DOI: http://dx.doi.org/10.12775/SE.2020.0001

ARTYKUŁY I ROZPRAWY

Bogdan Hojdis, Adam Cankudis

Cyfrowe edycje literackie i naukowe jako element cyfrowej humanistyki

Uniwersytet im. Adama Mickiewicza w Poznaniu, kontakt:
hojdis@amu.edu.pl, ORCID ID: 0000-0003-2130-1462;
kontakt: adacan@amu.edu.pl, ORCID ID: 0000-0002-0783-3457

Sztuka Edycji 1/2020
ISSN 2084-7963 (print)
ISSN 2391-7903 (online)
s. 7–15

Joris van Zundert, badacz i programista z Wydziału Studiów Literackich Instytutu Historii Holandii, napisał przed trzema laty, że „tekstologia i edytorstwo naukowe w swym cyfrowym kształcie należą do szerszego pola humanistyki cyfrowej, pola samego w sobie zbudowanego na interdyscyplinarności. […] W tym miejscu humanistyka cyfrowa nabywa swej innowacyjnej siły lub przynajmniej – obietnicy tej siły”[1]. Inna badaczka od lat związana z humanistyką cyfrową, Elena Pierazzo, stwierdziła wprost, że „cyfrowe edycje naukowe uważa się za jeden z klejnotów koronnych humanistyki cyfrowej”[2]. Cyfrowe edycje niewątpliwie stanowią istotny element humanistyki cyfrowej, ale nie są z nią, jak widać, tożsame. Czym jest zatem owo szerokie pole, zwane też poszerzonym polem (*expanded field*) czy wielkim namiotem (*big tent*)[3] humanistyki cyfrowej?

Jedną z pierwszych konferencji uniwersyteckich w Polsce, a jak twierdzą organizatorzy – pierwszą[4], na temat humanistyki cyfrowej zorganizowano w październiku 2012 roku na Uniwersytecie Marii Curie-Skłodowskiej w Lublinie. Konferencja została podsumowana publikacją o znaczącym tytule *Zwrot cyfrowy w humanistyce*[5].

Podobnym punktem zwrotnym w zachodnim świecie akademickim stała się publikacja *A Companion to Digital Humanities*[6] z 2004 roku. Jak piszą redaktorzy

tomu – Susan Schreibman, Ray Siemens, John Unsworth – we wstępie: „[w tej kolekcji] po raz pierwszy zostało zebrane szerokie grono teoretyków i praktyków […], by rozważyć humanistykę cyfrową jako dyscyplinę samą w sobie”[7] (tłumaczenie własne). Można powiedzieć, że zwrot właśnie się dokonał – zespół praktyk, teorii, metodologii został określony dyscypliną (nawet jeśli tylko rozważaną) i przypieczętowany terminem „humanistyka cyfrowa” (*Digital Humanities*), która to nazwa po raz pierwszy pojawia się właśnie w tytule tej publikacji[8]. Do dziś toczą się dyskusje, czy humanistyka cyfrowa jest faktycznie osobną dyscypliną, czy też po prostu zestawem nowych praktyk i narzędzi stosowanych do tych samych od wieków zadań i problemów humanistyki. Niewątpliwie na naszych oczach dokonuje się jakaś istotna zmiana, niezależnie, czy widzimy ją jako zwrot, czy ewolucyjną ciągłość, przejawiająca się w sposobie, w jaki dziś pracują humaniści (i edytorzy), oraz w narzędziach, których używają. Jak konstatują redaktorzy publikacji *Digital Scholarly Editing*:

> Oczywiste jest, że coś się radykalnie zmienia w świecie edycji: sposób, w jaki pracujemy, narzędzia, których używamy, by wykonać tę pracę, i pytania badawcze, na które próbujemy odpowiedzieć – wszystko to się zmieniło, niekiedy nie do poznania, w porównaniu do starszego, opartego na druku, trybu pracy[9].

Do dziś toczą się dyskusje, czy humanistyka cyfrowa jest osobną dyscypliną

I dalej zadają ważkie pytanie, czy zmiana, która się dokonuje, to tylko zmiana powierzchowna, czy też jest to zmiana istotna, dzięki której będzie można wykonać prace, które do tej pory były niewykonalne:

> Czy my po prostu wlewamy „stare wino do nowych butelek”, czy robimy coś, co nigdy nie zostało zrobione – a właściwie, nie było wykonalne – przedtem?[10]

Być może nigdy nie zostanie uzgodniona ostateczna, zadowalająca wszystkich definicja terminu, co przecież nie jest niczym zaskakującym, jeśli mówimy o pojęciach wyrosłych z praktyki, a nie z przemyślanej koncepcji teoretycznej. Nie bez przyczyny **w Polsce równolegle funkcjonują właściwie przeciwstawne opinie na temat funkcji humanistyki cyfrowej**:

> Humanistyka cyfrowa staje się dzisiaj nowym paradygmatem badań zdobywającym sobie coraz większe uznanie na świecie[11],

oraz

> Różnorodność projektów, którą obserwujemy na gruncie polskim, doskonale koresponduje z rozproszeniem charakterystycznym dla projektów z zakresu humanistyki cyfrowej na całym świecie. Podkreśla to tylko to, że cyfrowość nie jest wcale nowym paradygmatem badań, tylko narzędziem do realizacji przeróżnych celów badawczych często odmiennych dyscyplin[12].

Jeśli spojrzeć historycznie, to początków naszej humanistyczno-cyfrowej „dyscypliny” można doszukiwać się znacznie wcześniej niż w 2004 roku. Jak powszechnie się uważa[13], u jej podstaw legły prace nad konkordancją wyrazów z dzieł św. Tomasza z Akwinu prowadzone przez jezuitę Roberto Busa. W 1946 roku Busa postanowił stworzyć *Index Thomisticus*. W krótkim czasie zdał sobie sprawę, że do przetworzenia tekstów zawierających ponad dziesięć milionów słów będzie potrzebował maszyn. W 1949 roku rozpoczął poszukiwania możliwości rozwiązania problemu na amerykańskich uczelniach. Ostatecznie nawiązał współpracę z IBM, poniekąd dzięki szczęśliwemu zbiegowi okoliczności i swojemu uporowi, gdyż początkowo IBM oceniał projekt Busa jako niewykonalny (*sic!*). Rozpoczęło się żmudne przygotowanie tekstu – wstępne opracowanie, lematyzacja, korekta i kodowanie na kartach perforowanych (w drugiej połowie lat pięćdziesiątych zaczęto w projekcie wykorzystywać taśmy magnetyczne)[14].

