

# Obrazová data a strojové učení

Petr Žabička, Ján Bogár, Michal Tran - Moravská zemská knihovna v Brně

# Obsah

- Úvod
- Projekt PERO
- Identifikace obrázků na stránce
- Systém VISE



# Strojové učení a proces digitalizace

- Ořez, vyrovnání stránky, barevné podání
- Scelování dokumentů
- Struktura dokumentu
- OCR
- ...




# Strojové učení a zpřístupňování

- Vyhledávání obrázků
- Vyhledávání a podobnost částí stránek
- Identifikace obsahu obrázku
- ...



# OCR - projekt PERO



←  Digitální studovna Ministerstva obrany ČR

Hledat v celé digitální knihovně

Sbírký Procházet Informace English

Hledat v dokumentu


9 z 122 stránek

1 2 3 4 5 6 7 8 9 10 11 12

3

smě si naši byt na půdě  
malého domku a dva dny  
smě čekali co si bude dít!

Domček v  
Petrovaradině



Třetí den ráno smě udělali se

Muj Deňík od r 19 30/7 14

Nakladatelské údaje  
[S. I.], 1914-1915

Typ dokumentu  
Kniha

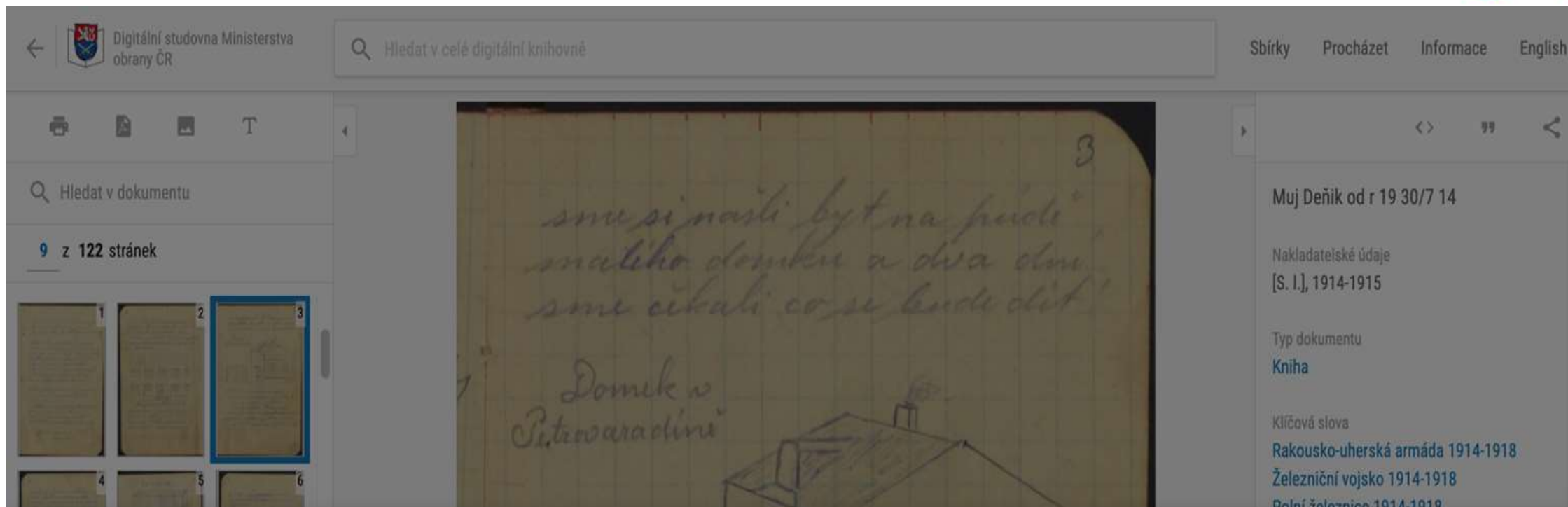
Klíčová slova  
Rakousko-uherská armáda 1914-1918  
Železniční vojsko 1914-1918  
Polské železnice 1914-1918  
Srbská fronta 1914-1915  
Východní fronta 1914-1918

Jazyk  
Čeština

Místo uložení  
Vojenský historický ústav Praha  
Signatura: XIII-10715 (př. č. 2862/2005)

Fyzický popis

# OCR - projekt PERO



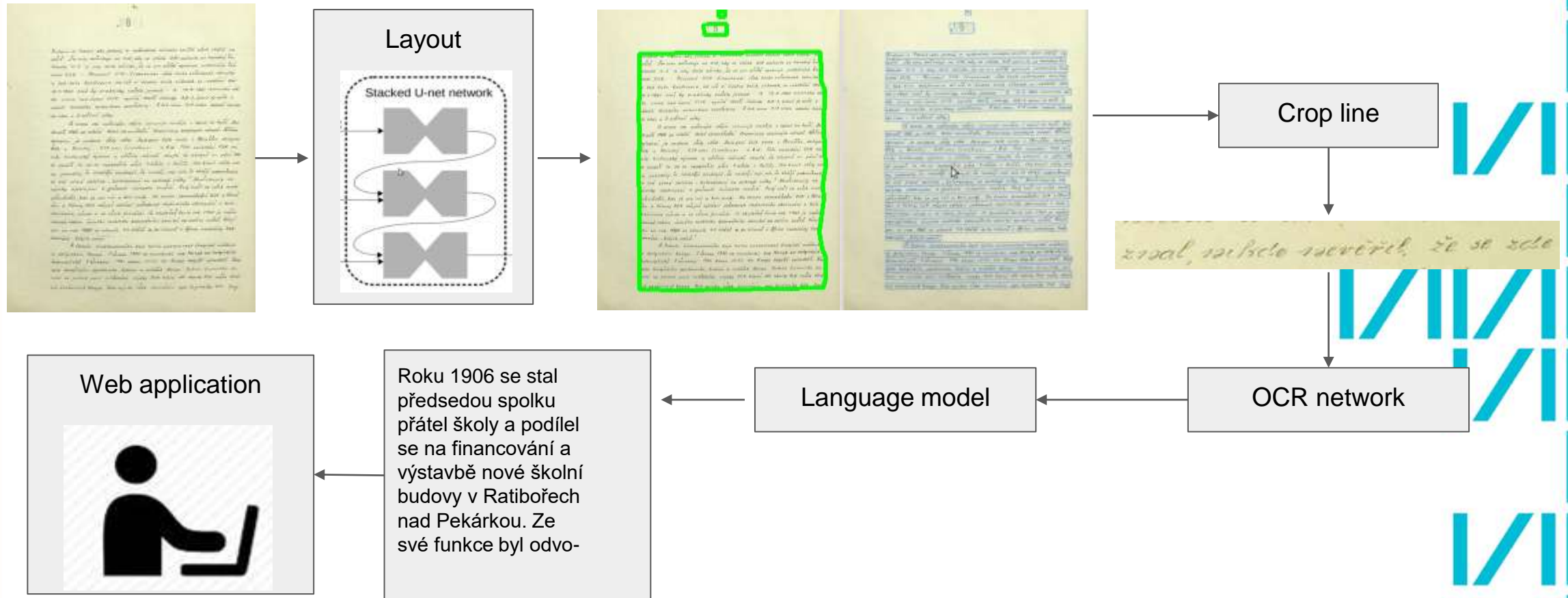
The screenshot shows a digital library interface. At the top left, there is a logo and the text "Digitální studovna Ministerstva obrany ČR". A search bar contains the text "Hledat v celé digitální knihovně". On the right, there are navigation links: "Sbírký", "Procházet", "Informace", and "English". Below the search bar, there are icons for print, download, and text. A search bar for the document says "Hledat v dokumentu". Below that, it says "9 z 122 stránek". A thumbnail gallery shows six pages, with the third page highlighted. The main view shows a handwritten page with the text: "sme si našli byt na půdě malého domku a dva dny sme čekali co se bude dít". Below the text is a drawing of a house with a chimney. The page number "3" is written in the top right corner. On the right side, there is a metadata panel with the following information: "Muj Deňik od r 19 30/7 14", "Nakladatelské údaje [S. I.], 1914-1915", "Typ dokumentu Kniha", and "Klíčová slova Rakousko-uherská armáda 1914-1918, Železniční vojsko 1914-1918, Polní železnice 1914-1918".

sme si našli byt na půdě malého domku a dva dny sme čekali co se bude dít!

