

Programsko orodje za avtomatizirano razpoznavanje in zakritje z zakonom varovanih podatkov (anonimizacijo)

Opis rešitve – Projekt e-ARH.si 2016-2020/2021

|  |  |
| --- | --- |
| **Operativni program:** | Operativni program za izvajanje Evropske kohezijske politike v obdobju 2014 -2020 |
| **Prednostna os:** | 11. Pravna država, izboljšanje institucionalnih zmogljivosti, učinkovita javna uprava, podpora razvoju NVO ter krepitev zmogljivosti socialnih partnerjev |
| **Prednostna naložba:** | 11.1 Naložbe v institucionalno zmogljivost ter učinkovitost javnih uprav in javnih storitev na nacionalni, regionalni in lokalni ravni za zagotovitev reform, boljše zakonodaje in dobrega upravljanja |
| **Specifični cilj:** | 11.1.3 Izboljšanje upravljanja in večja transparentnost v javni upravi z uvedbo novih orodij metod in interoperabilnih rešitev |
| **Naziv operacije:** | Razvoj slovenskega elektronskega arhiva e-ARH.si  |
| **Krajša oznaka:** | e-ARH.si: ESS 2016 – 2020 |
| **Odločitev o podpori:** | 11-1/3/MK/0 z dne 2.8.2016 (št. dokumenta 3032-16/2016/16) |
| **Naziv upravičenca:**  | Ministrstvo za kulturo, Arhiv Republike Slovenije |



13. julij 2021

Verzija dokumenta: 0.2

**Vsebina dokumenta**

[1. O anonimizaciji v SJAS 1](#_Toc77075917)

[2. Pridobitev in prilagoditev orodja za anonimizacijo 1](#_Toc77075918)

[3. Opis orodja za anonimizacijo 2](#_Toc77075919)

[4. Zaključek 4](#_Toc77075920)

**Verzije dokumenta in status**

|  |  |  |  |
| --- | --- | --- | --- |
| VERZIJA: | DATUM: | SPREMENIL: | STATUS |
| 0.1 | 8.7.2021 | Gregor Jenuš, Žiga Koncilija, Tatjana Hajtnik | Osnutek |
| 0.2 | 13.7. 2021 | Tatjana Hajtnik | Osnutek  |

**Kratice**

SJAS – Slovenska javna arhivska služba

OCR – Optical Character Recognition (slo. optična prepoznava znakov)

ZVDAGA – Zakon o varstvu dokumentarnega in arhivskega gradiva ter arhivih

DIP – Dostopni informacijski paket

eAG – Elektronsko arhivsko gradivo

AG – Arhivsko gradivo

ARS – Arhiv Republike Slovenije

# O anonimizaciji v SJAS

V okviru slovenske javne arhivske mreže (SJAS) se anonimizacija (zakritje z zakonom varovanih podatkov) izvaja na podlagi 65. člena Zakona o varstvu dokumentarnega in arhivskega gradiva in arhivih[[1]](#footnote-1), in sicer na dva načina; fizično na kopijah arhivskega gradiva in elektronsko na izvorno digitalnem arhivskem gradivu ali z uporabo digitalizatov arhivskega gradiva.

Slovenski arhivi pospešeno izvajajo digitalizacijo arhivskega gradiva, ki vsebuje z zakonom varovane kategorije podatkov. Digitalizati arhivskega gradiva se izdelujejo v formatih .pdf in .jpg oz. tiff. Anonimizacija se ne izvaja na formatih .pdf, kjer bi z optično prepoznavo znakov (OCR) in z uporabo plačljivih programskih rešitev lahko delo uporabnikov olajšali, ampak se izvaja na datotekah formatov jpg oz. tiff, kjer uporabniki z uporabo programskega orodja Slikar, ki je del programskega okolja MS Windows, prekrije nedostopne podatke, ki jih taksativno našteva drugi odstavek 65. člena ZVDAGA.

# Pridobitev in prilagoditev orodja za anonimizacijo

Glede na opravljeno poslovno analizo smo v okviru projektne naloge ugotavljali, da nobeno analizirano orodje v celoti ne ustreza našim zahtevam, zato smo se odločili, da poskušamo pridobiti orodje za anonimizacijo z izvedbo javnega razpisa. Januarja 2020 je bila tako na spletnem portalu javnih naročil objavljena razpisna dokumentacija *JN-20-004 Programska rešitev za avtomatizirano razpoznavanje in anonimizacijo z zakonom varovanih podatkov*.

