PIOTR TAFIŁOWSKI
Uniwersytet Marii Curie–Skłodowskiej
Instytut Bibliotekoznawstwa i Informacji Naukowej

Projekt „Spuścizny": elektroniczne publikacje ineditów

Na temat potrzeby elektronicznej publikacji źródeł historycznych pisze się w polskiej literaturze już od połowy lat dziewięćdziesiątych XX wieku[1]. Także w latach następnych pisano o elektronicznej edycji źródeł czy też zastosowaniu techniki komputerowej w pracy historyka[2], niestety z niezadowalającym do tej pory rezultatem praktycznym. Edycje elektroniczne nadal zliczyć można na palcach. Nadal sztandarowym przykładem tego typu projektów są „Teki Dworzaczka", wydane na CD (1995, 1997), a także dostępne w Internecie na stronach Biblioteki Kórnickiej PAN (http://www.bkpan.poznan.pl/biblioteka/teki.html)[3]. Był to jednak mały pierwszy krok, wykonany już kilkanaście lat temu, dziś można więc oczekiwać więcej.

Musimy pamiętać, że ogromna masa źródeł średniowiecznych i staropolskich zaginęła bądź została zniszczona, wiele z nich spłonęło w 1944 r. w Warszawie[4] i dziś treść zaledwie części z nich znana jest wyłącznie z wypisów dokonywanych przez badaczy. Te wypisy, pozostające często w rękopisach spoczywających w zbiorach prywatnych, należy chronić, informować o nich i je udostępniać. Dla przykładu ks. prof. Henryk R y b u s miał wiele takich wypisów z akt arcybiskupów gnieźnieńskich, z czego po Powstaniu Warszawskim została mu tylko mała część

---

[1] J.S. M a t u s z e w s k i, *Źródło–komputer–historyk*, KH, t. CII, 1995, z. 2, s. 79–84; J. W i s ł o c k i, *Centrum Elektronicznych Tekstów Historycznych* [w:] *Historia i komputery*, t. II, pod red. B. R y s z e w s k i e g o, Toruń 1997, s. 9–22.

[2] R. P r i n k e, *Fontes ex machina. Komputerowa analiza źródeł historycznych*, Poznań 2000; idem, *Elektroniczna edycja źródeł do dziejów Polski*, RH, t. LXXI, 2005, s. 235–238.

[3] A. B i e n i a s z e w s k i, R. P r i n k e, J. W i s ł o c k i, *Spuścizna Włodzimierza Dworzaczka*, „Pamiętnik Biblioteki Kórnickiej", t. XXIV, 1996, s. 169–182; A. B i e n i a s z e w s k i, R. P r i n k e, *Doświadczenia z pracy nad elektroniczną edycją „Tek Dworzaczka"* [w:] *Historia i Komputery*, t. II, s. 23–36.

[4] Cf. *Straty bibliotek i archiwów warszawskich w zakresie rękopiśmiennych źródeł historycznych*, t. I–III, pod red. P. B a ń k o w s k i e g o, Warszawa 1955–1957.

(jedna teczka)[5], a i ona gdzieś zaginęła po jego śmierci, jak powiedział mi kiedyś uczeń profesora, ksiądz Stanisław Grad.

Moja propozycja dotyczy wszelkiego rodzaju archiwaliów, źródeł i materiałów historycznych gromadzonych w prywatnych kolekcjach. Chodzi o umieszczenie w systemie informatycznym ineditów po zmarłych historykach. Praca opublikowana drukiem nawet kilkadziesiąt czy sto lat temu nie jest największym problemem w kwerendach, tekst wydrukowany w wielu egzemplarzach da się w końcu odnaleźć. Natomiast w rękopisach pozostają cały czas często bezcenne dla nauki gotowe prace, notatki, kartoteki, fiszki itp. Najczęściej możliwości skorzystania z tych materiałów są niezwykle ograniczone. Spoczywają one w różnych instytutach (głównie w archiwach PAN, a także uniwersytetów oraz w bibliotekach naukowych) czy też w zbiorach osób prywatnych (spadkobierców). Ale i to jeszcze nie jest największym zagrożeniem. Znaczna część tych rękopisów zostaje zagubiona lub całkowicie zniszczona, nie tylko wskutek działań wojennych, lecz także złych warunków przechowywania, ludzkich błędów, nieświadomości czy wręcz bezmyślności.