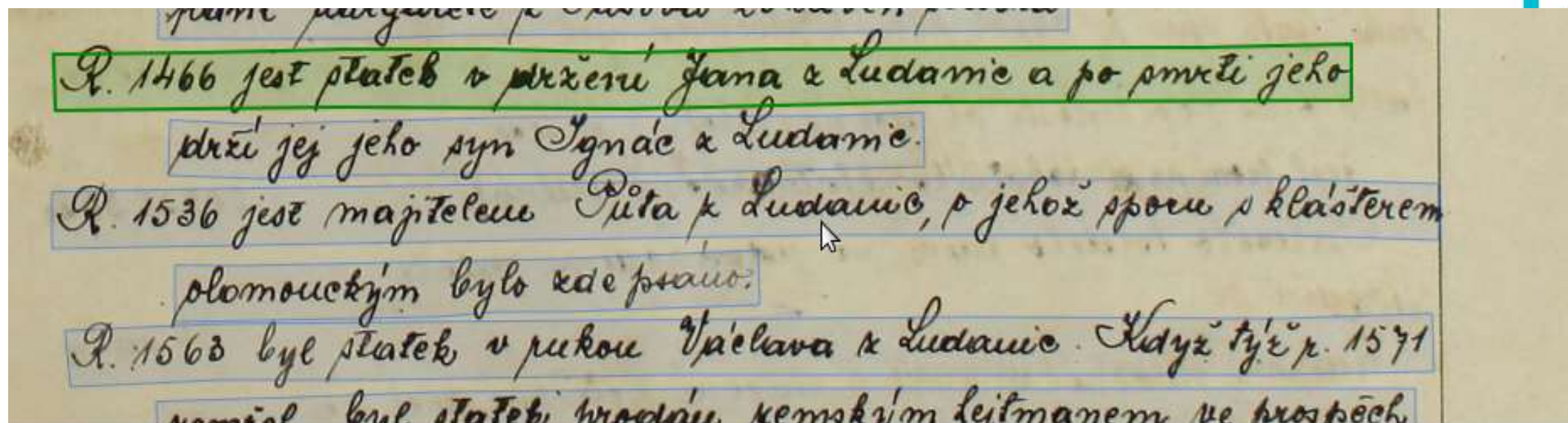
Muj Deňik od r 19 30/7 14. [S. I.], 1914-1915, s. 3. Dostupné také z: <http://www.digitalniknihovna.cz/dsmo/uuid/uuid:32f7ba5e-a323-11ea-95c0-001b63bd97ba>

# OCR - projekt PERO

## Řetězec procesu rozpoznání textu



## České kroniky 20. století



R. 1466 jest statek v držení Jana z Ludanic a po smrti jeho  
drží jej jeho syn Ignác z Ludanic.

R. 1536 jest majitelem Půta z Ludanic, o jehož sporu s klášteřem  
olomouckým bylo zde psáno.

R. 1563 byl statek v rukou Václava z Ludanic. Když týž r. 1571



## Periodika na mikrofilmech

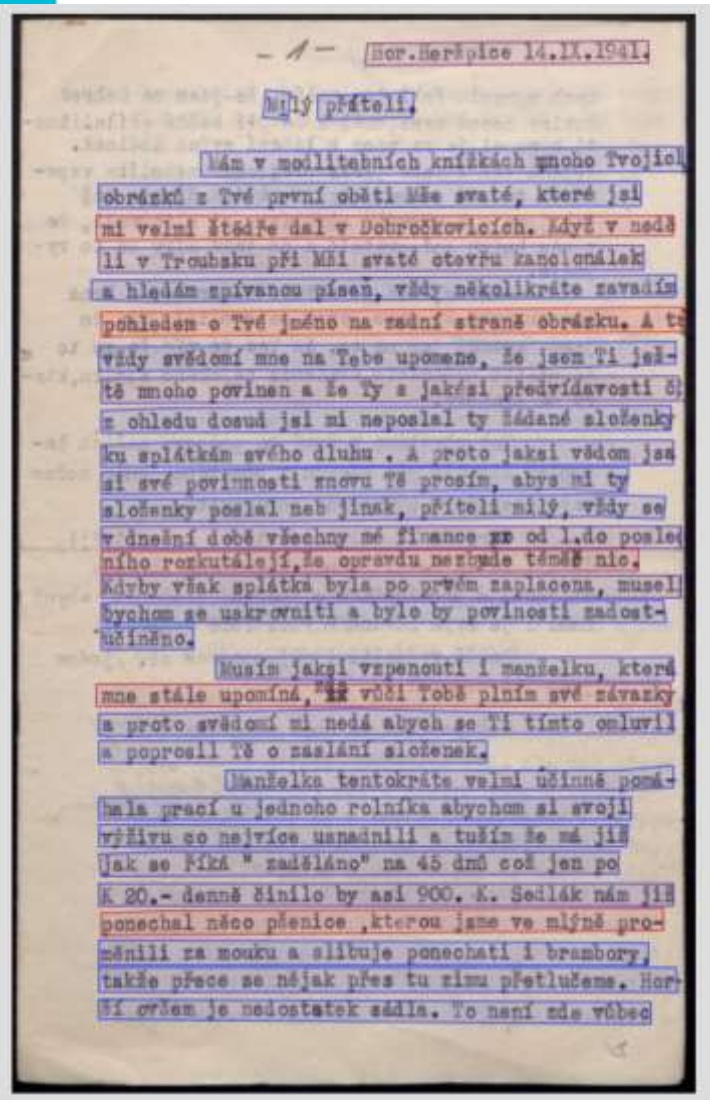
Socialisté italské mají za sebou veliký úspěch: vymožení amnestie pro odsouzené z posledních bouří. Co se nakřičely reakcionářské listy, že socialisté svou hlučnou agitací pro amnestii odsouzeným vlastně škodí, poněvadž prý vláda nemůže ustoupiti nátlaku z ulice -- vláda však daly socialistům za pravdu. Mocné hnutí v celé zemi přinutilo vládu k silným ústupkům. Nejsou sice amnestováni všichni političtí provinilci, ale téměř 3000 obětem vojenských i civilních soudů po milánských bouřích, kteří ni-

nakřičely

zence z posledních bouří. Co se nakřičely reakcionářské listy, že socialisté svou hluč-

