Biuletyn
Polonistyczny

08.08.2024

Dlaczego warto pracować z (meta)danymi? Workflow

W maju 2019 roku w Bibliotece Uniwersyteckiej w Warszawie odbyło się wydarzenie DARIAH Annual Event. Zadawano sobie wtedy pytania dotyczące rodzaju i ilości danych produkowanych i gromadzonych przez humanistów. Pytano o rodzaje danych, miejsca ich przechowywania oraz o to, kto jest ich właścicielem. Zastanawiano się nad specyfiką i złożonością danych humanistycznych, a także nad tym, co czyni je wyjątkowymi. Próbowano także definiować i konceptualizować termin „dane", sytuując je obok bardziej tradycyjnego nazewnictwa źródeł, z których korzystają i które tworzą nauki humanistyczne. Krótki zapis tego wydarzenia oraz kilka wypowiedzi o danych humanistycznych można obejrzeć pod tym linkiem.

Humaniści produkują dane

Dziś, po 5 latach od tego wydarzenia, ciągle warto podkreślać fakt, że humaniści pracują z danymi badawczymi. Proponuję pojemną formułę danych, w której mieszczą się wszelkiego rodzaju informacje i materiały zebrane, obserwowane, mierzone, opracowane, wygenerowane lub przetworzone podczas prowadzenia badań naukowych. Mogą one przyjmować różną postać i obejmować szeroki zakres formatów i typów. Mowa tu zarówno o danych ilościowych, będących wartościami liczbowymi, które mogą być mierzone i analizowane statystycznie, jak i o nienumerycznych danych jakościowych, które opisują właściwości, kategorie i cechy. Wśród danych zaobserwujemy te surowe, zebrane bezpośrednio z obserwacji; nieustrukturyzowane, będące najczęściej zestawem swobodnie wytworzonych tekstów; czy przetworzone – a więc poddane obróbce lub analizie polegającej na ujednoliceniu, harmonizacji albo wzbogaceniu.

Przykładami danych tworzonych i badanych w humanistyce mogą być:

- teksty literackie – wszelkie formy piśmiennictwa umożliwiające analizę obejmującą badanie stylu, tematyki, struktury lub kontekstu historycznego i kulturowego,
- słownictwo – różnego rodzaju rejestry, indeksy (także te rzeczowe) czy bibliografie załącznikowe, pozwalające na badanie używanej terminologii lub zależności między aktorami danego pola badawczego,
- dokumenty historyczne – archiwalia, pamiętniki lub korespondencja, pozwalające na odtwarzanie sieci zależności społecznych lub historycznych,

- wizualia – obrazy, fotografie lub rzeźby pozwalające badać techniki artystyczne i symbolikę (np. dzięki użyciu technologii *digital twin*),
- dźwięki i muzyka – nagrania lub partytury oraz badania obejmujące analizę np. kompozycji,
- materiały etnograficzne – różnego rodzaju notatki z obserwacji, transkrypcje i nagrania wywiadów, umożliwiające badania materialnego i niematerialnego dziedzictwa kulturalnego,
- dane wizualne – mapy i ilustracje, będące samodzielnym przedmiotem badań lub narzędziem wspomagającym komunikację i prezentację efektów,
- dane cyfrowe – bazy danych, teksty *digital born* i metadane, które mogą być analizowane z wykorzystaniem narzędzi humanistyki cyfrowej, aby odkrywać wzorce i zależności nieuchwytne dla tradycyjnych metod.

Powszechność metadanych

Spośród wszystkich rodzajów danych chcę się skupić na metadanych. Tradycyjnie definiowane są one jako dane o danych lub podstawowe informacje stanowiące opis całego zbioru. Dla przykładu: elementy opisu bibliograficznego, takie jak autor, tytuł i opis fizyczny mogą w środowisku cyfrowym funkcjonować jako egzemplifikacja metadanych. To, co wpływa na użyteczność i przydatność metadanych, można streścić jednym słowem – struktura. Metadane są zestandaryzowane i schematyczne, dzięki czemu nadają się do czytania zarówno przez ludzi, jak i przez maszyny, a to z kolei oznacza, że spełniają wymóg interoperacyjności, który jest jednym z czterech filarów zasad FAIR (o których więcej można przeczytać w moim wcześniejszym wpisie).