Znamienne, że narodzin *digital humanities* doszukujemy się w przetwarzaniu komputerowym wykorzystanym na potrzeby leksykograficzne (i ogólniej – językoznawcze). Językoznawcy leksykografowie mają bowiem na ogół do czynienia z ogromną masą materiału, który jest jednak policzalny i wydaje się konkretny, choć bywa skomplikowany (zważywszy na powiązania między elementami), przez co często trudny do analizy i prezentacji w tradycyjnej drukowanej formie. „Tylko komputerowy wykaz syntaktycznych korelacji może udokumentować, jaką myśl chciał autor wyrazić [tym] słowem” – przyznał Roberto Busa w przedmowie do *A Companion to Digital Humanities*[15].

Ostatecznie wydane drukiem w latach 1974–1980 dzieło zespołu Busa zajęło ponad sześćdziesiąt tysięcy stron w pięćdziesięciu sześciu tomach formatu encyklopedii[16]. Dzisiaj *Index* dostępny jest również online[17]. Zamierzenia Busa zdecydowanie wykraczały poza leksykografię, dlatego jeszcze w latach pięćdziesiątych rozpoczął nieukończony za życia projekt hermeneutyki komputerowej. Podstawą tej szerokiej koncepcji były jednak rudymentarne prace wykonane podczas przygotowywania *Index Thomisticus*.

Również w Polsce pierwsze przedsięwzięcia polegające na użyciu maszyn obliczeniowych w humanistyce były związane z językoznawstwem. Przykładem jest praca Marii Steffen-Batogowej dotycząca automatyzacji transkrypcji fonematycznej[18]. Przypomniano w niej, że już w 1969 roku Witold Doroszewski, jako pierwszy polski lingwista, wskazywał możliwości przekształcenia polskiego tekstu ortograficznego na zapis fonematyczny przez maszynę. Wiosną 1971 roku z powodzeniem udało się przeprowadzić próbne transkrypcje na komputerze Odra 1204 na podstawie reguł dostarczonych przez Steffen-Batogową, a zaprogramowanych i wykonanych przez Mieczysława Warmusa. Jak stwierdziła autorka, automatyczna transkrypcja fonematyczna jest podstawą rozwiązywania wielu zagadnień lingwistycznych, zwłaszcza natury statystycznej, w których wymagane jest przetworzenie obszernych zasobów tekstowych. Bez tej podstawowej pracy nie można podjąć wielu dalszych naukowych przedsięwzięć. Otwierają się nowe możliwości – automatyczna synteza mowy z tekstu ortograficznego i zapis mowy do tekstu ortograficznego. Trzeba bowiem uświadomić sobie, że nawet we współczesnych systemach przetwarzania języka naturalnego – czy to w systemach kognitywnych, czy sztucznej inteligencji – analizuje się wypowiedzi przekształcone do tekstu.

Oba wymienione projekty z początków cyfrowej ery w humanistyce pokazują charakterystyczny sposób użycia komputera w tej dziedzinie[19]. Widać wyraźnie, że nie intensywność komputacji, lecz złożoność danych – ich ilość i wielowymiarowość – są podstawowym problemem, który systemy informatyczne pozwalają rozwiązać. To kategoryzowanie, grupowanie, odnajdywanie powiązań między elementami zbioru wedle przyjętego modelu dziedziny są zadaniami, dla których w humanistyce wykorzystuje się moc procesorów i logikę zaklętą w algorytmach. Oba projekty mają też charakter rudymentarny – przeniesienie z formy „analogowej" do cyfrowej stanowi (lub może stanowić, jak w *Index Thomisticus*) punkt wyjścia dalszych przedsięwzięć – punkt konieczny, nawet „narzędziowy", co w rzeczywistości cyfrowej tak naprawdę oznacza możliwość ponownego wykorzystania materiału cyfrowego (bez uszczerbku dla oryginału) i wytworzonych narzędzi.

Niewątpliwie trwa u nas, wspomniane w opinii Marcina Werli i Macieja Maryla, rozproszenie projektów z zakresu humanistyki cyfrowej. Podobnie jak ich różnorodność i swoista atomizacja wynika z zastosowania cyfrowych narzędzi do rozmaitych celów badawczych, tak również cyfrowe edytorstwo można odnieść do literatury, badań naukowych i edukacji (w tym b-learningu oraz e-learningu) – choćby dlatego, że dzisiaj właściwie każde dzieło korzystające z tworzywa językowego powstaje najpierw w postaci cyfrowej, publikowane jest nierzadko zarówno w wersji cyfrowej, jak i analogowej; a ponadto coraz częściej jest wytworem pokolenia *born digitally* albo po prostu – produktem dla tej generacji przeznaczonym[20].

Należy również przypomnieć, że do zintensyfikowania dygitalizacji w ostatnich latach przyczyniły się rozmaite, nie do końca spójne i skoordynowane, legislacje oraz programy:

- zmiany z 2015 roku w ustawie o zasadach finansowania nauki, przez co środki MNiSW (tzw. DUN) na udostępnianie zasobów bibliotecznych w formie elektronicznej stały się jednym z trzech źródeł dofinansowania bibliotek naukowych,
- następne edycje programu Kultura Cyfrowa MKiDN,
- projekt OMNIS realizowany przez Bibliotekę Narodową.

Niestety, **efektem tego „narodowego skanowania" są niemal wyłącznie reprinty graficzne**. Dotąd trafiają one przede wszystkim do regionalnych bibliotek cyfrowych, w których często brak jednolitej deskrypcji (np. między RCIN i WBC), a zamiana danych graficznych na dane tekstowe spoczywa na barkach niewielkiej grupy lokalnych administratorów oraz informatyków. Szczęśliwie dzięki wysiłkom owych osób, wspartych oprogramowaniem do automatycznego rozpoznawania pisma (Optical Character Recognition – OCR), coraz więcej jest dokumentów, które istnieją i funkcjonują także w postaci tekstowej, przez co oprócz warstwy obrazu dostępna jest też warstwa tekstu. Ponadto przez lata znacząco poprawiała się jakość narzędzi OCR, choć należy zdawać sobie sprawę, że proces automatyczny nie może pozostać bez kontroli człowieka i żmudnej korekty wykonanej przez operatora czy redaktora. Z tego powodu wiele dokumentów dostępnych w sieci jest na etapie tzw. brudnego OCR, w którym może znajdować się sporo błędnych odczytań względem oryginalnego tekstu.