reakcionářské

# Strojopis

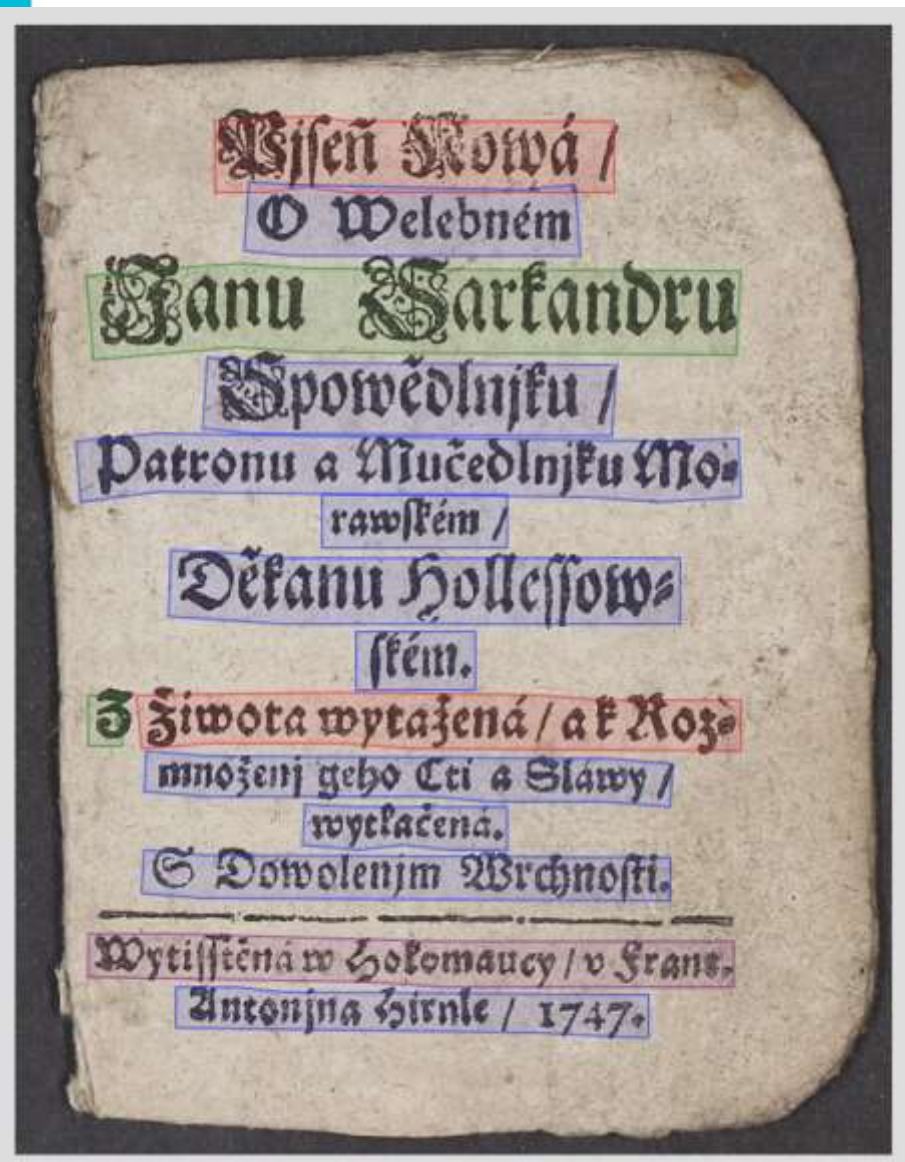


v dnešní době všechny mé finance od 1. do posledního rozkutálejí, že opravdu nezbude téměř nic.

Kdyby však splátka byla po prvním zaplacená, museli

finance xx od 1. do posledního rozkutálejí, že opravdu nezbude téměř nic.  
Kdyby však splátka byla po prvním zaplacená, museli

# Kramářské tisky



Pjšeň **N**owá/  
O Welebném  
Janu Sarkandru  
Spowědljku/  
Patronu a Mučedlnjku Mo=  
rawském/  
Děkanu Holleffow=  
lkém.

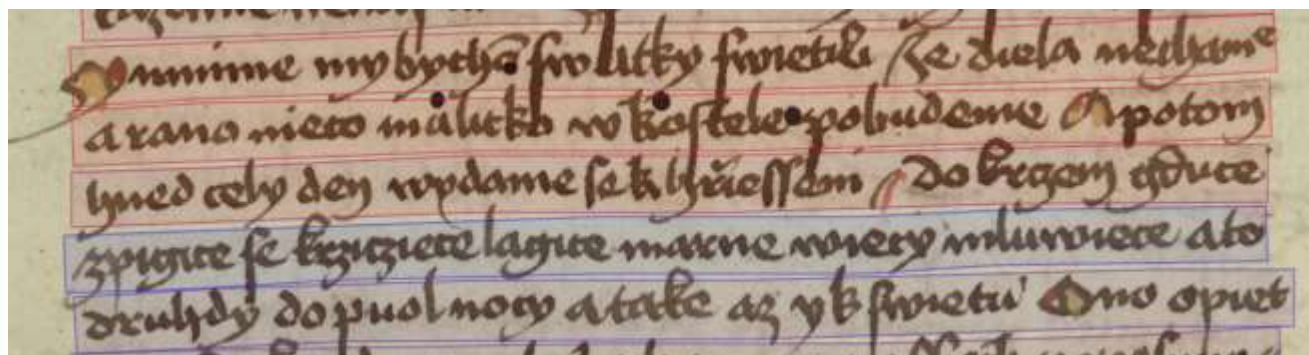
**Ž**iwota wytažená/ a k Roz=  
množenj geho Cti a Sláwy/  
wytlačena.

S Dowolenjm Wrchnosti.

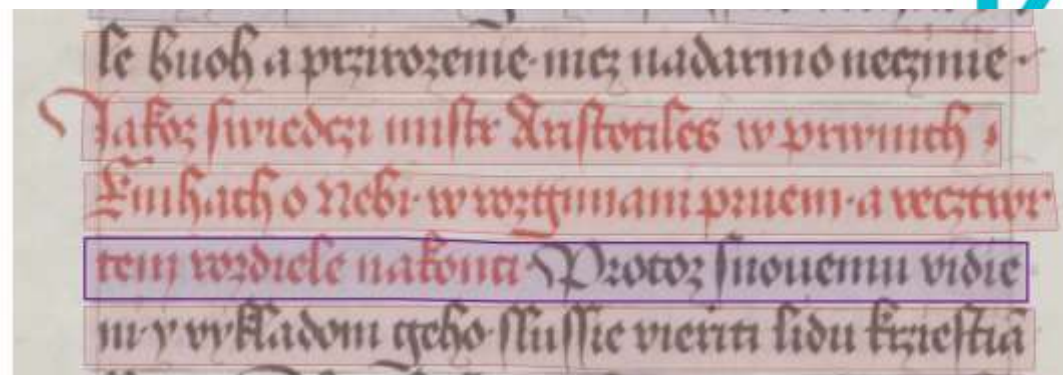
Z

Wytisštěná w Hoľomaucy/ v Frane.  
Antonjna Hirnle/ 1747.

