Jiří Hlaváček

METODICKÉ POSTUPY ARCHIVACE ORÁLNĚ-HISTORICKÝCH PRAMENŮ V DIGITÁLNÍM VĚKU

PŘEDMLUVA

Metoda orální historie poprvé významněji rozvířila „stojaté vody" české historiografie v roce 2000, kdy došlo k založení Centra orální historie (COH) při Ústavu pro soudobé dějiny Akademie věd České republiky (ÚSD AV ČR).[1] Dnes, téměř dvacet let od těchto pionýrských začátků, je zřejmé že orální historie na akademické půdě (i přes počáteční rozpaky a kritiku některých historiků[2]) urazila velký kus cesty a v současnosti ji lze považovat již za jednu z plně etablovaných a institucionalizovaných metod sběru dat v rámci humanitních věd. Současní zájemci z řad odborné a laické veřejnosti se mohou s touto interdisciplinární metodou, kterou lze nejsnáze definovat jako obraz minulosti popsaný slovy dobových aktérů, zevrubně seznámit nejen prostřednictvím množství domácích publikací (včetně podrobné příručky pro přípravu, vedení i analýzu rozhovorů[3]), ale mají také možnost orální historii studovat, resp. vzdělávat se v její teorii a praxi. Vysoké školy nabízí specializované semestrální kurzy nebo metodologické workshopy a pro opravdové zájemce vznikl dokonce samostatný navazující magisterský obor, který má dnes již více než sto padesát absolventů.[4] Stále častěji se s orální historií v posledních letech ve výuce setkávají i žáci základních a středních škol, kteří v rámci dějepisu realizují vlastní orálně-historické projekty. V neposlední řadě se vzpomínky pamětníků staly žádaným artiklem také pro média. Internet tak doslova zaplavily webové stránky, komunitní projekty a portály představující jakési databanky jedinečných vzpomínek, jejichž prostřednictvím se může kterýkoliv zájemce podrobně seznámit s životními příběhy tisícovek známých osobností, ale i tzv. obyčejných lidí.[5] Orální historie tak v českém kontextu právě zažívá éru, kterou lze

[1] Tato studie vznikla v rámci řešení projektu NAKI II s názvem *Virtuální asistent pro zpřístupnění historických audiovizuálních dat* (DG16P02R019) finančně podpořeného Ministerstvem kultury ČR.

[2] Viz např. Jiří Bílek, Rozpaky nad „*Oral History*". In: Historie a vojenství, 4, 2004, s. 121–122.

[3] Viz Miroslav Vaněk – Mücke, Pavel, *Třetí strana trojúhelníku: Teorie a praxe orální historie*. Praha 2015.

[4] Pracoviště Orální historie – soudobé dějiny vzniklo na Fakultě humanitních studií Univerzity Karlovy v roce 2008. Podrobnější informace lze nalézt na webových stránkách (http://ohsd.fhs.cuni.cz).

bez nadsázky nazvat jakýmsi „zlatým věkem". Rostoucí popularita této metody zároveň znamenala také abnormální nárůst počtu orálně-historických pramenů v archivních sbírkách různých institucí.

V souvislosti s neustálým rozvojem technologií, a zejména pak s trendem online uveřejňování rozhovorů prostřednictvím specializovaných webových rozhraní, jsou v současnosti na správce orálně-historických sbírek kladeny stále vyšší nároky, pokud jde o způsoby archivace, indexace a zpřístupňování pramenů. Na několika následujících stránkách se proto pokusím nastínit úskalí a hlavní zásady dobré praxe (*best practise*) v oblasti archivace orálně-historických pramenů v digitální podobě. V tomto ohledu budu vycházet nejen ze své dlouholeté zkušenosti se správou Digitálních sbírek COH ÚSD AV ČR, ale také z příkladů ze zahraničí (archivy, univerzity a nezisková sféra) a v neposlední řadě z poznatků, k nimž jsme s kolegy dospěli při vývoji univerzálního virtuálního rozhraní pro zpřístupnění historických audiovizuálních dat, na němž se ÚSD podílí v rámci grantového projektu ve společně s Matematicko-fyzikální fakultou Univerzity Karlovy (MFF UK) a Národním filmovým archivem (NFA). O tomto projektu se podrobněji zmíním v závěru.

