Digitalizace minulosti s nástroji budoucnosti [1]

Autor:

- POLANSKÝ, Petr [2]

Číslo:

- 2021, ročník 30, číslo 1 [3]

Rubrika:

- Trendy v knihovnách [4]

Klíčová slova:

- digitalizace [5], umělá inteligence [6]

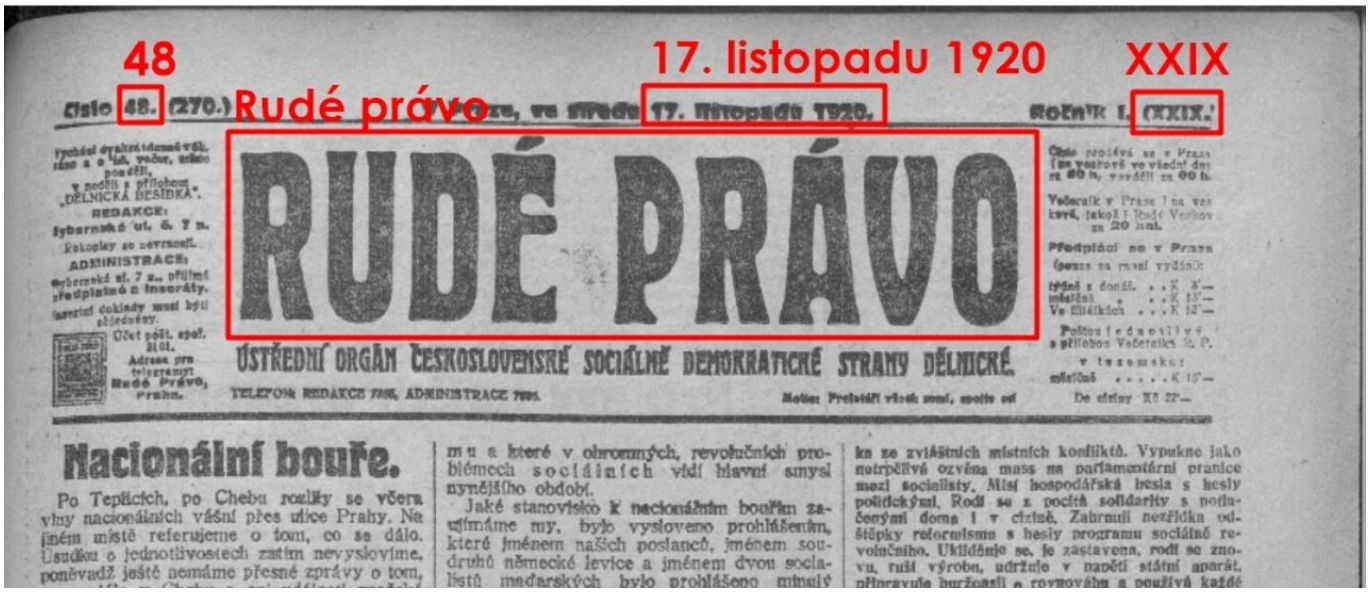
Digitalizace kulturního dědictví v posledních dvou dekádách zažívá nebývalý rozvoj. Vzniká velké množství iniciativ a programů, které mají za cíl digitalizovat historické dokumenty vedoucí ke zrodu velkého objemu digitálního obsahu. To je bezpochyby dobrá zpráva pro badatele a širokou veřejnost, jimž se zjednodušuje přístup k archiváliím, který je zejména v současné době obtížný.

Digitalizované dokumenty však v řadě případů nevyužívají veškerého potenciálu. Řeč je o detailní orientaci v logické struktuře dokumentu, která jde ruku v ruce s co nejpřesnějším vytěžením a následným vyhledáváním v textu dokumentu. Tyto možnosti, jsou-li dostupné, dramaticky urychlují vyhledávání v dokumentech a otevírají nové zdroje informací nejen pro zapálené badatele, ale také pro širokou veřejnost.