# Jan Hus, Vavřinec z Březové



mnime mybycha swatky swietili se diela nedani  
a rano nieto malitko w kostee pobudeme A potom  
hned cely den wydame sek hriešsem Do krczem gduce  
zpigice se krziciece agite marne wiery mluwiece a to  
druhdy do puo nocy a take az ys swietu Ono opiet



le buoh a przirozenie nez nadarmo necziie  
Jakoz swiedczy mistr Aristociles w prwnich  
Eiihach o Nebi w rozgiani pruem a vecztyr  
tem rozdziele nakonu Protoz Inouemu vidie  
ni y vykladom geho sluffie wieriti lidu krzieftia

# České kroniky 20. století

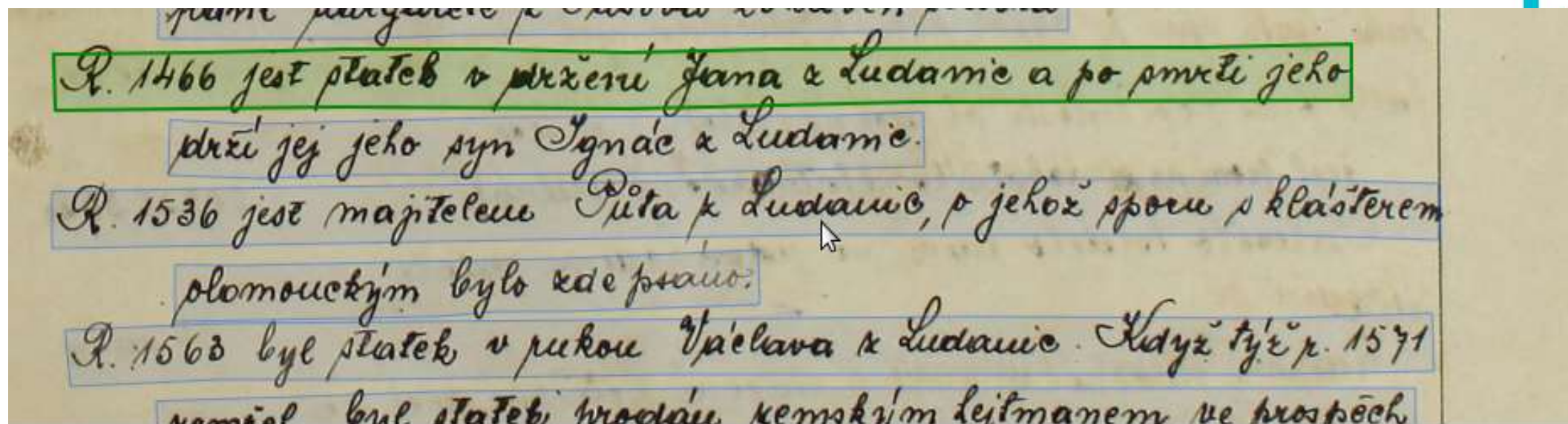
Moravoleň v horním závodě uvolnil této škole 2 nyní nepoužívané místnosti, které budou upraveny na dílnu pro práci s dřevem a s kovy.

ani projednávání komplexního plánu rozvoje školství v 3. pětiletce za účasti zástupců MNV a místních závodů však nevyřešilo ožehavý problém, to je zavedení 2 směnného vyučování. V příštích letech přibude na škole asi 100 dětí t. j. 3 třídy, pro které není učeben. Příslušné komise ONV a

Moravoleň v horním závodě uvolnil této škole 2 nyní nepoužívané místnosti, které budou upraveny na dílnu pro práci s dřevem a s kovy

ani projednávání komplexního plánu rozvoje školství v 3. pětiletce za účasti zástupců MNV a místních závodů však nevyřešilo ožehavý problém, to je zavedení 2 směnného vyučování. V příštích letech přibude na škole asi 100 dětí t. j. 3 třídy, pro které není učeben. Příslušné komise ONV a

## České kroniky 20. století



R. 1466 jest statek v držení Jana z Ludanic a po smrti jeho drží jej jeho syn Ignác z Ludanic.

R. 1536 jest majitelem Půta z Ludanic, o jehož sporu s klášterem olomouckým bylo zde psáno.

R. 1563 byl statek v rukou Václava z Ludanic. Když týž r. 1571

# Rukopisy

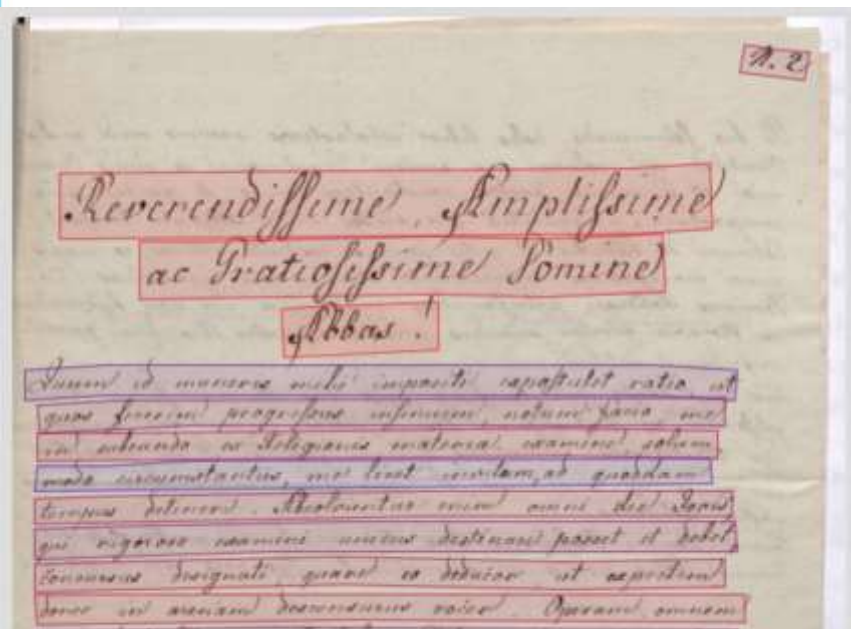
195  
hudbného a slavného jako Innocenc III., jenž řídil osudy téměř celé Evropy. Přední jeho starostí bylo způsobiti novou výpravu křížovou na osvobození Palaestiny; doba byla příhodná, neboť Saladin právě byl zemřel.  
Proto rozeslal posly po veškerém světě křesťanském, by hlásali kříž a vybírali peníze na novou výpravu. Ale poměry v Evropě nebyly utěšené, neboť

hudbného a slavného jako Innocenc III., jenž řídil osudy téměř celé Evropy. Přední jeho starostí bylo způsobiti novou výpravu křížovou na osvobození Palaestiny,

doba byla příhodná, neboť Saladin právě byl zemřel.

