EBIB Biuletyn EBIB, nr 6 (183)/2018,
Współpraca bibliotek z naukowcami
Artykuły

Małgorzata Rożniakowska-Kłosińska
Biblioteka Politechniki Łódzkiej
mroz@lib.p.lodz.pl

Otwarte dane badawcze w warsztacie pracy naukowca

Streszczenie: Artykuł zawiera podstawowe informacje dotyczące zagadnienia otwartych danych badawczych. We wprowadzeniu omówiono pojęcie otwartej nauki, a w dalszej części artykułu problematykę otwierania danych badawczych. Przedstawiono również przykładowe elementy, jakie powinny się znaleźć w *Planach zarządzania danymi*.

Słowa kluczowe: otwarta nauka, otwarte dane badawcze, otwarty dostęp, Horyzont 2020

Wprowadzenie

Otwarty dostęp do publikacji naukowych i otwarte dane badawcze to dwa istotne filary otwartej nauki. Jej nieformalną koncepcję zaproponował w 2011 r. Michael Nielsen (naukowiec, z wykształcenia fizyk kwantowy, a od roku 2008 również rzecznik otwartej nauki[1]): *to idea, która zakłada, że wszelkiego rodzaju wiedza naukowa powinna być otwarcie rozpowszechniana tak wcześnie, jak jest to praktyczne w procesie odkrywania*[2], [3].

Do wiedzy naukowej wszelkiego rodzaju Nielsen zaliczył m.in. artykuły w czasopismach, dane, kod (źródłowy) oraz idee i rozważania naukowe. Natomiast zawarty przez niego warunek wykonalności miał wskazywać, że bardzo często istnieją innego rodzaju czynniki np. prawne, etyczne lub społeczne, które trzeba wziąć pod uwagę[4]. Niemniej jednak należy podkreślić, że kiedy dostęp do publikacji naukowych i danych badawczych jest z jakichś powodów ograniczony prawnie, lokalizacyjnie czy też subskrypcyjnie, to komunikacja naukowa przestaje być efektywna i może ulec całkowitemu zahamowaniu – zwłaszcza w środowisku cyfrowym.

Otwarta nauka jest jednym z ważniejszych priorytetów Komisji Europejskiej, obok otwartej innowacyjności i otwartości na świat[5]. Intensywny rozwój technologii informatycznych i ich dostępność wpływa na kształtowanie się nauki obywatelskiej, na model współpracy naukowej oraz skalę realizowanych projektów badawczych i dzielenie się ich wynikami. Dlatego utworzenie Europejskiej Chmury dla Otwartej Nauki (*European Open Science Cloud*), o fazie implementacji przewidzianej na lata 2018–2019, jest jednocześnie wielopoziomową strategią rozwoju otwartości oraz wyzwaniem sprzętowym.

[1] NIELSEN, M. *Michael Nielsen* [online]. [Dostęp 22.09.2018]. Dostępny w: http://michaelnielsen.org/blog/michael-a-nielsen.
[2] *Otwarta nauka* [online]. [Dostęp 22.09.2018]. Dostępny w: https://pl.wikipedia.org/wiki/Otwarta_nauka.
[3] *Open science is the idea that scientific knowledge of all kinds should be openly shared as early as is practical in the discovery process*. Fragment z: NIELSEN, M. Re: *Definitions of Open Science?* Message to Peter Murray-Rust. 28.07.2011. E-mail [online]. [Dostęp 22.09.2018]. Dostępny w: https://lists.okfn.org/pipermail/open-science/2011-July/000907.html.
[4] Tamże, tłum. aut.
[5] *Open Innovation, Open Science, Open to the World – a vision for Europe* [online]. European Commission, Directorate-General for Research & Innovation, 2016. [Dostęp 22.09.2018]. Dostępny w: http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=16022.

