

Czyli migracja z MAK do Koha

W sierpniu 2012 roku została podpisana [umowa o współpracy Instytutu Józefa Piłsudskiego i Instytutu Naukowego w Nowym Jorku](#). Jej pierwszym owocem jest [wspólny katalog](#) zasobów bibliotecznych obu instytucji, który został uruchomiony w listopadzie 2012. W październiku 2012 roku podobna [umowa została podpisana z Polską Fundacją Kulturalną w Clark, NJ](#), której biblioteka wkrótce rozpocznie dodawanie swojej kolekcji do naszego katalogu. Wspólny katalog ma na celu ułatwić badaczom dostęp do zbiorów naszych instytucji tworząc jedno, docelowe narzędzie do przeszukiwania. Wspólny katalog powinien także usprawnić i przyspieszyć sam proces katalogowania zasobów dzięki "współkatalogowaniu", co jest szczególnie ważne w związku ze skromnymi środkami jakie nasze instytucje mogą przeznaczyć na ten cel. O ile nasze kolekcje nie są identyczne to jednak posiadają sporo duplikatów. Łatwo sobie wyobrazić sytuację, gdzie opisy bibliograficzne stworzone przez katalogujących jednej instytucji, będą mogły być wykorzystane przez drugą, oszczędzając w ten sposób czas i wysiłek.

Tutaj chciałbym przedstawić techniczną stronę łączenia bazy zbiorów biblioteki Instytutu Józefa Piłsudskiego i Instytutu Naukowego.

[Biblioteka Instytutu Piłsudskiego](#) od 2010 roku z powodzeniem korzysta z [zintegrowanego](#)

systemu bibliotecznego Koha

, który oferuje szereg modułów do opisu i zarządzania zbiorami, wypożyczania i obsługi czytelników, jak i posiada elektroniczny, dostępny w sieci katalog publiczny. Biblioteka Instytutu Naukowego z kolei korzystała do tej pory z oprogramowania Biblioteki Narodowej w Warszawie

programu MAK

. Katalog biblioteki Instytutu Naukowego dostępny był tylko lokalnie, na jednym komputerze, co zdecydowanie limitowało jego przydatność. Baza MAK, sięgająca swoimi korzeniami systemu operacyjnemu DOS, była co najmniej onieśmielająca dla pracowników Instytutu, ograniczając w ten sposób jej użytkowników do bibliotekarzy-stypendystów z Biblioteki Narodowej. Wybór Koha jako docelowego systemu dla obu bibliotek wydawał się więc zupełnie logiczny. Trudno wyobrazić sobie dwa bardziej różniące się wyglądem systemy niż Koha i MAK. Przedsięwzięcie połączenia obu baz zbiorów mogłoby na pierwszy rzut oka wydawać się zadaniem bardzo skomplikowanym. Na szczęście oba programy są kompatybilne ze

standardem MARC21

, formatem transferu opisów bibliograficznych. Dzięki temu mogliśmy wyeksportować z bazy MAK jako jeden plik wszystkie opisy bibliograficzne zbiorów i to w formacie, który mógł być teoretycznie niemal natychmiast zaimportowany do bazy Koha. Podobną operację przeprowadziliśmy w przypadku bazy zasobów bibliotecznych Instytutu Piłsudskiego w roku 2010, kiedy to Instytut przesiadł się z MAKa do Koha - mieliśmy więc już sporo doświadczenia i wiedzieliśmy czego się spodziewać.

Pierwszym krokiem po wyeksportowaniu rekordów bibliograficznych Instytutu Naukowego, była ich analiza. Do tego celu wykorzystaliśmy niezawodne narzędzie [MarcEdit](#), a zwłaszcza jego funkcję MARCValidator, która potrafi szybko zanalizować plik z rekordami bibliograficznymi i zidentyfikować dane, które zostały zakodowane niezgodnie ze standardem MARC21. Zgodnie z naszymi poprzednimi doświadczeniami, okazało się, że dane wymagały trochę masażu z naszej strony przed ich ostatecznym importem do wspólnego katalogu. Część problemów w rekordach Instytutu Naukowego wynikała ze szczególnych ustawień instalacji MAK, część z błędów katalogujących, co wcale, należy zaznaczyć, nie umniejszało ogólnej jakości rekordów, która ogólnie była bardzo wysoka. Niektóre operacje przeprowadzane na rekordach były w miarę proste (np. zmiana schematu kodowania znaków diakrytycznych na UTF-8), niektóre z kolei wymagały interwencji programisty. W większości przypadków MarcEdit i jego wszelakie funkcje globalnych edycji rekordów pozwoliły w miarę szybko dokonać koniecznych zmian. Za pomocą MarcEdit zostały także dodane obowiązkowe dane wymagane przez Koha i zostały stworzone pola kodujące informacje o egzemplarzach.

