

Projekt digitalizacije hemeroteke Hrvatskog povijesnog muzeja

Jelena Balog Vojak
Viši dokumentarist
Hrvatski povijesni muzej
j.balog@hismus.hr

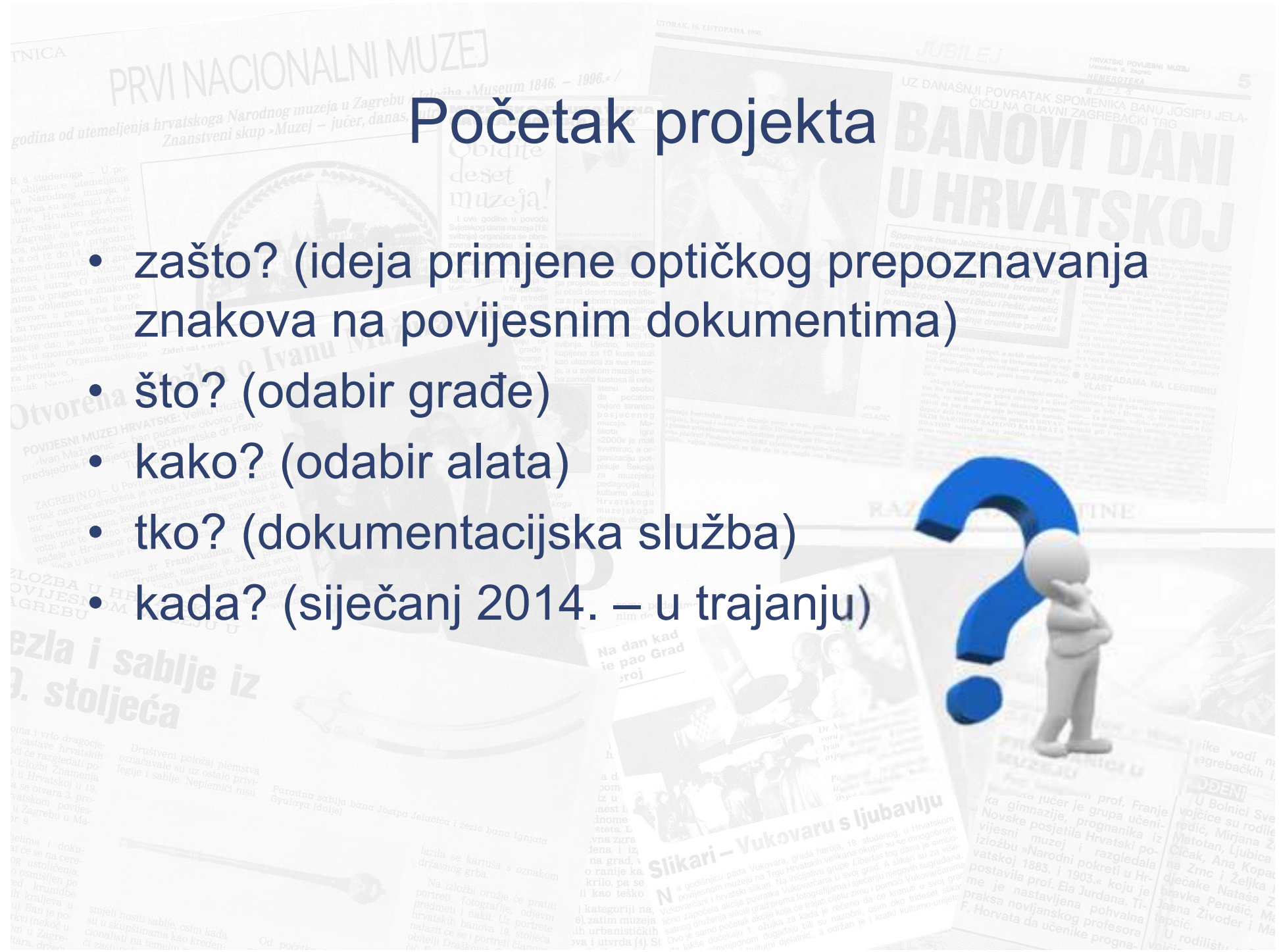
Zdenka Šinkić
Informatičar
Hrvatski povijesni muzej
z.sinkic@hismus.hr

ČETVRTI FESTIVAL HRVATSKIH DIGITALIZACIJSKIH PROJEKATA
Zagreb, 10. travnja 2014.