Pro digitalizaci je charakteristická nemalá pracnost spojená se získáváním úplných informací z dokumentů. V případě digitalizace tištěných knih je situace dobrá – logická struktura knih je poměrně přímočará a dobře čitelná. Podobně nástroje pro rozpoznávání tištěného textu (tzv. systémy OCR, optical character recognition) dosahují velmi dobrých výsledků a proces vytěžování textu je tak možné automatizovat.

Pokud ovšem budeme chtít zmapovat logickou strukturu u periodik, například novinových vydání, podmínky se výrazně komplikují (viz obr. 1). Je třeba vypořádat se s vícesloupcovým uspořádáním textu, textem proloženým různými obrázky a grafickými prvky a v neposlední řadě nejrůznějším dělením textu, kdy článek začíná na jedné straně a pokračuje na straně jiné. Zde je již potřeba porozumět nejen struktuře textu, ale také jeho významu, aby jednotlivé logické celky mohly být správně propojeny. Standardní nástroje OCR zde již částečně selhávají, protože nedokáží vrátit souvislý text článků v logickém sledu.

ISSN 1213-5828



[7]
Obr. 1: Příklad detekce textu z novinových článků (zdroj: Rudé právo, 17. 11. 1920, oddělení časopisů Knihovny Národního muzea, sign. Z 18 A 1 *[8], získáno 24. 3. 2021)*

Pro ručně psané texty je situace ještě komplikovanější, protože současné standardní nástroje pro rozpoznávání textu zpravidla poskytují velmi nepřesné a nespolehlivé výstupy. V praxi to znamená, že takové dokumenty je reálně nutné ručně přepisovat, což je ovšem časově a v konečném důsledku i finančně náročné.

Existuje cesta, jak tento nepříznivý stav změnit?

Umělá inteligence může zásadně změnit možnosti digitalizace

S dynamickým rozvojem oblasti strojového učení se pro digitalizaci dokumentů otevírají zcela nové možnosti. Řešení na bázi neuronových sítí je schopno vyhodnocovat dokumenty a jejich obsah na základě zkušenosti z předchozích zpracování, podobnosti a pravděpodobnosti shody se skutečným obsahem dokumentu. Umělá inteligence „dovozuje“, jaké slovo je napsáno rukopisem, kde na stránce je reklama, novinový článek nebo které části textu spolu logicky souvisí. Operátor v procesu digitalizace poté už nemusí provádět všechny činnosti ručně, ale může se soustředit na ty informace, které automatizované algoritmy identifikovaly s menší mírou jistoty.

Společnost EXON s.r.o. aktuálně pracuje na dvou produktech, které využívají metod umělé inteligence k zásadnímu zkvalitnění výstupů digitalizace historických dokumentů. Jedná se o řešení Kaitos [9] a InkCapture [10].

Kaitos – skutečné porozumění obsahu dokumentu

Řešení Kaitos dává paměťovým a dalším institucím do rukou nástroj, který zefektivní, zautomatizuje, zpřesní a „standardizuje“ podstatnou část dnes ručně či poloautomaticky prováděné práce. Kaitos pro vytěžování informací využívá neuronové sítě, díky kterým se zvýší nejen množství vytěžených informací, ale také spolehlivost jejich detekce.

Konkrétně má Kaitos za cíl automatizovat:

