Digitální obrat v českých humanitních a sociálních vědách

Radim Hladík
a kolektiv

STUDIA NOVÝCH MÉDIÍ

KAROLINUM

Digitální obrat v českých humanitních a sociálních vědách

Radim Hladík a kolektiv

———

Recenzovali:
Mgr. Marek Debnár, Ph.D.
Mgr. Martin Charvát, Ph.D.

Financováno Evropskou unií
NextGenerationEU

Národní plán obnovy

Publikace byla vydána za podpory Ministerstva školství, mládeže a tělovýchovy a Národního plánu obnovy v rámci projektu Transformace pro VŠ na UK (reg. č. NPO_UK_MSMT-16602/2022).

Vydala Univerzita Karlova
Nakladatelství Karolinum
Praha 2022
Redakce Václav Hozman
Grafická úprava Jan Šerých
Sazba DTP Nakladatelství Karolinum
Vydání první

ISBN 978-80-246-5193-4
ISBN 978-80-246-5393-8 (online : pdf)

Univerzita Karlova
Nakladatelství Karolinum

www.karolinum.cz
ebooks@karolinum.cz

OBSAH

Dějiny v datech

Digitální současnost

PŘEDMLUVA

Radim Hladík

Ohledně vymezení digitálních humanitních věd (*digital humanities*, DH) a komputačních sociálních věd (*computational social sciences*, CSS) panuje tolik názorů, že snahy o jejich definici tvoří samostatný žánr. Vzpírají se jasnému vymezení, neboť se samy průběžně vyvíjejí. Jejich dynamika, která se snaží držet krok s rychlými pokroky ve vývoji nových algoritmů, nepřeje tomu, aby digitální přístupy vnesly do humanitního a sociálněvědního výzkumu pevně nastolené výzkumné programy, s nimiž by institucionalizované obory mohly standardně spolupracovat.

Při úvahách o tom, jak digitální humanitní a sociální vědy představit českým výzkumníkům a studentům, by příklon k té či oné definici mohl napáchat více škody než užitku. Ještě méně užitečné by bylo pokoušet se přidat k již existujícímu dlouhému seznamu definici vlastní. Tím nemá být řečeno, že úctyhodná intelektuální práce, jež byla dosud v teorii digitálních humanitních věd vynaložena, přichází vniveč. Pro budoucí rozvoj tohoto oboru zůstávají takové kritické diskuse naprosto nezbytné. Tato kniha se však k představení digitálních humanitních věd vydává oklikou. Ve svém zaměření na demonstrativní případové studie vychází z úsloví, že důkaz pudinku je v jídle.

Autorky a autoři, kteří odpověděli na výzvu představit principy digitálního bádání širší akademické komunitě, přijali netriviální úkol ukázat čtenářům, jak oni samotní nebo jiní badatelé postupují při zodpovídání výzkumných otázek spojených s digitálními přístupy, aniž by použité metody popisovali technickým žargonem a for-

mulemi. Pokud jsou metody, jakkoliv by mohly být jejich výsledky zajímavé, prezentovány jako černá skříňka, nepůsobí přesvědčivě. Na druhou stranu, přísně zdokumentovaná metodologie může na netrénovaného čtenáře působit esotericky a nepřístupně. Proto zde dáváme nahlédnout do principů digitálního bádání tak, jako bychom servírovali pudink bez receptu, ale s otevřenými dveřmi do kuchyně.

Navzdory svému praktickému zaměření tato kniha nemá sloužit jako autoritativní učebnice nebo příručka k osvojení metod. Zvolený hybridní žánr inklinuje k didakticky pojatým původním případovým i přehledovým studiím, jejichž soubor cílí na konstrukci od technických detailů abstrahovaného intelektuálního modelu toho, co znamená provádět digitální výzkum na empirickém materiálu humanitní a sociálněvědní provenience. Pokud se autorskému kolektivu podařilo zadaný úkol splnit, čtenáři by si měli z četby vedle dílčích poznatků o demonstrativních výzkumných problémech odnášet především pochopení toho, jak digitální humanitní vědy uvažují o svých výzkumných postupech, jak formulují své otázky a problémy a jakými transformacemi procházejí jejich předměty bádání.

Induktivně vybudovaný intelektuální model digitálního výzkumu, který zde nahrazuje teoretické definice, nevyžaduje, aby digitální humanitní a sociální vědy vytvářely vlastní oborové teritorium. Jednotlivé studie fungují spíše jako špendlíky zapíchnuté do již známých map, kde označují místa specifického badatelského zájmu. Samotná šíře tematického spektra pokrytého v této knize pak dokládá, že digitální postupy nejsou vlastní žádné specializaci nebo tématu.

Variabilita uplatnění digitálních témat a přístupů se odráží i v členění jednotlivých kapitol. Bylo možné se v něm opřít i o tradiční oborové nomenklatury a seskupit příspěvky podle jejich mateřských disciplín. Zastoupeny jsou z této perspektivy lingvistika, lexikografie, sociologie, archeologie, historiografie, mediální studia, religionistika, literární věda i muzeologie. Konečná klasifikace se však opírá o jiná pravidla, která zdůrazňují interdisciplinární propojení různých technik a problémů.

Důvod, proč digitální humanitní vědy a komputační sociální vědy často vystupují nebo jsou vnímány jako svébytné specializace, může spočívat právě ve vzájemných vazbách mezi různými technikami a datovými modely než v existenci autonomních výzkumných

programů. Digitální sociolog i literární vědec se mohou od sebe inspirovat a sdílet zájem o analýzu textů navzdory tomu, že každý z nich se zabývá jiným empirickým materiálem. Digitální historik může využít nástrojů vyvíjených počítačovými lingvisty. Religionista i archeolog mohou k různým účelům využívat stejný geoinformační systém. Struktura knihy se tedy snaží sledovat tyto na první pohled skryté souvislosti, které mnohdy vycházejí spíše z povahy analyzovaných dat než z oborové příslušnosti.

První sekce, *Zdroje a nástroje*, obsahuje příspěvky, jež spojuje zaměření na tvorbu dat a nástrojů pro jejich zpřístupňování. Tyto datové platformy např. v podobě databází, korpusů či slovníků jsou označovány jako výzkumné infrastruktury, neboť podobně jako třeba laboratorní přístroje fungují jako základna pro další výzkum. Hranice mezi infrastrukturou pro výzkum a výzkumem samotným však nemusí být ostrá, jelikož vznik infrastruktur se často odvíjí od badatelské práce.

Kapitola o moderním popisu staročeské morfologie čtenáře upozorňuje na množství problémů, se kterými se musí potýkat jazykovědci při vymýšlení optimálních datových modelů pro starou češtinu. Zatímco obvykle uživatelé vidí pouze výsledné aplikace, kolektiv autorů vedený Borisem Lehečkou pro ně odkrývá množství dílčích kroků a designových rozhodnutí, které předchází finální prezentaci a poskytování dat.

Další kapitoly kromě způsobů budování infrastruktur demonstrují také možnosti jejich výzkumného využití. Vojtěch Malínek popisuje přerod souboru bibliografických záznamů o české literatuře do datové podoby a testuje nové možnosti, které digitalizovaná bibliografická data nabízejí. V elektronické podobě se využití bibliografické databáze nemusí omezovat pouze na vyhledávání primární a sekundární literatury, ale má bezprostřední význam i pro kvantitativně podložené studie o generační skladbě české spisovatelské obce nebo o ohlasu literárních děl.

Podobné propojení infrastruktur s výzkumem demonstruje Zuzana Laubeová se svými spoluautory, kteří vyzdvihují přínos digitálních korpusů pro lingvistické bádání. Počítačově zpracované korpusy lingvistům umožňují zkoumat řadu jazykových jevů, a to na širokém spektru uživatelů jazyka a v různých kontextech. Příkladem jazykového vývoje mohou být změny ve slovnědruhové klasifi-

kaci v mluvené češtině. Korpusová lingvistika otevírá prostor i pro studie, v nichž jazykové jevy mohou indikovat společenské procesy.

Analýza textových dokumentů vystupuje jako součást výzkumu v mnoha oborech. Dokument z hlediska digitálních humanitních a sociálních věd může nabírat různou podobu. Badatelé mohou analyzovat otevřené otázky v sociologických nebo psychologických dotaznících stejně jako středověké rukopisy nebo literární texty od poezie po romány. Otázce využití počítačů v analýze dokumentů se věnuje oddíl *Strojové čtení*.

Martin Hájek pomocí počítačových skriptů mapuje sémantický prostor biografických příběhů zaznamenaných v sociologických rozhovorech. Jeho kapitola se věnuje různým postupům pro rozbor vyprávění a formálnímu zachycení této všudypřítomné diskursivní formy. Ukázková studie odhaluje odlišně vnímané životní zkušenosti mužských a ženských vypravěčů, kteří vzpomínají na dobu státního socialismu a jejichž výpovědi jsou uloženy v orálněhistorických archivech.

Uplatněním digitálních metod v literární vědě se věnuje Richard Změlík. Právě z oblasti literární vědy se rekrutují již klasické výzkumy např. Franca Morettiho, jehož koncept vzdáleného či distančního čtení silně ovlivnil veřejnou představu o digitálních humanitních vědách, ale také vyvolal řadu polemik o schopnosti počítačů podílet se na interpretaci literárních textů. Čtenářům příspěvek nabízí praktickou ukázku toho, jak lze využít metody korpusové lingvistiky k daty řízenému výkladu sémantiky barevného spektra, s nímž pracují spisovatelé při vytváření fikčních světů. Práce s vícero korpusy a textovými metadaty (tedy daty, kterými popisujeme data) umožňuje také porovnávat autorsky podmíněnou práci s barvami nejen mezi jednotlivými spisovateli, ale i v kontrastu s kognitivními významy barev, které jsou zachyceny ve velkých jazykových korpusech.

Sémantické sítě jsou matice, které kvantifikují vazbu mezi slovy na základě jejich spoluvýskytu v definovaném textovém úseku. Za pomoci těchto sítí extrahovaných z korpusu antických textů se Vojtěch Kaše snaží sledovat frekvence a významové okolí morálních termínů, aby posoudil vzájemně konkurenční hypotézy, jež o kulturní evoluci náboženství postulují religionisté. Výsledky naznačují, že představy o rozvoji moralizujících náboženství v souvislosti s nástupem křesťanství nemají velkou empirickou oporu v dochovaných

textech, které místo toho dokládají silné propojení morálky a náboženství již ve starších obdobích.