Otwarte dane badawcze

Trudno jest jednoznacznie zdefiniować pojęcie danych badawczych. Z tym zadaniem przyjdzie się zmierzyć każdej jednostce prowadzącej badania, zwłaszcza na etapie tworzenia przez nią lokalnych procedur zarządzania i organizacji przepływu strumieni danych (nie tylko badawczych), powstających w ramach realizowanych grantów i projektów naukowych. Co więcej, definicje sformułowane przez specjalistów z obszaru nauk społecznych i humanistycznych, będą różnić się od tych, które stworzą przedstawiciele nauk medycznych lub inżynieryjno-technicznych. Wynika to przede wszystkim ze specyfiki poszczególnych dziedzin, w których mogą być stosowane różne techniki i narzędzia, chociażby do samego rejestrowania danych badawczych. Wykaz postaci wyników badań jest katalogiem otwartym oraz niezależnie od dyscypliny zmienia się dynamicznie. Obejmuje on na przykład: proste dane liczbowe i statystyki z eksperymentów, wyniki ankiet, dane z obserwacji, wizualizacje 2D i 3D, złożone modele matematyczne. Należy również zauważyć, że odmiennym metodom analizy i interpretacji podlega materiał empiryczny zebrany w badaniach jakościowych i ilościowych[6]. Stąd właśnie, w literaturze przedmiotu przytaczane jest tak szerokie spektrum definicyjne[7]. Przyjmuje się, że pierwszym kompleksowo przedstawionym pojęciem danych badawczych: *zarejestrowane materiały o charakterze faktograficznym, powszechnie uznawane przez społeczność naukową za niezbędne do oceny wyników badań naukowych*, posłużyła się w 1999 r. amerykańska jednostka rządowa *Office of Management and Budget*[8]. Definicję umieszczono w okólniku zawierającym ujednolicenie wymagań administracyjnych dotyczących przyznawania dotacji i zawierania umów z instytucjami szkolnictwa wyższego, szpitalami i innymi organizacjami *non-profit*. Wykluczono z niej: wstępne analizy, szkice publikacji i recenzje, treści dyskusji naukowych oraz obiekty fizyczne. Dodatkowo do danych badawczych nie zaliczono materiałów zawierających informacje handlowe, ani tych, które mogą być podstawą do uzyskania patentu. Wyłączono również dane personalne oraz medyczne, których ujawnienie byłoby nieuzasadnioną ingerencją w prywatność czy identyfikowałoby konkretną osobę, która wzięła udział w badaniu.

Komisja Europejska definiuje dane badawcze w podobny sposób. Są to: *informacje, w szczególności fakty, liczby, zebrane do analizy i uważane za podstawę do dalszego wnioskowania, dyskusji lub obliczeń*[9]. Zapis ten został umieszczony w dokumencie *Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020*, zawierającym przede wszystkim wytyczne niezbędne do realizacji otwartego dostępu do recenzowanych publikacji naukowych i danych badawczych,

[6] CISEK, S. *Zbiory danych badawczych online* [online]. 2014. [Dostęp 22.09.2018]. Dostępny w: https://www.slideshare.net/sabinacisek/cisek-zarzdzanie-inf-w-nauce-2014.

[7] STRZELCZYK, E. Otwarte dane badawcze – kolejny krok do otwierania nauki. W: SÓJKOWSKA, I., DERFERT-WOLF, L. (red.). *Bibliograficzne bazy danych: perspektywy i problemy rozwoju. III Konferencja Naukowa Konsorcjum BazTech, Kraków, 26–27 czerwca 2017* [online]. Stowarzyszenie EBIB, 2017. [Dostęp 20.09.2018]. Materiały Konferencyjne EBIB, nr 25. ISBN 978-83-63458-08-9. Dostępny w: http://open.ebib.pl/ojs/index.php/Mat_konf/article/view/599.

[8] *CIRCULAR A-110 REVISED 11/19/93 As Further Amended 9/30/99* [online]. Office of Management and Budget, 30.09.1999. [Dostęp 22.09.2018]. Dostępny w: https://georgewbush-whitehouse.archives.gov/omb/circulars/a110/a110.html.

[9] *Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020* [online]. European Commission, Directorate-General for Research & Innovation, 21.03.2017. [Dostęp 22.09.2018]. Dostępny w: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.

będących rezultatem projektów finansowanych ze środków publicznych w programie Horyzont 2020. Zatem otwarte dane badawcze to takie, do których każdy ma prawo dostępu, i standardem jest, że można je wykorzystywać, przetwarzać, powielać i rozpowszechniać w sposób nieodpłatny.

W latach 2016–2017 Komisja Europejska zrealizowała eksperymentalne działanie *Open Research Data Pilot* dla wybranych obszarów tematycznych programu Horyzont 2020. Od stycznia 2017 r. rozszerzono ten pilotaż jako działanie *Open Research Data by Default*.

