Szkice

Dane badawcze w literaturoznawstwie

Maciej Maryl, Marta Błaszczyńska,
Bartłomiej Szleszyński, Tomasz Umerle

TEKSTY DRUGIE 2021, NR 2, S. 13–44

DOI: 10.18318/td.2021.2.2 | Maciej Maryl – ORCID: 0000-0002-2639-041X
Marta Błaszczyńska – ORCID: 0000-0002-2377-4565
Bartłomiej Szleszyński – ORCID: 0000-0002-7758-1662
Tomasz Umerle – ORCID: 0000-0002-7335-0568

Maciej Maryl, dr, kierownik Centrum Humanistyki Cyfrowej i adiunkt w IBL PAN. WWW: http://maryl.org/. Kontakt: maciej.maryl@ibl.waw.pl

Marta Błaszczyńska, mgr, koordynatorka CHC IBL PAN, starsza specjalistka ds. otwartej nauki, doktorantka w SNS IFiS PAN. Kontakt: marta.blaszczynska@ibl.waw.pl

Bartłomiej Szleszyński, dr hab, prof. IBL PAN, kierownik zespołu Nowej Panoramy Literatury Polskiej, zastępca kierownika CHC IBL PAN. Kontakt: bartlomiej.szleszynski@ibl.waw.pl

Tomasz Umerle, dr, zastępca kierownika Pracowni Bibliografii Bieżącej i adiunkt w IBL PAN. Kontakt: tomasz.umerle@ibl.waw.pl

Wprowadzenie

Dane niepostrzeżenie stały się nieodłącznym elementem naszego życia. Zwykły kojarzyć się nam z czymś jednoznacznym, policzalnym, obiektywnym, jak dane i szukane w zadaniu matematycznym. W życiu codziennym napotykamy dane osobowe, biometryczne, statystyczne, sondażowe, dane o liczbie zachorowań na koronawirusa czy wyniki z okręgów wyborczych. Rozwój technologii cyfrowych sprawił, że każdego dnia my sami i nasze urządzenia – laptopy, smartfony, kamery, opaski, sensory, nadajniki GPS – generujemy ogromne ilości danych, które pieczołowicie zbierają (i często potem sprzedają) ogromne korporacje[1]. Ekspansja danych jest ściśle powiązana z rozwojem rozbudowanych narzędzi analitycznych do przetwarzania tych zbiorów i wykorzystywania ich w planowaniu działań w przeróżnych obszarach naszego życia. Bogate, złożone zbiory danych pozwalają nam budować modele, które coraz chętniej

1 Por. M. Maryl *Subnarracje metadanych*, „Teksty Drugie" 2014 nr 3, s. 179-193.

traktujemy jako reprezentację złożonej egzystencji ludzkiej i, co za tym idzie, klucz do zrozumienia rzeczywistości. Dość nieporadny sposób, w jaki nazywamy to zjawisko – dużo danych, *big data* – obrazuje pewną nieśmiałość, z jaką się z nim mierzymy.

A może to nie tyle nieśmiałość, ile poczucie swoistego zachwytu, zachłyśnięcia się rozległością i szczegółowością zbiorów danych, które William Davies opisuje za historykiem sztuki Julianem Sattalbrasem jako „wzniosłość danych" (*data sublime*)? Wzniosłością nazywa to swoiste oszołomienie powszechnością, totalnością i złożonością danych, które jednocześnie niepokoją i ekscytują[2]. I choć humanistyka jest zazwyczaj ustawiana w kontrze do podejść opartych na danych, uznawanych za redukcjonistyczne i upraszczające, to przecież tu właśnie najwyraźniej można dostrzec rolę, jaką humanistyka może odegrać, wprowadzając do tego zachwytu konieczną komplikację i namysł. I nie chodzi tylko o etyczny wymiar zbierania i udostępniania danych, lecz przede wszystkim o ich fundamentalnie arbitralną naturę, którą perspektywa humanistyczna pozwala nam dostrzec.

Na wagę tego spojrzenia zwraca uwagę Johanna Drucker w ważnym eseju *Humanities approaches to graphical display*[3], w którym stwierdza, że nie powinniśmy mówić o d a n y c h (*data*), ale o w z i ę t y c h (*capta*). Badaczka przeciwstawia podejścia realistyczne czy esencjalistyczne, zakładające, że obiektywne zjawiska są niezależne od obserwatora – a zatem są nam d a n e – podejściu konstruktywistycznemu, w którym dane są współzależne od obserwatora, a zatem w z i ę t e, współkreowane. Drucker zwraca uwagę, że wszystkie bardziej złożone koncepty, takie jak naród czy płeć kulturowa, są konstruowane, czyli operacjonalizowane – sami ustalamy wyznaczniki przynależności elementu do danej grupy[4]. Stanowisko konstruktywistyczne zwraca więc naszą uwagę na to, że dane w humanistyce to efekt interpretacji i przyjętych założeń.

Sposób, w jaki operacjonalizujemy dane, może być brzemienny w skutki. Weźmy choćby pierwszy z brzegu przykład powszechnie znanego wskaźnika produktu krajowego brutto (PKB), którym zwykliśmy mierzyć wzrost gospodarczy. Zapominamy przy tym, że jest to po prostu stan rozkręcenia

2 W. Davies *The data sublime*, „The New Inquiry" (blog), 12 January 2015, https://thenewinquiry.com/the-data-sublime/, akapit 25.

3 J. Drucker *Humanities approaches to graphical display*, „Digital Humanities Quarterly" 2011 vol. 5/1, http://digitalhumanities.org/dhq/vol/5/1/000091/000091.html#p3.

4 Tamże, akapity 11–12.

gospodarki bez względu na to, co tę gospodarkę rozkręca. Dla przykładu – ogromne inwestycje po przejściu tsunami mogą wpłynąć dodatnio na wzrost PKB, ponieważ wiele firm dostaje kontrakty na odbudowę zniszczonych terenów. Nie mówi to nam jednak nic o jakości życia obywateli, o dostępie do służby zdrowia, usług itd. Dlatego też ten wskaźnik jest bardzo krytykowany nawet przez takie organizacje jak OECD (Organizacja Współpracy Gospodarczej i Rozwoju), a wiele krajów, w tym Nowa Zelandia czy Szkocja, eksperymentuje ze wskaźnikami mierzącymi dobrostan obywateli. Refleksja humanistyczna, wychodząc od uświadomienia arbitralności wszelkich wskaźników, pozwala nam zdefiniować, jakie dane są nam potrzebne do zmierzenia konkretnego zjawiska.

Sceptycyzm humanistów wobec terminu „dane" można tłumaczyć właśnie niechęcią do uproszczeń i kwantyfikacji. Z kolei sama ta niechęć bywa często formułowana w sposób nazbyt upraszczający, jak choćby uwagi Stephena Marche'a z eseju *Literature is not data: against digital humanities*[5]. Badacz stwierdza, że algorytmy są faszystowskie, gdyż kreują iluzję nieuniknionego obiektywizmu oderwanego od ludzkiej rzeczywistości, a znaczenie jest czymś, co się wymyka, rozpada i wymaga „ręcznego" – czyli atechnicznego – podejścia. Zarazem wyparcie wszechobecności danych we współczesnym świecie skazywałoby nas na ignorancję i niemożność zrozumienia, jak funkcjonuje współczesna kultura. Jak celnie zauważa Ted Underwood w eseju wymownie zatytułowanym *Dear humanists: fear not the digital revolution*, humaniści powinni poznać metody cyfrowe nie dlatego, że mielibyśmy wszyscy z nich korzystać, lecz by „zrozumieć, dlaczego zaciera się granica między rozumowaniem jakościowym i ilościowym"[6]. Dane i metody ich przetwarzania – stwierdza Underwood – pozwalają na nowe sposoby interakcji z przeszłością i otwierają nowe przestrzenie dialogu z innymi dyscyplinami. Kluczem do rozwikłania tego sporu jest bowiem zrozumienie, że metody ilościowe w humanistyce stanowią tylko podstawę dla prac interpretacyjnych i nigdy ich nie zastąpią.

Nieporozumienia często wynikają z różnic w języku, jakim do opisu danych posługują się z jednej strony badacze, a dostawcy technologii z drugiej. Zdaniem Jennifer Edmond i Erzsébet Tóth-Czifry humaniści mają bardzo

5 S. Marche *Literature is not data: against digital humanities*, „Los Angeles Review of Books", 28 October 2012, https://lareviewofbooks.org/article/literature-is-not-data-against-digital-humanities/.

6 T. Underwood *Dear humanists: fear not the digital revolution*, „The Chronicle of Higher Education", 27 March 2019, https://www.chronicle.com/article/Dear-Humanists-Fear-Not-the/245987.

rozbudowany język pozwalający nazywać ich dane: źródła, literatura przedmiotu, dokumenty, bibliografie, edycje krytyczne, adnotacje, notatki itd.[7] Dane w humanistyce to zatem, by przywołać Miriam Posner, „konieczna sprzeczność", i „nawet jeśli tradycyjni humaniści nie nazywają swoich źródeł danymi, to mają bardzo naglące potrzeby związane z zarządzaniem danymi"[8]. Dlatego też należy przemyśleć relację między badaniami humanistycznymi a danymi i sposobami zarządzania nimi. Aby uniknąć nieporozumień, zdefiniujmy roboczo dane badawcze jako wszystkie materiały i źródła, które uczeni zbierają, wytwarzają, wzbogacają i wykorzystują na wszystkich etapach procesu badawczego.

Temat danych badawczych, coraz istotniejszy w humanistycznych przedsięwzięciach cyfrowych, jest stosunkowo nowym obszarem zarówno jeśli chodzi o praktykę, jak i refleksję teoretyczną. Powstałe raporty i zbiory wskazówek dotyczące gromadzenia i udostępniania danych pozostają wciąż na poziomie ogólności, co rodzi potrzebę ich dostosowania do specyfiki poszczególnych dyscyplin humanistycznych – w naszym przypadku do badań o literaturze. Obecny stan niedostosowania „ogólnohumanistycznej" refleksji o danych do specyfiki literaturoznawstwa rodzi liczne nieporozumienia.

Prowadząc warsztaty czy rozmawiając z badaczami literatury, spotkaliśmy się wielokrotnie z różnymi obawami, jakie budzi postulat zarządzania danymi badawczymi, w tym ich udostępniania. Nasi rozmówcy kojarzyli go z jednej strony z wymogiem publikowania prywatnych notatek lub brudnopisów tekstów naukowych, z drugiej zaś z postulatem dzielenia się unikatowymi elementami warsztatu badawczego czy pomysłami, które może wykorzystać ktoś inny. Nierzadko słyszeliśmy też, że temat danych badawczych nie dotyczy wielu przedsięwzięć literaturoznawczych – zwłaszcza tych o profilu interpretacyjnym. W naszym tekście podejmiemy próbę dostosowania refleksji o danych do specyfiki literaturoznawstwa, postaramy się także odczarować ten temat i rozwiać obawy z nim związane. W pierwszej części artykułu skupiamy się na typologii danych badawczych w literaturoznawstwie, w drugiej omawiamy typy publikacji danych w naszej dyscyplinie, by na koniec omówić sposoby postępowania z danymi w ramach projektu.