Programy OCR dość dobrze radzą sobie z drukami współczesnymi – głównie dwudziestowiecznymi i nowszymi. W najnowszych wersjach owych aplikacji zastosowano elementy uczenia maszynowego[21], „wytrenowano" oprogramowanie pod kątem różnych krojów pisma oraz układów występujących w dokumentach, dzięki czemu wiele materiałów, które jeszcze niedawno uznawano za trudne w automatycznym rozpoznawaniu, jest dzisiaj zamieniana na postać znakową z dobrym rezultatem. Należy przy tym pamiętać, że w procesie tym jednym z podstawowych problemów pozostaje odróżnienie tego, co jest informacją, od tego, co jest szumem, zakłóceniem, brudem. Przede wszystkim nie najlepszy stan zachowania skanowanego materiału może stanowić poważną przeszkodę w poprawnym rozpoznawaniu znaków. Projekty borykające się z takimi problemami, jak np. Narodowy Fotokorpus Języka Polskiego[22], mają opracowane całe katalogi „standardowych" błędów OCR, co pozwala im te uchybienia poprawiać automatycznie w znakowej wersji tekstu. Następnym problemem może być jakość wykonywanych skanów. Kurz, interferencje świetlne, niepoprawne nastawy urządzenia i/lub oprogramowania, zarysowania płyty skanera, nie mówiąc już o odciskach palców – potrafią znacząco utrudnić działanie OCR, a czasem dać efekt *false positives*, czyli rozpoznanie znaków, których w tekście faktycznie nie ma. W szczególności obrazy cyfrowe otrzymane z mikrofilmów ze względu na swoją niską jakość praktycznie nie nadają się, by poddać je procesowi OCR. Rodzi to potrzebę pilnej dygitalizacji druków wykonywanych na papierach kwaśnych, np. dziewiętnastowiecznej prasy. O ile bowiem manuskrypty i druki dawne poddawane są często pieczołowitej konserwacji, o tyle druki nowe już takiej ochrony nie mają. Jako że powstały na materiale ulegającym szybkiej biodegradacji, to wkrótce nie będzie czego skanować. Mimo dostępności mikrofilmów nie będzie można podjąć na tym materiale żadnych nowych przedsięwzięć z zakresu humanistyki cyfrowej.

> Programy OCR dość dobrze radzą sobie z drukami współczesnymi

Dostępność skanerów biurkowych, łatwość obsługi dołączanego do nich oprogramowania kusi, by dygitalizację wykonywać niskim nakładem sił i środków. Jakość tak otrzymanych skanów często pozostawia jednak wiele do życzenia, a trzeba pamiętać, że raz wykonana cyfrowa kopia oryginalnego dokumentu będzie używana przez dziesięciolecia, bowiem dysponenci środków finansowych rzadko zgadzają się na ponowne przeskanowanie opracowanej grupy dokumentów, słusznie zakładając, że sfinansowana i wykonana praca zostanie wykorzystana wielokrotnie. Nie należy więc oszczędzać na sprzęcie i kształceniu operatorów. Równie istotne jest działanie w ramach dobrze zorganizowanego systemu przepływu prac, w którym materiał źródłowy będzie odpowiednio przygotowany, a skany przejdą rzetelną weryfikację. Przykładem takiej dobrej praktyki jest proces dygitalizacji w ramach projektu Platformy Cyfrowej Biblioteki Kórnickiej. Celem przedsięwzięcia było m.in. opracowanie i zdygitalizowanie około sześciu tysięcy obiektów ze zbiorów Biblioteki Kórnickiej PAN, wybranych z kolekcji rękopisów średniowiecznych, dokumentów pergaminowych, rękopisów staropolskich, inkunabułów, szesnastowiecznych druków oraz wydawnictw kartograficznych. Ustalono zestaw i kolejność czynności, które objęły ocenę stanu zachowania obiektu, konieczne zabiegi i harmonogram ich wykonania, a w końcu – dokumentację przeprowadzonych prac. Każdy obiekt został wyposażony w metryczkę przedstawiającą wyniki następnych kroków. Pozwoli ona planować dalsze działania konserwatorskie już po zakończeniu projektu. Prymarne wobec dygitalizacji było czyszczenie obiektów przed skanowaniem, co niewątpliwie miało wpływ na dobrą jakość cyfrowych kopii. Być może powinien powstać ogólnopolski katalog dobrych praktyk dygitalizacyjnych, by unikać błędów przy ustalaniu ich na nowo w każdym projekcie. Niewątpliwie katalog takich praktyk powinien też zostać poddany pod rozwagę użytkowników, tj. badaczy humanistów. Jeśli bowiem w trakcie dygitalizacji pewne informacje zachowujemy, a inne musimy odrzucić, to warto zastanowić się nad przesłankami naszych decyzji.

Wspomniano, że OCR dobrze radzi sobie ze współczesnymi drukami, a zastosowanie uczenia maszynowego tę sytuację tylko poprawiło. Niestety, uczenie maszynowe wymaga znacznej liczby wzorców treningowych (*ground-truth*). W przypadku wczesnych druków oraz manuskryptów jest to więc technologia, której nie udaje się zastosować. W różnych ośrodkach trwają prace nad rozwiązaniem tego problemu. Na przykład na uniwersytecie w Uppsali w Szwecji prowadzone są eksperymenty nad automatycznym rozpoznawaniem pisma ręcznego z manuskryptów, z kolei Jacek Tlaga z Biblioteki Narodowej podejmuje próby ekstrakcji tekstu z polskich druków dawnych z wykorzystaniem głębokiego uczenia maszynowego, a jako materiał wzorcowy (*ground-truth*) wykorzystuje edycje dyplomatyczne, już