Predmet javnega naročila je bila informacijska rešitev, ki omogoča izvajanje anonimizacije gradiva bodisi digitaliziranega ali izvorno digitalnega.

Natančnejše tehnične specifikacije so bile, da mora rešitev omogočati:

* uvoz gradiva posameznega dokumenta ali skupine dokumentov za anonimizacijo (najmanj v formatih .pdf, .jpg oz. .tiff);
* izvedbo avtomatizirane optične prepoznave znakov (najmanj OCR);
* izdelavo in dodajanje iskanih nizov ali besed v bazo (geslovnik), ki so predmet željene anonimizacije;
* možnost uvoza pred-nastavljenih baz, oziroma kreiranje in urejanje novih (obstajati mora tudi možnost izvoza baz v druga okolja);
* izvedbo avtomatizirane anonimizacije posameznega dokumenta ali skupine dokumentov;
* ročno (s signiranjem) označbo besed, delov besedila, predmeta ali področja gradiva, ki so predmet anonimizacije. Označeni deli besedila naj se vnašajo v bazo (geslovnik), ki se stalno veča in uporabnikom s predvidevanjem besed pokaže, kateri deli gradiva so morebiti takšni, da jih je treba anonimizirati;
* uporabniški vmesnik mora uporabniku v fazi pregledovanja dokumenta oz. besedila omogočati, da predlagane iskalne nize ali besede za anonimizacijo sprejme ali jih zavrne.

Anonimizacija se v rešitvi izvede na podlagi pred-nastavljene baze iskalnih besed ali besedilnih nizov, po korenih besed oz. ročno, kjer s signiranjem označimo besedo, del besedila, predmeta ali področja arhivskega gradiva, ki uporabnikom arhivskega gradiva ne smejo biti dostopni. Uporabniku programske rešitve mora programski uporabniški vmesnik omogočati, da vidi, katere dele arhivskega gradiva je označil.

Poudarjeno je tudi bilo, da mora programska rešitev omogočiti, da se z izvozom pripravi uporabniška kopija gradiva, v kateri so podatki prekriti tako, da neposreden ali posreden vpogled ni omogočen (v skladu s četrtim odstavkom 65. člena ZVDAGA) in da povrnitev v izvorno stanje dokumenta nikakor ni mogoča. Dodatne zahteve za programsko rešitev so še bile, da mora:

* uporabniku programske rešitve v uporabniškem vmesniku biti dana možnost izbire videza oznake za anonimizacijo in izvoznega formata;
* omogočati tudi izvoz v več formatov hkrati, najmanj v .jpg oz. .tiff in pdf; slednji mora vključevati optično prepoznavo znakov (najmanj OCR);
* ohraniti obe različici dokumenta, ki sta predmet postopka anonimizacije – anonimizirana različica kot tudi delovna različica vključno s pripadajočo bazo iskalnih besed ali besedilnih nizov (geslovnik), v katero se po potrebi lahko še dodajajo dodatne zahteve za prekritje podatkov;
* omogočati tudi, da dela na delovni različici dokumenta več uporabnikov hkrati ali zaporedno.

V specifikaciji je bilo še navedeno, da mora programska rešitev delovati najmanj v okolju MS Windows, licenca programske rešitve mora biti trajna (časovno neomejena pravica do uporabe licenčne programske opreme v okviru kupljenih licenc) z vključenim licenčnim vzdrževanjem (uporaba najnovejših različic licenčne programske opreme) za dobo treh let po namestitvi programske rešitve v okolje za redno rabo naročnika ter možnostjo sklenitve kasnejšega licenčnega vzdrževanja.

Programska rešitev mora prav tako omogočati namestitev na naročnikovem strežniku (strežniška licenca) in sočasno uporabo vsaj desetih uporabnikov z delovnimi postajami v okolju MS Windows. Namestitev na naročnikov strežnik pomeni hkratno namestitev v testnem okolju in okolju za redno rabo (produkcijsko okolje) naročnika. Med izvedbo naročila mora izvajalec zagotoviti komunikacijo z naročnikom v slovenskem jeziku.

Ponudnik programske rešitve mora v okviru razpisnih pogojev zagotoviti pripadajočo dokumentacijo: navodila za namestitev in konfiguracijo programske rešitve, navodila za administratorje ter navodila za končne uporabnike. Ponudnik mora izvesti predstavitev in šolanje za uporabo programske rešitve v obsegu vsaj dveh delovnih dni za največ 10 uporabnikov.