Početak projekta

- zašto? (ideja primjene optičkog prepoznavanja znakova na povijesnim dokumentima)
- što? (odabir građe)
- kako? (odabir alata)
- tko? (dokumentacijska služba)
- kada? (siječanj 2014. – u trajanju)

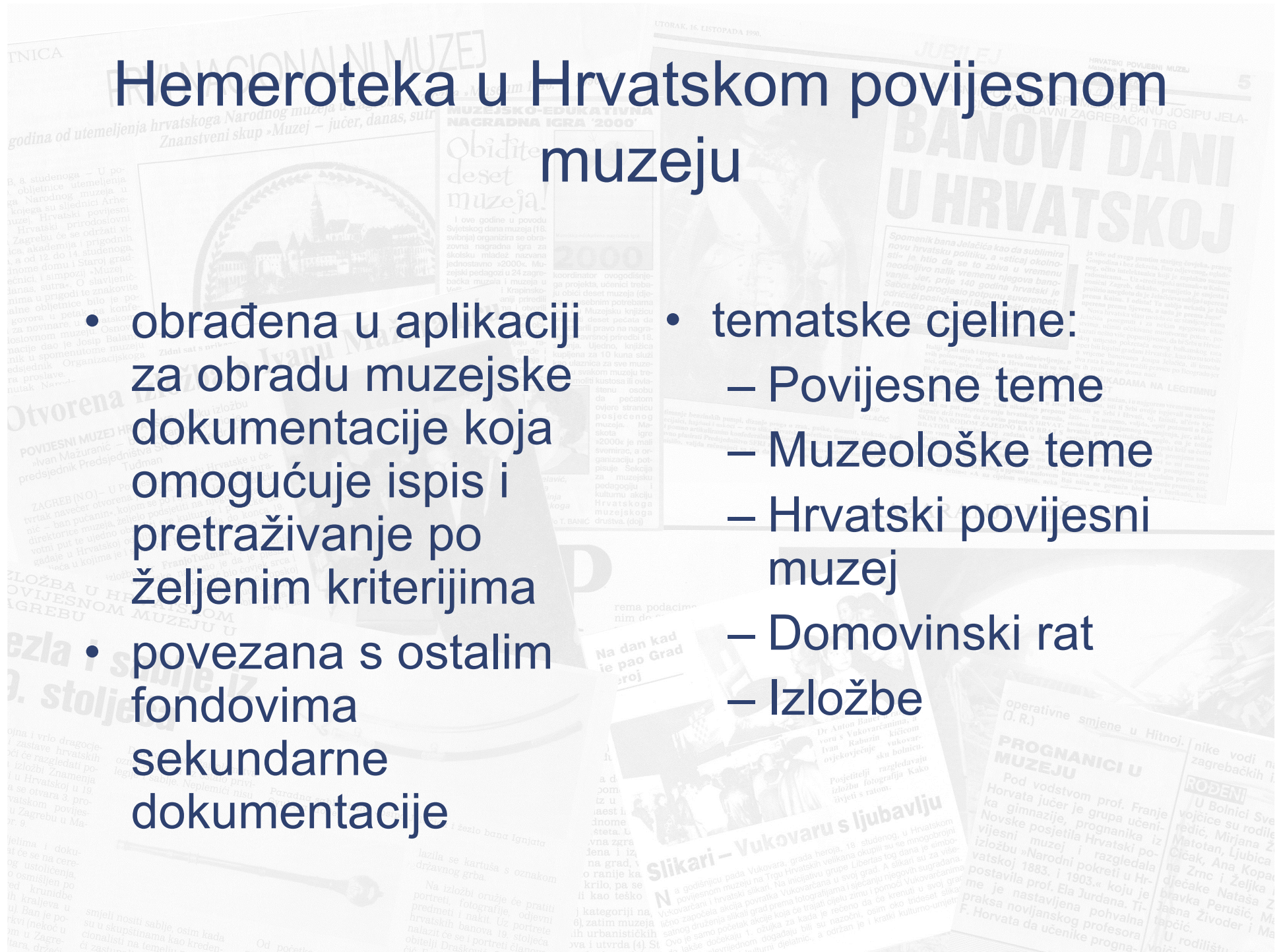


Hemeroteka u Hrvatskom povijesnom muzeju

- hemeroteka = zbirka izrezaka iz tekućih novina, časopisa ili koje druge tiskane građe, o određenim, unaprijed utvrđenim temama ili predmetima...
- hemeroteka je jedan od fondova sekundarne dokumentacije, te uz izdavačku i izložbenu djelatnost pripada u najbrojnije fondove sekundarne dokumentacije Hrvatskog povijesnog muzeja
- više od 60 godina
- oko 2.000 jedinica, djelomično skeniranih