DIGITÁLNÍ SBÍRKY

Geneze virtuálního rozhraní pro zpřístupňování orálně-historických pramenů a podmínky jejich využití

V České republice existuje řada archivů a sbírek nahrávek, jejichž obsahem jsou vzpomínky pamětníků a pamětnic různých historických událostí a období. Některé z těchto sbírek jsou jejich poskytovateli zpřístupňovány výhradně lokálně,[6] jiné prostřednictvím vzdáleného (ne)zabezpečeného přístupu.[7] Tyto archivy, virtuální badatelny či sbírky jsou někdy opatřeny transkripty rozhovorů a mnohdy i základními metadaty pro vyhledání v celé nahrávce či přepisu. Ve všech případech se však jedná o specifická webová rozhraní, která byla vytvořena (často za cenu nemalých nákladů) provozovatelem (institucí, univerzitou či zájmovým sdružením), a to zpravidla z veřejných prostředků (tj. prostřednictvím podpory v rámci různých grantových a dotačních projektů).[8] Při bližší analýze však dojdeme k závěru, že většina těchto rozhraní, i přes vynaložené náklady, nedisponuje žádným softwarovým řešením pro podporu návazného automatizovaného zpracování, ať už ve formě statistik, grafů, inteligentního či pokročilého (např. vícekriteriálního) vyhledávání.[9] Ohledně softwarových řešení v oblasti orálně-historických online archivů navíc prakticky neexistuje žádná volně dostupná dokumentace, s výjimkou některých výsledků v oblasti automatického rozpoznávání řeči a fonetiky.[10]

V Evropě je situace víceméně obdobná, velké archivy zde však (s ohledem na štědřejší možnosti financování) zpravidla mají propracovanou metodiku přístupu, nahrávky jsou přepsány, opatřeny klíčovými slovy, klasifikovány podle různých tezaurů nebo ontologií (např. s ohledem na místní jména, jména osob apod.), úseky nahrávek jsou označeny tematicky a mnohdy propojeny i s geografickým informačním systémem či publikační činností navázanou na uložené prameny. Všechny tyto informace jsou navíc uživatelům zobrazovány přehledným způsobem.[11] Za zajímavý počin lze v této souvislosti považovat španělskou iniciativu Dédalo, jejíž autoři již několik let soustavně prosazují vlastní vícejazyčné (tj. svým způsobem univerzální) open-source řešení pro sbírky rozhovorů.[12]

Mnohem sofistikovanější způsoby archivace a zpřístupňování rozhovorů jsou uplatňovány v zámoří, které lze považovat za kolébku orální historie.[13] Spojené státy americké a Kanada patří v tomto ohledu s velkým náskokem mezi přední vývojáře virtuálních badatelen a obecně lze říci, že věnují překotnému vývoji ve vztahu orální historie a digitálních technologií mnohem větší pozornost než orální historici na starém kontinentu.[14] Samotný vztah mezi orální historií a digital humanities lze sledovat ve třech základních rovinách: katalogizace rozhovorů a jejich indexace, nahrávky rozhovorů a jejich přepis, a konečně obsah a jeho datové mapování.[15] Jednoznačně největší orálně-historické sbírky představuje již od svého založení Archiv vizuální historie Nadace šoa Univerzity Jižní Kalifornie s bezmála

[5] Viz např. Paměť národa (http://www.pametnaroda.cz) nebo Skautské století (http://www.skautskyinstitut.cz/skautske-stoleti).

[6] Viz např. Orální historie českého divadla (http://vis.idu.cz) nebo Sbírka zvukových záznamů NFA (https://nfa.cz/cz/sbirky/sbirky-a-fondy/zvukove-nahravky).

[7] Viz např. Digitální sbírky COH ÚSD AV ČR (http://www.coh.usd.cas.cz/sbirky-rozhovoru/badatelna).

[8] Stručný přehled o stavu českých archivů orální historie viz např. Pavel Mücke, *The Pleasures and Sorrows of Czech Oral History (1990–2012)*. In: Oral History Journal of South Africa, 1, 2013, s. 111–130.

[9] Na rozdíl od většiny obdobných platforem vyvíjených v rámci tzv. digital humanities. Podrobněji k této problematice viz Česká asociace pro digitální humanitní vědy (https://www.czadh.cz/).

[10] Jan Vavruška – Jan Švec – Pavel Ircing: *Phonetic Spoken Term Detection in Large Audio Archive Using the WFST Framework*. In: Text, Speech, and Dialogue. Heidelberg 2013, s. 402–409.

[11] Viz např. sbírka rozhovorů „*Nucené práce 1939–1945*", jíž spravuje Freie Universität Berlin (http://www.zwangsarbeit-archiv.de).

[12] Podrobnější informace viz http://www.fmomo.org/dedalo.

[13] Viz např. americký webový portál „*The History Makers*" (https://www.thehistorymakers.org/digital-archives).

[14] Viz např. specializovaný webový portál „*Oral History in the Digital Age*" (http://ohda.matrix.msu.edu).