# Opis orodja za anonimizacijo

V okviru projektne naloge pridobitve orodja za anonimizacijo je bilo s strani izbranega ponudnika nameščeno orodje OpenText!BravaDesktop 16.6 in OpenText!Blazon Enterprise 16.6 oz. skupaj orodje za anonimizacijo v obliki 10 lokalnih licenc in eno strežniško licenco. Od tega so bile štiri licence nameščene na štiri virtualne računalnike v Arhivu Republike Slovenije, do katerih uporabniki dostopajo preko oddaljenega dostopa (Remote Desktop Protocol). Regionalno arhivi so dobili vsak po eno licenco (Zgodovinski arhiv Ljubljana, Pokrajinski arhiv Maribor, Zgodovinski arhiv Celje, Pokrajinski arhiv Koper, Pokrajinski arhiv v Novi Gorici in Zgodovinski arhiv Ptuj). Strežniška licenca OpenText!BravaBlazon se nahaja na strežniku v Arhivu Republike Slovenije.

Programska rešitev je produkt podjetja OpenText,[[2]](#footnote-2) ki je vodilno v svetu na področju sistemov za upravljanje z dokumenti oziroma informacijskimi sistemi za upravljanje z dokumenti (angl. Document Management System). Gre za rešitev, ki se jo namesti na računalnik oz. strežnik in velja za neomejeno število uporabnikov (t. i. seat license).

Postopek obdelave oz. anonimizacije se izvede po naslednjem vrstnem redu:

* uvoz dokumenta, ki je predviden za anonimizacijo (orodje podpira uvoz zelo velikega števila različnih formatov dokumentov (.pdf, .jpg, .png, .tiff itn.);
* izvedba »psevdoanonimizacije« (povratna oblika prekrivanja besed) in shranjevanje anonimizirane datoteke (v formatu .xrl ali .mrk);
* morebitno deljenje datotek drugim anonimizatorjem ter
* zaključek oz. »objava« anonimiziranega dokumenta (nepovratno anonimiziran dokument v željenem formatu).

Funkcionalnosti orodja za anonimizacijo so raznovrstne in so v grobem naslednje:

»Select« - označimo lahko predhodno anonimizirane dele dokumenta in jih dodatno oz. ponovno obdelamo (brišemo, kopiramo oz. premikamo).

»Redact Area« - ročno oblikujemo površino, na kateri se v dokumentu nahajajo občutljivi ali zaupni podatki, ki jih želimo anonimizirati.

»Redact Polygon« - ročno določimo eno ali več poljubnih večkotnih področij anonimizacije, vse ostale funkcionalnosti so podobne funkciji »Redact Area«.

»Redact Text« - ročno določimo področje anonimizacije v tekstu vendar brez slik.

»Allow Area (inside redaction)« - omogoča izrez območja, ki ga želimo znova razkriti oz. izključiti iz obstoječe anonimizacije.

»Redact Privacy Info« - namenjena je hitremu iskanju/anonimiziranju delno strukturiranih dokumentov (obrazcev) ter anonimizaciji podatkov kot so (številke socialnega zavarovanja, telefonske številke, e-poštni naslovi, datum rojstva ter ime in priimek).

»Find&Redact« - najdemo lahko in anonimiziramo več pojavitev iste besede ali besednih zvez. Predpogoj za uporabo funkcionalnosti je izvedena optična prepoznava znakov (OCR) v samem dokumentu. Podprta je uporaba iskanja z iskalnimi znaki, nadomestnimi znaki in t.i. »regularnimi izrazi«[[3]](#footnote-3). Znotraj tega funkcionalnost »Find Whole Words olnly« omogoča iskanje celotnih besed oz. besedilnih nizov. Če po besedilu iščemo besedilni niz v točno taki obliki kot je napisana, izberemo funkcionalnost »Match Case«.

»Redact Page« - omogoča hitro anonimizacijo celotne strani dokumenta z enim ukazom.

