INSTYTUT BADAŃ LITERACKICH POLSKIEJ AKADEMII NAUK
Institute of Literary Research Polish Academy of Sciences

DIGITAL HUMANITIES CENTRE

Dokumentacja kultury a cyfrowe zasoby archiwalne. Przypadek Polskiej Bibliografii Literackiej (PBL) i archiwum Telewizji Polskiej

dr Tomasz Umerle
mgr Beata Koper
mgr Cezary Rosiński

Kontakt:
tomasz.umerle@ibl.waw.pl
beata.koper@ibl.waw.pl
cezary.rosinski@ibl.waw.pl

www.ibl.waw.pl

ul. Nowy Świat 72, 00-330 Warsaw, Poland
phone/fax: (22) 826 99 45, (22) 65 72 895
e-mail: sekretariat@ibl.waw.pl

Plan wystąpienia

Plan wystąpienia:

1. Czym jest PBL? Skąd telewizja w PBL?
2. Przemiany pozyskiwania metadanych dot. obecności lit. w telewizji w ostatnich latach.
3. Wykorzystanie wewnętrznej bazy danych TVP:
   a. zdobycie dostępu,
   b. czyszczenie danych.
4. Wnioski płynące ze współpracy:
   a. istnienie trwałych i istotnych różnic w pojmowaniu roli metadanych,
   b. potrzeba działań promujących precyzyjne poszerzenie oddziaływania utrwalonych już w świecie nauki i *library and information services* idei otwartego dostępu do metadanych,
   c. wniosek roboczy: fakt pozyskiwania metadanych nieobjętych standardami wskazuje, iż coraz istotniejsze dla prac zespołów takich, jak PBL, będą kompetencje dot. czyszczenia “zanieczyszczonych danych” (*messy data*).

Zakres przedmiotowy PBL

PBL gromadzi metadane dotyczące polskiej literatury, teatru i filmu (w Polsce i za granicą) oraz recepcji literatury i teatru w Polsce. Rejestruje druki zwarte (analizuje każdego roku całość piśmiennictwa wydawanego w Polsce), ciągłe (ok. 850 tytułów czasopism każdego roku).

Prowadzi także uporządkowane elektroniczne zbiory danych dot. wydarzeń (ponad 25000), instytucji (prawie 6300) i osób uczestniczących w polskim życiu kulturalnym (ok. 70000) (niezależne od opisów bibliograficznych).

PBL wyróżnia szeroki zakres zbierania materiału - rejestruje również takie formy recepcji literatury, jak słuchowiska radiowe i przedstawienia teatru telewizji oraz programy radiowe i telewizyjne.

Dział “Telewizja”

PBL rejestruje przede wszystkim dwa typy informacji o obecności literatury w telewizji - metadane dotyczące Teatru Telewizji oraz, do roku 1996, audycji dokumentalnych poświęconych zagadnieniom kulturalnym, szczególnie literackim i teatralnym.

Hasła PBL dot. telewizyjnej recepcji literatury:

1. Zagadnienia ogólne (TV)

Zagadnienia wstępne (TV)

Teatr TV (omówienia, nie repertuar)

2. Repertuar TV

Spektakle fabularne TV

Audycje dokumentalne TV (do 1996)

3. Obce realizacje TV polskich utworów literackich

Sposoby gromadzenia metadanych

Źródłem danych, na podstawie których dokumentowano tę problematykę, była specjalistyczna prasa radiowo-telewizyjna (szczególnie tygodnik “Antena”), wspierana sporadycznie tradycyjnymi kwerendami archiwów telewizyjnych.

Lata 2016 i 2017 przyniosły nowe okoliczności opracowywania PBL. Po pierwsze, w roku 2003 (w roku ‘17 zespół opracowuje rocznik 2003 PBL) przestał ukazywać się wspomniany tygodnik “Antena”. Po drugie, realizacja grantu NPRH we współpracy z Centrum Humanistyki Cyfrowej IBL PAN pozwoliła zainwestować część zasobów w przetwarzanie danych cyfrowych.

Sposoby gromadzenia metadanych

W 2017 roku podjęto starania o dostęp do metadanych zasobów telewizyjnych gromadzonych przez TVP, które można by pozyskać w formie i skali umożliwiającej masowy import danych do bazy PBL.

Kontakt z TVP

Ośrodek Dokumentacji i Zbiorów Programowych Telewizji Polskiej zarządza dostępem do archiwaliów telewizyjnych (specjalna procedura opisuje posiadane zasoby, świadczone usługi, metody kontaktu itp.).

