



ARHIV REPUBLIKE SLOVENIJE
Zvezdarska 1, 1102 Ljubljana

T: 01 241 42 00, F: 01 241 42 76, E: ars@gov.si
www.arhiv.gov.si

Datum: 15. 1. 2020

Specifikacija zahtevanih funkcionalnosti programske rešitve za avtomatizirano razpoznavanje in anonimizacijo z zakonom varovanih podatkov

UVODNA POJASNILA

Marca 2016 je Vlada Republike Slovenije sprejela *Strategijo in izvedbeni načrt razvoja slovenskega elektronskega arhiva za obdobje 2016-2020 (dostopna je na spletnem naslovu)*.

https://www.gov.si/assets/organi-v-sestavi/Arhiv-RS/Projekt-e-ARH.si/1ead610374/Strategija_e-ARH_si_2016-2020_1.0.pdf

Strategija skupaj z izvedbenim načrtom do leta 2020 postavlja okvir za vzpostavitev stabilnega in zaupanja vrednega slovenskega elektronskega arhiva, s čimer bo omogočena uporaba in trajna ohranitev arhivskega gradiva v elektronski obliki kot dela slovenske, evropske in svetovne kulturne dediščine. Projekt s kratico imenujemo e-ARH.si in je trenutno sofinanciran s strani Evropske unije iz evropskih socialnih skladov za obdobje 2016 - 2020.

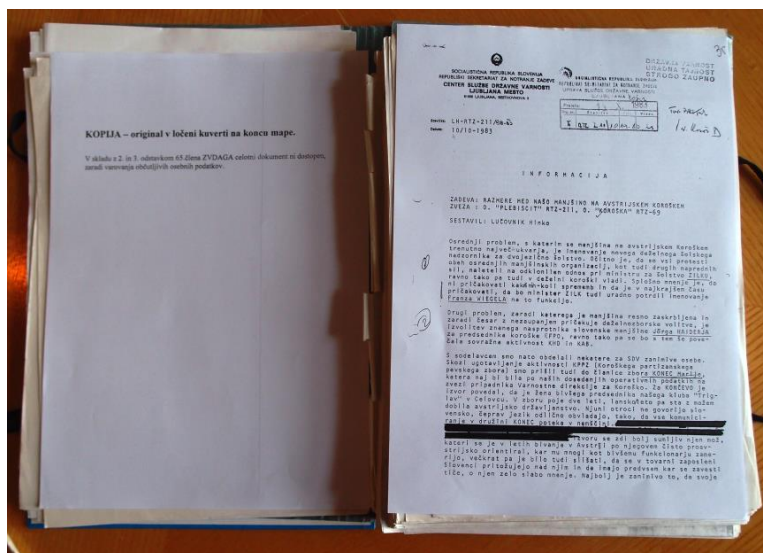
Eden izmed ciljev projekta je tudi pridobitev programske rešitve za anonimizacijo z zakonom varovanih podatkov.

SPLOŠNI PREGLED TRENUTNEGA STANJA IZVAJANJA ANONIMIZACIJE Z ZAKONOM VAROVANIH PODATKOV V ARHIVSKEM GRADIVU

V okviru slovenske javne arhivske mreže (SJAS)¹ se anonimizacija izvaja na podlagi 65. člena Zakona o varstvu dokumentarnega in arhivskega gradiva in arhivih (Ur. l. RS, št. 30/2006, 24/2914 – Odl. US in 54/2014) in sicer na dva načina; fizično na kopijah arhivskega gradiva in elektronsko na izvorno digitalnem arhivskem gradivu ali z uporabo digitalizatorov arhivskega gradiva.