Bistvena prednost orodja za anonimizacijo je tudi ta, da je mogoče v obdelan tekst navajati razloge za anonimizacijo. Razlogi za anonimizacijo se shranjujejo v bazo (shranjeno v datoteki »Reasons.ini«), v vmesnem koraku se lahko anonimizirano besedilo in razloge preverja, v končni različici dokumenta pa so nato vsi razlogi prekritja navedeni. Kode so trajno vidne tako takrat, kadar datoteko odprete na računalniku, kot v katerem koli natisnjenem dokumentu. Pri preverjanju anonimiziranih dokumentov pred končno objavo lahko uporabnik preverja posamezne razloge oz. po njih tudi išče po določenih filtrih, kar bistveno olajša preverjanje anonimiziranih dokumentov. Pred končno objavo anonimiziranega dokumenta, dokument še ročno pregledamo in na koncu dokument objavimo (»Finalize Redactions and Publish As«).

Veliko pozornost se posveča tudi preverjanju anonimizacij pred samo končno objavo. Ko v dokumentu uporabite funkcionalnosti za anonimizacijo, lahko kasneje vedno zaženete orodje za ročno preverjanje anonimizacij z uporabo funkcije »Verify«. Šele po preverjanju lahko psevdoanonimiziran dokument objavite kot dokončno anonimiziranega. Funkcionalnost omogoča preverjanje anonimiziranega teksta, slik ali celotnih dokumentov ob varnostnem zadržku, da se lahko preverja le lastne anonimizacije. Preverja se lahko po barvno določenih delih dokumenta ali po razlogih, kar je zelo uporabna funkcionalnost. Funkcionalnost preverjanja celotnega dokumenta omogoča sistematično in natančno pregledovanje dokumenta. Aplikacija si zapomni, kdaj ste prenehali z anonimizacijo, lahko se dodaja oz. ureja anonimizacije itn.

Programska rešitev prav tako omogoča skupno anonimizacijo dokumentov, kar je z vidika uporabnikov arhivistov pomembna funkcionalnost v primerih, kadar želimo intervencijo drugih uporabnikov (npr. pristojnih arhivistov). To se izvede na način, da se shrani originalno (npr. .pdf) in anonimizirano datoteko (.xrl ali .mrk), ki jo prejemnik nato prejme. Po prevzemu datoteke s funkcijo »Take ownership of all editable markup« novi uporabnik prevzame možnost popravljanja obstoječih anonimizacij.

Pred dokončno (nepovratno) objavo končnega anonimiziranega dokumenta, se lahko objavi osnutek trenutnega dokumenta za pregled v formatu .pdf, .tiff ali .cfs (slednji je privzet format orodja za anonimizacijo). Anonimizacije v tem načinu ostanejo prozorne in na ta način omogočajo pregled tega, kar bo anonimizirano v končni nepovratni obliki. Torej je zgolj razširljiv posnetek tega (psevdoanonimizacija) kar bo anonimizirano v končni obliki (nepovratna anonimizacija). S funkcijo »Finalize redactions and publish as« ustvarimo končno objavo. To pomeni, da so anonimizacije na koncu neizbrisljiv del datoteke .pdf.

Vse funkcionalnosti anonimizacije, ki temeljijo na iskanju po tekstu oz. označevanju teksta, so odvisne od optičnega prepoznavanja znakov (OCR). Če nad datoteko slednje ni izvedeno, aplikacija izvede OCR, katerega kvaliteta je odvisna od originalne kvalitete slike.

Glede na to, da v poslovni analizi nismo zaznali ustrezne programske rešitve za samodejno anonimizacijo (glede na naše zahteve, torej na osnovi pred-nastavljene baze besed), smo kot eno izmed tehničnih zahtev zahtevali tudi slednje. Vzpostavljena programska rešitev omogoča vzpostavitev anonimizacijskih akcij v določenem vrstnem redu. To se izvede z uporabo skript oz. makrov. Snemanje makrov skript se lahko izvede na dva načina; bodisi ročno ali pa se ukazne akcije enostavno posname, kar se izvede s funkcionalnostjo »Start script recording«. Ročno pa se skripte ureja v datotekah formata .xml in ob uporabi regularnih izrazov. Za lažje upravljanje datotek z makri, je bila izdelana posebna aplikacija BravaManager, ki iz obstoječih geslovnikov (npr. Excel datotek) prebere fraze in jih doda v poljubno datoteko z makri.

Izvajalec je, skladno s pogodbenimi določili, izdelal ustrezna navodila za namestitev in uporabo OpenText!Brave, navodila za administratorje za OpenText!Blazon, prav tako so bila izdelana video navodila za uporabo in .ppt prezentacije z navodili za uporabo orodja za anonimizacijo.