Istnieje zatem potrzeba elektronicznej edycji tekstów pisanych przez historyków, pozostających dotąd w rękopisach. Oprócz niepublikowanych artykułów wchodzą tu przecież w grę także choćby wspomniane wypisy, jak te bezcenne a utracone, sporządzane przez prof. Rybusa, czy regesty Ireny i Stanisława K u r a s i ó w do kolejnego tomu „Bullariów" (którego już nie będzie) czy „Kodeksu dyplomatycznego mazowieckiego", a także takie materiały jak fotografie, pocztówki, dawne czasopisma, druki ulotne, konspiracyjne, zbiory regionalne, korespondencja oraz pamiętniki. Są to bardzo cenne materiały, a ich utrata stanowi wielką szkodę dla naszej nauki.

Już choćby z tego wyliczenia widać, że spuścizny są całkowicie odmiennym typem „obiektów" niż źródła historyczne. By odwołać się tylko do wspomnianych już powyżej publikacji kórnickich, „Teki Dworaczka" są czym innym niż „Diariusze sejmowe I Rzeczypospolitej".

W ciągu ostatnich kilkudziesięciu lat diametralnie zmienił się sposób uprawiania nauki, a jeden z aspektów tych przemian interesuje mnie szczególnie. Otóż obecnie pracownicy naukowi piszą, używając komputerowych edytorów tekstu. Spuścizny po nich będą już w głównej mierze cyfrowe (*born digital*). Paradoksalnie powoduje to w pewnym sensie większe problemy[6]. Czy spadkobiercy powinni przekazywać do archiwów publicznych twarde dyski komputerów? Albo pliki skopiowane na jakiś nośnik (CD, *pen drive*)[7]? I co następnie te instytucje miały-

---

[5] H. R y b u s, *Kilka uwag o aktach arcybiskupów gnieźnieńskich*, ABMK, t. II, 1961, z. 1–2, s. 304–307; idem, *Regestry akt arcybiskupów gnieźnieńskich*, ABMK, t. III, 1961, z. 1–2.

[6] Cf. na temat problemów z danymi cyfrowymi w fizyce: M. R ó ż y c z k a, *Siadaj i badaj*, „Polityka" nr 16 (2803) z 16 kwietnia 2011, s. 66–67.

[7] Vide też P. S t a s i a k, *Sieć do wieczności*, „Polityka" nr 17 (2804) z 23 kwietnia 2011, s. 117–119.

by z nimi robić? W jaki sposób zabezpieczać je przed niebezpieczeństwem utraty w wyniku fizycznej degradacji nośnika lub niemożliwości odczytu wskutek postępu technologicznego?

Wydaje się, że sensowniejszym rozwiązaniem byłoby skonstruowanie systemu, do którego można by „wgrać” te pliki w trybie *online*, być może w ramach projektów Ośrodka Przetwarzania Informacji (opi.org.pl; nauka–polska.pl). W bazach OPI zgromadzono już tyle informacji na temat nauki polskiej, że dodanie kolejnych funkcjonalności, w tym udostępnianie ineditów, narzuca się samo przez się. Tak czy inaczej, prędzej czy później staniemy wobec konieczności rozwiązania tego problemu i skonstruowania odpowiedniego systemu elektronicznego.

Historycy posługujący się komputerami przygotowują własne bazy danych badając specyficzne problemy. Z tych samych materiałów źródłowych mogą być tworzone różne zestawy informacji, ponieważ poszczególni badacze mogą koncentrować się na różnych problemach i inne typy danych będą dla nich istotne. Np. z ksiąg metrykalnych dla jednego badacza istotny będzie wiek zawierania pierwszego małżeństwa, dla innego stosowanie tytulatury, dla jeszcze innego sposób zapisywania daty w tychże aktach itd. Będziemy więc potrzebowali systemu pozwalającego na łączenie tych różnych informacji z różnych baz danych, nawet jeśli tworzone były one przy pomocy różnego oprogramowania. Pozostałe funkcjonalności postulowanego systemu podsumowane zostaną na zakończenie tego tekstu.