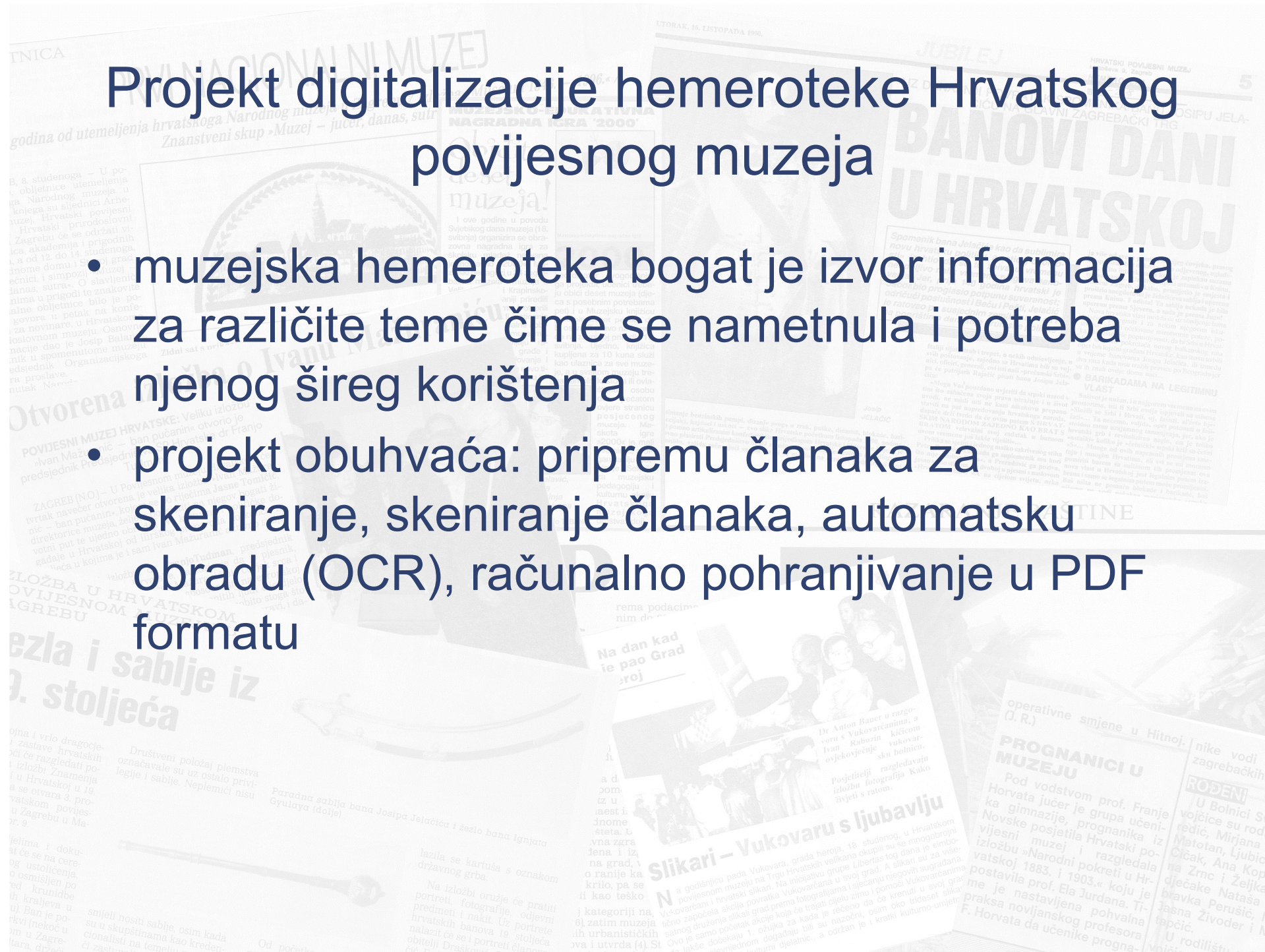
Hemeroteka u Hrvatskom povijesnom muzeju

