Zarządzanie danymi badawczymi

Poradnik dla naukowców
i data stewardów

Wersja 1.0

Opracowanie: Wojciech Fenrich, Natalia Gruenpeter, Krzysztof Siewicz, Jakub Szprot

Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego

Poradnik opracowany w ramach kursów z zarządzania danymi badawczymi udostępnionych na platformie edukacyjnej Navoica (navoica.pl).

Ministerstwo Edukacji i Nauki

NARODOWE CENTRUM NAUKI

Zadanie realizowane przez Narodowe Centrum Nauki na podstawie zlecenia Ministra Edukacji i Nauki dot. krajowej koordynacji partnerstwa European Open Science Cloud w latach 2022-2023.

Spis treści

Wprowadzenie do otwartej nauki

Otwarta nauka ma swoje początki w ruchu na rzecz otwartego dostępu do publikacji naukowych (Open Access), którego celem była zmiana modelu komunikacji naukowej. Komunikacja naukowa obejmuje działania prowadzące od utrwalenia wyników badań naukowych przez ich weryfikację aż do publikacji, rozpowszechnienia i wykorzystania. Komunikacja naukowa przebiega w określonym kontekście historycznym i kulturowym oraz z wykorzystaniem właściwych dla niego technik i środków dystrybucji, np. listów, publikacji

w periodykach naukowych, konferencji naukowych, forów i list dyskusyjnych czy repozytoriów.

Jednym z najważniejszych środków komunikacji naukowej są recenzowane publikacje, w szczególności artykuły w czasopismach, a w niektórych dziedzinach także publikacje książkowe. W tradycyjnym modelu wydawniczym naukowcy zainteresowani publikacją w określonym czasopiśmie przenosili majątkowe prawa autorskie na wydawcę, który następnie mógł czerpać z publikacji korzyści finansowe i ograniczać możliwość jej rozpowszechniania lub ponownego wykorzystywania. Zapoznanie się z treścią artykułu wymagało opłacenia dostępu, indywidualnie (przez zakup papierowego egzemplarza czasopisma lub wykupienie dostępu do wydania elektronicznego) bądź poprzez instytucję naukową. W modelu subskrypcyjnym dostęp ten opłacany jest przez biblioteki uniwersyteckie i inne instytucje naukowe, czasami zrzeszone w konsorcja, które wspólnie negocjują warunki i ceny.

Wykorzystanie w nauce elektronicznych środków komunikacji pozwoliło na szybką dystrybucję treści i ich bezpłatne powielanie, jednak dominujące modele wydawnicze przez długi czas utrwalały finansowe, techniczne i prawne bariery w dostępie do publikacji. Z czasem dzięki ruchowi działającemu na rzecz otwartego dostępu wypracowano rozwiązania pozwalające na znoszenie powyższych barier.

Od początku XXI wieku środowiska naukowe działające na rzecz otwartego dostępu do publikacji naukowych w coraz większym

stopniu uzgadniały swoje postulaty. Jednym z istotnych forów dyskusji na temat kształtu komunikacji naukowej stała się konferencja Budapest Open Access Initiative zorganizowana w 2001 r., której rezultatem była podpisana w lutym 2002 r. deklaracja Budapest Open Access Initiative. Przyjęto w niej wspólną definicję otwartego dostępu:

> Przez „otwarty dostęp" rozumiemy dostępność treści za darmo i w publicznym Internecie, co pozwala każdemu czytać, ściągać, kopiować, rozprowadzać, drukować, przeszukiwać, zamieszczać odnośniki do pełnych wersji tekstów, indeksować, przekazywać jako dane do oprogramowania oraz używać w dowolnym innym, zgodnym z prawem celu – bez barier finansowych, legalnych czy technicznych, innych niż te związane z uzyskaniem dostępu do samego Internetu. Jedynym ograniczeniem kopiowania i dystrybucji treści oraz jedyną rolą, jaką w tym obszarze odgrywa prawo autorskie, powinno być zapewnienie autorom kontroli nad integralnością ich utworów oraz prawa do odpowiedniego uznania ich autorstwa i cytowania ich prac[1].

Podobne rozumienie otwartego dostępu przyjęto w Deklaracji z Bethesdy (Bethesda Statement on Open Access Publishing, 2003) oraz Deklaracji Berlińskiej w sprawie otwartego dostępu

[1] Tarkowski A., Bednarek-Michalska B., Siewicz K. i in., *Przewodnik po otwartej nauce*, Warszawa 2009, s. 83, https://depot.ceon.pl/bitstream/handle/123456789/65/przewodnik-po-otwartej-nauce.pdf [data dostępu: 17.07.2023].

do wiedzy w naukach ścisłych i humanistyce (Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, 2003)[2]. Zarówno budapeszteńska, jak i berlińska inicjatywa zaowocowały dalszymi działaniami na rzecz otwartej nauki.

Od otwartego dostępu do otwartej nauki

Z czasem postulaty dotyczące otwartego systemu komunikacji naukowej zaczęły wykraczać poza publikacje naukowe i obejmować inne rezultaty badań (w tym dane badawcze), a także procesy, procedury i metody prowadzenia badań czy materiały edukacyjne. W coraz większym stopniu praktyki otwartej nauki dotyczą zatem sposobu prowadzenia badań, a nie tylko sposobu upowszechniania wyników. W ślad za tymi zmianami podążają inicjatywy na rzecz reformy systemu ewaluacji działalności naukowej, których celem jest odejście od niewłaściwego wykorzystania wskaźników dot. czasopism i publikacji, takich jak Journal Impact Factor (JIF) i indeks Hirscha. Wiele inicjatyw mających na celu reformę systemów ewaluacji nauki podkreśla znaczenie otwartej nauki i postuluje wpisanie jej praktyk, w tym otwartego udostępniania danych, w kryteria oceny pracy naukowców lub kryteria oceny wniosków grantowych. Ważnym elementem ruchu na rzecz otwartej nauki są ponadto postulaty angażowania obywateli w procesy naukowe, np. w projekty nauki obywatelskiej.

Obecnie otwarta nauka staje się powszechnie akceptowanym i rekomendowanym modelem komunikacji naukowej, a wiele

[2] Por. Tamże.

międzynarodowych inicjatyw dąży do ujednolicenia polityk otwartości oraz promowania wspólnych standardów ułatwiających współpracę i dzielenie się wynikami badań.

Rekomendacja UNESCO w sprawie otwartej nauki

W 2021 r. UNESCO przyjęło rekomendację dot. otwartej nauki, która wyznacza międzynarodowe zasady, wartości i cele w zakresie dostępności rezultatów badań. Zaakceptowany przez 193 kraje dokument powstał w wyniku prac komitetu doradczego UNESCO ds. otwartej nauki, który rozpoczął pracę w 2020 r., oraz szerokich konsultacji z różnymi grupami interesariuszy.

Otwarta nauka została w rekomendacji UNESCO określona jako „nowy paradygmat", który, opierając się na podstawowych zasadach wolności akademickiej, integralności badawczej i doskonałości naukowej, włącza w działalność badawczą praktyki dotyczące replikowalności i przejrzystości, a także dzielenia się wiedzą i współpracy.

> Otwarta nauka to koncepcja łącząca różne ruchy i praktyki mające na celu udostępnianie wiedzy naukowej w sposób otwarty i możliwy do ponownego wykorzystania dla wszystkich, zacieśnienie współpracy naukowej i dzielenie się informacjami z korzyścią dla nauki i społeczeństwa, a także otwarcie procesów wytwarzania i oceniania wiedzy naukowej czy

komunikację z podmiotami społecznymi spoza tradycyjnie rozumianej społeczności naukowej[3].

Obejmuje ona wszystkie dyscypliny naukowe i aspekty praktyk naukowych, w tym nauki podstawowe i stosowane, nauki przyrodnicze i społeczne oraz nauki humanistyczne, i opiera się na następujących filarach.

Otwarta wiedza naukowa

Obszar wskazany jako pierwszy obejmuje otwarty dostęp do publikacji naukowych, otwarte dane badawcze, otwarte zasoby edukacyjne, otwarte oprogramowanie i kod źródłowy oraz otwarty sprzęt.

Infrastruktury otwartej nauki

W zakres tego obszaru wchodzą wspólne infrastruktury badawcze służące potrzebom różnych społeczności. Chodzi tutaj zarówno o infrastruktury wirtualne, jak i fizyczne, w tym sprzęt potrzebny do prowadzenia badań, laboratoria i ich wyposażenie, zasoby obliczeniowe, platformy publikacyjne, repozytoria, archiwa, inne systemy gromadzące zasoby naukowe.

Otwarte angażowanie aktorów społecznych

Angażowanie obywateli w badania może obejmować takie działania jak rozwijanie modeli finansowania typu *crowdfunding*

[3] Rekomendacja UNESCO ws. otwartej nauki, https://www.unesco.org/en/open-science. Por. Open Science Toolkit, https://www.unesco.org/en/open-science/toolkit [data dostępu: 17.07.2023].

i nowych sposobów pozyskiwania informacji, wiedzy, treści czy pomysłów opartych na crowdsourcingu, promowanie wolontariatu czy nauki obywatelskiej. U podstaw wszystkich tych działań leży współpraca pomiędzy środowiskiem naukowym a obywatelami, uwzględniająca otwieranie procesu badawczego i zwiększanie inkluzywności nauki.

Otwarty dialog z innymi systemami wiedzy

Dialog w tym zakresie powinien uwzględniać wiedzę ze źródeł marginalizowanych oraz wzmacniać wzajemne relacje i komplementarność pomiędzy różnymi epistemologiami. U podstaw leży poszanowanie praw człowieka, przy zachowaniu praw wszystkich ludzi do sprawiedliwego udziału w korzyściach płynących z wykorzystania wiedzy. Istotnymi punktami odniesienia są tutaj przyjęta przez UNESCO Deklaracja o Różnorodności Kulturowej z 2001 r. oraz Deklaracja ONZ o Prawach Ludności Rdzennej z 2007 r.

Otwarte dane badawcze

Otwarte dane badawcze zdefiniować można jako dane dostępne za pośrednictwem Internetu, które można wykorzystywać bez ponoszenia opłat oraz bez istotnych ograniczeń technicznych i prawnych.

Definicja otwartych danych badawczych przyjęta w rekomendacji UNESCO w sprawie otwartej nauki podkreśla znaczenie standardowych rozwiązań, które ułatwiają wymianę, odczyt i ponowne wykorzystanie danych.

> Otwarte dane badawcze obejmują m.in. dane cyfrowe i analogowe, zarówno surowe, jak i przetworzone, oraz towarzyszące im metadane, jak również zapisy liczbowe, zapisy tekstowe, obrazy i dźwięki, protokoły, kod analizy i procedury, które mogą być wykorzystywane w sposób otwarty, ponownie wykorzystywane, zachowywane i redystrybuowane przez kogokolwiek, pod warunkiem uznania autorstwa. Otwarte dane badawcze są dostępne w aktualnym, przyjaznym dla użytkownika, czytelnym dla ludzi i maszyn oraz umożliwiającym dalsze działania (*actionable*) formacie, zgodnie z zasadami dobrego zarządzania i opieki nad danymi, w szczególności zasadami FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*) oraz podlegają regularnym działaniom w zakresie ich kuratorowania i utrzymywania[4].

Warto zauważyć, że definicja ta obejmuje nie tylko samą otwartość, która w potocznym rozumieniu może być sprowadzona do udostępnienia danych w Internecie bez konieczności wnoszenia opłat za dostęp przez korzystających z danych czy podejmowania innych działań, np. związanych z uwierzytelnieniem czy autoryzacją. W ujęciu UNESCO bardzo istotne znaczenie mają standardowe rozwiązania pozwalające na swobodne wykorzystywanie danych. W praktyce oznacza to znoszenie finansowych, prawnych i technicznych barier, przy czym w wypadku danych szczególne znaczenie mają kwestie

[4] Tamże.

techniczne, a także aktywne kuratorowanie zasobów danych i ich właściwe przechowywanie.

Właściwe zarządzanie danymi badawczymi i ich otwarte udostępnianie jest elementem polityk otwartości wielu instytucji, w tym Komisji Europejskiej w ramach programu Horyzont Europa czy Narodowego Centrum Nauki. U podstaw decyzji o przyjęciu wymogu udostępniania danych leży dostrzeżenie wartości otwartych danych badawczych oraz ich naukowego, gospodarczego i społecznego znaczenia.

Udostępnianie danych badawczych wiąże się z szeregiem korzyści dla naukowców, instytucji naukowych oraz całego społeczeństwa.

1. Możliwość weryfikacji wyników badań

System komunikacji naukowej w obecnym kształcie oparty jest głównie na publikacjach, w których prezentowane są wyniki badań naukowych wraz z opisem sposobu ich uzyskania. Otwarte udostępnianie danych stanowiących podstawę publikacji (tzw. *underlying data*) umożliwia innym weryfikację wniosków, a także daje wgląd w szczegóły procesu badawczego. To z kolei daje sposobność wykrycia błędów, nieścisłości czy przypadków nierzetelności naukowej, zarówno na etapie recenzji prowadzonej przez redakcję czasopisma naukowego, do którego zgłoszony został manuskrypt, jak i później, kiedy z artykułem zapoznają się czytelnicy.

2. Wyższa jakość danych i transparentność procesów badawczych

Odpowiedzialne zarządzanie danymi wiąże się z prowadzeniem badań w sposób zaplanowany, staranny i powtarzalny. Zarówno dobre praktyki w zakresie zarządzania danymi, jak i plany związane z ich udostępnianiem mogą wpłynąć na większy rygor w przygotowywaniu, opisywaniu i dokumentowaniu danych badawczych, co rzutuje na ich jakość. Otwarte dane traktować można ponadto jako „wizytówkę" autorów lub zespołu badawczego.

3. Zwiększenie efektywności badań naukowych

Otwarte udostępnianie danych zgodnie z dobrymi praktykami, zasadami FAIR i właściwymi standardami pozwala na uniknięcie powielania badań, prowadząc w rezultacie do oszczędzania czasu, wysiłku i środków finansowych. Zasoby te mogą zostać przeznaczone na inne prace, przyczyniające się do rozwoju nauki, związane np. z formułowaniem nowych pytań badawczych czy interpretacją otrzymanych wyników.

4. Zwiększenie widoczności badań i promocja dorobku

Dostępność danych wpływa na zwiększenie widoczności i oddziaływania dorobku naukowego badaczy i instytucji naukowych. Dane łatwe do znalezienia i dostępne są częściej cytowane i ponownie wykorzystywane w dalszych badaniach.

5. Wspieranie współpracy naukowej

Otwarte udostępnianie danych może również stymulować współpracę między naukowcami, przyczyniając się do szybszej wymiany informacji i rozwoju nauki. Dane mogą zostać ponownie przeanalizowane i wykorzystane przez badaczy z innych dziedzin, a także różnych instytucji czy krajów.

6. Wykorzystanie otwartych danych w dydaktyce akademickiej

Otwarte dane mogą być wykorzystywane zarówno w celach naukowych, jak i dydaktycznych. Dostęp do nich stwarza ponadto możliwość rozwijania otwartych zasobów edukacyjnych, które mogą być ponownie wykorzystywane i wzbogacane przez wykładowców.

7. Wykorzystanie otwartych danych w projektach nauki obywatelskiej

Angażowanie obywateli w proces zbierania, opisywania czy analizowania danych udostępnianych w sposób otwarty może przyczynić się do nowych odkryć naukowych i szybszego opracowania dużych zbiorów danych. Inicjatywy z zakresu nauki obywatelskiej zwiększają zaufanie do nauki poprzez pokazanie procesu badawczego oraz upodmiotowienie obywateli w tym procesie.

8. Wspieranie współpracy z podmiotami spoza środowiska akademickiego

Dane udostępnione w sposób otwarty na wolnych licencjach sprzyjają rozwijaniu współpracy środowiska akademickiego z innymi podmiotami z otoczenia społeczno-gospodarczego.

9. Znaczenie otwartych danych dla społeczeństwa

Otwarte udostępnianie danych jest korzystne nie tylko dla badaczy i instytucji naukowych. Mogą one służyć całemu społeczeństwu i różnym grupom zawodowym, w zależności od dziedziny. Z otwartych danych w różnych kontekstach korzystać mogą:

- organy administracji publicznej,
- organizacje pozarządowe,
- nauczyciele i edukatorzy,
- specjaliści z różnych dziedzin,
- dziennikarze i fact-checkerzy,
- przedsiębiorcy,
- wszyscy obywatele.

W ogólnym ujęciu otwarte dane badawcze mogą służyć dwóm celom: rozwiązywaniu problemów społecznych i gospodarczych oraz zwiększaniu zaufania do nauki. Założenia te wpisać można w ogólne zasady i wymagania obowiązujące naukowców, w szczególności dotyczące odpowiedzialności czy upowszechniania i wykorzystywania wyników. Zostały one zapisane w Europejskiej Karcie Naukowca:

„Naukowcy powinni być świadomi tego, że są odpowiedzialni wobec swoich pracodawców, grantodawców i innych organów publicznych lub prywatnych, a także, z przyczyn etycznych, wobec ogółu społeczeństwa. W szczególności, naukowcy, których badania finansowane są z funduszy państwowych, są również odpowiedzialni za efektywne wykorzystanie pieniędzy podatników”[5].

„Zgodnie z ustaleniami zawartymi w ich umowach wszyscy naukowcy powinni zapewnić, by wyniki ich badań były rozpowszechniane i wykorzystywane, np. ogłaszane, przekazywane innym środowiskom naukowym lub, w stosownych przypadkach, skomercjalizowane. W szczególności od starszych pracowników naukowych oczekuje się przejęcia inicjatywy w zapewnieniu, by badania naukowe były owocne, zaś ich wyniki wykorzystywane komercyjnie i/lub udostępniane ogółowi społeczeństwa przy każdej nadarzającej się sposobności”[6].

[5] Europejska Karta Naukowca, https://www.uw.edu.pl/wp-content/uploads/2016/06/europejska-karta-naukowca-i-kodeksu-postepowania-przy-rekrutacji-pracownikow-naukowych.pdf [data dostępu: 17.07.2023].

[6] Tamże.

Polityki otwartości

Ważnymi narzędziami kształtowania otwartej komunikacji naukowej są poliyki otwartości przyjmowane przez instytucje finansujące i prowadzące badania naukowe, a także kraje. Polityka otwartości to zbiór zasad określających sposób udostępniania rezultatów badań. Może zostać przyjęta na szczeblu krajowym lub instytucjonalnym oraz obejmować różne zakresy zagadnień związanych z otwartą nauką, w szczególności otwarty dostęp do publikacji naukowych oraz wymogi związane z zarządzaniem danymi badawczymi i ich otwartym udostępnianiem.

Polityki otwartości mogą różnić się pod względem stopnia rygoru – w ścisłym, węższym znaczeniu polityka określa obowiązki

nałożone na badaczy przez grantodawców lub pracodawców; w szerszym znaczeniu politykę otwartości można rozumieć także jako zbiór wytycznych i rekomendacji.

Komisja Europejska

Wdrażanie polityki otwartości Komisji Europejskiej rozpoczęło się od pilotażowego programu otwartego dostępu do recenzowanych publikacji naukowych (artykułów naukowych), który następnie został przyjęty jako obowiązkowy. Kolejnym krokiem było objęcie polityką danych badawczych – zarówno wymogów związanych z tworzeniem planu zarządzania danymi, jak i wymogów otwartego udostępniania danych. Aktualnie w programie Horyzont Europa przyjęto kompleksową politykę otwartości, obejmującą publikacje naukowe, dane badawcze, innego typu rezultaty badań, a także działania na rzecz nauki obywatelskiej czy innych form angażowania w projekty osób i grup spoza środowiska naukowego.

Publikacje naukowe

Recenzowane publikacje naukowe muszą zostać udostępnione w sposób otwarty w chwili publikacji, zgodnie z poniższymi zasadami:

- elektroniczna kopia opublikowanej wersji nadająca się do odczytu maszynowego lub ostateczny recenzowany manuskrypt zaakceptowany do publikacji powinny zostać zdeponowane w zaufanym repozytorium publikacji naukowych;

- publikacja powinna zostać udostępniona z wykorzystaniem wolnych licencji pozwalających na ponowne wykorzystanie, domyślnie na licencji CC BY; dla monografii i innych długich tekstów dopuszczalne są licencje ograniczające komercyjne wykorzystanie i tworzenie utworów zależnych: CC BY-NC, CC BY-ND, CC BY-NC-ND;
- za pośrednictwem repozytorium należy podać informacje o wynikach badań lub innych narzędziach i instrumentach potrzebnych do weryfikacji wniosków zawartych w publikacji naukowej.

Metadane publikacji muszą być udostępnione na podstawie oświadczenia o przekazaniu do domeny publicznej, czyli Creative Commons Public Domain Dedication (CC0) lub równoważnej, zgodne z zasadami FAIR. Metadane muszą zawierać informacje na temat:

- publikacji: autor (autorzy), tytuł, data publikacji, miejsce publikacji;
- źródła finansowania: nazwa projektu grantowego, akronim i numer;
- warunków licencji;
- trwałych identyfikatorów: publikacji, autora (autorów) oraz, jeżeli to możliwe, organizacji.

W stosownych wypadkach metadane muszą zawierać trwałe identyfikatory innych wyników badań lub wszelkich innych narzędzi oraz instrumentów potrzebnych do weryfikacji wniosków z publikacji.

Dane badawcze

W ramach programu Horyzont Europa obowiązkowe jest odpowiedzialne zarządzanie danymi zgodnie z zasadami FAIR, co przekłada się na konieczność stworzenia planu zarządzania danymi badawczymi i zaplanowania otwartego udostępniania danych według ogólnej zasady *as open as possible, as closed as necessary*.

Wymogi w zakresie planu zarządzania danymi badawczymi:

- sporządzenie planu we wszystkich projektach, w których wytwarzane bądź ponownie wykorzystywane są dane;
- złożenie planu zarządzania danymi badawczymi w terminie wskazanym w umowie grantowej (zwykle w ciągu sześciu miesięcy od rozpoczęcia realizacji projektu);
- aktualizowanie planu zarządzania danymi, kiedy wystąpią istotne zmiany oraz w połowie realizacji projektu (jeżeli trwa on dłużej niż 12 miesięcy) i przed końcową oceną.

Wymogi w zakresie udostępniania danych badawczych:

- zapewnienie otwartego dostępu zgodnie z zasadą *as open as possible, as closed as necessary*;
- deponowanie danych w zaufanym repozytorium tak szybko, jak to możliwe, bądź w terminach określonych w planie zarządzania danymi;
- deponowanie danych powiązanych z publikacjami naukowymi najpóźniej do dnia publikacji i w sposób zgodny ze standardami przyjętymi przez społeczność naukową;

- udostępnienie danych badawczych z wykorzystaniem licencji CC BY lub CC0 bądź równoważnych;
- zamieszczenie w repozytorium informacji o innych rezultatach lub narzędziach i instrumentach potrzebnych do weryfikacji bądź ponownego wykorzystania danych;
- opracowanie metadanych zgodnie z zasadami FAIR i udostępnianie ich z wykorzystaniem licencji CC0 bądź równoważnej.

W niektórych konkursach w programie Horyzont Europa uwzględniono ponadto sytuacje zagrożenia publicznego, w których na żądanie instytucji przyznającej środki należy zapewnić natychmiastowy otwarty dostęp do wszystkich wyników badań na otwartych licencjach lub – jeżeli zapewnienie takiego dostępu byłoby sprzeczne z uzasadnionymi interesami beneficjentów – udzielić niewyłącznych licencji osobom prawnym, które potrzebują wyników badań, aby zaradzić sytuacji kryzysowej.

W niektórych programach bądź konkursach wskazane mogą być dodatkowe wymogi. Uzupełniające informacje znaleźć można w dokumentach udostępnionych przez Komisję Europejską[7].

[7] Por. Annotated Model Grant Agreement, https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/aga_en.pdf, Horizon Europe Programme Guide, https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf [data dostępu: 17.07.2023].

Narodowe Centrum Nauki

Polityka otwartości Narodowego Centrum Nauki obejmuje recenzowane artykuły naukowe oraz dane badawcze[8].

Publikacje naukowe

W zakresie publikacji punktem odniesienia jest Plan S. Polityka dotyczy głównie artykułów w recenzowanych czasopismach, ale także np. recenzowanych materiałów konferencyjnych. Nie obejmuje natomiast jeszcze monografii, rozdziałów monografii i recenzowanych utworów zbiorowych.

NCN definiuje otwarty dostęp do publikacji nie tylko jako darmowy dostęp do publikacji w Internecie, ale także możliwość powielania, rozpowszechniania i dowolnego wykorzystania treści przez czytelnika, zgodnie z warunkami licencji praw autorskich CC-BY.

Ścieżki publikacyjne

1. W czasopismach lub na platformach otwartego dostępu zarejestrowanych lub będących na etapie rejestracji w Directory of Open Access Journal (DOAJ).

---

[8] Polityka Narodowego Centrum Nauki dot. otwartego dostępu do publikacji, https://www.ncn.gov.pl/sites/default/files/pliki/zarzadzenia-dyrektora/zarzadzenieDyr-38_2020.pdf. Por. Instrukcja do Polityki NCN dot. otwartego dostępu do publikacji, https://www.ncn.gov.pl/sites/default/files/pliki/2021_10_instrukcja_open_access_NCN.pdf [data dostępu: 17.07.2023].

2. W czasopismach subskrypcyjnych (hybrydowych) pod warunkiem, że Version of Record (VoR) lub Author Accepted Manuscript (AAM) zostanie bezpośrednio przez wydawcę lub autora opublikowane w otwartym repozytorium w momencie ukazania się publikacji on-line (bez embarga czasowego). Repozytorium musi być zarejestrowane w Open Directory of Open Access Repositories (OpenDOAR), a VoR lub AAM posiadać unikalny stały identyfikator (np. DOI, URN, UUID, Handle lub inne). Jeśli wersja zdeponowana w repozytorium i wersja opublikowana są osobnymi wersjami (AAM i VoR), powinny mieć osobne identyfikatory.

3. W czasopismach objętych licencją otwartego dostępu w ramach tzw. umów transformacyjnych, które muszą być zarejestrowane w rejestrze prowadzonym przez Efficiency and Standards for Article Charges (ESAC Registry). Przykładem krajowych umów transformacyjnych są programy pilotażowe podpisane z wydawcami przez Wirtualną Bibliotekę Naukową.

W czasopismach transformacyjnych, tzw. *transformative journals* (TJ). Czasopisma transformacyjne muszą spełniać kryteria, które znajdują się w wytycznych dot. wdrożenia Planu S i umożliwić autorom publikowanie ich oryginalnych artykułów naukowych w otwartym dostępie. Ta ścieżka publikacyjna (3) obowiązuje tylko, gdy praca została przyjęta do druku lub opublikowana do grudnia 2024 r.

Koszty

Koszty związane z procesem publikacyjnym, tzw. Article Processing Charges (APC), są kwalifikowalne w przypadku ścieżek pierwszej i trzeciej. Opłaty za publikacje w ścieżce drugiej (czasopisma hybrydowe) są kosztami niekwalifikowalnymi dla projektu i nie mogą pochodzić ze środków NCN.

Prawa autorskie i licencje

Publikacje, które zawierają wyniki będące efektem realizacji projektu, muszą być udostępnione (również te w wersji AAM lub VoR) na licencji Creative Commons Uznanie autorstwa, CC BY. Narodowe Centrum Nauki zaleca domyślne korzystanie z licencji Creative Commons Attribution CC BY 4.0, dopuszcza jednak publikowanie w czasopismach, w ramach umów transformacyjnych, na licencji CC-BY-SA, a w przypadkach uzgodnionych z NCN na licencji CC-BY-ND.

Dane badawcze

Narodowe Centrum Nauki wprowadziło wymóg sporządzania planów zarządzania danymi badawczymi w 2019 r.