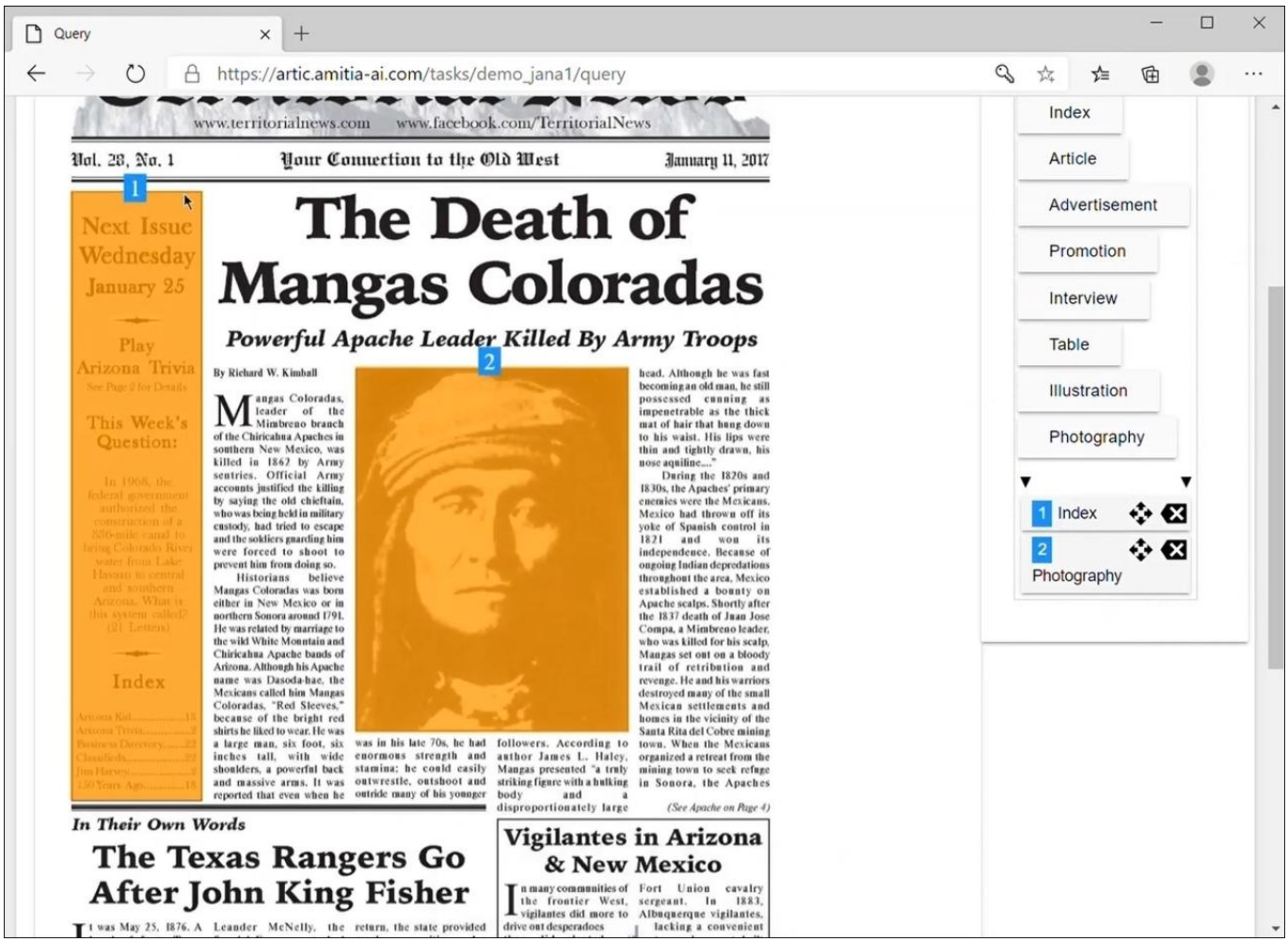
- předzpracování skenu dokumentu (detekce textu a dalších objektů, korekce nežádoucích vad obrazu ze skenování, rozdělení pravé a levé stránky, ořez a narovnání apod.);

ISSN 1213-5828

- zařazení digitalizovaných předloh do předem definovaných tříd (např. obsahová stránka knihy, strana knihy s ilustrací, titulní strana novin apod.);
- rozpoznání pozice logických bloků dokumentů (záhlaví, zápatí, grafický element v textu apod.);
- vytěžení textového obsahu a přiřazení textů k logickým blokům v dokumentu;
- uchování popisných metadat o digitalizovaném dokumentu pro rozšířené možnosti vyhledávání;
- uložení vytěžených informací do standardních formátů (generování balíčků PSP, angl. producer submission packages) pro následný export do digitální knihovny.

Na obr. 2 je znázorněn postup strojového učení. Předlohy je možné rozřadit do různých tříd, přesně identifikovat jednotlivé bloky a indexovat pro budoucí vyhledávání.