- obrađena u aplikaciji za obradu muzejske dokumentacije koja omogućuje ispis i pretraživanje po željenim kriterijima
- povezana s ostalim fondovima sekundarne dokumentacije
- tematske cjeline:
 - Povijesne teme
 - Muzeološke teme
 - Hrvatski povijesni muzej
 - Domovinski rat
 - Izložbe



Projekt digitalizacije hemeroteke Hrvatskog povijesnog muzeja

- muzejska hemeroteka bogat je izvor informacija za različite teme čime se nametnula i potreba njenog šireg korištenja
- projekt obuhvaća: pripremu članaka za skeniranje, skeniranje članaka, automatsku obradu (OCR), računalno pohranjivanje u PDF formatu

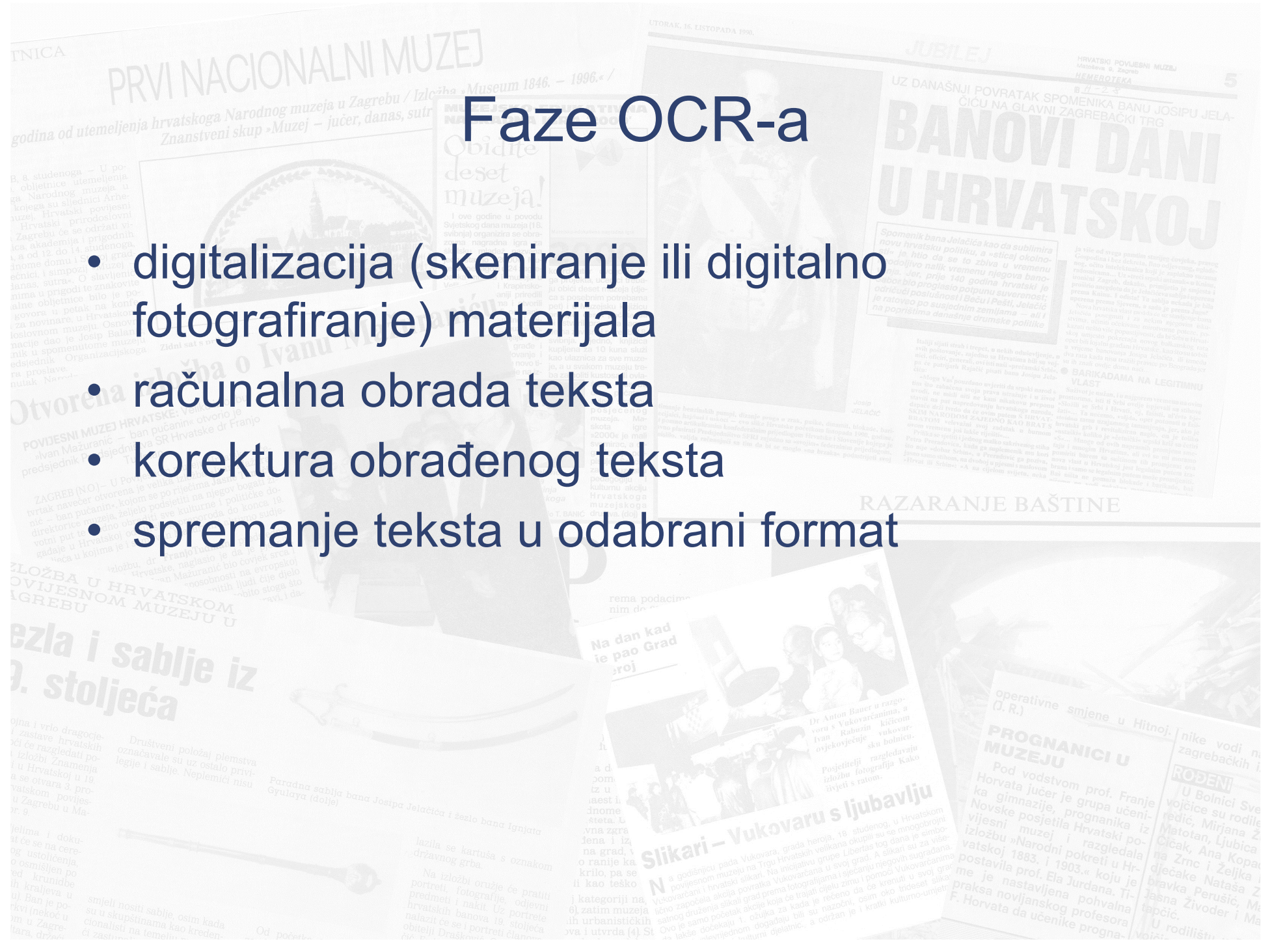


Optičko prepoznavanje znakova

- Optičko prepoznavanje znakova (eng. **Optical Character Recognition, OCR**) je računalna tehnologija koja omogućuje pretvorbu skeniranih papirnatih dokumenata, PDF dokumenata i slikovnih dokumenata snimljenih pomoću digitalnog fotoaparata u formate koji se mogu uređivati
- OCR sustav se sastoji od skenera ili fotoaparata i sofisticiranog softvera

Faze OCR-a

- digitalizacija (skeniranje ili digitalno fotografiranje) materijala
- računalna obrada teksta
- korektura obrađenog teksta
- spremanje teksta u odabrani format



Digitalizacija

- opremu čine dva skenera koje Muzej posjeduje u svrhu digitalizacije građe koju vršimo sami unutar Muzeja



Epson GT 2000(A3)



Epson V750 PRO (A4)

- rezolucija 600 dpi
- manji dio fonda već digitaliziran u PDF i JPG formatu

Digitalizacija - računalna obrada teksta