istniejące ręczne anotacje i transkrypcje. W planach jest także współpraca z projektami pracującymi nad korpusami i słownikami dawnej polszczyzny. Tekst jako taki oczywiście nadaje się do przeglądania przez człowieka. Zdygitalizowane i przetworzone do tekstu dokumenty, czy to skorygowane, czy nie, nadają się też do pełnotekstowego przeszukiwania, jakkolwiek wyniki zależą w dużej mierze od inwencji szukającego, ponieważ mechanizm wyszukiwawczy dopasowuje (często z uwzględnieniem metaznaków) wyszukiwany ciąg do udostępnionego tekstu według zasady znak do znaku. Korzystając z dostępnych w systemach narzędzi przeszukujących, możemy wszak odnaleźć fragmenty odpowiadające wzorcowemu ciągowi. System informatyczny nie rozumie jednak znaczenia przetwarzanych znaków[23]. Występujące obok siebie trzy litery m, a, m (poprzedzone i zakończone spacją) nie stanowią dla komputera miejsca, w którym kryje się potencjalnie czasownik „mieć" w pierwszej osobie liczby pojedynczej lub rzeczownik rodzaju żeńskiego „mama" w dopełniaczu liczby mnogiej. Zatem w efekcie wyszukiwania tekstowego użytkownik nie znajdzie tego miejsca, gdy wpisze do wyszukiwarki słowo „mieć" lub „mama". A przecież to dopiero wierzchołek góry lodowej – pozostają osoby, miejsca, daty, wydarzenia, pojęcia, wzajemne powiązania fragmentów wewnątrz tekstu oraz powiązania z tekstami innymi (intertekstualne). Aby było możliwe wyszukiwanie oparte na znaczeniu (również leksemu w określonym kontekście), potrzebna jest dodatkowa praca, polegająca na kodowaniu znaczeń w tekście.

Systemy i pomysły rozwiązania tego problemu istnieją nie od dziś. Najpowszechniejszą bodaj metodą jest anotowanie tekstu w takim czy innym standardzie. W środowisku humanistycznym powszechnie stosuje się format Text Encoding Initiative (TEI), który jest zestawem reguł i zaleceń zdefiniowanych przez konsorcjum TEI w wytycznych, a najnowszą wersją jest propozycja piąta (TEI: P5 Guidelines)[24]. Standard TEI wyrósł z potrzeby uzgodnienia sposobu, w jaki zapisywane są informacje o dokumencie i jego treści, żeby zwiększyć interoperacyjność, głównie możliwości bezproblemowej wymiany danych między projektami, ale także wspomóc w praktyce tworzenie tekstów i ujmowanie danych oraz wesprzeć niezależne od aplikacji przetwarzanie danych. Aby zrealizować te funkcjonalności, przyjęto następujące cele:

– dostarczenie standardowego formatu do wymiany danych,
– dostarczenie pomocy w kodowaniu tekstu w tym formacie,
– wsparcie dla wszelkiego rodzaju sposobów kodowania cech wszystkich rodzajów tekstów, które studiowane są przez badaczy,
– niezależność od aplikacji.

W wyniku przyjętych założeń podjęto różne decyzje projektowe: zaimplementowanie standardu w języku XML (od wersji P4, wcześniej był to SGML), wybór Unicode jako standardu kodowania znaków, zdefiniowanie szerokiego zestawu znaczników umożliwiającego kodowanie różnych spojrzeń na tekst, w tym kodowania alternatywne tych samych cech tekstowych oraz mechanizm pozwalający na rozszerzanie schematu przez użytkownika[25]. Dzięki temu TEI pozwala nadać opisywanej treści strukturę, czyniąc tekst możliwym do przetwarzania maszynowego. Bez owej struktury, z punktu widzenia systemu informatycznego, tekst będzie ciągiem nieskategoryzowanych znaków. Dopiero oznaczając fragmenty tekstu informacją semantyczną, jesteśmy w stanie poddać tekst zautomatyzowanej komputerowej obróbce.

Dla przykładu, w tekstach, zwłaszcza dawniejszych czy też stylizowanych, spotyka się różne sposoby tekstowego zapisu dat:

– dwunastego stycznia Roku Pańskiego MDLXXVIII,
– 4 września roku Pańskiego 1927,
– w noc stanu wojennego Anno Domini 1981,
– (przyjechał na) Święta Wielkanocne[26].

Większość osób nie będzie miała problemu ze zrozumieniem i z umieszczeniem na osi czasu powyższych fragmentów określających daty, choć ostatni przykład może sprawić nieco trudności ze względu na ruchomość świąt wielkanocnych i wyjęty z kontekstu stanie się niemożliwy do precyzyjnego odczytania. Bez utworzenia specjalizowanego algorytmu rozpoznającego powyższe zapisy dla systemu informatycznego fragmenty te pozostaną zwykłym ciągiem tekstowym, bez specjalnego znaczenia. Dopiero wprowadzenie metainformacji interpretującej fragment jako datę oraz przedstawienie tej daty w ustandaryzowany sposób da możliwość przetwarzania tych ciągów jako jednostek informacyjnych. Opisując powyższe fragmenty tekstu w formacie TEI, należałoby użyć znacznika *date* wraz z atrybutem *when* i datą zapisaną w standardzie ISO 8601:

– <date when="1578-01-12">dwunastego stycznia Roku Pańskiego MDLXXVIII</date>
– <date when="1927-09-04">4 września roku Pańskiego 1927</date>
– <date when="1981-12-13">w noc stanu wojennego Anno Domini 1981</date>

– przyjechał na <date when=”1999-04-04”>Święta Wielkanocne</date>

Tak zakodowana data może być teraz automatycznie ekstrahowana, odnajdywana, umieszczana na osi czasu, porównywana czy grupowana (według roku, miesiąca, dnia) z innymi datami, może stać się zmienną w obliczaniu interwałów czasowych itd. Osobom niezaznajomionym z takimi językami znacznikowymi, jak XML czy HTML, należy wyjaśnić, że programy interpretujące, renderujące dokument otagowany, przetwarzają znaczniki (tj. poddają interpretacji narzuconej przez znacznik otoczony nim tekst), ale ukrywają postać źródłową znacznika. Dzięki temu czytelnik ma dostęp do tekstu oryginalnego niezdeformowanego informacjami technicznymi.