Wymogi w zakresie planu zarządzania danymi badawczymi:

- na etapie składania wniosku wymagane jest sporządzenie planu zarządzania danymi badawczymi, który podlega ocenie opisowej;
- plan zarządzania danymi jest bezpośrednio związany z planem badawczym, w trakcie realizacji projektu może

podlegać zmianom, nie ma obowiązku informowania Narodowego Centrum Nauki o zmianach i aktualizacjach;
- na etapie oceny raportu końcowego plan zarządzania danymi podlega eksperckiej ocenie merytorycznej; jeśli plan będzie niekompletny, zostanie odesłany do uzupełnienia lub korekty.

Wymogi w zakresie udostępniania danych badawczych:
- dane powiązane (podstawowy zestaw danych) z opublikowanymi artykułami powinny być udostępniane w otwartym repozytorium;
- tam gdzie to możliwe, dane badawcze powinny być udostępniane zgodnie z warunkami oświadczenia o przekazaniu do domeny publicznej Creative Commons Public Domain Dedication (CC0);
- dane powinny być udostępniane zgodnie ze standardami cytowania danych zawartych w Declaration of Data Citation Principles by FORCE 11 oraz na zasadach zawartych w TOP Guidelines;
- wszystkie publikowane metadane muszą spełniać wytyczne podane przez OpenAIRE i zawierać adnotację o finansowaniu ze środków projektu (Narodowe Centrum Nauki, numer projektu);
- w pozostałym zakresie o udostępnieniu decyduje badacz, mając na uwadze ograniczenia formalnoprawne lub inne wskazane w uzasadnieniu.

Zasady FAIR

Akronim FAIR określa wymogi, jakie powinny spełniać dane badawcze. Zasady te zostały opisane w artykule „FAIR Guiding Principles for scientific data management and stewardship”[9] i stanowią obecnie ważny punkt odniesienia, przywoływany także w politykach otwartości.

Findable – możliwe do znalezienia

F1. (meta)dane mają przypisany trwały i unikalny identyfikator
F2. dane są opisane za pomocą bogatych metadanych (opisanych w R1)
F3. metadane w oczywisty sposób zawierają identyfikator danych, które opisują
F4. (meta)dane są zarejestrowane lub zaindeksowane w miejscu, którego zasoby można przeszukiwać

Accessible – dostępne

A1. (meta)dane są dostępne z wykorzystaniem standardowego protokołu komunikacyjnego
A1.1 protokół ten jest bezpłatny, otwarty i może być uniwersalnie implementowany
A1.2 protokół ten pozwala w razie konieczności na uwierzytelnienie i autoryzację
A2. metadane są dostępne, nawet jeśli dane już nie są dostępne

[9] Por. M. Wilkinson, M. Dumontier, I. Aalbersberg i in., *The FAIR Guiding Principles for scientific data management and stewardship*, „Scientific Data” 3, 2016, https://doi.org/10.1038/sdata.2016.18 [data dostępu: 17.07.2023].

Interoperable – interoperacyjne

I1. (meta)dane wykorzystują formalne, dostępne, wspólne i szeroko stosowane języki reprezentacji wiedzy
I2. (meta)dane wykorzystują słowniki zgodne z zasadami FAIR
I3. (meta)dane zawierają odpowiednie odwołania do innych (meta)danych

Reusable – możliwe do ponownego wykorzystania

R1. (meta)dane są bogato opisane za pomocą adekwatnych i istotnych atrybutów
R1.1. (meta)dane są udostępniane z określoną i dostępną licencją na wykorzystanie
R1.2. (meta)dane są związane ze szczegółową dokumentacją pochodzenia (*provenance*)
R1.3. (meta)dane są zgodne z dziedzinowymi standardami odpowiedniej społeczności naukowej

Najważniejszym celem wdrażania zasad FAIR jest zwiększenie ponownego wykorzystania danych. Wytwarzanie i zbieranie danych badawczych jest kosztowne i czasochłonne, a istniejące dane mogą posłużyć jako podstawa do dalszych badań czy innowacji.

Istotnym kontekstem zasad FAIR jest możliwość maszynowego odczytu i przetwarzania danych i metadanych (czyli danych opisujących właściwe dane). W praktyce oznacza to, że należy zadbać o taki opis danych i takie miejsca ich przechowywania, aby wyszukiwanie i ponowne wykorzystywanie danych możliwe

było przy zerowej lub minimalnej interwencji człowieka. Szczegółowe wytyczne do zasad FAIR skupiają się właśnie na technicznych aspektach tych procesów i uwzględniają cechy protokołów wymiany danych czy systemów informatycznych służących do przechowywania i udostępniania treści naukowych.

Zasady FAIR można jednak ujmować także jako zbiór wytycznych dla badaczy chcących we właściwy sposób zarządzać danymi. Stanowią one ponadto istotne elementy polityk otwartości instytucji finansujących badania naukowe, m.in. Komisji Europejskiej, która przygotowała wzór planu zarządzania danymi oparty właśnie na zasadach FAIR.

Findable – możliwe do znalezienia

Dane są możliwe do znalezienia, kiedy są dobrze opisane, posiadają trwały identyfikator i zamieszczone zostały w serwisach, których zasoby można przeszukiwać.

Co zrobić, aby ułatwić innym znalezienie danych?

- Opatrzyć je trwałym identyfikatorem, najlepiej DOI (Digital Object Identifier).
- Opisać je metadanymi, najlepiej zgodnymi ze standardami konkretnej dziedziny i dyscypliny.
- Zdeponować je w odpowiednim serwisie, najlepiej w repozytorium, którego zasoby można przeszukiwać.

Accessible – dostępne

Dane są dostępne, kiedy ludzie i maszyny mogą uzyskać do nich dostęp na jasnych zasadach i za pomocą standardowych, otwartych protokołów, bądź otrzymać informacje na ich temat w sytuacji, w której same dane nie mogły zostać udostępnione w sposób otwarty. W praktyce oznacza to, że dane nie muszą być otwarte, aby być FAIR.

Czasami dane nie mogą zostać udostępnione, np. z uwagi na ochronę danych osobowych, kwestie bezpieczeństwa czy kwestie etyczne. Aby zadbać o zgodność z zasadami FAIR, należy zapewnić dostęp do metadanych i, jeżeli to możliwe, wskazać szczególne warunki dostępu do danych, które mogą np. uwzględniać konieczność uzyskania specjalnej zgody bądź wykazania odpowiedniej afiliacji.

Co zrobić, aby zapewnić dostępność danych?

- Udostępnić metadane, nawet jeżeli nie można udostępnić samych danych.
- Wskazać warunki dostępu do danych, jeżeli nie są udostępnione w sposób otwarty, np. poprzez zapewnienie możliwości uwierzytelnienia lub autoryzacji.
- Korzystać z szeroko rozpowszechnionych i standardowych rozwiązań technicznych w udostępnianiu danych.

Interoperable – interoperacyjne

Interoperacyjność (meta)danych to możliwość łączenia ich z innymi (meta)danymi, wykorzystywania w wielu różnych systemach komputerowych i analizowania przy użyciu różnorodnego oprogramowania. Dla zapewnienia interoperacyjności kluczowe znaczenie ma wykorzystanie standardów, np. standardowych i otwartych formatów plików czy standardów metadanych.

Co zrobić, aby zapewnić interoperacyjność danych?

- Udostępnić dane w standardowym formacie, najlepiej otwartym.
- Skorzystać z odpowiednich standardów metadanych, np. z właściwych słowników kontrolowanych.
- Wskazać powiązania z innymi danymi bądź publikacjami.

Reusable – możliwe do ponownego wykorzystania

Możliwość ponownego wykorzystania danych to najważniejszy cel zasad FAIR. Osiągnięcie go jest możliwe, kiedy dane są dobrze i rzetelnie opisane, w szczególności posiadają dokumentację pozwalającą określić kto, kiedy i w jaki sposób je wytworzył lub zebrał. Pozwala to ocenić wiarygodność i rzetelność danych, a także adekwatność ich ponownego wykorzystania w odniesieniu do zakładanych celów. Innym aspektem jest zapewnienie możliwości ponownego wykorzystania od strony prawnej poprzez korzystanie z wolnych licencji.

Co zrobić, aby zapewnić możliwość ponownego wykorzystania danych?

- Przygotować i udostępnić wraz z danymi odpowiednią dokumentację.
- Korzystać ze standardowych wolnych licencji, które mają postać możliwą do odczytu maszynowego, np. z licencji Creative Commons.

Niektóre z opisanych wyżej działań związane są z wyborem odpowiedniego miejsca przechowywania i udostępnienia danych, najlepiej przystosowanego do tego repozytorium. Inne działania w pełni zależą od autorów i osób przygotowujących dane do udostępnienia.

Narzędzia wspierające udostępnianie danych zgodnie z zasadami FAIR

Urzeczywistnienie zasad FAIR, w szczególności poprzez stworzenie otwartego ekosystemu zgodnych z zasadami FAIR obiektów cyfrowych i serwisów, wymaga zarówno inwestycji w rozwój odpowiedniej infrastruktury, jak i działań szkoleniowych, informacyjnych i promocyjnych. Jednym z istotnych elementów jest tworzenie narzędzi wspierających udostępnianie danych zgodnie z zasadami FAIR. Są to zarówno narzędzia samooceny (*self-assessment*) zaprojektowane tak, aby wspierać badaczy w podejmowaniu decyzji dotyczących sposobu przygotowania i udostępnienia danych badawczych, jak również narzędzia informatyczne, analizujące stopień zgodności konkretnego zestawu danych z zasadami FAIR.

FAIR-Aware

FAIR-Aware (https://dans.knaw.nl/en/about/) pomaga ocenić poziom wiedzy na temat zasad FAIR i lepiej zrozumieć, w jaki sposób zgodność z zasadami FAIR może wpłynąć na oddziaływanie danych badawczych. Narzędzie jest niezależne od dyscypliny, dzięki czemu korzystać z niego mogą badacze ze wszystkich dziedzin – zarówno w celach edukacyjnych, jak i podczas realizacji projektu, np. na etapie tworzenia planu zarządzania danymi badawczymi, kiedy podjąć należy decyzje co do sposobu i miejsca udostępniania danych. FAIR-Aware wykorzystać mogą także trenerzy prowadzący szkolenia lub przygotowujący materiały szkoleniowe. Narzędzie rozwijane jest przez DANS (Data Archiving and Networked Services), holenderskie centrum kompetencji w zakresie przechowywania i udostępniania danych badawczych, rozwijające i prowadzące odpowiednią infrastrukturę.

FAIR Data Self Assessment Tool

FAIR Data Self Assessment Tool (https://ardc.edu.au/resource/fair-data-self-assessment-tool/) pomaga ocenić i zrozumieć wpływ różnych decyzji i warunków przechowywania i udostępniania danych na stopień zgodności z FAIR. Narzędzie opiera się na samoocenie, jednak ma nieco inny charakter niż FAIR-Aware. W większym stopniu nadaje się do oceny stopnia FAIR konkretnych danych udostępnionych w określony sposób, można jednak skorzystać z niego także przed udostępnieniem danych. Narzędzie rozwijane jest przez Australian Research Data Commons (ARDC), australijską organizację zajmującą się

rozwijaniem infrastruktury służącej do przechowywania i udostępniania danych, a także prowadzeniem działań zwiększających kompetencje w tym zakresie.

F-UJI Automated FAIR Data Assessment Tool

F-UJI (https://www.f-uji.net/) to narzędzie służące do automatycznej oceny obiektów cyfrowych pod kątem zgodności z zasadami FAIR. Dostępne jest w formie oprogramowania oraz serwisu internetowego, w którym stopień zgodności z FAIR oceniany jest po wpisaniu trwałego identyfikatora zbioru danych. W wyniku analizy generowany jest raport, który zawiera krótkie streszczenie z punktacją i wizualizacją wyników oraz 16 parametrów przypisanych do poszczególnych zasad FAIR. Twórcy narzędzia zastrzegają, że jest ono wciąż rozwijane i oparte na wstępnych danych.

Poradniki i wytyczne

Oprócz narzędzi wspierających udostępnianie danych zgodnie z zasadami FAIR wiele instytucji opracowuje także kompleksowe materiały edukacyjne, przewodniki czy rekomendacje.

How to FAIR – strona edukacyjna

W ramach przewodnika udostępnione zostały materiały edukacyjne, wywiady i quizy, które przybliżają różne aspekty zasad FAIR (https://howtofair.dk/). Strona stworzona przez Danish National Forum for Research Data Management przy wsparciu Danish e-Infrastructure Cooperation (DeiC).

FAIR Cookbook – przewodnik

FAIR Cookbook to przewodnik na temat wdrażania rozwiązań zgodnych z zasadami FAIR w naukach o życiu (https://faircookbook.elixir-europe.org/).

FAIR Software – rekomendacje

Zasady FAIR można odnieść do różnych typów obiektów cyfrowych, również do oprogramowania. Strona FAIR Software (https://fair-software.nl/) zawiera w tym zakresie pięć rekomendacji wraz ze wskazówkami i dodatkowymi informacjami opracowanymi przez Netherlands eScience Center oraz DANS.

Planowanie i organizacja

Planowanie pracy z danymi badawczymi można oprzeć na tzw. cyklu życia danych badawczych, który ukazuje kolejne etapy działań związanych z danymi, podejmowanych podczas realizacji projektu badawczego.

Jako model stanowi on pewne uproszczenie i uogólnienie. Korzystanie z niego pozwala jednak na zaplanowanie kolejnych kroków z uwzględnieniem dobrych praktyk, wytycznych czy obowiązków wynikających z polityk otwartości, a w pewnych wypadkach także na uniknięcie ewentualnych problemów związanych z udostępnianiem danych. Przykładem może być konieczność zadbania o kwestie prawne na etapie planowania i zbierania danych (np. uzyskania odpowiednich zgód bądź

podpisania odpowiednich umów), aby w kolejnych etapach możliwe było udostępnienie danych badawczych z uwzględnieniem zasad FAIR.

W materiałach szkoleniowych dotyczących zarządzania danymi badawczymi można natrafić na różne modele cyklu życia danych badawczych. Różnice między nimi często dotyczą etapu planowania, który bywa pomijany. Czasami też kolejne kroki opatrzone są nieco innymi nazwami lub ułożone są w innej kolejności, np. udostępnianie (*sharing*) bywa określane jako publikowanie (*publishing*), długoterminowe przechowywanie (*preserving*) rozumiane bywa także jako archiwizacja (*archiving*), przetwarzanie i analizowanie danych ujmowane bywają jako osobne etapy.

Różne sposoby konceptualizacji cyklu życia danych badawczych pokazują, że niektóre etapy mogą być ze sobą ściśle powiązane, nachodzić na siebie lub, w pewnych okolicznościach i w pewnym zakresie, przeplatać się, dlatego tak ważne jest zastrzeżenie, że posługujemy się tutaj pewnym modelem.

Inna kwestia istotna dla właściwego rozumienia cyklu życia danych badawczych to jego relacja w stosunku do projektu naukowego. Chociaż dane są zbierane i wytwarzane w ramach konkretnych projektów, cykl życia danych wykracza poza czas realizacji projektu i uwzględnia potencjał ponownego wykorzystania danych. Niektóre zadania projektowe również wymagają uwzględnienia dłuższej perspektywy czasowej, choć

w odniesieniu do innych celów, np. archiwizacji danych czy utrzymania trwałości projektu.

Dane uwzględniające zasady FAIR i udostępniane w sposób otwarty powinny być przygotowane w taki sposób, który umożliwia ich ponowne wykorzystanie w przyszłości, zarówno przez samych autorów, jak i przez innych badaczy. Dlatego model ten zwykle przedstawiony jest właśnie jako cykl – dane wytworzone, opracowane i udostępnione w ramach jednego projektu mogą zostać wykorzystane w kolejnym, a co za tym idzie, zainicjować nowy cykl życia danych.

W poradniku przyjmujemy poniższy model cyklu życia danych badawczych.

1. Planowanie

Jest to ważny etap zarządzania danymi, ponieważ wymaga identyfikacji wielu kwestii w tym zakresie, m.in. aspektów prawnych i etycznych, wymogów instytucji finansującej czy spraw związanych z organizacją pracy. Na tym etapie sporządza się plan zarządzania danymi badawczymi.

2. Wytwarzanie i zbieranie danych

Na tym etapie toczą się prace nad wytwarzaniem i zbieraniem danych według odpowiedniej metodologii i z uwzględnieniem właściwych standardów. Może to obejmować także wyszukiwanie istniejących danych, które zostaną ponownie przeanalizowane bądź połączone z danymi wytwarzanymi w ramach projektu.

3. Opracowanie i analiza danych

Właściwa praca z danymi, najściślej związana z realizacją projektu, uwzględniać może różne działania, w zależności od dziedziny i specyfiki projektu. Część czynności wykonywanych na tym etapie ma charakter techniczny, np. wprowadzanie danych do bazy danych, transkrypcja nagrań, zmiana formatów plików. Inne w większym stopniu wymagają ingerencji w surowe dane bądź tworzenia ich kolejnych wersji, np. zestawów danych poddanych anonimizacji lub pseudonimizacji. Etap ten obejmuje także analizę i interpretację danych, łączenie danych z różnych źródeł czy przygotowanie opracowań, publikacji i raportów na podstawie danych.

4. Przechowywanie danych

Przechowywanie danych istotne jest na każdym etapie realizacji projektu, od zbierania danych aż do ich archiwizacji po zakończeniu projektu. Dlatego też przechowywanie danych rozpatruje się w dwóch perspektywach czasowych, które mogą wymagać różnych narzędzi i zasad. Krótkoterminowe przechowywanie związane jest bezpośrednio z realizacją zadań przewidzianych w projekcie, natomiast długoterminowe przechowywanie uwzględniać powinno dłuższą perspektywę, np. wymogi instytucji finansującej w zakresie archiwizacji danych czy wybór formatów, które będą możliwe do odczytu za kilkanaście lub kilkadziesiąt lat.

5. Udostępnianie danych

Udostępnienie danych wymaga określenia miejsca i sposobu udostępnienia, a także warunków, na których dane zostaną udostępnione. Może również wymagać selekcji danych. Warto na tym etapie pamiętać o zasadach FAIR.

6. Ponowne wykorzystanie danych

Etap wykraczający poza czas realizacji konkretnego projektu naukowego. Udostępnione dane mogą być wykorzystane przez innych badaczy, na przykład ponownie przeanalizowane bądź połączone z innymi danymi. Mogą być także wykorzystywane w celach pozanaukowych.

Plany zarządzania danymi

Plan zarządzania danymi (ang. Data Management Plan, DMP) to jeden z najważniejszych elementów świadomego zarządzania danymi badawczymi. W największym skrócie jest to formalny dokument opisujący to, co dziać się będzie z danymi w ramach całego cyklu ich życia.

Refleksja dotycząca źródeł danych oraz tego, co dzieje się z nimi w trakcie realizacji i po zakończeniu projektu, nie zaczęła się rzecz jasna wraz z tak rozumianymi planami zarządzania danymi. Często jednak prowadziła ona do przyjmowanych ad hoc ustaleń, w wypadku których trudno było odtworzyć powody podjęcia konkretnych decyzji czy szczegóły zastosowanych rozwiązań i wrócić do nich po pewnym czasie.

Plan zarządzania danymi jest tymczasem dokumentem, do którego można i należy wracać. Nie wszystkie kwestie dotyczące zarządzania danymi można przewidzieć w momencie sporządzania jego pierwszej wersji, stąd każdorazowa zmiana planu badań, pojawienie się nowej infrastruktury czy instrumentów badawczych, mogące wpływać na charakterystykę danych i sposób zarządzania nimi, powinny prowadzić do aktualizacji planu zarządzania danymi.

Sporządzenie planu zarządzania danymi coraz częściej stanowi wymóg instytucji finansujących. Niekiedy realizacja tego wymogu konieczna jest już na etapie przygotowania wniosku grantowego (tak jest np. w wypadku Narodowego Centrum Nauki), a niekiedy na początku realizacji projektu (np. sześć miesięcy od jego rozpoczęcia dla projektów finansowanych ze środków Komisji Europejskiej).

Wymogowi sporządzenia planu zarządzania danymi towarzyszy zwykle oczekiwanie, że dane badawcze co do zasady zostaną udostępnione w sposób otwarty co najmniej w zakresie pozwalającym na weryfikację tez zawartych w artykułach naukowych stanowiących rezultat projektu. Odstępstwa od tej zasady – zarówno te „na plus", jak i „na minus" – należy każdorazowo uzasadnić w ramach planu zarządzania danymi.

Wymagania dotyczące planów zarządzania danymi bywają też elementem polityk instytucji prowadzących badania. W związku z tym dokumenty tego rodzaju mogą dotyczyć zarówno danych powstałych w ramach projektu finansowanego ze środków

zewnętrznego grantodawcy, jak i wytworzonych w ramach badań realizowanych przez pracowników naukowych czy doktorantów z danej instytucji.

Instytucje wymagające planów zarządzania danymi bardzo często tworzą ich własne szablony, stanowiące podstawę dla właściwych planów przygotowywanych przez badaczy. Takie szablony zwykle składają się z pytań dotyczących zarządzania danymi, jakie powinien zadać sobie badacz. Każdy zestaw danych jest na swój sposób wyjątkowy, stąd w wypadku zarządzania danymi trudno (dużo trudniej niż np. w odniesieniu do publikacji) o gotowe rozwiązania, które można stosować w każdej lub w większości sytuacji. W tych aspektach, w których udało się jednak wypracować standardy – standardowe słowniki, metodologie, metadane czy dominujące w danym obszarze badań repozytoria specjalistyczne – tym bardziej należy z nich korzystać.

Wielość istniejących szablonów zarządzania danymi ma swoje wady i zalety. Jej zaletą jest to, iż każda instytucja zyskuje dzięki temu możliwość utworzenia szablonu dobrze przystosowanego do własnych potrzeb. Z drugiej jednak strony może to powodować trudności w realizacji kilku projektów finansowanych z różnych źródeł, z których każdy posiada własne zasady zarządzania danymi, wynikające z nieco odmiennych oczekiwań instytucji finansujących.

Wzorcowe rozwiązania stowarzyszenia Science Europe

Aby ukierunkować tę różnorodność, stowarzyszenie Science Europe opracowało zestaw sześciu kluczowych elementów (zwanych wymaganiami), jakie zawierać powinien wzorcowy plan zarządzania danymi. Składają się na nie:

1. Opis danych oraz zbieranie lub ponowne wykorzystanie istniejących danych:
    a. W jaki sposób nowe dane zostaną zebrane lub wytworzone, oraz/lub jak zostaną ponownie wykorzystane dane, które już istnieją?
    b. Jakie dane (np. pod względem rodzaju, formatów i wielkości) zostaną zebrane lub wytworzone?
2. Dokumentacja i jakość danych:
    a. Jakie metadane i dokumentacja (np. dotycząca metodologii zbierania danych oraz sposobu ich organizacji) będą towarzyszyć danym?
    b. Jaki środki kontroli jakości danych zostaną wykorzystane?
3. Przechowywanie oraz kopie zapasowe danych w trakcie procesu badawczego:
    a. W jaki sposób dane i metadane będą przechowywane w trakcie procesu badawczego i jak będą tworzone kopie zapasowe?

b. W jaki sposób w trakcie badań zostanie uwzględniona kwestia zabezpieczeń oraz ochrony danych wrażliwych?

4. Wymogi prawne i etyczne, kodeksy postępowania:

    a. Jeśli dojdzie do przetwarzania danych osobowych, w jaki sposób zostanie zapewniona zgodność z ustawodawstwem dotyczącym danych osobowych oraz zabezpieczania danych?

    b. W jaki sposób zarządzane będą inne kwestie prawne, takie jak prawa własności intelektualnej? Jakie przepisy będą mieć tu zastosowanie?

    c. W jaki sposób zostaną uwzględnione możliwe kwestie etyczne oraz zgodność z kodeksami postępowania?

5. Udostępnianie danych oraz ich długoterminowe przechowywanie:

    a. W jaki sposób i kiedy dane zostaną udostępnione? Czy możliwe są ograniczenia w udostępnianiu danych lub istnieją powody do wprowadzenia embarga?

    b. W jaki sposób zostaną wybrane dane, które będą przechowywane? Gdzie dane będą przechowywane długoterminowo (np. w repozytorium danych lub archiwum)?

c. Jakie metody lub narzędzia związane z oprogramowaniem będą konieczne do uzyskania dostępu do danych i ich wykorzystania?

d. W jaki sposób zostanie zapewnione przypisanie unikalnego i trwałego identyfikatora (takiego jak DOI) do każdego zbioru danych?

6. Obowiązki oraz zasoby związane z zarządzaniem danymi:

   a. Kto (jaka rola, stanowisko lub instytucja) będzie odpowiadać za zarządzanie danymi (tj. *data stewardship*)?

   b. Jakie zasoby (np. finansowe i czasowe) zostaną przeznaczone na zarządzanie danymi i zapewnienie ich zgodności z zasadami FAIR?

Kolejność powyższych sześciu wymagań może być zmieniana w ich konkretnych implementacjach, ważne jednak, by każdy szablon planu zarządzania danymi powstały na ich bazie uwzględniał je wszystkie.

Wychodząc od powyższych wymagań, stowarzyszenie Science Europe opracowało również przykładowy szablon planu zarządzania danymi.

Zalety tworzenia planu zarządzania danymi

Wśród korzyści, jakie dają plany zarządzania danymi, twórcy przewodnika do zarządzania danymi badawczymi opracowanego przez organizację CESSDA ERIC wymieniają[10]:

1. Uzyskanie narzędzia pozwalającego na myślenie o danych z wyprzedzeniem oraz planowanie. Dzięki DMP możemy już na wczesnym etapie realizacji projektu zdać sobie sprawę z możliwych problemów i zastanowić się, w jaki sposób zostaną one rozwiązane. Możemy też od początku zbierać dane i dokumentować je w sposób, który później ułatwi ich archiwizację i ponowne wykorzystanie.
2. Łatwiejsze zarządzanie projektem. DMP to dokument, w którym opisane są wszystkie istotne kwestie związane z danymi w projekcie i którego istnienie na wszystkich etapach realizacji projektu ułatwia dotarcie do wielu potrzebnych informacji.
3. Łatwiejsze określenie potrzebnego budżetu. Dzięki DMP łatwiej jest uniknąć pułapek związanych z niedoszacowaniem kosztów wytworzenia wysokiej jakości danych. DMP może w tym pomóc zarówno na etapie wniosku grantowego, jak i na późniejszych etapach realizacji projektu, kiedy zachodzi konieczność szczegółowego rozplanowania wydatków w czasie.

10 Opracowanie na podstawie CESSDA Data Management Expert Guide, https://dmeg.cessda.eu/Data-Management-Expert-Guide/1.-Plan/Benefits-of-data-management [data dostępu: 17.07.2023].

4. Większa zgodność danych z zasadami FAIR. DMP pozwala z wyprzedzeniem przemyśleć, w jaki sposób w toku realizacji projektu zapewnić możliwość odnalezienia, dostępność, interoperacyjność, możliwość ponownego wykorzystania danych.
5. Klarowne wskazanie odpowiedzialności. Opracowując DMP, pokazuje się instytucji finansującej, instytucji macierzystej oraz partnerom, że do kwestii zarządzania danymi podchodzi się poważnie, a więc że poważnie podchodzi się też do kwestii wydatkowania środków na badania.

Narzędzia wspierające tworzenie planów zarządzania danymi

W sieci dostępne są serwisy wspomagające tworzenie planów zarządzania danymi. Można wśród nich wymienić dwa serwisy dostępne za pośrednictwem EOSC Marketplace:

- Argos (https://argos.openaire.eu/),
- DMPonline (https://dmponline.dcc.ac.uk/).