Nie wiem jeszcze oczywiście dokładnie, jak taki system powinien wyglądać. Niniejsza praca jest jedynie szkicem, zarysem problematyki i traktować ją należy jako głos w dyskusji. Nie jestem w stanie przedstawić gotowych rozwiązań, a jedynie pewne propozycje. Postulowany system należy dopiero zaprojektować, ale najpierw trzeba podjąć prace nad gromadzeniem spuścizn i stworzeniem modelu opisu tych materiałów. Trzeba zastanowić się, co będziemy chcieli udostępniać: rękopisy, notatki, regesty, wyciągi, czy także jakieś inne kolekcje, np. fotografie, rysunki, widokówki, numizmaty, medale, znaczki pocztowe itd. Dopiero kiedy udzielimy sobie odpowiedzi na tego typu podstawowe pytania, będziemy mogli przystąpić do prac nad systemem informatycznym.

Oczywiście wszelkiego typu materiały rękopiśmienne, od których rozpocząłem niniejsze rozważania, należałoby wpierw przepisać i wraz z zeskanowanymi (sfotografowanymi) źródłami ikonograficznymi opisać za pomocą języka XML, np. w EAD na wyższym poziomie (jako przykład można tu podać archiwalne inwentarze elektroniczne udostępniane w Internecie przez Archiwum Główne Akt Dawnych w Warszawie: http://agad.archiwa.gov.pl/metodyka/inwentarze.html)[8].

---

[8] Vide też R.T. P r i n k e, *Terra rubrica — terra electronica: Najkrótsza historia metatekstu* [w:] *Nuntius Vetustatis: prace ofiarowane Profesorowi Jerzemu Wisłockiemu*, pod red. A. B i e n i a s z e w -

Należałoby także odpowiedzieć sobie na pytanie, kto powinien się zająć realizacją takiego projektu. Być może Ośrodek Przetwarzania Informacji w ramach wspomnianych już przedsięwzięć. Zapewne wdrożenie go wymagać też będzie współpracy ze strony Ministerstwa Nauki i Szkolnictwa Wyższego oraz Polskiej Akademii Nauk.

Natomiast co do konkretnych propozycji wdrożeniowych, można wskazać możliwość zaprojektowania odpowiedniej mapy wiedzy[9] lub systemu ekspertowego z bazą wiedzy, którego najpotężniejszym przykładem jest obecnie system CYC. Jemu to, jako najlepszej moim zdaniem propozycji, poświęcona zostanie dalsza część tekstu.

Tzw. systemy ekspertowe powstały w toku badań nad rozwojem sztucznej inteligencji jako dziedziny nauki. Są to programy komputerowe, które rozwiązują złożone problemy wymagające dużego wysiłku intelektualnego, robiące to równie dobrze jak człowiek będący ekspertem w danej dziedzinie. Istnieją funkcjonujące systemy ekspertowe dziedzinowe, np. w zakresie medycyny, przemysłu czy techniki i obecnie nie ma już powodów czy barier, które nie pozwalałyby na skonstruowanie tego typu systemu dla nauk historycznych.

Podstawowym ograniczeniem systemów ekspertowych była stosunkowo niewielka liczba faktów, w jakie mogły być one wyposażone — zazwyczaj zaledwie od kilkuset do kilku tysięcy reguł. Pozwalało to co prawda komputerowi na „myślenie", ale w obrębie tylko jednej dziedziny. Dlatego też w 1984 r. Douglas L e n a t podjął się stworzenia projektu CYC — systemu, który nie ograniczałby się tylko do wybranej dziedziny, lecz także posiadał informacje o całym otaczającym nas świecie, która umożliwiłaby komputerom prowadzenie wnioskowania na wzór ludzkiego. Bazę tę określono mianem „zdrowego rozsądku". Autor projektu już wcześniej ustalił, że ilość reguł, jakie powinien zawierać „zdrowy rozsądek", stanowi liczbę rzędu 100 mln, chociaż inni badacze twierdzą, że powinna ona wynosić aż 500 mln. Nazwa projektu pochodzi od angielskiego słowa *encyclopedia*, gdyż początkowo zakładano, że CYC będzie zawierał wiedzę i definicje o charakterze encyklopedycznym. Jednak obecnie definicje wprowadzone do systemu są o wiele bardziej szczegółowe niż hasła encyklopedyczne. Prace zaplanowano na dziesięć lat, ale trwają one nadal. Do roku 2010 udało się wprowadzić do systemu ponad 5 mln reguł.