Zgodnie z wytycznymi zawartymi w *Open Research Data Pilot* uczestnicy byli zobowiązani do:

- przygotowania i aktualizowania *Planu zarządzania danymi* (*Data Management Plan*),
- zdeponowania danych w repozytoriach danych badawczych,
- określenia zasad swobodnego wykorzystywania danych (w tym licencje CC-BY lub oświadczenia CC0)[10],
- określenia, jakich narzędzi należy użyć w celu weryfikacji danych surowych (lub dostarczenie takich narzędzi).

Komisja Europejska zastosowała życzliwe podejście do podmiotów, które zdecydują się na wzięcie udziału w działaniu pilotażowym:

- *Plan zarządzania danymi* nie podlegał ocenie i nie był wymagany na etapie składania wniosku,
- możliwość wycofania się bez konsekwencji i na każdym etapie projektu,
- możliwość wybrania danych badawczych do otworzenia zgodnie z filozofią: tak otwarte, jak to możliwe, tak zamknięte, jak to konieczne (*as open as possible, as closed as necessary*)[11].

Działaniem *Open Research Data by Default* zostały objęte natomiast wszystkie obszary tematyczne programu Horyzont 2020. Zachowano również możliwość decyzji *opt-out*. W obu działaniach koszty zarządzania danymi badawczymi są kwalifikowalne.

Oczywistym jest, że każde dane badawcze, zarówno te pochodzące z eksperymentu lub zebrane w wyniku obserwacji, są nierozerwalnie związane z publikacją. Dlatego jeśli publikacje, które powstały w ramach finansowania ze środków publicznych, mają być otwarte, to niewątpliwie słuszna jest koncepcja otwierania danych badawczych *by default*. Można w tym miejscu zaryzykować stwierdzenie, że publikacja jest tak naprawdę otwarta, jeśli towarzyszące jej dane badawcze (surowe lub/i oczyszczone), są także bez ograniczeń dostępne.

Zarządzanie danymi badawczymi i ich otwieranie

Zrozumienie korzyści płynących z otwartej nauki jest pierwszym krokiem do podjęcia konkretnych działań związanych z otwieraniem publikacji naukowych i danych badawczych. Dzięki zastosowaniu otwartości następuje ograniczenie kosztów powielania badań, zwięk-

[10] Recommended conformant licenses. W: *Conformant Licenses* [online]. [Dostęp 22.09.2018]. Dostępny w: https://opendefinition.org/licenses.
[11] *Guidelines to the Rules…, dz. cyt.*

szenie ich wydajności, co przekłada się na potencjalnie szybsze odkrycia i tym samym następuje przyspieszenie postępu naukowego. Kiedy dostęp do literatury nie będzie obwarowany płatnymi subskrypcjami, ulegnie zwiększeniu widoczność dorobku naukowego, co wpłynie na stopień oddziaływania badań i wzrost liczby cytowań. Otwarty dostęp do danych badawczych ma także wpływ na ilość cytowań publikacji z nimi powiązanej[12]. Możliwość weryfikacji i odtworzenia danych badawczych (*reproducibility*), ich ponownego wykorzystania (*re-use*), w tym również komercyjnie, wspomaga podniesienie jakości późniejszych badań. Dodatkowo, otwieranie dostępu do danych badawczych wpływa na zwiększenie prawdopodobieństwa nawiązywania kontaktów z innymi grantodawcami i naukowcami, również spoza bazowej dziedziny, działających na przykład na styku różnych dyscyplin. Wśród pracowników naukowych konieczne jest także zwiększenie świadomości, że dzięki udostępnieniu danych badawczych, mogą stać się współautorami publikacji, w której te dane zostały wykorzystane jako część innego projektu badawczego.

Zbiory danych (*datasets*) wygenerowane w procesie naukowym ulegają zmianie w trakcie realizacji całego projektu badawczego od postaci surowej (*raw data*), po częściowo oczyszczoną i uporządkowaną, aż do postaci finalnej, podlegającej publikacji. Dokumentowanie prowadzonych badań jest częścią warsztatu pracy każdego naukowca. Narzędziem wspomagającym tę natywną działalność jest, wcześniej już wspomniany, *Plan zarządzania danymi*. Europejskie instytucje finansujące coraz częściej wymagają od grantobiorców przygotowywania takich planów jako warunku przystąpienia do konkursu. Co więcej, ich merytoryczna jakość może wpływać na końcową ocenę projektu, i w efekcie na otrzymanie, bądź nie, funduszy na badania. W *Planie zarządzania danymi* grantobiorca musi wskazać, w którym repozytorium będą deponowane dane badawcze, ale osobą decyzyjną w tej kwestii może być także grantodawca. *Plan zarządzania danymi* powinien przykładowo zawierać[13]:

- informacje o charakterze administracyjnym: opis projektu, uczestnicy projektu, grantodawca, rozpoznanie polityk instytucjonalnych w zakresie przetwarzania danych badawczych,
- informacje o charakterystyce powstałych podczas projektu zbiorów danych badawczych, wytyczne odnośnie metodologii i standardów, które będą użyte podczas rejestracji danych,
- informacje o sposobie dokumentacji procesu badawczego i wyborze standardu metadanych,
- informacje o tym, jak będą przestrzegane prawa własności intelektualnej, ochrona prywatności itp.,
- zasady tworzenia kopii zapasowych i długoterminowego przechowywania,
- zasady zarządzania bezpieczeństwem danych badawczych, np. wg normy ISO/IEC 27001,
- zasady ponownego wykorzystania danych badawczych, bariery techniczne i prawne,
- informacje dotyczące kosztów przechowywania danych badawczych (infrastruktura informatyczna),
- informacje o podziale odpowiedzialności uczestników projektu badawczego w zakresie gromadzenia, przetwarzania i udostępniania danych badawczych.

[12]PIWOWAR, H.A., VISION, T.J. Data reuse and the open data citation advantage Vision. *PeerJ* [online]. 2013, 1:e175. [Dostęp 22.09.2018]. Dostępny w doi: https://doi.org/10.7717/peerj.175.
[13]*Checklist for a Data Management Plan. v.4.0* [online]. Edinburgh: Digital Curation Centre, 2014. [Dostęp 22.09.2018]. Dostępny w: http://www.dcc.ac.uk/webfm_send/1279.

Deponowane dane badawcze powinny spełniać cztery kryteria określane akronimem FAIR[14]:

- *Findable*, czyli łatwo znajdowalne i wyszukiwalne,
- *Accessible*, czyli dostępne,
- *Interoperable*, czyli interoperacyjne,
- *Re-usable*, czyli możliwe do ponownego wykorzystania.

Efektywne poszukiwanie repozytorium danych badawczych może być wykonane w zasobach *Registry of Research Data Repositories* https://www.re3data.org[15]. Oprócz repozytoriów ogólnodostępnych, np.:

- *ZENODO* https://zenodo.org,
- *figshare* https://figshare.com,
- *RepOD* Repozytorium Otwartych Danych https://repod.pon.edu.pl,

znajdziemy także wybór repozytoriów dziedzinowych, np.:

- *Crystallography Open Database* http://www.crystallography.net/cod,
- *DRYAD* https://datadryad.org,
- *CancerData.org* https://www.cancerdata.org

oraz liczne repozytoria instytucjonalne, np.:

- Uniwersytetu w Edynburgu – *DataShare* https://datashare.is.ed.ac.uk,
- Uniwersytetu Oksfordzkiego, Bodleian Libraries – *DataBank* https://databank.ora.ox.ac.uk,
- Uniwersytetu w Cambridge https://www.repository.cam.ac.uk.

Podsumowanie

Z punktu widzenia polskiego naukowca otwieranie danych badawczych jest trudniejsze – organizacyjnie i technicznie – niż publikowanie np. artykułów w otwartym dostępie. Który zbiór danych badawczych i dlaczego powinien być chroniony długoterminowo i czy w razie udostępnienia nie zostanie naruszona własność intelektualna? Jak duże przestrzenie dyskowe, które narzędzia do wersjonowania i konwersji formatów należy wybrać? Jakie standardy metadanych zwiększą widoczność danego zbioru danych badawczych w internecie? Te i tym podobne pytania, już wkrótce, będą musieli zadawać sobie prowadzący badania naukowcy m.in. na uczelniach, gdy zostaną zobligowani do znalezienia na nie odpowiedzi w celu stworzenia *Planu zarządzania danymi* wymaganego przez grantodawców.