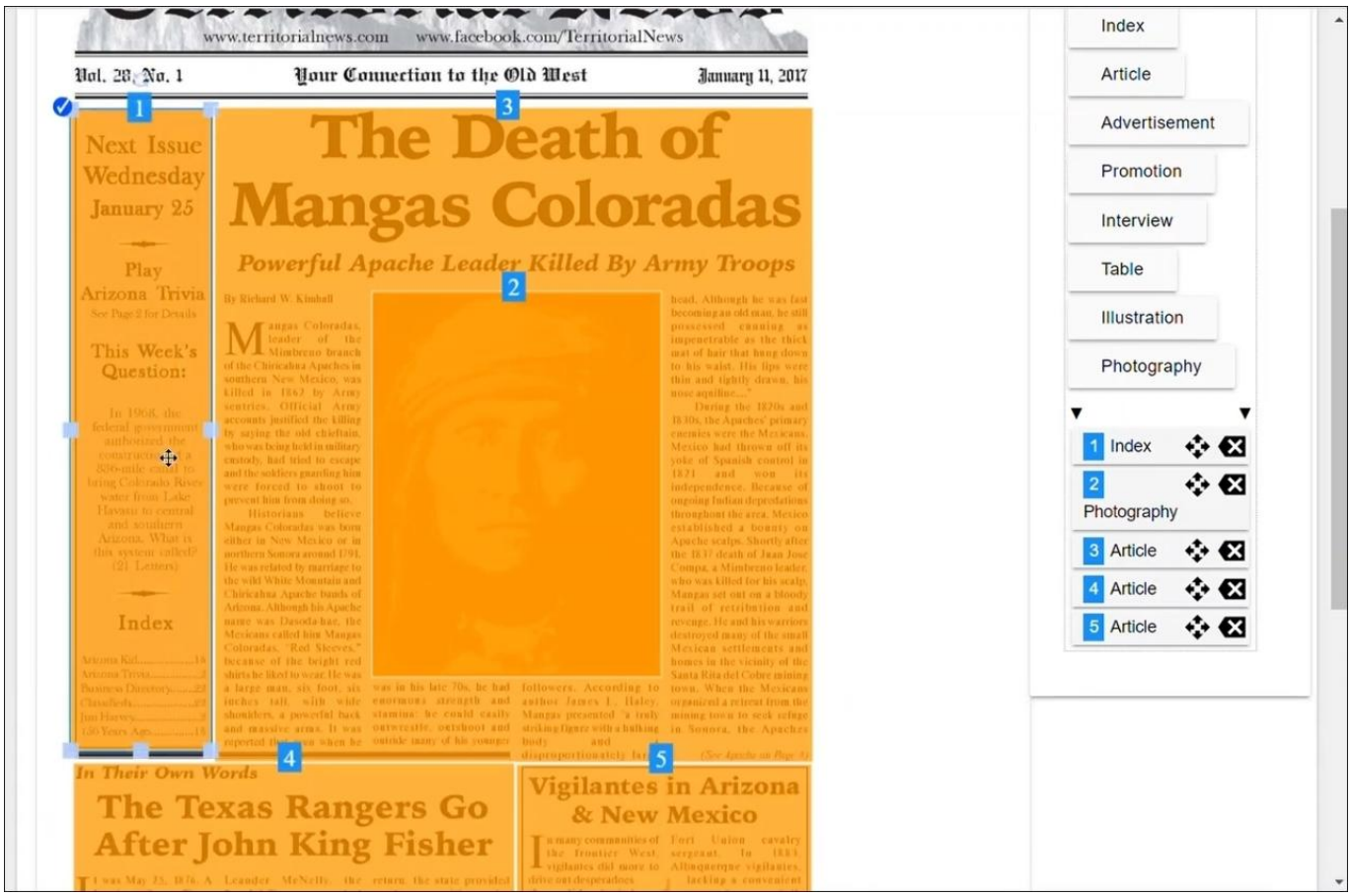
ISSN 1213-5828



[11]
Obr. 2: *Postup anotace dat I (zdroj:* MarkIt *[12]*| Amitia, video z kanálu zaměstnance *[12]*, získáno 25. 3. 2021)

Další postup je znázorněn na obr. 3.

