PAM Pokrajinski arhiv Maribor

Moderna arhivistika

Časopis arhivske teorije in prakse
Journal of Archival Theory and Practice

ISSN 2591-0884
https://doi.org/10.54356/MA

Letnik 6 (2023), št. 2 / Year 6 (2023), No. 2

Maribor, 2023

Prejeto / Received: 19. 05. 2023

1.02 Pregledni znanstveni članek 1.02 Review article

https://doi.org/10.54356/MA/2023/VRNY7665

EKSTRAKCIJA METAPODATKOV S POMOČJO STROJNEGA UČENJA

Ivančica SABADIN
študentka 3. stopnje programa Arhivske znanosti, Alma Mater Europaea, ECM, Maribor, Slovenija
ivancica.sabadin@almamater.si

Izvleček:

Namen prispevka je raziskovati tehnike ekstrakcije metapodatkov s pomočjo strojnega učenja. Uporabljena je bila metoda pregleda literature iz podatkovnih baz ProQuest, Scopus in Emerald Insight. Rezultati so pokazali, da so tehnike strojnega učenja že uveljavljene na področju ekstrakcije metapodatkov iz znanstvene literature. Najboljše rezultate so pokazale rešitve, ki združujejo analizo postavitve dokumenta in proces ekstrakcije metapodatkov. Glede na raziskave o ekstrakciji metapodatkov s pomočjo strojnega učenja lahko sklepamo, da je treba dodatno analizirati orodja in modele strojnega učenja GROBID, CERMINE, XTRACT, BERT, Mask R-CNN in BiLSTM. Na podlagi izkušnjah ekstrakcije metapodatkov iz znanstvene literature je treba prilagoditi modele za ekstrakcijo metapodatkov iz arhivskega gradiva.

Ključne besede:

ekstrakcija metapodatkov, strojno učenje, nadzorovano učenje, modeli strojnega učenja, obdelava naravnega jezika

Abstract:

Metadata extraction using machine learning

The aim of this paper is to explore metadata extraction techniques using machine learning. The method used was a literature review of the ProQuest, Scopus and Emerald Insight databases. The results showed that machine learning techniques are already well established in the field of metadata extraction from scientific literature. The best results were shown by solutions that combined document layout analysis and metadata extraction processes. Based on the research on metadata extraction using machine learning, it can be concluded that further analysis of the machine learning tools and models GROBID, CERMINE, XTRACT, BERT, Mask R-CNN and BiLSTM is needed. Based on the experience with metadata extraction from scientific literature, the models should be adapted for metadata extraction from archival material.

Key words:

metadata extraction, machine learning, supervised learning, machine learning models, natural language processing

1. UVOD

Metapodatki imajo v procesu hrambe arhivskega gradiva pomembno mesto. »*Metapodatki so podatki o drugih podatkih, ki opisujejo vsebino, strukturo in okoliščine nastanka dokumentarnega in arhivskega gradiva, njegovo upravljanje in uporabo*« (UVDAG, 2017, 2. čl.). Pomembnost metapodatkov je večkratna. Arhivom metapodatki omogočajo opisovanje ter zagotavljanje avtentičnosti, celovitosti in uporabnosti arhivskega gradiva. Za uporabnike so metapodatki enako pomembni, ker omogočajo dostopnost do gradiva in dodajanje kontekstualnih podrobnosti o gradivu. Metapodatki so tudi bistveni za nove tehnologije, ki se vse več uporabljajo v arhivih, npr. umetna inteligenca (angl. Artificial Intelligence) in povezani podatki (angl. Linked Data), saj pozitivno vplivajo na strojno berljivost gradiva.

Metapodatke iz arhivskega gradiva pridobivamo s procesom ekstrakcije, ki je v osnovi enostaven za človeka, vendar je glede na število elektronskih dokumentov, ki vsakodnevno nastajajo, preobremenjujoč. Zato je proces ekstrakcije metapodatkov treba avtomatizirati z uporabo orodij in metod, ki morajo pokazati določeno inteligenco. Na ta način lahko ekstrahiramo velike količine metapodatkov in tudi pridobimo točne rezultate oziroma zmanjšamo možnost napake. Sistem mora biti takšen, da lahko sam ugotovi, ali so zaznani sumljivi ali napačni podatki. »*V zadnjih letih je ekstrakcija metapodatkov postala pomembna naloga na vseh področjih. Ta pomembnost je posledica večje uporabe digitalnih vsebin, saj se je vse več ljudi navadilo pridobivati informacije po internetu*« (Bouabdallah et al. 2021, str. 1).

Pri strojni ekstrakciji metapodatkov se moramo zavedati postavitve (layout) dokumentov oziroma arhivskega gradiva. Pomembna lastnost arhivskega gradiva je heterogenost, kar pomeni, da imajo dokumenti med seboj različno postavitev, ki je odvisna od zvrsti in vsebine ter ustvarjalca gradiva. Obstaja veliko raziskav na področju ekstrakcije metapodatkov iz znanstvenih člankov, ki so lahko pomembne tudi za arhive. Še posebej zato, ker so znanstveni članki, podobno kot arhivsko gradivo, heterogeni glede na njihovo postavitev.

»*Metode strojnega učenja omogočajo robustno in prilagodljivo avtomatično ekstrakcijo metapodatkov ter se lahko uporabljajo za vse vrste dokumentov*« (Han et al. 2003, str. 1). Strojno učenje uporablja različne tehnike in postopke, za ekstrakcijo metapodatkov pa so zanimivi sistemi nadzorovanega učenja, ki uporabljajo tehnike SVM (angl. Support Vector Machines), CRF (angl. Conditional Random Fields) in HMM (angl. Hidden Markov Model), kakor tudi sistemi, ki temeljijo na transformatorjih, kot je BERT (angl. Bidirectional Encoder Representations from Transformers).

Za potrebe raziskave, ki je predmet prispevka, je bila opravljena metoda pregleda literature, dostopne v podatkovnih bazah ProQuest, Emerald Insight in Scopus. Pri iskanju sta bili uporabljeni ključni besedi »*metadata extraction*« ter »*machine learning*«, in sicer s pomočjo Boolovega logičnega operatorja AND. V prvem nizu je bilo identificiranih 470 zadetkov. Po pregledu in upoštevanju naslednjih meril je v raziskavo zajetih 20 člankov: objava v angleškem ali slovenskem jeziku; dostopnost besedila v celoti; tematska ustreznost glede na področje proučevanja; strokovni in znanstveni članki, konferenčni zborniki ter doktorske disertacije. Z dodatnim iskanjem po seznamih literature je bilo najdenih devet virov, ki so vključeni v analizo. Dodatno je vključena dokumentacija za posamezna programska orodja in modele strojnega učenja, priročnik o strojnem učenju v slovenskem jeziku, Uredba o varstvu dokumentarnega in arhivskega gradiva, Pravilnik o enotnih in tehnoloških zahtevah in spletišče Mednarodnega arhivskega sveta. Lastnosti posameznih programskih orodij in modelov strojnega učenja so bili pregledani z metodo analize. Z metodo osredotočanja so bile opredeljene lastnosti

orodij, pomembne za ekstrakcijo metapodatkov s pomočjo strojnega učenja. Programska orodja in modeli strojnega učenja so opisani z deskriptivno metodo.

