koji se odnose na semantičke tehnologije i koji se mogu koristiti za izvore objavljivane na internetu. Takvi standardi, koje održava W3C[43], jesu RDF,[44] OWL,[45] SPARQL, RDFa, SKOS i RDFS[46].

---

[39] DOI – Digital Object Identifier.

[40] *Handle System* je tehnologija za dodelu, upravljanje i prepoznavanje stalnih identifikatora digitalnih objekata i drugih izvora na Internetu.

[41] RDFa – The Resource Description Framework in Attributes.

[42] Semantički jezik za baze podataka koji omogućava pretraživanje i rukovanje pohranjenim podacima.

[43] Dostupno na: http://www.w3.org/.

Predlažemo da se, povodom LOD, razmotre dva standarda: RDFa, za prikazivanje RDF tripleta[47] na veb-sajtu, i SPARQL, za pretraživanje dostupnih informacija u RDF bazi. Oba standarda održava W3C. Očigledno je da postoje i druge opcije koje se mogu uzeti u obzir. Ipak, izgleda da su ova dva standarda najprikladnija za opšte namene.

RDFa je standard koji pruža mogućnost ugradnje RDF tripleta u HTML, XHTML ili XML dokumente. On omogućava jednostavan način prikazivanja izvora i informacija u obliku otvorenih podataka. Karakteristike RDFa mogu se koristiti ograničeno (što omogućava veoma jednostavnu implementaciju – *RDFa Lite*) ili u potpunosti, kada zahteva više stručnosti (*RDFa Core*). Kao rezultat, semantičke informacije mogu se izdvojiti sa veb-sajta (na primer, digitalne biblioteke) pomoću automatizovanih alata i da ih, zatim, dalje procesuiraju. RDFa već koristi 21% ispitanika istraživanja, što je 62% onih koji koriste semantičke tehnologije.

Da bi se omogućio napredniji pristup izvorima, preporučuje se ugradnja SPARQL interfejsa za očuvanje podataka. SPARQL je jezik upita za RDF i opšti način za pristupanje informacijama koje se skladište u RDF (koristi ga 8% ispitanika, što je 23% onih koji koriste semantičke tehnologije). U cilju pružanja SPARQL interfejsa (krajnja tačka), neophodno je imati RDF skladište podataka, koji je vrsta baze podataka za RDF triplete. Ovakvo skladište podataka se može napraviti, na primer, od informacija dostupnih na internetu u RDFa standardu.

## Lingvistički izvori

Preporučeni: TEI, CMDI[48] ili LMF[49].

Za otkrivanje, pronalaženje i ponovnu upotrebu lingvističkih podataka, važno je da se podaci skladište u predvidljivom formatu. Ima mnogo elemenata koji mogu biti sačuvani u vezi sa lingvističkim izvorima, pri čemu težište stavljamo na korpuse i rečnike koji mogu da budu od pomoći pri unapređenju OCR tehnika tokom procesa digitalizacije.

TEI format je više semantički, nego prezentacioni: tačno su označeni semantika i tumačenje svakog taga i atributa, što je oko 500 različitih tekstualnih komponenti i koncepata (reč, rečenica, karakter, osoba i tako dalje) i svi se zasnivaju na jednoj ili više akademskih disciplina, za šta su dati primeri. *TEI Lite* je XML format datoteke za razmenu tekstova. To je podesiv izbor iz

[44] RDF – Resource Description Framework.

[45] OWL – Web Ontology Language.

[46] RDFS – RDF Semantics.

[47] Dostupno na: http://infosys3.elfak.ni.ac.rs/nastava/attach/SemantickiWebKursPredavanja/SparQL%20i%20Triplestore.pdf .

[48] CMDI – Component MetaData Infrastructure.

[49] LMF – Lexical Markup Framework.

opsežnog skupa elemenata, dostupnih u kompletnim TEI smernicama. TEI nudi alate kao što su ODD i ROMA, koji pomažu korisniku u izboru podskupa iz TEI repertoara. Za lingvističke izvore, dostupan je specijalno korisnički orijentisan *TEI Corpus*. TEI je takođe već prisutan u oblasti kulturnog nasleđa. Stoga je korisno da razmotrimo njegovu upotrebu, naročito među onima koji ga već koriste u druge svrhe.