Proto rozeslal posly po veškerém křesťanském, by hlásali kříž a vybírali peníze na novou výpravu. Ale poměry v Evropě nebyly utěšené, neboť

# Rukopisy

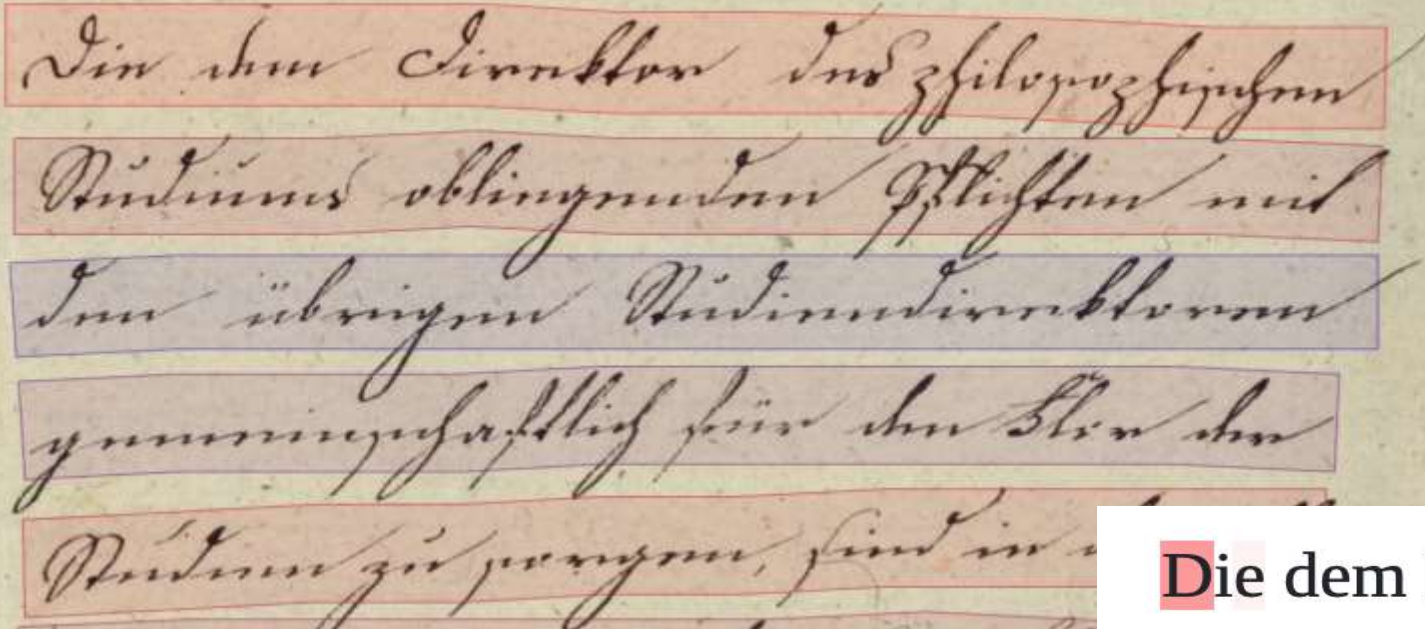


impendit Dominus Professor Löcher ne mora amplius, concursibus ex Historia Naturali, Diplomatica, et Numismatica absolutis, trahatur. Interea omne a studiis currentibus tempus reliquum materia revolvenda concesso. Quod studia concernit currentia, nactus sum duces ingenio

impendit Dominus Professor Löcher ne mora amplius, concursibus ex Historia Naturali, Diplomatica, et Numismatica absolutis, trahatur. Interea omne a studiis currentibus tempus reliquum materia revolvenda concesso. Quod studia concernit currentia, nactus sum duces ingenio



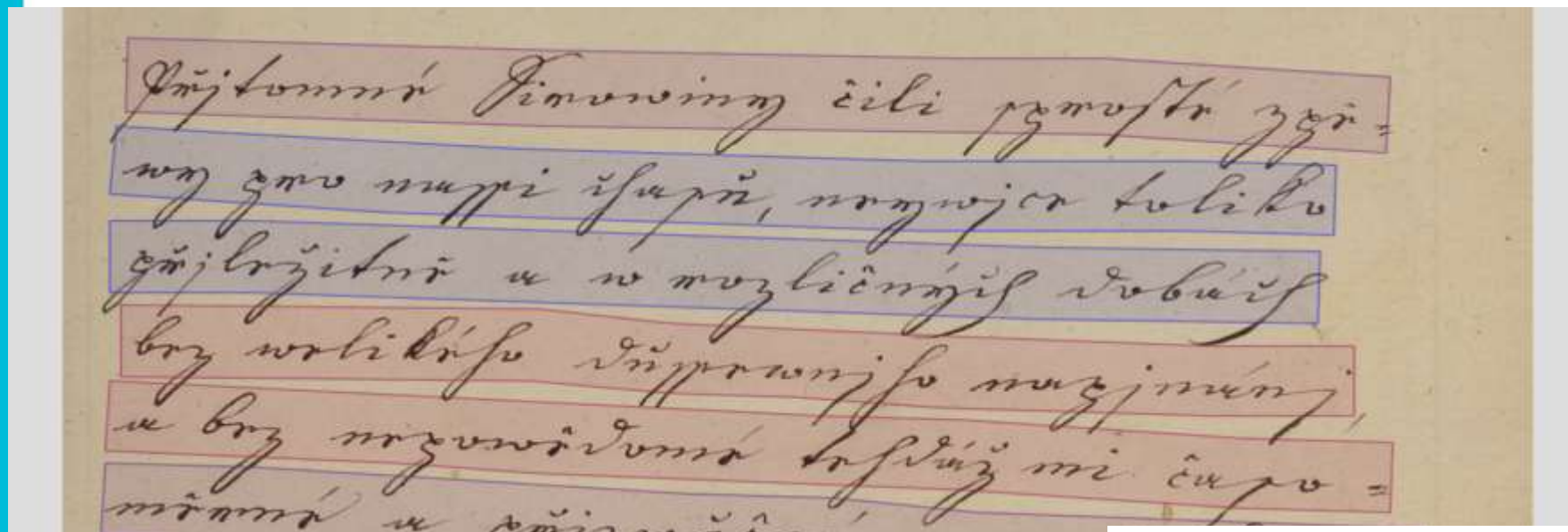
# Kurent



Die dem Direktor des philosophischen  
Studiums obliegenden Pflichten mit  
den übrigen Studiendirektoren  
gemeinschaftlich für den Fler der  
Studien zu sorgen, sind in der all,

Die dem Direktor, des philosophischen  
Studiums obliegenden Pflichten mit  
den übrigen Studiendirektoren  
gemeinschaftlich für den Fler der  
Studien zu sorgen, sind in der all,

# Kurent



Přjtomné Sirowiny čili sprofté zpě,,  
wy pro naši chasu, neywjce toliko  
přjležitně a w rozličných dobách  
bez welikého duffewnjho napjmáej  
a bez nepowědomé tehďáž mi čapo=

# PERO - důležité odkazy

- Jádru OCR - pero-ocr python balíček <https://github.com/DCGM/pero-ocr>
- Webová aplikace pro kontrolu a opravy - pero\_ocr\_web
  - Běží na <https://pero-ocr.fit.vutbr.cz>
  - Zdrojové kódy [https://github.com/DCGM/pero\\_ocr\\_web](https://github.com/DCGM/pero_ocr_web)
- OCR API pro hromadné zpracování
  - <https://pero-ocr.fit.vutbr.cz/api>
- Informace o projektu - <https://pero.fit.vutbr.cz/>



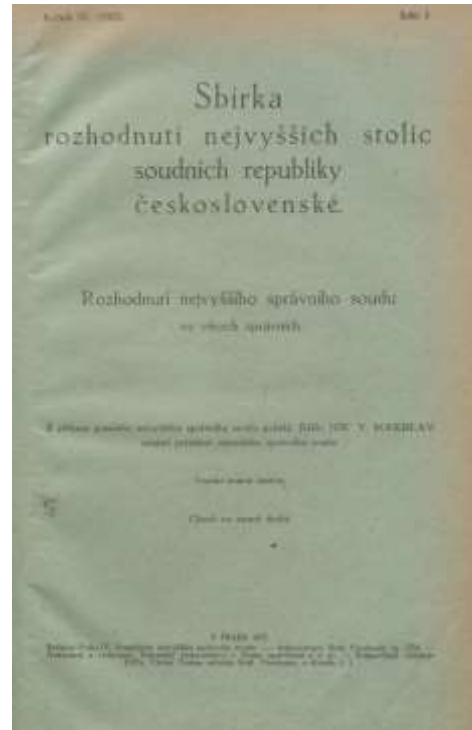
# Identifikace obrázků na stránce

- Cílem je automaticky detekovat pozici obrázků v naskenovaných dokumentech
- Proč je to těžké?