7 J. Edmond, E. Tóth-Czifra *Open data for humanists, a pragmatic guide*, DARIAH 2018, s. 1, DOI: 10.5281/zenodo.2657248.

8 M. Posner *Humanities data: a necessary contradiction*, Miriam Posner's Blog, 25 June 2015, https://miriamposner.com/blog/humanities-data-a-necessary-contradiction/.

Identyfikacja i typologia danych w badaniach literackich

Pojęcie danych, kojarzące się z suchymi zestawami cyfr, intuicyjnie może się wydawać obce badaniom literackim, których przedmiotem są wszak dzieła sztuki, zdające wymykać się liczbowym opisom, a których głównymi formami publikacji są książki i artykuły, same niepozbawione często walorów artystycznych. Interpretacyjna praca literaturoznawcza z pozoru nie wytwarza niczego, co mogłoby pełnić funkcję danych. Gdy jednak przyjrzeć się bliżej, dostrzeżemy, że zdecydowana większość przedsięwzięć literaturoznawczych danymi się posługuje i je wytwarza – są nimi choćby zarówno indeksy, którymi opatrzone są książki literaturoznawcze, jak i bibliografie, sporządzane do artykułów w czasopismach – ich miejsce w proponowanej przez nas typologii danych oraz możliwe wykorzystania omówimy w dalszej części artykułu.

Innymi słowy, wytwarzamy dane od zawsze. Dlaczego więc dopiero teraz mówimy o danych badawczych w literaturoznawstwie? Wszystko to za sprawą technologii cyfrowych, które zmuszają nas do innego myślenia o naszych wytworach i źródłach. W czasie przedcyfrowym w zasadzie wszystkie zebrane materiały i działania służyły produktowi końcowemu w formie artykułu i monografii, druk zaś pozostawał podstawowym nośnikiem wszystkich treści. Możliwości, jakie niesie cyfrowe środowisko pracy, nie tylko pozwalają nam przemyśleć, czym są dane badawcze, lecz także powoli nas do tego zmuszają. Na przykład bibliografia od zawsze służyła do tego, by poznać i zlokalizować konkretne teksty. Odkąd jednak możemy zbierać cyfrowe dane bibliograficzne w standardowych formatach, sama bibliografia staje się narzędziem badawczym dla historyków kultury, którzy na jej podstawie mogą śledzić choćby ewolucję gatunków literackich[9]. Podobnie transkrypcje tekstów źródłowych przygotowane w ramach projektu interpretacyjnego można wydać drukiem (lub nie), ale udostępnione w formie cyfrowej mogą posłużyć komuś innemu do badań lub włączenia tekstu do szerszego korpusu językowego. Albo najprostszy przykład – w każdym projekcie przeprowadzamy kwerendy, generujemy liczne skany czy zdjęcia materiału źródłowego. Czyż nie byłoby wszystkim wygodniej, gdyby inni badacze mogli potem skorzystać z tego wysiłku, zamiast ponawiać nasze działania?

9 Na przykład Franco Moretti zebrał dane bibliograficzne z różnych źródeł, by postawić tezy o ewolucji brytyjskich gatunków powieściowych. Zob. nota metodologiczna w tegoż *Wykresy, mapy, drzewa*, przeł. T. Bilczewski, A. Kowalcze-Pawlik, Wydawnictwo UJ, Kraków 2016, s. 39-41.

Kwestia danych w literaturoznawstwie jest dziś najbardziej wyrazista w przedsięwzięciach z komponentem cyfrowym, tam bowiem zarówno efekt badawczy, jak i pozyskane materiały wyrażane są w postaci kodu programistycznego. Dzięki temu mogą być one także cyfrowo agregowane, analizowane i przetwarzane, dając duży potencjał ich ponownego wykorzystania. W dalszej części artykułu przywołamy przykłady danych w różnorakich (m.in. historycznoliterackich, edytorskich, bibliograficznych) cyfrowych projektach literaturoznawczych. Należy jednak podkreślić, że nawet przy najbardziej tradycyjnym podejściu do badań literackich nie ma ucieczki do czasów przedcyfrowych – realia funkcjonowania współczesnej nauki wymuszają na badaczach i instytucjach wprowadzanie wszystkich tekstów do cyfrowych baz i zestawień, co sprawia, że właściwie nie istnieje element dorobku naukowego, który nie zostawia „śladu cyfrowego" – a zatem każdy dorobek ma reprezentację w postaci danych cyfrowych.

Ze względu na zindywidualizowaną kulturę pracy wśród literaturoznawców panuje przekonanie, że ich danych badawczych nie sposób oddzielić od kontekstu wytworzenia i użycia, czyli od ram jednostkowego, często monograficznego, projektu. Obawy budzi zwłaszcza kwestia udostępnienia czy publikacji materiałów roboczych, sporządzanych często na własny użytek. Postrzegane jest to zwykle jak ingerencja w obszary zastrzeżone dla prywatności naukowców, próbę upublicznienia tego, co niedoskonałe, nieskończone, nienadające się do pokazania. Dodatkowo badacze boją się publikować takie materiały w obawie przed zawłaszczeniem ich ciężkiej pracy lub wykorzystaniem niezrealizowanych jeszcze pomysłów badawczych związanych z tym materiałem. Nic bardziej błędnego!

Należy z całą mocą podkreślić, że w refleksji o danych badawczych oraz postulatach ich otwartości czy odpowiedniego opisywania nie chodzi o zmuszanie do publikacji ani prywatnych notatek naukowców, ani jakichkolwiek materiałów, których publikacja może się autorom wydać niekomfortowa czy zagrozić ich dalszym dokonaniom. Wręcz przeciwnie, celem publikacji danych jest, po pierwsze, umożliwienie głębszego kontaktu z dziełem badawczym dzięki udostępnieniu materiałów źródłowych czy zestawień, które nie mieszczą się w monografii czy artykule. Po drugie, publikacja danych ułatwi innym badaczom korzystanie z tych wyników i materiałów we własnych pracach, oczywiście przy odpowiednim oznaczeniu ich pochodzenia, gdyż zestawy danych, które np. publikujemy w cyfrowych repozytoriach, mają również atrybucję autorską. Zebrane i opisane przez nas dane nie muszą być otwarte i powszechnie dostępne, możemy je udostępniać w ograniczonym zakresie

i za każdorazową zgodą. Nie chodzi zatem o wykradanie notatek z biurek literaturoznawców ani tym bardziej o przejmowanie ich cennych pomysłów, lecz przede wszystkim o umiejętność zidentyfikowania, jakie dane powstają w trakcie przedsięwzięcia naukowego, jakie mogłyby zostać opublikowane i przydać się w przyszłości, oraz takie ich opisanie i umiejscowienie, by mogły zostać ponownie wykorzystane i zacytowane.

Podczas zorganizowanych jesienią 2020 roku w IBL PAN warsztatów poświęconych danym badawczym pracowaliśmy razem z literaturoznawcami z całej Polski nad identyfikacją danych badawczych, którymi posługują się w swej pracy. Tabela zawiera przegląd projektów i danych, które wykorzystują (dla lepszego efektu dydaktycznego i większej przejrzystości uprościliśmy tematy projektów).

Tabela. Przykłady danych wykorzystywanych w projektach literaturoznawczych przez uczestników warsztatów w IBL PAN

Projekt	Typy danych
Recepcja literatury polskiej w jidysz	Zdigitalizowany korpus prasy, adnotacje dotyczące osób, postaci i miejsc
Analiza prac laureatów Olimpiady Języka Polskiego	Zdigitalizowany korpus prac konkursowych
Rekonstrukcja językowego obrazu świata w hiszpańskiej poezji	Zdigitalizowany korpus tekstów, dane słownikowe
Analiza metafor religijnych	Korpus źródeł, statystyczne zestawienia terminologiczne
Analiza powieści młodopolskich	Zdigitalizowany korpus tekstów, dane słownikowe, kod programistyczny
Badanie efektywności nauczania języka z wykorzystaniem technik teatralnych	Kolekcja nagrań recytacji
Semantyka powieści Jane Austen	Korpus powieści z adnotacjami dotyczącymi np. emocji

Analiza motywu Wisły w kulturze	Korpus tekstów źródłowych, znaczniki miejsc, do których odnoszą się teksty, mapy
Kairska scena muzyczna	Teksty utworów, wywiady

W badaniach ilościowych dane publikuje się głównie po to, aby umożliwić weryfikację wyników i replikację eksperymentów. W literaturoznawstwie, jak się zdaje, służą one głównie pogłębieniu kontekstu badań (np. jako suplement artykułu) lub dalszemu wykorzystaniu przez innych badaczy materiałów opracowanych w ramach projektu. Powyższe zestawienie dość dobrze pokazuje, że danymi badawczymi są po prostu systematycznie potraktowane materiały źródłowe i adnotacje wytwarzane w trakcie badań. Dane te publikowane są często jako apendyks do tekstu naukowego, jak choćby dokumentacja wariantów tekstów z powieści Jennifer Egan[10] czy lista wszystkich miejsc (wraz ze współrzędnymi geograficznymi) pojawiających się w twórczości Christine de Pizan[11]. Warto nadmienić, że te zbiory danych przydają się innym badaczom – wspomnianą listę lokalizacji pobrano prawie dwieście razy (co wydaje się całkiem dobrym wynikiem, biorąc pod uwagę specjalistyczną problematykę). Opublikowane dane mogą też oczywiście same w sobie stanowić efekt projektu literaturoznawczego, jak to się dzieje w wypadku korpusu powieści europejskiej[12]. Inne przykłady opublikowanych zbiorów danych humanistycznych można znaleźć np. w repozytorium Humanities Commons[13].

Istnieje wiele kryteriów formalnych, którymi można się posłużyć, klasyfikując dane badawcze. Po pierwsze, ze względu na medium podstawowe[14], w którym są wyrażone: słowo, obraz, dźwięk, albo ich współwystępowanie (np. wywiad, notatnik, nagranie przedstawienia, fotografia). Po drugie, ze względu na formę publikacji, co w przypadku materiałów cyfrowych może

10 M.P. Eve *Data appendices for „Textual scholarship and contemporary literary studies: Jennifer Egan's editorial processes and the archival edition of Emerald City"*, DOI: 10.5281/zenodo.3253829.

11 D.J. Wrisley *The literary geographies of Christine de Pizan* (geo-data), DOI: 10.5281/zenodo.35350.

12 C. Schöch, L. Burnard *COST-ELTeC/ELTeC-fra: release with 100 novels encoded at level 1*, DOI: 10.5281/zenodo.3878650.

