Biuletyn
Polonistyczny

20.06.2024

Metadane jako klucz do szczęścia komunikacyjnego

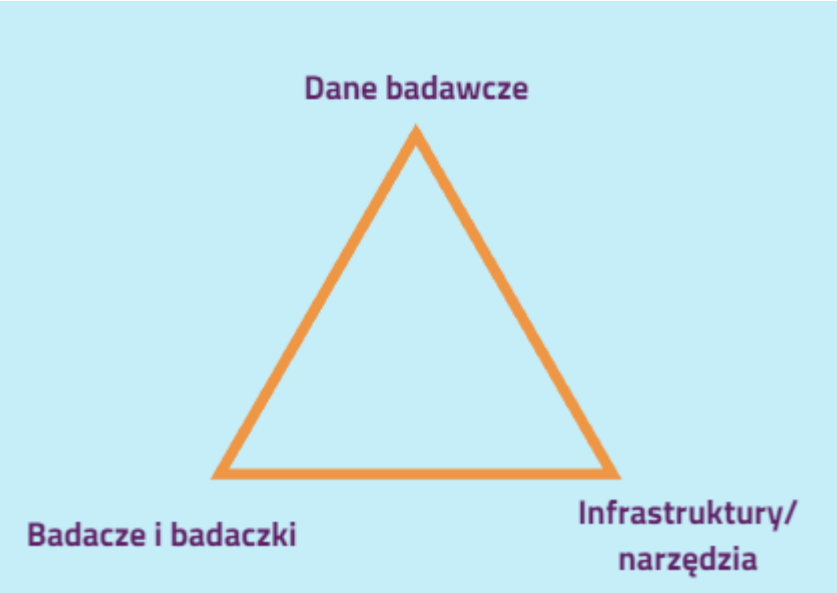
Jeśli przyjmiemy, że nauka to nie tylko proces zdobywania wiedzy, ale także nieustannie powiększający się zasób, który składa się z wyników badań, możemy z pewną śmiałością stwierdzić, że aktywność naukowa ma sens tylko wtedy, gdy mamy dostęp do efektów badań. Tylko znając aktualne osiągnięcia, możemy produkować nową, wartościową wiedzę. Skoro zatem — w trosce o jakość osiągnięć — wyniki badań powinniśmy oprzeć na wcześniejszych odkryciach innych badaczy, to zjawiskiem kluczowym dla nauki staje się komunikacja.

Szacuje się, że od 1996 roku opublikowano przynajmniej 64 miliony prac naukowych. W samym 2022 roku przybyło ich ponad 5 milionów. Przyjmując za punkt odniesienia ministerialną listę składającą się z 56 dyscyplin naukowych i artystycznych, a także pętając założenia logiki szkolną arytmetyką, można uznać, że biegłość w danej przestrzeni naukowej wymaga znajomości prawie miliona stu pięćdziesięciu tysięcy prac. Ograniczając ten zasób tylko do polskojęzycznych publikacji z roku 2021, do podziału zostaje nam około 104 000 tekstów, co i tak daje imponującą liczbę niemal tysiąca ośmiuset sześćdziesięciu rozpraw.

Nie będzie odkryciem stwierdzenie, że taka ilość wiedzy jest nie do przetworzenia przez żadnego badacza i żadną badaczkę. Jak zatem umiejętnie dobierać źródła, żeby utrzymać zdolność efektywnego tworzenia potrzebnej nowej wiedzy? Odpowiedzią jest skuteczna komunikacja.

Krajobraz komunikacji naukowej w humanistyce (z procesem badawczym w tle)

Wraz ze zwiększeniem produkcji wiedzy w parze idzie zestaw zjawisk, które stawiają naukowców i naukowczynie w nowych sytuacjach. Zapośredniczenie prac w środowisku cyfrowym, przekształcenia profilu badawczego ze względu na konieczność nabycia nowych kompetencji, konieczność ewaluacji pracy naukowej i wreszcie zmiany w finansowaniu wynikające z rozwoju modelu projektowego trwale zmieniają krajobraz komunikacji.



Rys. 1. Trójkąt komunikacji naukowej

Oglądane w pewnym uproszczeniu relacje towarzyszące procesom badawczym można wpisać w figurę trójkąta. Jednym wierzchołkiem są badacze i badaczki wraz z ich celami, specyfiką środowiska oraz zewnętrznymi czynnikami, które wywierają wpływ na tę grupę. Drugim wierzchołkiem są infrastruktury i wytwarzane przez nie narzędzia, zarówno w wymiarze materialnym, oznaczającym budynki, przyrządy i obiekty, ale także w ujęciu cyfrowym, związanym z e-infrastrukturami, jak zasoby i usługi, które o ile w zasadzie do prowadzenia badań nie są potrzebne, o tyle w praktyce są już niezbędne. Trzecim wreszcie – dane badawcze rozumiane w najszerszy możliwy sposób jako wszelkiego rodzaju teksty kultury, adnotacje, dane kultury literackiej, ale także jako literatura przedmiotu czy dokumentacja procesu badawczego. Do tego grona można zaliczyć także metadane.