Tworzenie polityk instytucjonalnych otwartego dostępu[16], zawierających również procedury gromadzenia, przetwarzania i udostępniania danych badawczych, ma z założenia należeć do uczelnianych pełnomocników do spraw otwartego dostępu (OD). Bibliotekarze mając kilkunastoletnie (licząc od ogłoszonej w 2002 r. Deklaracji Budapeszteńskiej) doświadcze-

[14] *The FAIR Data Principles* [online]. [Dostęp 22.09.2018]. Dostępny w: https://www.force11.org/group/fairgroup/fairprinciples.
[15] Wszystkie odwołania do stron internetowych zawierają dane aktualne w dniu 22.09.2018 r.
[16] *Kierunki rozwoju otwartego dostępu do publikacji i wyników badań naukowych w Polsce* [online]. 2015, s. 12. [Dostęp 26.09.2018]. Dostępny w: https://www.gov.pl/documents/1068557/1069061/20180413_Kierunki_rozwoju_OD_wersja_ostateczna.pdf/fc65e84c-8de0-3163-d1a4-e13a49fe1071.

nie w dziedzinie otwartego dostępu i budowy repozytoriów publikacji naukowych, mają wystarczająco wysokie kwalifikacje do objęcia powyższego stanowiska.

Artykuł powstał w ramach projektu: „Otwieramy naukę – udział Polski w międzynarodowych obchodach Open Access Week" – zadanie finansowane w ramach umowy 868/P-DUN/2018 ze środków Ministra Nauki i Szkolnictwa Wyższego przeznaczonych na działalność upowszechniającą naukę.

Bibliografia:

1. *Checklist for a Data Management Plan. v.4.0* [online]. Edinburgh: Digital Curation Centre, 2014. [Dostęp 22.09.2018]. Dostępny w: http://www.dcc.ac.uk/webfm_send/1279.
2. *CIRCULAR A-110 REVISED 11/19/93 As Further Amended 9/30/99* [online]. Office of Management and Budget, 30.09.1999. [Dostęp 22.09.2018]. Dostępny w: https://georgewbush-whitehouse.archives.gov/omb/circulars/a110/a110.html.
3. CISEK, S. *Zbiory danych badawczych online* [online]. 2014. [Dostęp 22.09.2018]. Dostępny w: https://www.slideshare.net/sabinacisek/cisek-zarzdzanie-inf-w-nauce-2014.
4. *The FAIR Data Principles* [online]. [Dostęp 22.09.2018]. Dostępny w: https://www.force11.org/group/fairgroup/fairprinciples.
5. *Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020* [online]. European Commission, Directorate-General for Research & Innovation, 21.03.2017. [Dostęp 22.09.2018]. Dostępny w: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.
6. *Kierunki rozwoju otwartego dostępu do publikacji i wyników badań naukowych w Polsce* [online]. 2015, s. 12. [Dostęp 26.09.2018]. Dostępny w: https://www.gov.pl/documents/1068557/1069061/20180413_Kierunki_rozwoju_OD_wersja_ostateczna.pdf/fc65e84c-8de0-3163-d1a4-e13a49fe1071.
7. NIELSEN, M. *Michael Nielsen* [online]. [Dostęp 22.09.2018]. Dostępny w: http://michaelnielsen.org/blog/michael-a-nielsen.
8. NIELSEN, M. Re: *Definitions of Open Science?* Message to Peter Murray-Rust. 28.07.2011. E-mail [online]. [Dostęp 22.09.2018]. Dostępny w: https://lists.okfn.org/pipermail/open-science/2011-July/000907.html.
9. *Open Innovation, Open Science, Open to the World – a vision for Europe* [online]. European Commission, Directorate-General for Research & Innovation, 2016. [Dostęp 22.09.2018]. Dostępny w: http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=16022.
10. *Otwarta nauka* [online]. [Dostęp 22.09.2018]. Dostępny w: https://pl.wikipedia.org/wiki/Otwarta_nauka.
11. PIWOWAR, H.A., VISION, T.J. Data reuse and the open data citation advantage Vision. *PeerJ* [online]. 2013, 1:e175. [Dostęp 22.09.2018]. Dostępny w doi: https://doi.org/10.7717/peerj.175.
12. Recommended conformant licenses. W: *Conformant Licenses* [online]. [Dostęp 22.09.2018]. Dostępny w: https://opendefinition.org/licenses.
13. STRZELCZYK, E. Otwarte dane badawcze – kolejny krok do otwierania nauki. W: SÓJKOWSKA, I., DERFERT-WOLF, L. (red.). *Bibliograficzne bazy danych: perspektywy i problemy rozwoju. III Konferencja Naukowa Konsorcjum BazTech, Kraków, 26–27 czerwca 2017* [online]. Stowarzyszenie EBIB, 2017. [Dostęp 20.09.2018]. Materiały Konferencyjne EBIB, nr 25. ISBN 978-83-63458-08-9. Dostępny w: http://open.ebib.pl/ojs/index.php/Mat_konf/article/view/599.