Editor monografie se věnuje dalšímu typu dokumentů, kterým jsou odborné vědecké texty. Jeho kapitola hledá vztah mezi genderem autorů, počtem autorů i citovaností sociologických článků na straně jedné a tématem článků na straně druhé. Analýza se opírá o modelování témat, což je technika, kterou digitální humanitní vědy často používají pro automatické zjišťování námětu textů. Tematický model odhaluje, že vnitřní rozrůzněnost sociologie není náhodná, ale odvíjí se od charakteristik autorů a jejich způsobu práce. Nalezená asociace mezi citovaností a tématy článků také upozorňuje na nedostatečnou vypovídací hodnotu agregátních časopiseckých statistik, jež jsou v hodnocení vědy často užívanými metrikami.

Dějiny v datech je název oddílu zaměřeného na příspěvky využívající digitální metody pro studium historie. Michal Frankl uvažuje o rozšíření perspektivy, kterou historikům holokaustu nabídla blogová platforma EHRI Document Blog. Schopnost internetového média kombinovat různé typy dat, začleňovat do textu nejen obrazy, ale i interaktivní mapy a jiné vizualizace stimuluje nové formy historiografického bádání. Flexibilita blogu vede k intenzivnější komunikaci mezi historiky a poskytuje prostor pro prezentaci dílčích pokroků v bádání, o nichž lze dále diskutovat. Zpětná vazba vede k cizelování publikovaných studií předtím, než jsou závěry publikovány v tradičních časopisech. Provoz blogu však také přináší technické výzvy, jež vyžadují, aby badatelé reflektovali povahu dat, se kterými pracují.

Přehled několika způsobů, jimiž mohou digitální metody přispět k medievistickému výzkumu, sepsali Jan Škvrňák a Adam Mertel. Historická data lze stejně jako ta současná podrobit zkoumání pomocí geografických informačních systémů nebo sociálních sítí. Propojení obou perspektiv pak dovoluje sledovat takové výzkumné otázky, jako zda geografická blízkost ovlivňovala sociální vazby středověké šlechty nebo jak pozice vládce v sociální síti aristokracie indikuje politický smír. Na jednu stranu mezi historickými a současnými daty neexistuje zásadní rozdíl z hlediska metod, kterými mohou být analyzována. Na straně druhé je při práci s historickými daty potřeba brát pečlivě v úvahu jejich fragmentárnost a zkreslení vyplývající z jejich nenáhodné dostupnosti a zachovalosti. Autoři

proto dávají prostor i úvahám o tom, jaké okolnosti musí mít badatelé na paměti, když historické prameny transformují do podoby digitálních historických dat.

Dvě kapitoly zaměřené na výzkum minulosti pocházejí z oboru archeologie. V tomto oboru se digitální obrat v rámci humanitních a sociálních věd projevuje možná nejmarkantněji. Přispívá k tomu skutečnost, že archeologové pracují v terénu, a tak nové technologie hojně uplatňují nejen při analýze, nýbrž i při sběru dat. Příklady těchto činností přinášejí Martin Kuna a David Novák, kteří čtenářům přibližují, jak digitální archeologie přetváří krajinu do počítačem zpracovatelného objektu. Základním datovým modelem pro digitalizaci krajiny je geografický informační systém, který eviduje prostorová data v různých vrstvách a rozměrech a v němž mohou být zachyceny výskyty archeologických lokalit a jakékoli další body zájmu. Prostorová data mohou být obohacena i leteckým laserovým skenováním terénu, díky němuž lze analýzou výškopisných modelů v krajině hledat dosud neznámé stopy minulosti. Archeologové tak mohou získávat poznatky o sídlištních strukturách, o prostorovém rozptýlení historických populací nebo o ekonomických systémech v různých obdobích.

Trojrozměrnými vizualizacemi archeologických dat se zabývá kolektiv autorů v čele s Jiřím Ungerem. Grafická znázornění archeologických objektů mohou sahat od drobných předmětů až po celá města a využití 3D modelů se v archeologii rychle rozšířilo. Pokročilé modely vyžadují práci vysoce odborných a interdisciplinárních týmů, které zahrnují i počítačové vědce a umělce. Ani technická sofistikovanost však nepřinesla, jak autoři upozorňují, kýžený objem nových poznatků, jež by bylo možné z vizualizací odvodit. Zcela nezastupitelné místo však počítačem vykreslené modely získaly pro komunikaci archeologického poznání širší veřejnosti. I přes své ukotvení v archeologii nese tato kapitola důležité poselství i pro další obory, neboť se nebrání epistemologickým úvahám o obecném významu vizualizací pro vědecký výzkum. Počítačová grafika dává humanitním a sociálním vědcům k dispozici mnoho nástrojů k prezentaci jejich dat a může posilovat snahy o popularizaci jejich vědecké práce. Zdánlivá samozřejmost vizualizací však skýtá i nástrahy v situacích, kdy data, nad nimiž vizualizace stojí, jednoznačná zdaleka nejsou. Proto se autoři věnují i otázce vizualizace výzkumné nejistoty.

V posledním oddílu jsou shromážděny příspěvky, jejichž náplň jde nad rámec reflexe digitálního obratu v odborných disciplínách. Reagují na aktuální svět, který sám v překotném tempu prochází vlastní digitalizací. Oddíl *Digitální současnost* otevírá kapitola Dina Numerata o zdraví a zdravotnictví v digitální éře. Podobně jako jiné oblasti života i péče o zdraví se více a více uchyluje k novým technologiím. Povinností sociologů je tyto proměny monitorovat a Dino Numerato předkládá výběrový, a přesto rozsáhlý výčet společenských pojmů, které pod vlivem digitalizace zdraví nabírají zcela nový obsah. Tradiční vztah mezi medicínskými experty a pacienty narušuje dostupnost informací na internetu, který pacienti navíc využívají ke sdílení svých osobních zkušeností. Informovanost, důvěra a kontrola mezi pacienty a experty hledají novou rovnováhu na pozadí digitalizace samotné medicíny, která díky velkým souborům dat, bioinformatice a strojovému učení testuje nové způsoby péče a léčby jednotlivců i celých populací.

Rozpad veřejnosti do polarizovaných názorových skupin patří k nejvýraznějším politickým fenoménům digitální současnosti: Podílí se na něm sociální sítě ve významu platforem, na nichž se ve virtuálním prostoru setkávájí a sebeprezentují jednotlivci. Termín sociální sítě však také označuje odborný koncept, s jehož pomocí sociální i humanitní vědci zkoumají vzájemné vazby různých společenských aktérů. Matouš Pilnáček a jeho kolegové aplikují analýzu sociálních sítí na výzkum politické polarizace tématu migrace v internetových diskusích mezi čtenáři zpravodajského portálu. Kapitola detailně vysvětluje, co obnáší analytický přístup založený na síťové perspektivě. Aby hrubá data z čtenářských komentářů převedli do relační formy, museli v nich výzkumníci indentifikovat vztahy pomocí hodnocení komentářů, která si diskutující mezi sebou udělují. Po automatizovaném sběru dat a jejich transformaci mohou již výzkumníci přikročit k využití grafových statistik, aby vyhodnotili míru názorové polarizace na diskusním fóru. Výsledky napovídají, že v diskusích o migraci neexistují dvě, nýbrž tři skupiny diskutujících, přičemž hlavním impulsem pro polarizaci debaty jsou komentáře postavené na aktivně protiimigrační rétorice.

Nina Wančová shrnuje poznatky o nových způsobech prezentování muzejních artefaktů v digitální podobě. Virtuální expozice neposkytují pouze další výstavní prostor, ale vedou k nekonvečnímu

promýšlení funkcí a poslání muzejních sbírek. Digitalizované exponáty a jejich popisy v podobě metadat už nejsou pouze výsledkem badatelské a kurátorské práce, ale mohou se stát i infrastrukturami pro další výzkum, jak o tom primárně pojednávaly úvodní příspěvky této knihy. Kapitola Niny Wančové mimo jiné podává zprávu o zkušenostech s vývojem a uplatněním českého softwarového nástroje pro digitální muzea INDIHU Exhibition. Při vývoji tohoto softwaru bylo nutné nalézt optimální řešení zohledňující jak kurátorské cíle, tak uživatelská očekávání. Výsledná asembláž multimediálních formátů nenahrazuje virtuálním návštěvníkům fyzickou muzejní expozici, ale poskytuje jim svébytný styl setkávání se se sbírkovými předměty.

Závěrečná studie Silvie Cinkové a dalších autorů završuje směřování knihy k orientační mapě českých digitálních humanitních a sociálních věd. Po metodologické stránce se příspěvek opírá o latentní sémantickou analýzu, která podobně jako modelování témat třídí dokumenty podle obsahové příbuznosti. Jednotlivé kroky této metody jsou detailně popsány a nastíněny jsou i cesty jejího rozvíjení dodatečným zpracováním dat pomocí shlukové a síťové analýzy. Náplní explorativního výzkumu jsou studijní katalogy českých vysokých škol, v nichž autoři hledají předměty, jejichž výuka může být relevantní pro digitální humanitní vědy. Ačkoliv studie metodologicky doplňuje spíše sekci *Strojové čtení*, její substantivní téma představuje navýsost aktuální přehled o přítomnosti výuky digitálních metod v českém terciárním vzdělávání. Závěry autorů jsou v zásadě optimistické – přestože předmětů, které by se explicitně věnovaly digitálním humanitním vědám, je poskrovnu, pro dnešní studenty již existuje poměrně široká nabídka kurzů digitálních dovedností, které lze v digitálním výzkumu uplatnit. Poslední kapitola tak přináší empirickou oporu pro leitmotiv celé knihy, kterým je přesvědčení, že bez ohledu na to, zda si digitální humanitní vědy a komputační sociální vědy získají v Česku robustní institucionální zázemí, digitální obrat nevyhnutelně rezonuje napříč obory.