2. REŠITVE ZA EKSTRAKCIJE METAPODATKOV

»*Ekstrakcija je proces, ki vključuje avtomatično zaznavanje strani, ki vsebujejo metapodatke, pridobivanje metapodatkov in njihovo označevanje kot ustrezne kategorije*« (Tang, 2006, str. 12). Prve rešitve ekstrakcije metapodatkov so omogočale ekstrakcijo le tehniških metapodatkov. Eden od najstarejših sistemov za ekstrakcijo metapodatkov je Infoharness iz leta 1999 (Skluzacek, 2022).

Na spletni strani Mednarodnega arhivskega sveta (ICA)[1] je arhivom na voljo seznam orodij za avtomatično ekstrakcijo metapodatkov, in sicer so navedeni: Exiftool[2], NLZN Metadata Extractor Tool[3], PhotoMe[4] in AsTiffTagViewer[5]. Navedena orodja so praviloma namenjena ekstrakciji tehničnih metapodatkov iz slikovnih datotek, kot so datum in čas nastanka, oblika, velikost, ločljivost in podobno. S Pravilnikom o enotnih tehnoloških zahtevah so za vsako zvrst gradiva določeni obvezni metapodatki. Za besedilne in mešane podatke so na primer obvezni metapodatki: enolična identifikacijska oznaka, naslov ali kratek opis vsebine, datum (prejetja, nastanka), rok hrambe in navedba subjekta (avtor, pošiljatelj ali prejemnik) (PETZ, 2020). Za zagotavljanje ustreznih metapodatkov, ki ne sodijo samo v tehnične, ampak tudi v vsebinske, strukturne in druge kategorije metapodatkov, ki so nujne za zagotavljanje avtentičnosti, celovitosti in uporabnosti gradiva, potrebujemo bolj kompleksna orodja in modele s področja umetne inteligence oziroma strojnega učenja.

3. METODE STROJNEGA UČENJA ZA EKSTRAKCIJO METAPODATKOV

Strojno učenje je podpodročje umetne inteligence. »*Cilj strojnega učenja je oblikovanje in razvoj algoritmov, ki sistemom omogočajo uporabo empiričnih podatkov, izkušenj in treniranja, da se razvijajo in prilagajajo spremembam v svojem okolju*« (Hu in Hao, 2013, str. 3). »*Izraz strojno učenje se nanaša na avtomatizirano zaznavanje smiselnih vzorcev v podatkih*« (Shalev-Shwartz in Ben-David, 2014, str. VII). Strojno učenje je lahko nadzorovano ali nenadzorovano, odvisno od označevanja vzorcev. »*Metode strojnega učenja za ekstrakcijo metapodatkov običajno spadajo v nadzorovano učenje*« (Tang, 2006, str. 14). Nadzorovano učenje za analizo potrebuje nabor objektov z oznakami razredov.

Na področju ekstrakcije metapodatkov se tudi uporabljajo modeli obdelave naravnega jezika (angl. Natural Language Processing, v nadaljevanju NLP). »*Obdelava naravnega jezika ali NLP je na splošno opredeljena kot avtomatična manipulacija naravnega jezika, kot sta govor in besedilo, s programsko opremo*.« (Brownlee, 2017, str. 2). Glede na to, da število digitalnih dokumentov narašča prehitro za zmožnosti človeške obdelave, se za področje NLP vse več uporabljajo modeli strojnega učenja, kot je BERT, ki je opisan v poglavju 3.4.1.

---

1 *ICA – International Council on Archives.*

2 *Več: http://owl.phy.queensu.ca/~phil/exiftool/.*

3 *Več: http://meta-extractor.sourceforge.net/.*

4 *Več: https://www.photome.de/.*

5 *Več: http://www.awaresystems.be/imaging/tiff/astifftagviewer.html.*

3.1 Nadzorovano učenje in ekstrakcija metapodatkov

Na področju ekstrakcije metapodatkov se uporabljajo naslednje metode nadzorovanega učenja: metoda podpornih vektorjev (angl. Support Vector Machines – SVM), skriti model Markova (angl. Hidden Markov Model – HMM) in pogojna naključna polja (angl. Conditional Random Fields – CRF). Flynn (2014) na podlagi rezultatov raziskav ugotavlja, da se pri tehnikah SVM in HMM zmanjšuje učinkovitost naraščanja heterogenosti zbirke.

Metoda podpornih vektorjev (SVM) s pomočjo algoritmov določa ravnino, ki najbolje razdeli nabor podatkov v dva razreda. Na področju ekstrakcije podatkov se lahko uporabi postopek klasifikacije informacij na način, da se za vsak zapis preveri, ali pripada posameznemu metapodatkovnemu elementu ali razredu. SVM lahko uporabimo za ekstrakcijo metapodatkov, ker se le-ta lahko pretvori v klasifikacijski problem. Tang (2006) pravi, da je za ekstrakcijo metapodatkov treba uporabiti večrazredni SVM, v katerem je vsak metapodatkovni element en razred.

Tehnike SVM in ekstrakcije značilnosti so uporabljene v raziskavi ekstrakcije metapodatkov iz znanstvenega članka »*Automatic Document Metadata Extraction using Support Vector Machines*«. Za potrebe raziskave je uporabljeno orodje SVM_Light. Pridobljeni rezultati so boljši kot v raziskavah, ki so za ekstrakcijo uporabljale metodo HMM. Proces ekstrakcije metapodatkov je bil razčlenjen na dva podproblema: 1.) klasifikacija vrstic in 2.) identifikacija delov večrazrednih (angl. multi-class) in večavtorskih (angl. multi-author) vrstic (Han et al. 2003).

Skriti model Markova (HMM) je tehnika verjetnosti z dvema komponentama: skriti proces obsega naključne spremenljivke, pri katerih so spremembe stanja skrite, proces observacije pa obsega naključne spremenljivke, pri katerih lahko spremljamo spremembe stanja. V primeru ekstrakcije metapodatkov so metapodatki lahko skrite spremenljivke, posamezni metapodatkovni elementi (naslov, čas nastanka …) pa observacijske spremenljivke. Skriti model Markova se uporablja na področju prepoznavanja govora in genov.

