

Další krok ke snazšímu archivnímu výzkumu - staré inventáře Vojenského historického archivu opatřené nově fulltextovým vyhledáváním

27.8.2023 - Tomáš Kykal, Marek Fišer | Vojenský historický ústav Praha

Na digitálním zpřístupnění svých archivních inventářů začal VHA pracovat už v letech 2011-2012, kdy vyhotovil jejich skeny.

Na webových stránkách se objevily se zpožděním až v roce 2015, a to v podobě prostých PDF souborů bez prohledávatelné textové vrstvy. Možná se čekalo více, ale i takovéto řešení tenkrát přineslo badatelům nezanedbatelný užitek. Pracovníci archivu se pokoušeli o zpracování obrazů metodou optického rozpoznávání znaků (OCR), avšak tehdejší dostupné programy nebyly schopny často nezřetelně propsané písmo několikáté kopie inventáře přečíst natolik kvalitně, aby mělo vůbec smysl v získané textové vrstvě vyhledávat. Zatím se bez ní bylo třeba obejít. Radikální změnu přinesla poslední dekáda překotného vývoje počítačových neuronových sítí a strojového učení. Ve Vojenském historickém ústavu Praha (VHÚ) jsme při řešení našich digitalizačních záměrů narazili na jedinečný projekt PERO OCR, vyvíjený týmem dr. Michala Hradiše na Fakultě informačních technologií Vysokého učení technického v Brně. Záhy jsme zjistili, že jejich vysoce přesná technologie pro čtení různých typů tištěného i rukopisného písma, kterou hojně využíváme, se výtečně hodí i ke zpracování nekvalitních strojopisných předloh, jakými jsou třeba právě zmíněné inventáře. Ideální námět pro další společný projekt.

Pro první várku vybrali archiváři VHA nejžádanější inventáře napříč různými obdobími – od první do druhé světové války. Vzhledem k nedostačující kvalitě deset let starých skenů bylo třeba ještě pořídit nové digitální obrazy. Celkem 29 svazků o bezmála 7100 stranách následně prošlo v péči VHÚ rozpoznávacím procesem v editoru PERO OCR. Výsledným produktem pro VHA mělo být tzv. dvouvrstvé PDF, jež pod obrazovou vrstvou skrývá fulltextově prohledávatelnou textovou složku. Tým Michala Hradiše vyšel tomuto přání vstříc a vytvořil pro PERO zvláštní nástroj pro tvorbu takovýchto souborů. První nově zpracované inventáře již na webu VHA nahradily své starší verze. Jsou též volně ke stažení a pro rozlišení jsou označeny ikonou stránky s lupou.

Pro lepší přehled o tom, jaké archivní inventáře prozatím byly „přečteny perem OCR“, uvádíme jejich seznam:

První čs. odboj

- Vladimír VELLA – Bohuslava KRÁLOVÁ, *Československá národní rada 1916–1918*, Praha 1992, 6 sv., 1596 s. (tento inventář nikdy neměl žádný rejstřík)

Československá armáda v letech 1918–1939

- Rudolf SANDER – Marie SALABOVÁ – Božena KAUTSKÁ – Jaroslav BOUČEK, *Vojenská kancelář presidenta republiky 1919–1939*, Praha 1975–1991, 5 sv., 1390 s.
- Václav SLUKA, *Generální inspektorát branné moci 1919–1924*, Praha 1978, 4 sv., 1242 s.
- Václav SLUKA, *Památník osvobození 1919–1939*, Praha 1976, 2 sv., 536 s.
- Václav SLUKA, *Velitelství I. sboru 1935–1939*, Praha 1982, 338 s.

- Marta FABŠIČOVÁ – Václav SLUKA, *Velitelství II., III. a IV. sboru 1935–1939* . Praha 1994, 195 s.

Druhý čs. odboj a fondy a sbírky z let 1939–1945

- Zdeněk VALIŠ – Jiřina DITTRICHOVÁ, *Polský Legion 1939–1941*, Praha 1975 (1981), 37 s.
- Dana NÝVLTOVÁ, *Čs. vojenská správa NV – Francie 1939–1945*, Praha 1965, 31 s.
- Dana NÝVLTOVÁ – Květa KOCOURKOVÁ, *Ministerstvo národní obrany Londýn 1940–1946*, Praha 1974, 196 s.
- Dana NÝVLTOVÁ, *Vojenská kancelář presidenta republiky – Londýn, Košice 1940–1945*, Praha 1967, 81 s.
- Dana NÝVLTOVÁ – František KUTINA, *Studijní skupina – Štáb pro vybudování branné moci Londýn 1940–1945*, Praha 1974, 53 s.
- Ivan ŠTOVÍČEK – Libuše VICHROVÁ, *Československé letectvo – Velká Británie 1940–1946*, Praha 1965, 389 s.
- Danuše NÝVLTOVÁ, *Soupis militárií uložených ve Studijním ústavu ministerstva vnitra 1939–1945*, Praha 1967–1969, 3 sv., 739 s.“
- Ivan ŠTOVÍČEK, *Revoluční gardy 1945*, Praha 1966, 262 s.

Mimo zveřejnění v příslušné sekci na webu VHA jsou uvedené inventáře paralelně přístupné také na webu Digitální studovna MO. V jejím prostředí je velkou výhodou možnost prohledávání celého obsahu či vícesvazkových souborů najednou. Rovněž lze výběr pro řešerši zúžit jen na sbírku Digitální zdroje VÚA-VHA Praha, do níž jsou inventáře zařazeny.

V současnosti se připravuje další dávka, která brzy rozšíří stávající portfolio o další frekventované inventáře. Podaří-li se touto cestou zvýšit efektivitu a komfort při práci s těmito archivními pomůckami, budeme spokojeni.

Vedení projektu: Tomáš Kykal (za VHÚ), Július Baláž (za VÚA-VHA); Michal Hradiš (za FIT VUT); digitalizace: VÚA-VHA; kompletace dat: Vladislav Kusko (VÚA-VHA); zpracování v PERO OCR a postprodukce: Tomáš Kykal; tvorba PDF: Karel Beneš (FIT VUT); tvorba metadat, kompletace datových balíčků a zveřejnění na DSMO: Marek Fišer (VHÚ).

Kontakt pro Vaše podněty a připomínky: ,

<https://www.vhu.cz/dalsi-krok-ke-snazsimu-archivnimu-vyzkumu-stare-inventare-vojenskeho-historickeho-archivu-opatrene-nove-fulltextovym-vyhledavanim>