Zvolené osy – infrastruktury, texty, historická data a současné problémy –, kolem nichž se jednotlivé příspěvky soustředí, zachycují jen vybrané rozměry mnohorozměrného procesu, kterým je přejímání digitálních metod a jejich aplikování ve výzkumu v humanitních a sociálních vědách. Kromě již předestřené oborové příslušnosti

mohou být příspěvky propojeny i metodologickými kritérii. Čtenáři si mohou všimnout, že některé příspěvky spojuje zájem o analýzu sociálních sítí, jiné využívají různé varianty maticových operací, frekvenční statistiky, vizualizace nebo spoléhají na geoinformační systémy či přehledy literatury. Knihu lze tedy číst vícero způsoby a sestavit si vlastní řazení příspěvků.

I přes svůj široký záběr a tvárnost vlastní organizace zanechává kniha mnoho slepých míst. Určitě v ní schází příspěvky psychologické, ekonomické, antropologické, muzikologické, stylometrické a další typy příspěvků. Cílem knihy však není poskytnout vyčerpávající přehled všech metod a oborů, nýbrž pouze nastínit komplexní charakter digitálního obratu v českých humanitních a sociálních vědách. Pokud budeme vnímat digitální obrat ve výzkumu v celé jeho složitosti a mnohostrannosti, lépe pochopíme i potíže s jeho nestabilními teoretickými definicemi. Ponechme tentokrát velké teorie stranou a pokusme se přemýšlet nad tím, zda ve svém portfoliu výzkumných otázek nemáme takové, kde bychom pomoc digitálních metod uvítali. Přednosti i limity těchto metod se nejvýrazněji zhmotní právě při jejich kritickém použití. Je-li důkaz pudinku v jídle, pak vám při četbě přeji dobrou chuť.

ZDROJE A NÁSTROJE

MOŽNOSTI VYUŽITÍ BIBLIOGRAFICKÝCH DATASETŮ JAKO ZDROJŮ PRO KVANTITATIVNÍ VÝZKUM V HUMANITNÍCH VĚDÁCH: PŘÍPAD ČESKÉ LITERÁRNÍ BIBLIOGRAFIE

Vojtěch Malínek

Bibliografie fungují po dlouhá léta jako jeden ze základních zdrojů vědeckých informací v humanitních vědách.[1] S nástupem digitálních technologií a zejména s rychlým rozvojem příslušných softwarových nástrojů však získávají nové možnosti využití: přestávají být pouhým zdrojem referencí o existující literatuře či primárních textech, ale nově začínají být využívány jako komplexní datasety, které mohou poskytnout podkladová data pro kvantitativní či statistický výzkum nejen v oblasti humanitních oborů, jako je v našem konkrétním případě literární věda.

Literární věda či jí příbuzné obory v tomto ohledu zaujímají specifickou pozici i mezi humanitními disciplínami jako takovými: poněvadž je základním předmětem jejího výzkumu text nebo jeho fyzická reprezentace (kniha, článek či v poslední době samozřejmě též jejich digitální verze), mohou bibliografické databáze sloužit jako vhodný nástroj pro mapování, deskripci a analýzu zvoleného literárního či libovolně jinak definovaného pole. Na základě bibliografických dat totiž lze sledovat nejen „pouze" cirkulaci či recepci určitého textu, ale lze mapovat i složitější vztahy a vazby mezi jednotlivými účastníky literárního dění, ať již jde o jednotlivé osoby, umělecké skupiny, korporace, periodika, nakladatele a nakladatelství atd., nebo jednotlivá díla, témata či výzkumné směry. V mnoha

1 Studie vznikla jako výstup projektu Česká literární bibliografie (LM2018136), podpořeného Ministerstvem školství, mládeže a tělovýchovy České republiky v rámci jeho aktivit na podporu výzkumných infrastruktur.

ohledech se kvantitativní výzkum bibliografických dat může inspirovat existujícími bibliometrickými postupy a nástroji, zároveň však otevírá možnost analýzy značně širšího spektra badatelských otázek a problémů, které nemusejí primárně souviset s mapováním vědeckého provozu a citační metrikou.

Nástup kvantitativního výzkumu v literární vědě je spojován především se jménem Franca Morettiho, jehož některé publikace byly přeloženy též do češtiny,[2] či jeho spolupracovníků ze Stanford University, resp. Stanford Literary Lab (zejména Mathew Lee Jockers).[3] Byť Morettiho práce byly v poslední době opakovaně kritizovány především kvůli nedostupnosti analyzovaných datových souborů a s tím související obtížné verifikovatelnosti jeho závěrů (zejména K. Bode, zde odkazy na další literaturu),[4] nelze popírat, že kvantitativnímu a statistickému výzkumu v oboru literární vědy či šířeji humanitních věd obecně významně napomohly otevřít cestu. Morettiho impulsů, dále stimulovaných rychlým rozvojem výpočetní techniky a softwarových nástrojů, se chopila celá řada následovníků, kteří výsledky jeho bádání dále rozvíjejí či inovativně využívají pro řešení svébytných badatelských problémů zasahujících řadu nejrůznějších oblastí literární vědy. Tento trend je zřetelný, i pokud jde o vytěžování bibliografických dat, která jsou předmětem této stati a kterým se v posledních několika letech začíná dostávat zasloužené pozornosti. Jmenujme v této souvislosti alespoň Macieje Maryla a jeho kvantitativní výzkum polské literatury po roce 1989[5] či Roberta Pétera a jeho bádání o anglické literatuře 18. století a zejména návazný projekt, v jehož rámci vznikl software AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts), na jehož platformě jsou vyvíjeny nástroje právě pro kvantitativní analýzy textových a bibliografických dat.[6] Zřejmě nejsystematičtěji je v současnosti kvantitativní výzkum bibliografických dat provozován při Helsinki Computational History Group (Mikko Tolonen, Leo Lahti

---

2 F. Moretti: *Grafy, mapy, stromy*; či F. Moretti: *Distant Reading*.

3 M. L. Jockers: *Macroanalysis*. Sledování aktivity Stanford Literary Lab je možné zejména prostřednictvím série online publikovaných tzv. „pamfletů", viz *Pamphlets*.

4 K. Bode: The Equivalence of „Close" and „Distant" Reading, s. 77–106.

5 M. Maryl: Literary Transition in Poland Viewed Through Bibliographical Data (1989–2000). Textová podoba studie je momentálně v tisku – srov. M. Maryl: Operationalising the Change, (v tisku).

6 R. Péter – Z. Szántó – J. Seres – V. Bilicki – G. Berend: AVOBMAT, s. 43–55.

a spol.). Tento tým provádí systematický komparativní výzkum založený na analýze bibliografických datasetů finské a švédské národní knihovny, English Short Title Catalogue a Heritage of Printed Book Database.[7] Ve své studii z počátku roku 2019 finští badatelé své výzkumné metody označují přímo jako „bibliographic data science" a poukazují na značný potenciál, který takto vymezená disciplína skrývá.[8]

V českém prostředí i české literární vědě je zatím kvantitativní výzkum bibliografických dat v samotných počátcích. Dílčí studie k otázce cirkulace menších národních literatur v Evropě připravil Ondřej Vimr.[9] Možnosti kvantitativně zaměřeného výzkumu jsou logicky prověřovány při výzkumné infrastruktuře Česká literární bibliografie (dále též jako ČLB). Cílem tohoto textu je právě na jejím příkladu ukázat, jaké možnosti pro kvantitativní a statistický výzkum bibliografických dat její zdroje nabízejí.

SOUČASNÝ STAV ČESKÉ LITERÁRNÍ BIBLIOGRAFIE

Bibliografické pracoviště, které by zpracovávalo oborovou analytickou bibliografii, bylo plánováno jako integrální součást akademického Ústavu pro českou literaturu již v přípravných dokumentech pro jeho založení (1947).[10] Za více než sedmdesát let existence ČLB při ní vznikl komplex bibliografických informačních zdrojů a souvisejících databází (biografická báze České literární osobnosti, báze literárních cen, excerpovaných časopisů atp.), které souvisle zpracovávají materiály od počátků novodobého českého písemnictví v poslední třetině 18. století až po nejaktuálnější současnost. Svými parametry (počet záznamů, rozsah zpracovaného období, metodická úroveň, rychlost a aktuálnost zpracování i neomezená přístupnost) patří ČLB k předním pracovištím svého druhu minimálně v evropském měřítku. Tuto její pozici lze mj. doložit i skutečností, že s platností od roku 2016 byla ČLB zařazena na Cestovní mapu velkých infrastruktur ČR jako

7 Více o aktivitách skupiny zde: https://www.helsinki.fi/en/researchgroups/computational-history.
8 L. Lahti et al.: Bibliographic Data Science and the History of the Book 1500–1800, s. 5–23.
9 O. Vimr: Exploring Big Data Approaches to International Literary Flows (v tisku).
10 *11. 6. 1947. Založení Ústavu pro českou literaturu v dokumentech*, s. 10.

jedna z pěti infrastruktur pro humanitní vědy v ČR a tuto pozici pak následně obhájila i pro léta 2020–2022.[11]

Aktivity výzkumné infrastruktury Česká literární bibliografie v posledním desetiletí významně překročily „pouhé" rutinní zpracování oborové bibliografické databáze a zřetelně směřují k vývoji a adaptaci softwarových nástrojů a utilit pro práci s bibliografickými daty a výzkumu těchto datasetů. Proto byly její snahy napřeny zejména k maximální homogenizaci a unifikaci jejích z historických důvodů poměrně disparátních dat na společném datovém standardu, kterým se stal mezinárodní knihovnický výměnný formát MARC21 pro bibliografické záznamy, resp. MARC21 pro autority pro záznamy autoritní/biografické. MARC21 je v současnosti univerzálně používaným standardem nejen v českém knihovnictví, ale též v řadě dalších zemí zejména v Evropě a Severní Americe. Jde o formát definovaný již před dvaceti lety, který navazuje na starší MARCové formáty, jejichž kořeny můžeme hledat dokonce v letech šedesátých. Byť jsou okolo budoucnosti a vhodnosti MARC21 vedeny obsáhlé diskuse, prozatím se jej v katalogizační praxi nepodařilo uspokojivě nahradit formátem jiným.