Rolę boków w tym trójkącie spełniają poszczególne zależności. I tak, w relacji między badaczkami i badaczami a danymi badawczymi na pierwszy plan wybija się oczywiście tworzenie wiedzy, ale towarzyszy mu także produkcja wszelkiej maści dodatków, notatek, pozostałości. Z drugiej strony to dane badawcze mają wpływ na losy naukowców i naukowczyń, zarówno ze względu na funkcję efektów prac w badaniu wpływu czy parametryzacji, jak i obecność naukową w dyskursie, będącą następstwem widoczności (i otwartości) samych danych. W relacji między badaczami i badaczkami a infrastrukturą kluczowe wydają się cele naukowe obu grup oraz partnerstwo projektowe. Dodatkowym czynnikiem są potrzeby, wyrażane z jednej strony, z drugiej zaś zaspokajane. Trójkąt domyka relacja danych badawczych i infrastruktur. Z jednej strony „dobre” dane są potrzebne, by wytworzyć skuteczne narzędzia, które to – z drugiej strony – będą pomagały radzić sobie z danymi. Ta cykliczna struktura narzędzi żywiących się danymi i danych wyrastających z narzędzi obarczona jest adekwatnością i nieadekwatnością, które rozciągają się na osi bezużyteczności i doskonałego dopasowania. Warto zwrócić

uwagę, że fundamentem skutecznej współpracy danych i narzędzi jest ich otwartość i dostępność.

I w samym środku tego trójkąta są metadane.

Metadane a skuteczna komunikacja

Może wydać się nieco zaskakujące, że w centrum komunikacji naukowej znajduje się tak techniczna kwestia jak metadane. Tym bardziej, jeśli zredukujemy ich znaczenie do postaci ustrukturyzowanych informacji stosowanych do opisu zasobów lub obiektów. Ale już nawet w tym uproszczeniu kryje się istota metadanych, która polega na tłumaczeniu światu rzeczy. Dlatego warto o metadanych myśleć w towarzystwie takich pojęć, jak zrozumienie, widoczność i wreszcie otwartość.

Skuteczna komunikacja to zrozumiałe zakomunikowanie światu efektów pracy naukowej. Przykładami metadanych, które opisują takie zasoby, jak zbiory danych, ale również pojedyncze teksty naukowe, są słowa kluczowe, deskryptory, hasła przedmiotowe oraz abstrakty. O ile skuteczniej prezentowałyby się słowa kluczowe wpisane w słowniki kontrolowane, które grupowałyby synonimy lub odzwierciedlały relacje hiperonimiczne i hiponimiczne. Przykładem uniwersalnego słownictwa, które rozlewa się na cały zasób, są deskryptory i hasła przedmiotowe, umożliwiające kwerendy tematyczne i skuteczną selekcję materiału.

Skuteczna komunikacja to bogactwo opisu i wykorzystanie jak największej liczby informacji przybliżającej osiągnięcia badawcze. Szczegółowe kategorie, które są dostosowane do specyficznych aktywności, można przełożyć na mechanizmy filtrowania treści i na tej podstawie umożliwić filtrowanie treści, pozwalające przeprowadzić kwerendy w serwisach typu discovery. Skrupulatne, wręcz pedantyczne, opracowanie metadanych przekłada się na znaczne zwiększenie widoczności.

Skuteczna komunikacja to umożliwienie innym skorzystania ze swoich osiągnięć. Ponowne użycie jest kluczową kategorią dla budowania wiedzy nie tylko dlatego, że efekty prac badawczych zostają wykorzystane przez innych naukowców i naukowczynie, ale przede wszystkim z tego powodu, że środowisko w ogóle ma szansę dowiedzieć się o ich istnieniu. Stosowanie otwartych metod publikacji, zapisywanie metadanych w zgodzie z formatami i standardami, tworzenie dokumentacji i opatrywanie zasobów stosownymi licencjami to dobre praktyki badawcze, które wzmacniają wymianę myśli w środowisku naukowym.