Serwisy te zawierają liczne szablony planów zarządzania danymi, ułatwiając zalogowanym użytkownikom wypełnienie odpowiedniego z nich oraz udostępnienie go innym osobom zaangażowanym w przygotowanie i/lub realizację projektu. Serwis Argos pozwala dodatkowo powiązać gotowy plan z innymi rezultatami projektu badawczego, takimi jak np. zbiór danych.

Część z planów zarządzania danymi przygotowanych przy użyciu tych serwisów jest przez ich autorów udostępniana publicznie wszystkim zainteresowanym. Lektura takich planów może być bardzo pouczająca, pozwala bowiem sprawdzić, jak inni rozwiązali problemy, przed którymi aktualnie stoimy my sami. Ważne jednak, by była to lektura krytyczna – wśród udostępnionych planów nie brakuje bowiem i takich, które prezentują niską jakość.

Plany zarządzania danymi w politykach otwartości Komisji Europejskiej i Narodowego Centrum Nauki

Plan zarządzania danymi badawczymi jako formalny dokument

Plan zarządzania danymi badawczymi wymagany jest przez wiele instytucji finansujących badania naukowe, które mogą formułować w tym zakresie różne szczegółowe wymagania dotyczące np. elementów planów zarządzania danymi, terminów przygotowania i przedłożenia planów bądź zasad ich aktualizacji.

Poniżej omówimy polityki i wytyczne dwóch instytucji: Komisji Europejskiej (w ramach programu Horyzont Europa) oraz Narodowego Centrum Nauki. Obie odwołują się do ogólnych wytycznych Science Europe dotyczących planów zarządzania danymi badawczymi.

Komisja Europejska, program Horyzont Europa

Obowiązkowym elementem polityki otwartości w programie Horyzont Europa jest odpowiedzialne zarządzanie danymi badawczymi zgodnie z zasadami FAIR, w szczególności poprzez sporządzanie planów zarządzania danymi i zapewnienie otwartego dostępu do danych badawczych zgodnie z zasadą „tak otwarte, jak to możliwe, tak zamknięte, jak to niezbędne" oraz zgodnie z zasadami określonymi w umowie grantowej.

Plany zarządzania danymi badawczymi w programie Horyzont Europa

- Uznawane są za podstawę odpowiedzialnego zarządzania danymi badawczymi.
- Stanowią integralną część metodologii zarządzania projektem, a dobre zarządzanie danymi badawczymi usprawnia pracę, oszczędza czas, zwiększa bezpieczeństwo danych i wpływa na ich jakość.
- Pomagają badaczom we właściwym zarządzaniu nie tylko danymi, ale też publikacjami i innymi rezultatami badań, zarówno cyfrowymi, jak i niecyfrowymi. Dotyczy to:
    - oprogramowania stworzonego podczas projektu,
    - procedur,
    - protokołów,
    - materiałów, takich jak próbki, linie komórkowe, przeciwciała,
    - innych rezultatów.
- Powinny być żywymi dokumentami, aktualizowanymi i wzbogacanymi w toku realizacji projektu, np. w związku

ze zmianami, pojawieniem się nieoczekiwanych okoliczności, nowych metod badawczych, bądź w związku z formalnymi zmianami, np. w konsorcjum.

- Dobrą praktyką jest udostępnianie planów w sposób otwarty, jeżeli to możliwe.
- Plany mogą być też publikowane w odpowiednich czasopismach lub służących do tego platformach (np. RIO, https://riojournal.com/), lub deponowane w repozytoriach (DMP online, https://dmponline.dcc.ac.uk/).
- Pełny plan zarządzania danymi badawczymi nie jest wymagany na etapie składania wniosku o finansowanie.
- Korzystanie z wzoru udostępnionego na stronie Komisji Europejskiej jest rekomendowane, ale nie obowiązkowe.

Elementy planu zarządzania danymi wskazane w przewodniku po programie Horyzont Europa[11]:

- opis danych,
- standardy i metadane,
- nazwy i trwałe identyfikatory zestawów danych,
- kuratorowanie (*curation*) oraz przechowywanie i zabezpieczanie (*preservation*) danych,
- udostępnianie danych,
- zarządzanie innymi rezultatami,
- zasoby, zadania i koszty zarządzania danymi.

[11] Podstawowe wskazówki dotyczące sporządzania planów zarządzania danymi badawczymi zawarto w przewodniku po programie Horyzont Europa, Horizon Europe, Programme Guide, Version 3.0 01 April 2023, por. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf [data dostępu: 17.07.2023].

Narodowe Centrum Nauki

Narodowe Centrum Nauki wymaga złożenia planu zarządzania danymi badawczymi w formie załącznika do wniosku o finansowanie[12].

Elementy planu zarządzania danymi wskazane w wytycznych:

- opis danych oraz pozyskiwanie lub ponowne wykorzystanie dostępnych danych,
- dokumentacja i jakość danych,
- przechowywanie i tworzenie kopii zapasowych podczas badań,
- wymogi prawne, kodeksy postępowania,
- udostępnianie i długotrwałe przechowywanie danych,
- zadania związane z zarządzaniem danymi oraz zasoby.

[12] Aby ułatwić pracę nad planami, NCN przygotowało wytyczne dla wnioskodawców w języku polskim https://www.ncn.gov.pl/sites/default/files/pliki/regulaminy/wytyczne_zarzadzanie_danymi.pdf i w języku angielskim https://ncn.gov.pl/sites/default/files/pliki/regulaminy/wytyczne_zarzadzanie_danymi_ang.pdf [data dostępu: 17.07.2023].

Repozytoria danych badawczych

Zgodnie z definicją zawartą w Annotated Grant Agreement dla projektów Horizon Europe repozytorium to internetowe archiwum, w którym badacze mogą deponować cyfrowe rezultaty działalności badawczej i zapewnić do nich (otwarty) dostęp. W zależności od rodzaju gromadzonych materiałów wyróżnić można repozytoria publikacji naukowych oraz repozytoria danych badawczych. Repozytoria można podzielić na typy w zależności od kilku kryteriów.

Zakres gromadzonych treści

Pierwszym (i chyba najważniejszym) z tych kryteriów jest zakres treści gromadzonych w repozytorium danych. Korzystając z tego kryterium, możemy wyróżnić:

- Repozytoria specjalistyczne, które gromadzą dane wąsko określonego rodzaju i często posiadają bardzo konkretne wymogi dotyczące typu deponowanych plików oraz sposobu ich dokumentacji. Taka specyfika pozwala repozytoriom specjalistycznym na opatrywanie danych metadanymi znakomicie przystosowanymi do ich charakteru. Przykładem repozytorium specjalistycznego może być repozytorium PDB (ang. Protein Data Bank), gromadzące wyłącznie dane dotyczące struktur białkowych.
- Repozytoria dziedzinowe, które są poświęcone danym o szerszym spektrum i służą badaczom z określonej dziedziny, maksymalnie kilku dziedzin. Przykład może tu stanowić repozytorium Pangaea, gromadzące informacje z zakresu nauk o Ziemi i środowisku, czy też polskie repozytoria: CLARIN (gromadzące dane z zakresu językoznawstwa) oraz Repozytorium Danych Społecznych (w ramach którego działają Archiwum Danych Jakościowych oraz gromadzące społeczne dane ilościowe Polskie Archiwum Danych Społecznych).
- Repozytoria instytucjonalne, które gromadzą dane powiązane z konkretną instytucją. To, na czym w szczegółach może polegać takie powiązanie, określa

samo repozytorium. Repozytoria instytucjonalne są zwykle zainteresowane zbiorami danych wytworzonymi przez pracowników naukowych instytucji, a niektóre z nich dopuszczają również zbiory wytworzone przez doktorantów i/lub studentów oraz zbiory będące rezultatem projektu, w którym instytucja prowadząca serwis uczestniczy w charakterze partnera, członka konsorcjum lub lidera.

- Repozytoria ogólnego przeznaczenia (zwane również repozytoriami sierocymi), które są gotowe przyjąć każdy rodzaj danych badawczych. Nazwa „repozytoria sieroce" bierze się z tego, iż winny do nich trafiać jedynie te zbiory danych, w wypadku których nie istnieją odpowiednie repozytoria specjalistyczne, dziedzinowe czy instytucjonalne. Przykładem takiego repozytorium może być prowadzone przez ICM UW Repozytorium Otwartych Danych RepOD (https://repod.icm.edu.pl/). Pełni ono jednak również funkcję repozytorium instytucjonalnego szeregu polskich instytucji naukowych. Jednym z największych repozytoriów tego rodzaju jest z kolei repozytorium Zenodo (https://zenodo.org/) dostępne w ramach EOSC Marketplace, które jest prowadzone przez CERN i stanowi element infrastruktury OpenAIRE. Repozytorium to, dzięki bardzo dobrej integracji z innymi serwisami, oferuje też pewne dodatkowe usługi, np. przypisywanie identyfikatorów DOI do poszczególnych wydań oprogramowania znajdującego się w największym repozytorium kodu GitHub.

Weryfikacja danych

Repozytoria różnią się również co do zakresu prowadzonej weryfikacji danych. Może je cechować[13]:

1. Brak weryfikacji: zbiory danych publikowane są w takiej postaci, w jakiej zostały przygotowane i dostarczone do repozytorium przez użytkownika.
2. Podstawowy poziom weryfikacji: repozytorium oferuje elementarne sprawdzenie poprawności elementów zbioru, a także uzupełnienie metadanych lub dokumentacji w podstawowym zakresie.
3. Rozszerzony poziom weryfikacji: repozytorium oferuje konwersję do nowych formatów oraz rozszerzenie dokumentacji.
4. Weryfikacja na poziomie danych: repozytorium oferuje dodatkowo sprawdzenie i edycję samych danych.

Wysoki poziom weryfikacji danych cechuje zwykle repozytoria o charakterze specjalistycznym oraz dziedzinowym. Jego zapewnienie jest z kolei bardzo trudne – jeśli nie niemożliwe – w przypadku repozytoriów ogólnego przeznaczenia, które muszą być przygotowane na przyjęcie bardzo szerokiego spektrum danych.

---

[13] CoreTrustSeal Standards and Certification Board, CoreTrustSeal Requirements 2023-2025 (V01.00), Zenodo, 2022, https://doi.org/10.5281/zenodo.7051012 [data dostępu: 17.07.2023].

W repozytoriach można również spotkać się z pewnymi ograniczeniami dotyczącymi wielkości deponowanych zasobów. Ograniczenia te mogą dotyczyć wielkości pojedynczego pliku wchodzącego w skład zbioru danych. Na przykład w repozytoriach opartych na oprogramowaniu Dataverse limit ten jest konfigurowalny i wynosi zwykle kilka gigabajtów (np. Harvard Dataverse – 2,5 GB, RepOD – 5 GB). Limit taki, choć może sprawiać nieco kłopotów po stronie osoby deponującej dane (w wypadku większych wielkości wymusza bowiem utworzenie wielu mniejszych plików archiwum), to jednak ułatwia zarazem pobieranie danych użytkownikom z odległych zakątków świata oraz tym, którzy dysponują słabym łączem internetowym. Mogą również istnieć ograniczenia dotyczące wielkości całego zbioru – np. w repozytorium Zenodo wynosi ono 50 GB, a w repozytorium Figshare 20 GB. Niektóre repozytoria posiadają też limity łącznej wielkości zbiorów danych zdeponowanych przez pojedynczego użytkownika (pojedyncze konto). Część repozytoriów dopuszcza ponadto możliwość zdeponowania danych o rozmiarze przekraczającym przyjęte limity, co wymaga jednak od użytkownika podjęcia dodatkowych działań, np. skontaktowania się z obsługą repozytorium lub wniesienia opłaty za dodatkową przestrzeń dyskową.

Rozproszone dane z perspektywy instytucji

Repozytorium instytucjonalne – w szczególności takie, któremu towarzyszy odpowiednia polityka regulująca kwestie otwartego

dostępu do danych badawczych i ich deponowania – często jest również wykorzystywane w charakterze źródła danych dla innych systemów w obrębie instytucji, np. odpowiadających za gromadzenie i prezentację dorobku jej jednostek czy pracowników, a także za wewnętrzne procedury ewaluacyjne.

Wspomniana polityka powinna uwzględniać to, iż część danych – zgodnie z dobrymi praktykami w zakresie zarządzania danymi badawczymi – mimo wszystko nie zostanie zdeponowana w repozytorium instytucjonalnym, ale w repozytorium specjalistycznym lub dziedzinowym. W tej sytuacji powinna istnieć możliwość rejestracji w obrębie infrastruktury instytucjonalnej tych zbiorów danych, które zostały zdeponowane poza repozytorium instytucjonalnym. Pozwala to uwzględnić rozproszony dorobek w postaci zbiorów danych zdeponowanych w najbardziej odpowiednich do tego celu serwisach specjalistycznych lub dziedzinowych.

Kryteria wyboru repozytoriów i bazy repozytoriów

Poszukiwanie najlepszego repozytorium dla swoich danych najlepiej rozpocząć wśród repozytoriów specjalistycznych i dziedzinowych. Dzięki odpowiednim schematom metadanych oraz zasadom kuratorowania danych są one w stanie zaoferować warunki pozwalające na maksymalizację zgodności z zasadami FAIR.

Trzeba przy tym pamiętać, że repozytoria odpowiadają jedynie za część zagadnień związanych z udostępnianiem danych w sposób zgodny z zasadami FAIR. Repozytorium powinno np. umożliwić opis zasobu przy użyciu standardowych metadanych i zapewnić do nich maszynowy dostęp, ale jeżeli użytkownik takiego opisu nie dostarczy, dane go pozbawione nie będą udostępnione zgodnie z zasadami FAIR. Ujmując rzecz jeszcze inaczej, również w repozytorium umożliwiającym udostępnienie danych zgodnie z zasadami FAIR można udostępnić zbiory danych w sposób niezgodny z zasadami FAIR.

O zgodność taką łatwiej w repozytoriach specjalistycznych i dziedzinowych, mogących zapewnić również wyższy poziom weryfikacji danych. W ich wypadku więcej zadań spada na personel repozytorium, który w większym stopniu odpowiada za zapewnienie zgodności z zasadami FAIR.

Znalezienie odpowiedniego repozytorium ułatwią bazy gromadzące informacje o istniejących serwisach. Na szczególną uwagę zasługują:

- Baza Registry of Research Data Repositories (https://www.re3data.org/) – gromadząca informacje o repozytoriach danych badawczych z całego świata. Baza dostępna jest za pośrednictwem EOSC Marketplace. Za jej prowadzenie odpowiada partnerstwo, w skład którego wchodzą Berlin School of Library and Information Science, Helmholtz Open Science Office, KIT Library, Purdue University Libraries oraz DataCite.

- Baza OpenDOAR (https://v2.sherpa.ac.uk/opendoar/) – znaleźć w niej można informacje o repozytoriach danych badawczych, ale też o repozytoriach publikacji czy bibliotekach cyfrowych.
- Katalog Open Access Directory prowadzony przez Simmons University (https://oad.simmons.edu/oadwiki/Data_repositories).

Część instytucji finansujących badania wymaga, aby dane stanowiące rezultat projektu zdeponować w zaufanym repozytorium. Instytucja finansująca może przy tym sama określić warunki, jakie musi spełnić zaufane repozytorium (dobrym wyborem wydaje się tu repozytorium certyfikowane), lub skorzystać z istniejących kryteriów dla takich repozytoriów (np. kryteria opracowane przez Science Europe).

Składają się na nie:

1. Zapewnienie trwałego i unikalnego identyfikatora (PID):
    a. Umożliwienie odnalezienia i identyfikacji danych.
    b. Umożliwienie wyszukiwania, cytowania i pobierania danych.
    c. Zapewnienie wsparcia dla wersjonowania danych.
2. Metadane:
    a. Umożliwienie odszukania danych.
    b. Umożliwienie odniesienia się do odpowiednich powiązanych informacji, takich jak inne dane i publikacje.

c. Zapewnienie i utrzymanie publicznie dostępnej informacji dotyczącej danych, nawet w przypadku danych nieopublikowanych, chronionych, wycofanych lub usuniętych.
d. Wykorzystanie standardów metadanych, które są szeroko akceptowane (przez społeczność naukową).
e. Zapewnienie maszynowej dostępności metadanych.

3. Dostęp do danych i licencje dotyczące wykorzystania:
    a. Umożliwienie dostępu do danych pod dobrze określonymi warunkami.
    b. Zapewnienie autentyczności i integralności danych.
    c. Umożliwienie pobrania danych.
    d. Dostarczenie informacji dotyczących licencji i zgód (najlepiej w formie dostępnej maszynowo).
    e. Zapewnienie poufności i poszanowania praw twórców i osób badanych.
4. Przechowywanie:
    a. Zapewnienie trwałości danych i metadanych.
    b. Transparentność misji, zakresu, polityk przechowywania oraz planów (w tym dotyczących zarządzania, długookresowej trwałości finansowania, okresu przechowywania danych oraz planów zapewnienia ciągłości działania).

Organizacja danych i zarządzanie ich wersjami

Dobra organizacja folderów i plików z danymi badawczymi polega przede wszystkim na przyjęciu określonej konwencji i konsekwentnym trzymaniu się jej zasad. Celem jest tu przede wszystkim uniknięcie zamieszania i konfuzji u potencjalnych odbiorców danych (a więc również nas samych, jeśli np. zechcemy kiedyś powrócić do opracowanych jakiś czas temu danych), do czego mogłyby prowadzić nazwy plików nieniosące żadnej informacji zrozumiałej dla odbiorców lub przypadkowe czy arbitralne stosowanie wielu różnych konwencji.

Trzeba przy tym pamiętać, że aby konwencja nazewnicza mogła spełnić swoje zadania, należy poinformować o jej zasadach tych,

którzy mają się do niej stosować. W trakcie realizacji projektu będą to przede wszystkim członkowie zespołu, stąd właściwym miejscem na omówienie zasad takiej konwencji będzie plan zarządzania danymi, który powinien być im znany. Z kolei w wypadku danych już zarchiwizowanych i udostępnionych szerokiemu gronu odbiorców właściwym miejscem na opisanie przyjętej konwencji jest dokumentacja zbioru danych.

Dobre praktyki w zakresie organizacji folderów i plików

W zakresie organizacji plików warto skorzystać z rad, jakie znaleźć można na stronie Uniwersytetu w Cambridge[14].

1. Należy korzystać z folderów w taki sposób, aby informacje na określony temat znajdowały się w jednym miejscu.
2. Warto trzymać się istniejących procedur i nie tworzyć rozwiązań ad hoc. Jeśli nasz zespół lub instytucja posiadają już sprawdzone procedury dotyczące organizacji plików i ich nazewnictwa, spróbujmy wykorzystać je w pierwszej kolejności, a alternatywnych rozwiązań szukajmy dopiero w razie pojawienia się dobrych powodów.
3. Nazewnictwo folderów powinno odpowiadać obszarowi zadań, którego dotyczy ich zawartość. Złym pomysłem jest natomiast nazywanie folderów w sposób związany

[14] Opracowanie na podstawie University of Cambridge, Data Management Guide, https://www.data.cam.ac.uk/data-management-guide/organising-your-data [data dostępu: 17.07.2023].

z pojedynczymi badaczami. Skład zespołu może się zmieniać, podobnie jak mogą zmieniać się obszary odpowiedzialności poszczególnych osób w projekcie, stąd tego typu nazewnictwo może w pewnym momencie okazać się mylące, w szczególności w sytuacji współdzielenia folderów przez kilkoro członków projektu oraz dla nowych osób, które dołączają do zespołu realizującego badanie.

4. Warto dbać o spójność i trzymać się raz wybranej konwencji. Pomoże w tym jej jasne określenie na jak najwcześniejszym etapie procesu badawczego.
5. Folderom warto nadać odpowiednią hierarchię. Na jej szczycie powinny znajdować się foldery obejmujące szerokie kwestie, zaś w miarę schodzenia w dół hierarchii ich zakres tematyczny powinien ulegać zawężeniu.
6. Należy oddzielić od siebie pliki, nad którymi wciąż pracujemy od tych, które osiągnęły już ostateczną postać. Pozwoli to uniknąć przypadkowych zmian w tych plikach, które nie powinny już być zmieniane.
7. Pomocne będzie założenie osobnego folderu na pliki, z którymi właśnie pracujemy, i raz na jakiś czas (np. co tydzień lub miesiąc) przeglądanie jego zawartości i przenoszenie gotowych plików w osobne miejsce.
8. Należy oczywiście zadbać o tworzenie kopii zapasowych plików – zarówno tych znajdujących się na dyskach lokalnych naszych komputerów, jak i tych dostępnych w chmurze/lokalizacjach sieciowych.

9. Co jakiś czas należy oszacować, czy dany plik jest nadal potrzebny, czy też może spełnił już swoje zadanie i nie ma potrzeby jego dalszego przechowywania.

Konwencje nazewnicze

Uzgodnienie konwencji nazewniczej i jej konsekwentne stosowanie przede wszystkim pozwala uczestnikom projektu na właściwą identyfikację plików i łatwe poruszanie się między nimi. Pomaga również zapobiegać kłopotom związanym z kontrolowaniem różnych wersji dokumentu. Na wstępie warto uzgodnić kilka ogólnych zasad dotyczących:

- słowników stosowanych w ramach nazw plików, tak by wszyscy używali ich w tym samym znaczeniu;
- interpunkcji, tak by określone symbole (takie jak np. podkreślnik „_”) czy wielkie litery mogły być stosowane w konsekwentny sposób;
- formatu daty, np. YYYYMMDD;
- kolejności plików, tak aby pliki na ten sam temat pojawiały się obok siebie;
- liczby zer wiodących (01, 001, 0001 itd.), jeśli w nazwach plików pojawiają się liczby.

Jak konwencja nazewnicza może wyglądać w praktyce? Przyjrzyjmy się przykładowej konwencji nazewniczej przyjętej w Cambridge[15].

[15] Zob. TILS Document Naming Convention, https://www.data.cam.ac.uk/files/gdl_tilsdocnaming_v1_20090612.pdf [data dostępu: 17.07.2023].

Weźmy przykładową nazwę pliku zawierającego wskazówki dotyczące konwencji nazewniczej:

GDL_TILSDocNaming_V1_20090612.docx

Poszczególne elementy nazwy tego pliku oddzielone są od siebie za pomocą podkreślników „_”.

Pierwszy, trzyliterowy element określa, z jakiego typu dokumentem mamy do czynienia. Litery „GDL” informują, że są to wskazówki (*guidelines*), o czym wiemy dzięki temu, że skrót ten pochodzi z krótkiego słownika prefiksów. Znaleźć w nim można również szablony (TEM – od angielskiego *template*), prezentacje (PRE – od *presentation*) czy raporty (REP – od *reports*).

Kolejny element to skrócony tytuł dokumentu, którego pełna wersja to „TILS Document Naming Convention”.

Następny element określa z kolei wersję dokumentu; w tym wypadku jest to jego pierwsza wersja.

Ostatni element nazwy to data utworzenia, rewizji lub akceptacji pliku. Data ta została podana w formacie YYYYMMDD, który pozwala na chronologiczne posortowanie listy plików z uwzględnieniem tej daty.

Opcjonalnymi elementami, które mogą zostać dodane po dacie, są informacja na temat statusu dokumentu (np. „DR” dla draftu,

czyli wersji roboczej dokumentu czy „FIN” dla jego wersji finalnej), a także informacja w postaci inicjałów osoby edytującej dokument.

Dzięki stosowaniu tej konwencji na podstawie nazwy pliku:

GDL_EmailManagement_V1_20081120_DR_NR.docx

możemy stwierdzić, że mamy do czynienia z dokumentem zawierającym wskazówki dotyczące zarządzania e-mailami i że jest to jego pierwsza wersja robocza z 20 listopada 2008 r., edytowana przez osobę o inicjałach NR.

Wersjonowanie

Jak widać, podstawowe informacje dotyczące wersji pliku zawrzeć można już w jego nazwie. Dodatkowo dokument da się rozbudować o tabelę zawierającą informację o wersjach, która określać będzie datę powstania danej wersji, osobę lub osoby, które ją wytworzyły, a także informacje o zmianach względem poprzednich wersji.

Należy przy tym zaznaczyć, iż tym, co powinno powodować inkrementację numeru wersji, jest pojawienie się nowej tzw. rewizji (*revision*) pliku, a nie każde zapisanie minimalnych zmian na dysku. Pomiędzy poszczególnymi rewizjami może powstać wiele wersji pliku, które będą się różnić w minimalnym stopniu. Dopiero gdy zmiany te osiągną poziom pewnej skończonej

całości (zwanej właśnie rewizją), powinniśmy zwiększyć numer wersji zawarty w nazwie pliku oraz jego tabeli wersji.

Nieco inaczej wygląda natomiast wersjonowanie samych zbiorów danych. W wypadku repozytoriów bazujących na oprogramowaniu Dataverse, drobna zmiana w metadanych zbioru powoduje powstanie tzw. małej wersji, gdzie inkrementacji ulega druga cyfra numeru wersji (dostajemy więc wersję 1.1., 1.2 itd.).

W wypadku modyfikacji plików z danymi zmienia się już jednak pierwsza cyfra numeru wersji (2.0, 3.0, 4.0 itd.) oraz sugerowane cytowanie zbioru, które również zawiera informację o konkretnej wersji zbioru.

W obu wypadkach zmianie nie ulega identyfikator DOI przypisany do zbioru danych.

Nieco inaczej podchodzi się do wersjonowania repozytorium Zenodo. Tu każdy zbiór danych posiada dwa numery DOI: jeden przypisany do konkretnej wersji zbioru i drugi przypisany do abstrakcyjnej „koncepcji" (*concept*) zbioru danych. Ten drugi identyfikator DOI obejmuje niejako wszystkie wersje zbioru danych, a adres URL zbudowany na jego podstawie kieruje zawsze do najnowszej wersji zbioru.

Korzystając z jakichś danych, powinniśmy w tym wypadku zawsze precyzyjnie wskazać wersję danych, z której korzystamy, i w cytowaniu posłużyć się numerem DOI przypisanym do tej konkretnej wersji. Dopiero gdybyśmy chcieli odnieść się do

wszystkich wersji zbioru, a więc do jego „koncepcji”, powinniśmy posłużyć się cytowaniem wykorzystującym szerszy identyfikator DOI.

Techniczne i finansowe ograniczenia w udostępnianiu danych badawczych

Przy opracowywaniu nazw plików istotne jest unikanie w nich znaków specjalnych, spacji oraz znaków diakrytycznych. Jest to szczególnie ważne w wypadku nazw plików przygotowywanych do archiwizacji i udostępnienia. W związku z tym, że nie wiemy kto, na jakim sprzęcie i na jakim systemie operacyjnym będzie próbował pobrać i otworzyć nasze dane, nie wiemy również, czy osoba ta będzie w stanie np. wpisać z klawiatury polskie „ogonki”, ani jak polskie znaki zostaną wyświetlone na jej ekranie; nie wiemy też, czy używany przez nią system operacyjny będzie pozwalać na umieszczanie w nazwach plików znaku spacji. (Z podobnych przyczyn nazwy plików powinny być zwięzłe, tak aby zminimalizować ryzyko użycia nazwy zbyt długiej dla danego systemu plików). Problematyczne mogą być nawet kropki (często sprawiające kłopot skryptom pisanym przez mniej doświadczonych programistów), a także warianty myślnika/znaku minusa. Tych ostatnich jest bowiem wiele i trudno niekiedy zorientować się, który z nich został użyty w nazwie jakiegoś konkretnego pliku i jak można go wprowadzić z klawiatury.

W tej sytuacji najlepiej ograniczyć się do jak najuboższych środków, pozostawiając sobie do dyspozycji litery angielskiego alfabetu, cyfry oraz znak podkreślnika „_”.