Univerzální využitelnost formátu MARC21 i existence sítí umožňující výměnu informací na něm založených podmínily rozhodnutí, aby do tohoto standardu byly převedeny i jednotlivé databáze České literární bibliografie a Střediska literárněvědných informací Ústavu pro českou literaturu AV ČR.[12] S rozsáhlými konverzemi bylo započato roku 2012, po dokončení konverze knihovních katalogů (2014) plynule navázala konverze bibliografických bází, ukončená v závěru roku 2016, a konverze biografické báze České literární osobnosti.

Chronologicky i historicky nejstarší vrstvu dat ČLB představuje tzv. Retrospektivní bibliografie české literatury. Tato bibliografie byla ve formě lístkové kartotéky pořizována zejména v 50. a 60. letech minulého století, kdy na její přípravě pracoval tým čítající až 40 osob, který ročně vyprodukoval až 100 000 excerpčních lístků. Tato bibliografie byla koncipována jako generální článková bibliografie evidující materiály k české i světové literatuře a literární kul-

---

11 *Cestovní mapa České republiky velkých infrastruktur pro výzkum, experimentální vývoj a inovace pro léta 2016 až 2022*, s. 99.

12 M. Topor: Dějiny jednoho oddělení, s. 217–234.

tuře z periodického tisku v českých zemích vydávaného od počátků novodobého českého písemnictví v poslední třetině 18. století (nejstarší excerpta zpracovávají články z roku 1771) až do konce druhé světové války.[13] Po dílčím útlumu v 70. a 80. letech pokračovalo zpracování Retrospektivní bibliografie i po roce 1989 komplexní redakcí celé kartotéky a souběžně byly zahájeny úvahy o jejím převodu do elektronické podoby, komplikovaném zejména značným rozsahem celé kartotéky (přes 1,6 mil. lístků).

Převod se podařilo realizovat v letech 2009–2012, kdy byla celá kartotéka digitalizována a zpřístupněna prostřednictvím speciálně vyvinutého systému RETROBI.[14] V průběhu projektu „Digitalizace lístkové kartotéky Retrospektivní bibliografie české literatury" byly jednotlivé lístky seskenovány a následně též obohaceny o OCR přepisy jednotlivých obrázků, v nichž je následně možno fulltextově vyhledávat. Součást aplikace RETROBI pak mj. tvoří podkladová databáze, která kromě evidence digitalizovaných dokumentů umožňuje též zpracování jednotlivých záznamů ve strukturované textové podobě v proprietárním formátu blízkému standardu MARC21. Zároveň software RETROBI disponuje několika nástroji pro úpravu a přepis kartotéčních lístků. Standardně je samozřejmě možné každý jednotlivý záznam ručně rozepsat do jednotlivých polí databázové struktury. Efektivnější se však z následných zkušeností zdají být nástroje pro tzv. hromadnou editaci, které umožňují zvolený atribut přidat hromadně zvolené skupině lístků. Příslušné nástroje přitom též disponují kontrolními mechanismy, které ověřují, zda je do daného pole možné zapsat jen právě jednu hodnotu, nebo hodnot více, popř. zda už příslušná hodnota není v záznamu vyplněna. Skupinu lístků, jimž má být zvolený atribut doplněn, lze přitom definovat různými kritérii, např. pomocí fulltextového vyhledávání v OCR přepisech jednotlivých lístků jako rešerši v databázi nebo jako sekvenci jednotlivých lístků řazených dle pořadí v původní kartotéce.[15]

Intenzivněji využity byly tyto nástroje během řešení grantového projektu INDIHU,[16] v jehož rámci byly testovány možnosti zapo-

---

13 K metodice zpracování Retrospektivní bibliografie srov. E. Macek: *Bibliografie české beletrie a literární vědy* a D. Řehák: K historii a struktuře Retrospektivní bibliografie, s. 387–395.

14 *Retrospektivní bibliografie české literatury 1775–1945*.

15 V. Malínek: Digitalizace lístkové kartotéky Retrospektivní bibliografie české literatury.

16 *INDIHU*.

jení dat Retrospektivní bibliografie do zastřešujícího centrálního vyhledávače v discovery systému VuFind, který by měl z jednoho místa umožnit přístup k obsahu heterogenních databází několika humanitněvědných ústavů Akademie věd ČR zapojených do projektu. Jako společná datová struktura byl přitom zvolen formát Dublin Core. Od roku 2016 tak byla databáze RETROBI obohacena o údaje o hlavním autorovi dokumentu (celkem 529 623 lístků, z toho u 502 905 byl zároveň doplněn i perzistentní identifikátor dané osoby v souboru národních autorit), zdrojovém dokumentu, v němž byl daný článek publikován (792 358 lístků, tj. bezmála 50 % celku), roku vydání (celkem 1 076 533 lístků, tj. přes 67 % kartotéky) a o pole osoba-předmět (592 294 lístků, z toho 449 498 včetně identifikátoru v bázi národních autorit) tak, aby mohly být zpřístupněny ve společném vyhledávači, vyvíjeném během řešení projektu. Následně byl připraven i konverzní skript, který data z RETROBI umožnil převést též do formátu MARC21, čímž došlo k jejich přímému propojení s ostatními bibliografickými bázemi České literární bibliografie.

Návazná bibliografická báze pro období po druhé světové válce, nazývaná podle někdejší řady bibliografických ročenek *Česká literární věda*[17] či pracovně označovaná též jako „současná" bibliografie, pokrývá období od roku 1945 do nejaktuálnější přítomnosti. Existuje již v plně databázové podobě, avšak z historických důvodů je mnohem roztříštěnější než bibliografie retrospektivní, neboť sestává z několika dobou vzniku i databázovou strukturou původně samostatných částí.

Zpracování současné bibliografie v databázové podobě bylo v ÚČL zahájeno roku 1990 v systému ISIS. Vedle aktuální produkce byla do databázové podoby v první polovině 90. let převedena též lístková excerpta z 80. let, připravovaná původně pro publikaci formou tištěné bibliografické ročenky, a databáze českého literárního exilu, kterou v polovině 90. let zpracoval a následně též pod názvem *Česká literatura v exilu 1948–1989* knižně publikoval tehdejší biblio-

17 Jednotlivé ročenky vycházely pod názvem *Česká literární věda. Bohemistika* od roku 1964, kdy vyšel první svazek za rok 1962, až do roku 1992, kdy vyšel poslední svazek pro rok 1980. Na jejich redakci se podílela zejména trojice Emanuel Macek, Boris Mědílek a Věra Vladyková.

graf ÚČL František Knopp.[18] Uskutečněn byl i později zastavený pokus o přepis lístků retrospektivní bibliografie do databázové podoby formou ručního přepisu, během něhož bylo převedeno přes 95 000 lístků.

Tato nejstarší databázová vrstva byla pořízena v relativně jednoduché databázové masce, která měla z dnešního pohledu řadu limitujících omezení: mj. většinou nepočítala s možností zápisu více hodnot do stejného pole či pro věcný popis nabízela jen pole pro zápis osobních jmen, věcných termínů a názvů literárních děl. Řada selekčních prvků tak byla zapisována bez dalšího pouze jako otagované hodnoty do anotace jednotlivého záznamu, v některých případech v jiném pádu než nominativu, což následně problematizovalo možnosti vyhledávání. Pro věcný popis pak byl používán specifický číselný třídník (tzv. tematické skupiny), původně vyvinutý k základnímu rozčlenění záznamů ve starších tištěných bibliografických ročenkách.

Protože původně definovaná databázová struktura se postupem času stále zřetelněji ukazovala jako nedostačující, byla roku 1997 nahrazena novou, která svou strukturou v zásadě odpovídala i současným standardům pro bibliografické databáze. Nová databázová maska umožňovala uložení neomezeného počtu opakovaných hodnot do stejného pole a nabízela detailnější a pročleněnější formulář pro zápis dat. V této nové databázové struktuře byla současná bibliografie následně pořizována až do roku 2012. Vedle zpracování aktuální produkce byly v závěru devadesátých let do databázové podoby ručně přepsány též zmiňované knižní bibliografické ročenky *Česká literární věda*, pokrývající období od roku 1961 do roku 1980.

Systém ISIS, v němž postupně bylo pro literárněvědnou bibliografii pořízeno cca 430 000 záznamů, však s přibývajícími lety jak technologicky, tak morálně přestával vyhovovat aktuálním potřebám, a proto bylo roku 2012 učiněno rozhodnutí o převodu databází Ústavu pro českou literaturu do systému Aleph, fungujícím na již zmiňovaném mezinárodním výměnném formátu MARC21. Konverze bibliografických dat ČLB z ISISu do MARC21 představovala náročnou operaci, která ve výsledku trvala přes dva roky. Významný problém znamenalo už jen namapování tří základních databázových

18 F. Knopp: *Česká literatura v exilu 1948–1989*.

formátů (kniha/část knihy/článek) ve dvou různých databázových strukturách (starší z roku 1990 a mladší z roku 1997) na jednotný formát MARC21. Zároveň ale konverze znamenala unikátní příležitost pro vyčištění, unifikaci a obohacení dat o některé potřebné údaje. Velmi náročné bylo zejména jednocení historicky odlišných způsobů věcného popisu (zmiňované tematické skupiny, MDT a klíčová slova) na jednotnou základnu, kterou se stal soubor věcných termínů národních autorit, resp. slovně vyjádřená hesla. Operace předpokládala vzájemné namapování jednotlivých klasifikačních systémů a v případě staršího databázového formátu též náročné a z velké části ručně provedené rozdělení nijak nečleněného rejstříku tagovaných hodnot z anotace o cca 60 000 položek do jednotlivých logických kategorií věcného popisu (osoby, korporace, geografické termíny atp.).