[15] Michael Frisch – Doug Lambert, *Mapping approaches to oral history content management in the*

56 000 nahrávkami ve 40 jazycích.[16] Tento archiv je přístupný na půdě kalifornské univerzity a dále prostřednictvím cca 40 přístupových bodů rozmístěných po celém světě, včetně Německa, České republiky, Polska a Maďarska.[17] Bohužel, jeho software je proprietární, smluvně licencovaný a jeho zdrojový kód nelze tedy využít jako východisko pro vývoj nových obdobně zaměřených webových rozhraní.[18]

Jak již bylo zmíněno v úvodu, pro systematický rozvoj orální historie a její aplikaci (nejen) v rámci české historiografie bylo na počátku ledna 2000 zřízeno při ÚSD AV ČR specializované pracoviště COH. Vedle základního výzkumu nedávné minulosti realizovaného prostřednictvím této metody se centrum dlouhodobě zaměřuje právě na archivaci a správu sbírek rozhovorů pořízených v rámci projektů řešených výzkumnými pracovníky ústavu či jeho spolupracovníky. Na tomto místě je třeba podotknout, že COH nemá statut archivu dle archivního zákona, a proto se na něj nevztahují všechny požadavky, které tento zákon stanovuje. Z hlediska legislativy jsme tedy především subjektem, který shromažďuje osobní a citlivé údaje, a pro tuto činnost jsme také od Úřadu pro ochranu osobních údajů získali příslušnou akreditaci. Nelze proto vyloučit, že níže popsaný způsob archivace orálně-historických pramenů může v některých ohledech kolidovat s archivní praxí. Tento text je tak nutné považovat spíše za určitou formu případové studie než za konkrétní manuál metodických postupů.

Významným mezníkem pro archivaci audio-vizuálních pramenů v COH se stal rok 2012, kdy bylo rozhodnuto, že s ohledem na prostorové možnosti centra i převažující způsob práce badatelů s archivními prameny, budou všechny rozhovory nadále archivovány již pouze v elektronické podobě, a to v tzv. Digitálních sbírkách COH. Za tímto účelem bylo třeba navrhnout databázové a posléze i webové rozhraní, které by umožnilo rozhovory nejen archivovat, ale zároveň také dálkově zpřístupňovat badatelům. První provizorní verze sbírek byla spuštěna již v roce 2012 a její jádro tvořilo upravené prostředí online výukové platformy MOODLE.[19]

digital age. In: Oral History in the Digital Age. [online] [03-03-2019]. Http://ohda.matrix.msu.edu/2012/07/mapping.

[16] USC Shoah Foundation Visual History Archive (https://sfi.usc.edu).

[17] V České republice je tento archiv přístupný (spolu s dalšími třemi archivy) v Centru vizuální historie Malach MFF UK (https://ufal.mff.cuni.cz/malach).

[18] Problematika orální historie a digital humanities je zpravidla pojednávána prostřednictvím online textů, tištěné publikace jsou v tomto případě spíše ojedinělé. Přehledně je tato problematika shrnuta v kolektivní monografii Douga Boyda a Mary Larson. Viz Doug Boyd – Mary Larson, *Oral History and Digital Humanities: Voice, Access, and Engagement*. New York 2014.

[19] Modular Object Oriented Dynamic Learning Environment (Modulární objektově orientované dynamické prostředí pro výuku) je balíček pro tvorbu výukových systémů a elektronických kurzů na internetu. Podrobnější informace na webu https://moodle.org.

V roce 2015 se ústavu podařilo získat od AV ČR finanční příspěvek v rámci programu *Strategie AV 21*, který byl investován do nového systému v podobě upravené a šifrované verze redakčního publikačního systému WordPress,[20] na němž běží rozhraní Digitálních sbírek dodnes. Z technického hlediska je provoz Digitálních sbírek zajišťován na ústavním serveru s jádrem Intel Xeon a data jsou automaticky zrcadlena na záložní pevný disk.

Digitální sbírky jako takové vznikly především za účelem shromažďovat a dále zpřístupňovat sbírky rozhovorů pořízené v souladu s metodologickými postupy a etickými pravidly orálně-historického výzkumu. Cílem je umožnit registrovaným badatelům z řad akademické obce (tj. výzkumným pracovníkům, pedagogům, studentům a zaměstnancům archivů či muzeí) dále využívat již vzniklé orálně-historické prameny, a to v rámci vlastních společenskovědních výzkumů. Sbírky, které v současnosti čítají více než 3 000 zvukových či audiovizuálních záznamů (tj. rozhovorů) roztříděných podle jednotlivých grantových projektů, tedy nejsou (na rozdíl např. od popularizačně-edukativního portálu *Paměť národa*, jehož provozovatelem je sdružení Post Bellum) primárně určeny zájemcům z řad široké laické veřejnosti.[21] Významnou roli v procesu zpřístupňování orálně-historických rozhovorů hraje současná legislativa. Badatel zaregistrovaný v rozhraní Digitálních sbírek COH se proto zavazuje řídit zákonem č. 110/2019 Sb., o zpracování osobních údajů, zákonem č. 89/2012 Sb., občanským zákoníkem, ve znění pozdějších předpisů, a nařízením (EU) 2016/679 o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů (GDPR) a dále dodržovat zásady Etického kodexu České asociace orální historie[22] a stanovený citační úzus COH.[23] Po vyplnění badatelského listu v elektronické podobě je uživateli vytvořen profil a přes webové rozhraní dočasně vzdáleně zpřístupněny požadované rozhovory.[24] Po uplynutí lhůty jednoho měsíce je přístup k pramenům badateli z bezpečnostních důvodů automaticky odebrán a pro opětovné nahlížení je třeba znovu vyplnit badatelský list.