Udostępniając dane badawcze, warto też pamiętać o technicznych ograniczeniach stawianych przez poszczególne repozytoria. Ograniczenia te dotyczyć mogą formatów plików akceptowanych przez dane repozytorium czy wielkości poszczególnych plików i zbiorów. Część repozytoriów dopuszcza deponowanie większych danych po nawiązaniu kontaktu z obsługą lub uiszczeniu dodatkowej opłaty, ale w takich wypadkach warto zawsze uwzględnić nie tylko wygodę własną czy koleżanek i kolegów z danego obszaru geograficznego, ale też sytuację tych osób, które mogą być zainteresowane danymi, ale dysponują słabymi łączami i mieszkają w odległych zakątkach świata. Dla nich pobranie większej liczby mniejszych plików może okazać się znacznie łatwiejsze niż pobranie jednego czy kilku plików rozmiaru kilkudziesięciu czy kilkuset gigabajtów.

W wypadku danych o naprawdę dużych rozmiarach, wynoszących kilka terabajtów i więcej, dostęp do nich polegający na tym, że użytkownik pobiera takie dane i analizuje je lokalnie, jest bardzo problematyczny. Rozmiar taki przekracza wielkość dostępnej przestrzeni dyskowej na lokalnym komputerze, co powoduje po stronie użytkownika konieczność zapewnienia macierzy dyskowej, na którą dane takie można by pobrać. Jest to więc nie tylko kłopot techniczny, lecz także źródło dodatkowych kosztów.

Pobieranie danych trwałoby jednak bardzo długo: kilka dni, tygodni, a nawet miesięcy. W takiej sytuacji o wiele sensowniejszy wydaje się model dostępu, w którym dane zdeponowane w repozytorium można przeanalizować, korzystając

z infrastruktury obliczeniowej zintegrowanej z systemem przechowywania danych (np. macierzą S3) wykorzystywanym przez repozytorium. Dostęp do danych odbywa się w takich wypadkach pośrednio, dzięki zdalnemu dostępowi do maszyn służących do prowadzenia analiz obliczeniowych.

Z kolei z perspektywy osoby zainteresowanej wykorzystaniem takich danych ważne jest, by odpowiednio wcześnie ustalić, na jakich zasadach finansowane są takie obliczenia. Możliwe, że ich przeprowadzenie będzie wymagać odpowiedniego grantu obliczeniowego lub pokrycia kosztów takich obliczeń. W tym ostatnim wypadku warto więc zadbać o ich ujęcie w kosztorysie projektu.

Przechowywanie danych badawczych

Bezpieczne przechowywanie danych ma kluczowe znaczenie w procesie prowadzenia badań w sposób efektywny i odpowiedzialny. Jest to istotne zarówno na etapie realizacji projektu, kiedy dane są na bieżąco wytwarzane, opisywane i analizowane, jak i po jego zakończeniu, kiedy należy zadbać o archiwizację rezultatów badań. Te dwie perspektywy czasowe niosą za sobą różne wyzwania i potrzeby, które należy wziąć pod uwagę, planując pracę z danymi.

Kryteria wyboru miejsca przechowywania danych

Co należy wziąć pod uwagę?

- Istotne ramy prawne, polityki instytucji prowadzących lub finansujących badania, zobowiązania wynikające z umów.
- Bezpieczeństwo przechowywania danych.
- Zasadniczy cel związany z konkretnym etapem realizacji projektu badawczego, którym może być w szczególności:
    - bieżąca praca z danymi,
    - otwarte udostępnianie danych,
    - długoterminowe przechowywanie danych.
- Techniczne uwarunkowania, wynikające np. z rozmiaru czy typu danych bądź oprogramowania wykorzystywanego w bieżącej pracy z danymi.
- Koszty i zasoby potrzebne do przechowywania danych.

Miejsca przechowywania danych

Urządzenia przenośne

Urządzenia przenośne nie są zalecane do przechowywania danych badawczych, ponieważ wykazują podatność na utratę i zniszczenie. Co do zasady korzystanie z nich nie jest dobrą praktyką, ale mogą być przydatne w niektórych sytuacjach, np. podczas pracy w terenie lub jako tymczasowe miejsce przechowywania kopii zapasowej, zwłaszcza kiedy inne rozwiązania nie są dostępne. Urządzenia przenośne nie powinny być traktowane jako jedyne miejsca przechowywania danych.

Jeżeli korzystanie z nich jest konieczne, należy wybrać produkt wysokiej jakości, korzystać z niego zgodnie z zaleceniami producenta i regularnie sprawdzać jego działanie.

Urządzenia przenośne Zewnętrzne dyski twarde, dyski flash i płyty CD		
Zalety	**Wady**	**Środki ostrożności dot. (wrażliwych) danych osobowych**
Umożliwiają łatwe przenoszenie danych i plików bez konieczności przesyłania ich przez Internet. Może to być szczególnie pomocne podczas pracy w terenie. Tanie rozwiązanie.	Mogą zostać zgubione, uszkodzone, ukradzione itd., dlatego nie są bezpiecznym miejscem przechowywania danych. Nie są odpowiednie do długoterminowego przechowywania danych lub przechowywania kopii głównych plików. Nie umożliwiają automatycznej kontroli wersji.	Należy używać ich w połączeniu z szyfrowaniem i ochroną silnym hasłem.
Rekomendacje Używaj do tymczasowego, krótkotrwałego przechowywania danych niewrażliwych, np. w terenie, lub do przenoszenia danych i plików, gdy transmisja online nie jest możliwa. Używaj w połączeniu z szyfrowaniem i silną ochroną hasłem, zwłaszcza podczas pracy z poufnymi informacjami. Przeprowadzaj regularne kontrole, aby upewnić się, że urządzenie działa i że pliki są dostępne. Nie używaj do długoterminowego przechowywania danych ani kopii wzorcowych plików.		

Chmura

Usługi przechowywania danych w tzw. chmurze polegają na korzystaniu z zewnętrznych serwerów, do których dostęp uzyskuje się przez Internet z dowolnego miejsca i urządzenia. Rozwiązanie to ułatwia szybkie udzielanie dostępu do plików większej grupie osób, a także zapisywanie zmian i śledzenie aktywności poszczególnych użytkowników. Korzystanie z takich rozwiązań wymaga jednak analizy uwzględniającej przede wszystkim kwestie bezpieczeństwa. Warto pamiętać, że dane przechowywane w chmurze w istocie znajdują się na serwerze zlokalizowanym w konkretnym miejscu, co może mieć szczególne znaczenie w kontekście przechowywania i przetwarzania danych osobowych.

Wiele organizacji prowadzących badania naukowe korzysta z komercyjnych usług, które można wykorzystać nie tylko do przechowywania danych, ale także do bieżącej komunikacji, dydaktyki, spotkań online czy działań promocyjnych. Polityka instytucji może uwzględniać zalecane rozwiązania w odniesieniu do konkretnych celów, a odpowiednie działy wsparcia IT dostarczają informacji na temat warunków i ram korzystania z konkretnych usług.

Chmura np. Google Drive, OneDrive, Dropbox, instytucjonalne ownCloud, Nextcloud, Open Science Framework, Tresorit		
Zalety	**Wady**	**Środki ostrożności dot. (wrażliwych) danych osobowych**
Automatyczne kopie zapasowe. Często także automatyczna kontrola wersji.	Nie wszystkie usługi są bezpieczne. Mogą nie być odpowiednie dla danych wrażliwych zawierających dane osobowe obywateli UE. Niewystarczająca kontrola nad miejscem przechowywania danych i częstotliwością tworzenia kopii zapasowych. Bezpłatne usługi dostawców komercyjnych (np. Dysk Google, Dropbox) mogą rościć sobie prawa do wykorzystywania treści, którymi zarządzasz, i udostępniania ich do własnych celów. Dane mogą zostać utracone, jeśli konto zostanie zawieszone lub przypadkowo usunięte, lub jeśli dostawca zakończy działalność.	Należy zaszyfrować wszystkie (wrażliwe) dane osobowe przed przesłaniem ich do chmury. Jest to szczególnie ważne, jeśli nie wiesz, w których krajach znajdują się serwery używane do przechowywania i tworzenia kopii zapasowych. Należy pozostać w zgodzie z europejskimi przepisami dotyczącymi ochrony danych. (Szyfrowanie powinno być na tyle silne, aby uniemożliwić odczytanie danych, a klucz należy przechowywać w bezpieczny sposób w Europejskim Obszarze Gospodarczym).

Rekomendacje

Korzystaj z usług w chmurze, aby zapewnić zdalny i łatwy dostęp do danych i innych informacji wszystkim osobom zaangażowanym w projekt.
Przeczytaj warunki korzystania z usługi. Szczególną uwagę należy zwrócić na prawa do korzystania z treści przysługujące usługodawcy.
Wybierz – jeśli to możliwe – europejskie, krajowe lub instytucjonalne usługi w chmurze, które przechowują dane w Europie, np.:

- B2DROP (EUDAT) to przykład europejskiego rozwiązania do przechowywania danych w chmurze;
- SWITCHdrive (SWITCH) to szwajcarskie rozwiązanie;
- DataverseNL (Data Archiving and Networked Services) to przykład usługi dla holenderskich naukowców, która umożliwia przechowywanie i udostępnianie danych zarówno podczas realizacji projektu, jak i po jego zakończeniu.

Nie traktuj usług w chmurze jako jedynego rozwiązania służącego do przechowywania i tworzenia kopii zapasowych.
Nie używaj do niezaszyfrowanych (wrażliwych) danych osobowych.

Komputery

Służbowe komputery i laptopy są stosowane zarówno do bieżącej pracy nad danymi, np. z wykorzystaniem odpowiedniego oprogramowania, jak i do przechowywania danych, w szczególności tymczasowych kopii roboczych. Co do zasady komputery i laptopy nie powinny być jedynym ani podstawowym miejscem przechowywania danych, a pliki stanowiące rezultaty bieżących prac należy kopiować i przenosić w inne bezpieczne miejsca. Można to robić po zakończeniu konkretnego etapu prac lub w terminach określonych w zasadach tworzenia kopii zapasowych, dostosowanych do specyfiki projektu.

Urządzenia lokalne Komputery i laptopy		
Zalety	**Wady**	**Środki ostrożności dot. (wrażliwych) danych osobowych**
Pełna kontrola nad plikami. Łatwa ochrona przed nieautoryzowanym dostępem.	Jeśli dane i pliki są przechowywane tylko na jednym urządzeniu, są narażone na utratę, np. w wyniku awarii, kradzieży, nadpisania lub usunięcia plików z powodu błędu człowieka. Dostęp do danych i plików ma tylko osoba dysponująca komputerem.	Należy zabezpieczyć komputer hasłem i rozważyć zaszyfrowanie dysku twardego.

Rekomendacje

Korzystanie z komputerów stacjonarnych i laptopów osobistych jako podstawowego sposobu przechowywania i uzyskiwania dostępu do danych i plików jest odpowiednie tylko w przypadku projektów, w których bierze udział bardzo niewiele osób (najlepiej tylko jedna) i w których dane i pliki nie będą musiały być często przenoszone między komputerami osobistymi. Jeśli planujesz pracować na danych na różnych (lokalnych) komputerach, np. na laptopie w domu i komputerze stacjonarnym w biurze:

- upewnij się, że zawsze pracujesz na najbardziej aktualnej wersji plików, np. za pomocą oprogramowania do wersjonowania lub zgodnie z przyjętymi zasadami kontroli wersji;
- upewnij się, że zawsze tworzona jest kopia zapasowa najnowszej wersji.

Dyski sieciowe

Dane mogą być przechowywane z dużym powodzeniem również na podłączonych bezpośrednio do sieci komputerowej dyskach znajdujących się np. w rozproszonych jednostkach organizacji. Dzięki możliwości zastosowania odpowiedniego oprogramowania mogą być to zarówno pojedyncze dyski, jak i systemy dyskowe udostępnione na komputerach osobistych lub serwerach.

Dyski sieciowe Dyski współdzielone na serwerach instytucji naukowej lub serwerach NAS (Network Attached Storage)		
Zalety	**Wady**	**Środki ostrożności dotyczące (wrażliwych) danych osobowych**
Centralne przechowywanie danych. Dostęp współdzielony, możliwy zdalny dostęp dla wszystkich osób zaangażowanych w projekt. Możliwość centralnego zarządzania kopiami zapasowymi i ich automatycznego tworzenia.	Wymagane są wyższe środki bezpieczeństwa, aby zapobiec nieautoryzowanemu dostępowi oraz przypadkowemu usunięciu lub manipulacji danymi. Dostęp dla zewnętrznych partnerów projektu może być utrudniony lub niemożliwy. Wyższy koszt.	Należy korzystać w połączeniu z odpowiednią strategią bezpieczeństwa w celu ochrony danych przed nieautoryzowanym dostępem.

Rekomendacje

Używaj w projektach realizowanych we współpracy z wieloma osobami, które potrzebują dostępu do danych.
Używaj w połączeniu z odpowiednią strategią bezpieczeństwa w celu ochrony danych przed nieautoryzowanym dostępem.
Używaj w połączeniu ze ścisłymi zasadami wersjonowania.
Zaplanuj długoterminowe archiwizowanie danych, które są kompletne i zostały przeanalizowane. W ten sposób może zostać zwolniona cenna przestrzeń dyskowa.
Korzystaj z kontroli dostępu i uprawnień, aby upewnić się, że nie każdy ma dostęp do wszystkiego, jeśli nie jest to konieczne (np. dostęp do plików głównych jest bardziej ograniczony niż dostęp do plików roboczych).

Tabele opracowane na podstawie przewodnika CESSDA Data Management Expert Guide zamieszczonego na stronie: https://dmeg.cessda.eu/Data-Management-Expert-Guide/4.-Store/Storage [data dostępu: 17.07.2023].

Przechowywanie danych podczas realizacji projektu

Przechowywanie danych podczas realizacji projektu powinno uwzględniać bieżące potrzeby związane m.in. z warunkami zbierania danych (np. pracą w terenie, wykorzystaniem określonej aparatury lub sprzętu), z opracowywaniem i analizowaniem danych (np. współpracą z innymi członkami zespołu) czy z ochroną szczególnego typu danych (np. danych osobowych, danych osobowych wrażliwych, danych objętych klauzulą poufności). Ogólne zasady przechowywania danych mają na celu zminimalizowanie ryzyka związanego z utratą, uszkodzeniem lub nieuzasadnioną zmianą danych. Chodzi w szczególności o zapobieganie sytuacjom takim jak awaria, zniszczenie czy utrata sprzętu, które mogą poważnie zagrozić realizacji projektu. Powtórne zebranie utraconych danych często nie jest bowiem możliwe z uwagi na specyfikę badań albo – jeżeli jest możliwe – będzie wymagało dodatkowych środków finansowych i czasu.

Inne ryzyko może wiązać się z przedwczesnym lub nieplanowanym udostępnieniem danych, np. w wyniku uzyskania dostępu do danych przez osoby pozbawione odpowiednich uprawnień. Może chodzić tutaj zarówno o dane, które nie mogą zostać udostępnione z powodów prawnych lub etycznych, jak również o dane, które w dalszej perspektywie mają zostać udostępnione. Chociaż jedną z praktyk otwartej nauki jest prowadzenie otwartych notatników badawczych umożliwiających dzielenie się wynikami badań jak najszybciej, niemalże w chwili ich uzyskania, wciąż powszechne jest priorytetowe traktowanie publikacji naukowej i nieudostępnianie danych przed publikacją.

Strategia przechowywania danych badawczych wymaga określenia miejsc przechowywania danych oraz procedur związanych z kopiowaniem, modyfikowaniem, wersjonowaniem, usuwaniem, a także udzielaniem dostępu do danych. Może ona obejmować ponadto ustalenie różnych poziomów ochrony w zależności od możliwych ryzyk związanych z ujawnieniem, uszkodzeniem czy utratą danych. Wiele instytucji prowadzących badania naukowe opracowuje i wdraża polityki polegające na klasyfikacji danych pod tym kątem[16].

---

[16] Por. University of Cambridge, https://help.uis.cam.ac.uk/service/security/data-sec-classes; Harvard University, https://security.harvard.edu/data-security-levels-research-data-examples University of California, https://security.berkeley.edu/resources/how-classify-research-data [data dostępu: 17.07.2023].

Przygotowując plan zarządzania danymi badawczymi, w pierwszej kolejności należy zapoznać się z wytycznymi instytucji zatrudniającej badaczy, która może rozwijać i udostępniać narzędzia służące do przechowywania danych, formułować w tym zakresie zalecenia bądź oferować wsparcie. Organizacje prowadzące badania naukowe mogą także korzystać z komercyjnych usług, które w takiej sytuacji są zwykle jednym z rekomendowanych rozwiązań. Jeżeli w ramach realizacji projektu rozważane jest korzystanie z innych usług o charakterze komercyjnym, należy zapoznać się z zasadami i warunkami przyjętymi przez ich dostawcę, przeanalizować obowiązujące badaczy regulacje oraz wziąć pod uwagę ewentualne ryzyka.

Bezpieczeństwo w bieżącej pracy z danymi

Surowe dane

Surowe dane, stanowiące podstawę wszelkich dalszych prac, powinny podlegać szczególnej ochronie. Należy je przechowywać w oddzielnej lokalizacji i zabezpieczyć przed zmianami, aby nie zostały nadpisane, zmienione czy skasowane; można skorzystać z ustawień plików „tylko do odczytu”. Dalsze prace należy prowadzić na kopiach, dokumentując kolejne etapy badań, w tym procedury i metody.

Kontrola wersji

W toku realizacji projektu powstają często kolejne wersje plików z danymi, które mogą być przetwarzane i wzbogacane o inne dane bądź poddawane analizom zgodnie z przyjętymi założeniami metodologicznymi. Kontrola wersji pełni istotną funkcję w zapewnieniu bezpieczeństwa plików, ponieważ chroni je przed usunięciem i nadpisaniem, a także zapewnia integralność danych. Dobrym rozwiązaniem może być osobne przechowywanie kopii głównych plików i tymczasowych kopii roboczych oraz przyjęcie ścisłych zasad wersjonowania i synchronizowania plików w różnych lokalizacjach.

Kopie zapasowe

Regularne tworzenie kopii zapasowych jest dobrą praktyką, pozwalającą zapobiegać utracie danych. Standardowym i rekomendowanym rozwiązaniem jest zasada 3-2-1: należy przechowywać trzy kopie plików na dwóch różnych nośnikach, w tym jeden w innej lokalizacji geograficznej. Zarówno stworzenie planu zarządzania danymi badawczymi, jak i jego późniejsza realizacja wymaga odpowiedzi na następujące pytania związane z bezpieczeństwem danych:

- W jaki sposób tworzone będą kopie zapasowe?
- Gdzie przechowywane będą kopie zapasowe?
- Jak często tworzone będą kopie zapasowe danych?
- Ile kopii będzie tworzonych?
- W jaki sposób dane zostaną odzyskane w wypadku ich utraty?

- Kto będzie odpowiedzialny za tworzenie kopii zapasowych i odzyskiwanie danych?

Szyfrowanie danych

Jednym ze środków zabezpieczających przed niepożądanym ujawnieniem danych jest ich szyfrowanie. Należy stosować bezpieczne algorytmy z kluczem publicznym (klucz szyfrujący – publiczny jest inny niż klucz deszyfrujący – prywatny), przy czym należy pamiętać o przechowywaniu klucza prywatnego w bezpiecznym miejscu, niedostępnym dla osób nieuprawnionych. Dane można również umieszczać na uprzednio zaszyfrowanych partycjach dyskowych, przesyłając je szyfrowanymi protokołami komunikacji sieciowej. Zarówno szyfrowanie danych, jak i bezpieczna komunikacja sieciowa powinny być realizowane przez specjalistyczne oprogramowanie, najlepiej wybrane przez dział IT instytucji naukowej w zależności od jej możliwości operacyjnych.

Długoterminowe przechowywanie danych

Określenie zasad długoterminowego przechowywania danych wymaga odpowiedzi na poniższe pytania:

- Jakie dane powinny być zachowane po zakończeniu projektu?
- Z jakich powodów należy je zachować?
- Jak długo, gdzie i w jaki sposób należy je przechowywać?

Przechowywanie danych po zakończeniu realizacji projektu powinno uwzględniać zobowiązania wynikające z umów

grantowych bądź polityk instytucji prowadzących i finansujących badania, a także dobre praktyki i standardy przyjęte w konkretnej dziedzinie lub obszarze badań. Szczególne znaczenie w tym kontekście mają dane stanowiące podstawę publikacji, potrzebne do weryfikacji zawartych w nich twierdzeń, jednak wymogi mogą obejmować także inne dane badawcze oraz dokumenty związane z realizacją projektu. Zgodnie z wytycznymi Narodowego Centrum Nauki dotyczącymi planu zarządzania danymi badawczymi dane surowe i przetworzone powinny być przechowywane przez okres odpowiedni dla danej dyscypliny i zastosowanej metodologii. W rozumieniu NCN uzasadniony okres przechowywania danych to minimum 10 lat.

Taka perspektywa wiąże się z koniecznością uwzględnienia zachodzącego z czasem procesu degradacji danych (*data rot*, *bit rot*) oraz ryzyka wyjścia z użytku określonych nośników danych, formatów plików czy oprogramowania służącego do ich odczytywania. Z tych powodów długoterminowe przechowywanie danych nie może ograniczać się do zwykłego składowania danych. Zapewnienie bezpieczeństwa i integralności danych wymaga zaplanowanych i systematycznych działań, które wiążą się z konkretnymi kosztami. W zakresie danych przeznaczonych do udostępnienia z pomocą przychodzą repozytoria danych, które posiadają i realizują własne polityki długoterminowego przechowywania danych, uwzględniające powyższe czynniki; w szczególności mogą rekomendować deponowanie plików w określonych formatach oraz regularnie sprawdzać sumy kontrolne i w razie niezgodności – odzyskiwać kopie zapasowe przechowywane w innej lokalizacji.

Własne strategie i rozwiązania techniczne zapewniać powinny także instytucje prowadzące badania, zwłaszcza w zakresie danych, które nie są przeznaczone do udostępniania i muszą być objęte szczególną ochroną, np. danych osobowych.

Długoterminowe przechowywanie danych wiąże się także z selekcją danych. Zachowanie wszystkich danych może być niemożliwe ze względów finansowych. Ilość wytwarzanych i zbieranych danych stale wzrasta, co przekłada się na coraz wyższe koszty ich przechowywania, tworzenia kopii zapasowych i prowadzenia aktywnej polityki zapewniającej bezpieczeństwo danych.

Oprócz wymogów instytucji finansujących i prowadzących badania naukowe pod uwagę należy wziąć także inne kryteria. Przy selekcji danych przeznaczonych do długoterminowego przechowywania warto odpowiedzieć na poniższe pytania:

- Czy dane mają szczególną wartość naukową bądź historyczną, wynikającą np. z zebrania ich w wyjątkowych okolicznościach? Czy aktualny stan wiedzy pozwala sądzić, że dane będą miały znaczenie w przyszłości?
- Czy dane są unikatowe?
- Czy dane mają potencjał w zakresie ponownego wykorzystania? Czy mogą skorzystać z nich inni badacze lub osoby spoza środowiska naukowego?
- Czy dane są wysokiej jakości? Czy są dobrze udokumentowane i opisane?

Przygotowanie danych do udostępnienia

Repozytoria danych badawczych, niekiedy zwane również archiwami danych, są miejscem pierwszego wyboru dla udostępnienia danych. Jeśli w ramach planu zarządzania danymi przewidujemy inny sposób udostępniania naszych danych, decyzja taka powinna zostać w tym dokumencie odpowiednio uzasadniona.

Najważniejsza funkcja repozytoriów polega na tym, iż gromadzą one dane badawcze (z określonej dziedziny lub z różnych dziedzin), umożliwiając dostęp do nich i ich ponowne wykorzystanie. Dane zdeponowane w repozytorium cechuje dostępność długoterminowa, co znaczy, iż będą one bezpiecznie przechowywane przez długi czas (najlepiej wieczyście), ale też –

w wypadku wybranych repozytoriów – że podejmowane będą dodatkowe działania (np. konwersja na nowe formaty) mające na celu zapewnienie dostępności danych również po wielu latach.

Repozytoria zwykle zapewniają przypisanie do zbiorów danych trwałych identyfikatorów i umożliwiają szeroką wymianę możliwie precyzyjnych metadanych z innymi systemami (takimi jak wyszukiwarki czy agregatory) ułatwiającymi odnalezienie informacji o znajdujących się w nich zasobach.

Wiele repozytoriów umożliwia również śledzenie informacji dotyczących wykorzystania zbiorów danych (takich jak np. cytowania) oraz ich powiązań z innymi rezultatami pracy badawczej, takimi jak artykuły, monografie czy inne zbiory danych.

Repozytoria oferują też szereg funkcji możliwych do wykorzystania w sytuacji, w której z określonych powodów decydujemy się na opóźnienie dostępności danych (embargo) lub ich udostępnienie określonym osobom (np. recenzentom procedowanej równolegle publikacji) jeszcze przed ich opublikowaniem.

Wybór danych do udostępniania

Jeżeli w ramach projektu badawczego powstały duże ilości danych, możemy stanąć przed dylematem, które z nich – w obliczu ograniczonego budżetu, jaki jest do naszej dyspozycji –

udostępnić szerokiemu gronu odbiorców. Aby to ustalić, na początek powinniśmy zadać sobie trzy pytania:

- Do udostępnienia jakich danych jesteśmy zobowiązani? Źródłem takich zobowiązań najczęściej będą polityki instytucji naukowej, instytucji finansującej lub czasopisma, regulujące kwestie zarządzania danymi badawczymi. Do udostępnienia określonych zasobów mogliśmy też zobowiązać się w planie zarządzania danymi.
- Jakich danych udostępnić nam nie wolno? Ograniczenia w tym zakresie mogą wynikać np. z obowiązujących przepisów prawa powszechnego lub z zawartej przez nas umowy.
- Na udostępnienie jakich danych nas nie stać? Jeśli odpowiednie opracowanie i udokumentowanie danych wykracza poza nasz budżet, siłą rzeczy takich danych nie udostępnimy.

Odpowiedzi na powyższe trzy pytania mogą pomóc w rozstrzygnięciu dylematu dotyczącego tego, jakie dane udostępniać, a jakie nie, w odniesieniu do części z nich. Trzeba jednak pamiętać, że odpowiedzi te powstają zawsze w konkretnym momencie, a uwzględnione przy ich udzielaniu okoliczności mogą ulec zmianie. Prawo może zostać zmienione, umowa może zostać aneksowana, a w przyszłości mogą też pojawić się dodatkowe możliwości pozyskania środków, które będzie można przeznaczyć na opracowanie danych. Sprawia to, że te dane, których w aktualnej sytuacji nie możemy udostępnić, warto traktować jak zasób, który być może będziemy mieli okazję

udostępnić w przyszłości i w związku z tym zachować tak same dane, jak i ich istniejącą dokumentację.

W odniesieniu do tych danych, w wypadku których odpowiedzi na powyższe pytania nie wystarczą do podjęcia decyzji o ich (nie)udostępnianiu, można posłużyć się kilkoma dodatkowymi kryteriami:

- Po pierwsze, powinniśmy zadać sobie pytanie o wartość naukową lub historyczną zebranych przez nas danych. Jeśli istnieją powody, by sądzić, iż mamy do czynienia z danymi o wysokiej wartości, z pewnością warto je udostępnić.
- Po drugie, powinniśmy zastanowić się, jak unikatowe są nasze dane. Jeśli dane podobne do naszych są nieliczne lub nie istnieją, jest to bardzo dobry powód, by je udostępnić.
- Po trzecie, powinniśmy rozważyć, czy istnieje możliwość ponownego zebrania lub wytworzenia takich samych lub zbliżonych danych. Jeśli nie, udostępnienie naszych danych może być jedynym sposobem umożliwienia przeprowadzenia jakichś analiz w przyszłości.
- Po czwarte, powinniśmy rozważyć to, jaki był koszt wytworzenia lub zebrania naszych danych. Jeśli był on wysoki, a dzięki udostępnieniu danych będzie można uniknąć konieczności ich gromadzenia po raz kolejny, stanowi to doskonały powód, by udostępnić dane.