ISSN 1213-5828



[13]
Obr. 3: Postup anotace dat II (zdroj: MarkIt | Amitia, video z kanálu zaměstnance *[12], získáno 25. 3. 2021)*

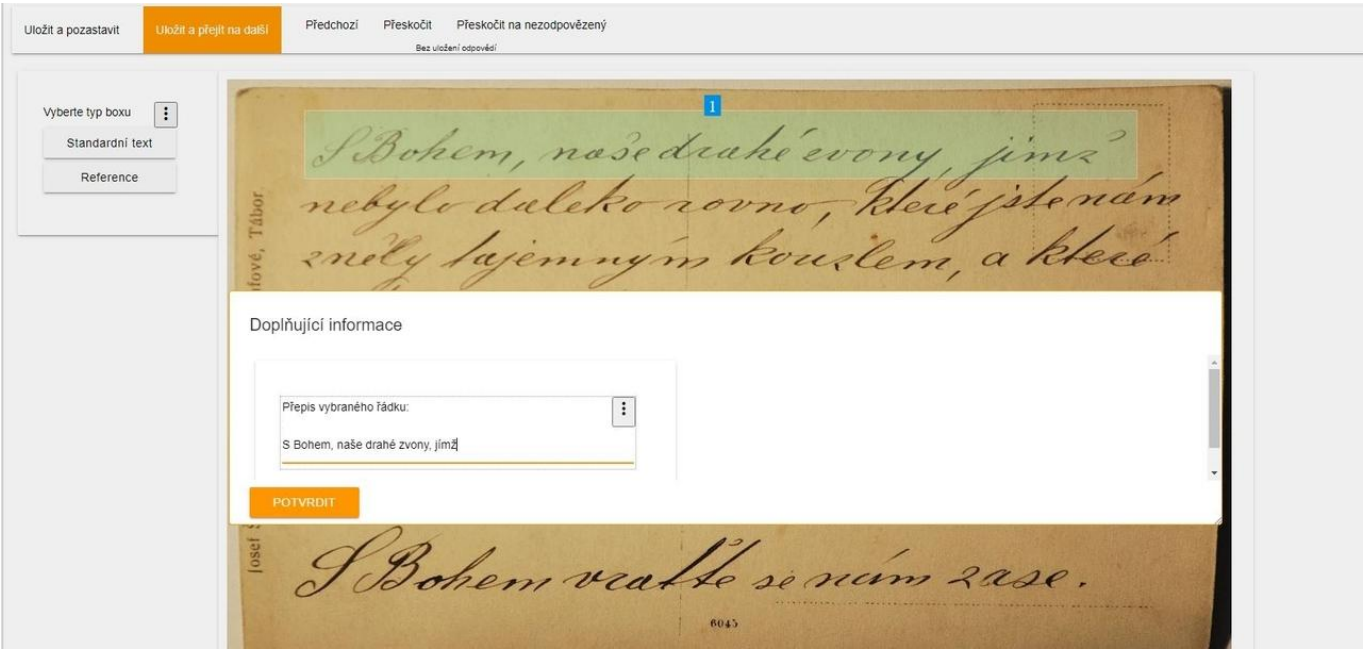
Ve videu, z něhož pocházejí obr. 2 a 3, je možné si celý proces prohlédnout:

InkCapture – rozpoznávání ručně psaného písma

Nástroj InkCapture je zaměřen na rozpoznávání ručně psaného písma a efektivní vyhledávání v ručně psaných textech. Mohlo by se zdát, že rozpoznávání písma je už dobře zvládnutá oblast – jsou přeci běžně dostupná zařízení, na která je možné psát a v reálném čase převádí psané písmo do textu, se kterým je možné dále pracovat. Tato zařízení mají k dispozici například informaci o tahu pera, rychlosti pohybu, tlaku jednotlivých tahů a další informace, které jsou pro rozpoznání textu velmi cenné.