[20] WordPress je svobodný open source redakční publikační systém napsaný v PHP a MySQL a vyvíjený pod licencí GNU GPL. Podrobnější informace na webu https://cs.wordpress.org.

[21] Správce Digitálních sbírek si nicméně vyhrazuje právo ve výjimečných případech a se souhlasem vedoucí COH umožnit přístup do sbírek i badatelům z řad laické veřejnosti bez institucionální podpory a zaměstnancům/zaměstnankyním státní správy, pokud náležitě doloží a vymezí účel nahlížení, a pokud tento účel nebude v rozporu s informovanými souhlasy (či smlouvami o poskytnutí práv) pamětníků.

[22] Etický kodex COHA viz http://www.coha.cz/wp-content/uploads/2018/05/Eticky_kodex-COHA_2018.pdf.

[23] Dle následujícího vzoru: Rozhovor s [NARÁTOR] vedl [TAZATEL], [DATUM]. Sbírka [NÁZEV], Digitální sbírky COH ÚSD AV ČR.

[24] Badatelský list viz http://www.coh.usd.cas.cz/wp-content/uploads/badatelsky_list_ds_coh.doc.

V zabezpečené sekci Digitálních sbírek jsou badatelům primárně zveřejňovány pouze transkripty rozhovorů, průvodní informace (protokoly o rozhovoru nebo karty narátora) a informovaný souhlas. Preferována je varianta zpřístupnění omezeného počtu konkrétních rozhovorů, celé sbírky jsou tak zveřejňovány pouze výjimečně, a to vždy s ohledem na účel nahlížení. Originální nahrávka rozhovoru je dostupná jen v případě, pokud není k dispozici její přepis. K tomuto opatření bylo přistoupeno z několika důvodů. Za prvé, z dosavadní několikaleté praxe je zřejmé, že badatelé preferují práci s transkriptem a o přístup k samotnému prameni často nejeví zájem. Za druhé, zpřístupňování malých datových souborů snižuje celkové hardwarové nároky na provoz sbírek. V neposlední řadě toto opatření slouží jako prostředek minimalizace rizika zneužití dat či neoprávněného zveřejňování orálně-historických rozhovorů v online prostoru. Nicméně platí, že správce sbírek je oprávněn na požádání dočasně zveřejnit konkrétní zvukový či audiovizuální záznam prostřednictvím šifrovaného webového rozhraní, pokud to není v rozporu s požadavky informovaného souhlasu pamětníka. Všechny rozhovory i další prameny je také možné po vyplnění badatelského listu bez omezení studovat v elektronické podobě v rámci osobní návštěvy COH v budově ÚSD ve Vlašské v Praze.

METODIKA ARCHIVACE

Současný stav, formální náležitosti a nové metadatové schéma pro orálně-historické sbírky

V této části se zaměřím na konkrétní metodické zásady a postupy archivace orálně-historických pramenů, tak jak jsou uplatňovány v praxi Digitálních sbírek COH v posledních několika letech. Orálně-historický pramen je specifický svým obsahem i formou. Jedná se o audio či video záznam, na němž je fixována autentická interakce mezi tazatelem a dotazovaným (tj. narátorem, resp. pamětníkem), jejímž předmětem je zpravidla subjektivní reflexe lidské zkušenosti s určitou historickou událostí či dobou. Tento vždy nutně minimálně zvukový, ale v posledních letech stále častěji také obrazový, záznam bývá posléze zpravidla doplněn o doslovný nebo edičně upravený transkript a průvodní dokumentaci v podobě protokolu o rozhovoru a narátorem signovaného informovaného souhlasu se zpracováním osobních a citlivých údajů.[25] Volitelně pak mohou být doplňkem rozhovoru také různé další materiály jako kopie fotografií, archivních pramenů či jiných dokumentů získaných od pamětníků během rozhovoru.

[25] COH doporučuje používat v rámci orálně-historických výzkumů aktuální vzor informovaného souhlasu uveřejněný na webu České asociace orální historie (http://www.coha.cz/wp-content/uploads/2018/10/Informovany_souhlas_GDPR_vzor.pdf).