Formaty, dokumentacja i opis danych

Po ustaleniu, jakie dane zamierzamy udostępnić, należy zadbać o nadanie im określonego formatu oraz o ich odpowiednią dokumentację i opis. Na tym etapie należy sprawdzić, czy repozytorium, na które się zdecydowaliśmy, nie ma w tym zakresie określonych preferencji, oczekiwań lub nawet wymagań.

Wiele repozytoriów preferuje np. otwarte formaty danych, w wypadku których nie występują zwykle trudności z ich otwarciem i/lub analizą za pomocą darmowego oprogramowania. Przy wykorzystaniu formatów tego rodzaju, ze względu na otwarty charakter ich dokumentacji, nawet po wielu latach nie powinno również być problemu z ich przekonwertowaniem na inne, nowe formaty – również takie, które obecnie jeszcze nie istnieją. Warto zaznaczyć, że udostępnianie danych w formatach otwartych nie wyklucza bieżącej pracy z wykorzystaniem – często bardzo popularnych – formatów zamkniętych. Istotne jest to, aby w momencie przygotowania danych do udostępnienia przekonwertować je na formaty otwarte i udostępnić w takiej postaci. Niektóre repozytoria dopuszczają nawet możliwość zdeponowania danych w dwóch formatach: zamkniętym (ale często popularnym) oraz otwartym. Takie rozwiązanie ułatwia wykorzystanie danych zarówno tym osobom, które preferują popularne obecnie formaty zamknięte i dysponują odpowiednim oprogramowaniem umożliwiającym ich analizę, jak i osobom, którym w przyszłości może być łatwiej prowadzić analizy w oparciu o formaty otwarte.

Należy przy tym zaznaczyć, że na co dzień pracować możemy z takimi formatami, jakie najlepiej pasują do naszych potrzeb, nawet jeśli są to formaty zamknięte. Ważne natomiast, by w momencie przygotowywania danych do udostępnienia zadbać o dostępność formatów otwartych i dokonać odpowiednich konwersji.

W odniesieniu do obrazów, nagrań audio oraz wideo dodatkowo należy preferować te formaty, które nie wymuszają stratnej kompresji. W momencie archiwizacji powinniśmy dbać przede wszystkim o jakość udostępnianych materiałów, stratna kompresja wpłynie zaś na nią w negatywny sposób. Zwłaszcza w wypadku materiałów wideo brak kompresji może jednak sprawić trudności innego rodzaju: nieskompresowany materiał może być bardzo duży, a przez to trudny do pobrania przez osoby dysponujące słabszym łączem. W takiej sytuacji dobrym pomysłem może być zdeponowanie danych w dwóch wersjach – skompresowanej oraz nieskompresowanej. Dodanie tej pierwszej ułatwi dostęp do danych wszystkim tym, którzy mogliby mieć trudność z pobraniem wersji nieskompresowanej. O wersji skompresowanej można też myśleć jako o swoistej próbce, która pozwoli zainteresowanym na wstępną selekcję danych, których bardziej szczegółowa analiza odbywałaby się już przy użyciu nieskompresowanych plików.

Na szczególna uwagę zasługuje format kompresji archiwów ZIP. Choć nie brak formatów przewyższających go pod pewnymi względami (np. 7-Zip czy RAR charakteryzują się znacznie większym stopniem kompresji), to jeśli decydujemy się na

zgromadzenie większej liczby plików w obrębie jednego lub wielu plików archiwum, warto zdecydować się właśnie na format ZIP. Przemawia za nim przede wszystkim to, iż jest dostępny niemal na każdym komputerze bez konieczności instalowania dodatkowego oprogramowania. Dzięki temu niezależnie od tego, czy osoba zainteresowana naszymi danymi będzie je otwierać na swoim prywatnym komputerze czy też na komputerze w miejscu pracy lub bibliotece (gdzie może nie mieć możliwości instalowania dodatkowego oprogramowania), będzie w stanie to zrobić i sprawdzić ich zawartość.

Deponowanym danym powinny również towarzyszyć odpowiednie metadane, czyli dane opisujące dane. Jeśli zdecydowaliśmy się na wybór repozytorium specjalistycznego lub dziedzinowego, serwis taki prawdopodobnie dobrał już odpowiedni dla nas standard metadanych.

Zbiór danych powinna również uzupełniać dokumentacja zawierająca wszystkie informacje niezbędne do zrozumienia i właściwej interpretacji udostępnianych danych. Dobrą praktyką w tym zakresie jest umieszczanie w zbiorze danych dodatkowego pliku README.txt, zawierającego podstawowe informacje dotyczące udostępnianych danych.

Jeśli udostępniane dane nie są kompletne – np. w związku z umownymi ograniczeniami nie było możliwe udostępnienie jakiegoś ich fragmentu – w dokumentacji należy zawrzeć informację dotyczącą tego, co i z jakich powodów nie znalazło się w dostępnym zbiorze danych.

Z kolei jeśli mamy dobre racje ku temu, by jakichś danych nie udostępniać w ogóle, warto skorzystać z możliwości oferowanej przez niektóre repozytoria i utworzyć tzw. ciemny depozyt (*dark deposit*). Jest to rekord w repozytorium pozbawiony plików z danymi, ale posiadający metadane. Rekord taki posiada zwykle własny trwały identyfikator, co umożliwia jego poprawne zacytowanie. Dzięki temu informacja o danych (choć nie one same) trafia do szerokiego obiegu.

Trwałe identyfikatory (Persistent Identifiers) i zasady ich stosowania

Rodzaje trwałych identyfikatorów

Historia trwałych identyfikatorów sięga wielu lat wstecz. Trwałe identyfikatory nie potrzebują w zasadzie ani komputerów, ani łączącej je sieci Internet. Trwałymi identyfikatorami są również jak najbardziej „analogowe" identyfikatory, takie jak znane i popularne do dzisiaj identyfikatory ISBN i ISSN. Trwały identyfikator to bowiem nic innego jak długotrwałe odniesienie do jakiegoś obiektu. W wypadku numerów ISBN i ISSN tym, do czego odnoszą się te identyfikatory będą – w dużym uproszczeniu – książki oraz czasopisma.

Osobną grupę trwałych identyfikatorów stanowią identyfikatory cyfrowe, którym towarzyszą pewne dodatkowe usługi cyfrowe i które umożliwiają określone działania nie tylko ludziom, ale i maszynom. Cyfrowymi identyfikatorami będą więc np. identyfikator ORCID, odnoszący się do naukowców, czy też

ROR, odnoszący się do instytucji naukowych. W obu wypadkach identyfikatorom tym towarzyszą serwisy zawierające informacje o – odpowiednio – danym naukowcu (wraz z jego bibliografią) oraz instytucji (wraz z informacją o jej alternatywnych nazwach oraz innych identyfikatorach). Informacje takie może uzyskać człowiek, ale też – dzięki odpowiednim endpointom API – również maszyny.

Należy przy tym zwrócić uwagę na dwie kwestie. Po pierwsze, trwałość identyfikatora jest czymś, co w praktyce zależy od tego, czy istnieje ośrodek gromadzący informacje o identyfikatorze oraz obiekcie, którego on dotyczy, i umożliwiający aktywne zarządzanie tymi informacjami (np. aktualizację metadanych lub powiązanego odniesieniem identyfikatora adresu URL, jeśli te ulegną zmianie). O trwałych identyfikatorach można więc myśleć w praktyce jak o czymś długotrwałym, przy czym owa długotrwałość wymaga odpowiedniej dbałości po stronie odpowiedzialnych za nie podmiotów. Po drugie, identyfikator może, ale nie musi odnosić się do obiektu cyfrowego. Innymi słowy, cyfrowy jest identyfikator, ale niekoniecznie jego odniesienie. Często obiekty oznaczane rzeczywiście będą mieć charakter cyfrowy, jednak – jak pokazują przykłady cyfrowych identyfikatorów ROR oraz ORCID – cyfrowe identyfikatory mogą również oznaczać instytucje czy osoby.

W świecie nauki problemy z trwałością odniesień stały się dobrze widoczne wraz z nastaniem ery Internetu. W miarę jak w globalnej sieci zaczęło pojawiać się coraz więcej treści ważnych z perspektywy naukowej i w miarę jak treści te zaczęły być coraz

częściej cytowane, badacze zaczęli też napotykać problem niedziałających odnośników. Przykładem mogą tu być adresy URL, pod którymi kiedyś znajdowały się wprawdzie jakieś zasoby, ale które po pewnym czasie przestawały działać np. z powodu przeniesienia zasobu w inne miejsce czy też czysto technicznej zmiany struktury adresów na stronie. Problem ten dotyczył treści zamieszczanych bezpośrednio na stronach internetowych, ale też elektronicznych wersji artykułów naukowych czy linków do plików zawierających dane badawcze. Zasób plikowy mógł zostać usunięty lub przeniesiony, czasopismo mogło zmienić wydawcę, a co za tym idzie, zostać udostępnione w nowym miejscu – tego typu okoliczności mogły spowodować, że adresy URL (zawarte często w setkach, a nawet tysiącach przypisów i pozycji bibliograficznych) mogły w jednej chwili przestać działać.

Problemy te rozwiązują właśnie cyfrowe trwałe identyfikatory. Umożliwiają one trwałe powiązanie ciągu znaków z zasobem niezależnie od tego, pod jakim adresem URL taki zasób aktualnie jest dostępny. Jeśli zasób zmienia swoją lokalizację wyrażoną adresem URL, wystarczy adres taki zaktualizować w serwisie powiązanym z cyfrowym identyfikatorem.

W praktyce z punktu widzenia danych badawczych ważne są dwa identyfikatory: handle oraz DOI.

Handle to rodzaj trwałego identyfikatora dla zasobów internetowych. Posiada on centralny rejestr pozwalający na rozwiązywanie adresów URL opartych na identyfikatorach handle

w taki sposób, by prowadziły one do aktualnej lokalizacji danego zasobu. Nad globalnym rejestrem niezbędnym do poprawnego funkcjonowania identyfikatora handle pieczę sprawuje fundacja DONA. Z identyfikatora handle korzystają np. niektóre repozytoria działające na bazie oprogramowania Dataverse (np. https://data.cimmyt.org/).

Na systemie identyfikatora handle nabudowany jest z kolei najpopularniejszy w świecie naukowym identyfikator, a mianowicie DOI (Digital Object Identifier). Na identyfikator DOI składają się trzy elementy:

- nazwa, czyli unikalny, zbudowany według określonych zasad ciąg znaków,
- metadane, czyli informacje o obiekcie, do którego odnosi się dany identyfikator,
- adres URL kierujący do strony z informacjami o tym obiekcie.

Pieczę nad tym identyfikatorem sprawuje z kolei DOI Foundation. Same identyfikatory są nadawane przez agencje rejestrujące. Z perspektywy akademickiej najważniejsze z nich to Crossref (z grubsza rzecz biorąc odpowiadający za rejestrację numerów DOI dla publikacji naukowych) oraz DataCite (odpowiadający za rejestrację identyfikatorów dla zbiorów danych badawczych).

Wzbogacenie systemu handle o metadane pozwala też na tworzenie i udostępnianie na ich bazie dodatkowych serwisów i usług. Infrastruktura DataCite umożliwia np. wyszukanie zbiorów, które otrzymały numery DOI. Dzięki wymianie metadanych

z ekosystemem Crossref można uzyskać informacje o artykułach naukowych cytujących dany zbiór.

Rejestracja numeru DOI dla zbioru danych wymaga dostarczenia do infrastruktury DataCite metadanych opisujących zbiór danych w formacie zgodnym ze standardem opracowanym i stale aktualizowanym właśnie przez DataCite. Metadane te wykorzystują z kolei inne identyfikatory, np. zawierając w nich informację o autorach zbiorów, możemy też wskazać identyfikator ORCID każdego z nich. Opisując afiliacje albo wprowadzając informację o instytucji finansującej, możemy posłużyć się identyfikatorem ROR. Z kolei opisując relacje łączące deponowany zbiór z innymi zbiorami danych i publikacjami naukowymi, możemy posłużyć się numerami DOI tych powiązanych zasobów.

Jak korzystać z trwałych identyfikatorów

Uzyskanie identyfikatora DOI dla zbioru danych z perspektywy naukowca w praktyce sprowadza się do trafnego wyboru repozytorium, które powinno zatroszczyć się o resztę. Naukowiec nie musi o nic wnioskować, a cały proces sprowadza się do krótkiej „rozmowy" serwera repozytorium z serwerem agencji rejestrującej. W jej trakcie serwer repozytorium przekazuje metadane zbioru danych (opracowane uprzednio przez osobę deponującą i/lub personel repozytorium) w formacie zrozumiałym dla serwera DataCite, który zachowuje je po swojej stronie, przypisuje numer DOI i przesyła do repozytorium informację o pomyślnym przebiegu całej operacji. Od tej chwili możemy

korzystać z dobrodziejstw wszystkich usług udostępnianych przez DataCite, w tym tych zintegrowanych w ramach serwisu https://commons.datacite.org/.

Z drugiej strony korzystanie z trwałych identyfikatorów oznacza ich właściwe cytowanie. Sam identyfikator może występować w dwóch postaciach:

- w postaci zwykłej, np. DOI: 10.18150/1234567,
- w postaci linku, np. https://doi.org/10.18150/1234567.

Większość stylów cytowania zaleca obecnie korzystanie z drugiej postaci, która umożliwia natychmiastowe przejście do strony z informacjami o zasobie (tzw. landing page publikacji czy zbioru). Użycie identyfikatora w tej postaci powoduje, że zbędne staje się umieszczanie w cytowaniu bezpośredniego adresu URL oraz informacji o dacie dostępu do zasobu (jako że numer DOI jest jego stałym identyfikatorem), ponieważ dostęp taki będzie dzięki identyfikatorowi zapewniony długoterminowo i niezależnie od adresu URL, pod jakim zlokalizowany jest obiekt w danym czasie.

Obecnie nie zaleca się natomiast stosowania w cytowaniach identyfikatora DOI w postaci wykorzystującej adres „dx.doi.org”.

Funkcje trwałych identyfikatorów

Trwały identyfikator daje nam przede wszystkim pewność co do tego, do czego lub do kogo się odnosi oraz rękojmię tego, że w przyszłości będzie to ten sam obiekt. Podobnych gwarancji nie

daje nam np. imię i nazwisko czy adres URL (pod którym jutro może się znajdować coś innego niż dzisiaj).

W konsekwencji unikamy problemu „martwych” odnośników w cytowaniach: wiemy, że będą one aktywne również w przyszłości. Dużo łatwiej jest nam również identyfikować cytowania do danego zasobu: o ile stylów cytowania jest wiele, o tyle ciąg znaków tworzący DOI jest identyczny w każdym z nich. Dzięki jego wyodrębnieniu możemy zaś bardzo łatwo ustalić, co zostało w danym miejscu zacytowane.

Identyfikatory wzbogacone odpowiednimi metadanymi dają nam również możliwość korzystania z usług opartych na takich metadanych. Możliwość taka dostępna jest również maszynom, które łącząc się z odpowiednim API, mogą na jego bazie oferować dodatkowe usługi.

Przykładowo, w ramach repozytorium RepOD na stronie zbioru danych wyświetlana jest informacja dotycząca liczby cytowań tego zbioru. Aby zaktualizować informacje o cytowaniach, repozytorium okresowo łączy się z odpowiednim API DataCite. Sam DataCite jest zaś w stanie dostarczać te informacje dzięki temu, że wymienia się metadanymi z infrastrukturą agencji Crossref, która posiada m.in. informacje o cytowaniach zawartych w publikacji opatrzonych numerem DOI.

Otwarte formaty

Jednym z powodów wielozadaniowości współczesnych komputerów jest to, że bity używane do przechowywania informacji mogą być interpretowane zasadniczo dowolnie, w zależności od kontekstu i użytego oprogramowania.

Ta sama zasada działa również w drugą stronę: te same dane możemy zapisywać za pomocą różnych ciągów bitów. W tym kontekście, w odniesieniu do pojedynczych znaków mówimy o ich „kodowaniu", a w odniesieniu do całych plików mówimy o „formatach danych".

Format pliku jest więc w najprostszym ujęciu konwencją dotyczącą interpretacji znaczenia bitów podczas ich zapisu na nośnik pamięci lub odczytu z niego. Niektóre z tych konwencji są publicznie dostępne i możliwe do użycia przez każdego – i wtedy, w przybliżeniu, mówimy o formacie otwartym. Wykorzystanie innych podlega różnego rodzaju ograniczeniom, z uwagi np. na niedostępność specyfikacji formatu (w tym objęcie tajemnicą przedsiębiorstwa), niewystarczające uwzględnienie przy jego projektowaniu potrzeb interoperacyjności czy konieczność uzyskania licencji na „własność intelektualną” związaną z danym formatem – i wtedy mówimy o formacie zamkniętym.

Poniżej przyjrzymy się wybranym definicjom formatów otwartych. Zidentyfikujemy też główne korzyści, jakie przynoszą formaty otwarte w zakresie interoperacyjności, długoterminowej

archiwizacji danych oraz ich ponownego wykorzystania, a także przedstawimy przykłady otwartych formatów.

Wybrane definicje formatów otwartych

Otwarte formaty często bywają definiowane na użytek dotyczących ich polityk, przyjmowanych przez różne podmioty. Polityki takie mogą np. obligować administrację publiczną danego państwa do używania wyłącznie otwartych formatów plików.

Open Government Directive (2009, USA)

Przyjęta w 2009 r. przez rząd Stanów Zjednoczonych Dyrektywa Otwartego Rządu[17] definiuje pojęcie „otwarty format" jako format:

- niezależny od platformy,
- odczytywalny maszynowo,
- udostępniony publicznie bez ograniczeń, które uniemożliwiałyby ponowne wykorzystanie informacji.

Poprzez „platformę" zazwyczaj rozumie się łącznie zarówno sprzęt, system operacyjny, jak i oprogramowanie.

[17] Open Government Directive, https://obamawhitehouse.archives.gov/open/documents/open-government-directive [data dostępu: 17.07.2023].

Open Standards Principles (2012, UK)

Z kolei przyjęte w 2012 r. przez rząd Wielkiej Brytanii Zasady Otwartych Standardów[18], zaktualizowane następnie w roku 2015 i 2018, rozwijają listę cech, jakie są wymagane od otwartych formatów, ujmowanych w szerszym kontekście otwartych standardów. Otwarty standard wybrany do użytku przez rząd musi:

- spełniać potrzeby użytkowników,
- zapewniać dostawcom równy dostęp do kontraktów rządowych,
- wspierać elastyczność i zmianę,
- wspierać zrównoważone koszty,
- być wybrany w oparciu o przemyślane decyzje,
- być wybrany, określony i wdrożony z wykorzystaniem uczciwych i transparentnych procesów.

Otwarte standardy powinny być wystarczająco dojrzałe i popularne, tzn. posiadać wielu użytkowników i odpowiednio duże zaplecze w postaci rozwijającej je społeczności.

The Linux Information Project (2004, Internet)

Definicja zaproponowana przez Projekt Informacyjny Linuxa[19] rozróżnia z kolei formaty otwarte i wolne: format otwarty to „format, który został opublikowany w taki sposób, aby każdy mógł

---

[18] Open Standards Principles, https://www.gov.uk/government/publications/open-standards-principles/open-standards-principles [data dostępu: 17.07.2023].

[19] Linux Information Project, http://linfo.org/, por. http://www.linfo.org/free_file_format.html [data dostępu: 17.07.2023].

go czytać i badać, ale który może być lub nie być chroniony patentami, prawami autorskimi lub innymi ograniczeniami dotyczącymi użytkowania". Natomiast format wolny to „format, który jest jednocześnie 1) opublikowany w taki sposób, aby każdy mógł go czytać i badać w całości oraz 2) nie jest objęty żadnymi prawami autorskimi, patentami, znakami towarowymi ani innymi ograniczeniami, dzięki czemu każdy może go używać bezpłatnie do dowolnego zamierzonego celu". Odpowiada to rozróżnieniu na otwarte i wolne oprogramowanie.

Zmiany formatów z zamkniętych na otwarte

Warto zauważyć, że przynależność danego formatu do grupy formatów otwartych lub zamkniętych może zmieniać się w czasie. Przykładowo format PDF zaczynał swe istnienie jako format zamknięty, kontrolowany przez firmę Adobe. Jednak w roku 2008, gdy specyfikacja PDF została zaakceptowana jako ISO 32000-1 przez Międzynarodową Organizację Normalizacyjną (ISO), stał się otwartym standardem. W wyniku tego procesu standaryzacji specyfikacja formatu PDF zaczęła być publicznie dostępna i niezależna od Adobe, co oznacza, że inni producenci oprogramowania mogą implementować obsługę plików PDF zgodnie ze specyfikacją ISO 32000-1. Należy jednak zauważyć, że niektóre części standardu PDF, takie jak Adobe XML Forms Architecture, pozostają pod pełną kontrolą firmy Adobe.

Bardzo istotnym z praktycznego punktu widzenia przypadkiem tego typu są formaty tworzone i utrzymywane przez firmę Microsoft w ramach pakietu oprogramowania MS Office.

Początkowo formaty te były zamknięte, co miało istotne konsekwencje rynkowe w związku z popularnością systemu operacyjnego Windows. W 2006 r. Microsoft opublikował jednak Microsoft Open Specification Promise (OSP), a razem z nią upublicznił specyfikacje niemal wszystkich swoich formatów (z wyjątkiem np. bazodanowych MDB i ACCDB). OSP nie jest formalną licencją, ale raczej obietnicą niepodejmowania kroków prawnych wobec podmiotów korzystających z tych wybranych formatów w ich niezmienionej postaci. Należy zauważyć, że implementacje w zmienionej postaci nie są objęte tą obietnicą. W rezultacie starsze formaty popularnych dokumentów Microsoft, takie jak DOC dla dokumentów tekstowych, XLS dla arkuszy kalkulacyjnych i PPT dla prezentacji, bywają uważane za „zamknięte, ale opublikowane". Natomiast nowsze wersje tych formatów, takie jak DOCX, XLSX i PPTX, są już uznawane za otwarte.

Zalety formatów otwartych

1. Interoperacyjność

Jak widzieliśmy na przykładzie Microsoftu i MS Office, cechą charakterystyczną formatów zamkniętych jest ich przynależność do konkretnego wytwórcy oprogramowania, który może być (i często jest) powiązany z jedną platformą sprzętową czy systemem operacyjnym – w wypadku formatów MS Office sprzed OSP był to system Windows. Natomiast z formatami otwartymi można pracować z wykorzystaniem różnego sprzętu, systemów operacyjnych i oprogramowania. Jest to ogromna zaleta

w kontekście zarządzania danymi badawczymi, ponieważ umożliwia ponowne wykorzystanie danych bez konieczności wykonywania konwersji danych, wymagającej często dużych nakładów czasu i zasobów.

2. Długoterminowa archiwizacja danych

Różnorodność platform staje się jeszcze istotniejszą okolicznością, jeśli uwzględnimy nie tylko aktualną sytuację, ale również jej zmianę w czasie. Ogólnodostępność specyfikacji formatów otwartych połączona z gwarancjami ich trwałości sprawia, że możemy być pewni, że w dającej się przewidzieć przyszłości formaty otwarte będą albo powszechnie używane, albo przynajmniej powszechnie zrozumiałe i możliwe do łatwej konwersji na otwarte standardy, które pojawią się w przyszłości. Natomiast wykorzystanie nieużywanych już formatów zamkniętych może albo być niemożliwe, albo wymagać dużego nakładu pracy i środków, a także posiadania zasobów niedostępnych dla indywidualnych badaczy czy zespołów badawczych.

3. Ponowne wykorzystanie danych

Wskazane wyżej zalety wspierają ponowne wykorzystanie danych. Im większa interoperacyjność zgromadzonych danych, tym więcej użytkowników może z nich korzystać na różnorodnych platformach. Im lepiej zadbano o długoterminową archiwizację, tym dłużej w przyszłości dane będą mogły być odczytywane, analizowane i wykorzystywane.

Rozwój technologii w warunkach rynkowych może jednak prowadzić do sytuacji, w której przynajmniej przez jakiś czas dany zamknięty format jest de facto standardem (jak w przywołanym wyżej przypadku formatu PDF). Zarządzając danymi badawczymi, warto konwertować formaty zamknięte na formaty otwarte, ale w niektórych wypadkach dobrą praktyką będzie udostępnienie zarówno formatów otwartych, jak i zamkniętych. Wiąże się to z tym, że osoby zainteresowane pracą z udostępnionymi danymi mogą mieć przygotowane procedury i narzędzia uwzględniające właśnie zamknięte – ale powszechne w danym obszarze – formaty. Udostępnienie obu rodzajów formatów może być pomocne w zapewnieniu optymalnego ponownego wykorzystania danych zarówno w krótkiej, jak i w dłuższej perspektywie.

Wybrane formaty wolne, otwarte i zamknięte

Nawiązując do przytoczonego wyżej rozróżnienia na formaty wolne, otwarte i zamknięte, przedstawiamy w poniższej tabeli wybrane przykłady tych formatów.

Rodzaj danych	Formaty wolne	Formaty otwarte, ale nie wolne	Formaty zamknięte
Tekst	• TXT • CSV • XML • HTML • LaTeX • Open Document (ODT, ODS, ODP)	• Office Open XML (DOCX, XSLX, PTTX) • RTF (chroniony prawem autorskim, dokumentacja dostępna publicznie, ale format pozostaje pod kontrolą Microsoft)	• MS Office (DOC, XLS, PPT)
Dźwięk	• MP3 (algorytmy używane do kompresji są przedmiotem patentów w niektórych krajach) • ALAC (w przeszłości zamknięty format Apple'a) • FLAC • WAV (podtyp RIFF)	• AAC (chroniony prawem autorskim, dokumentacja dostępna publicznie, ale format pozostaje pod kontrolą Via Licensing)	• WMA
Obraz	• PNG • SVG • JPEG (kompresja stratna) • GIF (w przeszłości chroniony patentami)	• PSD (natywny format Adobe Photoshop, dokumentacja dostępna publicznie) • SWF (grafika wektorowa)	• CDR (natywny format CorelDraw, dokumentacja niedostępna)
Film	• AV1	• MP4	• AMV

Archiwa (skompresowane lub nie)	• ZIP (starsze wersje w domenie publicznej) • GZIP (.gz) • 7Z • TAR	• ZIP (nowsze wersje chronione patentami)	• RAR
Inne	• DAE (modele 3D) • GLTF (modele 3D, występuje również w rozwinięciu GLB dla plików binarnych)	• GEDCOM (chroniony prawem autorskim, powszechnie używany do wymiany danych genealogicznych, dokumentacja publicznie dostępna)	• DWG (natywny format dla AutoCAD, popularnego programu używanego w projektowaniu wspomaganym komputerowo)

Metadane w kontekście standardów dyscyplinowych

Definicja, rodzaje i funkcje metadanych

Metadane – czyli dane o danych albo dane opisujące dane – stanowią bardzo ważny element zarządzania danymi badawczymi. Terminem pokrewnym do metadanych – niekiedy nawet traktowanym zamiennie – jest dokumentacja, rozumiana jako zbiór wszystkich tych informacji, które są niezbędne do poprawnego zinterpretowania danych oraz ich ponownego wykorzystania.

Termin „metadane” stosowany jest jednak zwykle w tych kontekstach, w których dane o danych są dobrze ustrukturyzowane

i maszynowo dostępne. Termin „dokumentacja” dotyczy natomiast tych przypadków opisu danych, w których zawarte w opisie informacje są w mniejszym stopniu ustrukturyzowane i przeznaczone przede wszystkim dla ludzi, a nie dla maszyn.