Zautomatyzowane przetwarzanie pozwala m.in. przeszukiwać dokument według kategorii i informacji rekodowanej do standardu[27], specjalnie formatować wybrane kategorie informacji, wreszcie – wydobywać określone semantycznie fragmenty, aby je analizować, „przenieść” do innej edycji bądź utworzyć czasową wariację edycji przez czytelnika. Oczywiście wiele z tych operacji zależy od aplikacji działającej na dokumencie TEI, ale dokument anotowany w tym formacie udostępnia aplikacji ustaloną strukturę, która czyni przetwarzanie możliwym. TEI jest obecnie zaimplementowany w metajęzyku XML, dlatego do edycji, przetwarzania i prezentacji dokumentów w tym formacie można wykorzystać standardowe narzędzia dla XML-a (edytory, procesory, XML-owe bazy danych, systemy prezentacyjne). Powstały również takie wyspecjalizowane narzędzia jak np. TEI Publisher[28], który jest systemem edycyjno-publikacyjnym, stworzonym na podstawie systemu bazodanowego eXist-db. Właśnie dzięki standaryzacji opisu dokumentów można budować aplikacje niezależne od konkretnych projektów edycyjnych i korzystać w różnych projektach z tych samych programów szerokiego zastosowania, co znacznie przyspiesza tworzenie cyfrowych edycji oraz zmniejsza koszty ich wytworzenia.

Niewątpliwie format TEI koncentruje się na tekście – od pojedynczych inskrypcji po kolekcje dzieł. Podobnie zatem jak XML (również wyłącznie tekstowy), sam w sobie nie daje możliwości osadzenia obrazu (np. skanu dokumentu źródłowego), pozwala jednak na załączenie obrazu przez linkowanie, a nawet synchronizację pozycji między obrazem a tekstem. Odmienne podejście proponuje wywodzący się ze środowisk GLAM[29] standard IIIF[30], który najbardziej znany jest ze sprawnego dostarczania wysokiej rozdzielczości obrazów w środowisku webowym. Przy tradycyjnej metodzie serwowania obrazów w usłudze WWW plik z obrazem przesyłany jest z serwera do klienta w całości. Przesyłanie bardzo dużych obrazów o wysokiej jakości zajmuje dużo czasu, obciąża tym samym łącza i wymaga przetwarzania zarówno na serwerze, jak i komputerze klienta. Użytkownik otrzymuje sporą porcję danych, często tylko po to, by pobieżnie obejrzeć obraz i porzucić dużą ilość otrzymanych danych, które wpływają na obciążenie i responsywność urządzenia klienckiego (również w przyszłości, bo przeglądarki internetowe zazwyczaj zapisują pobrane pliki w pamięci podręcznej). Ponadto użytkownik otrzymuje dostęp do kompletnego pliku i techniczną możliwość rozpowszechniania go poza kanałem dozwolonym przez właściciela, co potencjalnie może prowadzić do naruszenia praw autorskich. Rozwiązaniem problemu stosowanym w tradycyjnym podejściu jest utworzenie wielu wersji tego samego obrazu o różnych parametrach technicznych, które są przechowywane na serwerze. Użytkownik może wybrać, mniej lub bardziej świadomie, wersję spełniającą jego oczekiwania. Jest to podejście mało elastyczne z punktu widzenia użytkownika, producentów treści oraz w odniesieniu do wydajności serwera. W IIIF mechanizm jest inny: użytkownik otrzymuje tylko taki fragment obrazu, który w danym momencie jest mu potrzebny (a dokładniej – którego w imieniu użytkownika zażąda od serwera przeglądarka internetowa). Wydobycie owego fragmentu z wysokiej jakości pliku graficznego odbywa się bezpośrednio na serwerze. Do użytkownika wysyłana jest więc ograniczona, niezbędna dla określonego fragmentu ilość danych.

Niewątpliwie format TEI koncentruje się na tekście – od pojedynczych inskrypcji po kolekcje dzieł

Standard IIIF stosuje się w celu:

– zapewnienia badaczom zunifikowanego i szerokiego dostępu do zasobów cyfrowych obrazów na świecie,
– zdefiniowania zestawu wspólnych, aplikacyjnych interfejsów programistycznych, które będą wspomagać interoperacyjność między repozytoriami obrazów,
– rozwijania, udoskonalania i dokumentowania takich współdzielonych technologii, jak serwery obrazów i klienckie aplikacje webowe, które pozwalają użytkownikom oglądać, porównywać, przekształcać i anotować cyfrowe obrazy wysokiej jakości[31].

Aby zrealizować te założenia, IIIF zbudowano z czterech głównych interfejsów programistycznych (Application Programming Interface – API):

– image API,
– presentation API,
– search API,
– authentication API.

Image API określa, co można zrobić z obrazem: jaki rozmiar obrazu pobrać, który fragment, w jakim formacie, czy obraz lub jego fragment przekształcić (np. obrócić, zrobić lustrzane odbicie, zmienić przestrzeń kolorystyczną). Co ważne, żądania do serwera przesyłane są z wykorzystaniem standardowych protokołów HTTP/HTTPS, a polecenia są konstruowane tak, że z punktu widzenia klienta wyglądają jak zwykłe adresy URL (po stronie serwera wykorzystywana jest technologia RESTful API).

Dzięki prezentacyjnemu API możliwe jest z kolei np. tworzenie standardowych kolekcji jako sekwencji obrazów – podobnie jak stron w książce czy obrazów na wystawie w muzeum – i to niezależnie od miejsca ich udostępnienia (miejsca w rozumieniu kolekcji IIIF czy instytucji udostępniającej)[32] pod warunkiem wszakże, że udostępnianie pozostaje zgodne ze standardem IIIF. Możliwe jest tworzenie aplikacji porównujących obrazy, nakładających obraz na obraz (tworzenie tzw. blendów), tworzenie kolekcji kuratorskich pokazujących inne spojrzenie na tę samą zawartość[33], włączanie do obiektu cyfrowego warstw tekstowych (np. OCR tekstu), anotowanie i opisywanie obiektów graficznych bezpośrednio na stronie[34] czy tworzenie tzw. storiiies, czyli prezentacji prowadzących użytkownika po elementach obiektu z jednoczesnym objaśnianiem tych elementów[35]. Wszystkie instrukcje budujące warstwę prezentacyjną obiektu są tekstowymi instrukcjami w języku JavaScript dla przeglądarki internetowej, mają więc minimalny wpływ na wielkość pliku zawierającego cyfrową postać obiektu. Ponadto korzystają ze znanego i z powszechnie stosowanego w przeglądarkach języka programowania. W tej platformie programistycznej stosowane są dobrze już wdrożone, współczesne technologie webowe: wspomniany JavaScript, JSON, JSON-LD (JSON for Linking Data), RESTful API i inne, bardziej specjalizowane, ale otwarte, jak np. Open Annotation Data Model. Po stronie klienta nie wymaga się więc żadnych wtyczek czy dodatkowego oprogramowania – wystarczy w miarę aktualna przeglądarka stron internetowych.