Metadane są także podstawą mechanizmów wyszukiwawczych w Internecie – czy to w wyszukiwarkach internetowych, czy w serwisach typu discovery, które wykorzystują poszczególne informacje zgromadzone w metadanych w taki sposób, aby efektywnie odnajdywać oraz filtrować zasoby (tak działa np. GoTriple). Do metadanych jesteśmy już przyzwyczajeni, bo towarzyszą nam na co dzień. Najwyższy czas, by z tego dobrobytu skorzystać w badaniach humanistycznych.

Powszechność metadanych to oczywiście nie tylko ogólna praktyka publikacji treści w sieci, ale przede wszystkim sposób działania instytucji, które są pierwszym wyborem humanistów przy przeprowadzaniu kwerend. Mowa oczywiście o katalogach bibliotek narodowych, które udostępniają metadane w domenie publicznej lub na otwartych licencjach, wykorzystując struktury informacji zapośredniczone w formatach bibliotecznych (np. deskryptory w określonych polach formatu MARC21). W praktyce oznacza to możliwość wpisania tytułu poszukiwanego tekstu do dedykowanego pola, aby zwiększyć skuteczność wyszukiwania. Nikt nie chce błądzić, wpisując w pole tytuł nazwisko autora. Podobny mechanizm możemy zaobserwować w bazach wzorcowych, np. VIAF lub Wikidata. Ta ostatnia jest największym zbiorem danych i zawiera ponad 110 milionów elementów.

Metadane są bogate w informacje. Na badaczy i badaczki czekają gotowe zasoby, które wprawdzie niekoniecznie muszą spełniać wszystkie oczekiwania, ale z pewnością pozwalają na to, aby nie tworzyć zbiorów od zera. Poprzez bogactwo możemy rozumieć zarówno wielość atrybutów, którymi opisane są elementy danych (przykład może stanowić plik graficzny przedstawiający podpis Olgi Tokarczuk lub nagranie audio z wymową jej imienia i nazwiska w ojczystym języku, dostępne w jej haśle na Wikidacie), a także powiązanie danych między zbiorami za sprawą struktur Linked Open Data (LOD) oraz stałych identyfikatorów (PIDs). Posiadając jeden identyfikator, będziemy mogli – żeby sparafrazować piosenkę otwierającą serial *Pokemon* – złapać je wszystkie.

Metadane wreszcie są łatwo dostępne i wcale nietrudno po nie sięgnąć. Duża część zbiorów udostępniana jest na otwartych licencjach, a o wykorzystaniu i ponownym użyciu informują nas powszechnie przyjęte standardy (np. FAIR data lub 5-star data). Dodatkową zachętą są masowe operacje, które można realizować poprzez otwarte protokoły, udostępniane przez wiele instytucji. Skoro znalazłem już wszystkie teksty na interesujący mnie temat w katalogu Biblioteki Narodowej, to, zamiast kopiować wszystkie tytuły po kolei, mogę skorzystać z interfejsu programistycznego (API) BN (początkujących użytkowników odsyłam do dokumentacji serwisu).

