EBIB Biuletyn EBIB, nr 5 (194)/2020,
Otwarte dane badawcze
Artykuły

Wojciech Fenrich
w.fenrich@icm.edu.pl
Natalia Gruenpeter
n.gruenpeter@icm.edu.pl
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
Uniwersytet Warszawski

Dziedzinowe repozytoria otwartych danych badawczych

Streszczenie: W artykule przedstawiono genezę, założenia, cele i planowane rezultaty projektu *Dziedzinowe repozytoria otwartych danych badawczych*. Autorzy charakteryzują dane opracowywane i udostępniane w ramach przedsięwzięcia. Omówiono techniczne aspekty związane z opracowaniem oprogramowania dla tworzonych repozytoriów oraz ich funkcjonowanie, jak również działania informacyjne i szkoleniowe skierowane do pracowników polskich instytucji akademickich.

Słowa kluczowe: otwarte dane badawcze, repozytoria danych badawczych

Wprowadzenie

W ostatnich miesiącach i latach na kwestię zarządzania danymi badawczymi coraz intensywniej zwracają uwagę instytucje finansujące badania naukowe oraz rządy. Odpowiedniego zarządzania danymi badawczymi oraz ich udostępnienia w zgodzie z zasadami FAIR[1] oczekuje Komisja Europejska, finansując projekty w ramach programu Horyzont 2020. Stworzenia i realizacji planu zarządzania danymi badawczymi od 2019 r. wymaga również Narodowe Centrum Nauki. Kwestia dostępu do danych badawczych jest też obecna w ramach trwających prac nad projektem „Ustawy o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego”[2] oraz „Programu otwierania danych na lata 2021–2027”[3].

Projekt

W krajobraz ten wpisuje się realizowany od sierpnia 2018 r. projekt *Dziedzinowe repozytoria otwartych danych badawczych* (DRODB) finansowany ze środków Programu Operacyjnego Polska Cyfrowa. Łączna kwota dofinansowania wynosi 4 998 889 PLN, z czego 4 230 559,76 PLN to środki

[1] Zgodnie z zasadami FAIR dane powinny być możliwe do odnalezienia (Findable), dostępne (Accessible), interoperacyjne (Interoperable) i możliwe do ponownego użycia (Reusable) [przyp red.].
[2] *Projekt ustawy o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego*. W: Rządowe Centrum Legislacji [online]. 24.08.2020. [Dostęp 10.11.2020]. Dostępny w: https://legislacja.rcl.gov.pl/projekt/12337400.
[3] *Projekt uchwały Rady Ministrów w sprawie Programu otwierania danych na lata 2021-2027*. W: Ministerstwo Cyfryzacji [online]. 30.10.2020. [Dostęp 10.11.2020]. Dostępny w: https://mc.bip.gov.pl/projekty-aktow-prawnych-mc/projekt-uchwaly-rady-ministrow-w-sprawie-programu-otwierania-danych-na-lata-2021-2027.html.

wspólnotowe, a 768 329,24 PLN środki z budżetu państwa. Liderem przedsięwzięcia jest Uniwersytet Warszawski, gdzie za jego realizację odpowiadają Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego oraz Instytut Studiów Społecznych im. Profesora Roberta Zajonca. Partnerami są Instytut Filozofii i Socjologii Polskiej Akademii Nauk oraz Uniwersytet Adama Mickiewicza w Poznaniu, gdzie projekt realizowany jest na Wydziale Chemii. W ramach inicjatywy przewidziane jest opracowanie i udostępnienie danych badawczych w postaci 400 zbiorów z zakresu nauk społecznych, 200 zbiorów krystalograficznych oraz 20 zbiorów z zakresu innych nauk. Całkowity rozmiar danych planowanych do udostępnienia przekracza 3 terabajty. Dane opracowane w ramach działań projektowych są sukcesywnie udostępniane z preferencją dla modelu otwartego wykorzystującego wolne licencje. Jedynie w przypadku części danych społecznych konieczne będzie udostępnienie zasobów w bardziej restrykcyjnych modelach z uwagi na istniejące ograniczenia o charakterze etycznym oraz prawnym. Dostęp do danych będzie możliwy zarówno za pomocą przeglądarki internetowej i graficznego interfejsu użytkownika, jak i za pośrednictwem interfejsów programistycznych (API).