- nakon što su skenirani, dokumenti prelaze u drugu fazu digitalizacije odnosno šalju se na „prepoznavanje“ uz pomoć OCR tehnologije
- odabir programa koji ima najvjerniju mogućnost reproduciranja dokumenta u digitalni format prihvatljiv za daljnju upotrebu

Pokušajima do pravog izbora...

• i2OCR



[Home](#) [Features](#) [FAQ](#) [About](#)

Vašem(to

KULTUA

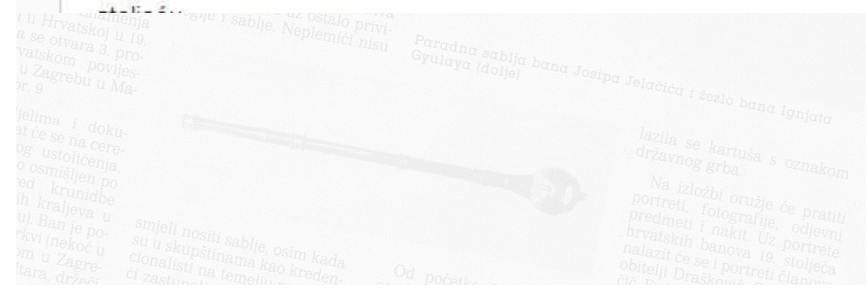
IZLOŽBA HRVATSKOGA POVJESNOG MUZEJA U MESTROWĆEVU PAVILJONU

Taman kad je formalno i stvarno dokinut Muzej revolucije naroda Hrvatske, eksponaū pxipali Hrvatskom povijesnom muzeju, a za Meštrovićev paviljon na Tr hrvatskih velikana u &grebu trazila se najprikladnija namjena i dokazivalo vlasništvo A na hrvatskim se prostorima dogodilo novi rat. Ne svjetski, ali još jednom strašan, nama treći samo u ovome



Slikari – Vukovaru s ljubavlju

Na godišnjicu pada Vukovara, grada heroja, 18. studenog, u Hrvatskom povijesnom muzeju na Trgu Hrvatskih velikana okupili su se mnogostrojno ično započela akcija povratka Vukovarčana u svoj grad. A slikari su za vijećatno društvo postavili sliku koja će trajati cijelu zimu i poručiti Vukovarčanima. Ovo je samo početak akcije koja za kada je rečeno da će krenuti u svoj grad. Ova je akcija namijenjena mladima i onima koji su se uključili u ovaj, grad. Ova je akcija namijenjena mladima i onima koji su se uključili u ovaj, grad. Ova je akcija namijenjena mladima i onima koji su se uključili u ovaj, grad.