Pogojna naključna polja (CRF) so verjetnostni grafični model, ki se uporablja za klasifikacijo in je bil uporabljen v raziskavi ekstrakcije in delitve referenčnih informacij ter znanstvenih publikacij. Namen raziskave je bilo ustvarjanje semantičnih metapodatkov za knjižnice in repozitorije, ki se ukvarjajo z znanstvenimi publikacijami (Groza, Grimmes in Handschuh, 2012). »*Rešitev, uresničena z uporabo CRF, je dosegla dobre rezultate pri poskusnem ocenjevanju, pokazala se je z uporabo enkratno natreniranega drobilca [chunker, op. p.][6] na več testnih naborih podatkov*« (Groza, Grimmes in Handschuh, 2012, str. 17). »*Zelo priljubljen model v skupini diskriminativnih modelov so pogojna naključna polja (CRF). CRF združuje prednosti klasifikacije in grafičnega modeliranja ter povezuje zmožnost modeliranja večdimenzionalnih, zelo odvisnih podatkov z zmožnostjo uporabe velikega števila vhodnih lastnosti za napovedovanje. CRF lahko obravnavamo kot diskriminativno različico HMM*« (Tkaczyk, 2015, str. 15). CRF je med drugim osnova orodij za ekstrakcijo metapodatkov GROBID in ParsCit. ParCit je odprtokodno orodje, ki na podlagi tehnike strojnega učenja CRF razčlenjuje reference in pridobiva citate iz znanstvenih člankov (Councill, Giles in Kan, 2008). Glede na dosedanje raziskave je metoda strojnega učenja CRF pokazala dobre rezultate pri ekstrakciji metapodatkov iz gradiva s heterogeno vsebino.

[6] *Izraz vzet po Piškur, K. (2015). Popravljanje vejic v slovenskih besedilih z orodjem LanguageTool (https://dokumen.tips/documents/popravljanje-vejic-v-slovenskih-besedilih-z-orodjem-kur-popravljanjevejic.html?page=1)*

3.2 Metrike strojnega učenja

Obstoječe raziskave za merjenje rezultatov strojnega učenja na področju ekstrakcije metapodatkov uporabljajo metrike *preklic*, *preciznost* in *F1-mera*. Kot pišeta Karakatič in Fister (2022, str. 97), uporabljajo metrike štiri vrednosti, »*ki kažejo količino instanc glede na rezultate klasifikacije*«: TP (angl. True Positives) in FP (False Positivies) označujeta število pravilno ali napačno pozitivnih klasificiranih instanc, TN (angl. True Negatives) in FN (angl. False Negatives) pa sta število pravilno in natančno negativnih klasificiranih instanc. »*Metrika priklic (angl. recall) nam pove delež pozitivnih instanc, ki so pravilno klasificirane v pozitivni razred*« (Karakatič in Fister, 2022, str. 100). »*Druga metrika je preciznost (angl. precision) in nam pove, kolikšen delež instanc, klasificiranih v pozitivni razred, je dejansko pripadnikov pozitivnega razreda.*« (Karakatič in Fister, 2022, str. 101). Prednost F-metrike je združevanje preciznosti in priklica v eno število, ker nam omogoča, da takoj opazimo nizko vrednost katerekoli metrike. F1-mera prikazuje »*obe vrednosti enako uteženi in imata uravnoteženo vlogo pri izračunu*« (Karakatič in Fister, 2022, str. 102).

3.3 Orodja

Za ekstrakcijo metapodatkov s pomočjo strojnega učenja so v raziskavah uporabljena različna orodja. Dejstvo je, da v dostopni literaturi ni na voljo veliko prispevkov o ekstrakciji metapodatkov s pomočjo strojnega učenja iz arhivskega gradiva. Zato so v nadaljevanju opisana orodja, ki se uporabljajo za ekstrakcijo metapodatkov iz znanstvene literature. To področje je dobro raziskovano, izkušnje in rezultati raziskav pa se lahko uporabijo tudi na področju arhivistike. Orodja so večinoma odprtokodna in brezplačna, kar nam omogoča dodaten razvoj in prilagajanje situacijam v arhivski znanosti in stroki.

3.3.1 CERMINE

CERMINE ali »Content Extractor and Miner« je Javina knjižnica in spletna aplikacija[7], ki omogoča ekstrakcijo metapodatkov iz znanstvenih publikacij v obliki PDF. Orodje je dostopno na podlagi licence GNU Affero General Public License, verzija 3. Orodje za ekstrakcijo metapodatkov uporablja več tehnik strojnega učenja, kot so metoda podpornih vektorjev, »k-means« in pogojna naključna polja. Orodje CERMINE je izbrano v raziskavi o ekstrakciji metapodatkov iz nemških znanstvenih člankov, ker »*omogoča zanesljivo ekstrakcijo za dokumente z zahtevno postavitvijo in omogoča zbiranje informacij o geometrični strukturi dokumenta, kot sta položaj besedila in slog pisave*« (Bouabdallah et al. 2021, str. 3). Kakor so zapisali Tkaczyk in ostali (2015), je »*CERMINE ... zasnovan kot univerzalna rešitev, zato lahko dobro obdeluje veliko najrazličnejših struktur publikacij, namesto da bi bil popoln pri obdelavi samo omejenega števila struktur dokumentov. To smo dosegli z uporabo nadzorovanih in nenadzorovanih algoritmov strojnega učenja, ki smo jih trenirali na velikih različnih naborih podatkov. Ta odločitev je prinesla tudi večjo možnost nadgradnje sistema in njegovo sposobnost prilagajanja novim, prej neznanim postavitvam dokumentov*« (Tkaczyk et al. 2015, str. 318).

[7] *Več: http://cermine.ceon.pl/index.html.*

Avtorji v nadaljevanju opisujejo način delovanja orodja CERMINE in ustvarjanje hierarhične strukture dokumenta, ki poleg vsebine ohranja informacije o načinu prikazovanja strukturnih elementov znotraj datoteke. Vsak strukturni element vsebuje informacije o vsebini in okviru, ki ga omejuje (Tkaczyk et al. 2015, str. 320).

Tkaczyk in ostali (2015) še navajajo, da postopek ekstrakcije metapodatkov z orodjem CERMINE vsebuje naslednje korake:

1. Analiza strukture dokumenta
    a. Ekstrakcija znakov – ekstrakcija posameznih znakov
    b. Segmentacija strani – združevanje znakov v besede, vrstice in besedilna polja
    c. Določanje vrstnega reda branja za vse strukturne nivoje
2. Razvrščanje vsebine
    a. Začetna razvrstitev besedilnih polj – razvrstijo se v eno od štirih kategorij: metapodatki, reference, telo in drugo
    b. Razvrstitev metapodatkovnih entitet – določanje ustrezne kategorije za metapodatkovne entitete (naslov, avtor …)
3. Ekstrakcija metapodatkov
    a. Razvrstitev metapodatkovnih entitet, kot je opisano v predhodnem koraku
    b. Ekstrakcija metapodatkov – pridobivanje jedrnih informacij iz označenih polj
4. Ekstrakcija bibliografije
    a. Ekstrakcija referenčnih nizov – razdelitev vsebine referenčnih polj na posamezne referenčne nize
    b. Razčlenitev referenc – ekstrakcija metapodatkov iz referenčnih nizov.
5. Evaluacija – opravljena je evaluacija naslednjih korakov: segmentacija strani, razvrščanje začetnih in metapodatkovnih polj ter razčlenitev referenc.