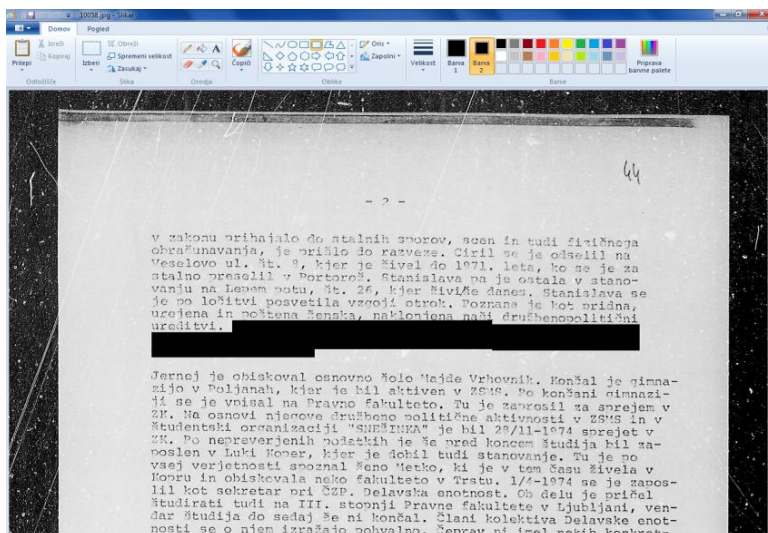
'Fizična' anonimizacija se izvaja na kopijah arhivskega gradiva z uporabo flomastrov (markerjev). Dokumente je potrebno fizično kopirati in izločiti v zapečatene kuverte, kopije pa na mestih, kjer je to zahtevano prekriti in ponovno kopirati, da se prepreči prosojnost besedila (Slika 1).

¹ Slovensko javno arhivsko mrežo sestavljajo Arhiv Republike Slovenije in regionalni arhivi (Pokrajinski arhivi v Kopru, Mariboru in v Novi Gorici ter Zgodovinski arhivi v Celju, Ljubljani in na Ptujju).



Slika 1: Primer izvedbe anonimizacije na fizičnem arhivskem gradivu.

Slovenski arhivi (posebno Arhiv Republike Slovenije) 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 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 (Slika 2).



Slika 2: Primer izvedbe anonimizacije na elektronskem arhivskem gradivu z uporabo programskega orodja Slikar v okolju MS Windows. Uporabniki anonimizacijo izvajajo s pomočjo vstavljanja objektov, s katerimi prekrivajo z zakonom varovane podatke.

Postopka anonimizacije (fizična in digitalna), ki sta trenutno v uporabi v SJAS, sta zamudna in hkrati tudi neučinkovita. Čeprav anonimizacija v digitalnem svetu prinaša več svobode in manevrskega prostora, ter uporabniku prihrani čas neprestanega fizičnega kopiranja dokumentov, ima svoje slabosti. Uporabniki morajo tako konstantno shranjevati dve obliki dokumentov – anonimiziranega in neanonimiziranega. Še posebno velja to, kadar je gradivo pogostejše v uporabi ali se izkaže, da je oseba, na katero se podatki nanašajo, umrla.

NAMEN DOKUMENTA

Dokument predstavlja vsebinski in tehnični okvir oz. specifikacijo zahtev za pridobitev programske rešitve za avtomatizirano razpoznavanje in anonimizacijo z zakonom varovanih podatkov.

PREDMET JAVNEGA NAROČILA

Predmet javnega naročila je programska rešitev, ki bo omogočala izvajanje anonimizacije arhivskega gradiva bodisi digitaliziranega ali izvorno digitalnega najmanj v formatih pdf, jpg ali tif z uporabo optične prepoznavne znakov (najmanj OCR) ter vzpostavitev baze digitaliziranega in anonimiziranega gradiva ter besednih nizov, ki so bili predmet anonimizacije.

Programska rešitev mora 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, možnost prijave napake v licenčni programski opremi ponudniku) za dobo treh let po namestitvi programske rešitve v okolje za redno rabo naročnika ter možnostjo sklenitve kasnejšega licenčnega vzdrževanja.

Programska rešitev mora 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.

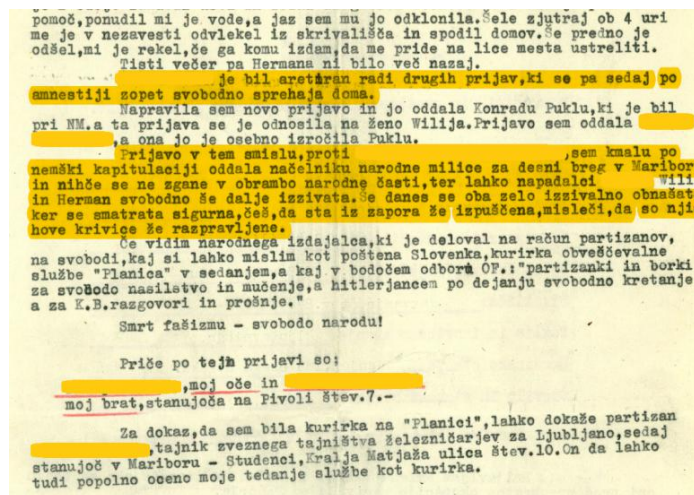
Med izvedbo naročila mora izvajalec zagotoviti komunikacijo z naročnikom v slovenskem jeziku.