Izvedenih je bilo več predstavitev in šolanj za uporabo programske rešitve. Aprila 2020 je bila izvedena predstavitev in presoja orodja glede ustreznosti izdelka. Junija 2020 je bila izvedena testna delavnica za deset udeležencev iz vseh arhivov SJAS. Oktobra 2020 pa je bilo na online delavnici končnim uporabnikom predstavljeno delovanje in osnovne funkcionalnosti orodja za anonimizacijo, kar je zajemalo postopke od uvoza datotek, do izvedbe različnih tipov anonimizacije ter prikaza izmenjave datoteke med več anonimizatorji.

Na šestih individualnih delavnicah je bilo izvedeno izobraževanje »superanonimizatorja«, na katerem se je, poleg pregleda osnovnih funkcionalnosti orodja, prikazalo še uporabo spletnega orodja za pomoč pri kreiranju regularnih izrazov, kreiranje skript ter uporabo orodja na primerih. Opredelilo se je tipične primere rabe orodja za anonimizacijo (samostalniške sklanjatve, spol in število, glagolski čas, pridevniki z izpeljankami) ter posebnosti anonimizacije za arhivsko gradivo v digitalni obliki (pogoste besede in besednih zvez ter opredelitev tipičnih primerov postopka anonimizacije). Naprednega uporabnika se je prav tako usposobilo za uporabo funkcionalnosti programa, namenjenega uporabi šifrantov pri anonimizaciji gradiva in seznanilo s postopkom kreiranja skript ter njihovo uporabo pri običajni in samodejni anonimizaciji.

# Zaključek

Projekt pridobitve orodja za anonimizacijo je šel skozi osnovne faze pridobivanja programske rešitev od izvedbe poslovne analize, v okviru katere smo analizirali obstoječe rešitve, do izdelave tehnične specifikacije, v kateri smo opredelili ključne funkcionalnosti, do izvedbe javnega naročila. Izbran izvajalec je nato izvedel namestitev, izobraževalne delavnice ter uvedel orodje v produkcijo. Poslovna analiza je pokazala, da obstoječa orodja ne omogočajo vseh zahtevanih funkcionalnosti, vključno s TACITO, ki je specializirana programska oprema za avtomatizirano anonimizacijo zgolj sodnega gradiva (to je praviloma z v naprej znano strukturo vsebine), v primerih anonimizacije drugega gradiva, pa se učinkovitost bistveno zmanjša. S pridobitvijo orodja za anonimizacijo (OpenText!Brava in OpenText!Blazon) smo v največji meri zagotovili zahtevanim funkcionalnostim. Torej uvoz gradiva posameznega dokumenta ali skupine dokumentov za anonimizacijo, izvedbo avtomatizirane optične prepoznave znakov, izdelavo in dodajanje iskanih nizov ali besed v bazo besed. Obstaja tudi možnost uvoza prednastavljenih baz besed oziroma kreiranje in urejanje novih ter tudi možnost njihovega izvoza. Orodje za anonimizacijo omogoča izvedbo avtomatizirane anonimizacije posameznega dokumenta ali skupine dokumentov ter tudi ročno (s signiranjem) označbo besed, del besedila, predmeta ali področij gradiva, ki so predmet anonimizacije. S signirjem označeni deli besedila se lahko vnašajo v bazo (geslovnik), ki se stalno veča in uporabnikom s predvidevanjem besed pokaže, kateri deli gradiva so morebiti takšni, da jih je treba anonimizirati, uporabniški vmesnik pa uporabniku v fazi pregledovanja dokumenta oz. besedila omogoča, da predlagane iskalne nize ali besede za anonimizacijo sprejme ali jih zavrne. Na ta način smo uspešno dosegli cilje zastavljene naloge v okviru projekta e-ARH.si in pridobili ustrezno rešitev za izvajanje anonimizacije. Orodje, navodila in usposobljen anonimizator, ki lahko nudi pomoč, so tako pri nas na voljo tudi drugim zainteresiranim strankam.

1. Ur. l. RS, št. 30/2006 in 54/2014. [↑](#footnote-ref-1)
2. https://www.opentext.com/ [↑](#footnote-ref-2)
3. Ang. »regular expression«. Regularni izrazi so sekvenca znakov, ki določajo iskalni vzorec. Običajno take vzorce uporabljajo določeni iskalni algoritmi. [↑](#footnote-ref-3)