3.3.2 GROBID

GROBID[8] (GeneRatiOn of Bibliographic Data) je Javino odprtokodno orodje (Apache 2 licenca) za ekstrakcijo metapodatkov iz znanstvenih člankov. Deluje na operacijskih sistemih Linux in MacOS. Od leta 2008 razvija orodje Patrice Lopez s francoskega inštituta za računalniško znanost French Institute for Research in Computer Science and Automation. »*Orodje uporablja CRF-tehniko strojnega učenja za samodejno ekstrakcijo in prestrukturiranje vsebine iz neobdelanih in heterogenih virov v enotne dokumente, usklajene s standardom TEI*[9] *(Text Encoding Initiative)*« (Romary in Lopez, 2015, str. 2). Glede na uporabniško dokumentacijo ima GROBID več prednosti: hitri rezultati na manj zmogljivih računalnikih, razširljivost in zanesljivost, hitra obdelava datotek PDF, modularnost in ponovna uporaba modelov strojnega učenja, samodejno ustvarjanje podatkov za treniranje modela. GROBID omogoča ustvarjanje strojno berljivih, strukturiranih in predvidljivih vsebin s tehnikami rudarjenja po besedilu, ekstrakcije informacij in semantične analize znanstvenih publikacij (GROBID dokumentacija, b. d.).

---

[8] *Več: https://grobid.readthedocs.io/en/latest/.*

[9] *Konzorcij, ki razvija standard za prikazovanje podatkov v strojno berljivi obliki. Več informacij je dostopno na: https://tei-c.org/.*

Orodje GROBID je bilo primarno razvito za ekstrakcijo metapodatkov iz znanstvenih člankov. Glede na to, da je orodje odprtokodno in omogoča usklajevanje z drugimi CRF-modeli strojnega učenja, je treba dodatno raziskovati možnost uporabe orodja za ekstrakcijo metapodatkov iz arhivskega gradiva. Leta 2013 so Lipinski in ostali opravili raziskavo o delovanju orodij za ekstrakcijo metapodatkov iz znanstvenih člankov. Analizirana so bila odprtokodna orodja, ki se lahko vključijo v projekte, izdelane po meri, z izhodnimi podatki v strojno berljivi obliki, kakršen je na primer XML[10] (angl. eXtensible Markup Language). Od sedem analiziranih orodij je najboljše rezultate pokazal prav GROBID.

3.3.3 XTRACT

Xtract[11] je orodje za ekstrakcijo metapodatkov iz heterogenih zbirk. Primarno se uporablja za ekstrakcijo iz znanstvenih zbirk, vendar lahko uporabnik ustvari lasten ekstraktor, kot funkcijo Python[12] ali skripto Bash shell[13] (Skluzacek, 2022). Orodje »*uporablja oddaljeno in distribuirano računalniško infrastrukturo*« (Skluzacek, 2022, str. 61). Xtract omogoča ekstrakcijo metapodatkov iz različnih vrst datotek. Skluzacek (2022) ugotavlja, da Xtract uporablja različne modele strojnega učenja za določanje vrste datotek ter je že v osnovi projektiran na način, da omogoča uporabo na strojnem učenju temelječih ekstraktorjev po meri. Prednost orodja je, da ima uporabnik možnost sam ustvariti ekstraktorje, glede na svoje potrebe.

3.3.4 Druga orodja

V literaturi so dostopne informacije o orodjih, ki niso primarno ustvarjena za ekstrakcijo metapodatkov, temveč gre za knjižnice in zbirke algoritmov strojnega učenja, ki pa med drugim omogočajo ekstrakcijo metapodatkov.

Weka[14] ali »Waikato Environment for Knowledge Analysis« je zbirka algoritmov strojnega učenja. Orodje je razvito na novozelandski univerzi University of Waikato in je na voljo z licenco GNU. »*WEKA omogoča dostop do več standardnih tehnik strojnega učenja v integriranem okolju, ki uporabniku omogoča hitro preizkušanje različnih tehnik na kateremkoli naboru podatkov*« (Flynn, 2014, str. 109). Weka je Javino orodje, dostopno za operativne sisteme Linux, Windows in Macintosh. Po namestitvi orodja lahko uporabnik izbere enega izmed petih uporabniških vmesnikov: Explorer, Knowledge Flow, Experimenter, Workbench in Simple CLI ali ukazna vrstica. Weka je primerno orodje za uporabnike, ki še nimajo veliko izkušenj s strojnim učenjem, saj ponuja pakete vtičnikov z že narejenimi algoritmi strojnega učenja. Uporabniki, ki so strokovnjaki na področju strojnega učenja, se morajo pred uporabo orodja seznaniti s podatkovnimi strukturami, ki so dobro razložene v dokumentaciji. Orodje Weka je bilo uporabljeno v raziskavi o vplivu klasifikacije dokumentov na postopek ekstrakcije metapodatkov iz heterogenih zbirk. Na isti osnovni zbirki dokumentov je s pomočjo orodja Weka opravljeno podrobno testiranje algoritmov strojnega učenja s ciljem klasifikacije dokumentov. Čeprav je pravilnost zaznavanja bila 80-odstotna, se uspešnost močno razlikuje med različnimi metapodatkovnimi elementi (Flynn, 2014). Orodje Weka je

---

[10] *Oblika zapisa, ki se uporablja za prenos podatkov. Več informacij dostopno na: https://www.w3.org/XML/.*

[11] *Več: https://github.com/xtracthub.*

[12] *Programski jezik, več informacij dostopno na: https://www.python.org/.*

[13] *Program, ki omogoča posredovanje ukazov operativnemu sistemu, ki jih izvede.*

[14] *Več: https://www.cs.waikato.ac.nz/ml/weka/.*

uporabil Teregowda (2021) v raziskavi o uporabi računalniških tehnik pri iskanju v digitalnih knjižnicah. V raziskavi je poudarjena ekstrakcija metapodatkov, ki so v nadaljevanju osnova za citiranje, povezovanje in razvrščanje dokumentov. Predlagani model ekstrakcije metapodatkov (naslov, avtor, naslovi oddelkov, konteksti citatov in citati) uporablja nadzorovan klasifikator stojnega učenja, ki je narejen z orodjem Weka.

V doktorski disertaciji »*Bibliographic reference analysis in archival data using supervised machine learning and grammatical features*« so uporabljeni algoritmi strojnega učenja, dostopni v knjižnici **Scikit-learn**[15] (Philips, 2021). Čeprav v raziskavi ni bila opravljena ekstrakcija metapodatkov, so bili uporabljeni modeli strojnega učenja za ekstrakcijo tradicionalnih in slovničnih značilnosti, kot so: frekvenca besed v naslovu, frekvenca besed z velikimi ali malimi črkami in število besed. »*Scikit-learn je modul za Python, ki združuje širok nabor najsodobnejših algoritmov strojnega učenja za srednje obsežne nadzorovane in nenadzorovane probleme*« (Pedregosa et al. 2011, str. 2826). Scikit-learn je dostopen z licenco BSD[16] za operativne sisteme Windows, Linux in macOS. Prednost uporabe programskega jezika Python je približevanje strojnega učenja splošnim uporabnikom.