13 https://hcommons.org/deposits/?facets[genre_facet][]=Data+set.

14 Odwołujemy się tu do rozróżnienia na media podstawowe i szczegółowe proponowanego przez Grzegorza Godlewskiego w: tegoż *Słowo – pismo – sztuka słowa. Perspektywy antropologiczne*, Wydawnictwa UW, Warszawa 2008, s. 277-279.

się odnosić zarówno do typu pliku (np. csv, txt, pdf, jpg, avi), jak i sposobu jego zdeponowania (np. plik w repozytorium lub dane udostępniane wyłącznie przez platformę projektu)[15]. Wreszcie, można je podzielić ze względu na strukturę na: ustrukturyzowane (baza danych), częściowo-ustrukturyzowane (np. XML) i nieustrukturyzowane (tekst)[16].

W ramach podsumowania tej części rozważań chcielibyśmy zaproponować typologię danych badawczych w literaturoznawstwie opracowaną na podstawie materiału zebranego podczas wspomnianych już warsztatów w IBL PAN. Odwołuje się ona nie tyle do specyfiki formalnej, ile do typów materiałów najczęściej wykorzystywanych w naszej dyscyplinie.

1. **Tekst kultury**[17] – konkretny egzemplarz utworu, wyrażonego znakami słownymi czy graficznymi, takiego jak wydane dzieło, rękopis, afisz teatralny, dwu- lub trójwymiarowa reprezentacja artefaktu (np. skan), nagranie.

2. **Metadane** – opis fizyczny i informacje o danym egzemplarzu utworu bądź dokumentu, jego wydaniu, czy też dane techniczne pliku cyfrowego.

3. **Adnotacje** – notatki, komentarze, uwagi, aparat krytyczny czy znaczniki TEI (*Text Encoding Initiative*) dodane do tekstu w procesie analizy i interpretacji przez badaczy lub wygenerowane mechanicznie (np. dane frekwencyjne, biogramy, n-gramy, nazwy własne, geolokacja miejsc w utworze). Adnotacje tym się różnią od metadanych, że wynikają z interpretacji własnej badaczy i przyjętych przez nich założeń metodologicznych, a nie wyłącznie ze standardu opisu np. bibliograficznego.

4. **Dane kultury literackiej** – zestawienia informacji o życiu literackim, takie jak kalendaria, listy osób, zdarzeń, dat, zestawienia statystyczne (np. nakładów, recepcji), wyniki ankiet, słowniki terminów i pojęć literackich.

---

15 N. Harrower, M. Maryl, T. Biro, B. Immenhauser *Sustainable and FAIR data sharing in the humanities: Recommendations of the ALLEA Working Group E-Humanities*, Berlin 2020, DOI: 10.7486/DRI.tq582c863, s. 14-15.

16 Por. C. Schöch *Big? Smart? Clean? Messy? Data in the humanities*, „Journal of Digital Humanities" 2013 no. 2/3, http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/.

17 Stosujemy tu szerokie rozumienie tekstów kultury za Stefanem Żółkiewskim, który definiował je jako „wszelkie struktury kodowe właściwe danej kulturze, a realizujące określone elementy jednego lub więcej systemu znaków funkcjonującego w tejże kulturze", w: tegoż *Teksty kultury*, PWN, Warszawa 1988, s. 23. Tekstem nazwiemy zatem każdy semantyczny wytwór kultury niezależnie od medium, w jakim został wyrażony lub utrwalony, np. tekst literacki, film fabularny, obraz, rytuał.

5. **Literatura przedmiotu** – interpretacje, podręczniki, świadectwa odbioru i inne opracowania lub wypowiedzi naukowe dotyczące przedmiotu badań.

6. **Dokumentacja procesu badawczego** – metodologia, notatki, systemy klasyfikacji pozwalające zrozumieć dane projektowe i sposób ich wytworzenia, treść kwestionariuszy, notatki ze spotkań, lista członków zespołu.

Skoro zidentyfikowaliśmy i zdefiniowaliśmy dane badawcze, z jakimi mamy do czynienia, kolejnym krokiem będzie omówienie sposobów, dlaczego warto je publikować. Zanim się tym jednak zajmiemy, przyjrzymy się przemianom komunikacji naukowej, które zwracają naszą uwagę na kwestię danych i umożliwiają nowe sposoby ich publikacji.

Formy publikacji danych literaturoznawczych

Forma publikacji determinuje jej odbiór, czyli sposób, w jaki czytelnicy wejdą w interakcję z tekstem. Odnosi się to także do możliwości ponownego wykorzystania materiałów. Wszystkie formy wypowiedzi artystycznej czy użytkowej – a zatem i dyskurs naukowy – ewoluują pod wpływem powiązanych ze sobą przemian kulturowych i technologicznych. Dotyczy to zarówno gatunków wypowiedzi, jak i ich podstawy technologicznej[18]. Do zrozumienia tej ewolucji przydatna będzie zaproponowana przez Jaya Davida Boltera i Richarda Grusina koncepcja remediacji, zakładająca, że nowe medium nigdy nie jest absolutną nowością, odtwarza bowiem w sobie stare medium, uzupełnione o nowe możliwości[19]. Badacze rozróżniają dwie strategie remediacji, w zależności od tego, czy obecność medium próbuje się przed widzami ukryć (bezpośredniość) czy też przeciwnie, uwydatnić (hipermedialność)[20]. Odnosząc te koncepcje do digitalizacji literatury, Maciej Maryl zauważa, że w pierwszym wypadku mamy do czynienia ze swoistymi cyfrowymi reprintami, ponieważ medium elektroniczne zachowuje kluczowe cechy tekstu drukowanego[21]. Chodzi tu zatem o proste skany czy e-booki utworu. W od-

18 Por. M. Maryl *Życie literackie w sieci. Pisarze, instytucje, odbiorcy wobec przemian technologicznych*, Fundacja Akademia Humanistyczna i Wydawnictwo IBL PAN, Warszawa 2015, s. 197-220.

19 J.D. Bolter, R. Grusin *Remediation: understanding new media*, MIT Press, Cambridge, MA 2000.

20 Tamże, s. 272-273.

21 M. Maryl *Reprint i hipermedialność – dwa kierunki rozwoju literatury ucyfrowionej*, w: *Tekst (w) sieci*, t. 2, red. A. Gumkowska, WAiP, Warszawa 2009, s. 86.

niesieniu do komunikacji naukowej możemy mówić o wszystkich formach traktujących medium elektroniczne wyłącznie jako środek dystrybucji, a nie formułowania wypowiedzi, takich jak artykuły czy monografie w formatach PDF czy EPUB. Druga strategia remediacji, polegająca na maksymalnym wykorzystaniu możliwości nowego medium, stara się łączyć różne środki przekazu w sposób wcześniej nieosiągalny, jak elektroniczne edycje naukowe, które łączą aparat naukowy z wykorzystaniem hipertekstu, wariantywności czy wyszukiwania i statystyk tekstowych[22].

Możemy zatem wyróżnić dwa podstawowe sposoby publikowania danych badawczych: jako pliki i jako platformy. W pierwszym wypadku chodzi o (1) zamieszczanie plików w repozytoriach, księgarniach, stronach wydawnictw, w drugim – tworzenie specjalnego interfejsu pozwalającego na dostęp do danych i interakcję z zebranym materiałem. W tej drugiej grupie wyróżniamy (2) monografie multimedialne, (3) edycje elektroniczne i (4) prace dokumentacyjne. Omówimy kolejno te typy publikacji.

Deponowanie pliku

W poprzedniej części artykułu, prezentując przykłady danych badawczych w literaturoznawstwie, podawaliśmy odnośniki do dwóch repozytoriów: Zenodo i Humanities Commons. Zdeponowanie danych badawczych w formie plików jest najprostszym i najtańszym sposobem ich publikacji i może być zarówno wykonane w otwartym repozytorium, jak i na zamkniętym dysku twardym danej jednostki. W tym kontekście za zdeponowanie pliku uznajemy także publikację pliku z tekstem przez wydawcę czy czasopismo. Mechanizm jest ten sam – użytkownicy pobierają dany plik i następnie otwierają go za pomocą odrębnego oprogramowania, takiego jak edytor tekstu, arkusz kalkulacyjny, programy do odczytu PDF itp. Deponujący mają do wyboru różne formaty plików, wspomniane wyżej, których zastosowanie wyznacza możliwości korzystania z tekstu. Na przykład plik PDF zachowuje formatowanie i paginację utworu, co może być wygodne dla niektórych czytelników, ale trudniej z niego korzystać na czytnikach czy skopiować tekst lub dane do dalszego przetwarzania.

Dane badawcze można publikować w ten sposób także jako aneks do artykułu (jak w przywoływanych wcześniej przykładach) czy książki. Ciekawy przypadek stanowi monografia Teda Underwooda *Distant horizons. Digital*

22 Tamże, s. 88-90.

evidence and literary change (2019). Autor opublikował w repozytoriach dane i kod programistyczny, które posłużyły mu do uzyskania wyników opisywanych w książce[23]. Czytelnicy mogą zatem z jednej strony zapoznać się z materiałami, których nie udało się umieścić w wydaniu książkowym, z drugiej – zweryfikować uzyskane wyniki na podstawie materiału źródłowego. Publikowanie danych pozwala więc przekroczyć ograniczenia zamkniętych form publikacji.

Monografie multimedialne

Zagadnienia związane z danymi w odniesieniu do multimedialnych monografii literaturoznawczych omówimy na przykładzie naukowych kolekcji cyfrowych na platformie Nowa Panorama Literatury Polskiej (NPLP.PL). W perspektywie koncepcji remediacji monografia multimedialna stanowi przeciwieństwo pliku zdeponowanego. Środowisko cyfrowe dostarcza w tym wypadku narzędzi, dzięki którym można zestawiać ze sobą teksty wyrażone w różnych mediach (takie jak zdjęcia, reprodukcje dzieł sztuki, skany dokumentów, mapy statyczne oraz interaktywne), dopuszczając przy tym inne formy interakcji z materiałem, np. struktury sieci hiperłączy pozwalających na zerwanie z narracją liniową i umożliwiających użytkownikom/czytelnikom modelowanie porządku obcowania z treścią[24]. Każda z dziewięciu kolekcji obecnych na platformie w chwili pisania tego artykułu ma nieco inny charakter i założenia (jest inną „naukową opowieścią cyfrową"), a zarazem stanowi łącznie dającą się łatwo przeszukiwać bazę wiedzy o literaturze i kulturze polskiej, której poszczególne jednostki (artykuły, hasła, mapy) mogą być również wykorzystywane jako autonomiczne teksty na konkretne tematy (np. hasło *Kobiety i miłość* w *Leksykonie Lalki*[25] z kolekcji *PrusPlus* czy opatrzony

23 Zob. szersze omówienie w: M. Maryl *Computational monograph: reading and writing distant horizons* [Ted Underwood, Distant Horizons, 2019], „JLTonlline" 2020 vol. 14 (2), urn:nbn:de:0222-004455.