Na obr. 4 je pro ukázku znázorněna identifikace textu a extrakce informací z rukopisu z roku 1917.

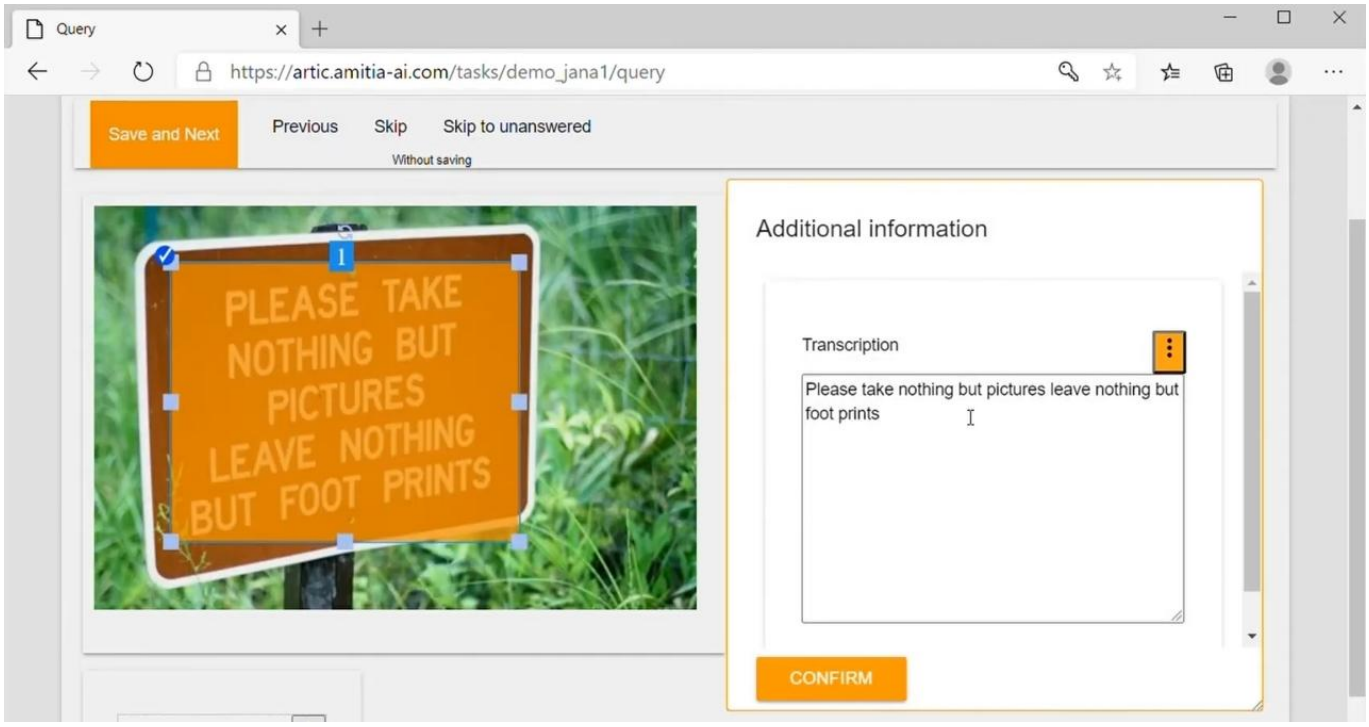
ISSN 1213-5828



[14]
Obr. 4: Anotace dat z ručně psaného písma (zdroj: Jan Sommer, Flickr [15], získáno 25. 3. 2021)

Na obr. 5 je znázorněna identifikace textu a extrakce informací z textu na dopravním značení.