---

s k i e g o i R.T. P r i n k e, [Dokument elektroniczny], Kórnik: Biblioteka Kórnicka PAN, 1998; tryb dostępu: http://www.bkpan.poznan.pl/biblioteka/JW70/terra.htm. Cf. także połączenie faksymile cyfrowego i transkrypcji kodowanej jako XML/TEI; w ten sposób jest obecnie udostępniana rękopiśmienna spuścizna Izaaka Newtona: www.newtonproject.sussex.ac.uk.

[9] Y. L o z o w i c k, *Mapy wiedzy i biblioteki* [w:] *Archiwa elektroniczne* [Dokument elektroniczny], Warszawa: AGAD, 2009; tryb dostępu: http://www.agad.archiwa.gov.pl/electro/Mapy_wiedzy_Yaacov_Lozowick.pdf. Toż samo w: „Przegląd Informacyjno–Dokumentacyjny", t. XXXIX, 2009, nr 3 (306), s. 86–91.

Warto wymienić przykładowe zastosowania systemu CYC:

a) dokonywanie maszynowych tłumaczeń;
b) analiza, rozumienie i tłumaczenie tekstów naturalnych;
c) semantyczne integrowanie baz danych;
d) tworzenie tezaurusów (z dziedziny techniki i medycyny);
e) wyszukiwanie informacji;
f) automatyczne adnotacje;
g) sprawdzanie spójności wiedzy;
h) integrowanie heterogenicznych baz danych;
i) prowadzenie symulacji, które wykorzystują ograniczenia zawarte w ontologii;
j) dzielenie się wiedzą przez niezależnie pracujące grupy, sprzedawanie towarów i usług przez Internet;
k) budowanie modeli użytkownika danego systemu oraz wykorzystywania go do badań;
l) modelowanie użytkowników programów i urządzeń technicznych;
ł) używanie w inteligentnych interfejsach programowych, które reagują na intencje użytkownika;
o) integrowanie informacji;
p) symulacje inteligentnych zachowań postaci w grach komputerowych;
r) inteligentna symulacja rzeczywistości wirtualnej
s) zastosowanie w wojskowości (dobrym przykładem była próba stworzenia na podstawie systemu CYC programu, który miał za zadanie doradzać prezydentowi USA w sytuacjach kryzysowych i sprawach militarnych — oficjalnie jednak pomysł nie został zrealizowany).

System CYC, podobnie jak inne systemy ekspertowe, składa się z kilku podstawowych elementów. W tym przypadku są to: baza wiedzy (VKLB — *Very Large Knowledge Base*), mechanizm wnioskujący, język reprezentacyjny wiedzy (CYCL), podsystem przetwarzania języka naturalnego, szyna integracji semantycznej i zestaw narzędzi dla rozwoju systemu. Baza CYC składa się z kilku tysięcy mikroteorii, podzielonych ze względu na dziedzinę wiedzy, poziom uszczegółowienia informacji itp.

W systemie CYC osoby, które nie potrafią posługiwać się językiem CYCL, mogą również używać języka naturalnego (CYC–NL). System jest w stanie przetłumaczyć polecenia z języka angielskiego na język CYCL, co bardzo ułatwia pracę. Dzięki CYC–NL możliwa jest analiza zdań złożonych i wieloznacznych. Wymaga to oczywiście od systemu posiadania określonej wiedzy, ale na obecnym etapie nie stanowi to już problemu.