Potencjał metadanych

Jeśli metadane pochodzą z wiarygodnego źródła, możemy zaufać skuteczności procesów identyfikacyjnych. Niech za przykład posłuży nam tradycyjny zapis bibliograficzny, który znalazłem w książce Tomasza Mizerkiewicza pt. *Po tamtej stronie tekstów. Literatura polska a nowoczesna kultura obecności*: „Gombrowicz W., *Kosmos*, Kraków 2004”. Ograniczenia tradycyjnej bibliografii sprawiają, że wszystkie elementy składające się na opis książki są przedstawione za pomocą ciągów znakowych, a co za tym idzie, wszelka ich identyfikacja spoczywa na odbiorcy. Polski czytelnik wyposażony w kompetencje literaturoznawcze bez problemu zidentyfikuje imię autora kryjące się pod inicjałem oraz wskaże na mapie miejsce wydania. Trudności zaczynają się, gdy odbiorca nie ma narzędzi do przeprowadzenia poprawnej identyfikacji. A co, jeśli mowa tu o (wymyślonym przeze mnie na poczekaniu) Wiktorze Gombrowiczu, który zbudował swoją tożsamość debiutanta literackiego na życiu w cieniu wielkiego pisarza i postanawia się w tekście o tytule *Kosmos* zmierzyć z tymi uwarunkowaniami. A gdy polskojęzyczny zbiór poezji regionalnej zostanie po raz pierwszy wydany w Paryżu, to będzie to stolica Francji czy wieś w województwie kujawsko-pomorskim?

Konieczność samodzielnej identyfikacji wymusza skorzystanie z mechanizmów wyszukiwawczych, które sortują wyniki według bliżej nieokreślonej trafności. W rzeczywistości *big data* Wiktor Gombrowicz może nie pojawić się na pierwszej stronie wyników wyszukiwarki Google. Dlatego tak istotne jest zapośredniczenie identyfikacji w pracy instytucji, naukowców i naukowczyń oraz zespołów badawczych, którym możemy zaufać. Dodatkowym zyskiem jest szeroki wachlarz

tych wszystkich atrybutów, którym opatrzono już wcześniej interesujące nas obiekty. Przykład bibliografii cyfrowej dla *Kosmosu* Witolda Gombrowicza można znaleźć tutaj.

Możliwości, które metadane oferują środowisku naukowemu to oczywiście nie tylko identyfikacja. Metadane mogą posłużyć jako sformalizowany język umożliwiający skodyfikowanie pytań badawczych. Co to w praktyce oznacza? Istotne z perspektywy procesu badawczego kwestie można zapisać w taki sposób, aby posłużyły jako cechy warunkujące eksplorację i analizę pewnych zagadnień. Jeśli interesuje mnie nałożenie typologii przemocy na teksty literackie, mogę w tabeli przypisać do każdego tekstu terminy pochodzące z tej typologii. Jeśli chcę określić dynamikę wydawniczą konkretnego zbioru tekstów, każdemu z miejsc wydania mogę przypisać koordynaty geograficzne, a następnie wizualizować ten proces na mapie z wykorzystaniem osi czasu. Jeśli zamierzam z dużego zestawu tekstów wyselekcjonować podzbiór zgodny z moimi zainteresowaniami badawczymi, mogę oznaczać, czy materiał jest zgodny z daną tematyką.

Sformalizowanie pytań badawczych w postaci metadanych nie tylko wpływa na trwałość tych elementów procesu badawczego, ale także pozwala te pytania (albo elementy, dzięki którym można udzielić na nie odpowiedzi) osadzić w interoperacyjnym środowisku, umożliwiającym przetwarzanie i wizualizację treści w taki sposób, aby dodatkowo wspomagać wyciąganie wniosków.

Przepis jest następujący: skonceptualizuj, zapisz w tabeli, przetwórz do takiej postaci, która pozwoli zobaczyć więcej.

Schemat dla badań opartych na (meta)danych

Praca z danymi i metadanymi także wymaga zastosowania struktury. Poniżej dzielę się schematem dla badań opartych na danych, który wytworzyłem wraz z grupą naukowców i naukowczyń podczas przygotowywania tekstu, w którym badaliśmy przemiany dychotomii miasto/wieś w polskiej literaturze pięknej w latach 1864-1939. Tekst dostępny jest tutaj, a korpus, który stworzyliśmy, znajduje się pod tym adresem.