Každý orálně-historický záznam v digitálním archivu či sbírce by se tedy v ideálním případě (tj. v souladu s metodologií orálně-historických výzkumů) měl skládat minimálně ze čtyř komponent, jimiž jsou audio či video záznam rozhovoru (multimédiální soubor), jeho přepis (textový soubor), protokol o rozhovoru (textový soubor) a informovaný souhlas (obrázkový či pdf soubor). Základní pojítko mezi těmito čtyřmi soubory představuje v rámci databáze osoba konkrétního narátora (tj. pamětníka, respondenta nebo informátora), resp. unikátní přírůstkové číslo, které mu je při ukládání do sbírek automaticky přiděleno. Právě tato signatura je klíčovým údajem, na nějž jsou navázány všechny další informace a soubory. Signatura je vždy tvořena na základě kombinace čísel a písmen, a to podle následujícího klíče:

Vzorová signatura: N0001-01-01-CSNT-M01;
N0001: přírůstkové/pořadové číslo narátora;
01: číslo rozhovoru s tímto narátorem;
01: část rozhovoru/setkání s narátorem (v případě, že je jeden rozhovor rozdělen do více souborů či se uskutečnil v rozmezí několika dní);
CSNT: zkratka utvořená na základě názvu projektu, v jehož rámci byl rozhovor s narátorem pořízen;[26]
M01: označuje typ souboru (M – materiál, T – transkript, A – audio, V – video, O – ostatní) a jeho pořadové číslo.[27]

Klasický protokol (či záznam) o rozhovoru, který v rámci orální historie plní roli jakéhosi průvodního dokumentu každého interview uloženého ve sbírkách, byl od roku 2016 v COH nahrazen tzv. kartou narátora. Karta narátora obsahuje, obdobně jako protokol, všechny důležité informace o rozhovoru (tj. základní údaje o narátorovi a rozhovoru, charakteristické mimoslovní projevy, zjevné omyly a neobvyklé události), navíc je však rozšířena o stručný biogram pamětníka, a především pak o tzv. kondenzovaný rozhovor (tj. shrnující protokol jeho obsahu). Jedná se v podstatě o tematický popis obsahu rozhovoru, vedený po chronologické linii, v němž jsou zdůrazněny klíčové momenty interview, v případě potřeby doplněné o přímé citace z transkriptu (včetně časové značky citovaných pasáží či klíčových momentů).

Specifická pravidla je třeba dodržovat také při zhotovování transkriptu. Každý přepis uložený ve sbírkách by měl být opatřen úvodní hlavičkou, která obsahuje všechny základní informace potřebné pro identifikaci rozhovoru. Vzorová hlavička v Digitálních sbírkách COH vypadá následovně:

[26] V tomto případě se jedná o grantový projekt „*Česká společnost v období normalizace a transformace: životopisná vyprávění*“, který byl v COH ÚSD AV ČR řešen v letech 2011–2015.

[27] Zkratky byly zvoleny s ohledem na možnost mezinárodního využití (tj. material, transcript, audio, video, others).

PŘEPIS ROZHOVORU: PŘÍJMENÍ, JMÉNO NARÁTORA/NARÁTORKY
Signatura: *kombinace číslic a písmen dle specifického klíče*
Jméno a příjmení narátora/ky: *včetně akademických titulů a iniciál užívaných v přepisu*
Ročník narození: *ve formátu RRRR (rok)*
Datum rozhovoru: *ve formátu DD. MM. RRRR (den, měsíc, rok)*
Místo rozhovoru: *místo (adresa), název (např. instituce), město či obec*
Jméno a příjmení tazatele/ky: *včetně akademických titulů a iniciál užívaných v přepisu*
Délka rozhovoru: *ve formátu HH:MM:SS (hodiny, minuty, sekundy)*
Počet stran přepisu: *číslo*
Projekt: *výzkumný projekt, případně institucionální podpora (číslo grantu, poskytovatel)*
Informovaný souhlas: *ano, ne, případné požadavky ze strany narátora*
Poznámka: *jakékoliv podstatné informace související s formou či obsahem přepisu*

V rámci přepisu je také důležité dodržovat jednotná pravidla pro úpravu a formátování, která jsou nutným předpokladem pro následné bezproblémové automatizované zpracování prostřednictvím dalších aplikací, včetně synchronizace zvukové či audiovizuální stopy s textem přepisu. Z tohoto důvodu při zhotovování přepisu zásadně užíváme pouze podobu doslovného transkriptu (tj. přepis včetně tzv. vatových slov, přeřeknutí apod.) a jakékoliv informace, které nejsou součástí zvukového (či obrazového) záznamu (resp. nejsou z něj zřejmé jako např. nonverbální projevy a emoce) uvozujeme v textu transkriptu hranatými závorkami.[28] Jednotlivé konce odstavců přepisu poté označujeme časovou stopou, která usnadňuje čtenáři dohledávání pasáží v původním audio(vizuálním) záznamu. Pro ilustraci uvádíme příklad úryvku jednoho z transkriptů:

JH: Vzpomínáte si, co jste dělal v noci z 20. na 21. srpna 1968? [01:01:00]
HB: Ano, to vím naprosto přesně. [smích] To se totiž nedá zapomenout. No, jak bych to řekl. Víte, tam byl jeden takový člověk, jmenoval se – teď to prosím na chvíli vypněte. [ZÁZNAM POZASTAVEN]. Takže tenhle dotyčný, on měl funkci [NESROZUMITELNÉ] a ten to vše řídil. [01:01:32]

Poté co datové soubory splní všechny vstupní požadavky a obdrží jedinečné identifikátory v podobě signatur jsou dále obohaceny o základní metadata, která