24 Podstawy tego, jak zdefiniowano naukową kolekcję cyfrową na platformie Nowa Panorama Literatury Polskiej i jakie dodatkowe kwestie wynikają z takiej formy publikacji, zob. B. Szleszyński, K. Niciński, A. Kochańska *Jak przekazywać naukową wiedzę w Internecie. (Na marginesach kolekcji „PrusPlus" w Nowej Panoramie Literatury Polskiej)*, „Napis" 2015 nr 21, DOI: 10.18318/napis.2015.1.24.

25 G. Borkowska *Kobiety i miłość*, „PrusPlus", Nowa Panorama Literatury Polskiej, http://nplp.pl/artykul/kobiety-i-milosc/ (28.01.2021).

interaktywną mapą artykuł *Nazwy geograficzne w poezji Mickiewicza*[26] z *Atlasu Literatury Romantyzmu*).

Kwestie monografii cyfrowej najlepiej będzie pokazać na przykładzie naukowej kolekcji cyfrowej *Sienkiewicz Ponowoczesny* i towarzyszącej jej książki opublikowanej prymarnie w wersji cyfrowej[27]. W obu przypadkach na treść składa się 12 rozpraw dotyczących twórczości i biografii pisarza oglądanych poprzez narzędzia interpretacyjne kojarzone z ponowoczesnością. O ile książka stanowi klasyczną monografię wieloautorską, o tyle kolekcja cyfrowa jest wzbogacona o dodatkowe materiały (dane) oraz struktury hiperłączy. Jej narracja opiera się na wykorzystaniu licznych materiałów wizualnych i na mechanizmach interakcyjnych. To one stanowią główny zbiór danych wytworzonych w projekcie; są wśród nich pozyskane z Muzeum Narodowego w Kielcach i zarządzanego przez niego pałacu w Oblęgorku reprodukcje obrazów i eksponatów (opatrzone opisami/metadanymi), wykonane podczas projektu zdjęcia pałacu z zewnątrz oraz w każdym z czterech pomieszczeń wykorzystanych w naukowej opowieści, wreszcie zdjęcia obiektów (jak pomniki Henryka Sienkiewicza czy tablice pamiątkowe) wykonane przez zespół projektowy. Inny typ danych powstałych w ramach projektu to mapy – zarówno statyczne, jak i interaktywne – dynamiczne[28]. Wszystkie te dane są oczywiście wpisane w zaplanowany kontekst narracji naukowej, jednak, tak jak eksponaty z muzeum czy mapy, są również autonomicznymi materiałami, podlegającymi także ewentualnej dalszej interpretacji. Warto zauważyć, że jakkolwiek zidentyfikowane, opisane i odpowiednio przechowywane, dane te nie mogą zostać udostępnione na otwartej licencji – umowa zawarta z Muzeum Narodowym w Kielcach ogranicza ich wykorzystanie do platformy NPLP.PL – w tym przypadku możliwość ponownego użycia ogranicza się do innych kolekcji na tej platformie. Podczas licznych kwerend projektu po-

26 D. Siwicka *Nazwy geograficzne w poezji Mickiewicza*, „Atlas Literatury Romantyzmu", Nowa Panorama Literatury Polskiej, http://nplp.pl/artykul/nazwy-geograficzne-w-poezji-mickiewicza/ (28.01.2021).

27 Publikacje stanową efekt projektu „Sienkiewicz ponowoczesny – laboratorium cyfrowe" i można je zobaczyć pod następującymi linkami: https://nplp.pl/pobierz-ebooka-sienkiewicz-ponowoczesny/; https://nplp.pl/kolekcja/sienkiewicz-ponowoczesny/.

28 Więcej o kolekcji powstałej podczas projektu „Sienkiewicz ponowoczesny – laboratorium cyfrowe", jej strukturze i wykorzystywanych materiałach wizualnych zob. B. Szleszyński *Kilka uwag o intymności kolekcji cyfrowych przy okazji prac nad projektem „Sienkiewicz ponowoczesny – laboratorium cyfrowe"*, „Sztuka Edycji" 2019 nr 15/1, red. A. Markuszewska, s. 111-121, DOI: 10.12775/SE.2019.0011.

wstała również obszerna dokumentacja fotograficzna – została ona opisana i „zmagazynowana" na dyskach służbowych, by ewentualnie posłużyć podczas kolejnych projektów.

Formą wzbogacenia kolekcji cyfrowej w porównaniu z książką jest ponadto dodatkowa strukturyzacja. Wszystkie artykuły zostały napisane w taki sposób, by ich poszczególne fragmenty mogły funkcjonować również poza liniowym porządkiem artykułu, jako teksty autonomiczne. Przypisanie ich do kategorii umożliwia zupełnie inny porządek lektury, oparty na zagadnieniach tematycznych – np. wybierając kategorię „Cielesność" i podkategorię „Choroba i nerwy", zobaczymy zbiór 10 fragmentów z różnych artykułów, które im odpowiadają. Można zatem powiedzieć, że w ten sposób kolekcja oprócz tradycyjnej monografii naukowej oferuje także narzędzie do zapoznania się z różnymi wątkami tematycznymi w biografii i twórczości Henryka Sienkiewicza.

W kolekcji *Sienkiewicz ponowoczesny*, w dziale „O tworzeniu kolekcji"[29], zespół projektowy opublikował część dokumentacji i notatek z kolejnych faz eksperymentu (podtytuł projektu „laboratorium cyfrowe" traktowany był bardzo serio), odsłaniających jego przebieg, zmieniające się założenia, odrzucone rozwiązania i niewykorzystane ostatecznie dane. Znalazły się tam m.in. wytyczne dla autorów tekstów, pierwsza koncepcja kolekcji, przykłady dokumentacji fotograficznej (wraz z nieprzetworzonymi cyfrowo zdjęciami wnętrz pałacu), slajdy z prezentacji dotyczących tworzenia struktury treści czy próbki różnych rozwiązań graficznych związanych z interaktywnym sterowaniem. Celem było podzielenie się uzyskaną wiedzą o tworzeniu naukowych kolekcji cyfrowych, pokazanie złożoności projektu i wieloaspektowości działań, wreszcie po trosze udokumentowanie procesu „produkcyjnego". Oczywiście jest to inna sytuacja niż potencjalne dzielenie się notatkami bądź szkicowymi wersjami swoich tekstów, ale istota dylematu stojącego przed wykonawcami była podobna, ponieważ zdecydowano się na udostępnienie *know-how* wypracowanego w ramach projektu.

Innym przykładem publikacji danych badawczych w monografii cyfrowej jest kolekcja *Atlas Literatury Zagłady*[30]. W przeciwieństwie do *Sienkiewicza Ponowoczesnego* nie ma ona liniowej, książkowej wersji; pomyślana została od razu jako cyfrowa sieć powiązanych ze sobą wpisów – fragmentów świadectw Zagłady z getta warszawskiego, ustrukturyzowanych wedle trzech możliwych porządków: topograficznego, osobowego i czasowego, wzbogaconych

29 http://nplp.pl/o-tworzeniu-kolekcji-komentarz/.

30 http://nplp.pl/kolekcja/atlas-zaglady/.

o niemal tysiąc map statycznych poświęconych jednostkom topograficznym, siedemnaście map interaktywnych pokazujących topografię miejsc związanych z każdym z autorów/autorek świadectw na przestrzeni czasu, biogramy autorów/autorek świadectw, opisy specyfiki okresów, na które została podzielona linia czasu, oraz zdjęcia archiwalne. Te same treści mogą więc być przeglądane, zestawiane i wyszukiwane pod najróżniejszymi kątami, co czyni z kolekcji bardziej narzędzie dla badaczy do przeszukiwania najróżniejszych rodzajów danych, które mogą ich wspomóc w pracy analityczno-interpretacyjnej. Monografia cyfrowa otwarta jest zatem na współpracę z instytucjami dziedzictwa kulturowego – czy będą to obrazy i eksponaty z Muzeum Narodowego w Kielcach, zdjęcia z Żydowskiego Instytutu Historycznego czy grafiki z XIX-wiecznej prasy ze zbiorów biblioteki IBL PAN, zostają one nie tylko włączone w nurt naukowych opowieści i zinterpretowane, ale i udostępnione cyfrowo szerokiej publiczności.

Cyfrowa edycja naukowa

Kolejny typ publikacji danych literaturoznawczych stanowią naukowe edycje tekstów literackich i okołoliterackich. W wypadku standardowej edycji drukowanej prezentuje się daną wersję tekstu, uznaną za tekst krytyczny, wraz z komentarzem naukowym, indeksami, niekiedy ilustracjami (czasami rolę ilustracji odgrywają zeskanowane wybrane strony rękopisu czy pierwodruku). Druk narzuca linearność sposobu publikacji, utrudniając prezentację różnych warstw aparatu krytycznego czy wariantów, emendacji i koniektur. Edycja cyfrowa pozwala, przynajmniej częściowo, przezwyciężyć te trudności dzięki nielinearnej, hiperlinkowej strukturze połączeń między tekstem a aparatem krytycznym i różnymi wersjami tekstu, dzięki możliwościom zestawiania różnych wersji tekstu czy uruchamiania przez użytkownika wybranych, dowolnie nakładanych na siebie warstw aparatu krytycznego (może np. wyświetlić w danym tekście jedynie osobowe jednostki indeksowe lub wyłącznie te fragmenty, które modyfikowano w różnych wariantach tekstu). W trakcie prac nad edycją na wielu etapach powstają dane, takie jak skany rękopisów, transkrypcja, materiał ilustracyjny czy komentarz edytorski. Wszystkie te dane muszą być tak czy inaczej zebrane podczas realizacji projektu – po jego zakończeniu potencjalnie można je udostępnić, co z opisywanych już wcześniej względów jest stosunkowo proste w środowisku cyfrowym.

Możliwości wykorzystania danych w edycji cyfrowej omówimy na podstawie stworzonej w IBL PAN platformy do naukowych edycji cyfrowych TEI.NPLP.PL.