Pokušajima do pravog izbora...

- **ONLINE OCR**



[Home](#) [Workspace](#) [Contact Us](#)

Welcome

Username:

grickavjestica

Free credits

25

[Workspace](#)

[Settings](#)

[Buy credits](#)

[Log out](#)

What is new ?

New Feature !

OCR Web Service

[learn more...](#)

Email OCR

[learn more...](#)

OnlineOCR.net is a Free Online OCR (Optical Character Recognition software) service that allows you to convert scanned PDF and

Online OCR

Recognition languages:

- | | | | |
|--|-------------------------------------|-------------------------------------|------------------------------------|
| <input type="checkbox"/> ENGLISH | <input type="checkbox"/> DUTCH | <input type="checkbox"/> ITALIAN | <input type="checkbox"/> RUSSIAN |
| <input type="checkbox"/> BRAZILIAN | <input type="checkbox"/> ESTONIAN | <input type="checkbox"/> LATIN | <input type="checkbox"/> SERBIAN |
| <input type="checkbox"/> BULGARIAN | <input type="checkbox"/> FINNISH | <input type="checkbox"/> LATVIAN | <input type="checkbox"/> SLOVAK |
| <input type="checkbox"/> BYELORUSSIAN | <input type="checkbox"/> FRENCH | <input type="checkbox"/> LITHUANIAN | <input type="checkbox"/> SLOVENIAN |
| <input type="checkbox"/> CATALAN | <input type="checkbox"/> GERMAN | <input type="checkbox"/> MOLDAVIAN | <input type="checkbox"/> SPANISH |
| <input checked="" type="checkbox"/> CROATIAN | <input type="checkbox"/> GREEK | <input type="checkbox"/> POLISH | <input type="checkbox"/> SWEDISH |
| <input type="checkbox"/> CZECH | <input type="checkbox"/> HUNGARIAN | <input type="checkbox"/> PORTUGUESE | <input type="checkbox"/> TURKISH |
| <input type="checkbox"/> DANISH | <input type="checkbox"/> INDONESIAN | <input type="checkbox"/> ROMANIAN | <input type="checkbox"/> UKRAINIAN |

Output formats:

- Adobe PDF (pdf)
- HTML 4.0 (html)
- MS Excel (xls)
- MS Word (doc)
- RTF Word (rtf)
- Text Plain (txt)

- Convert to Black and White (it is recommended for photos)
- Multipage document
- Demonstration mode

[How to use - Video Tutorial](#)

Datoteka nije odabrana.

You can upload more than one file at once by placing the files in a ZIP archive. Max file size: 100 MB
You can also use our service via [Email OCR](#)