Na co powinniśmy uważać, żeby wyszło dobrze

Na całe szczęście nie musimy wymyślać koła od nowa. Nie musimy także wyważać drzwi, bo są już szeroko otwarte. W środowisku cyfrowym istnieją zasady i zestawy dobrych praktyk, które wspomagają skuteczną komunikację. Warto zasygnalizować trzy z nich: FAIR, 5-star data i 7-star data.

Za akronimem FAIR kryje się zapewnienie danym znajdowalności (***F****indability*), dostępności (***A****ccessibility*), interoperacyjności (***I****nteroperability*) i ponownego wykorzystania (***R****e-usability*). Zasady akcentują potencjał maszynowego wykorzystania danych, które powinny być łatwe do znalezienia zarówno przez ludzi, jak i przez programy, aplikacje oraz algorytmy. Wgląd w dane i możliwość ich integracji z innymi zasobami przekładają się na ostateczny cel, którym jest możliwość skorzystania z gotowego już zasobu. FAIRyfikacja danych istotna jest nie tylko przy przygotowaniu planu zarządzania danymi, ale i w każdej innej aktywności badawczej, której produktem są zasoby.

Standard 5-star Open Data, którego autorem jest Tim Berners-Lee, współtwórca usługi WWW, pozwala określić, w jakim wymiarze publikowane dane są dostępne dla użytkowników Internetu, czy są to dane nieustrukturyzowane czy ustrukturyzowane, czy ich użycie wiąże się z koniecznością wykupienia dostępu do narzędzi bądź usług, czy stosowane są mechanizmy identyfikacji zasobów w sieci (URI) i wreszcie czy dane można osadzić w kontekście innych zasobów.

Za dodanie dwóch kolejnych gwiazdek odpowiedzialne są fińskie środowiska akademickie i biblioteczne. Szósta gwiazdka akcentuje wymiar zrozumiałości danych i możliwości ich ponownego wykorzystania dzięki zapewnieniu schematów i dokumentacji, siódma – skupia się na zaufaniu jakości danych poprzez szansę na ich zweryfikowanie i określenie pochodzenia. Tylko zrozumiałe dla nas dane, którym możemy zaufać (np. poprzez rangę instytucji je wytwarzających) będziemy chętnie wykorzystywać do dalszej pracy.

W grupie raźniej

Wielką zaletą tworzenia wiedzy w środowisku cyfrowym jest konieczność współpracy. Figura naukowca zamkniętego w bibliotece i otoczonego jedynie swoim księgozbiorem ustępuje formatowi współpracy między jednostkami i organizacjami o czasem bardzo różnych kompetencjach. Do głosu dochodzi skuteczność komunikacji.

Na zakończenie chciałbym przedstawić formułę cyklu komunikacyjnego, który został wykorzystany w projekcie Dariah.lab we współpracy Instytutu Badań Literackich PAN i konsorcjum naukowego CLARIN-PL. Zaprezentowana formuła jest jednym z przykładów efektywnej współpracy między badaczami i badaczkami a instytucjami (z danymi w tle), ale z łatwością można wyobrazić sobie równie efektywne procesy.

Rys. 2. Cykl komunikacyjny

W naszej współpracy punktem wyjścia były literaturoznawcze teksty naukowe opublikowane w otwartym dostępie. Długotrwała praca przy zasobach bibliograficznych doprowadziła nas do zdefiniowania potrzeby automatycznego uzupełniania metadanych dla tekstów ze względu na dołączane do artykułów materiały, takie jak notki biograficzne, słowa kluczowe czy abstrakty. Podczas pracy

z danymi źródłowymi byliśmy w stanie zbudować typologię elementów obecnych w plikach PDF, a także wskazać ich lokalizację. Następnym krokiem było przygotowanie usługi, która umożliwiła ich identyfikację i ekstrakcję. Tak przetworzony materiał umożliwił nam skorzystanie z kolejnej dedykowanej usługi, która sugeruje słowa kluczowe na podstawie treści abstraktów, zarówno w wymiarze ekstraktywnym (słów występujących w tekście), jak i deskryptywnym (gdy słowo kluczowe jest spoza treści abstraktu). Celem tych działań już niebawem będzie uzupełnienie serwisu European Literary Bibliography o metadane będące efektem wykorzystania metod przetwarzania języka naturalnego, w tym generowania słów kluczowych i topic modelingu.

Pierwodruk: Metadane jako klucz do szczęścia komunikacyjnego. *Otwarta Humanistyka*, https://doi.org/10.58079/wkc5

Linki:

- https://doi.org/10.58079/wkc5

Słowa kluczowe: Open Linked Data, metadane, komunikacja naukowa, FAIR data, Polska Bibliografia Literacka

Autor: Cezary Rosiński