Repozytoria

W ramach projektu uruchomione zostały trzy repozytoria danych badawczych: repozytorium ogólnego przeznaczenia „RepOD", który to serwis przejął funkcje pilotażowej wersji repozytorium, prowadzonej wcześniej przez ICM UW w oparciu o oprogramowanie CKAN, repozytorium danych krystalograficznych „MX-RDR" oraz repozytorium danych społecznych „RDS". W obrębie tego ostatniego funkcjonują odrębne kolekcje dla danych ilościowych i danych będących wynikiem badań o charakterze jakościowym. Infrastruktura ta służy do udostępniania opracowywanych w ramach projektu danych badawczych. Każde z repozytoriów jest też otwarte na nowych użytkowników oraz dane wytworzone i opracowane poza projektem. Aby zdeponować dane, konieczne jest jedynie założenie konta w repozytorium. Każdy zbiór danych udostępniony w repozytoriach otrzymuje swój numer DOI. Zarówno zdeponowanie danych, jak i ich pobranie są bezpłatne. W każdym z trzech serwisów przed opublikowaniem zbioru jest on weryfikowany przez redaktora repozytorium lub administratora konkretnej kolekcji, jeśli zbiór deponowany jest w jej obrębie. Na mocy porozumienia z Uniwersytetem Warszawskim, zainteresowane instytucje mogą też tworzyć w repozytorium RepOD swoje kolekcje instytucjonalne i nimi zarządzać.

Oprogramowanie

Repozytoria funkcjonują w oparciu o otwarte oprogramowania Dataverse. W ramach projektu bazowa wersja tego oprogramowania została zmodyfikowana i rozszerzona o dodatkowe funkcjonalności (np. funkcjonalność embarga, pozwalającą opóźnić udostępnienie plików wchodzących w skład zbioru danych, czy zmiana sposobu określania licencji udostępnianych zasobów, która została zestandaryzowana i dopasowana do polskich warunków prawnych). Wersja ta stanowi podstawę dla repozytorium ogólnego przeznaczenia RepOD. W toku projektu wersja ta została poddana dalszym modyfikacjom z myślą o repozytoriach dziedzinowych. Rozszerzone zostały istniejące już schematy metadanych (generyczny schemat „Citation" oraz schemat dziedzinowy dla nauk społecznych), dodano też zupełnie nowy schemat metadanych dla nauk krystalograficznych. Opraco-

wano też dodatkowe narzędzia specyficzne dla danej dziedziny, takie jak moduł automatycznej analizy danych krystalograficznych, będący elementem oprogramowania repozytorium MX-RDR. Przeprowadzono też szereg prac związanych z podniesieniem jakości kodu oprogramowania, które – choć niewidoczne dla końcowego użytkownika repozytorium – znacząco ułatwią jego dalsze modyfikacje, np. na potrzeby repozytoriów dziedzinowych, które mogą powstać w przyszłości.

Serwisy towarzyszące

Repozytoriom towarzyszy również witryna projektu DRODB zawierająca bieżące informacje dotyczące realizacji zadań projektowych, w szczególności informacje o przeprowadzonych i planowanych szkoleniach z zakresu zarządzania danymi badawczymi i ich udostępniania. Można też znaleźć na niej trzy broszury informacyjne, z których pierwsza poświęcona jest selekcji i przygotowaniu danych do udostępnienia, druga – korzystaniu z zasobów udostępnionych w repozytoriach danych, a trzecia – prawnym aspektom otwierania dostępu do danych badawczych. Uruchomiono również witrynę informacyjną repozytorium RepOD, na której można znaleźć dodatkowe informacje dotyczące funkcjonowania repozytorium oraz poradnik wyjaśniający prawne aspekty deponowania danych badawczych.