Standardy TEI i IIIF nie są ze sobą sprzeczne ani konkurencyjne. Formaty te mają inne cele i w praktyce w edycjach cyfrowych bywają stosowane jednocześnie: TEI jako opis, naukowa interpretacja, semantyzacja dokumentu tekstowego, a IIIF jako format dla obrazu stanowiącego faksymile opisywanego dokumentu. Oba formaty w tych rolach znakomicie się sprawdzają. Trzeba przyznać, że w IIIF drzemie dużo większy niż w TEI potencjał popularyzatorski. Obraz, choćby skan dokumentu, wzbogacony dodatkową narracją objaśniającą, będzie dla niespecjalistów na pewno atrakcyjnym obiektem, a zakończenie prac nad IIIF dla audiowizualnych danych na pewno tę atrakcyjność jeszcze podniesie. Z drugiej strony trudno sobie wyobrazić zamknięcie w ramach obiektu graficznego zaawansowanego aparatu krytycznego i zaprezentowanie nierzadko skomplikowanych powiązań między tekstami.

*

W 2018 roku na Wydziale Filologii Polskiej i Klasycznej UAM rozpoczęto budowę Poznańskiej Platformy Humanistyki Cyfrowej, której cele należało zdefiniować, ponieważ dla polonistów, slawistów, klasyków, filmoznawców i teatrologów jawiła się przede wszystkim jako alternatywa dla tradycyjnych publikacji (analogowych). Humanistyka cyfrowa bywa bowiem w środowisku uniwersyteckim rozumiana różnie, ale dominują dwa nurty.

Pierwszy nurt dotyczy badania szeroko rozumianych tekstów kultury, które wykorzystują możliwości cyfrowych mediów (multimedialność, natychmiastowość, zaangażowanie odbiorcy we współtworzenie). Zazwyczaj zwraca się uwagę, że w tym nurcie lokują się badania Web 2.0, a właściwie jego kontentu, współtworzonego przez użytkowników świadomie (np. na portalach społecznościowych) lub mimowolnie (np. zachowań zarejestrowanych przez programy śledzące). W obu wypadkach ilość cyfrowych danych przyrasta lawinowo (*big data*) i wymaga specjalistycznej analizy, którą postrzega się jako służebną wobec rozmaicie ukierunkowanego marketingu. Namysł nad tradycyjnymi publikacjami wyników badań wykonanych przy pomocy cyfrowych narzędzi dla humanistów[36] oraz ogląd narzędzi (np. Google projekt Ngram Viewer) skłaniają do konkluzji, że cyfryzacja badań nad twórczością człowieka – zwłaszcza tą, której tworzywem jest język – również zależna jest od wytworzenia i udostępnienia dużej ilości danych, ale głównie tekstowych (*textual big data*).

Drugi nurt humanistyki cyfrowej wyznacza korzystanie z narzędzi cyfrowych: statystycznych, obliczeniowych, informatycznych, prezentacyjnych i publikacyjnych – w tym do wizualizacji wyników – oraz do analizy i eksploracji

danych, w końcu do badania tekstów wytworzonych cyfrowo bądź wtórnie zdygitalizowanych.

Chociaż dla funkcjonowania i rozwoju humanistyki cyfrowej w naszym kraju wytworzenie *Polish textual big data* jest po prostu niezbędne, to oczywiście nie sprosta temu zadaniu jeden wydział czy uczelnia. Dlatego ze względów praktycznych postrzegamy humanistykę cyfrową jako cyfrowe instrumentarium i metodologię do udostępniania oraz analizy tekstów w postaci cyfrowej. Będziemy więc próbowali zbudować platformę ze standardowym zestawem narzędzi do pracy z tekstami cyfrowymi (choć nadal ustalamy, jaki powinien to być zestaw), a namysł nad metodologią pracy badawczej ma stanowić istotny element przedsięwzięcia. Powinno zostać zaimplementowane wyszukiwanie semantyczne treści opublikowanych materiałów, a nie wyłącznie tekstowe („Rzecki" musi się odróżniać od „Dworzeckiego" czy „Międzyrzeckiego", również jako postać z *Lalki* Prusa). Konieczne więc będzie oznaczanie tekstu w sposób semantyczny. Zostanie do tego wykorzystany TEI – omówiony już standard elektronicznej reprezentacji tekstu zawierający informację o jego treści. Wykorzystamy też IIIF do serwowania obrazów (głównie faksymile). Platforma powinna zatem umożliwić:

- zapisanie zeskanowanych dokumentów,
- wprowadzenie ustandaryzowanych metainformacji o dokumencie źródłowym i obiekcie cyfrowym,
- przetworzenie obrazów na potrzeby serwowania z użyciem IIIF oraz pod kątem dalszej obróbki,
- OCR (w trybie wsadowym),
- wstępne przetwarzanie tekstu – automatyczną i ręczną korektę,
- zautomatyzowane przetwarzanie tekstu i podstawową anotację, w tym:
- modernizację pisowni i normalizację diachroniczną (w tej formie teksty jako alternatywy dla oryginału powinny trafić do bazy dla wyszukiwarki),
- przetwarzanie języka naturalnego (NLP), rozpoznawanie jednostek nazwanych (NER), np. osób, miejsc wraz z odwołaniem do baz autorytatywnych,
- rozpoznanie i oznaczenie podstawowych struktur dokumentu (nagłówków, artykułów, akapitów),
- weryfikację przetwarzania automatycznego i anotację ręczną.