Jeśli chodzi o sposób opisu wiedzy, to podstawę w systemie CYC stanowią tzw. ramy. Każda z nich ma zdefiniowany rachunek predykatów oraz możliwość

rozbudowania go domyślnymi zmiennymi. Ramy te posiadają mechanizmy dziedziczenia. Polega to na przypisaniu każdej regule wielkiej ilości „szufladek" (liczba ta może być teoretycznie nieskończona). Np. zdanie „Wszyscy studenci Informacji Naukowej i Bibliotekoznawstwa lubią książki" sprawi, że każda ramka „informacja naukowa i bibliotekoznawstwo" w szufladce „student" odziedziczy „książka" w szufladce „lubi".

Oprócz ram istnieje także możliwość opisu poprzez stosowanie ograniczeń, czyli *constraint language*. W przypadku zastosowania ograniczenia zdanie „Bogdan lubi ludzi, którzy mają złote karty kredytowe" nie przypisze wszystkim ramkom ludzi szufladki „Bogdan ich lubi", ale ogranicza szufladkę „lubi" w ramce Bogdana do osób, które mają złote karty kredytowe.

System posiada dwadzieścia różnych mechanizmów wnioskowania, zależnie od dziedziny. By sprawdzić, czy system dobrze rozumie daną dziedzinę, daje mu się, tak jak w przypadku człowieka, do przeanalizowania tekst, na temat którego następnie zadaje się mu pytania. Poprawne odpowiedzi oznaczają, że wiedza systemu jest wystarczająca. Czas, w którym przebiega rozumowanie, zależy od reguły, jakiej ono dotyczy. W systemie wyróżnia się reguły używane tylko wtedy, gdy się do nich odwołujemy (*if–needed rule*) i takie, które opisują wszystko co się da (*if–added rule*). Nad zachowaniem spójności i wydajności systemu w trakcie dodawania nowych danych czuwa podsystem Truth Maintenance System.

Mówiąc o procesie wnioskowania należy także zauważyć, że CYC odróżnia rzeczy indywidualne od ich zbiorów, własności zewnętrzne rzeczy od wewnętrznych, a także zdarzenia od procesów.

CYC nie posiada jednolitej struktury jeśli chodzi o zaimplementowaną wiedzę, lecz składa się z kilku podsystemów. Podsystemy te mogą współpracować ze sobą przy rozwiązywaniu problemów. Jeśli podsystem, który posiada wiedzę ogólną nie jest w stanie odpowiedzieć na pytanie z jakiejś dziedziny, zwraca się do podsystemu specjalistycznego. Współpracę tę określa się mianem *The Cycic Friends Network*. Oba te podsystemy, zwane agentami, mogą również korzystać z informacji zawartych w Internecie.

Na bazie systemu CYC wyrosło kilka projektów — OpenCYC, CYCSecure, CYC Answers i Research CYC. OpenCYC to udostępniona bezpłatnie w 2002 r. wersja CYC, którą można pobrać ze strony producenta (www.cyc.com). Jego zadaniem jest prezentowanie zalet systemu i rozpowszechnienie go. Przykładowo, wersja 1.0 OpenCYC posiadała około 300 tys. pojęć i około 3 mln faktów, które dotyczą tych pojęć. Składała się z modułu wnioskującego CYC Interference Engine, przeglądarki bazy CYC Knowledge Base Browser, a także z narzędzi pozwalających na posługiwanie się językiem naturalnym w trakcie pracy z systemem, dokumentacji systemów i kilku programów demonstracyjnych. CYCSecure to program symulujący ataki na sieci komputerowe. Składa się z trzech elementów: programu, który symuluje sieć komputerową, bazy wiedzy zawierającej informacje o bezpieczeństwie sieci i analizatora możliwości ataku na symulowaną sieć. Od roku 2006

program jest w sprzedaży. CYC Answer to program zarządzający wiedzą i odpowiadający na pytania. Research CYC to dostępna również od 2006 r. wersja przeznaczona dla celów badawczych. W momencie jej udostępnienia zawierała około 300 tys. koncepcji, około 3 mln reguł i ponad 26 tys. relacji.