Komplikovaný problém představoval též převod údajů pro bibliografickou citaci, neboť struktura dat v ISIS zapisovala každou její část do samostatného pole, kdežto MARC21 ji spojuje do pole jediného – množství možných variant struktury těchto citací předpokládalo vypracování podrobných pravidel postupu konverze, která nutně musela počítat s řadou dílčích pravidel a výjimek. Největším úspěchem, kterého se podařilo v průběhu konverze dosáhnout, však nesporně bylo propojení dat ČLB se souborem národních autorit a s tím související doplnění perzistentních identifikátorů k jednotlivým selekčním termínům. Byť byla kritéria propojení nastavena velmi defenzivně (absolutní shoda záhlaví, resp. absolutní shoda jména osoby, pokud je toto včetně veškerých evidovaných variant v souboru národních autorit jedinečné), dosáhla celková úspěšnost propojení 55,8 %. Vedle toho byl do jednotlivých záznamů doplněn např. i kód ISSN pro jednoznačné odlišení stejnojmenných časopisů. Veškerá historická data ze systému ISIS pak byla na počátku roku 2017 v relativně jednotné struktuře přesunuta do systému Aleph.

V něm začala být již od poloviny roku 2012 zpracovávána současná bibliografie a díky zisku několika grantových projektů i další datové celky: v letech 2012–2015 bibliografie období 1945–1960, jejíž vypracování konečně propojilo oba hlavní databázové soubory ČLB v jeden souvislý a chronologicky propojený celek; od roku 2016 bylo zahájeno zpracování analytické bibliografie literárního samizdatu a konečně v polovině roku 2017 začala být zpracovávána též bibliografie českého literárního internetu. V závěru roku 2021 tak databáze

České literární bibliografie obsahovaly již přes 660 000 databázových bibliografických záznamů a dalších více než 1,6 milionu excerpčních lístků v bibliografii retrospektivní.

Úspěšnost celé konverze lze podtrhnout konstatováním, že od momentu importu dat do ostré databáze na přelomu let 2016 a 2017 až do současnosti nebyla nalezena žádná zásadní chyba, která by si vynutila potřebu opravy konvertovaných dat. Data ČLB jsou samozřejmě dále upravována a redigována již přímo v Alephu, v němž je možno pro tyto účely využít jednotných rejstříků pro celou databázi či specifických nástrojů pro úpravu dat. Nejzávažnější úpravy se přitom momentálně týkají propojování údajů o jednotlivých osobách na soubor národních autorit či harmonizace bibliografických citací. Počet těmito opravami dotčených záznamů se přitom např. jen za rok 2019 pohyboval okolo 137 000.

Harmonizace databázové struktury na bázi mezinárodního výměnného formátu MARC21 nejen významně napomáhá integritě dat a údržbě databáze, ale též usnadňuje orientaci v ní koncovým uživatelům a zejména otevírá možnosti pro využití datasetů ČLB pro kvantitativní a statistický výzkum. Přestože excerpční základna ČLB pro jednotlivá období nebyla definována úplně jednotně (před rokem 1945 byly podchyceny i materiály ke světové literatuře, pro období 1961–2015 chybí záznamy beletrie, knižní publikace jsou podchyceny až od roku 1961), lze na základě jejích dat zodpovědět řadu odborných otázek.

ČLB i proto v zimě 2020 spustila vlastní implementaci discovery systému VuFind, v němž jsou uživatelům nabídnuta veškerá data z jednoho místa. Zároveň byl během projektu OP VVV Česká literární bibliografie – Český literární internet: data, analýzy, výzkum, řešeného v letech 2017–2021 vyvinut tzv. Statistický a analytický modul (SAM), který uživatelům nabízí možnost online přípravy datových analýz ze zdrojů ČLB, ať již formou vizualizací na časové ose, nebo formou přehledových grafů, žebříčků a tabulek.

Hlavní výsledek digitalizace Retrospektivní bibliografie a konverze bibliografie současné pro koncového uživatele nesporně tkví v usnadnění přístupu k datům ČLB a jejich významném zkvalitnění. Pro zpracovatelský tým ČLB jsou však snad ještě důležitější získané zkušenosti se zpracováním a analýzou jejích dat, které během této doby načerpal. Během kontrolních prací bylo totiž potřeba zpracovat

řadu dílčích výpisů a rešerší chybných záznamů jako podklady pro kontroly či následné opravy. Pro tyto účely byly používány pokročilé textové editory typu PsPad či Notepad++, tabulkové procesory či nejnověji též skripty v programovacím jazyku R a Python. Pro vyhledávání dat si pak bylo třeba osvojit práci s regulárními výrazy či jazykem SQL. Věříme přitom, že vyvinuté postupy mohou být dále adaptovány a rozvíjeny a bude je možno využít nejen pro interní kontrolu integrity dat v bázi jako dosud, ale – poněvadž se prováděné operace typově velmi podobají – stejně tak i pro datový výzkum nad zdroji ČLB.

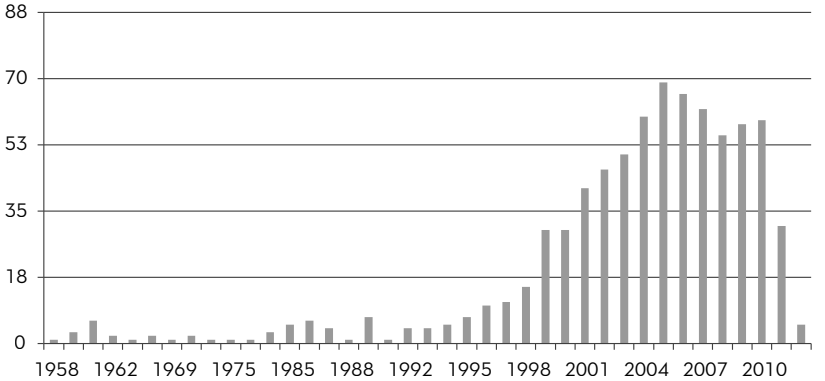
OBECNÉ PROBLÉMY KVANTITATIVNÍHO VÝZKUMU ZALOŽENÉHO NA BIBLIOGRAFICKÝCH DATASETECH

Byť kvantitativní výzkum založený na bibliografických datech může vzbuzovat dojem maximálně exaktního výzkumu s jednoznačnými a zejména verifikovatelnými závěry, při bližší analýze zjistíme, že zdaleka nejde o metodu samospásnou a nelze ji přijímat zcela bez výhrad. V následujících pasážích se proto pokusíme přiblížit hlavní rizika a omezení, s nimiž je potřeba při kvantitativním výzkumu bibliografických dat počítat. Vycházíme přitom stále zejména ze zkušeností vzniklých při zpracování dat České literární bibliografie, většinově by však mělo jít o obecné charakteristiky, které je nutno vzít v potaz při jakékoli kvantitativní analýze bibliografických datasetů.

Jednou z určujících charakteristik jakéhokoli bibliografického datasetu je pochopitelně *vymezení množiny zpracovávaných dat*. Stav a záběr excerpce či katalogizace samozřejmě klíčovým způsobem ovlivňuje, jaká data jsou ve zvoleném datasetu zastoupena: ze samé podstaty bibliografické práce je jasné, že databáze bude vždy odrážet situaci zpracování v určitém konkrétním časovém momentu a množina zpracovaných dokumentů nikdy nebude zcela úplná, popř. že může v datech do budoucna dojít k dílčím posunům (zpracování nových přírůstků, atribuce šifer či pseudonymů, různě motivované korekce či opravy atp.). Nejspolehlivější co do úplnosti podchycených publikací zřejmě mohou být knižní katalogy národních knihoven či přímo národní bibliografické databáze či národní souborné katalogy

z posledních desetiletí, které se mohou opřít o právo povinného výtisku. U bibliografií článkových či analytických je však situace mnohem komplikovanější, poněvadž pro jejich potřeby prakticky není možné zpracovat veškerý existující materiál, a to zejména pro jeho nepřeberné množství. Prakticky každá analytická bibliografie je tak nutně bibliografií výběrovou, je proto třeba si ověřit, jak byl pro vybraná data v průběhu času definovaný excerpční záběr, zda nedošlo k výpadkům excerpce, cenzuře zpracovávaného materiálu, popř. je vhodné též prověřit kvalitu zpracování daného období či konkrétního časopiseckého titulu atp. Nezpracovávala-li ČLB po určité období časopisy z oblasti sci-fi, neposkytne datová analýza jejích zdrojů o sci-fi žádné vypovídající údaje: o sci-fi nebudou v bázi k nalezení žádná data, což není nutně chybou zpracovatelů. Zjednodušeně řečeno: nepřítomnost dat neznamená automaticky jejich neexistenci. Důležité je přitom pochopitelně i oborové zaměření databáze, které ovlivňuje jak výběr zpracovávaného materiálu, tak způsob jeho popisu. Stejný článek popíše jinak bibliografie literární, jinak např. teatrologická a jinak historická. Nejde přitom nutně o chybu, protože každá z bází předpokládá poněkud odlišného uživatele a postupuje dle svébytných zásad.

Neméně důležitým faktorem pak je používaná *struktura dat a normy popisu*. Používaná struktura dat i způsob vyplňování jednotlivých polí se pochopitelně mohly v průběhu času měnit a leckdy se i hojně měnily. Rozdíl pak může být dán i technologicky: záznamy pořízené původně pro lístkovou kartotéku a tištěnou bibliografii budou obvykle obsahovat chudší informace nebo atypické situace nebudou popisovat jednotně, řada údajů navíc mohla zapadnout či mohla být zkreslena během přepisů dat či v průběhu konverzí do databázové podoby (chybné opisy, překlepy atp.). V případě záznamů vzniklých již přímo v elektronické podobě lze naopak spíše počítat s detailnější strukturou a větší mírou unifikace. Ale i v případě databází se pochopitelně může struktura dat v průběhu času měnit (změny katalogizačních pravidel, využití nového pole pro zachycení nově se vyskytujícího typu údajů, přesun zápisu zvoleného typu informace mezi poli, nebo dokonce změna katalogizačního softwaru či datového formátu atp.), byť k této situaci v praxi nedochází příliš často. V případě ČLB je ne zcela jednotná struktura patrná zejména u věcného popisu, jehož normy – jak již bylo zmíněno výše – prošly v průběhu času několika změnami. Ukázat si to můžeme na příkladu



Obr. 1 Počet výskytů deskriptoru „rozbory děl" v ČLB dle jednotlivých let (stav k roku 2021)

formálního deskriptoru „rozbory děl". Ten byl zařazen mezi množinu cca 20 předdefinovaných hodnot pro zápis formy/žánru dokumentu v systému ISIS, proto se v ČLB u materiálů zpracovaných v období 1997–2012 vyskytuje poměrně hojně. A naopak: protože se dané heslo nevyskytuje v souboru národních autorit, nevyskytuje se od roku 2012 ani v datech ČLB. Jeho nepřítomnost však neznamená, že by rozbory děl přestaly vznikat, jen je pro tento typ materiálů využívána jiná hodnota formálního popisu (srov. obr. 1).