Ponadto zostaną opracowane wspomagane automatycznie metody konwersji, aby można było umieścić na platformie tekst anotowany w innym formacie niż TEI. Dokumenty będą dostępne w ramach kolekcji, ale dostęp do tekstu, wypracowanych w ramach projektów informacji autorytatywnych oraz metainformacji, możliwy będzie za pomocą odpowiedniego API. Zdygitalizowanie tekstu, umieszczenie na platformie i anotowanie zgodnie ze standardem powinno również umożliwić ponowne użycia tego samego tekstu źródłowego w innych badaniach. W zależności bowiem od dziedziny czy zakresu badań tekst źródłowy będzie nabierał innych, nowych znaczeń. Zapewniając długotrwałe udostępnianie i archiwizowanie danych, platforma będzie wspierać umieszczanie opisanych metainformacjami zasobów źródłowych w repozytoriach otwartej nauki. Na potrzeby poszczególnych projektów uruchamianych na platformie będą też opracowywane narzędzia i aplikacje do pracy z kolekcjami – statystyczne, analityczne, do przetwarzania tekstu czy obrazu, wizualizacji danych oraz narzędzia wspierające Crowdsourcing. Oczywiście, zgodnie z zasadą *re-use*, takie wytworzone instrumentarium będzie dostępne dla innych projektów.

Nie wiadomo, w jakim zakresie uda się zrealizować przedstawione tu zamierzenia, ale z pewnością praca przy tworzeniu, opracowywaniu i publikowaniu online cyfrowych tekstów, zwłaszcza literackich, to nowa jakość edytorstwa.

Key Words: modern scholarly, digital humanities, digital tools, digital material, philological textual criticism, digital editions

Abstract: In the modern scholarly community there are two predominant ways of understanding of what the digital humanities means: 1) research into cultural texts in a broad sense when exploiting the potential of digital media, 2) using digital tools for analysis and data mining, including texts which are digitally born or were secondarily digitized. In both ways, texts in a digital form, if supposed to be useful on the digital humanities field as a research material, have to fulfill qualitative and quantitative conditions. Accessible (preferably on-line) collections should be made of characters, not (only) of images and should be formed with help of uniform set of tags and of course should be correct from the philological point of view. In turn, maximizing of those collections will allow for not only diversity of scholarly inquiries, but also reliability of research, and statistically significant results.

Therefore, literary part of the digital material is prepared with two criteria in mind: traditional, philological textual criticism and semantic elaboration of strings of characters. Standardization of this semantic elaboration is still work-in-progress and is a challenge for humanists and computer scientists. Such digital editions are becoming a new paradigm regarding to model critical editions of literary works.

[1] Tłumaczenie własne. Tekst oryginalny: „Textual scholarship in its digital fashion belongs to the broader field of Digital Humanities, itself a field built on interdisciplinarity, where many skills and theories of the realms of computer technology and those of scholarship intersect, and thus where many new interfaces and interactions arise between those skills and the fields they are tied to. This is where Digital Humanities acquires its innovative power, or at least the promise of that power"; J. van Zundert, *Barely Beyond the Book?*, w: *Digital Scholarly Editing. Theories and Practices*, eds. M. J. Driscoll, E. Pierazzo, Cambridge 2016, s. 83–84.

[2] Tłumaczenie własne. Tekst oryginalny: „Digital scholarly editions are considered to be one of the crown jewels of Digital Humanities [...]"; E. Pierazzo, *What Future for Digital Scholarly Editions? From Haute Couture to Prêt-à-Porter*, „International Journal of Digital Humanities" 2019, 1.2, s. 209, DOI 10.1007/s42803-019-00019-3 (dostęp: 14.12.2019).

[3] Na język polski ów metaforyczny termin należałoby tłumaczyć raczej jako „wielki parasol". *Big tent* na określenie humanistyki cyfrowej pojawił się w temacie przewodnim dorocznej konferencji ADHO („The Alliance of Digital Humanities Organizations") w 2011 roku i, w nawiązaniu do konferencji, został użyty także w antologii pod redakcją Matthew K. Golda *Debates in the Digital Humanities* z 2012 roku (edycja cyfrowa, dostęp: 2.11.2019, DOI 10.5749/9781452963754). W następnej edycji *Debates in the Digital Humanities* z 2016 roku redaktorzy tomu, Matthew K. Gold i Lauren F. Klein, określają humanistykę cyfrową jako *expanded field*, podając w wątpliwość, wraz z innymi autorami i krytykami, czy zakres otwartości i równoprawności konotowany przez poprzednio używaną metaforę jest możliwy do osiągnięcia (edycja cyfrowa, dostęp: 2.11.2019, DOI 10.5749/9781452963761). W najnowszej edycji *Debates* z 2019 roku podtrzymują metaforę rozszerzonego pola (edycja cyfrowa, dostęp: 2.11.2019, DOI 10.5749/9781452963785).

[4] A. Radomski, R. Bomba, *Wstęp. Zwrot cyfrowy w humanistyce*, w: *Zwrot cyfrowy w humanistyce. Internet – nowe media – kultura 2.0*, pod red. A. Radomskiego i R. Bomby, Lublin 2013, s. 8.

[5] *Zwrot cyfrowy w humanistyce*.

[6] *A Companion to Digital Humanities*, eds. S. Schreibman, R. Siemens, J. Unsworth, Oxford 2004 (edycja cyfrowa, dostęp: 2.11.2019).

[7] Tłumaczenie własne. Tekst oryginalny: „This collection marks a turning point in the field of digital humanities: for the first time, a wide range of theorists and practitioners, those who have been active in the field for decades, and those recently involved, disciplinary experts, computer scientists, and library and information studies specialists, have been brought together to consider digital humanities as a discipline in its own right [...]"; *A Companion to Digital Humanities*; S. Schreibman, R. Siemens, J. Unsworth, *The Digital Humanities and Humanities Computing. An Introduction*, w: ibidem.

[8] A. Nacher, *Poza cyfrowość w zwrocie cyfrowym – od humanistyki cyfrowej do spekulatywnej komputacji*, w: *Zwrot cyfrowy w humanistyce*, s. 89. Za ojca terminu uważa się Johna Unswortha, jednego z redaktorów naukowych antologii *A Companion to Digital Humanities*.

[9] Tłumaczenie własne. Tekst oryginalny: „It is evident that something is radically changing in the scholarly editing world: the way we work, the tools we use to do such work and the research questions to which we try to give answers – all of these have changed, in some case beyond recognition, with respect to the older print-based workflow"; M. J. Driscoll, E. Pierazzo, *Introduction. Old Wine in New Bottles?*, w: *Digital Scholarly Editing. Theories and Practices*, s. 3.

[10] Tłumaczenie własne. Tekst oryginalny: „Are we simply putting 'old wine in new bottles', or are we doing something which has never been done – indeed, never been doable – before?"; ibidem.