[28] Doslovné transkripty nejsou v rámci české orální historie příliš rozšířeným řešením. Jejich hlavním nedostatkem je časová a finanční náročnost, proto jsou často nahrazovány pouze obsahovým shrnutím ve formě klíčových slov a časové stopy. Badatel tak získává jakýsi tematický časový itinerář rozhovoru, který však lze pro následné automatické zpracování využít jen velmi obtížně.

umožňují vyhledávání v korpusu všech záznamů, a to na třech samostatných úrovních (projekt, narátor, rozhovor). Celkem je v rámci Digitálních sbírek v současnosti možné data třídit a vyhledávat prostřednictvím 20 metadatových položek, z nichž je však pouze 8 dostupných registrovaným badatelům (jedná se o tato metadata: signatura, narátor, projekt, rozhovor, datum rozhovoru, místo rozhovoru, tazatel, poznámka), zatímco zbývajících 12 slouží výhradně pro potřeby správce (tj. pro katalogizační účely).

Vedle těchto formálních metadat jsou rozhovory, resp. jejich přepisy, dále indexovány také prostřednictvím obsahových metadat (tzv. kódů), které informují uživatele o tom, co konkrétně je obsahem daného orálně-historického pramene. Obsahová metadata jsou rozhovorům přiřazována manuálně na základě obsahové analýzy textu a s ohledem na oborový tezaurus (tj. slovník obsahující témata, pojmy, osobnosti, události, místa apod. včetně synonym a antonym).[29] Vytváření obsahových metadat představuje tvůrčí a časově velmi náročnou činnost, která by v ideálním případě měla být vždy výsledkem spolupráce mezi tazatelem, který rozhovor vedl (protože právě on je tím, kdo nejlépe zná konkrétní okolnosti vzniku pramene), a správcem, jenž zná specifika kategorizace metadat v rámci dané sbírky. Proces tvorby obsahových metadat probíhá v Digitálních sbírkách COH průběžně od roku 2016, i přesto však bylo na počátku roku 2019 výše popsaným způsobem indexováno jen necelých 10% ze všech archivovaných rozhovorů.

Problémem výše uvedené metadatové struktury je její přílišná generalizace. Omezené množství základních metadat se totiž dlouhodobě ukazuje jako nedostatečné pro efektivní vyhledávání (resp. filtrování dat), a to jak z pozice kurátora sbírek, tak z pohledu badatele. Při řešení v úvodu zmiňovaného projektu tzv. virtuálního asistenta jsme proto byli nuceni ve spolupráci s kolegy z NFA navrhnout zcela nové metadatové schéma, které by zároveň mohlo fungovat jako univerzální standard pro katalogizaci a indexaci orálně-historických sbírek. Výsledkem našich společných diskusí je níže uvedené třístupňové hierarchicky uspořádané schéma vzorových metadat, v nichž může uživatel (správce či badatel) vyhledávat buď napříč všemi úrovněmi, nebo v každé úrovni zvlášť. Toto schéma je od začátku roku 2019 testováno v rámci orálně-historických sbírek ÚSD a NFA, včetně vytváření nového tezauru pro obor audiovizuální kultury a historiografie. První úroveň obsahuje základní informace o narátorovi. Druhá rovina faktické údaje o rozhovoru samotném. Třetí úroveň slouží pro zaznamenání dat technické povahy. Níže jsou uvedeny aktuálně využívané kategorie metadat v rámci jednotlivých úrovní:

1. ÚROVEŇ: Narátor
signatura; příjmení; jméno; titul; pohlaví; datum narození; projekt; informovaný

[29] Tezaurus je v současnosti vytvářen v Microsoft Access Database.

souhlas; poznámka; klíčová slova (v kategoriích: profese, období, zaměření, charakteristika); výstupy; materiály; kontaktní údaje.[30]

2. ÚROVEŇ: Rozhovor
název souboru; formát rozhovoru (tj. narativní, polostrukturovaný, ostatní apod.); transkript (tj. doslovný, redigovaný, orientační, ne apod.); datum rozhovoru; délka rozhovoru; místo rozhovoru (včetně propojení na Geografický informační systém a Google Maps); jméno a příjmení tazatele; klíčová slova; kódy; jazyk rozhovoru (ISO kód); poznámka.

3. ÚROVEŇ: Technické údaje
originální záznam; typ souboru/materiálu (tj. audio, video, přepis, obrázek, ostatní); formát souboru; velikost souboru; původce záznamu; licence; poznámka.

Ve většině případů systém nabízí odpověď z předem definovaných možností výběru, jejichž cílem je především intuitivnost celého rozhraní, která má minimalizovat selhání lidského faktoru při zadávání vstupních dat. Proces vkládání je zároveň maximálně automatizován, přičemž některá data (zejména ta z třetí úrovně) je systém schopen abstrahovat ze vstupních dat nezávisle, tj. bez zásahu uživatele.