W tym rozumieniu plik README utworzony na bazie określonego szablonu będzie więc np. elementem dokumentacji, ale nie metadanych. Szablony takich plików sugerują bowiem jedynie dość ogólny sposób ustrukturyzowania informacji w ich obrębie, same pliki są zaś traktowane przez oprogramowanie repozytorium czy zewnętrzne serwisy tak samo lub bardzo podobnie jak inne pliki w zbiorze danych. W konsekwencji, jeśli plik README zawiera tytuł zbioru czy dane o jego autorach, informacje takie będą łatwe do zauważenia i przetworzenia dla człowieka, ale nie dla maszyn, np. zewnętrznych serwisów agregujących informacje o danych badawczych. Ta sama informacja o tytule czy autorach może też jednak stanowić element metadanych i w takich przypadkach będzie łatwo dostępna również dla komputerów.

Dokumentacja może być też zagnieżdżona w poszczególnych plikach z danymi, np. niektóre formaty plików tabelarycznych mogą pozwalać na zawarcie w nich informacji o typie danej zmiennej. W takim wypadku zewnętrznym serwisom również trudno będzie do niej dotrzeć. Ta sama informacja może jednak stanowić element metadanych w standardzie DDI, a wtedy maszyny nie powinny mieć problemu z jej wydobyciem.

Sposób, w jaki metadane są ustrukturyzowane, jest określony przez ich standard. Standard określa zarówno to, jakie informacje mogą zostać ujęte w jego obrębie, jak i wymagania dotyczące ich struktury. Może on np. określać, które elementy opisu metadanowego są obowiązkowe, a które opcjonalne, a także jaka powinna być ich liczebność.

Standard metadanych może przykładowo określać, że zbiór danych powinien mieć co najmniej jeden tytuł oraz co najmniej jednego autora, a autor taki może – choć nie musi – posiadać jakiś identyfikator. Identyfikator może z kolei być określonego typu, standard zaś określać będzie listę dopuszczalnych typów. Wśród nich może się znajdować np. identyfikator ORCID, zaś standard określać będzie, że dany autor może posiadać tylko jeden identyfikator ORCID – a jednocześnie dopuszczać, by autor posiadał kilka identyfikatorów różnych typów. Autor może też posiadać afiliację, a instytucja, przy której jest afiliowany – dokładnie jeden identyfikator ROR. Zbiór danych może być też związany z grantem (których może być wiele). Grant taki jest przyznany przez jakąś instytucję, która z kolei posiada nazwę oraz identyfikator ROR. Grant może mieć też – dokładnie jeden – tytuł oraz numer grantu.

W największym uproszczeniu schematy metadanych można podzielić na ogólne i dziedzinowe.

Ogólne informacje dotyczące zbioru danych, takie jak te wymienione powyżej, stanowią domenę ogólnych (generycznych) schematów metadanych. W wypadku danych badawczych

najważniejszym schematem tego rodzaju jest standardowy schemat DataCite.

Istotność tego schematu wynika z ogromnej popularności trwałego identyfikatora DOI w światowym ekosystemie danych badawczych. Aby uzyskać numer DOI, repozytorium musi dostarczyć do infrastruktury DataCite informacje w postaci zgodnej ze standardem określonym przez tę organizację. Innymi słowy, metadanowy opis każdego zbioru danych posiadającego numer DOI musi spełniać minimalne wymogi określone przez ten standard.

Wynika z tego jednak, że standard ten musi być ogólny, aby pozwalać na opisanie każdego zbioru danych. W konsekwencji zawarta w nim charakterystyka zbioru będzie mieć przede wszystkim charakter formalny: łatwo można w niej zawrzeć informacje o tytule, autorach, źródle finansowania czy identyfikatorach, ale jedynie w minimalnym stopniu i w słabo ustrukturyzowany sposób (np. za pośrednictwem pola „Opis") można uwzględnić w niej informacje o samych danych, celach i założeniach badania, w ramach którego powstały, czy wykorzystanych urządzeniach i oprogramowaniu.

Na uwzględnienie tego rodzaju szczegółowych informacji pozwalają z kolei dziedzinowe standardy metadanych, opracowane dla poszczególnych dyscyplin.

Standardy dyscyplinowe

To, jakie konkretnie informacje będą istotne w przypadku określonego zbioru danych, zależy w dużej mierze od dziedziny i dyscypliny, w ramach których został on wytworzony. Inne informacje będą istotne dla geologa, a inne dla językoznawcy czy socjologa; stąd właśnie bierze się znaczenie odpowiedniego doboru dyscyplinowych standardów metadanych. Standardy takie pozwalają opisać dane w sposób zarówno ziarnisty, jak i dobrze ustrukturyzowany. W tym poradniku nie prezentujemy szczegółowo poszczególnych dziedzinowych i dyscyplinowych standardów metadanych – zwracamy jednak uwagę na miejsca, gdzie można je odnaleźć.

Witryna Digital Curation Centre

Prowadzona przez brytyjskie Digital Curation Centre lista dyscyplinowych standardów metadanych dostępna jest na stronie organizacji[20]. Prezentuje ona schematy metadanych w podziale dziedzinowym i zawiera zarówno informacje o samych standardach, jak i ich możliwych rozszerzeniach, narzędziach używanych do ich utrwalenia i przechowywania oraz konkretnych przypadkach użycia.

[20] Disciplinary Metadata, https://www.dcc.ac.uk/guidance/standards/metadata [data dostępu: 17.07.2023].

Serwis Research Data Alliance

Zbliżona co do zakresu witryna koordynowana jest również przez organizację Research Data Alliance[21]. Umożliwia ona przeglądanie standardów metadanych w ujęciu alfabetycznym i tematycznym, zawiera również szereg informacji dodatkowych dotyczących dostępnych narzędzi, wzorców mapowania pomiędzy różnymi schematami metadanych, przypadkach użycia oraz organizacjach utrzymujących i wspierających poszczególne standardy metadanych. Szereg informacji zamieszczonych w witrynie jest również dostępny za pośrednictwem interfejsu programistycznego API.

Serwis FAIRsharing

Dostępny w ramach EOSC Marketplace serwis FAIRsharing prowadzony jest przez Uniwersytet Oksfordzki[22]. Gromadzi on informacje na temat standardów metadanych, repozytoriów i baz danych oraz polityk czasopism naukowych i instytucji finansujących badania naukowe, ze szczególnym uwzględnieniem rozwiązań mających na uwadze zasady FAIR. Badacze mogą wykorzystywać FAIRsharing w celu identyfikacji standardów, baz danych lub repozytoriów oraz ich zgodności z politykami instytucji finansujących lub wydawców naukowych.

---

[21] Metadata Standards Catalog, https://rdamsc.bath.ac.uk/ [data dostępu: 17.07.2023].

[22] Serwis dostępny jest na stronie https://fairsharing.org/ [data dostępu: 17.07.2023].

Ponowne wykorzystanie danych badawczych

Wykorzystanie danych związane jest z pierwotnym kontekstem ich wytworzenia, np. na potrzeby konkretnego projektu badawczego przez zespół realizujący ten projekt. Ponowne wykorzystanie to sytuacja, w której z danych udostępnionych w określonym miejscu, zwykle w repozytorium, korzysta ktoś inny[23]. Istnieją jednak też inne możliwości, które komplikują proste rozróżnienie na wykorzystanie i ponowne wykorzystanie w oparciu na wspomnianych kryteriach. Na przykład dane zdeponowane w repozytorium

[23] Por. Pasquetto I.V., Randles B.M., Borgman C.L., *On the Reuse of Scientific Data*, „Data Science Journal", 16(0), 2017, https://doi.org/10.5334/dsj-2017-008 [data dostępu: 17.07.2023].

mogą zostać (ponownie) wykorzystane przez ich autora lub współautora w innym, późniejszym projekcie. Istnieją też organizacje, które w systematyczny i ciągły sposób gromadzą i archiwizują dane na potrzeby całych społeczności naukowych. Mogą być to organizacje powołane do zbierania danych, których wytworzenie jest bardzo kosztowne i wymaga specjalnej infrastruktury, np. danych astronomicznych. Inny charakter mają organizacje zbierające dane o kluczowym znaczeniu dla administracji publicznej, np. urzędy statystyczne, które również mogą udostępniać swoje zasoby społeczności naukowej. Dane zbierane i opracowywane w tych kontekstach cechuje zwykle wysoka jakość i spójność z uwagi na standardy przyjęte przez powyższe organizacje.

Czynniki wpływające na ponowne wykorzystanie danych

Ponowne wykorzystanie danych zebranych przez innych badaczy w ramach innego projektu naukowego wymaga osobnego podejścia, w szczególności oceny danych pod kątem ich przydatności, adekwatności i znaczenia dla projektu. Kwestie te ujmować można bądź jako zbiór postulatów i wytycznych, bądź jako diagnozę stanu faktycznego. Ta druga perspektywa przyjęta została w badaniach, których rezultaty zaprezentowane zostały w artykule „Seeing oneself as a data reuser: How subjectification activates the drivers of data reuse in science”[24].

[24] Por. LaFlamme M., Poetz M., Spichtinger D., *Seeing oneself as a data reuser: How subjectification activates the drivers of data reuse in science*, „PLOS ONE” 17(8), 2022, https://doi.org/10.1371/journal.pone.0272153 [data dostępu: 17.07.2023].

Stanowią one dobry punkt wyjścia do omówienia kluczowych czynników wpływających na ponowne wykorzystanie danych.

Czynniki zależne od projektu

1. Charakterystyka danych

Charakterystyka danych obejmuje m.in. ich zgodność z zasadami FAIR, czyli łatwość znalezienia danych, ich dostępność, interoperacyjność i możliwość ponownego wykorzystania, a także widoczność. To warunki niezbędne do wykorzystania zastanych danych, choć różne rozwiązania mogą wpływać na stopień wdrożenia zasad FAIR. Czynniki te mają charakter obiektywny.

2. Zaufanie do danych

Czynnik ten do pewnego stopnia ma charakter subiektywny, a badacze mogą podejmować różne decyzje i działania mające na celu określenie wiarygodności, rzetelności, dokładności i kompletności danych, które potencjalnie mogliby wykorzystać. W praktyce może to być zarówno pobieżne sprawdzenie danych, jak również przeprowadzenie ponownych analiz pod kątem reprodukowalności.

Duże znaczenie mają w tym kontekście instytucje gromadzące i udostępniające dane, odgrywające rolę pośredników, którzy mogą w różnym stopniu zaświadczać o rzetelności danych. Z punktu widzenia badaczy zainteresowanych ponownym wykorzystaniem danych z konkretnego źródła istotne będzie więc sprawdzenie procedur deponowania danych w konkretnym

repozytorium, archiwum czy centrum danych; im większy zakres prowadzonej weryfikacji danych, tym więcej podstaw do tego, aby uznać je za godne zaufania. Co więcej, na dane zdeponowane w określonym miejscu, w ustalonej formie i wersji, można powołać się poprzez cytowanie. Dane pozyskane w inny sposób (np. poprzez korespondencję z autorem) wymagają niestandardowego opisu i zwykle nie mogą zostać udostępnione innym.

Podobnie pod uwagę można wziąć także polityki i praktyki czasopisma naukowego, w którym opublikowany został artykuł powiązany z danymi. Innym ważnym czynnikiem może być zaufanie do samej instytucji, przy której afiliowani są badacze, którzy wytworzyli dane.

Warto zastrzec, że powyższe kwestie nie mają charakteru wytycznych, a raczej odzwierciedlają różne strategie oceniania danych.

3. Adekwatność danych w stosunku do celów projektu

W zakres tego czynnika wchodzą kwestie merytoryczne, a punktem odniesienia są tu pytania badawcze. Dane znalezione poprzez ich powiązanie z istotnymi dla projektu artykułami naukowymi bądź poprzez wyszukiwanie w repozytoriach czy wyszukiwarkach danych, np. przy zastosowaniu słów kluczowych, mogą wiązać się z badanym problemem, a mimo wszystko okazać się nieprzydatne z różnych powodów. Kwestia adekwatności danych najściślej wiąże się z projektem

badawczym i każdorazowo wymaga oceny uwzględniającej cel i specyfikę projektu.

4. Zasoby i umiejętności potrzebne do ponownego wykorzystania danych

Ostatni czynnik wiąże się z zasobami, umiejętnościami oraz zakresem wsparcia zespołu w projekcie badawczym. Może chodzić tutaj m.in. o instytucjonalny dostęp do baz, które nie są w pełni otwarte, możliwość pracy z wykorzystaniem odpowiedniego oprogramowania, dodatkowe narzędzia ułatwiające pracę z danymi czy szkolenia i wsparcie ze strony data stewardów.

Czynniki niezależne od projektu

Oprócz powyższych czynników można wskazać także czynniki niezależne od projektu, związane z szeroko rozumianym systemem komunikacji naukowej oraz ewaluacji działalności naukowej. Chodzi tutaj o stosunek badaczy do ponownego wykorzystywania danych, normy i dobre praktyki przyjęte w konkretnej dyscyplinie czy obszarze badań oraz o system wymogów i zachęt. Czynniki te mają w istocie charakter polityczny, instytucjonalny i kulturowy; mogą wpływać na ponowne wykorzystanie istniejących danych na kilka sposobów, np. premiowanie przygotowania danych zgodnie z zasadami FAIR zwiększa zasób dobrze opracowanych i godnych zaufania danych, z których inni będą chętniej korzystać, natomiast coraz częstsze przypadki ponownego wykorzystania danych mogą

wpłynąć na upowszechnienie się tych praktyk i zmianę stosunku do nich poszczególnych badaczy. Jednym ze sposobów wpływania na zmiany w tym zakresie są nagrody za ponowne wykorzystanie danych przyznawane przez różnego typu instytucje, np. organizacje prowadzące badania naukowe.

Rzetelne wykorzystanie istniejących danych

Podejmując decyzję o ponownym wykorzystaniu danych warto kierować się jasnymi wskazówkami, np. listami kontrolnymi. Można opracować własną listę bądź skorzystać z list udostępnionych przez inne instytucje[25].

Dostęp do danych a ich ponowne wykorzystanie

„Uzyskanie dostępu do danych nie jest jednoznaczne z możliwością wykorzystania ich w dowolny sposób. Korzystanie z nich musi odbywać się zgodnie z przypisanymi im ograniczeniami prawnymi. Z danych udostępnionych bez wskazania konkretnej licencji bądź udostępnionych z wyraźnym zastrzeżeniem praw można korzystać jedynie w granicach swobód określonych prawem. Chodzi tu przede wszystkim o przepisy prawa autorskiego oraz ustawy o ochronie baz danych – dane badawcze mogą podlegać prawu autorskiemu, prawu *sui generis* do baz danych bądź obu tym reżimom jednocześnie. Konsekwencją takiej ochrony jest wąski zakres dopuszczalnego

[25] Przykładem może być lista kontrolna udostępniona na stronie Harvey Cushing/John Hay Whitney Medical Library, https://library.medicine.yale.edu/research-data/reuse-data [data dostępu: 17.07.2023].

ponownego wykorzystania (dozwolony użytek i jego okrojony odpowiednik w ustawie o ochronie baz danych, obejmujący użytek osobisty, użytek w celach dydaktycznych lub badawczych oraz cele państwowe: bezpieczeństwo wewnętrzne oraz postępowania sądowe lub administracyjne)”[26].

Z danych udostępnionych na określonej licencji można korzystać zgodnie z zakresem swobód określonych w tej licencji.

Cytowanie danych

W zakres rzetelnego wykorzystania danych wchodzi także ich odpowiednie cytowanie. „Ogólne zasady cytowania danych określone zostały w dokumencie Joint Declaration of Data Citation Principles[27], zgodnie z którym dane badawcze uznaje się za pełnoprawne rezultaty badań. Cytowanie danych pozwala na szybką i jednoznaczną identyfikację ich źródła, ułatwia weryfikację twierdzeń zawartych w publikacjach, a także sprzyja ponownemu wykorzystaniu danych. Powinno ponadto ułatwiać dostęp do danych wraz z powiązanymi z nimi metadanymi, dokumentacją, kodem i innymi materiałami niezbędnymi do rzetelnego korzystania z danych. Informacje, które należy uwzględnić w cytowaniu, to: autor lub autorzy, tytuł zbioru, rok

[26] J*ak korzystać z zasobów w repozytoriach danych*, oprac. N. Gruenpeter, Warszawa 2022, http://pon.edu.pl/dane-materialy [data dostępu: 17.07.2023].

[27] Joint Declaration of Data Citation Principles, https://www.force11.org/datacitationprinciples [data dostępu: 17.07.2023].

i miejsce udostępnienia, np. nazwa repozytorium lub archiwum, wersja, trwały identyfikator”[28].

Replikacja i reprodukcja badań

Replikacja polega na odtworzeniu badania, czyli jego ponownym przeprowadzeniu z zachowaniem warunków, metod i narzędzi zastosowanych w oryginalnym badaniu. Celem jest ustalenie, czy uzyskane wyniki powtórzą się, a w konsekwencji, czy wnioski z badania rzeczywiście mogą zostać uznane za zasadne. Replikacja najczęściej przeprowadzana jest przez inny zespół badawczy, co odróżnia ją od prostego powtórzenia, w podstawowym znaczeniu rozumianego jako ponowne przeprowadzenie badania przez ten sam zespół przy użyciu tych samych narzędzi i technik, np. w badaniach empirycznych.

Inne istotne rozróżnienie dotyczy replikacji i reprodukcji, które podajemy za przewodnikiem The Turing Way[29]. Przyjęte w nim definicje uwzględniają kontekst nauk obliczeniowych.

Replikacja (*replication*) ujęta jest jako przeprowadzenie takich samych analiz na różnych zestawach danych. Wynik uznać można za replikowalny (*replicable*), jeżeli kolejne analizy dają jakościowo podobne odpowiedzi.

---

[28] Tamże.

[29] Przewodnik dostępny na stronie: https://the-turing-way.netlify.app/ [data dostępu: 17.07.2023].

Reprodukcja (*reproduction*) rozumiana jest jako przeprowadzenie tych samych analiz na tych samych zbiorach danych. Wynik uznać można za reprodukowalny, (*reproducible*) jeżeli kolejne analizy stale dają tę samą odpowiedź.

Przewodnik uwzględnia także inne pojęcia oparte na rozróżnieniu sytuacji, w których przeprowadzane są różne analizy na takich samych oraz na różnych danych.

Stabilność (*robust*) wyników można uzyskać, kiedy ten sam zestaw danych poddawany jest różnym analizom w celu uzyskania odpowiedzi na to samo pytanie badawcze, a analizy te dają jakościowo podobny lub identyczny wynik. W zakresie badań prowadzonych z wykorzystaniem metod obliczeniowych działanie to pozwala pokazać, że wynik nie jest uzależniony od wybranego języka programowania.

Uogólnienie (*generalisable*) wyników jest rezultatem połączenia zreplikowanych i stabilnych ustaleń, które nie są zależne od konkretnego zestawu danych ani od konkretnej procedury analizy danych.

Dalsze rozróżnienia terminologiczne z zakresu szeroko rozumianego replikowania badań mogą odnosić się do zakresu, sposobu czy metod przeprowadzenia kolejnego badania oraz uwzględniać uwarunkowania specyficzne dla określonych dziedzin, dyscyplin czy obszarów badań.

Reprodukowalność w programie Horyzont Europa

Przewodnik po programie ramowym Horyzont Europa[30] uwzględnia wskazówki dotyczące reprodukowalności badań naukowych, które obejmują zarówno kwestie związane z metodologią badań, jak i z otwartą nauką. Należy w szczególności:

- precyzyjnie i jednoznacznie określić cel i przebieg badania oraz stosowaną metodologię;
- określić sposób postępowania z tzw. negatywnymi wynikami, jeśli takie wystąpią, aby inni mogli korzystać z rezultatów badania;
- zaplanować wyszukiwanie i sprawdzenie istniejących wyników i zastanych danych, aby upewnić się, że nie są one niepotrzebnie dublowane;
- określić sposób korzystania z preprintów i prerejestracji badań;
- szczegółowo określić kroki, które zostaną podjęte, aby proces badawczy i narzędzia były transparentne i dostępne w trakcie badania i po jego zakończeniu;
- określić działania mające na celu zapewnienie wiarygodności i jakości procesu i rezultatów projektu (np. recenzowanie, dzielenie się wiedzą, niezależne testowanie, nadzór, mechanizmy kontroli jakości);

[30] Horizon Europe, Programme Guide, s. 49–50, https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf [data dostępu: 17.07.2023].

- zaplanować wykorzystanie DMP w pełnym możliwym zakresie, aby wyszczególnić zasoby i materiały leżące u podstaw gromadzenia i analizy danych;
- upewnić się, że dane są zgodne z zasadami FAIR, aby inni mogli je znaleźć i ponownie wykorzystać w celu odtworzenia wyników;
- określić sposób zapewnienia stabilnej analizy statystycznej, którą można powtórzyć (moc testu, solidne techniki eksperymentalne, otwarte oprogramowanie);
- określić, jakie „wspólne zasoby" istotne dla badań i innowacji będą wytworzone w ramach projektu, w tym bazy wiedzy, metodologie, ramy ewaluacji, ontologie, otwarte repozytoria itp.;
- wprowadzić rozwiązania mające na celu walidację, przedstawienie, zapewnienie interoperacyjności, zwiększenie skali i replikowalności wyników działań w zakresie badań naukowych i innowacji;
- przemyśleć, czy w ramach projektu powstaną cyfrowe kopie wyników, np. cyfrowe bliźniaki, wirtualne reprezentacje, cyfrowe schematy, które zwiększają prawdopodobieństwo ponownego wykorzystania i reprodukowalności.

Narzędzia komunikacji naukowej sprzyjające replikowalności i reprodukowalności

Artykuł naukowy z opisem metod badawczych i wnioskami nie zawiera wszystkich informacji pozwalających na zreplikowanie badań, dlatego tak ważny jest rozwój nowych narzędzi

komunikacji naukowej, pozwalających na dzielenie się wynikami badań oraz metodami. Należą do nich repozytoria danych badawczych, w których udostępniać można dane wraz z dokumentacją, a także inne serwisy służące do prowadzenia notatników laboratoryjnych, udostępniania opisów metod, procedur i narzędzi badawczych czy oprogramowania. Częste jest równoczesne korzystanie z wielu rozwiązań, np. artykuł opublikowany w czasopiśmie naukowym może zawierać odniesienia do repozytorium, w którym udostępnione zostały dane, oraz do serwisu gromadzącego opis protokołów i procedur. W takich wypadkach ważne jest powiązanie poszczególnych zasobów poprzez trwałe identyfikatory.

Rzetelne ponowne przeprowadzenie badania wymaga dobrze przygotowanej dokumentacji udostępnionej wraz z danymi badawczymi. Powinna ona zawierać wszystkie niezbędne informacje na temat badania, w wypadku badań społecznych dotyczące np. doboru i rekrutacji uczestników badania, przekazywanych im instrukcji, warunków przeprowadzenia badania, metod i technik badawczych, sposobów analizowania informacji czy oprogramowania. Zakres i charakter tych informacji jest ściśle związany ze specyfiką badań.

Narzędzia wspierające kontrolę jakości udostępnionych danych

Na poziomie ogólnym najważniejszym narzędziem wspierającym kontrolę jakości udostępnianych danych jest samo repozytorium, w którym są one udostępnione. Jest to szczególnie dobrze

widoczne w wypadku tych repozytoriów specjalistycznych i dziedzinowych, które oferują gruntowną weryfikację danych przez personel repozytorium.

W zapewnieniu odpowiedniej jakości danych pomagają również zalecenia dotyczące sposobu przygotowania danych mających trafić do konkretnego repozytorium. Zalecenia tego rodzaju (a niekiedy nawet wymagania, które trzeba spełnić, aby opublikować dane w określonym repozytorium) dotyczą najczęściej rodzaju danych możliwych do udostępnienia w danym serwisie, wymaganych lub sugerowanych formatów plików, nazewnictwa plików i katalogów, kształtu dokumentacji towarzyszącej plikom z danymi czy licencji, jakimi mają zostać objęte dane. Przykłady takich zaleceń znaleźć można na stronach repozytorium Dryad (https://datadryad.org/stash/best_practices), Zenodo (https://about.zenodo.org/policies/) czy na stronie informacyjnej Polskiego Archiwum Danych Społecznych (https://pads.org.pl/index.php/dokumenty/).

W wypadku repozytoriów specjalistycznych i dziedzinowych zalecenia mogą obejmować również użycie konkretnych narzędzi służących weryfikacji danych, charakterystycznych dla danych określonego typu lub opracowanych w określonym formacie. Przykłady zaleceń dotyczących narzędzi umożliwiajacych przygotowanie i/lub walidację danych znaleźć można na stronach

specjalistycznego repozytorium PDB[31] czy repozytorium ogólnego przeznaczenia Dryad[32].

Nawet repozytoria ogólnego przeznaczenia oferują jednak pewne wbudowane narzędzia pozwalające w ogólnym sensie sprawdzić poprawność zdeponowanych danych. Takim narzędziem jest np. suma kontrolna generowana dla każdego pliku przesłanego do repozytorium. Dzięki temu zarówno osoba przesyłająca plik do repozytorium, jak i pobierająca go z niego po wygenerowaniu sumy kontrolnej tego pliku na swoim lokalnym komputerze za pomocą odpowiedniego algorytmu (często stosowane są tu algorytmy skrótu MD5 lub algorytmy z rodziny SHA) może upewnić się, że plik znajdujący się u niej na dysku oraz plik znajdujący się w repozytorium to dwa egzemplarze tego samego pliku. Repozytoria oparte na oprogramowaniu Dataverse dla każdego pliku tabelarycznego, który przejdzie automatyczną analizę, wyliczają dodatkową sygnaturę UNF, która jest identyczna dla dwóch plików o takiej samej zawartości tabelarycznej, nawet jeśli pliki te różnią się od siebie formatem czy kolejnością niektórych elementów. Więcej informacji na temat UNF można znaleźć na stronach przewodnika oprogramowania Dataverse[33].

---

[31] RCSB Protein Data Bank, https://www.rcsb.org/docs/general-help/deposition-resources-#validation-services [data dostępu: 17.07.2023].

[32] Dryad News, Frictionless Data and Dryad join forces to validate research data, https://blog.datadryad.org/2021/08/09/product-update-frictionless-data-and-dryad-join-forces-to-validate-research-data/ [data dostępu: 17.07.2023].

[33] Universal Numerical Fingerprint (UNF), https://guides.dataverse.org/en/latest/developers/unf/index.html [data dostępu: 17.07.2023].

Fakt, iż repozytoria są istotnym narzędziem wspierania jakości danych, dostrzegli twórcy kryteriów oceny serwisów ubiegających się o przyznanie uznanego certyfikatu CoreTrustSeal. Kwestia jakości jest w nich powiązana przede wszystkim ze standardami dotyczącymi formatów akceptowanych przez repozytorium, stosowanych przez nie schematów metadanych, treści metadanych oraz ich powiązania z innymi obiektami cyfrowymi. Repozytorium ubiegające się o certyfikat powinno też określić standardy dla danych, metadanych oraz dokumentacji (korzystając z istniejących wzorców lub opracowując własne), a także ustalić sposób postępowania w wypadku, gdy takie standardy nie są spełnione (np. rozstrzygnąć, czy poprawki dokonywane są przez samą obsługę repozytorium, czy też zbiór zwracany jest do poprawy przez osobę deponującą dane). To, że w wypadku jakiegoś konkretnego zbioru danych nie uda się usunąć problemów związanych z jego jakością, nie musi od razu oznaczać, że staje się on bezużyteczny. Ważne jednak, by został odpowiednio pod tym względem udokumentowany.