Clowder je odprtokodni repozitorij, ki organizira datoteke in metapodatke v mape, podatkovne nabore, zbirke in prostore. Osnovna enota je podatkovni nabor (angl. dataset), ki vsebuje datoteke, mape in metapodatke. Metapodatki so lahko zbrani ročno ali s pomočjo ekstraktorjev. Clowder se uporablja na različnih področjih, kot so raziskovanje znanstvenega gradiva, kulturna dediščina, vodenje po digitalnih razstavah, informatika s področja kliničnih raziskav, geoznanost, spremljanje stanja okolja, digitalna humanistika in družbene vede. Sistem Clowder omogoča avtomatično ekstrakcijo metapodatkov, ki temelji na neodvisnih ekstraktorjih, slednji pa se izvajajo kot zunanji procesi in komunicirajo s Clowderjem preko zunanjega vodila in vmesnika API (angl. Application Programming Interface) (Marini et al. 2018). Orodje omogoča implementacijo zunanjega ekstraktorja metapodatkov, ki lahko temelji na strojnem učenju. Prednost Clowderja je, da omogoča razširitev z različnimi modeli, ki tudi lahko delujejo na podlagi strojnega učenja.

Orodje **GATE**[17] ali »General Architecture for Text Engineering« je omenjeno v anketi o vrednotenju delno avtomatskih orodij za ustvarjanje metapodatkov. Orodje je specializirano za analizo besedil z velikim številom različnih uporabnikov, kot so podjetja in raziskovalni instituti. GATE vključuje več komponent, vendar je za področje ekstrakcije metapodatkov najbolj zanimiv odjemalec GATE Developer ali razvojno okolje za področje obdelave jezika, ki se uporablja za ekstrakcijo informacij (Park in Brenza, 2015). V dokumentaciji orodja je navedeno (b. d.), da »GATE Developer« med drugim tudi uporablja tehniko strojnega učenja, in sicer metodo podpornih vektorjev. Za ekstrakcijo informacij se uporablja GATE-ov sistem ANNIE (angl. a Nearly-New Information Extraction System)[18].

Javino orodje **Apache Tika**[19] je dostopno s strani Apache Software Foundation. Tika omogoča samodejno zaznavanje datotek, ekstrakcijo besedil in objektov ter ekstrakcijo metapodatkov. Pri ekstrakciji metapodatkov omogoča Tika izbiro izmed

---

15 *Več: https://scikit-learn.org/stable/index.html.*

16 *BSD ali Berkley Software Distribution je odprtokodna licenca, ki omogoča uporabo, spreminjanje in distribucijo programske opreme.*

17 *Več: https://gate.ac.uk/ie/.*

18 *Več informacij je dostopno na: https://gate.ac.uk/sale/tao/splitch6.html#x9-1190021.*

19 *Več: https://tika.apache.org/.*

standardnih metapodatkovnih modelov (na primer: Dublin Core[20]) ali ustvarjanje lastne metapodatkovne sheme. Orodje tudi omogoča zaznavanje jezika. Prednost Tike je velika skupnost uporabnikov, ki je omogočila razvoj več kot 200 ekstraktorjev. Tiko je uporabil Skluzacek (2022) v raziskavi o samodejni ekstrakciji metapodatkov. Iz dokumentacije in literature ni razvidno, ali Tika za ekstrakcijo metapodatkov uporablja tehnike strojnega učenja, kljub temu pa obstaja povezava med Tiko in strojnim učenjem, in sicer s projektom Apache Mahout[21], katerega cilj je uporaba strojnega učenja za zagotavljanje skupnega filtriranja, združevanja v gruče in kategoriziranja.

3.4 Modeli

Poleg programskih orodij, ki s pomočjo strojnega učenja ekstrahirajo podatke, je dostopnih tudi več modelov, ki so zmožni opraviti proces ekstrakcije metapodatkov. Zelo dobre rezultate je pokazal model BERT, o katerem od leta 2019 naprej obstaja vse več raziskav. Poleg BERT-a so v nadaljevanju opisani tudi drugi modeli strojnega učenja, ki se uporabljajo za ekstrakcijo metapodatkov.

3.4.1 BERT

BERT ali »Bidirectional Encoder Representations from Transformers« je Googlov model strojnega učenja, narejen za potrebe boljšega razumevanja jezika. BERT je prvič predstavljen leta 2018 in je danes dostopen kot del platforme Google Cloud. Osnova BERT-a je dvosmernost, kar pomeni, da lahko istočasno bere besedilo v obe smeri – z leve proti desni in z desne proti levi. BERT je predtreniran za dve pomembni nalogi: maskirano modeliranje jezika (angl. Masked Language Modeling – MLM) in predvidevanje naslednjega stavka (angl. Next sentence prediction) (Devlin et al. 2019).

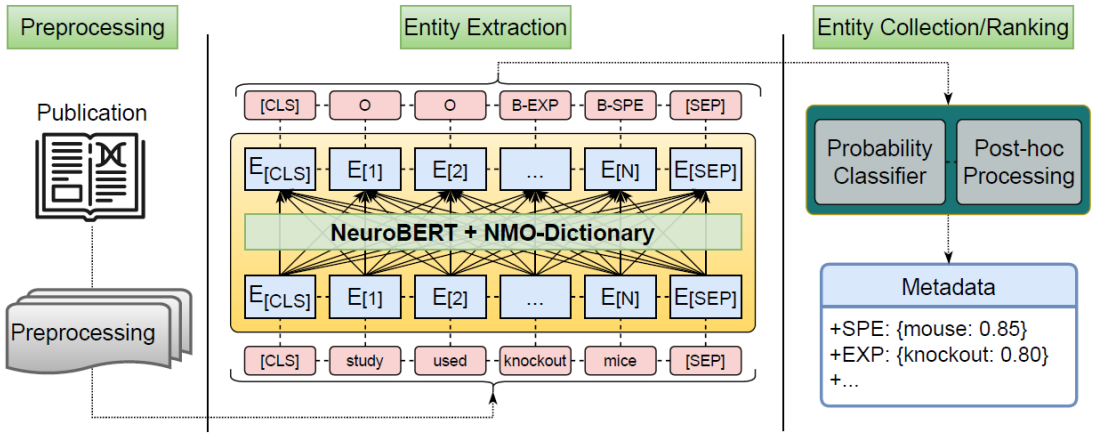
Model so uporabili v raziskavi ekstrakcije povezav in metapodatkov iz grških književnih besedil iz 19. stoletja (Christou in Tsoumakas, 2021).