Szkolenia z zarządzania danymi badawczymi

W ramach projektu organizowane są szkolenia, których celem jest przybliżenie dobrych praktyk w zakresie zarządzania danymi badawczymi i otwartego udostępniania danych. Zagadnienia te stały się dla polskich badaczy szczególnie istotne w roku 2019, kiedy Narodowe Centrum Nauki wprowadziło wymóg sporządzania planu zarządzania danymi badawczymi przy składaniu wniosków grantowych oraz zapowiedziało wprowadzenie polityki otwartości. Program spotkań nawiązuje do warsztatów z zarządzania danymi badawczymi, które zespół działającej w ICM UW Platformy Otwartej Nauki prowadził już od roku 2015. Na program szkolenia składają się: krótka prezentacja założeń projektu oraz części praktyczna i prawna. Część praktyczna obejmuje wprowadzenie do tematu, omówienie korzyści związanych z udostępnianiem danych i wymogów instytucji finansujących badania naukowe (Narodowe Centrum Nauki i Komisja Europejska w ramach programu Horyzont 2020) oraz przybliżenie dobrych praktyk zarządzania danymi z uwzględnieniem zasad FAIR. Część prawna służy zarysowaniu szerszego tła niezbędnego do zrozumienia otoczenia prawnego, które należy uwzględnić, planując zarządzanie danymi badawczymi. Otwarte udostępnianie danych omawiane jest z uwzględnieniem praw własności intelektualnej, ochrony danych osobowych, a także kwestii związanych z komercjalizacją wyników badań naukowych. Prezentacja uwzględnia także podstawowe informacje na temat wolnych licencji i ich stosowania w odniesieniu do danych.

Pierwsze warsztaty zorganizowane w ramach projektu odbyły się pod koniec 2019 r. w Poznaniu i Toruniu. W sumie wzięło w nich udział prawie 40 osób: naukowców, bibliotekarzy akademickich oraz pracowników biur obsługi projektów badawczych. Stacjonarna forma szkoleń ma charakter warsztatowy, dlatego liczba uczestników pojedynczego szkolenia była ograniczona do maksymalnie 25 osób. Część wykładowa, obejmująca wymienione wcześniej zagadnienia, uzupełniona była dwoma ćwiczeniami praktycznymi. Pierwsze polegało na lekturze i ocenie przykładowych planów

zarządzania danymi, drugie – na dyskusji uwzględniającej różne role i interesy podmiotów zaangażowanych w realizację projektu badawczego, m.in. naukowców i ich pracodawców, instytucji finansujących badania czy komercyjnych partnerów przedsięwzięcia. Ćwiczenia przeprowadzane były odpowiednio po częściach praktycznej i prawnej. Ze względu na pandemię COVID-19 warsztaty zaplanowane na pierwszą połowę 2020 r. nie odbyły się w formie stacjonarnej. Szkolenia wznowione zostały z końcem marca 2020 r. w formie zdalnej, która wprawdzie ograniczyła możliwość przeprowadzenia bezpośredniej dyskusji oraz ćwiczeń o charakterze warsztatowym, ale pozwoliła na dotarcie do większej liczby osób zainteresowanych tematyką danych badawczych. W siedmiu kolejnych szkoleniach online, organizowanych co miesiąc od marca do września 2020 r., udział wzięło ponad 450 osób z całej Polski. Po każdym szkoleniu uczestnicy zachęcani są do wypełnienia ankiet ewaluacyjnych, które dla prowadzących oraz dla całego zespołu projektu stanowią ważne źródło informacji o oczekiwaniach, problemach i wątpliwościach uczestników, a często także o wyzwaniach, jakie stoją przed nimi bądź zatrudniającymi ich instytucjami.

Podsumowanie

Realizacja projektu potrwa do połowy roku 2021. W jej wyniku społeczność naukowa zyska zarówno trzy serwisy umożliwiające udostępnianie danych badawczych zgodnie z najlepszymi światowymi praktykami, jak i szeroki zasób zbiorów danych badawczych do ponownego wykorzystania i analizy. Uczestnicy prowadzonych w ramach przedsięwzięcia szkoleń zyskają wiedzę i kompetencje w zakresie zarządzania danymi badawczymi, dzięki którym będą mogli w pełni wykorzystać możliwości stwarzane przez nowe repozytoria. Projekt pomyślany został w taki sposób, by jego efekty mogły oddziaływać jeszcze długo po jego zakończeniu: powstała infrastruktura przez następne lata służyć będzie udostępnianiu kolejnych porcji danych (w tym również tych, które nie były opracowywane w ramach samego projektu), a powstałe oprogramowanie zachowuje charakter otwarty i pozostaje do dyspozycji wszystkich podmiotów zainteresowanych jego dalszym udoskonalaniem lub wykorzystaniem w obecnej postaci.

FENRICH, W., GRUENPETER, N. Dziedzinowe Repozytoria Otwartych Danych Badawczych. *Biuletyn EBIB* [online]. 2020, nr 5 (194), Otwarte dane badawcze. [Dostęp 17.11.2020]. ISSN 1507-7187. Dostępny w: http://open.ebib.pl/ojs/index.php/ebib/article/view/698.