PBL nie interesują (przynajmniej nie w pierwszej kolejności) materiały audiowizualne, a ich metadane.

Stąd robocze założenie było takie, aby wykorzystać jakiekolwiek posiadane przez TVP metadane służące do wewnętrznych procedur archiwizacji zasobów.

Kontakt z TVP

Takich usług Ośrodek standardowo nie świadczy - brak informacji o metadanych, sposobie ich przechowywania i udostępniania.

Konieczne było opisanie potrzeb naukowych zespołu PBL w indywidualnych rozmowach (telefonicznych i mailowych) z pracownikami TVP.

Spotkaliśmy się z otwartością i chęcią współpracy oraz - w gruncie rzeczy - zainteresowaniem nietypowym zapotrzebowaniem na zbiory TVP. Owa "nietypowość" wiąże się z faktem zapotrzebowania na "jedynie" metadane, a nie "treść" zasobów.

Wskazano wewnętrzną bazę danych TVP jako najlepsze źródło poszukiwanych informacji.

Pozyskanie metadanych

Ostatecznie sformułowano zamówienie - za pozyskane metadane uiściliśmy opłatę - którego wynikiem były cyfrowo udostępnione metadane 700 spektakli i programów TVP (spektakli Teatru TV i audycji dokumentalnych poświęconych literaturze, teatrowi, filmowi, pisarzom i aktorom) za lata 1989-2016 w formie tabelarycznej (określono format csv; otrzymano dane w szczątkowo tabelarycznej formie w html-u).

Czyszczenie danych

Struktura otrzymanych danych w formacie html zawierała jedynie szczątkowy kod html, który nie dawał możliwości standardowej procedury przekształcenia danych do tabeli, m.in. z powodu braku polecenia <table> w pliku.

Czyszczenie danych

Stan wyjściowy

Kod źródłowy stanu wyjściowego

Usunięcie formatowania i wyszukiwanie wyrażeniami regularnymi

Dane wymagały czyszczenia i wtórnego ustrukturowania.
Pozbyto się nieużytecznego formatowania html oraz uporządkowano podział danych na kolumny (dane zawierały błędy, szczególnie kategorie w kolumnie wartości).

Wyszukiwanie wyrażeniami regularnymi pozwoliło na identyfikację tych kategorii, które znalazły się w kolumnie wartości i następnie na przeniesienie ich do odrębnych wierszy.

Prace w arkuszu kalkulacyjnym

Ujednolicony tekst bez formatowania zaimportowano do arkusza kalkulacyjnego, który umożliwił zdefiniowanie separatora o stałej szerokości i utworzenie dwóch odrębnych kolumn.

Tabelę należało ograniczyć do wersów niepustych (powstałych wskutek przenoszenia kategorii do nowych wierszy) poprzez filtrowanie.

Tabelę, która składała się z właściwej liczby kolumn i zbyt dużej liczby wierszy (na jedną kategorię przypadało wiele wierszy wartości) zmodyfikowano dzięki funkcjom umożliwiającym powielanie pustych wierszy kategorii, złączanie tekstów poszczególnych wierszy wartości w jedną komórkę.

Zdefiniowanie zakresu kategorii i uzyskanie postaci tabeli

Pozyskane metadane zawierały kategorie, które były niepotrzebne z perspektywy Polskiej Bibliografii Literackiej. Usunięto m.in. kategorie umożliwiające zbieranie informacji o cechach fizycznych nośników, prawach autorskich, przeznaczeniu materiałów i tych rodzajach współautorstwa, które wykraczały poza podstawowe funkcje poszczególnych programów.

Zdefiniowanie zakresu kategorii i uzyskanie postaci tabeli

Dla dalszych prac istotne było to, aby każdy zapis miał identyczne kategorie i ich liczbę, dlatego przy zapisach, które nie posiadały kategorii z bazowej listy, należało takowe uzupełnić.

Mimo takich możliwości jak tabele przestawne lub funkcje dodatku Power Query, tabelę zawierającą nazwy kategorii w nagłówku najłatwiej uzyskać poprzez konwersję tekstu na tabelę w edytorze tekstu.

Dalsze prace przygotowujące do importu

[Tabela z danymi (wersja 1 - wielowartościowe pola)]

[Tabela z danymi (wersja 2 - wartości w wydzielonych kolumnach)]

Dalsze prace przygotowujące do importu

Stan początkowy vs. efekt prac

Wnioski z prezentowanego przypadku

1. Istnieją trwałe i istotne różnice w pojmowaniu roli metadanych - perspektywa instytucji gromadzących zasoby (biblioteki, archiwa, muzea etc.) a perspektywa dokumentacji kultury rozumianej jako praktyka gromadzenia metadanych.