Do opisu struktury tekstów wykorzystuje ona międzynarodowy standard TEI (*Text Encoding Initiative*) wykorzystująca znaczniki XML specjalnie opracowane na potrzeby edycji cyfrowych[31] – można dzięki nim zakodować różne poziomy aparatu krytycznego, m.in. odmiany i warianty tekstu, właściwości artefaktu (np. zniszczenia rękopisu) czy rodzaje didaskaliów w tekście dramatycznym[32]. Każdy tekst opublikowany na tej platformie można pobrać zarówno w formacie XML (ze znacznikami), jak i jako „czystą" transkrypcję, pozbawioną aparatu krytycznego. Dzięki wykorzystaniu standardu TEI metadane tekstu oraz jego strukturę można przetwarzać, zestawiać i porównywać z innymi, podobnymi tekstami oznaczonymi w TEI. Staje się to jeszcze łatwiejsze, gdy mówimy o tekstach określonego rodzaju literackiego, którego opisowi służy odrębny podstandard, np. przeznaczony do opisu struktury dramatycznej TEI Drama. O tym, że możliwość ponownego wykorzystania danych nie jest czysto potencjalna, przekonuje platforma DraCor[33], zbierająca ponad dziesięć obszernych korpusów dramatów z różnych epok i języków, powstałych w różnych projektach, których wspólną cechą jest to, że wszystkie zostały oznaczone właśnie w podstandardzie TEI Drama. Dzięki tym oznaczeniom oraz oprogramowaniu na stronie możliwe jest zestawianie różnorakich statystyk i wykresów dla każdego z ponad tysiąca opracowanych wcześniej dramatów.

Prace dokumentacyjne

Opracowania dokumentacyjne – w tym bibliografie, katalogi (archiwów, bibliotek), słowniki, encyklopedie, leksykony czy kalendaria – odgrywają istotną rolę w badaniach literaturoznawczych. Ich funkcją są zarówno zachowanie, ochrona dziedzictwa kulturowego i naukowego, jak i dostarczanie wiedzy oraz informacji dla szeroko rozumianej działalności badawczej. Procesy dokumentowania kultury, w tym literatury, można rozumieć jako *de facto* praktyki gromadzenia, organizowania i udostępniania danych badawczych. Jest tak przede wszystkim dlatego, że efekty prac dokumentacyjnych

31 Opisy znaczników TEI oraz wiele innych dokumentów dotyczących tego standardu oznaczania można znaleźć na stronie konsorcjum TEI: https://tei-c.org/.

32 Więcej o cyfrowym aparacie krytycznym opartym na standardzie TEI zob. K. Niciński *Obecność przypisu w edycji cyfrowej – rekonesans*, „Napis" 2019 nr 25, DOI: 10.18318/napis.2019.1.13.

33 https://dracor.org/ Więcej o projekcie w: F. Fischer et al. *Programmable Corpora: Introducing DraCor, an infrastructure for the research on European drama*, Proceedings of DH2019: „Complexities", Utrecht University 2019, DOI:10.5281/zenodo.4284002.

od zawsze miały charakter ustrukturowany (katalog biblioteczny, archiwalny itp.) lub częściowo ustrukturowany (bibliografia adnotowana, słownik biobibliograficzny itp.). Współczesne technologie cyfrowe pozwalają na tworzenie nowoczesnych usług (bibliograficzne bazy danych, katalogi cyfrowe, e-słowniki itp.) umożliwiających sprawne przeszukiwanie, pobieranie i ponowne wykorzystywanie takich zasobów.

O ile tworzeniem słowników, kalendariów czy katalogów zajmują się często specjaliści bądź specjalistyczne zespoły, o tyle wytwórcami czy, częściej, przetwórcami danych bibliograficznych są praktycznie wszyscy badacze na co dzień, tj. nie generują nowych danych, ale je wykorzystują (przetwarzają) w swojej pracy. Gromadzenie, organizowanie i wykorzystywanie danych bibliograficznych to proces bezpośrednio związany z każdą działalnością badawczą. Istnieją dwa kluczowe aspekty takich interakcji z danymi bibliograficznymi: 1) formatowanie informacji bibliograficznej w obrębie własnej publikacji naukowej (literatura cytowana), 2) opatrywanie własnych publikacji metadanymi. Oba te aspekty wytwarzania i przetwarzania informacji bibliograficznej wpływają na jakość danych badawczych.

Współcześnie spotykamy trzy główne formy publikacji efektów prac dokumentacyjnych:

1. Publikacja (drukowana lub cyfrowa, w tym zdigitalizowana np. w postaci skanów drukowanej publikacji) w formie książki (słownik, encyklopedia, monografia bibliograficzna, bibliografia itp.) lub, rzadziej, artykułu (cykliczne kalendaria, wyimki ze słowników czy encyklopedii albo inne publikacje przyczynkarskie).
2. Usługi cyfrowe – powstałe oryginalnie w formie cyfrowej lub będące efektem retrospektywnej konwersji bazy danych, najczęściej dostępne (w różnym stopniu) online przy użyciu publicznego interfejsu (katalog archiwum, biblioteki, słownik online, kolekcje bibliograficzne publikowane za pomocą menedżerów bibliografii itp.).
3. Zestawy danych (*datasets*) w formie plików deponowanych w archiwach czy repozytoriach danych (por. wyżej), które są odpowiednio przygotowane i udokumentowane, tak aby umożliwiać dalsze ich wykorzystywanie w celu weryfikacji badań dotychczasowych czy też nowych eksploracji.

We wszystkich tych przypadkach sposób publikacji wpływa bezpośrednio na możliwości dalszego wykorzystywania danych, stąd całość procesów dokumentacji powinna uwzględniać zgodność wytworzonych danych ze standardami FAIR (o których mowa w kolejnej części artykułu) lub innymi

dobrymi praktykami stosowanymi do określonego typu danych. Dla słownika wydanego w formie e-booka lub w formie usługi cyfrowej może to być np. opatrzenie tekstu omawianymi wyżej tagami TEI (aby ułatwić ekstrakcję danych, tworzenie narzędzi na podstawie korpusu itp.), dla bibliografii – zastosowanie międzynarodowo rozpoznawanych formatów danych.

Informacja bibliograficzna w publikacjach naukowych jest tradycyjnie ujmowana w formie przypisów oraz bibliografii załącznikowej. W ostatnich dekadach dane bibliograficzne zawarte w publikacjach – przede wszystkim naukowych – podlegają formalizacji nie tylko w zakresie wykorzystywanych czy rekomendowanych stylów cytowania, lecz także dobrych standardów formatowania tekstów pozwalających na wydobywanie z nich cytowanej literatury (np. wyodrębniane są pola na cytowaną literaturę). Informacje te służą do gromadzenia danych o odniesieniach międzytekstowych, a wykorzystują ją m.in. nowoczesne serwisy informacji naukowej (w szczególności tzw. indeksy cytowań), podmioty zajmujące się ewaluacją działalności naukowej oraz grono badaczy z zakresu szeroko rozumianej bibliometrii, analityki kulturowej, historii nauki. Informacje te służą z jednej strony do monitorowania i badania efektywności prac naukowych, z drugiej zaś wspierają inicjatywy z zakresu otwartego dostępu[34], gdyż w ekosystemie komunikacji naukowej, w którym część pełnych tekstów publikacji jest niedostępna, metadane stanowią narzędzie częściowego otwierania do nich dostępu (tzn. informują o nich, pozwalają gromadzić o nich wiedzę również w sytuacji, gdy dostęp do pełnego tekstu jest zamknięty lub ograniczony).

Z efektów prac dokumentacyjnych korzysta większość badaczy literatury, publikowanie ich jako danych badawczych umożliwia natomiast ich specyficzne użycie, np. analizę wspomaganą komputerowo, badania na danych, analizy ilościowe. Wśród istotnych i szeroko spopularyzowanych form ponownego wykorzystania badawczego efektów prac dokumentacyjnych znajdują się choćby analizy sieciowe słowników, zarówno biograficznych, jak i językoznawczych[35]. W wypadku danych biograficznych pozwalają one

34 Na przykład platforma Open Citations (https://opencitations.net/), która agreguje informacje z otwartych i zamkniętych naukowych baz danych, umożliwiając dostęp do cytowań.

35 Zob. np. C. Garrido, C. Gutierrez *Dictionaries as networks: identifying the graph structure of Ogden's basic English*, w: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, The COLING 2016 Organizing Committee, Osaka 2016, https://www.aclweb.org/anthology/C16-1336/, s. 3565–3576; I. Sánchez-Berriel, O.S. Suárez, V.G. Rodríguez, J.P. Aguiar *Network analysis techniques applied to dictionaries for identifying semantics in lexical Spanish collocations*, w: *Lexical collocation analysis. quantitative methods*

na odkrywanie zależności między osobami, instytucjami czy miejscami na podstawie ilościowych analiz wielu biogramów. Dla przykładu analizy sieciowe *Austriackiego słownika biograficznego* umożliwiły identyfikację i wizualizację najbardziej popularnych miejsc podróży i emigracji ujętych w nim postaci czy też najbardziej popularnych miejsc pobierania nauki[36]. Badania na korpusie *Fińskiej biografii narodowej* pozwoliły zrekonstruować i porównać długość życia oraz miejsca urodzeń i śmierci bohaterów biogramów na przestrzeni kilku stuleci, a także wskazać najpopularniejszych pracodawców i najchętniej uczęszczane uniwersytety[37].

W odniesieniu do danych bibliograficznych powszechnie znane są badania bibliometryczne nad komunikacją naukową, również w humanistyce, tam bowiem dane pozwalają zarówno uchwycić relacje między badaczami, ich wzajemne oddziaływania, jak i zrozumieć modele pracy naukowej charakterystyczne dla wybranych (sub)dyscyplin, krajów czy regionów. Dla przykładu badania z zakresu ewaluacji badań naukowych[38] pokazują, że około 25 procent publikacji badaczy humanistów z Polski, Czech czy Słowacji opublikowane jest w języku angielskim, podczas gdy w Danii, Finlandii, Flandrii czy Norwegii jest to powyżej 60 procent. Jednocześnie w krajach Europy Środkowej rośnie udział artykułów w ogólnej liczbie publikacji naukowych, przy spadku liczby monografii oraz rozdziałów w książkach zbiorowych. Z kolei badania wykorzystujące sieci cytowań pozwalają choćby zrozumieć sposoby kształtowania się nowych dyscyplin badawczych, takich jak humanistyka cyfrowa. Wstępne badania w 2019 roku[39] wykazały, że jedynie jedna trzecia publikacji z tej dziedziny (zindeksowanych w serwisie Dimensions) zawierała

in the humanities and social sciences, ed. by P. Cantos-Gómez, M. Almela-Sánchez, Springer, Cham 2018, s. 39-57, DOI: 10.1007/978-3-319-92582-0_3.

36 Á.Z. Bernád, M. Kaiser *The biographical formula: types and dimensions of biographical networks*, w: *BD-2017 Biographical Data in a Digital World 2017. Proceedings of the Second Conference on Biographical Data in a Digital World 2017*, ed. by A. Fokkens et al., Linz 2018, s. 45-52.

37 P. Leskinen, E. Hyvönen, J. Tuominen *Analyzing and visualizing prosopographical linked data based on biographies*, w: *Biographical Data in a Digital World 2017. CEUR Workshop Proceedings*, RWTH Aachen 2018, http://ceur-ws.org/Vol-2119/paper7.pdf, s. 39-44.