# Co není text ještě nemusí být obrázek

- Grafické elementy, artefakty, pozadí



# Co obsahuje text stále může být obrázek

- text v obrázcích, překryv

An advertisement for Oriflame Beauty mascara. It features a large image of the mascara wand and tube. The text is in Czech. At the bottom left, there is a small image of a Nivea product.

Pokud vám rodinný rozpočet právě teď dovozuje koupit si jen jedno zkrášlovadlo, vyberte si voděodolnou řasenku. Nalíčené řasy dodají tváři výraz a upravený vzhled a je to otázka dvou minut, které se dají ukrást i v tom nejhorsím rameni skluzu. Složení, odolávající dešti, vám dodá jistotu, že se ani po srážce s podzimní plískanicí neproměníte ve smutnou panda.

**ZKUSTE:**  
Voděodolnou objemovou řasenku Oriflame Beauty (199 Kč) se silikonovými složkami, které nabídnou voděodolávající vlastnosti a prodlouží trvanlivost nalíčení.

á  
stice

vadlo.  
vsta: vo-  
ný make-up  
že. Díky  
nému  
adlu Nivea  
(104 Kč)  
úkladně,  
ně.

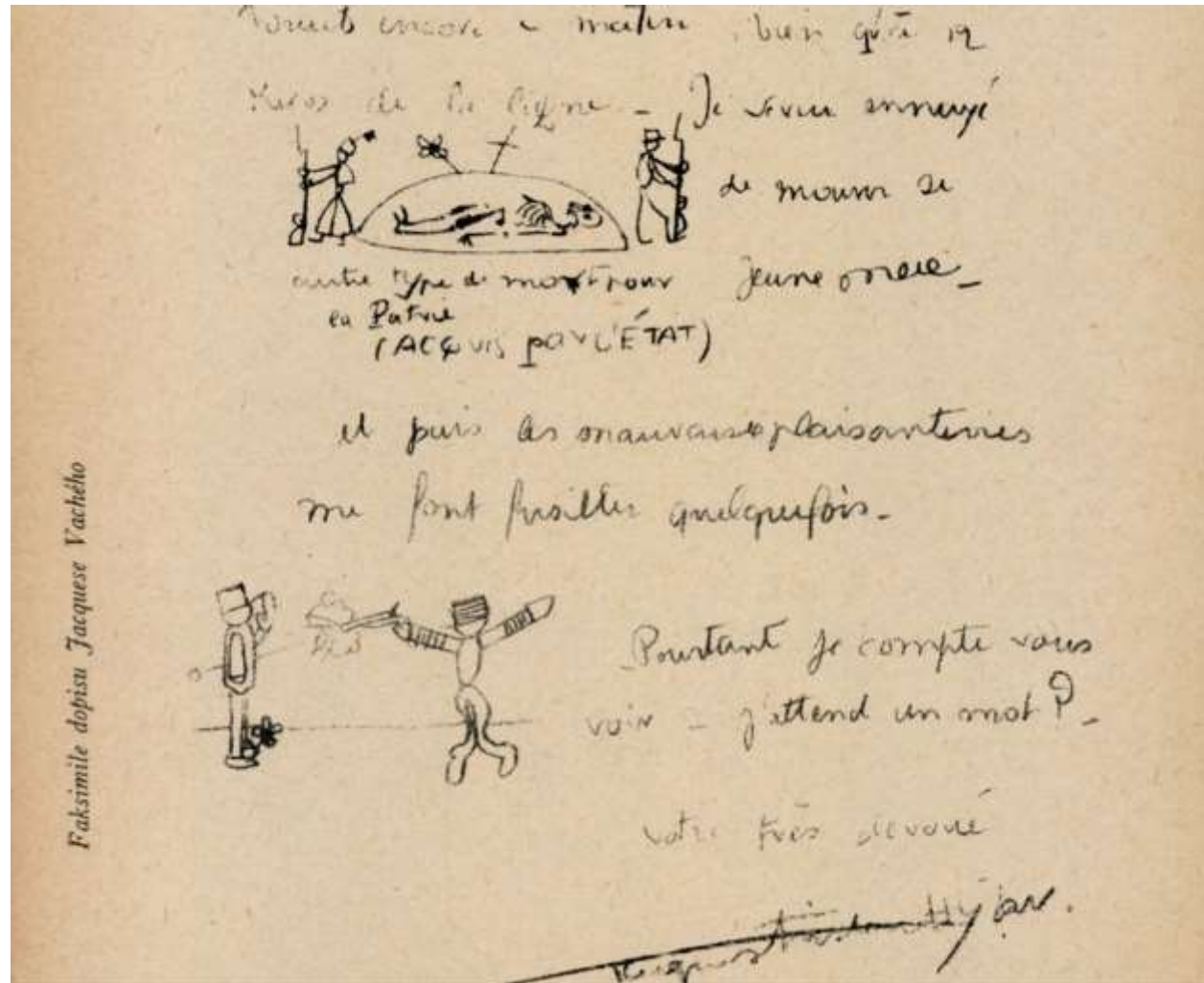
Oriflame  
BEAUTY  
WATERPROOF  
Beauty

(len) / Foto archiv linim

# Obrázky nemusí být ohraňčeny



# Obrázek a text mohou být vizuálně podobné







# Dostupná řešení - OCR od ABBYY (v ALTO)

- Dataset: 500 náhodných, manuálně oannotovaných stránek
- Výsledek:
  - Celkové IOU\*: 0.69
  - Sensitivity\*\*: 0.69
  - Precision\*\*: 0.36
  - 18% obrázků není detekováno vůbec
  - 57% detekovaných obrázků jsou falešná pozitiva\*\*\*

\* Intersection over union

\*\*Vypočteno pro práh citlivosti IOU=0.5

\*\*\*Neobsahují obrázek, nebo obsahují obrázek který už byl detekován, tzn. obrázky se překrývají.



OCR od ABBYY    Ruční anotace

# Detekce pomocí vlastního modelu strojového učení

Řešení: Segmentace pomocí konvoluční neuronové sítě

## Proč konvoluční sítě?

- Translační invariance: poloha obrázku na stránce není důležitá
- Množství volně dostupných implementací state of the art modelů (napr. ResNet, AlexNet, VGG,...)
- Možnost použít váhy předtrénované na velkém datasetu.



# Slepé uličky

- **Newspaper Navigator model**
  - Vyvinut v Library of Congress na detekci obrázků, nadpisů a dalších objektů ve starých novinách.
  - Založen na síti Faster-RCNN od Facebooku.
  - Při evaluaci (bez našeho trénování) měl mnohem horší výkon než OCR od ABBYY.
  - Domníváme se, že model je náchylný k overfittingu a špatně generalizuje na náš dataset.
- **Natrénování UNet sítě (bez předtrénovaných vah)**
  - Architektura využívaná k segmentaci v medicíně a při zpracování satelitních snímků.
  - Nepodařilo se nám dosáhnout dostatečné přesnosti, náš dataset byl zřejmě příliš malý na to, aby se model naučil všechny potřebné vizuální znaky (features).