Funkcionalne zahteve programske rešitve

Programska rešitev mora omogočati:

- **uvoz gradiva posameznega dokumenta ali skupine dokumentov za anonimizacijo** (najmanj v formatih pdf, jpg oz. tif) v programsko rešitev,
- **izvedbo avtomatizirane optične prepoznavne znakov** (najmanj OCR),
- **izdelavo in dodajanje iskanih nizov ali besed v bazo** (geslovnik ali vreča besed), ki so predmet anonimizacije; obstajati mora možnost uvoza prednastavljenih baz oziroma kreiranje in urejanje novih baz; obstajati mora tudi možnost izvoza baz;
- **izvedbo avtomatizirane anonimizacije posameznega dokumenta ali skupine dokumentov**;
- **ročno** (s signiranjem) **označbo besede, del besedila, predmeta ali področja arhivskega gradiva, ki so predmet anonimizacije**. S signiranjem označeni deli besedila naj se vnašajo v bazo (geslovnik ali vrečo besed), 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**.

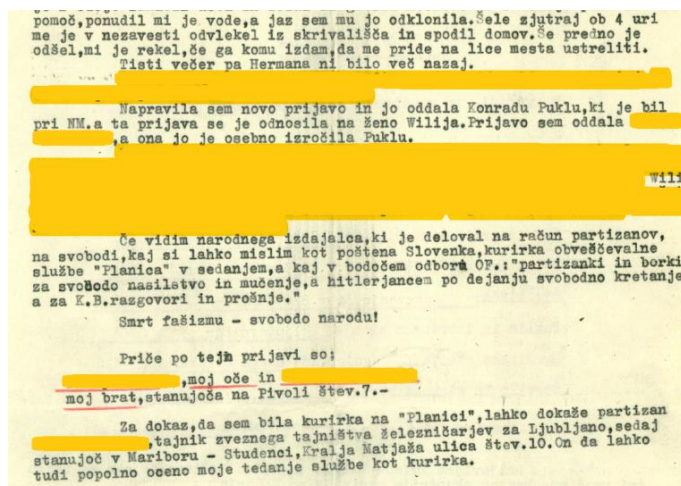
Anonimizacija se mora izvesti na podlagi pred-nastavljene baze iskalnih besed ali besedilnih nizov, po korenih besed oz. ročno, kjer s signiranjem označimo besedo, del besedila, predmeta ali področja arhivskega gradiva, ki uporabnikom arhivskega gradiva ne smejo biti dostopni. Uporabniku programskega rešitve mora programski uporabniški vmesnik omogočati, da vidi, katere dele arhivskega gradiva je označil - brez potrebe po podvojenemu odpiranju dokumentov (Slika 3).



Slika 3: Pregled označenih delov arhivskega gradiva v postopku anonimizacije.

Programska rešitev mora 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 četrtem odstavkom 65. člena ZVDAGA) in da povrnitev v izvorno stanje dokumenta nikakor ni mogoča. Uporabnik programske rešitve mora v uporabniškem vmesniku imeti možnost izbirati **videz oznake za anonimizacijo in izvoznega formata**. Programska rešitev mora omogočati izvoz v več formatov hkrati. Najmanj jpg oz. tif in pdf. Slednji mora vključevati optično prepoznavo znakov (najmanj OCR) (Slika 4).

Ohraniti se morata obe različici dokumenta, ki je bil predmet postopka anonimizacije – anonimizirana različica kot tudi delovna različica vključno s pripadajočo bazo iskalnih besed ali besedilnih nizov (geslovník ali vreča besed), v katero se po potrebi lahko še dodajajo dodatne zahteve za prekritje podatkov. Programska rešitev mora omogočiti, da na delovni različici dokumenta, dela več uporabnikov hkrati ali zaporedno.



Slika 4: Pregled anonimiziranega arhivskega gradiva.

Nefunkcionalne zahteve:

Ponudnik programske rešitve mora zagotoviti pripadajočo dokumentacijo: navodila za namestitvev in konfiguracijo programske rešitve, navodila za administratorje ter navodila za končne uporabnike (slednje mora biti v slovenskem jeziku).

Ponudnik mora izvesti predstavitev in šolanje za uporabo programske rešitve v obsegu vsaj dveh delovnih dni za največ 10 uporabnikov.