Prawne aspekty zarządzania danymi badawczymi

Udostępnianie danych badawczych może podlegać ograniczeniom wynikającym z okoliczności technicznych, organizacyjnych, finansowych, prawnych lub etycznych.

Okoliczności techniczne związane są głównie z formatami plików stosowanymi do zapisywania danych, a także protokołami i interfejsami wykorzystywanymi do ich przesyłania, a konkretnie z tym, czy dany format, protokół czy interfejs jest obsługiwany przez wykorzystywane przez użytkownika oprogramowanie i w jakim stopniu. Pośrednio są one zatem związane także z dostępnością takiego oprogramowania i możliwością jego ewentualnego dostosowania. Obsługa zapisu lub przesłania

danych w określony sposób uzależniona jest z kolei od dostępności oraz jakości dokumentacji (specyfikacji) odpowiedniego formatu, protokołu czy interfejsu dla osób tworzących dane oprogramowanie.

Niektóre specyfikacje są otwartymi standardami, czyli zostały przyjęte i opublikowane przez organizację standardyzującą w drodze formalnego procesu gwarantującego inkluzywność i niezależność, a także techniczną jakość oraz minimalizację ograniczeń prawnych implementacji standardu. Otwarte standardy mogą być bez większych przeszkód wdrożone w dowolnym oprogramowaniu, w tym w wolnym i otwartym oprogramowaniu, co znacząco wpływa na dostępność i możliwość wyboru oprogramowania przez użytkownika. Ujmując to szerzej, udostępnienie danych w otwartym standardzie nie ogranicza odbiorców danych co do sposobów technicznych, na jakie mogą z tych danych skorzystać. Oczywiście, nawet wtedy istnieć mogą inne ograniczenia techniczne lub podobne, wynikające np. ze złej jakości danych, niewystarczającego ich udokumentowania itd.

Niestety istnieje też sporo standardów zamkniętych lub nawet takich formatów, protokołów i interfejsów, których specyfikacje istnieją w najlepszym przypadku jedynie jako wewnętrzna dokumentacja producenta danego własnościowego oprogramowania. Skorzystanie z takich specyfikacji jest albo niemożliwe, albo ograniczone do wąskiego kręgu (czasem wyłącznie do producenta jednej aplikacji). Ewentualna obsługa zamkniętych standardów przez alternatywne oprogramowanie lub konwersja

do otwartego formatu często rodzi kolejne problemy techniczne. Udostępnianie danych w standardzie zamkniętym oznacza, że odbiorca danych będzie musiał posłużyć się konkretnym oprogramowaniem do ich odebrania lub przetwarzania, co nie zawsze musi odpowiadać jego potrzebom, a w skrajnych przypadkach może być niemożliwe (gdy np. nie jest w stanie zintegrować tego oprogramowania ze swoim systemem).

Do okoliczności technicznych możemy zaliczyć także uwarunkowania fizycznej infrastruktury służącej do przechowywania i udostępniania danych. Przy dużych zbiorach danych rośnie znaczenie fizycznego miejsca ich przechowywania i zastosowanych zabezpieczeń. Potrzeba, a jednocześnie trudność zapewnienia bezpieczeństwa danych może podpowiadać wykorzystanie infrastruktury rozproszonej (tzw. „chmury"). Tu, podobnie jak przy omówionych wyżej kwestiach związanych ze standardami i oprogramowaniem, duże znaczenie ma to, kto administruje taką infrastrukturą, jakie są możliwości sprawowania nad tym kontroli, a w skrajnych przypadkach, czy istnieje łatwy sposób na wydobycie danych i przeniesienie ich w inne miejsce.

Ograniczenia o charakterze organizacyjnym dotyczą koniecznej koordynacji pomiędzy różnymi interesariuszami zaangażowanymi w cały proces, a także działań potrzebnych do udostępniania, w tym usuwania ograniczeń wynikających z pozostałych okoliczności. Jak pokazujemy w innych częściach kursu, udostępnienie danych może wkraczać w interesy, obowiązki i kompetencje wielu różnych osób i jednostek. Jeżeli nie zostali oni włączeni w odpowiednie procesy (np. w ramach przyjętej

w instytucji polityki otwartości), to trudności organizacyjne mogą być niewątpliwie większe. Co istotne, niekiedy osoby, z którymi potrzebna jest koordynacja procesu udostępniania danych, mogą nie podlegać bezpośredniej „władzy" udostępniającego i jego polityce otwartości, co implikuje konieczność dodatkowych uzgodnień i może skutkować specyficznymi ograniczeniami organizacyjnymi. Z drugiej strony nawet hierarchiczna podległość nie wystarczy, jeżeli zaangażowane w proces udostępniania danych osoby nie mają wystarczającej wiedzy, kompetencji, a także motywacji do działania. Do wyzwań organizacyjnych należy wobec tego zaliczyć odpowiednie zarządzanie czy kierowanie zespołem oraz zapewnienie dobrej komunikacji i współpracy pomiędzy zespołami.

Ograniczenia finansowe wymagają prawdopodobnie najmniej wyjaśnień, gdyż związane są po prostu z koniecznością pokrycia kosztów przygotowania danych do udostępniania oraz kosztów związanych z ich przechowywaniem i przesyłaniem. Kluczowe jest przede wszystkim dobre oszacowanie tych kosztów w ramach projektu oraz w trakcie tzw. okresu trwałości. Ponadto, biorąc pod uwagę cele otwierania danych badawczych, należy zapewnić możliwość pokrywania tych kosztów w długim okresie, czy to przez pozyskanie nowego finansowania, czy to przez deponowanie danych w repozytoriach, które dają gwarancję długoterminowej archiwizacji.

Koszty zależą od wielu czynników, a odpowiednie zarządzanie procesem badawczym może pozwolić wpłynąć na ich wysokość. Istotne jest zwłaszcza wczesne identyfikowanie ewentualnych

trudności i planowanie oraz projektowanie całego procesu tak, aby jak najdłużej zachowywać możliwość wyboru różnych rozwiązań. W zarządzaniu kosztami otwierania danych badawczych warto posługiwać się koncepcją „kosztu całkowitego", na który składają się nie tylko bezpośrednie koszty wybranego rozwiązania (np. licencja na oprogramowanie użyte do przetwarzania danych), ale też ewentualne koszty jego zmiany (np. konwersji danych do innego formatu) wtedy, gdy pierwotnie wybrane rozwiązanie okaże się niewystarczające. W tym kontekście możliwość modyfikacji oprogramowania oraz możliwość sięgnięcia po wsparcie w jego rozwoju (istniejąca z definicji przy wolnym oprogramowaniu rozwijanym przez dobrze zorganizowane i rozwinięte społeczności, a niedostępna lub zależna od wybranego pakietu licencyjnego przy oprogramowaniu własnościowym), oraz swoboda wyboru różnych rozwiązań związana z otwartymi standardami nabierają istotnego znaczenia.

W dalszej części niniejszego rozdziału zajmować się będziemy ograniczeniami o charakterze prawnym i etycznym. Mówiąc w największym skrócie, chodzi tu o ograniczenia wynikające z tego, że udostępnienie danych lub określony sposób tego udostępnienia może naruszać czyjeś uprawnienia bądź powszechnie obowiązujące prawo. Szczególnym przypadkiem ograniczenia prawnego będzie sytuacja, gdy różne związane z danymi obowiązki są sprzeczne i nie jest możliwe spełnienie ich wszystkich naraz.

Ograniczenia o charakterze prawnym wynikają po pierwsze z faktu przysługiwania określonym osobom praw obejmujących udostępnianie określonych danych (skutkujących koniecznością ograniczenia lub zaniechania podejmowanych działań bądź pozyskania zgody tych osób). Zdecydowanie najwięcej uwagi poświęcimy poniżej prawom wyłącznym (skutecznym wobec każdego). Szczególnym przypadkiem tego typu ograniczeń mogą być patenty na tzw. wynalazki realizowane komputerowo, których naruszenie może być wynikiem korzystania z zamkniętych standardów. Zazwyczaj będą to jednak bardziej „typowe" kwestie związane z cudzymi prawami autorskimi czy dobrami osobistymi. Ograniczenia wynikające z praw wyłącznych mogą być zazwyczaj przedmiotem negocjacji stron, istotna jest wobec tego wiedza o zasadach prawidłowego formułowania umów, których przedmiotem są prawa takie jak np. prawa autorskie.

Ograniczenia prawne mogą być też „wykreowane" przez strony umowy, nawet jeżeli żadnej z nich nie przysługuje prawo wyłączne związane z danymi. Wynika to z zasady swobody umów, która pozwala każdemu zobowiązać się wobec innej osoby do określonego działania lub zaniechania. Do takiej sytuacji może dojść wtedy, gdy naukowcy otrzymują dane do badań od osoby lub instytucji, która uzależnia ich faktyczne wydanie od zawarcia umowy określonej treści.

Wreszcie, nawet gdy żadna osoba nie ma praw wyłącznych do danych i zostały one pozyskane bez dodatkowych zastrzeżeń, prawo może mimo to uzależniać ich udostępnienie od wypełnienia określonych obowiązków lub wprost zakazywać

udostępniania określonych kategorii danych. Przykładem takiej sytuacji jest ochrona danych osobowych. Tego typu ograniczeń nie można negocjować, można jednak podejmować działania w celu udostępnienia danych zgodnie ze związanymi z nimi obowiązkami. Przy czym tam, gdzie obowiązkiem administratora jest uzyskanie zgody na udostępnienie danych, możemy mówić o swoistych negocjacjach – z samej istoty „zgody" wynika bowiem konieczność ustalenia takiego sposobu udostępniania danych, który spotka się z akceptacją osoby, której dane dotyczą.

Przez ograniczenia o charakterze etycznym rozumieć tu będziemy przede wszystkim sformalizowane kodeksy etyczne obowiązujące badaczy. Mogą one wprowadzać ograniczenia udostępniania danych, których nie ma w systemie prawnym. Z drugiej jednak strony mogą one też stymulować naukowców do podejmowania dodatkowych wysiłków w celu usuwania istniejących ograniczeń (prawnych i innych) na drodze do udostępniania danych (np. wtedy, gdy nie ma prawnego obowiązku zbierania danych tak, aby było możliwe ich późniejsze udostępnienie). Przykładem są postanowienia Kodeksu etyki socjologa, który wzywa do jak najszerszego udostępniania wyników badań. Kodeksy mogą wiązać w sensie prawnym z uwagi na włączanie ich do wiążących ich umów (np. zatrudnienia na uczelni, dofinansowania od grantodawcy lub zlecenia od osoby wymagającej zastosowania danego kodeksu).

W dalszej kolejności mieszczą się tu różnego rodzaju normy postępowania przyjmowane przez badaczy kierujących się wartościami takimi jak rzetelność naukowa, ochrona osób

badanych itp. Przykładem takiej sytuacji jest powstrzymywanie się od udostępniania niektórych kolekcji gromadzonych w koloniach przez badaczy pochodzących z dominiów, bez ich odpowiedniego opisania i wyjaśnienia braku obiektywizmu przy tworzeniu tych kolekcji. Normy te nie muszą być zawsze skodyfikowane w postaci kodeksów etycznych. Mogą jednak nabrać prawnego znaczenia, gdyż przepisy prawa często odwołują się do takich pojęć jak „rzetelność naukowa" przy ocenianiu zakresu praw i obowiązków naukowców. Wobec tego takie lub inne normy etyczne mogą przełożyć się na prawną odpowiedzialność.

Dane badawcze jako przedmiot ochrony prawa własności intelektualnej

„Własność intelektualna" to niezbyt ścisły termin używany zazwyczaj dla wygody na zbiorcze określenie wielu różnych rodzajów praw na dobrach intelektualnych. Cechami wspólnymi tych praw jest to, że dotyczą one dóbr intelektualnych (niematerialnych) i są regulowane przez prawo prywatne (cywilne), czyli w największym skrócie prawo regulujące stosunki majątkowe i niemajątkowe między równorzędnymi podmiotami, osobami fizycznymi lub prawnymi.

Do praw „własności intelektualnej" zalicza się zwykle prawo autorskie, prawo *sui generis* do baz danych oraz prawa własności przemysłowej (patenty i wzory użytkowe, znaki towarowe, wzory przemysłowe itd.). Są one prawami wyłącznymi skutecznymi „wobec każdego", czyli przysługują określonej

osobie lub osobom, pozwalając im wykluczyć inne osoby z korzystania z danego dobra, na podobieństwo prawa własności. Bardzo blisko tych regulacji jest ochrona dóbr osobistych, jednak przynajmniej polska nauka prawa raczej nie zalicza tych dóbr do „własności intelektualnej".

W ramach „własności intelektualnej" dość zgodnie wymienia się natomiast ochronę tajemnicy przedsiębiorstwa oraz delikty nieuczciwej konkurencji, choć ich konstrukcja nie jest w ogóle oparta na koncepcji praw wyłącznych. Ochrona tajemnicy przedsiębiorstwa oznacza w największym skrócie, że przedsiębiorca, który podjął działania zabezpieczające określone wartościowe informacje przed ujawnieniem, może domagać się m.in. odszkodowania od osób, które weszły w ich posiadanie niezgodnie z prawem. Z kolei delikty nieuczciwej konkurencji to różne czyny, które naruszają interes innego przedsiębiorcy lub klienta, a przy tym są sprzeczne z prawem lub dobrymi obyczajami, np. wprowadzające w błąd oznaczenie przedsiębiorstwa lub inne przypadki podszywania się czy dezinformacji.

Ochrona danych osobowych z kolei to obszar traktowany osobno do „własności intelektualnej". Choć dane osobowe są również dobrami niematerialnymi, a RODO zawiera szczegółowe konstrukcje podobne do tych, które są wykorzystywane np. w prawie autorskim (wymagania dotyczące zgody), to jednak jest to zasadniczo odmienny reżim. Źródłem tej regulacji jest prawo administracyjne, z czego wynika, że osobom, których dane dotyczą nie przyznano żadnych prywatnych praw do ich danych. Administrator danych ma obowiązki nadzorowane i kontrolowane

przez organy władzy, choć oczywiście poszczególnym osobom przysługują w związku z tym określone uprawnienia, a ostatecznie chodzi o to, aby to ich interesy były zabezpieczone. Obowiązków tych zdecydowanie nie można wyłączyć w drodze umowy, a uprawnień osób przenieść na inne podmioty – co byłoby możliwe, gdyby skorzystać z konstrukcji praw prywatnych. Jednak choć ochrona danych osobowych nie zalicza się do „własności intelektualnej", to na końcu niniejszej części omówimy podstawowe zasady tej ochrony na zasadzie porównania i zestawienia. Zostaną one rozwinięte w osobnej części poświęconej wyłącznie danym badawczym zawierającym dane osobowe.

Nie wszystkie rodzaje praw „własności intelektualnej" mają znaczenie w odniesieniu do każdego zbioru danych. Najprawdopodobniej najczęściej spotykane ograniczenia wynikają z praw autorskich, praw *sui generis* oraz z ochrony danych osobowych (ta ostatnia w ogóle nie jest zaliczana do kategorii „własności intelektualnej"). Ograniczenia wynikające z praw własności przemysłowej (głównie patentów, ale też ochrony tajemnicy przedsiębiorstwa) mogą mieć znaczenie przede wszystkim w kontekście równolegle prowadzonej komercjalizacji wyników badań, w związku z czym poniżej poruszymy także ten aspekt.

Prawo autorskie chroni utwory rozumiane jako dobra niematerialne będące wytworem człowieka, cechujące się wkładem indywidualnej twórczości. Utworami są w szczególności utwory naukowe. Co istotne, wkład indywidualnej twórczości

wymagany do objęcia utworu ochroną jest niski. Chroniony utwór naukowy nie musi zawierać przełomowego odkrycia, utwór literacki może być po prostu kiczowaty. Z prawnego punktu widzenia jakość utworu nie ma znaczenia, decyduje bowiem jedynie to, czy (upraszczając) można w utworze odnaleźć indywidualne piętno jego autora (tzn., czy nie jest on efektem działań odtwórczych, prowadzących do powtarzalnych rezultatów).

Ochrona prawa autorskiego przysługuje automatycznie, nie wymaga zgłaszania do jakiegokolwiek urzędu ani badania stopnia twórczości (w odróżnieniu od np. ochrony patentowej). Dość niski próg twórczości oraz brak „urzędowego" potwierdzania ochrony oznacza, że jest dość prawdopodobne, że dane dobro intelektualne jest utworem, jednak nie jesteśmy tego w stanie nigdzie kategorycznie potwierdzić, wyłączając oczywiście scenariusz, w którym dojdzie do rozstrzygnięcia sądowego tej kwestii. Musi nam wystarczyć jedynie nasza własna ocena stopnia twórczości (a ocena jakiegokolwiek eksperta w danej dziedzinie będzie tylko prywatnym zdaniem takiego eksperta).

Mimo to nie wszystkie dane badawcze lub przynajmniej nie wszystkie części danych badawczych będą zawsze chronione prawem autorskim. Wyraźny przepis prawa autorskiego stanowi, że ochronie podlega tylko sposób wyrażenia, a nie idee zawarte w utworze (brak jest jednak wyraźnej definicji pozwalającej ustalić granicę między oboma tymi pojęciami). Na pewno nie są chronione suche fakty, informacje, odkrycia naukowe. Nie będą

więc chronione dane polegające na mechanicznej, odtwórczej rejestracji rzeczywistości. Jednocześnie jednak twórcza aranżacja, zbiór nawet niechronionych elementów będzie utworem objętym prawem autorskim. Oznacza to, że autorski dobór parametrów danych rejestrowanych z rzeczywistości może prowadzić do powstania chronionego prawem autorskim zbioru. Warto pamiętać jednak, że twierdzenie o przysługiwaniu ochrony prawa autorskiego musi być poparte wskazaniem osoby lub osób (ludzi), które wniosły do utworu swój własny wkład indywidualnej twórczości (czyli kto konkretnie dobrał parametry danych i dlaczego nie jest to wybór podyktowany jedynie zewnętrznymi czynnikami, takimi jak np. wymogi rzetelności określonego badania). Ponadto, choć próg wymaganej twórczości jest niski, to chroniona postać utworu musi być możliwa do obiektywnego doświadczenia za pomocą ludzkich zmysłów. Oznacza to, że osoba powołująca się na ochronę musi być w stanie wskazać, co konkretnie jest efektem twórczości (nie wystarczy więc enigmatyczne powoływanie się na „zestawienie danych").

Sui generis ochrona baz danych jest bardzo podobna do prawa autorskiego, jeżeli chodzi o konstrukcję (stąd określenie „swego rodzaju"). Inaczej jednak definiuje przedmiot ochrony i przesłanki jej udzielenia. Przedmiotem ochrony jest tu baza danych rozumiana jako zbiór danych lub jakichkolwiek innych materiałów i elementów zgromadzonych według określonej systematyki lub metody, indywidualnie dostępnych w jakikolwiek sposób, w tym środkami elektronicznymi, wymagający istotnego co do jakości lub ilości nakładu inwestycyjnego w celu sporządzenia, weryfikacji lub prezentacji jego zawartości. Podobnie zatem jak

w przypadku prawa autorskiego, które wymaga wkładu indywidualnej twórczości, ochrona *sui generis* będzie przysługiwała nie każdej bazie danych, a tylko takim, których sporządzenie, weryfikacja lub prezentacja wymaga istotnego nakładu inwestycyjnego. Samo rozumienie bazy danych jest natomiast dość szerokie i praktycznie każdy zbiór czegokolwiek może podlegać ochronie – o ile tylko wykazany zostanie nakład inwestycyjny. I tu znowu podobnie, o ile samo pojęcie nakładu inwestycyjnego może nie nastręczać szczególnych trudności, to już ustalenie, czy jest on „istotny", może być przedmiotem sporów. Dodatkowym ważnym kryterium ochrony jest możliwość wskazania podmiotu, który poniósł ryzyko nakładu inwestycyjnego przy tworzeniu bazy danych – temu bowiem podmiotowi przysługuje prawo wyłączne do bazy danych. Innymi słowy, brak ryzyka inwestycyjnego oznacza brak ochrony. Wobec automatyzmu ochrony i braku rejestracji ani urzędowego poświadczania ochrony *sui generis*, mamy tu także ten sam problem jak w przypadku prawa autorskiego – trudność w ustaleniu, czy dany przedmiot jest chroniony, czy nie. Omawiane wyżej kryteria nakładu inwestycyjnego i ponoszenia ryzyka inwestycyjnego musimy oceniać samodzielnie.

Ochrona prawa autorskiego i ochrona *sui generis* nie wykluczają się wzajemnie. Zbiór danych może podlegać obu tym reżimom, jeżeli spełnione są wszystkie odpowiednie kryteria, jednemu z nich bądź żadnemu. Co istotne, ochrona ta może działać na wielu poziomach zbioru danych – dla przykładu, poszczególne dane w zbiorze mogą być utworami, a zbiór jako całość może dodatkowo stanowić osobny utwór z uwagi na twórczość

wniesioną przy doborze i zestawieniu jego elementów. Zbiór taki może stanowić jednocześnie chronioną bazę danych. Mówimy wobec tego o ochronie kumulatywnej i wielopoziomowej. Kolejne reżimy prawne mogą dodatkowo komplikować tę kwestię.

Powyżej omówione zasady ochrony wynikają z przepisów ustawowych. Choć zostały one w dużym stopniu zharmonizowane na skutek prawa Unii Europejskiej oraz prawa międzynarodowego, to nadal szczegółowe rozwiązania mogą różnić się w zależności od kraju, którego prawu podlegałaby dana umowa bądź ustalenie odpowiedzialności w przypadku naruszenia. Ochrona danych osobowych z kolei wynika obecnie z rozporządzenia Unii Europejskiej (Rozporządzenie 2016/679 zwane w skrócie „RODO"), które od kilku lat w istotnym zakresie zastąpiło różne krajowe ustawy. W związku z tym, że rozporządzenia unijne są bezpośrednio skuteczne bez konieczności implementacji do prawa krajowego, mówimy tu o znacznie wyższym stopniu harmonizacji.

Powtórzmy, że ochrona danych osobowych nie jest zaliczana do prawa „własności intelektualnej", inna jest też konstrukcja tej ochrony – skupia się ona przede wszystkim na określeniu obowiązków administratora danych, których podstawowa egzekucja zapewniana jest drogą administracyjną.

Natomiast podstawowe zasady ochrony danych osobowych, wyznaczające m.in. zakres obowiązków administratorów danych, zostały ujęte w art. 5 RODO i są to:

1) wymaganie zapewnienia przetwarzania zgodnie z prawem, rzetelnie i przejrzyście dla osoby, której dane dotyczą;
2) zasada ograniczonego celu: dane osobowe muszą być zbierane w konkretnych, wyraźnych i prawnie uzasadnionych celach i nie przetwarzane dalej w sposób niezgodny z tymi celami (przy czym dalsze przetwarzanie m.in. do celów badań naukowych nie jest uznawane za niezgodne z pierwotnymi celami);
3) zasada minimalizmu: dane powinny być adekwatne, stosowne oraz ograniczone do tego, co niezbędne do celów, w których są przetwarzane;
4) wymaganie, aby dane były prawidłowe i w razie potrzeby uaktualniane (wymóg usuwania lub sprostowania nieprawidłowych danych);
5) ograniczenie przechowywania w formie umożliwiającej identyfikację osoby, której dane dotyczą, tylko przez okres niezbędny do realizacji celów, w których dane te są przetwarzane;
6) zasada integralności i poufności: wymóg przetwarzania w sposób zapewniający odpowiednie bezpieczeństwo danych osobowych, w tym ochronę przed niedozwolonym lub niezgodnym z prawem przetwarzaniem oraz przypadkową utratą, zniszczeniem lub uszkodzeniem, za pomocą odpowiednich środków technicznych lub organizacyjnych;
7) zasada rozliczalności: za przestrzeganie powyższych zasad odpowiedzialny jest administrator i musi on być w stanie wykazać ich przestrzeganie.

Prawa członków zespołu badawczego, prawa instytucji prowadzącej badania, prawa uczestników konsorcjum, prawa instytucji finansującej

W ramach jednego zbioru danych może dojść nie tylko do kumulacji ochrony i występowania różnych praw na różnych poziomach danego zbioru. Prawa te mogą ponadto przysługiwać różnym podmiotom. Zasady regulujące ochronę praw wyłącznych takich jak prawa autorskie lub prawa *sui generis* inaczej bowiem definiują tzw. pierwotny podmiot prawa, czyli tego, komu prawa te przysługują w momencie ich powstania. Ponadto część tych praw jest zbywalna, więc nawet po ich powstaniu na rzecz jednego podmiotu mogą zostać przeniesione (w całości lub w części). Z kolei obowiązki takie jak obowiązki administratora danych osobowych z definicji odnoszą się do wszystkich osób, których dane znajdują się w zbiorze. Zaangażowanie w proces badawczy kolejnych podmiotów (członków konsorcjum, instytucji finansującej) rodzi kolejne możliwości w zakresie przysługiwania tym podmiotom określonych praw do gromadzonych w ramach badań danych.

Przed omówieniem zasad nabywania praw autorskich konieczne jest wyjaśnienie, że w skład tych praw wchodzą dwa rodzaje uprawnień: autorskie prawa osobiste i autorskie prawa majątkowe. Różnią się one celem i zakresem. Autorskie prawa osobiste mają na celu ochronę głównie niemajątkowych interesów twórcy, czyli osoby wnoszącej do utworu wkład swojej indywidualnej twórczości, przed działaniami takimi jak

przypisanie sobie autorstwa utworu przez kogoś innego, wypaczenie sensu utworu, prezentowanie zmienionej wersji utworu jako oryginał. Z kolei autorskie prawa majątkowe mają na celu ochronę interesów majątkowych, z nich bowiem wynika konieczność uzyskiwania zgody podmiotu tych praw na kopiowanie i rozpowszechnianie utworu (a zgoda ta może być udzielona w zamian za wynagrodzenie).

Autorskie prawa osobiste zawsze powstają po stronie twórcy (twórców) utworu i są niezbywalne. Nie ma wyjątków od tej zasady w prawie polskim.

Autorskie prawa majątkowe z kolei powstają co do zasady po stronie twórcy (twórców) utworu, jednak od tej zasady istnieją wyjątki. Ponadto prawo autorskie przewiduje ułatwiony sposób nabywania przez pracodawcę autorskich praw majątkowych w przypadku utworów pracowniczych. Prawa te są zbywalne, więc niezależnie od tego komu przysługiwały na początku, mogą być następnie przenoszone na dowolne podmioty w drodze umów. Podlegają również dziedziczeniu, więc po śmierci twórcy przechodzą na spadkobierców wskazanych zgodnie z prawem spadkowym, o ile oczywiście nie zostały przez niego przeniesione wcześniej np. na wydawcę.

Wyjątkiem od zasady powstawania praw po stronie twórców utworu jest sytuacja pracowniczych programów komputerowych. Mianowicie prawa do programu stworzonego przez pracownika w ramach jego pracowniczych obowiązków powstają od razu po stronie pracodawcy. Na nieco podobnej zasadzie oparte są

prawa do utworów zbiorowych takich jak np. encyklopedie lub publikacje periodyczne. Przysługują one od razu ich wydawcy, choć prawa do poszczególnych części tych utworów, czyli haseł encyklopedycznych lub artykułów w czasopiśmie, podlegają już ogólnym zasadom.