Na koniec powróćmy do materii historycznej. W tradycyjnym systemie archiwalnym odpowiedzią na zapytanie użytkownika jest zazwyczaj zestaw dokumentów — linków do mniej lub bardziej relewantnych dokumentów w systemie cyfrowym, lub np. teczka dokumentów papierowych w systemie tradycyjnym. Kiedy poszukuję informacji związanych z moim nazwiskiem, otrzymuję zestaw dokumentów (papierowych lub elektronicznych), w których pojawia się ono w mniej lub bardziej interesującym mnie kontekście. Dopiero po przeczytaniu ich wszystkich mogę zadecydować, które z nich są dla mnie przydatne, które odpowiadają na moje pytanie. Korzystając z istniejących obecnie systemów nie dowiem się natomiast, że Piotr jest synem Jerzego, który z kolei ma dwóch braci — Krzysztofa i Waldemara, a wszyscy trzej są synami Władysława. Być może po przeczytaniu wszystkich dokumentów będę w stanie wydedukować sobie, jakie relacje zachodzą między Piotrem, Jerzym, Waldemarem, Krzysztofem i Władysławem, ale obecnie nie ma możliwości otrzymać taką wiedzę z systemu, ponieważ żaden system w tej chwili po prostu tego nie wie. Podobnie, jeśli szukam informacji na temat „praskiej wiosny” czy też „poznańskiego czerwca”, w obecnie funkcjonujących systemach mogę otrzymać kilka dokumentów dotyczących w jakiś sposób tych wydarzeń, nie ma natomiast sposobu, bym dowiedział się, czym one w swej istocie były w całym kontekście społeczno–historycznym[10].

Pierwszy krok do uzyskania interesującej mnie tu funkcjonalności zostały uczynione dzięki digitalizacji wspomnianych już „Tek Dworzaczka”. Wydawcy tego CD–ROM zastosowali oprogramowanie umożliwiające automatyczne generowanie tabel genealogicznych. Relacje między encjami czy osoby są tu obiektami logicznymi. Teraz należałoby pójść dalej i wydaje się, że przystąpienie do prac nad wdrożeniem systemu ekspertowego, obejmującego również spuścizny (wtórnie digitalizowane oraz *born digital*) oraz inedita, jest niezbędne.

W podsumowaniu można wyliczyć, jakie korzyści osiągniemy dzięki odpowiedniemu wdrożeniu informatycznemu:

1) Łatwy dostęp do wartościowych informacji, często niedostępnych w inny sposób.
2) Zminimalizowanie ryzyka bezpowrotnej utraty cennych materiałów historycznych, które zostaną zabezpieczone poprzez wykonanie ich kopii cyfrowych i będą w ten sposób udostępniane wszystkim zainteresowanym badaczom.

---

[10] P. T a f i ł o w s k i, *Archiwa 3.0 — sieć wiedzy* [w:] *Archives 2.0 : Warsztaty Warszawa 26 listopada 2008*, dokument elektroniczny, Lubań: Archnet: Naukowy Portal Archiwalny, 2009, tryb dostępu: http://adacta.archiwa.net/file/2009archives20/pdfy/Archiwa30_final.pdf.

3) Uzyskanie dodatkowych korzyści w postaci możliwości analizy komputerowej danych.
4) Możliwość dołączania do systemu kolejnych modułów, np. elektronicznych edycji źródłowych, wiedzy kontekstowej (komentarze do źródeł), poszerzających m.in. możliwości analizy komputerowej.
5) Możliwość integracji heterogenicznych danych, zapisywanych w różnych formatach, a także połączenia danych tekstowych z materiałami audiowizualnymi opisanymi szerokim zestawem metadanych.
6) Połączenie obiektów (nazw osobowych i geograficznych, obiektów architektonicznych, wydarzeń historycznych, nazw urzędów i godności, dat itd.) relacjami logicznymi.
7) Szkielet systemu, przygotowany dla nauk historycznych, byłby przydatny także dla dokumentów z innych dziedzin nauk humanistycznych i społecznych (jest to cecha charakterystyczna dla systemów ekspertowych z bazą wiedzy).