Workflow zakłada 3 przestrzenie: zasobów, metadanych oraz badań, a każda z nich zbudowana jest z dwóch elementów

Zasoby (Resources)

Punktem wyjścia pracy z danymi, podobnie jak w przypadku tradycyjnych badań, jest (1) pytanie badawcze. Na jego podstawie jesteśmy w stanie określić zakres materiału, który potrzebny jest do udzielenia odpowiedzi (RQ-based metadata design). Chodzi zarówno o zestaw niezbędnych elementów, jak i o ich cechy. Kolejnym krokiem jest (2) stworzenie kolekcji danych i metadanych z wykorzystaniem istniejących już zasobów (data collection & re-use). Temu elementowi przyświeca zasada ponownego użycia, która jest kluczowym aspektem w budowaniu progresu dla badań naukowych.

Metadane (Metadata)

W przestrzeni metadanych pierwsze miejsce zajmuje identyfikacja składowych zasobów poprzez (3) wykorzystanie stałych identyfikatorów i struktur LOD (PID and LOD enrichment). Takie podejście umożliwia automatyczną weryfikację oraz pozwala na zarysowanie szerszego kontekstu omawianych zagadnień. Niezbędna jest także (4) manualna weryfikacja oraz ewentualne uzupełnienia (metadata verification and completion). Należy pamiętać, że nie ma niewinnych danych, wszystkie powstają w konkretnym środowisku naukowym i społecznym, mają umożliwić realizację określonych celów oraz są tworzone z wykorzystaniem najróżniejszych metod (na ten temat wypowiadałem się szerzej w jednej z rozmów OPERAS-PL, z której nagrania można wysłuchać na YouTubie). Dodatkowo przestrzeń masowo przetwarzanych danych także zmaga się z błędami i ciągle potrzebuje oka ludzkiego, aby nieprawidłowości wyłapywać.

Badania (Research)

Ostatnia przestrzeń to przede wszystkim (5) wykorzystanie środowiska semantycznego i tworzenie ontologii, które pozwalają definiować relacje między elementami danych oraz komunikować je w sposób łatwy do przeczytania przez maszyny (semantic environment & ontology). Ostatnim krokiem jest (6) zastosowanie metod komputacyjnych, pozwalających eksplorować wytworzone zbiory z wykorzystaniem dedykowanego języka zapytań (computational literary research). Taka struktura pozwala na pisanie kwerend w grafowej bazie danych lub analizę sieci.

Zapośredniczenie własnych badań w porządku pochodzącym spoza domeny humanistyki, osadzenie dorobku naukowego w skomplikowanej technologii, a także konieczność nabycia specjalistycznych kompetencji, by badania przeprowadzić, mogą zniechęcać i przerażać. Doświadczenie badaczy i badaczek pokazuje jednak, że wypracowano już takie formy symbiozy między dyscyplinami, które umożliwiają wspólnotową realizację celów naukowych. Przykładem takich działań może być z jednej strony obecność w projektach badawczych osób definiowanych jako *data scientist*, łączników między światem nauki i technologii, posiadających kompetencje do wytwarzania algorytmów i aplikowania ich w celu odnajdywania wniosków i prawidłowości. Z drugiej strony to aktywność środowiskowa mająca na celu propagowanie skutecznych rozwiązań przy pracy z danymi, dzięki której powstająca dokumentacja i przykłady użycia, zestawy dobrych praktyk czy schematy lub potoki stają się drogowskazami (albo interaktywnymi mapami). Dzięki tym zabiegom nie musimy wyważać otwartych drzwi i możemy skorzystać z (częściowo) już wydeptanych ścieżek.

Wystąpienie konferencyjne na ten sam temat podczas konferencji ADHO Digital Humanities 2024 w Grazu można obejrzeć w serwisie YouTube (ENG).

Pierwodruk: *Dlaczego warto pracować z (meta)danymi? Workflow*, blog *OtwartaHumanistyka* (22 lipca 2024), https://otwartanauka.hypotheses.org/1350.

Linki:

- https://otwartanauka.hypotheses.org/1350#more-1350

Słowa kluczowe: humanistyka cyfrowa, otwarta humanistyka, metadane, dobre praktyki

Autor: Cezary Rosiński