Nadalje je uporabljen tudi v disertaciji »Enhanced Semi-Automated Metadata Extraction, Acquisition, and Management via Web Technologies and Machine Learning Models for NeuroMorpho.Org«, tj. za oblikovanje modela za ekstrakcijo metapodatkov s portala NeuroMorpho.org. Ustvarjen je sistem za ekstrakcijo metapodatkov (slika 1), ki v postopku predpriprave (angl. Preprocessing) označi začetek in konec besed ter ustvari zaporedje besed. V fazi ekstrakcije entitet preverja algoritem zaporedje besed in identificira pojavnost metapodatkovnih entitet. V zadnji fazi so predlagani izrazi razvrščeni na podlagi ocene ustreznosti (Bijari, 2022).

---

[20] *Dublin Core je metapodatkovna shema, ki določa osnovne metapodatkovne elemente. Vsebuje 15 osnovnih metapodatkovnih elementov. Več informacij je dostopnih na: https://www.dublincore.org/specifications/dublin-core/.*

[21] *Okvir, ki omogoča izvedbo aplikacij strojnega učenja. Več informacij: https://mahout.apache.org/.*



Slika 1: Struktura procesa za ekstrakcijo metapodatkov (Bijari 2022, str. 31)

Za označevanje zaporedja besed s približno 40.000 ciljnimi metapodatkovnimi oznakami je uporabljeno orodje DataTurks[22].

Model BERT je osnova jezikovnega modela Layout-MetaBERT, namenjen je ekstrakciji metapodatkov iz znanstvenih člankov. Okvir, ki temelji na Layout-MetaBERT-u, so imenovali LAME (angl. LAyout-Aware Metadata Extraction). »*Zaradi zagotavljanja dosledne kakovosti pri oblikovanju podatkov za učenje, ki temeljijo na strukturi, in gradnjo bolj izpopolnjenega modela za napredno ekstrakcijo metapodatkov predlagamo okvir LAyout-aware MEtadata extraction (LAME)*« (Choi et al. 2021, str. 2). LAME sestavljajo tri faze:

- prva faza – z orodjem PDFminer[23] se avtomatično analizira struktura dokumenta,
- druga faza – ustvarjanje velike količine metapodatkov, odvisnih od postavitve, na podlagi analize prve strani dokumenta,
- tretja faza – predtreniranje modela Layout-MetaBERT na podlagi treh velikosti modela Google BERT: drobni, majhni in osnovni.

Za potrebe okvira LAME je ustvarjen nabor podatkov iz 70 znanstvenih časopisov ali 65.007 dokumentov. Layout-MetaBERT osnovne velikosti je pokazal odlične rezultate pri ekstrakciji metapodatkov za še neznane publikacije z različnimi postavitvami.

3.4.2 Mask R-CNN

V prispevku »MexPub: Deep Transfer Learning for Metadata Extraction from German Publications« je narejena raziskava o ekstrakciji metapodatkov iz znanstvene literature s pomočjo modela Mask R-CNN (angl. Convolutional Neural Network), ki je del okvira MexPub. Model je treniran za dokumente formata PDF, ki so bili obdelani kot slike, zato ker v nemški znanstveni literaturi obstajajo velike razlike v postavitvi dokumentov. Model za ekstrakcijo metapodatkov uporablja tehnike globokega učenja in dostop, ki temelji na računalniškem vidu (computer vision-based). »*Naša metoda je dosegla povprečno točnost okoli 90 %, kar potrjuje njeno zmožnost natančnega pridobivanja metapodatkov iz različnih dokumentov PDF z zahtevnimi predlogami*« (Boukhers et al.

[22] *Več: https://github.com/DataTurks/DataTurks.*

[23] *Več: https://github.com/euske/pdfminer.*

2021, str. 1). Model Mask R-CNN za segmentacijo instanc objektov je namenjen zaznavanju objektov v slikah na ravni posameznih pikslov. V raziskavi je uporabljenih 100 nemških znanstvenih člankov, ki so imeli različno postavitev. Ker je večina metapodatkov na prvi strani, so opravili pretvorbo prve strani dokumenta iz formata PDF v JPG ter ročno določili področja, ki vsebujejo metapodatke. Zaradi težav z določanjem metapodatkov je bilo ustvarjenih 28 najpogostejših postavitev metapodatkov v dokumentu. Za vsako postavitev je bilo oblikovanih okoli 1600 sintetičnih dokumentov, kar je vplivalo na raznovrstnost vsebine in videza. Raznovrstnost je pomembna v primerih, ko en metapodatek (npr. naslov) zavzema več vrstic in pride do spremembe strukture. Rezultati validacije in testiranja na treniranju 1.5000 iteracij so pokazali točnost okoli 90 %, in sicer validacija 90,363 % in testiranje 90,167 % (Boukhers et al. 2021).

Po opravljenem testiranju je bila narejena primerjava s podobnim modelom GROBID, v ta namen so bili uporabljeni dokumenti s strukturo, ki se razlikuje od strukture, uporabljene za treniranje. Rezultati so prikazani na sliki 2. Izračunali so preciznost, preklic in F1-mero.

	MexPub			GROBID		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Title*	0.934	**0.947**	**0.940**	**0.965**	0.577	0.723
Author*	0.670	**0.851**	0.750	**0.982**	0.609	**0.752**
Journal*	**0.147**	**0.385**	**0.212**	0.000	0.000	0.000
Affiliation*	0.000	0.000	0.000	**1.000**	**0.118**	**0.210**
Abstract*	0.219	**0.833**	0.346	**0.972**	0.593	**0.739**
DOI	NaN	NaN	NaN	NaN	NaN	NaN
Address*	**0.125**	**1.000**	**0.222**	0.000	0.000	0.000
Email*	0.000	0.000	0.000	**1.000**	**0.500**	**0.667**
Date	0.250	1.000	0.400	NaN	NaN	NaN
Macro average	0.299	**0.574**	0.353	**0.703**	0.342	**0.442**
Micro average	0.558	**0.754**	**0.613**	**0.766**	0.447	0.559

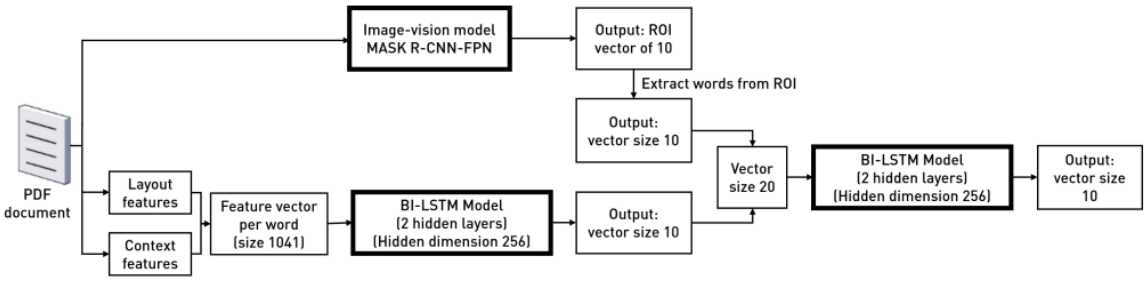
Slika 2: Primerjava okvira MexPub s programskim orodjem GROBID (Boukhers et al. 2021, str. 6)

Primerjava je pokazala različne rezultate delovanja orodji. GROBID ima večjo točnost in manjšo pojavnost lažno pozitivnih rezultatov, vendar ima precej nizek preklic, kar pomeni, da ima težave z ekstrakcijo pravilno pozitivnih vzorcev.

