

Marta Hoffman-Sommer
ICM UW
m.hoffman-sommer@icm.edu.pl

RepOD – Repozytorium Otwartych Danych na Uniwersytecie Warszawskim

Słowa kluczowe: repozytoria polskie, dane badawcze

Jesienią 2015 r. w Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego (ICM) uruchomiło pierwsze w Polsce ogólnodostępne repozytorium otwartych danych badawczych — RepOD (Repozytorium Otwartych Danych). RepOD jest przeznaczone do przechowywania i udostępniania danych ze wszystkich dziedzin nauki. Każdy, kto prowadzi w Polsce badania naukowe, może zdeponować swoje dane w repozytorium. W ten sposób zapewnia ich bezpieczną, długoterminową archiwizację. W przypadku RepOD jedynym warunkiem przyjęcia danych jest udostępnienie ich w sposób otwarty, to znaczy tak, by inni użytkownicy mieli do nich swobodny dostęp.



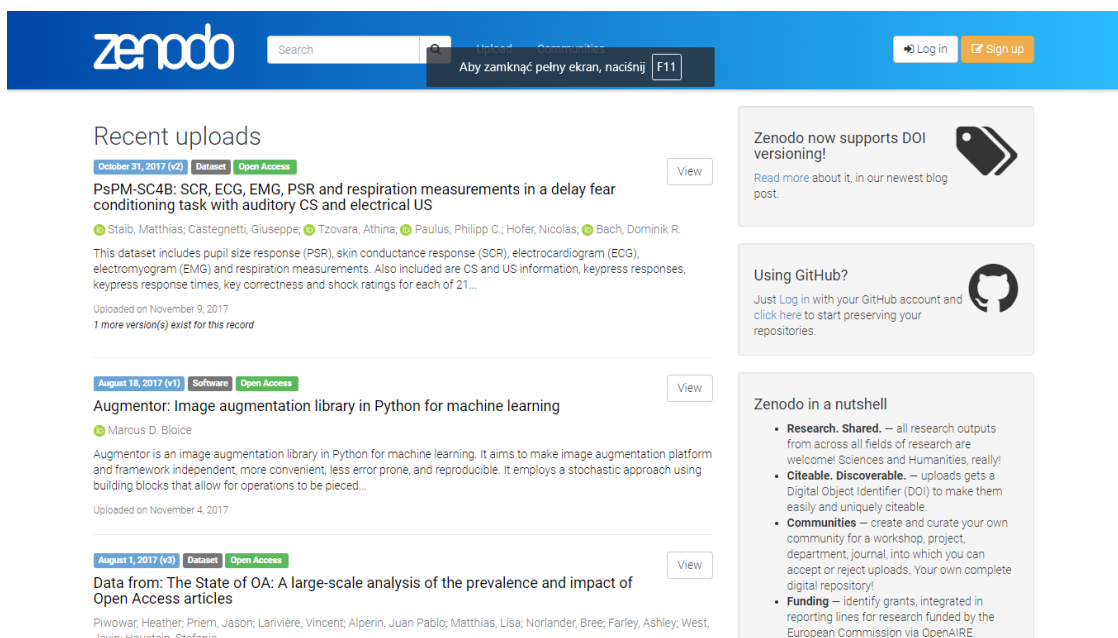
Il. 1. Zrzut ekranu RepOd.-

Źródło RepOD, ICM 2017. [Data dostępu 12 grudnia 2017]. Dostępne w: <https://repod.pon.edu.pl/pl/>.

W czasie prowadzenia badań naukowych gromadzone i wytwarzane są dane najróżniejszego rodzaju. Zaliczamy do nich nie tylko wyniki różnych pomiarów, ale również zdjęcia, filmy, wyniki symulacji komputerowych, modele matematyczne, powiązane z nimi programy komputerowe. Uzyskane dane naukowe najlepiej przechowywać w repozytoriach specjalistycznych, przeznaczonych dla bardzo konkretnego rodzaju danych — najbardziej znanym repozytorium tego typu jest GenBank

(<https://www.ncbi.nlm.nih.gov/genbank/>), w którym przechowywane są sekwencje genetyczne deponowane przez naukowców z całego świata. Repozytoria specjalistyczne mają wiele zalet: są technicznie przystosowane do przyjmowania danych w określonych formatach i ze ściśle określonym zestawem metadanych, dzięki czemu zapewniają interoperacyjność przechowywanych zbiorów, możliwość łączenia ich w większe zbiory. Zwykle są także dobrze znane wśród społeczności badaczy z danej dyscypliny.