Wnioski z prezentowanego przypadku

Jak zauważają Burnett, Ng i Park (w: *A Comparison of the Two Traditions of Metadata Development*), dwie tradycje tworzenia metadanych - biblioteczna i związana z zarządzaniem danymi - odpowiadają na inne potrzeby użytkowników.
Rozróżnienie to możemy tak naprawdę zastosować do opisu różnic w traktowaniu metadanych przez instytucje skoncentrowane na przechowywaniu i udostępnianiu zasobów oraz instytucje zainteresowane zarządzaniem metadanymi:

Podejście biblioteczne	**Zarządzanie danymi**
1. Znajdź materiał dzięki frazie wyszukiwawczej. 2. Zidentyfikuj zasób. 3. Wybierz zasób. 4. Pozyskaj zasób.	1. Jakie dane są dostępne? 2. Czy spełniają one moje potrzeby? 3. Jak je pozyskać? 4. Jak je przenieść do lokalnego systemu?

Wnioski z prezentowanego przypadku

Badacze i zbieracze metadanych o kulturze powinni zadbać o poszerzenie oddziaływania idei wymiany i otwartego dostępu do metadanych, które wywodzą się ze środowisk *library and information services* (wspomnijmy tutaj inicjatywy i narzędzia takie jak Crossref, nowo powstały Metadata 2020 czy szeroko rozumiany ruch Open Access).

Report from the Information Overload and Underload Workgroup (z 2016) wskazuje wśród rozwijających się trendów w tworzeniu i wykorzystaniu zasobów cyfrowych "zwiększanie użycia otwartych metadanych" i "wzbogacanie metadanych i ich standardów"

Wnioski z prezentowanego przypadku

2. Znaczenie metadanych i dostępu do nich nie jest oczywiste wśród ich posiadaczy; jednocześnie rośnie zapotrzebowanie na metadane.

Wnioski z prezentowanego przypadku

Nasz przypadek pokazuje, że w tej sytuacji

1. zacząć należałoby od przekonania instytucji dziedzictwa, iż "już" metadane ich zasobów mają dużą wartość, niezależnie od opisywanej przez nie treści,
2. przekonywać, iż postępujące udostępnianie zdigitalizowanych (archiwalnych, muzealnych, bibliotecznych) zasobów kultury nie zaspokoi potrzeb, które zaspokajają otwarte i wartościowe metadane.

Wnioski z prezentowanego przypadku

Na znaczenie tych ograniczeń w dostępie do metadanych wskazują także inne instytucje dokumentujące kulturę poprzez gromadzenie metadanych, na co wskazywała dyskusja podczas zorganizowanych w VII 2017 r. przez CHC IBL PAN i PBB IBL warsztatów “Nowa dokumentacja kultury” (uczestnicy: Instytut Sztuki PAN, Instytut Historii PAN, Instytut Teatralny, Polskie Centrum Informacji Muzycznej, Biblioteka i Ośrodek Informacji Filmowej PWSFTViT).

Wnioski z prezentowanego przypadku

3. Wniosek roboczy: fakt, iż część metadanych nie jest objęta standaryzacją i służy potrzebom wewnętrznym zwiększa nakłady pracy związane z przystosowaniem danych do wykorzystania w innym środowisku.

Wnioski z prezentowanego przypadku

Sygnalizujemy przy tej okazji chęć współpracy z podobnymi zespołami (badaczy i zbieraczy metadanych) chcącymi podzielić się doświadczeniami czyszczenia danych (np. przy użyciu Open Refine, który wykorzystujemy przy innych projektach, albo innych narzędzi open-access sprawdzonych przy zadaniach tego rodzaju i tej skali, jak różnego typu parsery etc.).

Dziękujemy za uwagę.
Kontakt:

tomasz.umerle@ibl.waw.pl
beata.koper@ibl.waw.pl
cezary.rosinski@ibl.waw.pl

Praca naukowa finansowana w ramach programu Ministra Nauki i Szkolnictwa Wyższego pod nazwą „Narodowy Program Rozwoju Humanistyki” w latach 2015-2018. Projekt: „Polska Bibliografia Literacka – laboratorium wiedzy o współczesnej kulturze polskiej” nr 0061/NPRH3/H11/82/2014.