ISSN 1213-5828



[16]
Obr. 5: Příklad anotace dat z textu dopravního značení (zdroj: MarkIt | Amitia AI [17], získáno 25. 3. 2021)

Dalším samostatným problémem je vyhledávání v analyzovaných dokumentech. Ať už hovoříme o detekci textů v tištěných knihách nebo v ručně psaných dokumentech, vždy bude detekce zatížena nějakou chybovostí. Při vyhledávání v dokumentech je pak zapotřebí jistá tolerance vyhledávacího algoritmu k chybám a také k nepřesnostem způsobeným už při samotném vzniku historického dokumentu. Je jistě žádoucí umět při vyhledávání jména „Cimrman“ detekovat také výskyty jmen „Cimrmann“ nebo „Zimmermann“, protože historicky mohlo být jméno zapsáno z nejrůznějších důvodů odlišně. To samo o sobě vyžaduje sofistikovanější algoritmy vyhledávání, než jsou v současné době využívané fulltextové metody, případně doplněné o lemmatizaci (vyhledávání pomocí základního tvaru slova). Nejvyšším stupněm takového vyhledávání je pak vyhledávání na základě sémantické vazby mezi slovy, kdy například pro zadané slovo „hajný“ očekáváme také nalezení slova „myslivec“ apod.

Projekt InkCapture nabídne pokročilé vyhledávání v ručně psaných dokumentech na základě podobnosti výrazů – nehledá se pouze přesná shoda, ale hledají se také slova podobná. Hledání probíhá na základě zadaného textu, ale také na základě obrazu (v dokumentu se hledá text vizuálně podobný zadanému obrazu).

Základní vlastností celého řešení přitom je učení, zdokonalování schopností na základě zpětné vazby z rozpoznávání textu a vyhledávání.

Budoucnost digitalizace – buďte její součástí

Jak nástroj Kaitos, tak nástroj InkCapture přináší nový přístup k získávání dat z digitalizovaných dokumentů. Díky neuronovým sítím mají potenciál významně snížit objem ruční práce spojený s důsledným vytěžováním informací z dokumentů. Současně se neustále zlepšují a zvyšují kvalitu a spolehlivost vytěžených dat.

Pokud i vy chcete pomoci tvořit budoucnost digitalizace, můžete se do našich projektů zapojit také a poskytnout pro vývoj nástroje InkCapture své historické dokumenty. Dokumenty vám profesionálně zdigitalizujeme a data z nich využijeme pro trénování neuronových sítí, které budou základem popisovaných produktů nové generace. Bližší informace o možnostech zapojení najdete na webu

ISSN 1213-5828

nástroje [10].

Redakční poznámka: Firma EXON se SKIP spolupracuje na realizaci konference Archivy, knihovny, muzea v digitálním světě.

Odkazy
[1] https://bulletinskip.skipcr.cz/vsechna-cisla/prohlizet-cisla/2021-rocnik-30-cislo-1/digitalizace-minulosti-s-nastroji-budoucnosti
[2] https://bulletinskip.skipcr.cz/vsechna-cisla/autori/polansky-petr
[3] https://bulletinskip.skipcr.cz/vsechna-cisla/prohlizet-cisla/2021-rocnik-30-cislo-1
[4] https://bulletinskip.skipcr.cz/vsechna-cisla/rubriky/trendy-v-knihovnach
[5] https://bulletinskip.skipcr.cz/vsechna-cisla/klicova-slova/digitalizace
[6] https://bulletinskip.skipcr.cz/vsechna-cisla/klicova-slova/umela-inteligence
[7] https://bulletinskip.skipcr.cz/sites/default/files/images/838/polansky1.jpg
[8] https://kramerius.nm.cz/view/uuid:ab2a030a-a993-4cd6-b552-9e31ecd04515?page=uuid:4819c261-8a13-11e9-9aeb-001b63bd97ba
[9] http://www.kaitos.eu/
[10] http://www.inkcapture.com/
[11] https://bulletinskip.skipcr.cz/sites/default/files/images/838/polansky2.jpg
[12] https://youtu.be/Hpdu1YwIFGQ
[13] https://bulletinskip.skipcr.cz/sites/default/files/images/838/polansky3.jpg
[14] https://bulletinskip.skipcr.cz/sites/default/files/images/838/polansky4.jpg
[15] https://www.flickr.com/photos/monudet/8272452443/
[16] https://bulletinskip.skipcr.cz/sites/default/files/images/838/polansky5.jpg
[17] https://www.amitia-ai.com/cs/data-annotation/

ISSN 1213-5828