»*Čeprav MexPub ne upošteva nobenih kontekstualnih značilnosti, bi lahko dosegel dobre rezultate na še neznanih dokumentih s popolnoma različno postavitvijo. Vendar pa je glavna omejitev tega modela njegova mala prilagodljivost za publikacije z bistveno drugačno postavitvijo. Zato predvidevamo, da bi model pokazal boljše rezultate z vključitvijo kontekstualnih/besedilnih značilnosti*« (Boukhers et al. 2021, str. 6).

3.4.3 Model Bi-LSTM

Model Bi-LSTM (BiDirectional Long Short-Term Memory) je uporabljen v raziskavi o ekstrakciji metapodatkov iz nemških znanstvenih publikacij. Upošteva vizualne in kontekstne lastnosti dokumentov. Model je uporabljen za potrebe multimodalnega pristopa k ekstrakciji metapodatkov na podlagi globokega učenja. V raziskavi je tudi uporabljeno že opisano orodje CERMINE.



Slika 3: Diagram multimodalnega pristopa k ekstrakciji metapodatkov (Bouabdallah et al. 2021, str. 3)

Struktura modela (slika 3) je sestavljena iz treh delov: modela NLP, modela računalniškega vida in klasifikatorja. Pomemben del je predobdelava, ki pripravlja podatke za vnos v model. Za predobdelavo je uporabljeno orodje CERMINE, ki omogoča zanesljivo ekstrakcijo besedila iz zahtevnih struktur in pridobivanje informacij o geometrijski strukturi dokumenta, kot so položaj posameznega besedila in slogi pisave. Po ekstrakciji se za vsako besedo opravi ekstrakcija strukturnih in kontekstnih lastnosti. Rezultat je vektor velikosti 1041, ki predstavlja vhodne podatke v model NLP. Za fazo NLP se uporablja model biLSTM. Osnova tega modela je dvosmernost, ki je pokazala najboljše rezultate za kontekstualno razumevanje. Za fazo računalniškega vida se uporablja model iz MexPub-a, ker je ta že treniran za ekstrakcijo metapodatkov iz nemške znanstvene literature. V končni fazi je še uporabljen klasifikator oziroma model biLSTM, katerega rezultat je distribucija verjetnosti za vsako besedo, ki lahko spada v 10 razredov (Bouabdallah et al. 2021).

Rezultati (slika 4) so pokazali, da je multimodalni pristop v večini pokazateljev nekoliko boljši kot MexPub in GROBID.

	Our approach	MexPub	GROBID
Overall	**0.846**	0.823	0.618
Abstract	**0.923**	0.910	0.821
Author	0.807	**0.824**	0.770
Email	0.844	**0.901**	0.624
Address	**0.870**	0.821	0.324
Journal	**0.835**	0.828	0.741
Affiliation	**0.679**	0.535	0.240
Title	**0.964**	0.942	0.812

Slika 4: Primerjava multimodalnega modela (F1-mera) z MexPub-om in GROBID-om (Bouabdallah et al. 2021, str. 7)

4. ZAKLJUČEK

Postopek pridobivanja metapodatkov je izziv, posebej za arhive, ki poleg gradiva v fizični obliki hranijo tudi elektronsko gradivo. Na podlagi metapodatkov lahko zagotovimo načela varne hrambe arhivskega gradiva. Metapodatki so tudi pomembni, ker je uporabnik danes navajen iskati gradivo na podoben način, kot išče informacije na spletu, oziroma uporablja določene ključne besede, kot so datum ustvarjanja dokumenta, avtor, številka ali kombinacijo le-teh.

Raziskave so pokazale, da je proces pridobivanja metapodatkov za človeka načeloma enostaven, vendar je tudi naporen in preobremenjujoč. Zato je treba najti

rešitve, ki delujejo na podlagi novih tehnologij in bodo omogočale zanesljivo ekstrakcijo metapodatkov. Tehnike in modeli strojnega učenja nam lahko pomagajo pri ekstrakciji metapodatkov, za to nalogo so posebej primerne tehnike nadzorovanega strojnega učenja in orodja CERMINE, GROBID in Xtract. Orodje CERMINE združuje več različnih tehnik strojnega učenja in je v raziskavah pokazalo dobre rezultate pri ekstrakciji metapodatkov iz znanstvene literature. GROBID temelji na tehniki strojnega učenja CRF. Leta 2022 je bilo razvito orodje Xtract, ki omogoča ekstrakcijo metapodatkov iz različnih vrst dokumentov, kar je posebej zanimivo za arhive. Obstajajo še druga orodja, kot so WEKA, Clowder, Tika in GATE, ki se tudi lahko uporabijo v postopku ekstrakcije metapodatkov.

V prihodnosti potrebujemo več raziskav o avtomatizaciji procesa ekstrakcije metapodatkov iz arhivskega gradiva s pomočjo sodobnih tehnologij. Pri tem nam lahko pomagajo obstoječe raziskave o delovanju orodij in modelov: CERMINE, GROBID, Xtract, BERT, Bi-LSTM in Mask R-CNN, s ciljem prilaganja obstoječih rešitev arhivskemu gradivu ali ustvarjanja lastnih orodij in modelov strojnega učenja za ekstrakcijo metapodatkov.

5. LITERATURA

Bijari, K. (2022). *Enhanced Semi-Automated Metadata Extraction, Acquisition, and Management via Web Technologies and Machine Learning Models for NeuroMorpho.Org*. Doktorska disertacija. Fairfax: George Mason University.

Bouabdallah, A, Gavilan J., Gerbl J. in Patumcharoenpol, P. (2021). *Multimodal Approach for Metadata Extraction from German Scientific Publications.* Pridobljeno 18. julija 2023 s spletne strani: https://arxiv.org/abs/2111.05736. DOI: https://doi.org/10.48550/arXiv.2111.05736.

Boukhers, Z., Beili, N., Hartmann, T., Goswami, P. in Zafar, A.. (2021). MexPub: Deep Transfer Learning for Metadata Extraction from German Publications. *2021 ACM/IEEE Joint Conference on Digital Libraries* (JCDL). URL: https://ieeexplore.ieee.org/document/9651740. DOI: https://doi.org/10.1109/JCDL52503.2021.00076.

Brownlee, J. (2017). *Deep Learning for Natural Language Processing*. San Juan: Machine Learning Mastery.