Jednak nie każdy zbiór danych znajdzie miejsce w repozytorium specjalistycznym. W każdej dziedzinie nauki zbierane są również dane nietypowe, niepasujące do funkcjonujących repozytoriów i baz specjalistycznych. Z myślą o nich prowadzone są repozytoria ogólne, w których można przechowywać dane zróżnicowane, w dowolnych formatach. Jednym z największych repozytoriów tego typu jest Zenodo (<https://zenodo.org/>), prowadzone w CERN i finansowane ze środków Komisji Europejskiej poprzez projekt OpenAIRE. Jest ono otwarte dla wszystkich europejskich naukowców.



II.2. Zrzut ekranu repozytorium Zenodo

Źródło Zenodo. CERN 2017. [Data dostępu 12 grudnia 2017]. Dostępne w: <https://zenodo.org/>.

Podobna idea przyświeca repozytorium RepOD, które działa na skalę bardziej lokalną, krajową. Dzięki temu z jednej strony jest bardziej odpowiednie dla danych o lokalnym charakterze, z drugiej zaś może zaoferować większe wsparcie osobom zainteresowanym deponowaniem danych. Pytania deponentów najczęściej dotyczą kwestii prawnych albo technicznych aspektów przygotowania zbiorów danych: formatów plików i sposobu ich opisu. Niejasności dotyczą tego, kto ma prawo zdecydować o przekazaniu danych do repozytorium — bezpośredni twórca, kierownik projektu naukowego, w ramach którego prowadzone były badania, czy jednostka naukowa, w której twórca danych jest zatrudniony. Także kwestia wyboru prawnego sposobu udostępnienia danych stwarza badaczom problemy: czy udostępnić dane bez licencji dla użytkowników (na zasadach

1 Wszystkie odesłania do internetu były aktualne 12 listopada 2017 r.

dozwolonego użytku), czy nadać im jakąś licencję, a jeżeli tak, to jak daleko idącą? Z kolei spośród aspektów technicznych najczęściej poruszane są kwestie dopuszczalnych formatów plików oraz zakresu metadanych i potrzebnej dokumentacji danych.

Obecnie RepOD przechowuje około 50 zbiorów danych. Przeważająca część dotyczy nauk przyrodniczych, choć reprezentowane są również nauki techniczne, medyczne i społeczne. Jest kilka zespołów naukowych, które regularnie deponują dane, ale większość użytkowników zdeponowała na razie pojedyncze zbiory danych. Często te pojedyncze zbiory są powiązane z konkretnymi artykułami naukowymi, które pojawiają się w czasopiśmie wymagających udostępniania danych. W niektórych dziedzinach nauki wiele czasopism wymaga, by wszystkie dane niezbędne do zweryfikowania lub replikacji publikowanych wyników zostały publicznie udostępnione w internecie. Część użytkowników właśnie ze względu na wymagania wydawców decyduje się zdeponować dane w RepOD.



II. 3. Zrzut ekranu RepOd

Źródło RepOD, ICM 2017. [Data dostępu 12 grudnia 2017]. Dostępne w: <https://repod.pon.edu.pl/pl/dataset>.

Coraz więcej badaczy w Polsce jest świadomych istnienia ogólnych repozytoriów danych. Oprócz działań wydawców naukowych wpływ na to mają również działania edukacyjne, prowadzone między innymi przez ICM UW. W ramach projektu OpenAIRE 2020 ICM przeprowadził w ostatnich dwóch latach ponad dziesięć warsztatów z zarządzania danymi badawczymi, które przygotowują uczestników do deponowania danych w repozytoriach. O otwieraniu danych mówi się też coraz więcej w instytucjach naukowych na terenie całego kraju. Dla dalszego rozwoju repozytorium duże znaczenie będą mieć właśnie z jednej strony działania edukacyjne i informacyjne w polskim środowisku naukowym, a z drugiej strony również przyszłe działania MNiSW względem otwartych danych w polskiej nauce.