38 E. Kulczycki, T.C.E. Engels, J. Pölönen et al. *Publication patterns in the social sciences and humanities: evidence from eight European countries*, „Scientometrics" 2018 no. 116, s. 463-486, DOI: 10.1007/s11192-018-2711-0.

39 G. Spinaci, G. Colavizza, S. Peroni *Preliminary results on mapping digital humanities research*, w: *AIUCD 2020. La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica*, Milano 2020, https://convegni.unicatt.it/aiucd-Spinaci_et_al.pdf.

wewnętrzne odniesienia (cytaty). Ponadto odkryto chociażby, że wiele czasopism z tej dziedziny ma wyraźny rys dyscyplinarny – dotyczy m.in. literaturoznawstwa ilościowego, lingwistyki komputerowej czy przetwarzania języka naturalnego – i publikujący tam autorzy odwołują się do literatury ważnej dla danej dyscypliny, a w mniejszym stopniu do innych rodzajów badań z zakresu humanistyki cyfrowej.

Innym ważnym nurtem badań na danych bibliograficznych są badania z pogranicza historii książki i analityki kulturowej (*cultural analytics*). W historii książki zawsze obecne były badania ilościowe; jednym z ich przykładów są prace Jana IJ. van der Meera dotyczące epoki stanisławowskiej na podstawie danych *Nowego Korbuta*[40]. Współcześnie można zauważyć coraz większe zainteresowanie kolekcjami danych bibliograficznych, jakimi są katalogi biblioteczne. Fińskie badania na danych z *English Short-Title Catalogue*[41] pozwalają określić najbardziej produktywne wydawnictwa i autorów w okresie wczesnego druku (wieki XV-XVIII), a także zrozumieć trendy publikacyjne istotne choćby dla bibliografii opisowej, np. to, że rozmiary książek historycznych były statystycznie większe w XVI niż w XVII i XVIII stuleciu.

Jak postępować z danymi badawczymi

Ucyfrowienie humanistyki uświadomiło wielu z nas, jak dużo materiałów zwyczajnie marnuje się po zakończeniu projektu. Często mogłyby przecież służyć kolejnym badaczom, którzy nie musieliby powielać naszych starań, w tym wyszukiwania materiałów czy czasochłonnej digitalizacji źródeł papierowych. Co więcej, jeśli nie porządkujemy danych w swoim przedsięwzięciu badawczym, może dojść do sytuacji, gdy sami nie będziemy mogli z nich w przyszłości skorzystać. Zostaną pochopnie skasowane, zgubią się w jednym z licznych folderów bądź stracimy do nich dostęp po zmianie sprzętu.

Takim sytuacjom ma zapobiec stosowanie zasad FAIR, czyli założenie, że dane powinny być możliwe do znalezienia (*findable*), dostępne (*accessible*), interoperacyjne (*interoperable*) i nadające się do ponownego wykorzystania (*reusable*). Grupa ekspertów przy Komisji Europejskiej opracowała specjalny raport *Turning FAIR into reality*, w którym formułuje rekomendacje dotyczące

40 J.I. van der Meer *Literary activities and attitudes in the Stanislavian Age in Poland (1764-1795)*, Rodopi B.V., Amsterdam–New York 2002.

41 L. Lahti, N. Ilomäki, M. Tolonen *A Quantitative Study of History in the English Short-Title Catalogue (ESTC), 1470-1800*, „LIBER Quarterly" 2015 vol. 25, no. 2, s. 87-116, DOI: 10.18352/lq.10112.

zasad FAIR[42]. Dokument analizuje bieżące praktyki i proponuje plan działań dla europejskich interesariuszy, by wdrażać zasady FAIR na różnych poziomach. Mają być one wykorzystywane przy tworzeniu Europejskiej Chmury Danych Badawczych (EOSC). Warto również podkreślić znaczenie zarządzania danymi badawczymi w kontekście wymagań w konkretnych konkursach i grantach badawczych. Jednym z ważnych przykładów jest oczywiście europejski program „Horyzont 2020", w którego ramach od 2014 roku wdrażana jest pilotażowa inicjatywa „Open Data Pilot"[43], obejmująca już wszystkie obszary programu. Europejska platforma „Open Research Europe"[44] ma na celu ułatwić udostępnianie publikacji i danych wytworzonych w projektach finansowanych z programów ramowych Unii Europejskiej. W 2021 roku będą tam opublikowane pierwsze wyniki pracy. Co bardzo istotne z perspektywy polskich humanistów, plan zarządzania danymi badawczymi (PZD) – dokument, który porządkuje kwestie związane z danymi zbieranymi, wykorzystywanymi i wytwarzanymi w projekcie – stał się także jednym z elementów wniosku o finansowanie projektu w Narodowym Centrum Nauki.

Ponieważ temat jest stosunkowo świeży, wielu humanistów oddolnie opracowuje rozwiązania i standardy, aby uniknąć bezwiednego kopiowania procedur przyjętych przez nauki ścisłe i uwzględnić specyfikę własnych dyscyplin. Europejskie konsorcjum DARIAH (Digital Research Infrastructure for the Arts and Humanities)[45] wiele swoich działań skupia na danych badawczych w humanistyce i naukach o sztuce. Jedną z grup roboczych działających w ramach konsorcjum jest Research Data Management Working Group[46], koncentrująca swoje prace właśnie na wyzwaniach, z jakimi borykają się przedstawiciele różnych dziedzin – w tym literaturoznawstwa – w zarządzaniu danymi. Inne grupy robocze skupiają się na konkretnych rodzajach danych (np. danych bibliograficznych w wypadku Bibliographical Data Working Group) lub na węższych zagadnieniach (np. kwestiach etycznych i prawnych, które opracowuje Ethics and Legality in the Digital Arts and Humanities).

---

42 Directorate-General for Research and Innovation (European Commission) *Turning FAIR into reality*, Brussels 2018, DOI: 10.2777/1524.

43 Zob. https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm.

44 Zob. https://open-research-europe.ec.europa.eu/.

45 Zob. https://www.dariah.eu/.

46 Zob. https://www.dariah.eu/activities/working-groups/research-data-management/.

O tym, że tematyka danych przestaje w humanistyce być sprawą marginalną, świadczy również raport ALLEA (ALL European Academies), zawierający rekomendacje dotyczące zarządzania danymi badawczymi i stosowania zasad FAIR w obszarze humanistyki. Autorzy zalecają, by humaniści zaczęli myśleć o materiałach i źródłach w swojej pracy naukowej właśnie w kategoriach danych badawczych. Podkreślają też rosnące znaczenie zasad FAIR w pracach instytucji dziedzictwa kulturowego, które tak często współpracują z humanistami[47]. Co być może najważniejsze w kontekście niniejszego tekstu, raport jasno stwierdza, że sposoby ich wprowadzenia w obszarze konkretnych dziedzin nie zostały jeszcze ustalone. Potrzebna jest zarówno refleksja badaczy-specjalistów o różnych kompetencjach i zainteresowaniach, jak i wymiana praktyk między dziedzinami[48].

Zasady FAIR i inne związane z danymi inicjatywy nie są zatem narzuconymi nakazami, lecz przydatnymi wskazówkami, które mogą trwale wpłynąć na obieg i dostępność źródeł w humanistyce. Wielość oddolnych i odgórnych inicjatyw jasno wskazuje, że zarządzanie danymi badawczymi staje się dla humanistów koniecznością. Nie nadążył za nią jeszcze system ewaluacji, który rzadko uwzględnia wartość badawczą związaną z opracowaniem i udostępnianiem danych[49]. Nie dziwi więc pewna podejrzliwość wobec propozycji dostosowania zasad FAIR do świata humanistyki. Może być widziana jako dodatkowy zestaw czasochłonnych obowiązków nakładanych na badaczy, a efektu związanych z nimi prac często nie można przedstawiać bezpośrednio jako swojego dorobku naukowego. Uwzględnienie zarządzania materiałami badawczymi w systemie ewaluacji jest zatem niezbędne, by wspólnie stworzyć ekosystem danych badawczych dla humanistyki.

Niezależnie od tego, czy decydujemy się na publikację danych badawczych wytworzonych w ramach projektu, praca z nimi wymaga systematycznego przemyślenia. Bardzo pomocne jest opracowanie na początku projektu planu zarządzania danymi badawczymi, zbierającego podstawowe założenia

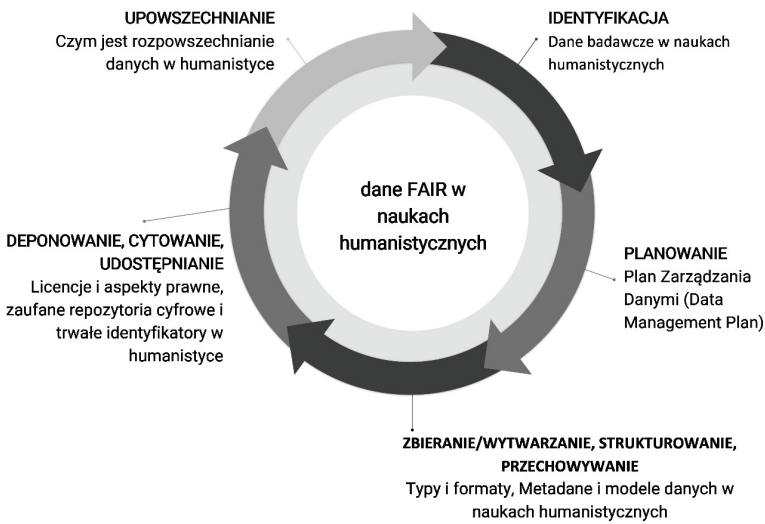
47 N. Harrower, M. Maryl, T. Biro, B. Immenhauser *Sustainable and FAIR data sharing in the humanities*, s. 3.

48 Tamże, s. 36.

49 Eksperci KE zachęcają do uwzględniania publikacji danych badawczych i innych działań z zakresu otwartej nauki w ewaluacji pracowników naukowych. Por. Working Group on Rewards under Open Science *Evaluation of research careers fully acknowledging Open Science practices: rewards, incentives and/or recognition for researchers practicing Open Science*, Publications Office of the European Union, Brussels 2017, https://publications.europa.eu/en/publication-detail/-/publication/47a3a330-c9cb-11e7-8e69-01aa75ed71a1.

metodologiczne i praktyczne dotyczące postępowania z danymi na kolejnych etapach prac. Stworzenie PZD przydaje się na wielu płaszczyznach: pozwala usystematyzować kwestię danych w projekcie, usprawnia koordynację i organizację działań w zespole, zapobiega utracie danych, przyczynia się do ułatwienia współpracy akademickiej oraz ułatwia ponowne wykorzystanie zgromadzonych materiałów. Jeśli zastanawiamy się nad opublikowaniem danych zebranych bądź wytworzonych w naszym projekcie, sporządzenie planu pozwoli nam na ustrukturyzowanie całego procesu oraz wczesne rozpoznanie i rozwiązanie potencjalnych przeszkód, usprawniając późniejsze działanie.