Pokud jde o standardy multimediálních souborů orálně-historických pramenů, v Digitálních sbírkách COH je jako výchozí formát pro zvuk preferován soubor typu wav (*Waveform Audio File Format*), jehož předností je ukládání v nekomprimované lineární podobě, což je nejvýhodnější pro případnou další konverzi. Samotné rozhovory jsou potom v rámci webového rozhraní zpřístupňovány vždy pouze v komprimované podobě, a to v podobě mp3 (MPEG Audio Layer 3) souborů s různou kvalitou datového toku (zpravidla však 320 kb/s). U audio-vizuálních souborů v současnosti není pro jejich variabilitu stanoven ideální výchozí formát, obdobně jako v případě zvuku však platí, že video je online zpřístupňováno vždy v komprimované podobě.

BUDOUCNOST (NEJEN) PRO ORÁLNÍ HISTORII

Virtuální asistent pro zpřístupnění historických audiovizuálních dat

Výše popsané nové metadatové schéma bylo navrženo jako součást projektu, jehož cílem je vytvoření tzv. virtuálního asistenta pro zpřístupnění historických

[30] U položek „jméno" a „příjmení" lze uvést libovolné množství „alias" (umělecká jména, rodná příjmení apod.). V případě „výstupů" a „materiálů" lze nahrát libovolné množství multimediálních souborů (audio, video, text, obrázky).

audiovizuálních dat, který by v budoucnosti mohl nahradit stávající, a pro naše účely již ne zcela vyhovující, rozhraní Digitálních sbírek COH na bázi WordPress. Záměr vývojářů z MFF UK je však mnohem ambicióznější. Vedle modifikovaného repozitáře[31] by měla být hlavním výsledkem projektu sada vzájemně propojených univerizálních nástrojů, které umožní zpracování archivů zvukových či audiovizuálních nahrávek a jejich následné zpřístupnění či využití jak ve vědeckém výzkumu, tak ve vzdělávání. Hlavní roli zde sehrávají jazykové technologie pro automatickou analýzu mluvené řeči a její následné zpracování automatickými metodami jazykové analýzy. Zpracovaní bude navíc podpořeno řadou dalších softwarových nástrojů (modulů) pro studie vycházející z materiálu samotného (tj. z rozhovorů) a z jeho kvantitativních i kvalitativních charakteristik. Tyto nástroje by měly umožnit extrakci částí zdrojových nahrávek, jejich přepisů, grafů, tabulek a zajistit i vizualizaci kvantitativních výsledků. Jednou z hlavních předností vyvíjených nástrojů je právě možnost jejich mezinárodního využití, a to vzhledem k jazykové nezávislosti řady plánovaných komponent.[32]

Výsledná platforma by se měla skládat celkem ze čtyř komponent: repozitáře, systému pro zpřístupnění, softwaru pro deponování nahrávek a programu pro jejich anotaci a exploataci. Ověření funkčnosti všech součástí systému, včetně úprav podle zjištěných nedostatků je prováděno ve spolupráci s Digitálními sbírkami COH a Sbírkou zvukových záznamů NFA. Prostřednictvím virtuálního asistenta by tak mělo být možné zpracovávat zdrojová data od malých souborů pro individuální vědecké projekty založené na metodě orální historie (např. závěrečné kvalifikační, tj. bakalářské, magisterské a disertační, práce) až po velké soubory již existujících nahrávek, které jsou výsledkem týmové práce skupin a institucí (grantové projekty), nebo pro rozsáhlé rešerše archivních institucí.

Podle ideálního scénáře by tedy měly být začátkem roku 2020 nahrávky nejprve vloženy za pomoci nástrojů pro deponování nahrávek do repozitáře s metadaty, kde budou opatřeny permanentními identifikátory pro citace a odkazování v publikacích i na internetu a jasně definovanými licencemi pro další použití respektujícími autorskou ochranu, ochranu osobních údajů i archivní zákon.[33] V následném kroku budou podle potřeby tyto nahrávky obohaceny o automatický přepis, případně o další údaje získané jak automaticky softwarovými nástroji pro zpraco-

[31] Jedná se o zabezpečené úložiště, které je v současnosti zpřístupněno registrovaným a přihlášeným uživatelům online v testovacím náhledu s několika vzorky dat (https://repositar.usd.cas.cz/xmlui).

[32] Využití metody strojového učení, která omezuje nutnou jazykovou expertizu na anotaci tzv. trénovacích dat (pro daný jazyk a doménu).

[33] Testovací repozitář vycházející ze softwaru DSpace 5je dostupný na webu COH ÚSD AV ČR (https://repositar.usd.cas.cz/xmlui).

vání mluvené řeči a její následnou analýzu (např. lematizaci, rozpoznávání jmen a dalších pojmenovaných entit, identifikaci tématu a podtémat, indexaci klíčovými slovy z existujících tezaurů apod.), tak pomocí manuální analýzy (za podpory příslušného softwaru, např. pro opravy chybné automatické transkripce, nebo anotace pomocí vlastní klasifikace podle potřeb daného konkrétního výzkumu). Obohacené nahrávky budou poté vloženy jednak do repozitáře za účelem jejich trvalého uchování a zároveň také do systému pro jejich zpřístupnění „koncovému" uživateli – badateli, výzkumníkovi, či pracovníkovi paměťové instituce. V tomto systému, který tvoří podstatu vyvíjeného virtuálního asistenta, pak proběhne také iniciální příprava (indexace podle přepisu a dalších atributů) a konverze dat pro efektivní přístup.