[11] R. Bomba, *Narzędzia cyfrowe jako wyznacznik nowego paradygmatu badań humanistycznych*, w: *Zwrot cyfrowy w humanistyce*, s. 57.

[12] M. Werla, M. Maryl, *Humanistyczne projekty cyfrowe w Polsce*, Poznań–Warszawa 2014, s. 6.

[13] S. Hockey, *The History of Humanities Computing*, w: *A Companion to Digital Humanities*; S. E. Jones, *The Emergence of The Digital Humanities* (wyd. cyfrowe, dostęp: 2.11.2019, DOI 10.4324/9780203093085); R. Busa, *The Annals of Humanities Computing. The Index Thomisticus*, „Computers and the Humanities" 1980, Vol. 14, No. 2, s. 83–90; G. Bolognesi et al., *The Work of Roberto Busa SJ. Open Spaces Between Computation and Hermeneutics*, „Anuario Filosófico" 2006, Vol. 39, No. 2, s. 465–476.

[14] R. Busa, *The Annals of Humanities Computing*.

[15] Tłumaczenie własne. Tekst oryginalny: „Only a computer census of the syntactic correlations can document what concepts the author wanted to express with that word"; R. A. Busa, *Foreword. Perspectives on the Digital Humanities*, w: *A Companion to Digital Humanities*.

[16] Informacja w dokumencie *Roberto Busa* na stronie wydawcy Frommann-Holzboog Verlag e.K. (dostęp: 13.11.2019).

[17] *Corpus Thomisticum. Index Thomisticus*, by R. Busa SJ and associates, WEB edition by E. Bernot and E. Alarcón, la versión española estará disponible en el futuro, http://www.corpusthomisticum.org/it/ (dostęp: 13.11.2019).

[18] M. Steffen-Batogowa, *Automatyzacja transkrypcji fonematycznej tekstów polskich*, Warszawa 1973.

[19] Należy mieć na uwadze, że lingwistyka komputerowa nie zawsze jest utożsamiana z humanistyką cyfrową. Podział ten będzie się uwidaczniał dopóty, dopóki humaniści cyfrowi nie uświadomią sobie korzyści, jakie dają im prace lingwistów i możliwości zautomatyzowanego przetwarzania tekstu. Próby pogodzenia obu środowisk podejmowano w przeszłości: „Językoznawstwo komputerowe zawsze rozwijało się niezależnie od informatyki humanistycznej i mimo wysiłków Dona Walkera w TEI Steering Committee nadal była ona odrębną dyscypliną. Walker i Antonio Zampolli z Instytutu Lingwistyki Komputerowej w Pizie bardzo starali się połączyć nauki humanistyczne z lingwistyką komputerową, ale z niewielkim skutkiem" (tłumaczenie własne, tekst oryginalny: „Computational linguistics had always developed independently of humanities computing and, despite the efforts of Don Walker on the TEI Steering Committee, continued to be a separate discipline. Walker and Antonio Zampolli of the Institute for Computational Linguistics in Pisa worked hard to bring the two communities of humanities computing and computational linguistics together but with perhaps only limited success"; S. Hockey, op. cit.).

[20] Ł. Gołębiewski, *Gdzie jest czytelnik?*, Warszawa 2012, s. 19.

[21] Przykładem jest opensource'owe oprogramowanie Tessearct OCR, w którego czwartej wersji wprowadzono silnik OCR oparty na sieci neuronowej o architekturze LSTM; *Tesseract OCR. Brief History*, w: *Readme.md* (dokument cyfrowy na stronach repozytorium projektu Tesseract w serwisie Github, dostęp: 13.11.2019).

[22] Fotokorpus jest dostępny online (dostęp: 13.11.2019), a informacja o sposobach radzenia sobie z błędami OCR pochodzi z rozmów własnych autorów z twórcami Fotokorpusu.

[23] Ciekawym sposobem pozwalającym zrozumieć ten problem jest eksperyment myślowy „Chiński pokój" sformułowany przez Johna Searla; B. Brożek, *Okropny sen filozofa*, w: B. Brożek, M. Heller, J. Stelmach, *Spór o rozumienie*, Kraków 2019, s. 61–62.

[24] *TEI: P5 Guidelines* (dokument cyfrowy na stronach internetowych konsorcjum Text Encoding Initiative, dostęp: 2.11.2019).

[25] *About These Guidelines*, w: *TEI: P5 Guidelines*.

[26] Przykłady zaczerpnięte z Korpusu Języka Polskiego PWN (dostęp: 2.11.2019).

[27] W odniesieniu do naszych przykładów zautomatyzowane przetwarzanie pozwoli wyszukiwać daty, miejsca ze skrótami lub z zaimkami na podstawie pełnej wartości itd.

[28] Strona WWW projektu: TEI Publisher. The Instant Publishing Toolbox, https://teipublisher.com/. TEI Publisher jest otwartym oprogramowaniem, udostępnianym na licencji GPL v.3.

[29] Akronim od *galleries*, *libraries*, *archives*, *museums.*

[30] Skrót od International Image Interoperability Framework; wymowa „tripl-aj-ef".

[31] About IIIF (dokument cyfrowy na stronach IIIF Consortium, dostęp: 2.11.2019).

[32] To właśnie dzięki IIIF możliwe jest funkcjonowanie agregatorów cyfrowych obiektów dziedzictwa, takich jak CBN Polona, Bibliotheca Europeana czy Biblissima.

[33] Do tworzenia takich kuratorskich kolekcji może posłużyć IIIF Curation Platform (http://codh.rois.ac.jp/icp/, dostęp: 2.11.2019).

[34] Przy odpowiedniej aplikacji takie pisanie może dawać wrażenie notowania na obiekcie. Oczywiście notatka nie ingeruje w oryginalny obraz i pozostaje w osobnej warstwie programowej – kanwie.

[35] Experiments in digital storytelling by Cogapp (http://storiies.cogapp.com/, dostęp: 2.11.2019).

[36] Zob. pracę F. Morettiego *Graphs, Maps, Trees – Abstract Models for a Literary History* z 2005 roku (wyd. polskie: *Wykresy, mapy, drzewa. Abstrakcyjne modele na potrzeby historii literatury*, tłum. T. Bilczewski i A. Kowalcze-Pawlik, Kraków 2016).