Podczas obróbki rekordów udało nam się także rozwiązać jedną z bardziej palących kwestii kodowania sygnatur. Do tej pory jedyną informacją podpowiadającą gdzie fizycznie znajduje się dana książka była nazwa działu kodowana w jednym z lokalnych pól rekordów bibliograficznych. Działy odpowiadały z kolei trzy-cyfrowym sygnaturom, które były naklejane na grzbiety książek. W ramach działów książki ułożone były alfabetycznie według pozycji głównej

(nazwisko autora lub tytuł). O ile opisy bibliograficzne podawały informacje w jakim dziale książka się znajduje, to brakowało im informacji jaką sygnaturę posiada dany dział. Co więcej ta informacja dawała tylko przybliżone pojęcie gdzie danej książki szukać, co w przypadku większych działów stanowiło problem. Dzięki opracowanym przez Marka Zielińskiego skryptom udało nam się nie tylko dodać sygnaturę do rekordów, ale także dodać do niej pierwszą literę pozycji głównej rekordu, co w znaczący sposób powinno usprawnić odnajdywanie książek na półkach.

Po zakończeniu normalizacji rekordów czekało nas kolejne, dosyć spore przedsięwzięcie porównania rekordów Instytutu Piłsudskiego i Instytutu Naukowego w poszukiwaniu duplikatów. Jedynymi danymi, które mogłyby nam pomóc w tym zadaniu był numer ISBN, choć i ten nie zawsze był unikalny (np. ISBN całości wydawnictwa w przypadku publikacji wielotomowych). Po przefiltrowaniu naszych rezultatów (usunięciu do ręcznej obróbki problematycznych rekordów) dokonaliśmy automatycznego połączenia zidentyfikowanych duplikatów. Pozostawiło to nadal grupę kilkuset rekordów, które nie posiadały numeru ISBN (większość z nich to książki opublikowane przed 1970, kiedy to praktyka przypisywania ISBN została wprowadzona), a które zostały przez nas zidentyfikowane jako potencjalne duplikaty na podstawie porównania pierwszych liter autora i tytułu. Te rekordy zostały przejrane manualnie, rekord po rekordzie, w poszukiwaniu duplikatów. W związku z bardzo dobrą jakością rekordów Instytutu Naukowego przyjęliśmy zasadę, że to one będą rekordami docelowymi, a duplikat Instytutu Piłsudskiego zostanie usunięty. Deduplikacja nie polegała jedynie na usunięciu jednej kopii rekordu, ale także na skopiowaniu z rekordu usuwanego informacji o egzemplarzu (przynależność, stan, sygnatura, etc.), haseł przedmiotowych, a także co ważne, numerów identyfikujących rekord w środowisku Koha, który to służył do ich nadpisania w katalogu przy ich imporcie do wspólnej bazy.

W ten sposób otrzymaliśmy dwa pliki z rekordami: jeden ze znormalizowanymi, unikalnymi rekordami Instytutu Naukowego i drugi z rekordami duplikatów, które miały zastąpić istniejące rekordy w Koha. Przed ich importem musieliśmy jednak dokonać pewnych zmian w ustawieniach Koha. Obie biblioteki zostały potraktowane jako niezależne oddziały w ramach Koha, co pozwoliło na stworzenie zdrowej bariery pomiędzy obiema instytucjami. Bibliotekarze jednej instytucji nie mogą dokonywać zmian w rekordach egzemplarzy drugiej instytucji, nie mają też dostępu do danych czytelników drugiej biblioteki. Zmodyfikowane zostały szablony do tworzenia nowych rekordów, jak i dodane zostały odpowiednie stałe informacje kodowane przez bibliotekarzy Instytutu Naukowego. O ile zrab prac został już wykonany, system nadal wymaga pewnych kosmetycznych zmian nad którymi będziemy w najbliższym czasie pracować.

Podsumowując nasze doświadczenia, przy projektach gdzie wymagana jest manipulacja sporych ilości rekordów bibliograficznych współpraca między bibliotekarzem i programistą wydaje się

nieodzowna. Bibliotekarz uzbrojony nawet w takie doskonałe narzędzia jak MarcEdit nie jest w stanie przeprowadzić wszelkich koniecznych operacji, zwłaszcza jeżeli chodzi o proces deduplikacji dwóch dużych zestawów rekordów.

Tomasz Kalata

Artykuł ukazał się 16 stycznia 2013 w *Blogu archiwistów i bibliotekarzy Instytutu Piłsudskiego*

Może Cię też zainteresować:

- [Humanistyka Cyfrowa w New York City](#)
- [Otwarte czasopisma naukowe a prawa autorskie](#)
- [Czy jesteś GLAM?](#)