Choi, J., Kong, H., Yoon, H., Oh, H. S., in Jung, Y. (2021). *LAME: Layout Aware Metadata Extraction Approach for Research Articles*. Pridobljeno 18. julija 2023 s spletne strani: https://arxiv.org/abs/2112.12353. DOI: https://doi.org/10.48550/arXiv.2112.12353.

Christou, D. in Tsoumakas, G. (2021). Extracting Semantic Relationships in Greek Literary Texts. *Sustainability, 13 (16).* URL: https://www.mdpi.com/2071-1050/13/16/9391. DOI: https://doi.org/10.3390/su13169391.

Councill, I. G., Giles, C. L. in Kan, M. K. (2008). ParsCit: an Open-source CRF Reference String Parsing Package. *Proceedings of the International Conference on Language Resources and Evaluation, LREC. (str. 661-667).* URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/166_paper.pdf.

Devlin, J., Chang M. W., Lee, K. in Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019. (str. 4171–4186)*. URL: https://aclanthology.org/N19-1423.pdf. DOI: **https://doi.org/10.48550/arXiv.1810.04805**

Flynn, P. K. (2014). *Document Classification in Support of Automated Metadata Extraction from Heterogeneous Collections.* Doktorska disertacija. Norfolk: Old Dominion University. DOI: 10.25777/vred-zd22

GATE – General Architecture for Text Engineering. (b. d.). Pridobljeno 2. avgusta 2023 s spletne strani: https://gate.ac.uk/.

GROBID Documentation. (b. d.). Pridobljeno 2. avgusta 2023 s spletne strani: https://grobid.readthedocs.io/en/latest/.

Groza, T., Grimnes, G., in Handschuh, S. (2012). Reference Information Extraction and Processing Using Conditional Random Fields. *Information Technology and Libraries. 31. (str. 6-20).* URL: https://ejournals.bc.edu/index.php/ital/article/view/2163. DOI: https://doi.org/10.6017/ital.v31i2.2163.

Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z. in Fox, E.A. (2003). Automatic document metadata extraction using support vector machines. *2003 Joint Conference on Digital Libraries, 2003*. Proceedings., Houston, TX, USA, (str. 37-48). URL: https://ieeexplore.ieee.org/document/1204842. DOI: https://doi.org/10.1109/JCDL.2003.1204842.

Hu, F. in Hao, Q. (2013). *Intelligent Sensor Networks: The Integration of Sensor Networks, Signal Processing and Machine Learning*. Boca Raton: CRC Press

International Council on Archives. (b. d.). Pridobljeno 2. avgusta 2023 s spletne strani: https://www.ica.org/en.

Karakatič, S. in Fister I. ml. (2022). *Strojno učenje: S Pythonom do prvega klasifikatorja*. Maribor: Univerza v Mariboru, Univerzitetna založba.

Lipinski, M., Yao, K., Breitinger, C., Beel, J. in Gipp, B. (2013). Evaluation of Header Metadata Extraction Approaches and Tools for Scientific PDF Documents. *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries. (str. 385-386).* URL: https://dl.acm.org/doi/10.1145/2467696.2467753. DOI: https://doi.org/10.1145/2467696.2467753.

Marini, L., Gutierrez-Polo, I., Kooper, R., Puthanveetil Satheesan, S., Burnette, M., Lee, J., Nicholson, T., Zhao, Y., in McHenry, K. (2018). Clowder: Open Source Data Management for Long Tail Data. PEARC '18: *Proceedings of the Practice and Experience on Advanced Research Computing (str. 1-8).* URL: https://dl.acm.org/doi/10.1145/3219104.3219159. DOI: https://doi.org/10.1145/3219104.3219159

Park, J. in Brenza, A. (2015). Evaluation of Semi-Automatic Metadata Generation Tools: A Survey of the Current State of the Art. *Information Technology and Libraries. 34 (3).* (str. 22-42). URL: https://ejournals.bc.edu/index.php/ital/article/ view/5889 (18. julij 2023). DOI:10.6017/ital.v34i3.5889

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M.,,Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. in Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research 12 (85).* (str. 2825–2830). URL: https://jmlr.org/papers/v12/pedregosa11a.html. **https://doi.org/10.48550/arXiv.1201.0490**

Philips, J. P. (2021). *Bibliographic reference analysis in archival data using supervised machine learning and grammatical features*. Doktorska disertacija. GreenVille: East Carolina University

Pravilnik o enotnih tehnoloških zahtevah (2020). Uradni list RS, št. 118.

Romary, L. in Lopez, P. (2015). *GROBID - Information Extraction from Scientific Publications. ERCIM News, 2015, Scientific Data Sharing and Re-use*, URL: https://hal.science/hal-01673305/

Shalev-Shwartz, S. in Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms.* New York: Cambridge University Press.

Skluzacek, T. J. (2022). *Automated metadata extraction can make data swamps more navigable.* Doktorska disertacija. Chicago: The University of Chicago. DOI: https://doi.org/10.6082/uchicago.4760

Tang, J. (2006). Template-based metadata extraction for heterogeneous collection. Doktorska disertacija. Norfolk: Old Dominion University. DOI: 10.25777/3w53-dq19

Teregowda, P. (2012). *Computational issues in digital library search engines. Doktorska disertacija.* University Park: The Pennsylvania State University.

Tkaczyk, Dominika. (2015). *New Methods for Metadata Extraction from Scientific Literature.* Doktorska disertacija. Varšava: University of Warsaw. DOI: https://arxiv.org/abs/1710.10201

Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., in Bolikowski, Ł. (2015). CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR) 18.* (str .317–335). URL: https://link.springer.com/article/10.1007/s10032-015-0249-8. DOI:10.1007/s10032-015-0249-8

Uredba o varstvu dokumentarnega in arhivskega gradiva (2017). Uradni list RS, št. 42.

SUMMARY

METADATA EXTRACTION USING MACHINE LEARNING

Ivančica SABADIN
PhD student of Archival Sciences at
Alma Mater Europaea – European Centre Maribor,
Slovenia
ivancica.sabadin@almamater.si

Metadata, or data about data, is important in the process of preserving archival material because it not only describes it, but also ensures its authenticity, integrity and usability. Metadata is extracted from archival material through an extraction process that is fundamentally simple for humans, but overwhelming given the number of electronic documents created on a daily basis. For this reason, the process of extracting metadata needs to be automated using tools and methods that have a degree of intelligence. Research has shown that machine learning techniques can provide reliable metadata extraction. Of particular interest for metadata extraction are supervised machine learning techniques such as SVM (Support Vector Machines), CRF (Conditional Random Fields) and HMM (Hidden Markov Model), and transformer-based systems such as BERT (Bidirectional Encoder Representations from Transformers).

When performing machine extraction of metadata, we need to be aware of the layout of the documents or archival material. An important characteristic of archival material is heterogeneity, which means that documents have different layouts depending on the type and content of the material and the creator of the material. There is a lot of research in the area of extracting metadata from scientific articles, which is also relevant to archives. This is because scientific articles, like archival material, are heterogeneous in terms of layout.