*Infrastuktura za komponentne metapodatke* (CMDI) razvijena je u okviru CLARIN projekta. Ona obezbeđuje okvir za opis i ponovno korišćenje nacrta metapodataka. Sastavni delovi opisa (*komponente* koje sadrže definicije polja) mogu da se grupišu u već gotov format opisa (*profil*). Oba se čuvaju i dele ostalim korisnicima preko Registra komponenti, da bi se promovisalo ponovno korišćenje. Tada se svaki zapis metapodataka prikazuje kao XML datoteka, uključujući i link koji je u vezi sa profilom na kom je zasnovan. Metapodaci se čuvaju u skladištenim repozitorijumima. CLARIN obezbeđuje centralni portal za otkrivanje izvora (*CLARIN Visual Language Obseravatory*). Povrh toga, CLARIN pravi specijalni softver dostupan za uređivanje CMDI zapisa (*Arbil*). CLARIN ima za cilj da u Evropi obezbedi infrastrukturu za istraživanje koja podrazumeva i biblioteke i javne arhive, a koja neće biti na raspolaganju grupama izvan tog domena, kao što su komercijalna preduzeća ili pojedinci.

*Okvir za leksičko označavanje* (LMF) je ISO 24613:2008 standard. Njegovi ciljevi su obezbeđivanje opšteg modela za kreiranje i korišćenje leksičkih izvora, sprovođenje razmene podataka između tih izvora, kao i omogućavanje spajanja većeg broja individualnih elektronskih izvora u obimne svetske elektronske izvore. Vrste pojedinačnih nivoa LMF-a su monolingvalni, bilingvalni ili multilingvalni leksički izvori. Ista specifikacija se koristi za male i velike, jednostavne i složene leksikone, kao i za pisane i govorne leksičke prezentacije. Lingvističke konstante, kao što su *feminine* ili *transitive*, LMF nije definisao, ali su zabeležene u *Registru kategorija podataka* (DCR)[50], koji je napravljen i održava se kao globalni izvor ISO/TC37, u skladu sa ISO/IEC 11179-3:2003. Ove konstante se koriste za poboljšanje nivoa strukturalnih elemenata. LMF je relativno nov, ali je već stekao prilično popularnost. Prema nekim lingvistima, standard nije dovoljno precizan. ISO je to rešio uvođenjem uputnih struktura za nekoliko poddomena.

## Pakovanje alata

Preporučeni: paket alata za ciljane operativne sisteme (barem za *MS Windows* i *Linux*).

Pakovanje je jedan od elemenata koji olakšava rukovanje i razumevanje novih alata. Prednost posedovanja specifičnih paketa za određene operativne sisteme je u tome što proces instalacije može biti automatski. Na primer, u slučaju *Linux* sistema, pakovanje omogućava sredstva za instalaciju ili apdejt (ažuriranje) softverskih paketa, uključujući i prečice i alate komandne linije. Takođe, dokumentacija se može automatski dodati (na primer, stranicama priručnika). To nije moguće bez softverskog paketa (iako je moguće jednostavno pokrenuti softver iz binarne datoteke, ali bez duboke integracije sa operativnim sistemom). Stoga se snažno preporučuje

[50] DCR – Data Category Registry.

upotreba tehnika pakovanja alata, kako bi se softver isporučio krajnjim korisnicima. *MC Windows* je, prema analizama ispitivanja, najpopularniji operativni sistem (87% ispitanika). *Linux* je na drugom mestu (49%). *Unix* i *MacOS* imaju, u proseku, po 10% popularnosti. Ovo ukazuje da treba podržati izgradnju softverskih paketa barem kod *MS Windows* i *Linux*, jer bi većina potencijalnih korisnika tada mogla da koristi automatski proces instalacije.

# 6. REZIME

Ovaj izveštaj pruža niz preporuka za specifične aspekte digitalizacije, kao što su kreiranje glavnih datoteka, onlajn isporuka digitalnih objekata ili napredne tehnologije i tehnologije za podršku. On, takođe, daje pregled postojećih preporuka i savremene prakse u procesima digitalizacije, posebno tekstualne/štampane građe. Sadrži i analizu ispitivanja sprovedenog među stručnjacima za poslove digitalizacije iz različitih institucija širom sveta. Preporuke su izrađene na osnovu postojećih preporuka i analize ispitivanja. To je učinjeno utvrđivanjem formata i standarda najpogodnijih za upotrebu u aktivnostima vezanim za digitalizaciju, imajući u vidu sve njihove prednosti i mane.