Obdobným případem může být i *doba excerpce*, tj. s jakým odstupem od svého vzniku byl záznam o daném textu pořízen. Pokud by byl např. článek publikovaný v roce 1950 zpracován obratem po svém vydání, asi bychom v jeho věcném popisu nenalezli hesla jako „stalinismus", „kult osobnosti" či „literární kánon". Pokud by ale totožný záznam vznikl o sedmdesát let později, asi daná hesla uživatele už tolik nepřekvapí. Z této skutečnosti však nelze dovozovat například to, že by na počátku 50. let procházel v Československu některý z uvedených pojmů odbornou diskusí. V záznamu je přítomen jen díky časovému odstupu excerpce od doby otištění daného textu a proměnám oborového diskursu v mezičase. Naopak, ne příliš často budou zpětně doplňovány a o nové hodnoty věcného popisu aktualizovány záznamy starší.

Zpracování rozsáhlých databázových setů s sebou nese *chyby či odchylky v zápisu dat*. Ty mohou být několika druhů: (a) překlepy

a zjevné omyly; (b) chybná databázová interpunkce; (c) hodnota jiného druhu zapsaná v neodpovídajícím poli; (d) náležitá hodnota zapsaná v jiném (pod)poli; (e) chybějící hodnota. Zejména poslední typ je přitom velmi obtížné, ba přímo nemožné odhalit. Interní nástroje jednotlivých databází jsou totiž s to identifikovat nanejvýš chyby formálního charakteru, věcné omyly lze strojovou kontrolou odhalovat jen velmi omezeně (např. lze odhalit chybnou identifikaci autora článku v momentě, kdy by měl dle zápisu být autorem dokumentu publikovaného ještě před jeho narozením atp.). Tyto chyby by v ideálním případě měla odhalit redakce či průběžné kontroly databáze, nicméně ne všechny takovéto chyby je možné identifikovat bez kontroly s dokumentem v ruce, která by zase zpracování dat neúměrně protahovala. Při jisté mechaničnosti bibliografické práce se sice takovýmto chybám nelze zcela vyhnout, avšak snad jen výjimečně by mělo jít o chyby, které by zásadně zkreslily podkladová data pro kvantitativní analýzu.

Konečná podoba záznamu je samozřejmě z velké části ovlivněna i *individuálním rukopisem jeho zpracovatele*. Zdaleka přitom nelze spoléhat na to, že by rukopis jednotlivých excerptorů byl jednotný. I přes snahu o maximální unifikaci je potřeba počítat s tím, že záznam zpracovává člověk a jeho individuální preference, zájmy, odborné znalosti i míra zkušeností s bibliografickou prací se mohou do konečné podoby záznamu významně promítnout. Bibliografický záznam je vždy výsledkem individuální práce jeho zpracovatele. Prakticky nikdy přitom neexistuje jediný možný a správný zápis. Relativně jednotný by měl být alespoň způsob vyplnění jmenného popisu (údaje o autorech, názvu, zdrojovém dokumentu atp.), neboť jeho údaje jsou většinou objektivně dány, mnohem variantnější však bude popis věcný (zjednodušeně řečeno klíčová slova a popis obsahu dokumentu), který je založen na analýze a interpretaci popisovaného textu, a je tak vždy významně podmíněn osobností zpracovatele záznamu. Individualita rukopisů by měla být oslabována průběžnou redakcí báze či vzájemnými konzultacemi jednotlivých zpracovatelů, avšak ani ty nemohou být univerzálním řešením.

Výraznou metodickou komplikaci pro kvantitativní analýzy bibliografických dat spatřujeme v otázce *disproporce významu jednotlivých podchycených záznamů*. Lákavá jednoznačnost kvantitativní analýzy s sebou totiž přináší určitá zjednodušení. Budeme-li každý záznam

brát jako rovnomocnou položku, musíme si být zároveň vědomi, že takto dáváme stejnou váhu např. rozsáhlé knižní monografii a krátké časové zprávě, studii v prestižním oborovém časopise i textu v zapadlém regionálním periodiku či článku v celostátním deníku s masovým nákladem a publikaci v samizdatovém časopise šířeném pouze v komunitě několika jednotlivců. Badatel by tak měl uvažovat o určitém vážení hodnoty záznamů či redukci přechodných typů, aby výsledek jeho bádání nebyl obdobnými faktory zkreslen. Asi není náležité tvrzení, že jedním z nejvýznamnějších prvorepublikových literárních publicistů byl Josef Trojan, i když mu pro enormní množství otištěných soudniček patří mezi nejčastějšími autory zastoupenými v Retrospektivní bibliografii druhé místo (tab. 1). Nutno však konstatovat, že tyto problémy představují pro kvantitativní bibliografický výzkum otázku budoucnosti, která bude předpokládat úzkou spolupráci s odborníky z oblasti datové vědy, sociologie, statistiky a souvisejících oborů.

Kvantitativní analýzy proto mohou badateli posloužit nikoli jako výsledek či cíl poznání, ale jako prostředek či podpůrný argument pro jeho výzkum, který musí být dále interpretován a verifikován. I při vědomí výše zmíněných rizik se však domnívám, že pro určité typy otázek má kvantitativní výzkum bibliografických dat nezastupitelný význam a může často rozhodujícím způsobem přispět k řešení daného badatelského problému, popř. může předkládané řešení podepřít argumenty, které by „klasickými"/„ne-kvantitativními" metodami nebyly získatelné. Adekvátní interpretace výsledků těchto analýz proto předpokládá poučeného badatele obeznámeného se zkoumaným problémem, nebo v ideálním případě oboustranný dialog producentů dat, kteří znají jejich strukturu, metody zpracování a s tím související potenciálně zkreslující faktory či omezení pro jejich interpretaci, a badatelů, kteří jsou tento pohled schopni doplnit o hlubší odbornou znalost dané problematiky a jsou s to výsledná data analyzovat a interpretovat v odpovídajícím kontextu.[19]

Navzdory výše řečenému vykazují bibliografické sety jako zdroj analýzy nesporné výhody. První z nich je určitě *množství dostupných dat*: bibliografické databáze vznikají longitudinální prací rozsáhlých specializovaných týmů a opětovné zpracování těchto dat pro jedno-

19 V. Malínek – T. Umerle – P. Wciślik: From a Reference Book to Research Data.

Tab. 1 Nejčastěji zastoupení autoři v Retrospektivní bibliografii

Novák, Arne	7 370	Vrchlický, Jaroslav	3 575
Trojan, Josef	6 105	Valenta, Edvard	3 439
Suchý, Lothar	5 327	Hora, Josef	3 332
Brtník, Václav	4 798	Eisner, Pavel	3 166
Vodák, Jindřich	4 558	Herrmann, Ignát	3 146
Píša, Antonín Matěj	4 090	Cháb, Václav	3 142
Bass, Eduard	4 080	Červinka, Vincenc	2 948
Neruda, Jan	3 988	Kuffner, Josef	2 631
Poláček, Karel	3 905	Fischer, Otokar	2 618
Weiner, Richard	3 756	Procházka, František Serafínský	2 589

rázové potřeby konkrétního datového výzkumu ať již jednotlivcem, nebo i jednorázově sestaveným projektovým týmem je v zásadě nemožné. Řada bibliografických databází je dnes již tak rozsáhlá, že se v nich individuální rukopisy rozmývají v rámci většího celku a řada dílčích disproporcí v takto velkém celku zaniká.

Množství dat jde ruku v ruce s *délkou časového záběru*: bibliografická data umožňují provádět longitudinální výzkumy a sledovat vybraný jev na dlouhé časové ose. Většina databázových katalogů oborových bibliografií v elektronické podobě začala vznikat v devadesátých letech minulého století, a dnes tedy nabízí již bezmála 30 let trvající kontinuální řadu. Řada institucí pochopitelně investovala značnou energii do retrospektivní katalogizace a převodu původních analogových dat pro předcházející období (lístkové katalogy, tištěné přehledy) do digitální podoby. Dostupnost a kvalita retrokonvertovaných dat může být samozřejmě s ohledem na řadu faktorů (společenská poptávka a využitelnost, finanční možnosti, technické podmínky atp.) významně rozdílná a zdrojová data nemusejí být zcela souměřitelná, avšak možnost sledovat danou otázku v delším časovém horizontu je pochopitelně lákavá. Data ČLB dnes např. souvisle pokrývají období 250 let a obdobně dlouhé období by bylo možné pokrýt též na základě existujících knihovních databází národního významu (katalogy největších knihoven s právem povinného výtisku, Česká národní bibliografie, Souborný katalog ČR atp.).

Významnou předností bibliografických datasetů je *obohacování záznamů o perzistentní identifikátory*, které umožňují propojovat daný dataset s jiným a kombinovat tak původem různorodé datové celky do větších souborů či klást si rozličné specializované otázky: např. propojením biografických či prosopografických bází a bibliografických datasetů lze získat data pro sociologické analýzy daných dat, propojením dat bibliografických a geografických lze zase sledovat cirkulaci určité osoby či fenoménu v časoprostoru atp.

Konečně je třeba zmínit i dobrou *dostupnost existujících bibliografických datasetů*: jednotlivé databáze jsou často bez omezení přístupné online prostřednictvím online katalogů pro jednotlivé dílčí dotazy. Přístupnost databází jako celku však bývá z technických či právních důvodů omezována např. maximálním počtem záznamů, omezením exportů jen na určitá pole, povinnou registrací, nutností uzavřít smlouvu o využití atp. Na druhou stranu i v otázce zpřístupnění bibliografických dat se prosazuje politika open access a např. finská, polská a od roku 2022 i česká národní knihovna vystavují svá data jako celek ve strojově zpracovatelném formátu k volnému využití.