W planowaniu zarządzania danymi pomocne jest spojrzenie na projekt jako na cykl badawczy, podzielony na mniejsze segmenty, odpowiadające kolejnym etapom prac (zob. schemat). Taki zamysł sprawdzi się zarówno w większym grancie, lecz także w mniejszym przedsięwzięciu; cykl danych w projekcie może też odzwierciedlać np. proces przygotowania artykułu naukowego.



Zagadnienia związane z danymi na poszczególnych etapach procesu badawczego na podstawie raportu *Sustainable and FAIR Data Sharing in the Humanities*[50]

50 N. Harrower, M. Maryl, T. Biro, B. Immenhauser *Sustainable and FAIR data sharing in the humanities*, s. 6.

Powyższy schemat posłuży nam za strukturę do omówienia konkretnych zagadnień w zarządzaniu danymi badawczymi. Systematyczna refleksja nad danymi na poszczególnych etapach projektu ułatwia opracowanie PZD. Warto jednak zaznaczyć, że plan zarządzania danymi pozostaje tylko planem i podlega modyfikacjom oraz aktualizacjom wraz z rozwojem projektu. Na przykład analiza wstępnie zebranych źródeł czy notatek może wskazać na potrzebę dalszej kwerendy i skorzystanie z innych materiałów, które nie zostały uwzględnione na etapie planowania. Tworząc pierwszy szkic PZD w początkowej fazie prac, możemy oczywiście nie znać odpowiedzi na niektóre pytania. Warto wówczas zaznaczyć w dokumencie, że konkretną decyzję zamierzamy podjąć na późniejszym etapie.

W dalszych częściach tekstu omawiamy kolejne etapy tworzenia PZD. Nie jest naszą intencją stworzenie szczegółowego poradnika, ten bowiem wymagałby odrębnego opracowania. Zwracamy za to uwagę na najważniejsze zagadnienia na poszczególnych etapach i odsyłamy do innych opracowań, które dostarczą bardziej praktycznych wskazówek.

Identyfikacja

Identyfikacja danych, czyli określenie, jakie dane zostaną zebrane bądź wytworzone w projekcie, oraz ich wstępny opis (w tym uszczegółowienie ich typu, formatu czy objętości) wypływają z refleksji nad pytaniami badawczymi oraz materiałami, które planujemy wykorzystać. Ten wstępny etap może być szczególnie problematyczny dla humanistów, ponieważ nie przywykli do myślenia o swoich źródłach w kategoriach danych. Opis nie wymaga podawania skomplikowanych informacji technicznych, a zwłaszcza we wstępnym planie najczęściej wystarczające są podstawowe szacunki, np. planowana liczba tekstów do opracowania.

Jak już wcześniej pisaliśmy, nie istnieje jedna uniwersalna typologia danych w badaniach o literaturze, lecz opisując rodzaj danych, można sięgnąć do propozycji podanych w pierwszej części naszego artykułu. Warto przemyśleć kwestię, jak różne dane pojawią się w projekcie, nie skupiając się wyłącznie na głównych materiałach badawczych. Na przykład jeżeli robimy kwerendę na temat przestrzeni w biografii i twórczości Henryka Sienkiewicza, oprócz tekstów literackich i listów godnymi uwagi danymi będą również metadane tych źródeł. Natomiast wybierając odpowiedni format dla naszych danych, warto się zastanowić nad jego interoperacyjnością, tj. czy dany format pozwoli użytkownikom na otwieranie i przetwarzanie plików

w różnych programach. Jeśli mamy wątpliwości, możemy skorzystać z istniejących opracowań[51].

Jednym z największym wyzwań w planowaniu danych jest dla wielu humanistów szacowanie objętości materiałów zbieranych czy wytwarzanych w projekcie. W jaki sposób kalkulować rozmiar danych i jaki poziom dokładności będzie akceptowalny? W wypadku niektórych gotowych schematów planu, np. w formularzu Narodowego Centrum Nauki, pola do umieszczania informacji są niewielkie, w pewnym sensie uniemożliwiając wysoki poziom uszczegółowienia co oznacza, że opis danych będzie krótki. Podczas przywoływanych już warsztatów o danych w IBL PAN uczestnicy zasugerowali, by w takich przypadkach przygotować dwie wersje dokumentu: krótszy plan na potrzeby konkursu oraz dokładniejszy dokument dla zespołu. Można też posłużyć się szacowaniem opisowym – jeśli nie wiemy, ile interesujących nas materiałów zgromadzimy podczas kwerendy, wystarczy opisać, w jakim zbiorze te kwerendy prowadzimy i jakiego typu dane będziemy zbierać, np. wszystkie utwory poetyckie opublikowane w miesięczniku „Twórczość" w latach 1945-1956.

Planowanie

Po wstępnym zidentyfikowaniu danych możemy przejść do planowania. Na tym etapie warto przemyśleć m.in. kwestię współpracy z instytucją, w której realizujemy nasz projekt, i uwzględnić istniejące w niej procedury, np. mechanizmy przechowywania oraz archiwizacji danych czy nadawania im identyfikatorów (DOI itp.). W projekcie należy wyznaczyć opiekuna/opiekunkę danych. Funkcję tę powinna pełnić osoba, która rozumie specyfikę zbieranych i wytwarzanych tekstów czy materiałów. Nie będzie jednak działać w pojedynkę; wsparciem służyć mogą poszczególne działy, a także procedury określone w dokumentacji danej instytucji, np. w polityce otwartości. W niektórych grantach istnieje możliwość uwzględnienia kosztów związanych z organizacją danych w kosztorysie projektu. W planowaniu potrzebnego czasu i koniecznych środków warto zastanowić się nad krokami, które muszą zostać podjęte, aby dane spełniały kryteria FAIR.

51 Na przykład zestawienie preferowanych i akceptowanych formatów dla różnych rodzajów danych w poradniku BUW. Zob. A. Książczak-Gronowska, M. Bogajczyk *Dane badawcze*, Biblioteka Uniwersytecka w Warszawie, Warszawa 2020, https://www.buw.uw.edu.pl//wp-content/uploads/2020/05/DANE-BADAWCZE-1.pdf, s. 18.

Część uczestników i uczestniczek warsztatów o danych badawczych w literaturoznawstwie w IBL PAN zwracała uwagę na rozproszenie ról w instytucjach, gdzie brakuje konkretnej osoby bądź osób do wspierania procesów związanych z danymi na każdym etapie projektu. Kwestiami dotyczącymi danych zajmuje się większa liczba pracowników, co rozmywa odpowiedzialność i sprawia, że badaczowi lub badaczce trudno zdecydować, do kogo zwrócić się z prośbą o pomoc w konkretnej sprawie. Na różnych etapach zaangażowane mogą być: dział informatyczny, biblioteka i archiwum, dział wsparcia badań lub projektów. Tworzone są nowe, specjalistyczne stanowiska pracy dla osób, których zadaniem jest wspieranie badaczy w sprawach związanych z zarządzaniem danymi badawczymi. W kontekście polskim najczęściej pomocy w opracowaniu PZD udzielają bibliotekarze i bibliotekarki; wraz z rozwojem ich stanowisk oraz kompetencji być może poszczególni pracownicy będą mogli pogłębiać wiedzę specjalistyczną dotyczącą danych w konkretnych dziedzinach, jak już się to dzieje w niektórych ośrodkach[52].

Zbieranie/wytwarzanie, strukturyzowanie, przechowywanie

W badaniach literackich często korzystamy z już istniejących źródeł. Warto opisać, w jaki sposób zamierzamy je pozyskać i jaki związek z celami projektu będą miały zebrane materiały. Opisując proces pozyskiwania i wytwarzania danych, warto od razu uwzględnić ewentualne plany digitalizacji. Jeżeli wiemy, z jakich narzędzi zamierzamy skorzystać (np. oprogramowanie do rozpoznawania tekstu), umieszczamy również tę informację wraz z krótkim uzasadnieniem.

Kluczowa, zwłaszcza w większych projektach, będzie struktura przechowywania danych. Nie istnieje jeden złoty system organizowania materiałów i dokumentacji czy nazewnictwa plików, natomiast powinien on być jasny dla wszystkich osób zaangażowanych w projekt i konsekwentnie stosowany. Gdy dłużej pracujemy nad jednym dokumentem, przydatne będzie także wersjonowanie, a w związku z tym dokładne oznaczanie kolejnych wersji plików. Dobrą praktyką może być stworzenie folderu „Archiwum", dokąd trafią starsze wersje dokumentów oraz ukończone pliki, z których już obecnie nie korzystamy.

52 Przykładem służy Uniwersytet Nauk Stosowanych w Utrechcie (HU University of Applied Sciences Utrecht), w którym każdy/a spośród 10 specjalistów ds. zarządzania danymi (data stewards) jest związany/a z innym obszarem wiedzy. Zob. https://bibliotheek.hu.nl/en/researchers/datamanagement/.

Tworząc zbiory danych badawczych, należy rozstrzygnąć kwestię standardu metadanych służących do ich opisu. W tym wypadku kluczowa jest decyzja o miejscu/sposobie publikacji danego zbioru, gdyż to ona determinuje taki standard. Przed utworzeniem zbioru warto przeanalizować standard metadanych wykorzystywany w repozytorium/archiwum, w jakim planujemy zdeponować dane, oraz ustalić dobre praktyki związane z opisywaniem danych w tym standardzie.

Choć w badaniach literaturoznawczych kwestie związane z danymi osobowymi pojawiają się rzadziej niż w naukach społecznych, potrzeba ich uregulowania może pojawić się np. przy przeprowadzaniu wywiadów lub ankiet czy nawet organizacji spotkań wymagających rejestracji uczestników i uczestniczek. W stworzeniu klauzuli informacyjnej dla osób wypełniających ankiety, wyborze miejsca przechowywania danych osobowych bądź podjęciu decyzji wokół ich anonimizacji czy pseudonimizacji zazwyczaj możemy liczyć na wsparcie inspektora ochrony danych w swojej instytucji.

Kwestie związane z własnością danych warto rozwiązać na początku projektu, zwłaszcza jeśli prowadzony jest we współpracy z innymi instytucjami – to z właścicielem danych będą mogły się kontaktować np. osoby zainteresowane ich ponownym wykorzystaniem. Istotny jest także wybór odpowiednich, najlepiej otwartych licencji (np. Creative Commons). Możemy opisać ograniczenia dotyczące ponownego wykorzystania danych (np. materiałów pochodzących od osób trzecich). Wybierając licencję dla własnego zestawu danych, warto szczególną uwagę zwrócić właśnie na ograniczenia prawne nałożone na materiały, których używamy. Mimo dobrych chęci nie możemy udostępnić na otwartej licencji skanów artykułów, które nie należą do nas, kiedy wszelkie prawa są zastrzeżone.