Tri tabelarna pregleda preporuka daju sažet prikaz izabranih opcija:

- u tabeli 20, dat je kratak pregled preporuka za dugoročno očuvanje, koji obuhvata formate za glavnu datoteku, metapodatke i OCR rezultate;
- u tabeli 21, dat je kratak pregled preporuka za onlajn isporuku, koji obuhvata formate za isporuku datoteke, za deskriptivne metapodatke i identifikaciju objekata;
- u tabeli 22, dat je kratak pregled preporuka za napredne tehnologije i tehnologije za podršku, koji obuhvata LOD, lingvističke izvore i pakovanje alata.

| **Aplikacija** | **Preporučeni** | **Alternativni** |
|---|---|---|
| Format glavne datoteke za digitalne fotografije | TIFF | JPEG2000 (JP2) |
| Format glavne datoteke za tekstualne dokumente | TEI, PDF/A | Kodirani standardni UTF-8 tekst |
| Format za opisne metapodatke | DCMES, MODS | MARC 21 |
| Format za strukturalne metapodatke | METS | Ne postoji |
| Format za administrativne metapodatke | PREMIS, MIX, TextMD | Ne postoji |
| Format za OCR rezultate | ALTO, PAGE | Kodirani standardni UTF-8 tekst |

Tabela 20. *Preporuke za dugoročno očuvanje*

| Aplikacija | Preporučeni | Alternativni |
|---|---|---|
| Format datoteke za isporuku | JPEG, PDF, JPEG2000 (JP2), ePUB, MOBI dobijen iz ePUB | Ne postoji |
| Format za opisne metapodatke | DCMES, MODS | Ne postoji |
| Identifikacija objekata | OAI identifikator, DOI | Ne postoji |

Tabela 21. *Preporuke za onlajn isporuku*

| Aplikacija | Preporučeni | Alternativni |
|---|---|---|
| *Linked Open Data* | RDFa, SPARQL | Ne postoji |
| Lingvistički izvori | TEI, CMDI, LMF | Ne postoji |
| Pakovanje alata | Paketi bar za *MS Windows* i *Linux* | Ne postoji |

Tabela 22. *Preporuke za napredne i pomoćne tehnologije*

Postoji još nekoliko zanimljivih aspekata do kojih se došlo zahvaljujući *Succeed* ispitivanju. Prvo, primetno je da još uvek ne postoji dovoljno aktivnosti za osiguranje kvaliteta OCR rezultata, kako za prepoznavanje znakova, tako i za elemente raspoređivanja (*layout*-a) (oko 60–70% ispitanika ne vrši provere kvaliteta OCR rezultata; videti crteže 28 i 27). Izuzetno je važno razmotrati ovo pitanje, s obzirom na značaj tekstualnih izvora u procesu istraživanja (na primer, u humanistici), kao i u procesu otkrivanja izvora (na primer, mašine za pretraživanje). Drugo, sasvim je jasno da je upotreba naprednih tehnologija (na primer, Prepoznavanja imenovanih entiteta, Geolokaliteta[51]) veoma ograničena. Više od 90% ispitanika uopšte ne koristi ove tehnike (videti crtež 25). Ovo je, svakako, aspekt digitalizacije koji bi trebalo unaprediti. Takođe, većina ispitanika ne upotrebljava LOD, 66% njih ne koristi ovu vrstu tehnologije u svojim procesima digitalizacije (videti crtež 23). Zanimljivo je i da je veoma mali broj ispitanika raspoložen da pribavi glavne datoteke onlajn (29%). Na kraju, ispitivanje je pokazalo da veći broj ispitanika (55%) ne koristi sisteme za upravljanje u procesima digitalizacije (videti crtež 30). To znači da se, u većini institucija, procesi i projekti digitalizacije sprovode ručno. Profesionalni sistemi upravljanja u procesima digitalizacije mogu da obezbede automatizovano i samoorganizovano okruženje za digitalizaciju i tako poboljšaju efikasnost i kvalitet rezultata.

*Prevela s engleskog i napomene dala*

*Ljubica Ljubišić*

[51] Named Entities Recognition, Geolocation.