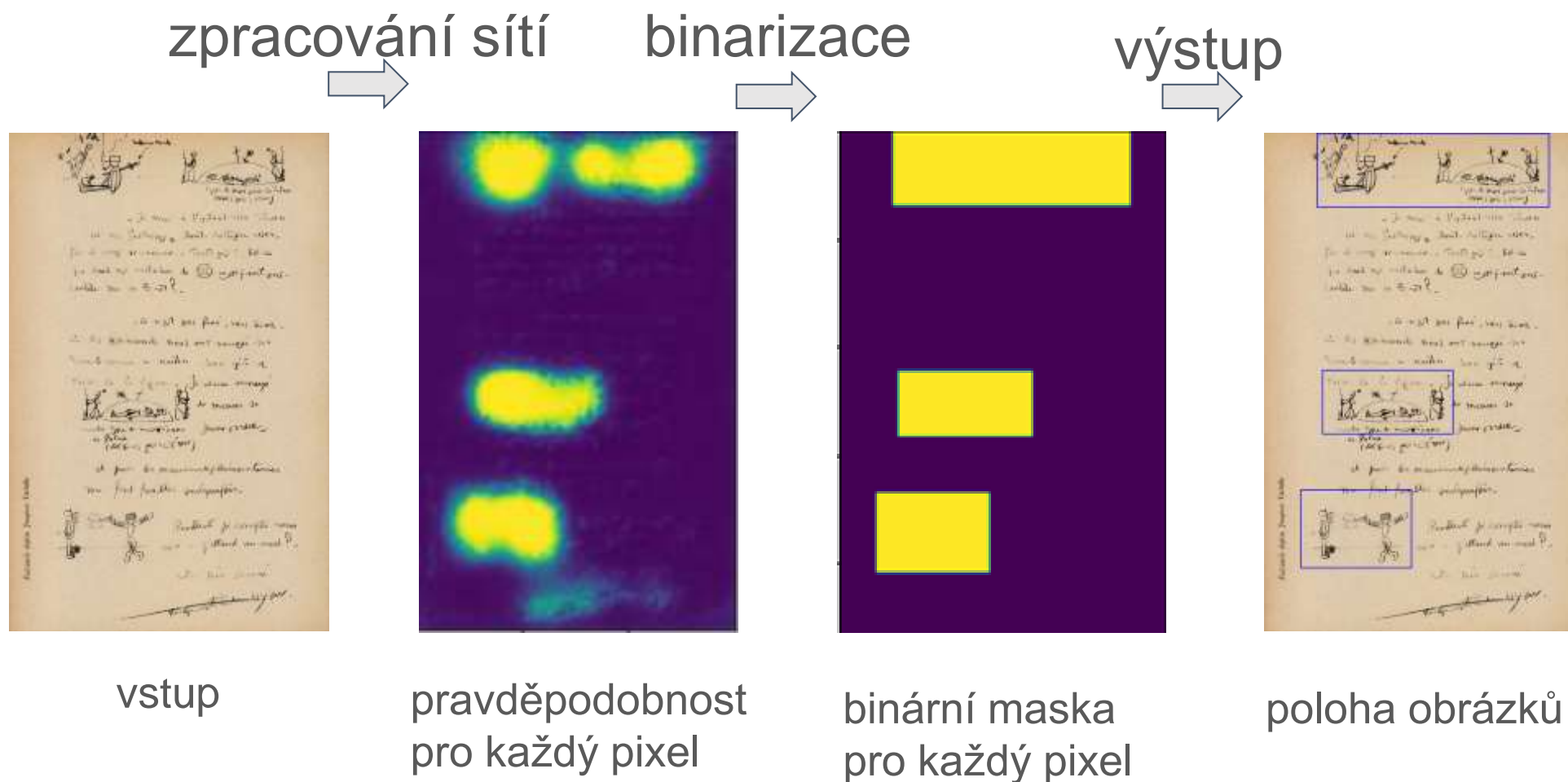
# Neuronová síť dhSegment

- Vyvinuta v 2018 specificky k zpracování skenovaných historických dokumentů
- Použita k detekci ornamentů a fotek na naskenovaných stránkách
- Inspirována segmentační sítí UNet
- Možnost využít předtrénované váhy z ResNet-50
- Implementována v Pythonu pomocí Tensorflow



# Neuronová síť dhSegment

- Natrénována na ~1000 manuálně anotovaných stránkách na notebooku bez GPU



# Neuronová síť dhSegment

- Výsledek:
  - Celkové IOU: 0.65
  - Senzitivity\*: 0.65
  - Precission\*: 0.4
  - 24% obrázků není detekováno vůbec
  - 53% detekovaných obrázků jsou falešná pozitiva\*\*
- Rychlost (notebook bez GPU):
  - 2.87s na stránku
  - 33 dní na milion stránek
- **Při vynaložení relativně malého množství práce je kvalita prakticky identická s OCR od ABBYY**

\*Vypočteno pro práh citlivosti IOU=0.5

\*\*Neobsahují obrázek, nebo obsahují obrázek který už byl detekován



# Odkazy

- dhSegment
  - Github <https://github.com/DCGM/pero-ocr>
  - Publikace: Oliveira, Sofia Ares, Benoit Seguin, and Frederic Kaplan. "dhSegment: A generic deep-learning approach for document segmentation." *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018, preprint at <https://arxiv.org/abs/1804.10371>
- NewspaperNavigator projekt: <https://github.com/LibraryOfCongress/newspaper-navigator>
- OCR od ABBYY: <https://www.abbyy.com/finereader-server/>
- Resnet: <https://arxiv.org/pdf/1512.03385.pdf>
- UNet: <https://arxiv.org/pdf/1505.04597.pdf>





# Vyhledávání obrázků podle podobnosti - VISE

- Vyhledávání příbuzných obrázků podle prostorové podobnosti (výřez)
- Periodika, staré ilustrace, grafiky, loga...
- Přesný i pro malé rozlišení (testováno na 1024x1024)
- Možnost kombinace daty identifikujícími obrázky na stránce
- 13 000 obrázků zaindexováno za 7 hodin



Datová sada obrázků (jpg, png)

Indexace / trénování  
vizuálních deskriptorů..



Zaindexovaná datová sada

Search ready





The image shows a page from the newspaper 'Svobodné slovo'. At the top, there are logos for 'ARMIN' and 'HYDRO'. Below them is an advertisement for 'BOH. KREJČÍK' with the headline 'Romány z českých dějin od Jos. Svátka'. To the right of this is an advertisement for 'Svatovítského pojistovacího ústavu'. Further right is a small circular advertisement for 'KURÁŽNÍ - DĚŽNÉ - POKRÁČ'. Below the 'ARMIN' advertisement is a large advertisement for 'Americké PULTY' by 'Jos. Jiroušek', featuring an illustration of a desk. To the right of the 'PULTY' advertisement is another advertisement for 'Kupujte - naše dámy tak rády'. At the bottom left, there are advertisements for 'Dr. Frant. Bažkovský' and 'Výběhy prádla in needstý'. At the bottom right, there is an advertisement for 'Kupujte - naše dámy tak rády' with an illustration of a woman. The page is filled with text and small illustrations, typical of a newspaper advertisement page.