MARC21 A JEHO MOŽNOSTI PRO KVANTITATIVNÍ VÝZKUM

Většina současných bibliografických databází či knihovních katalogů v ČR je zpracovávána v mezinárodním standardizovaném výměnném formátu MARC21. Nutno ovšem zdůraznit, že jednotnost standardu zdaleka neznamená jednotnost dat, a to ani v rámci téže instituce. Rozdíly v rámci jedné instituce mohou být nejčastěji dány změnou katalogizačních pravidel či formátů na národní úrovni v průběhu času (např. přechod z formátu UNIMARC na MARC21 na počátku nového milénia či naposledy přechod na pravidla RDA, nasazená v českém knihovnictví od roku 2015) či postupným doplňováním a rozvojem existujících pravidel. Výraznější rozdíly pak pochopitelně mohou panovat mezi jednotlivými institucemi a týkají se zejména norem věcného popisu: řada institucí si historicky vytvořila vlastní oborové tezaury či soubory selekčních termínů, které lépe odpovídají individuálním potřebám a zvyklostem toho kterého pracoviště nebo nastavení příslušného katalogizačního softwaru, zároveň jsou

však tyto tezaury navzájem obtížně převoditelné a slučitelné. V řadě případů jsou při jednotlivých databázích ve speciálních polích podchycovány i údaje nad rámec struktury MARC21. Velmi výrazná je tato disparátnost zejména u oborových článkových bibliografií, které často až doposud využívají proprietární oborové systémy věcného popisu založené na schématech zavedených ještě v dobách, kdy byly příslušné bibliografie publikovány formou tištěných přehledů. K dílčím odlišnostem může dojít na úrovni vyplňování jednotlivých polí, zejména pak pokud jde o jejich obligatornost či fakultativnost.

Obecně lze nicméně tvrdit, že zatímco bibliografie knižní jsou v rámci ČR s ohledem na existenci Souborného katalogu ČR, do něhož přispívají prakticky veškeré rozhodující české knihovny, vlivem poměrně striktních kontrol při importu dat relativně unifikované, situace bibliografií článkových je mnohem méně přehledná. Oborové článkové bibliografie jsou totiž z logiky věci pořizovány při oborových pracovištích, zejména ústavech AV ČR či jiných specializovaných institucích, a potřeba zohlednit oborová specifika a zvyklosti je v jejich případě pociťována mnohem silněji. Zároveň pak po zrušení oddělení analytické bibliografie ČR v Národní knihovně schází pracoviště, které by mohlo tyto aktivity koordinovat a metodicky zaštiťovat na národní úrovni. I tak by jeho možnosti při zajišťování aplikace navrhovaných pravidel do provozní praxe jednotlivých institucí byly poměrně omezené a nutně by se zredukovaly pouze na definici základního invariantu, který by s velkou pravděpodobností byl dále lokálně upravován a uzpůsobován.

Řada limitů komplikujících vytěžování bibliografických datasetů pro kvantitativní výzkum je dána přímo standardem MARC21, což je pochopitelné, poněvadž MARC21 je formátem určeným pro katalogizaci v knihovních databázích a nebyl vyvíjen cíleně pro datové analýzy. Primárně dokonce MARCové formáty navazují na katalogizaci knih prostřednictvím kartotéčních lístků, což je dodnes patrné i v datové struktuře MARC21 např. existencí pravidla o oddělování zápisu prvního (pole 100) a druhého a dalšího autora (700). V této souvislosti je třeba zároveň zdůraznit, že vedle samotné datové struktury definované standardem MARC21 mají na výslednou podobu dat vliv i katalogizační pravidla, jejich národní interpretace a zpřesnění či používaný knihovní systém, což uživatel při pohledu zvenčí nemusí vždy rozlišovat. Detailní analýza této problematiky

by vyžadovala samostatnou studii, pokusíme se zde proto poukázat na několik dle našeho mínění nejzávažnějších omezení spojených s formátem MARC21 a jeho aplikací v české praxi.

Obecně je třeba podtrhnout, že *ne všechna pole v MARC21 jsou při zpracování validována*, popř. vybavena perzistentním identifikátorem, což samozřejmě komplikuje odlišení homonymních hesel (např. v nakladatelských údajích). *Pro řadu údajů*, které mohou badatele zajímat, *nemá MARC21 definované pole*, mohou tak být nesystematicky zapisovány do poznámkových polí, popř. v bázi nejsou obsaženy vůbec. Některá pole používají *specifické kódovníky*, které nemusejí být laickému uživateli zcela srozumitelné. V určitých případech zároveň počítá MARC21 v závislosti na některých formálních faktorech s *využitím různých polí pro zápis hodnot stejné kvality*, což se týká mj. zápisu jazyků či zemí vydání a jmen autorů, které se popisují do více různých polí v závislosti na tom, je-li třeba zapsat hodnotu jedinou nebo dvě a více. Pro potřeby článkových bibliografií není uspokojivě řešen zápis bibliografické citace, která je zaznamenávána jako celek do jednoho společného, dále nečleněného podpole.

Problém může přinášet i zpracování údajů o odpovědnosti, z nichž může být obtížně strojově vydělitelná informace o původcích daného dokumentu, obzvláště v případě, kdy se jejich jméno uvedené v dokumentu významněji liší od reprezentativní podoby jména uvedené v autoritní bázi (varianty jmen, šifry, pseudonymy). Ne zcela intuitivní je např. vztah mezi místem a zemí vydání, neboť země vydání bývá uváděna dle situace v momentě katalogizace, tj. při současné katalogizaci starších fondů bude knize vydané roku 1970 v Leningradě přiřazena jako země vydání Rusko, knize vydané v Bratislavě roku 1900 Slovensko či Vratislavi roku 1930 Polsko, což pochopitelně může znesnadňovat datové analýzy.

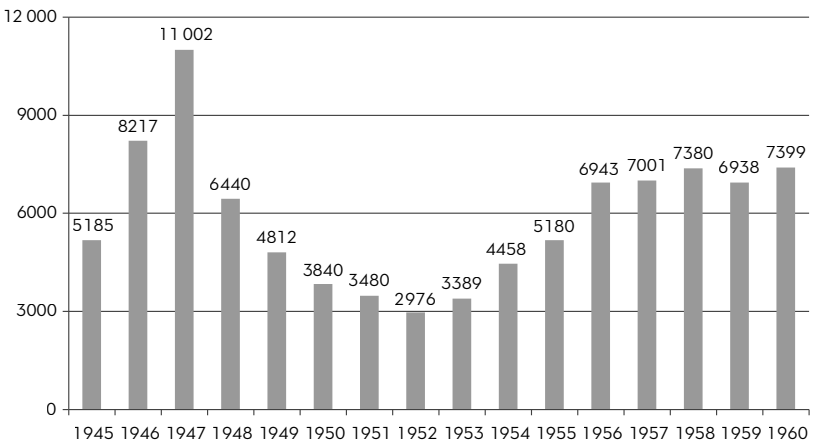
Pravidla RDA pak přinesla několik dalších komplikujících faktorů, daných primárně zásadou „piš, jak vidíš“: je-li daná hodnota v publikaci evidentně zapsána chybně (věcný omyl, překlep), je dle pravidel v této podobě zaznamenána i v databázi, správný údaj je uváděn pouze do poznámky. Pro češtinu jako flektivní jazyk pak toto pravidlo přináší komplikaci např. při zápisu míst vydání: pokud je toto v knize zapsáno v jiném než 1. pádě, stanovují pravidla takovýto zápis i v daném poli bibliografického záznamu, což už může znamenat problémy i při vyhledávání záznamů (rozdíl „Praha“

vs. „V Praze“). Ve většině případů však jde o záležitosti, kterých by si měl zkušenější datový analytik všimnout.

UKÁZKY MOŽNOSTÍ KVANTITATIVNÍ ANALÝZY BIBLIOGRAFICKÝCH DAT[20]

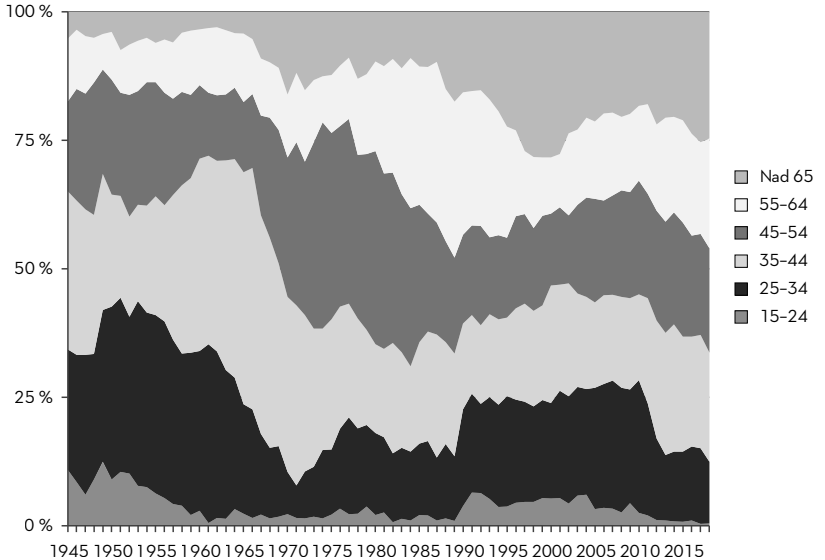
V následující části bude představeno několik praktických ukázek kvantitativní analýzy bibliografických dat a zároveň na těchto příkladech budou ukázány určité limity a možnosti takovéhoto výzkumu.