Korektura obrađenog teksta

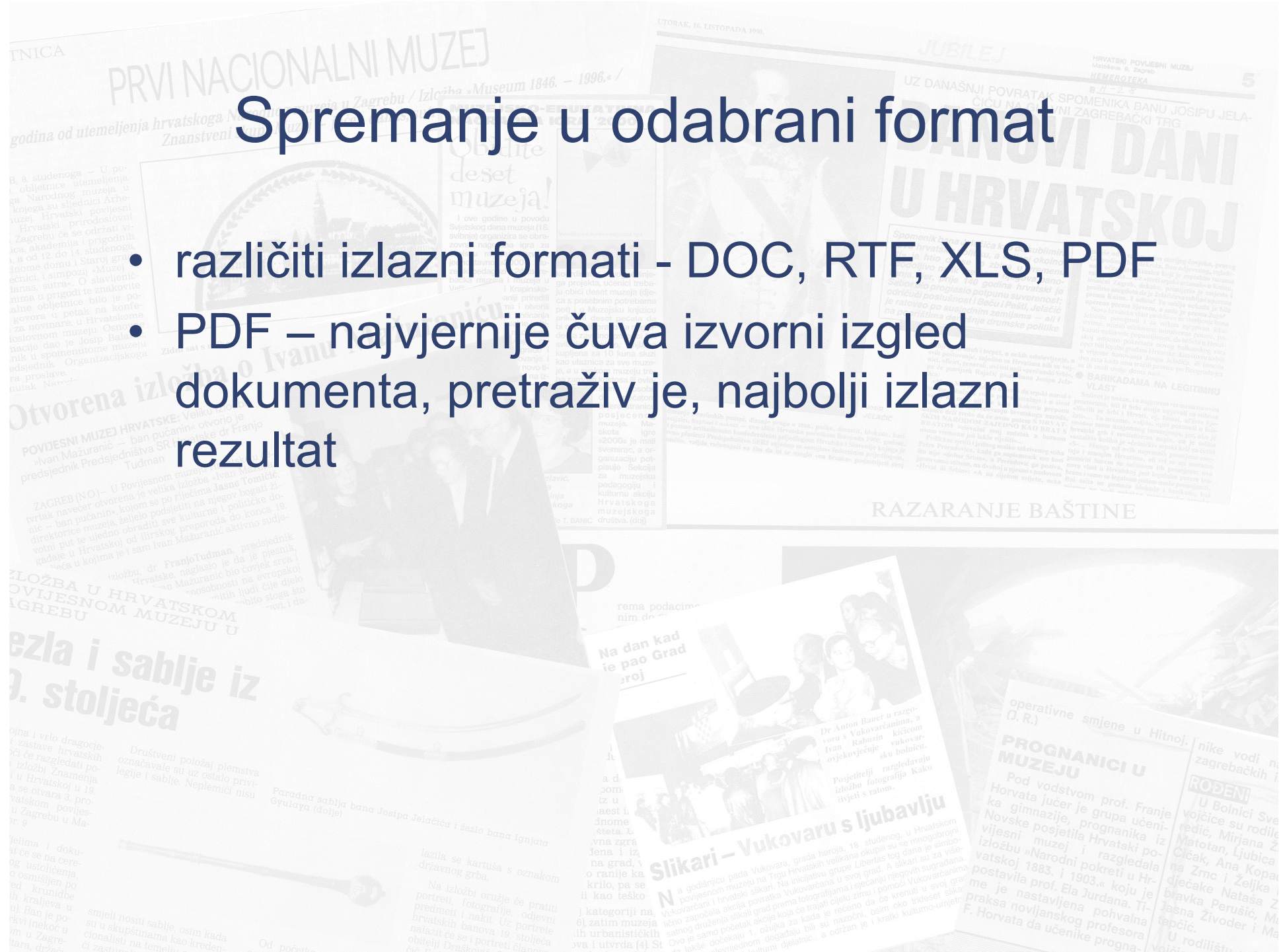
- brzina OCR-a impresionira: obrađuje jednu stranicu za nekoliko sekunda, što je 100 puta brže od profesionalnog daktilografa, a u isto vrijeme radi manje grešaka
- učinkovitost OCR-a 95% - 99% ipak podrazumijeva potrebu dodatne ručne korekture dokumenta

HRVATSKE POVIJESNE ZASTAVE

bRVrLTSK e POVI|6SNe ZA3TAV6

Spremanje u odabrani format

- različiti izlazni formati - DOC, RTF, XLS, PDF
- PDF – najvjernije čuva izvorni izgled dokumenta, pretraživ je, najbolji izlazni rezultat



Izazovi projekta

- koju građu odabrati – fond hemeroteke
- odabir besplatnog programa koji ima najvjerniju mogućnost reproduciranja
- odabir formata za pohranjivanje
- program koji se pokazao najučinkovitijim alatom nije mogao pročitati jedinice fonda hemeroteke koje su već ranije digitalizirane
- postojeće PDF datoteke pretvorili smo u slikovne datoteke pomoću Adobe Photoshopa

Pretraživanje prema sadržaju

- rezultati projekta kopirani u polje „sadržaj“ unutar aplikacije, koje je pretraživo po čitavom sadržaju

Sadržaj:

IZLOŽBE - Ostavština osječke slobodnozidarske lože
»Budnost« u Hrvatskome povijesnom muzeju u Zagrebu
TAJNA UDRUGA VIĐENIH GRAĐANA
Slike, fotografije, novine, časopisi, isprave, knjige, pokućstvo
te jedaći i crtači pribor i nacrti prate djelovanje osječke

- hemeroteca se sadržajno pretražuje te je značajno unaprijeđena njezina dostupnost i iskoristivost

i	Sadržaj (Inventarna knji	sadrži	budnost		
i		=			
i		=			