Hledat

# Porovnání obrázků inzerce deníku Svobodné slovo

## ARMIN

Je to nejlepší a nejlevnější domácí výrobci továrna mydel

### Jan Hubínek a syn, Praha-Vlt. Libeň.

Krásné ARMIN se praní nebýt a umývá se jeho náklonem.  
V dobrotě prodává kus 6 krejčovů.

**Pánům občasným cestníkům a zvláště zdarma vyřazená.**

## BOH. KREJCÍK,

velkoobchodní knihárna, v Praze - Vinohrady, Náměstí Republiky č. 23

## Romány z českých dějin od Jos. Svátka

romány a povídky. Třetího dílu. 24 K 30 h. ak. váz. 6 K 30 h. — Praha a Rim. Roman ze století XVI a K 40 h. ak. váz. 6 K 30 h. — Svatava. Období. Roman z dějiny Marie Terézie. 24 K 30 h. ak. váz. 1 K 50 h. — Pásmo v Praze. Roman ze století XVII. 24 K 30 h. ak. váz. 1 K 70 h. — Bílá Blahoborská. Roman ze století XVII. 24 K 30 h. ak. váz. 1 K 30 h. — Paměti katovské rodiny Mýdla v Praze. 24 K 30 h. ak. váz. 1 K 30 h. — Pád rodu Smolického. 24 K 30 h. ak. váz. 1 K 30 h. — Právě na cestě. 24 K 30 h. ak. váz. 1 K 30 h. — Povídky z dějin. 24 K 30 h. ak. váz. 1 K 30 h. — Seznam v Praze r. 1620. 24 K 30 h. ak. váz. 1 K 30 h. — Svědkové v Praze r. 1642. 24 K 30 h. ak. váz. 1 K 30 h. — Počátky Budova. 24 K 30 h. ak. váz. 1 K 30 h. — Zelená kouzla. 24 K 30 h. ak. váz. 1 K 30 h. — Vápen v Křivoklátě. 24 K 30 h. ak. váz. 1 K 30 h. — Astrolog. 24 K 30 h. ak. váz. 1 K 30 h.

## Vazby skvostné

pro potřeby do světa V. Olivy. — Každá vazba se odobrá také pro v učitelích po 40 hal.

**Prof. J. VODÁK**  
napsal v „Čas“ č. 16. srp. 1906 o románech Svátkových, že mají vlastnost, která je stala nad dějepisnou belletrii V. Henze Trěhákého.

**Prof. FR. SEKANINA**  
napsal v „Národním listu“ 6. března 1906 o Svátkově, že jeho „Mládí“ svým způsobem jest nejlepším ukázkou toho, jak se v Praze v mnoha dnech románech děje, jak se v Praze děje. Typ. Charolinská ul. č. 11. Vydavatel: Jaroslav Pránský.

## Dr. Frant. Bačkovský,

lékař v Praze, Ústí u nář. 56.

## Výbavy prádla pro nevěsty.

Krásná, světlá, jednoduchá, vhodná i pro svatební obřady a svatební hostiny. Každá sada obsahuje: 12 kusů prádla, 12 kusů rukavic, 12 kusů ponožek, 12 kusů šatek, 12 kusů kalhot, 12 kusů košilek, 12 kusů košilových kalhot, 12 kusů košilových kalhot, 12 kusů košilových kalhot.

**J. NOVÁK, Vodňanská ulice č. 34.**  
Zašle cenu.

## Dr. Frant. Bačkovský,

lékař v Praze, Ústí u nář. 56.

## Výbavy prádla pro nevěsty.

Krásná, světlá, jednoduchá, vhodná i pro svatební obřady a svatební hostiny. Každá sada obsahuje: 12 kusů prádla, 12 kusů rukavic, 12 kusů ponožek, 12 kusů šatek, 12 kusů kalhot, 12 kusů košilek, 12 kusů košilových kalhot, 12 kusů košilových kalhot.

**J. NOVÁK, Vodňanská ulice č. 34.**  
Zašle cenu.

## Bratři! První výrobní družstvo dělnictva krejčovského

činnost je se zabývá k získání lepšího blahobytu dělnictva v Praze, zejména v Praze. Každý výrobce modních látek. Obilná vzrůst. Rozhodně provedení. Za vlastní práce.

## Jarní mody

# BAZAR

Obilná vzrůst. Rozhodně provedení. Za vlastní práce.

## Malý seznamovatel.

Chmelovo maso	20	Jehněčí maso	30	Přes 10 000 pírní	20
Chmelovo maso	20	Jehněčí maso	30	Přes 10 000 pírní	20
Chmelovo maso	20	Jehněčí maso	30	Přes 10 000 pírní	20

## Veškeré - Giskové - Potřeby

Kuřičská ulice. — soc. dělnictva

## Společně hospodyňky

americké PULTY

## Hynek Gotwald

v Praze, Na příkopě 2.

Továrna na železný a mosazný nábytek a zboží látkové doporučené.

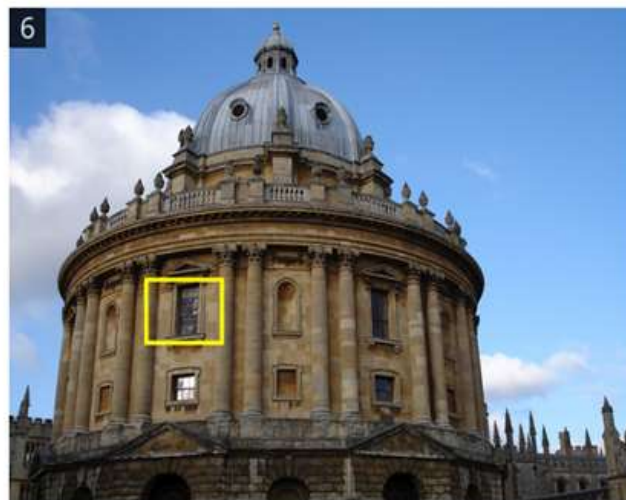
## AMERICKÉ PULTY

americké PULTY

## AMERICKÉ PULTY

americké PULTY

# Funkční i pro 3D prostorové obrázky



# Výhody VISE

- Velmi dobrá přesnost i pro obrázky s malým rozlišením
- Vyhledávání je rychlé
- Není příliš mnoho konkurenčních systémů
- Není potřeba grafické karty
- Systém je dále rozvíjen

# Nevýhody

- Nefunguje pro běžné bloky textů, vhodné spíše pro obrázky, ilustrace nebo větší texty jako tituly, nadpisy atd.
- Může nastat nepřesnost ve vyhledávání, např. pokud je výřez málo detailní nebo je špatná trénovací sada
- Nelze použít obrázky s vysokým rozlišením
- Nelze přidávat nové obrázky již k natrénované datové sadě
- Nelze použít JPEG2000



# Odkazy

- Abhishek Dutta, Relja Arandjelović, and Andrew Zisserman. 2021. VGG Image Search Engine. from <https://www.robots.ox.ac.uk/~vgg/software/vise/>
- Gitlab: <https://gitlab.com/vgg/vise>
- Oxford Visual geometry group: <https://www.robots.ox.ac.uk/~vgg/>



# Co dál?

- Existuje řada zajímavých projektů, aplikací, modelů
- Není snadné najít hotové řešení
- Specifika reálných datových sad
  - rozsah, variabilita
- Velký prostor pro další rozvoj



# Děkuji za pozornost!

## Dotazy?

Petr Žabička, Ján Bogár, Michal Tran - Moravská zemská knihovna v Brně