Nejjednodušším vstupním případem může být kvantifikace určitého jevu na časové ose, což můžeme demonstrovat na příkladu vývoje celkového počtu publikovaných článků dle jednotlivých kalendářních let evidovaných v ČLB pro období 1945–1960 (obr. 2). Tento graf má očekávatelný průběh, který kopíruje běžný narativ o vývoji české literatury v tomto období: po konci druhé světové války dochází k rychlému nárůstu publikovaných textů týkajících se české literatury s jednoznačným vrcholem v roce 1947. Od roku 1948 dochází ke zřetelnému obratu a následuje dramatický pokles s vrcholem v roce 1952, kdy bylo ve srovnání s rokem 1947 publikováno



Obr. 2 Počty záznamů v ČLB 1945–1960

20 Grafy v této kapitole vycházejí ze stavu databáze k 30. červnu 2019.

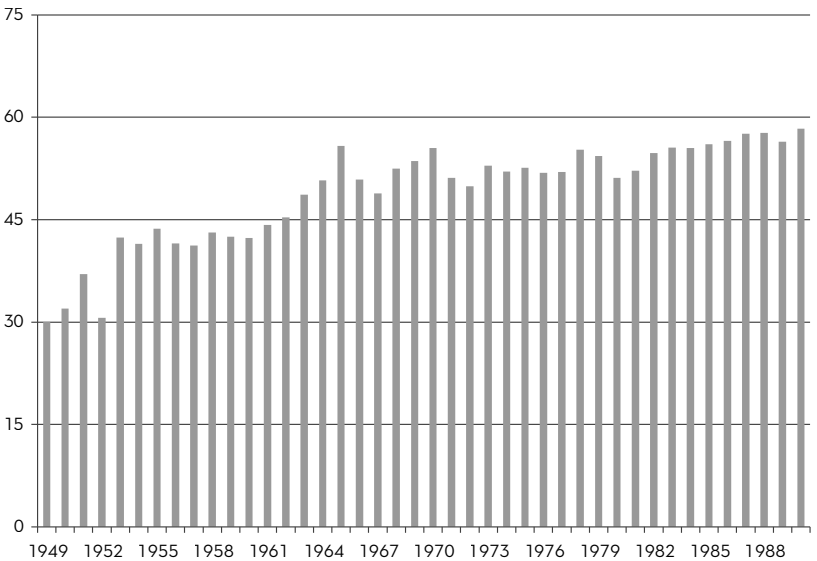


Obr. 3 Podíl autorů dle věku v bázích ČLB v období 1945–2018

pouze 27,05 % článků. Následujícího roku začíná postupné oživení vrcholící v roce 1956, kdy proběhl II. sjezd Svazu československých spisovatelů. V druhé polovině padesátých let pak dochází k určité stagnaci a počet publikovaných článků osciluje kolem hranice 7 000 záznamů.

Bibliografická data zároveň mohou posloužit jako podklad pro výzkum v oblasti literární sociologie. Běžnou součástí údajů o autorech zpracovávaných textů (stejně jako o osobách, jichž se texty týkají) totiž jsou jejich životopisná data, která nám umožňují sledovat věkové složení i u autorů textů podchycených v České literární bibliografii. Na příkladu obrázku 3 si můžeme ukázat podíl zastoupení jednotlivých věkových skupin v České literární bibliografii v průběhu času. Graf dle očekávání významně koreluje s demografickým profilem české populace v daném období, neboť významný podíl zaujímají silné ročníky narozené po obou světových válkách a „Husákovy děti", tj. generace let sedmdesátých.

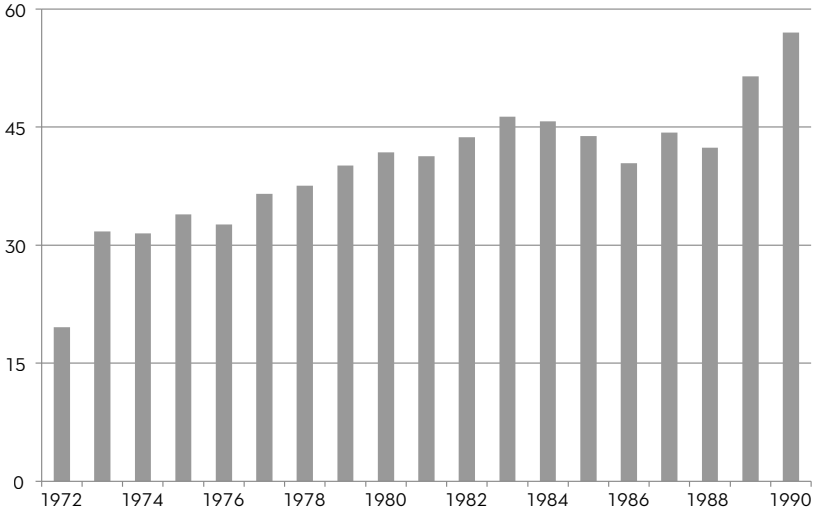
Zajímavější data můžeme získat, pokusíme-li se vytvořit obdobný graf pro alternativní komunikační okruhy, v nichž česká literatura fun-



Obr. 4 Exil – průměrný věk autorů

govala v průběhu druhé poloviny 20. století, tj. pro exil a pro samizdat. Zřetelně se na nich totiž ukazuje generační podmíněnost obou komunikačních okruhů, které „stárnou" spolu s jejich určující věkovou skupinou. U exilové literatury můžeme pozorovat určující roli exilové vlny z roku 1948 a následujících několika let, jejíž vliv oslabila až exilová vlna po roce 1968 (obr. 4). Samizdat se proti tomu ukazuje jako zřetelně spojený s generací narozenou ve 40. letech, jejíž dominance byla přerušena až v druhé polovině let osmdesátých nástupem mladých autorů narozených v letech šedesátých, pro něž se publikování v sílícím samizdatu v dané době už stávalo plnohodnotnou publikační alternativou vůči publikování v oficiálních periodikách (obr. 5).

Velké možnosti pro literární historiky otevírají datové analýzy pochopitelně při sledování recepce či reflexe jednotlivých účastníků literárního života. Jejich prostřednictvím lze plasticky ukázat, nakolik je ten který autor reflektován v průběhu času. Názorný může v tomto ohledu být graf srovnávající intenzitu reflexe J. Wolkera a F. Halase po roce 1945, který – jak ukázal bohužel v nepubliko-



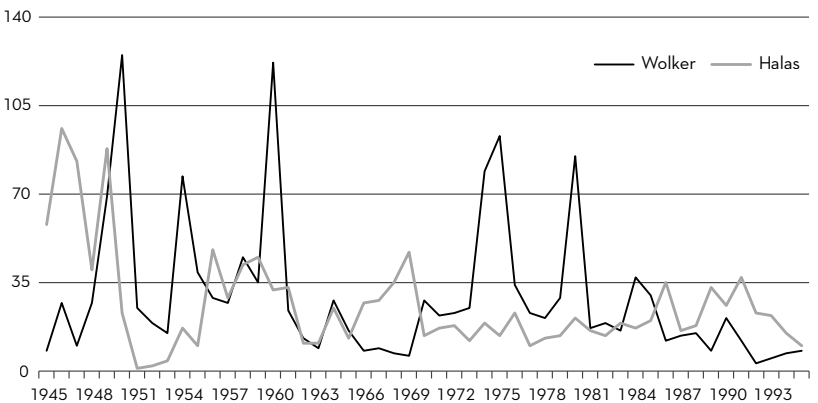
Obr. 5 Samizdat – průměrný věk autorů

vaném konferenčním příspěvku F. A. Podhajský[21] – velmi pregnantně dokládá, jak oba autoři v dané době fungovali jako vzájemné protiklady: v dobách utahování ideologických šroubů (přelom 40. a 50. let, přelom let 60. a 70.) stoupá frekvence textů o J. Wolkerovi, a naopak reflexe díla F. Halase klesá, zatímco Halas proti tomu vystupuje do popředí v ideologicky méně zatížených obdobích druhé poloviny 60. a 80. let, stejně jako v době dočasného uvolnění ideologizace kulturní politiky okolo roku 1956. Zřetelně z grafu vynikne i změna v množství textů o F. Halasovi vydaných za jeho života a v roce jeho smrti (1949) a především jejich dramatický pokles související s Halasovým „uvrhnutím v klatbu" po uveřejnění referátu L. Štolla *Třicet let bojů za českou socialistickou poezii* v lednu 1950, který definoval ideologickou doktrínu kulturní politiky nastupující komunistické diktatury. V případě Wolkerově naopak vidíme, jak silná byla v komunistické epoše potřeba připomínat jeho dílo a tvorbu v letech jeho životních výročí (narozen 1900, zemřel 1924), což je dále posíleno skutečností, že ta „kulatá" z nich často spadají do let razantní proměny kulturněpolitické situace (1950, 1974). Ve srovnání

21 F. A. Podhajský: The Revolution in the Collective Mind?.

s jinými kulatými jubilei též zřetelně vynikne slabší reflexe Wolkerových nedožitých sedmdesátých narozenin v roce 1970: můžeme se jen dohadovat, zda daného roku ještě režim nestihl Wolkerova jubilea využít k propagaci požadované ideologické linie, nebo – což i s ohledem na silné připomenutí 50. výročí Wolkerova úmrtí v roce 1974 vnímáme jako pravděpodobnější – byl Wolker v literárně-kulturním prostředí tentokrát zastíněn výročími jiných národních klasiků: B. Němcové (* 1820; 78 záznamů), K. J. Erbena († 1870; 38 záznamů) a zejména pak J. A. Komenského († 1670; 343 záznamů) a v politicko-ideologickém diskursu pak připomínkami stého výročí narození V. I. Lenina. Po roce 1989 můžeme naopak pozorovat zřetelné opadnutí tohoto protichůdcovství a reflexe obou autorů ztrácejí někdejší vzájemnou podmíněnost. Pro období let 1948–1989 však oba fungují jako určitý lakmusový papírek pro indikaci ideologizace, či naopak demokratizace dobové literární diskuse.

Množství dalších variant různých grafů, závislostí a srovnání je pochopitelně nepřeberné, nehledě k možnostem, které nabízejí analýzy sítí, vizualizace na mapě, analýzy pojmů (wordcloud) atp. Individuálním rozhodnutím každého badatele tak je nejen stanovení množiny zkoumaných dat či formulování badatelské otázky, ale i forma její vizualizace a prezentace. Zopakujeme přitom znovu, že data sama slouží především jako ukazatel trendů a tendencí, bez další kontextuální interpretace však sama o sobě mohou fungovat jen velmi omezeně.



Obr. 6: Srovnání recepce J. Wolker vs. F. Halas 1945–1995