„Środki kontroli jakości", o których wspomina się na tym etapie cyklu danych, mogą początkowo kojarzyć się bardziej z produkcją przemysłową, lecz w tym wypadku stanowią zestaw kryteriów minimalnych, jakie powinny spełniać nasze materiały. Na przykład skanując teksty źródłowe, powinniśmy ustalić nie tylko format zeskanowanych plików, lecz także jakość skanu (wyrażoną w dpi), odczytanie warstwy tekstowej oprogramowaniem OCR (można też przyjąć pożądaną dokładność odczytu wyrażoną w procentach). Tylko pliki spełniające te wymagania będą wykorzystane do badań. W zespole możemy też ustalić wspólny system notowania podczas kwerendy (tak by każda osoba mogła korzystać z materiałów zebranych przez współpracowników i współpracowniczki), jak również konkretny i konsekwentny system cytowań. Na przykład opracowując innowacyjne studia przypadku

w komunikacji naukowej[53], pracownicy IBL PAN posługiwali się wspólnie ustalonym cyklem pracy oraz schematem notatek, co pomagało sprawnie zebrać dane o konkretnych przykładach.

Sen z powiek badaczy spędza wizja utraty cierpliwie zebranych lub wytworzonych z trudem materiałów. Zapobiec temu może skrupulatne zaplanowanie przechowywania danych oraz metadanych. Kierując się zasadą „3-2-1", tworzymy trzy kopie danych, które przechowujemy na dwóch osobnych urządzeniach, a jedno z nich znajduje się poza miejscem pracy. Procedury techniczne (w tym tworzenie kopii zapasowych) są często regulowane na poziomie instytucji, zalecamy więc kontakt z działem odpowiedzialnym za systemy informatyczne i bezpieczeństwo danych. W zależności od kwestii związanych z prywatnością danych lub naszą chęcią kontynuacji pracy z ich wykorzystaniem możemy ustawić bardziej rygorystyczne ograniczenia w dostępie do naszych danych.

Deponowanie, cytowanie, udostępnianie

Jednym z najważniejszych elementów planu zarządzania danymi badawczymi jest kwestia udostępniania danych. Jeżeli się na to zdecydujemy, należy opisać proces selekcji danych, które zamierzamy zdeponować. Czasami z różnych przyczyn po zakończeniu projektu należy zniszczyć część materiałów, ponieważ zawierają dane wrażliwe czy stanowią zapiski o charakterze roboczym. Wówczas taka informacja również powinna zostać zawarta. Jeśli dane nie będą udostępniane, w planie warto zaznaczyć powody naszej decyzji, a także ewentualną możliwość późniejszego otwarcia danych, np. po wygaśnięciu praw autorskich źródeł lub po analizie materiałów w naszym kolejnym projekcie. Natomiast jeżeli w badaniu uczestniczyły osoby z zewnątrz i ich dane miałyby być udostępnione (np. transkrypcje z wywiadów, które z nimi przeprowadziliśmy), należy w tym celu zebrać odpowiednie zgody.

Aby tworzone zbiory danych badawczych mogły być ponownie wykorzystywane – zarówno w zespole badawczym, jak i poza nim – muszą być one opatrzone właściwymi metadanymi w taki sposób, aby można było je odnaleźć i zastosować w kolejnych działaniach badawczych. Przy deponowaniu

53 Prace prowadzono w ramach zadania poświęconego przemianom w pisaniu tekstów naukowych w humanistyce i naukach społecznych w pakiecie roboczym 6 (Work Package 6) projektu OPERAS-P (Preparing open access in the european research area through scholarly communication).

danych w repozytoriach czy archiwach danych będziemy mogli określić metadane danego zbioru choćby w zakresie licencji lub tematyki. Często możliwe jest stworzenie krótkiego narracyjnego opisu danego zestawu, przypisanie zbioru do projektu, w ramach którego powstał. Wreszcie – na co warto zwrócić szczególną uwagę – usługi deponowania danych badawczych pozwalają na uzyskanie tzw. trwałych identyfikatorów (PID, *persistent identifiers*), czyli trwałych odnośników do udostępnionego zbioru danych.

W planie powinno się wstępnie określić planowane miejsce lub miejsca udostępnienia danych, jak omawiane wyżej repozytoria danych, dyski twarde instytucji czy samodzielne platformy. Wśród kryteriów wyboru repozytorium znajdują się m.in.: umożliwienie trwałego przechowywania danych, przestrzeganie zasad FAIR, wykorzystanie trwałych identyfikatorów (np. DOI). Pomocne będzie stworzenie wewnętrznej procedury deponowania – jeżeli decydujemy się na publikowanie danych, czy będziemy robić to systematycznie w trakcie trwania projektu, czy zbiorczo po jego zakończeniu.

Upowszechnianie

Po opublikowaniu danych możemy udostępnić informację o nich potencjalnym odbiorcom. Wśród proponowanych działań można wskazać promocję w mediach społecznościowych i na stronach internetowych (własnej, instytucji czy zespołu) oraz zindywidualizowane informowanie innych badaczy i badaczek (poprzez wiadomość e-mail, na spotkaniach zespołów czy konferencjach naukowych). Dzieląc się informacją, że rozpowszechniliśmy dane, nie tylko umożliwiamy innym zapoznanie się z efektami naszej pracy (a może wręcz ich ponowne wykorzystanie), lecz także promujemy dobre praktyki wśród przedstawicieli naszej dziedziny.

Ciekawym rozwiązaniem alternatywnym wobec umieszczania zestawów danych w repozytoriach jest publikacja w tzw. *data journals*, czasopismach naukowych poświęconych właśnie danym badawczym (np. „Data in Brief", „Scientific Data", „Research Data Journal for the Humanities and Social Sciences", „Journal of Open Humanities Data"). Problemem zidentyfikowanym przez uczestników i uczestniczki warsztatów okazało się wybranie odpowiedniego *data journal*, by spełniał wymagania związane z punktacją oraz przypisaniem odpowiedniej dziedziny. Przydatne będzie również przemyślenie kwestii powiązania artykułu z wykorzystanymi do niego danymi, np. udostępnianie ich wspólnie lub cytowanie opublikowanego zbioru w artykule, by czytelnicy mogli się do niego odnieść.

Konkluzje

Gdy projektowaliśmy warsztaty z danych badawczych dla literaturoznawców, staraliśmy się przedstawić powyższe zagadnienia podczas trzech spotkań poświęconych kolejno identyfikacji, miejscu publikacji danych i sporządzaniu PZD. Już na pierwszych zajęciach po wprowadzającej prezentacji jeden z uczestników zabrał głos mniej więcej w takim tonie: dobra, dobra, nie musimy wdawać się w szczegóły, czy możecie państwo po prostu powiedzieć, jak sporządzić PZD, bo po to tu przyszedłem. Postawa ta wynika z przeświadczenia, że temat zarządzania danymi badawczymi oraz sporządzania planów to zbędna biurokracja narzucana dziś przez grantodawców. W tym tekście próbowaliśmy pokazać, że jest zupełnie inaczej.

Myślenie w kategoriach danych to nie radykalna rewolucja w naszej pracy – pamiętamy chyba korzystanie z fiszek? – lecz uzupełnienie naszej metodologii o systematyczne, ustrukturyzowane zasady posługiwania się materiałem badawczym, by ułatwić pracę sobie i innym. Nie postulujemy, by wszyscy badacze literatury stali się nagle specjalistami w zakresie danych badawczych, a tym bardziej znawcami skomplikowanych aspektów technicznych, którymi powinien się zajmować specjalistyczny personel w zespole czy jednostce badawczej. Chcielibyśmy raczej, by problematyka danych w literaturoznawstwie została zaakceptowana jako element procesu badawczego w naszej dyscyplinie, a docelowo, jako taka, została uwzględniona w procesie ewaluacji działalności naukowej literaturoznawców.

Zbierając zatem najważniejsze konkluzje, warto myśleć o swoich źródłach jako danych, by w pełni wykorzystać dostępne możliwości technologiczne oraz dopuścić ponowne ich wykorzystanie przez innych. Drogą do tego celu jest stosowanie ogólnie przyjętych standardów i wytycznych FAIR, które próbowaliśmy przybliżyć w tym artykule, kierując także do innych opracowań. Kolejne instytucje opracowują dziś wewnętrzne wytyczne i rekomendacje dotyczące danych badawczych. Chcieliśmy podzielić się naszym doświadczeniem, by ułatwić tworzenie takich dokumentów w naszej dyscyplinie i wpłynąć tym samym na zakres danych badawczych dostępnych dla literaturoznawców.

Informacja o finansowaniu pracy z grantu:

Część badań do tego artykułu przeprowadzono w ramach projektów:

- Preparing open access in the european research area through scholarly communication (OPERAS-P), projekt H2020 #871069.
- Shaping interdisciplinary practices in Europe (SHAPE-ID), projekt H2020 #822705.
- Transforming Research through Innovative Practices for Linked interdisciplinary Exploration (TRIPLE), projekt H2020 #863420.
- Skamandrycka triada na emigracji. Edycja listów Jana Lechonia, Kazimierza Wierzyńskiego i Mieczysława Grydzewskiego, publikacja książkowa i cyfrowa, grant NPRH, nr 0173/NPRH4/H1a/83/2015.
- Dramat polski. Reaktywacja/Kontynuacja, grant NPRH, IBL PAN, nr 0139/NPRH6/H11/85/2018.
- Sienkiewicz ponowoczesny – laboratorium cyfrowe, grant NPRH, nr rejestracyjny 2aH 15 0195 83.
- Open Badge Ecosystem for the Recognition of skills in Research Data management and sharing (OBERRED), projekt Erasmus+, nr 2019-1FR01-KA203-063056..

Autorzy serdecznie dziękują za wszystkie komentarze i sugestie. uczestnikom warsztatów *Dane badawcze w badaniach literackich*, zorganizowanych w IBL PAN jesienią 2020.

Abstract

Maciej Maryl, Marta Błaszczyńska, Bartłomiej Szleszyński, Tomasz Umerle
THE INSTITUTE OF LITERARY RESEARCH OF THE POLISH ACADEMY OF SCIENCES (WARSAW)
Research Data in Literary Studies

This article presents the methodological foundations and practical implications of research data use in literary studies. Drawing on their own research and consultations with Polish literary scholars, the authors propose a typology of research data in literary studies. They go on to discuss four types of data publication in this discipline: file depositing, multimedia monographs, digital scholarly editions and documentary publications. The final section of the article offers practical suggestions for handling data at various stages of a literary studies project.

Keywords

literary studies, research data, methodology, research