Závěr

Metoda orální historie je již od svých počátků v první polovině minulého století, kdy došlo ve Spojených státech k její profesionalizaci, významně provázána s technologickým pokrokem. Vedle audio-vizuálních záznamových zařízeních však společně s nástupem digitální éry po roce 2000 začíná v rámci orálně-historického výzkumu sehrávat stále větší roli nově také otázka návazného zpracování vzniklých pramenů, a to zejména v rovině možnosti jejich katalogizace, indexace a obsahového mapování. Charakteristickým rysem současného stavu v oblasti digital humanities je rychlost, již se vývoj softwarových řešení nezadržitelně žene kupředu. Ruku v ruce s tímto trendem jde problematika permanentního nárůstu množství dat, jak co do objemu, tak do obsahu. Orálně-historické prameny přibývají geometrickou řadou, ale rychlost jejich zpracování je významně omezena kapacitními možnostmi manuálního zpracování, tj. lidským faktorem. Automatizované nástroje typu virtuálního asistenta tak mohou v tomto případě kurátorům sbírek významně pomoci zefektivnit jejich práci. Zároveň je však třeba mít neustále na paměti, že spolehlivost softwaru (zejména pak toho analytického) není stoprocentní a jeho využití tak s sebou vždy nese potenciální riziko zkreslení či selekce dat, které může významně ovlivnit validitu výzkumů založených na rešerších orálně-historického materiálu.

Seznam literatury

Jiří Bílek, *Rozpaky nad „Oral History"*. In: Historie a vojenství, 4, 2004, s. 121–122.
Doug Boyd – Mary Larson, *Oral History and Digital Humanities: Voice, Access, and Engagement*. New York 2014.
Michael Frisch – Doug Lambert, *Mapping Approaches to Oral History Content Management in the Digital Age*. In: Oral History in the Digital Age. [online] [03-03-2019]. Http://ohda.matrix.msu.edu/2012/07/mapping.
Pavel Mücke, *The Pleasures and Sorrows of Czech Oral History (1990–2012)*. In: Oral History Journal of South Africa, 1, 2013, s. 111–130.
Miroslav Vaněk – Pavel Mücke, *Třetí strana trojúhelníku: Teorie a praxe orální historie*. Praha 2015.
Jan Vavruška – Jan Švec – Pavel Ircing: *Phonetic Spoken Term Detection in Large Audio Archive Using the WFST Framework*. In: Text, Speech, and Dialogue. Heidelberg 2013, s. 402–409.

ZUSAMMENFASSUNG

METHODISCHER LEITFADEN ZUR ARCHIVIERUNG MÜNDLICHER HISTORISCHER QUELLEN IM DIGITALEN ZEITALTER

Diese Studie beschäftigt sich mit aktuellen Möglichkeiten, die die virtuelle Internetschnittstelle für die Sammlung mündlicher historischer Quellen im digitalen Zeitalter bietet. Die Studie befasst sich mit der Entstehung des Zentrums für digitale Sammlungen mündlicher historischer Quellen am Institut für Zeitgeschichte der Tschechischen Wissenschaftsakademie und mit den Bedingungen für die Interviewauswertung. Im nächsten Teil werden die methodischen Verfahren zur Archivierung der Interviews, die unter Beachtung der Grundsätze der mündlich überlieferten Geschichten entstanden sind, detailliert besprochen. Abschließend präsentiert der Autor ein Projekt, das auf einer offenen Plattform beruht und wie der virtuelle Assistent für die Bewertung der audiovisuellen historischen Daten funktioniert. Dieser soll in der Zukunft als die Universallösung für das Management und für die Archivierung und Bereitstellung der Sammlungen verschiedenster audiovisueller Quellen dienen.

■ Překlad Artlingua, a. s.

SUMMARY

METHODOLOGICAL GUIDE TO ARCHIVING ORAL-HISTORICAL SOURCES IN THE DIGITAL AGE

The study focuses on the current possibilities of using a virtual web interface for collections of oral-historical sources in the digital age. Attention is paid to the genesis of the Digital Collections of Oral History Center at the Institute of Contemporary History of the Czech Academy of Sciences and the conditions for accessing interviews. In the next part, there are discussed in detail the methodological procedures for archiving interviews, which were acquired in accordance with the principles of oral history. In conclusion, the author presents a project of an open platform of Virtual Assistant for Accessing to Audiovisual Historical Data, which should offer a universal solution for managing, archiving and making available collections of various audiovisual sources in the future.

■ Překlad autor