Ułatwienie w nabywaniu praw do utworów pracowniczych polega natomiast na tym, że prawo do utworu (innego niż program komputerowy) stworzonego przez pracownika w ramach wykonywania przez niego obowiązków ze stosunku pracy nabywa pracodawca. Jest to nadal przeniesienie praw (nie powstają one od razu po stronie pracodawcy), a zakres tego przeniesienia wynika z „celu umowy o pracę i zgodnego zamiaru stron". Oznacza to w szczególności, że pracodawca może nabyć te prawa tylko w jakiejś części. Przeniesienie to nie wymaga jednak żadnych szczególnych postanowień umowy – wystarczy tylko fakt przygotowania utworu w ramach stosunku pracy. Oczywiście przy lakonicznych umowach o pracę mogą powstać wątpliwości co do zakresu nabycia, więc mimo wszystko wskazane jest rozwinięcie tych kwestii w postanowieniach umowy. Co istotne, omawiane ułatwienie dotyczy wyłącznie stosunków pracy, nie dotyczy zatem osób „samozatrudnionych", umów o dzieło ani umów zlecenia – nabywanie autorskich praw majątkowych do utworów stworzonych przez te osoby podlega zasadom ogólnym, które omówimy w dalszej części tego rozdziału.

Wyłomem od zasady ułatwionego nabywania praw przez pracodawcę jest przypadek pracowniczych utworów naukowych.

Mianowicie jeżeli do obowiązków pracowniczych pracownika instytucji naukowej należy tworzenie utworów naukowych, to prawa do takiego utworu pozostają co do zasady przy pracowniku. Prawo autorskie przyznaje jednak pracodawcy pierwszeństwo opublikowania takiego utworu, co oznacza, że z formalnego punktu widzenia autor powinien zwrócić się do pracodawcy o zgodę na publikację utworu. Pierwszeństwo takie wygasa po sześciu miesiącach od przyjęcia utworu. W praktyce pierwszeństwo to nie jest wykonywane – wręcz pracownicy często są zachęcani do publikowania w zewnętrznych wydawnictwach, jedynie z obowiązkiem podania afiliacji (koniecznej do prawidłowej oceny dorobku).

Prawo zezwala ponadto instytucji naukowej na nieodpłatne korzystanie „z materiału naukowego zawartego w utworze". Nie ma jednak jasności, czy oznacza to coś więcej niż ogólnie obowiązujące zasady braku ochrony dla informacji, idei i odkryć. Może jednak dawać podstawy do rozważań, na ile możliwe jest udostępnianie przez instytucję danych badawczych zawartych w pracowniczych utworach naukowych, nawet jeżeli pracownicy zachowają prawa do tych utworów i do doprecyzowania zakresu tego udostępniania np. w regulaminach zarządzania własnością intelektualną uczelni.

W przypadku danych badawczych, o ile uznamy ochronę danego zbioru z uwagi na istnienie wkładu indywidualnej twórczości, sporych trudności może dostarczyć próba ustalenia, czy mamy do czynienia ze „zwykłym" utworem pracowniczym, czy też pracowniczym utworem naukowym. Kolejnym krokiem mogącym

skutkować rozmaitymi interpretacjami będzie próba ustalenia, co jest „materiałem naukowym" zawartym w utworze, z którego zatrudniająca badacza instytucja może korzystać niezależnie od praw przysługujących pracownikowi, a co wymaga dodatkowych ustaleń i zawarcia umowy na korzystanie z praw autorskich badacza.

W porównaniu z prawem autorskim zasady określające przysługiwanie praw *sui generis* są o wiele prostsze. Powstają one na rzecz tego, kto poniósł ryzyko nakładu inwestycyjnego związane z bazą danych („producent"). Wydaje się wobec tego, że pomijając szczególne sytuacje, takim podmiotem będzie zazwyczaj instytucja prowadząca i organizująca badania naukowe (o ile oczywiście w ogóle działa w warunkach „ryzyka inwestycyjnego"). Istotą stosunku pracowniczego jest ograniczenie ryzyka pracownika, wobec czego zdecydowanie trudno będzie przypisać prawa *sui generis* zatrudnionym w ramach instytucji badaczom. Może natomiast pojawić się pytanie, czy ryzyko to nie jest tak naprawdę ponoszone przez podmiot finansujący badania, a nie przez instytucję naukową. Zależeć to będzie oczywiście od konkretnej relacji pomiędzy tymi dwoma podmiotami, rodzaju badań, sposobu zapewniania finansowania i towarzyszących temu warunków. Status producenta nabywa się z mocy prawa, nie można więc się co do tego „umówić", jednak umowa z grantodawcą może precyzować kwestie pomagające ustalić, kto ponosi ryzyko inwestycyjne. Ponadto w takich umowach często znajdują się klauzule „potwierdzające" lub wprost przenoszące prawa „własności intelektualnej" na instytucję prowadzącą badania.

W przypadku zaangażowania w projekt badawczy kilku instytucji pojawia się dodatkowe pytanie o to, której z nich przysługują prawa związane z danymi, czy też, czy dzielą się one między te instytucje. Do ustalenia tego służą te same zasady, które zostały już omówione powyżej, przy czym postanowienia łączących ich umów mogą dostarczać istotnych informacji pozwalających określić spełnienie wymagań koniecznych do nabycia praw bądź też wprost regulować przeniesienie tych praw na konkretną instytucję lub instytucje. Nie jest jednak prawnie wykluczone, że określone prawa będą przysługiwały kliku instytucjom wspólnie, podobnie jak nie jest wykluczone, że będą one współdzielone przez kilka osób. I tak, prawo autorskie posługuje się m.in. pojęciem utworu współautorskiego, czyli takiego, który powstał w wyniku zgodnego współdziałania więcej niż jednej osoby. Chodzi tu wyłącznie o ludzi, współtwórcą nie może być osoba prawna. Jednak w wyniku np. zasad dotyczących utworów pracowniczych lub umów zawartych z jednym ze współtwórców, osoba prawna (instytucja naukowa) może wejść w posiadanie jego udziału w prawach. W przypadku praw *sui generis* z kolei może dojść do ich powstania jednocześnie na rzecz kilku podmiotów, jeżeli one wszystkie w jakimś stopniu ponosiły (współdzieliły) ryzyko inwestycji. Przysługiwanie praw wspólnie oznacza, że udostępnienie i inne sposoby korzystania z danych wymagają jednomyślnej decyzji wszystkich uprawnionych.

Wykorzystywanie w badaniach danych objętych prawami osób trzecich

W niniejszej części omówimy zagadnienia związane z wykorzystywaniem w badaniach danych objętych prawami osób trzecich. Przez „osoby trzecie” rozumiemy tu wszystkie podmioty inne niż badacze oraz zatrudniająca je instytucja naukowa. Osobom takim mogą przysługiwać prawa do danych badawczych lub też wykorzystanie tych danych może wkraczać w ich prawa. Do praw tych zaliczyć należy na pewno omówione już powyżej prawa autorskie i prawa *sui generis*. Nie będziemy więc tu powtarzać zasad pozwalających na ustalenie, komu prawa te przysługują. Omówimy jednak podstawowe zasady dozwolonego użytku oraz wyjaśnimy zakres domeny publicznej jako wyznaczników sytuacji, w których korzystanie możliwe jest bez zgody osoby, której prawa przysługują. Następnie omówimy zasady zawierania umów z podmiotami praw wyłącznych, które potrzebne będą w sytuacjach wykraczających poza dozwolony użytek w odniesieniu do takich utworów i baz danych, które nie przeszły jeszcze do domeny publicznej.

Wykorzystując w badaniach dane objęte cudzymi prawami autorskimi, trzeba mieć świadomość różnicy ograniczeń wynikających z autorskich praw osobistych i majątkowych. Autorskie prawa osobiste nie są ograniczone w czasie (obowiązują nawet po śmierci twórcy) i chronią osobistą więź twórcy z utworem. Nie wynika z nich zakaz korzystania z utworu ani jego udostępniania, jednak wszystkie takie działania nie mogą naruszać tej więzi. Poza takimi oczywistymi kwestiami jak

zachowanie autorstwa utworu, niewprowadzanie w błąd co do tego autorstwa, oznaczanie autorstwa zgodnie z wolą twórcy (nazwiskiem, pseudonimem, anonimowo) należy dopilnować, aby nie doszło do innego naruszenia osobistej więzi twórcy z utworem. Nie można naruszać integralności utworu (np. udostępniać modyfikacje w sposób wprowadzający w błąd, że odbiorca ma nadal do czynienia z oryginałem). Twórca ma także prawo do nadzoru nad sposobem korzystania z utworu, dzięki czemu może przeciwstawiać się prezentacji go w innym kontekście. Innym przykładem autorskiego prawa osobistego jest prawo do pierwszego publicznego udostępnienia utworu – musi to być z twórcą uprzednio uzgodnione. Autorskie prawa osobiste nie dają jednak twórcy swobody blokowania korzystania z jego twórczości. Do ich naruszenia dochodzi bowiem tylko wtedy, gdy więź z utworem została przerwana w sposób obiektywny, a nie tylko w subiektywnym poczuciu jego twórcy. Charakter tej więzi musi być więc oceniany m.in. z uwzględnieniem rodzaju i przeznaczenia utworu, stosowanych w danym obszarze zwyczajów.

Autorskie prawa majątkowe z kolei obowiązują za życia twórcy, po czym wygasają po upływie 70 lat (niekiedy termin ten jest liczony od innej daty niż śmierć twórcy). Po tym czasie utwór należy do domeny publicznej, co oznacza, że można go swobodnie powielać, modyfikować i rozpowszechniać (z zachowaniem praw osobistych twórcy). Do domeny publicznej należą także m.in. materiały i dokumenty urzędowe (od momentu ich powstania), jak również utwory, które powstały w czasie, gdy nie obowiązywało prawo autorskie. Szczególnym przypadkiem przejścia do domeny publicznej są fotografie powstałe przed

1994 r., wobec których nie dopełniono wymaganych wtedy formalności. Prawo wymagało wtedy zastrzeżenia praw na odbitkach, wobec czego brak takiego zastrzeżenia oznacza, że takie fotografie nie są chronione prawem autorskim, nawet jeżeli nie minęło jeszcze 70 lat od śmierci ich autorów.

Prawa *sui generis* podlegają podobnej zasadzie – wygasają po 15 latach od roku, w którym baza danych została sporządzona (lub od roku jej pierwszego publicznego udostępnienia).

Dane badawcze zawierające utwory lub bazy danych należące do domeny publicznej można wykorzystywać swobodnie (oczywiście przy założeniu braku innych praw osób trzecich). Jeżeli natomiast przedmioty te są jeszcze chronione, to możliwy zakres takiego wykorzystywania wyznaczają ustawowe przepisy o dozwolonym użytku, a w przypadku, gdy korzystanie nie mieściłoby się w tym zakresie – konieczne jest uzyskanie zgody osoby uprawnionej.

Dozwolony użytek natomiast to ustawowy zakres swobody, w ramach której możliwe jest korzystanie z chronionego utworu bez zgody uprawnionego. Zakresu tego nie można modyfikować samodzielnie, np. w drodze umowy z uprawnionym (może on jednak udzielić zgody, licencji, której zakres będzie szerszy niż to, co można zrobić z jego utworem w ramach dozwolonego użytku). Wśród przypadków określonych w przepisach o dozwolonym użytku są takie jak cytowanie, dozwolony użytek naukowy, edukacyjny, prawo panoramy (osobny katalog przypadków zawiera ustawa o ochronie baz danych w odniesieniu do

praw *sui generis*). Każdy z nich został sprecyzowany w odpowiednim przepisie ustawy, co jednak nie oznacza, że nie pojawiają się tu wątpliwości interpretacyjne. Co istotne, obowiązek wykazania, że dane działanie mieści się w dozwolonym użytku leży po stronie tego, kto z utworu korzysta i to ta osoba ponosi konsekwencje pomyłki, gdyż wykroczenie poza ten zakres bez zgody twórcy oznacza naruszenie praw wyłącznych.

Dozwolony użytek naukowy pozwala instytucjom naukowym wykorzystywać cudze utwory na potrzeby badań naukowych (szczegółowe zasady są określone w art. 27 prawa autorskiego). Mowa tu jednak o badaniach, a już nie o udostępnianiu danych badawczych zawierających utwory podmiotom spoza kręgu instytucji naukowych. Oznacza to, że co do zasady otwieranie takich danych badawczych będzie musiało odbywać się w oparciu na wyraźnych umowach z uprawnionymi osobami. Prawo autorskie przewiduje dwa podstawowe rodzaje takich umów: przeniesienie autorskich praw majątkowych i licencja. Różnią się one przede wszystkim skutkiem i jego trwałością – przeniesienie powoduje, że podmiotem praw staje się nabywca, podczas gdy licencja to jedynie upoważnienie do korzystania z utworu, do którego uprawniony zachowuje swoje prawa. Istnieją tylko wyjątkowe sytuacje, w których można jednostronnie skutecznie zakwestionować lub odwrócić przeniesienie praw. Mianowicie twórca może odstąpić od umowy z uwagi na swoje „istotne interesy twórcze” (poza koniecznością wykazania „istotności”, jest to obwarowane dodatkowymi ograniczeniami w art. 56 prawa autorskiego).

Pozostałe sytuacje to nieprzystąpienie do rozpowszechniania utworu przez nabywcę, który się do tego zobowiązał, a także udostępnienie w nieodpowiedniej formie lub ze zmianami, którym twórca mógłby słusznie się sprzeciwić (tu także ciężar dowodu tych okoliczności jest po stronie twórcy). Umowa licencyjna także podlega tym przepisom, ale poza tym może zostać rozwiązana przez wypowiedzenie. Wypowiedzenie to proste oświadczenie jednej ze stron, niewymagające wykazywania specjalnych okoliczności, które powoduje, że umowa rozwiązuje się (zwykle po upływie określonego terminu). Ewentualne ograniczenia możliwości skorzystania z wypowiedzenia muszą zostać wyraźnie uzgodnione przez strony i będą skuteczne tylko, o ile nie czynią wypowiedzenia niemożliwym ani nie pozbawiają go sensu.

Niezależnie jednak od rodzaju zawartej umowy powinna ona wyraźnie określać, czego dotyczy (wskazanie utworu) oraz nie pozostawiać wątpliwości co do tego, jaki jest uzgodniony zakres korzystania z utworu (zasada specyfikacji pól eksploatacji czy osobnych sposobów korzystania, takich jak np. druk, udostępnianie online itd.). Umowa obejmuje bowiem tylko taki zakres, który został w niej wyraźnie uzgodniony. Milczenie stron co do niektórych innych kwestii (np. czas trwania umowy, terytorium) skutkuje z mocy prawa włączeniem wynikającej z ustawy regulacji domyślnej, która nie zawsze jest intuicyjna (np. wobec braku określenia terytorium obowiązywania licencji obejmuje ona wyłącznie kraj, w którym siedzibę ma licencjobiorca).

Poza omówionymi wyżej prawami autorskimi i prawami *sui generis*, do praw osób trzecich powiązanych z danymi badawczymi mogą należeć prawa na dobrach osobistych (prywatność, dobre imię, wszelkie inne emanacje szeroko pojętej godności ludzkiej). Prawa na dobrach osobistych są prawami wyłącznymi, lecz niezbywalnymi. Przysługują one zawsze konkretnej osobie, której dotyczą, i wygasają z chwilą jej śmierci. Korzystanie z danych stanowiących wkroczenie w dobra osobiste jest możliwe wtedy, gdy nie jest bezprawne (przez co rozumie się sprzeczność z prawem lub „zasadami współżycia społecznego"). Jest to więc zasadniczo odmienna konstrukcja niż ta przyjęta w odniesieniu do praw autorskich (majątkowych) czy praw *sui generis*. Wśród okoliczności wyłączających bezprawność jest jednak zgoda uprawnionego, co w praktyce prowadzi do podobnego rezultatu – tam, gdzie brak jest podstaw do powołania się na inną okoliczność, konieczne jest uzyskanie wyraźnej, dobrowolnej i świadomej zgody na udostępnianie danych, jeżeli to udostępnienie wkraczałoby w dobra osobiste (np. naruszało czyjąś prywatność). Innymi okolicznościami wyłączającymi bezprawność są działanie na podstawie wyraźnego przepisu prawa, wykonywanie własnego prawa (np. obrona własnego prawa własności), o ile nie jest nadużywaniem tego prawa, a także działanie w ochronie uzasadnionego interesu społecznego (np. społecznie uzasadniona krytyka). W praktyce przy udostępnianiu danych badawczych mogącym wkroczyć w cudze dobra osobiste istotna będzie zwłaszcza ta ostatnia okoliczność. Tam, gdzie brak będzie dobrych argumentów, że interes społeczny przemawia za otwartym udostępnieniem danych (tzn., że nie byłoby wystarczające

np. przetworzenie danych wewnętrznie i udostępnienie samej publikacji), konieczne będzie uzyskanie zgody osoby, której dobra osobiste mogłyby zostać naruszone. W kontekście prac naukowych raczej tylko w specyficznych sytuacjach będzie można wskazać wyraźne przepisy wyłączające bezprawność otwartego udostępniania danych, wkraczające w dobra osobiste, a trudno też byłoby argumentować, że takie udostępnianie jest działaniem w ramach własnego prawa przeważającego zawsze nad ochroną dóbr osobistych.

Do praw osób trzecich mogących mieć związek z danymi badawczymi zaliczyć trzeba ponadto prawa z patentów na wynalazki oraz prawa do wzorów użytkowych. Jest to jednak zwykle związek pośredni, gdyż wynalazki i wzory użytkowe to rozwiązania techniczne (określone produkty materialne lub procesy przemysłowe), dla których dane badawcze mogą stanowić zazwyczaj jedynie podstawę (badacz dokonuje wynalazku w oparciu na zebranych danych, w wyniku ich analizy oraz towarzyszących eksperymentów).

Bardziej bezpośrednia relacja może zachodzić pomiędzy danymi badawczymi a innymi rodzajami własności przemysłowej, takimi jak znaki towarowe czy wzory przemysłowe (których przedstawienia mogą znajdować się w bazach tworzonych przez badaczy, np. w ramach badań marketingowych dotyczących rozpoznawalności znanych marek). Tu jednak z kolei samo udostępnianie takich danych nie będzie zazwyczaj stanowiło korzystania z przedmiotu własności przemysłowej w sposób prowadzący do naruszenia powiązanych z nią praw.

Z prawami osób trzecich nie będą związane sytuacje, gdy wykorzystujemy cudzą tajemnicę przedsiębiorstwa, gdyż chroniona jest ona prawnym (ustawowym) zakazem wykorzystania, z wyjątkiem wskazanych w ustawie sytuacji, a nie żadnym dodatkowo wyodrębnionym prawem wyłącznym. Zakaz ten należy jednak oczywiście szanować, a dopuszczalny zakres wykorzystania tajemnic przedsiębiorstwa, w tym w obszarze badań naukowych, można regulować w umowach z dysponentami tych tajemnic (tzw. umowach o zachowaniu poufności, NDA). Umowy te kreują uprawnienia po stronie dysponenta tajemnicy i podobnie jak wszelkie inne umowy kreujące takie uprawnienia po stronie „osoby trzeciej", określają możliwy zakres korzystania z danych. W niniejszym rozdziale nie będziemy jednak szczegółowo wchodzić w kwestie ograniczeń wynikających z umów. Podlegają one ogólnym zasadom prawa zobowiązań, a przede wszystkim zasadzie swobody umów, zgodnie z którą każde postanowienie niestojące wbrew przepisowi powszechnie obowiązującemu będzie ważne.

Z formalnego punktu widzenia nie można również mówić o „prawach osób trzecich" tam, gdzie prawo reguluje zasady korzystania z zasobów informacyjnych instytucji publicznych. Zasady te, uregulowane obecnie w Ustawie o otwartych danych, a wcześniej w ustawie o ponownym wykorzystaniu informacji sektora publicznego, określają zakres możliwego wykorzystania danych publicznych, jednak nie przyznają żadnej instytucji publicznej praw wyłącznych (ponad te, które mogą jej przysługiwać, jeżeli spełnione zostaną wymagania prawa autorskiego, praw *sui generis* itd.). Punktem wyjścia Ustawy

o otwartych danych jest nie prawo wyłączne, lecz przysługujące każdemu prawo do ponownego wykorzystania informacji sektora publicznego. Przy czym informacje niedostępne publicznie udostępniane są na wniosek. Instytucje publiczne nie mają swobody ustalania zasad wykorzystania informacji, mogą jedynie nałożyć z góry określone w ustawie „warunki" (obowiązek informowania o źródle, czasie wytworzenia i pozyskania informacji; informowania o przetworzeniu informacji; warunki dotyczące odpowiedzialności za informacje; warunki dotyczące przetwarzania danych osobowych). Istnieją szczególne sytuacje, w których instytucja publiczna może odmówić przekazania danych lub nałożyć opłaty. Pewne kategorie informacji podlegają też innym przepisom niż wskazana wyżej ustawa.

Zarządzanie danymi badawczymi stanowiącymi dane osobowe

Podstawowe pojęcia z zakresu ochrony danych osobowych to przede wszystkim „dane osobowe" (wraz z ich szczególnym podzbiorem, „danymi wrażliwymi", omówionym w osobnej części), „administrator" oraz „przetwarzanie". Ich dokładne definicje legalne znajdują się w art. 4 RODO, natomiast poniżej omówimy ich kluczowe elementy.

RODO posługuje się szeroką definicją danych osobowych i uznaje, że są to „informacje o zidentyfikowanej lub możliwej do zidentyfikowania osobie fizycznej". Nie ma zatem żadnego zamkniętego katalogu takich informacji i zawsze zależy to od kontekstu, a zwłaszcza tego, jakie istnieją w danym momencie

możliwości identyfikacji osoby za pomocą związanych z nią informacji. Nie można zatem z góry wykluczać, że w danym zbiorze znajdują się dane osobowe tylko dlatego, że nie ma tam takich „podstawowych" danych jak imię, nazwisko czy adres. Czasem np. zestaw kilku charakterystycznych cech pozwala zidentyfikować osobę równie dobrze i powoduje, że do takich danych należy stosować wymagania wynikające z RODO.

Na podobnej zasadzie oparto definicję „przetwarzania", którym jest „operacja lub zestaw operacji wykonywanych na danych osobowych lub zestawach danych osobowych…". Zatem w zasadzie dowolna czynność mająca za przedmiot dane osobowe podlega przepisom RODO. Spośród takich mniej oczywistych czynności, które RODO wprost kwalifikuje jako przetwarzanie, warto wymienić zbieranie i pobieranie (z czego wynika m.in., że do przetwarzania dochodzi jeszcze przed wykorzystaniem danych do badań), a także usuwanie i niszczenie (co oznacza, że te procesy muszą również odbywać się z poszanowaniem wszystkich obowiązków administratora).

„Administrator" to ten, kto ustala cele i sposoby przetwarzania danych osobowych. Może to być osoba fizyczna lub prawna, ale też organ publiczny, jednostka lub jakikolwiek inny podmiot. Możliwe jest też „współadministrowanie" danymi osobowymi, a dzieje się to wtedy, gdy cele i sposoby ustalane są wspólnie przez kilka podmiotów. Administratorami nie będą więc wszyscy, którzy w jakimkolwiek stopniu zostali zaangażowani w badania wykorzystujące dane osobowe. W przypadku badań prowadzonych przez instytucję naukową będzie to ta instytucja, ale nie

poszczególni zatrudnieni w niej badacze. Administratorem nie będzie też przedsiębiorca, który na zlecenie takiej instytucji przechowuje dane lub zapewnia środki techniczne ich przetwarzania (o ile oczywiście nadal decyzje o celach i sposobach pozostają po stronie instytucji naukowej). Generalnie RODO nie zakazuje administratorowi angażowania innych podmiotów w przetwarzanie danych, jednak to na nim nadal spoczywają wszelkie obowiązki i to administrator będzie z nich rozliczany. Oznacza to m.in. konieczność dobrego doboru tych osób i zapewnienia, że zapewnią oni odpowiednie bezpieczeństwo danych (w przypadku powierzania osobom trzecim przetwarzania danych w imieniu administratora RODO wymaga ponadto zawarcia umowy regulującej wskazane w przepisach kwestie).

Jedną z podstawowych zasad RODO jest wymóg przetwarzania danych w oparciu na konkretnej podstawie prawnej. Katalog takich podstaw prawnych jest zamknięty i określony w art. 6 RODO (osobny katalog przewidziano dla danych wrażliwych). Zgodnie z tym przepisem przetwarzanie jest zgodne z prawem tylko wtedy, gdy zachodzi co najmniej jedna z sytuacji, które skrótowo przedstawiamy poniżej:

a) osoba, której dane dotyczą, wyraziła zgodę;
b) przetwarzanie jest niezbędne do wykonania (lub zawarcia) umowy z osobą, której dane dotyczą;
c) przetwarzanie jest niezbędne do wypełnienia obowiązku prawnego administratora;
d) przetwarzanie jest niezbędne do ochrony żywotnych interesów osoby fizycznej;

e) przetwarzanie jest niezbędne do wykonania zadania realizowanego w interesie publicznym lub w ramach władzy publicznej administratora;
f) przetwarzanie jest niezbędne do celów wynikających z prawnie uzasadnionych interesów realizowanych przez administratora lub stronę trzecią.

Oczywiście nie każda z tych sytuacji będzie mogła zachodzić w kontekście otwartego udostępniania danych badawczych zawierających dane osobowe. Na pewno może to być zgoda na takie udostępnianie, przy czym RODO wymaga, aby było to „dobrowolne, konkretne, świadome i jednoznaczne okazanie woli". Oznacza to, że proces pozyskiwania zgód należy tak zorganizować, aby następnie można było bez przeszkód udowodnić, że mają one wszystkie te cechy. Dodatkowe wymagania zawiera art. 7 RODO, m.in. w przypadku oświadczeń pisemnych, obowiązek przedstawienia zgody „w sposób pozwalający wyraźnie odróżnić je od pozostałych kwestii, w zrozumiałej i łatwo dostępnej formie, jasnym i prostym językiem". Osoby muszą być powiadomione o możliwości wycofania zgody przed jej wyrażeniem, a wycofanie musi być równie łatwe, jak jej wyrażenie. W kontekście dążenia do otwartego udostępniania danych badawczych trzeba też pamiętać o tym, że zgoda na sam udział w badaniu nie może być powiązana ze zgodą na udostępnianie danych.

Poza zgodą, alternatywną podstawą udostępniania może być uzasadniony interes administratora, przy czym to administrator musiałby zgromadzić i obronić argumenty za tym, że otwarte

udostępnianie jest faktycznie niezbędne. Dodatkowym wymogiem RODO w tym wypadku jest konieczność wykazania, że interesy lub podstawowe prawa i wolności osób (np. prawo do prywatności) nie są w tym wypadku nadrzędne. W chwili obecnej nie istnieją natomiast zawarte w powszechnie obowiązującym prawie „prawne obowiązki administratora", z których wynikałby nakaz otwartego udostępniania danych badawczych. Może się to jednak zmienić w przyszłości, co pozwoliłoby na udostępnianie danych bez zgody ani konieczności udowadniania, że jest to niezbędne w konkretnych przypadkach. Z kolei pozostałe wymienione powyżej sytuacje raczej nie będą miały miejsca w kontekście prowadzenia badań.

Istotna część obowiązków administratora danych to obowiązki informacyjne. Przepisy RODO podkreślają, że wszelkie wymagane informacje muszą być przekazywane „w zwięzłej, przejrzystej, zrozumiałej i łatwo dostępnej formie, jasnym i prostym językiem" (art. 12 ust. 1). Preferowana jest pisemna lub elektroniczna postać informacji.

RODO odróżnia obowiązki informacyjne mające zastosowanie w sytuacji, gdy dane zbierane są bezpośrednio od osoby, której dotyczą, od tych, które należy stosować, gdy dane są uzyskiwane od pośredników. W przypadku bezpośredniego zbierania (art. 13) należy podać informacje identyfikujące i kontaktowe administratora (ew. inspektora ochrony danych), cele i podstawę prawną przetwarzania (w tym wyjaśnienie prawnie uzasadnionych interesów, gdy administrator powołuje się na tę podstawę), a także informacje o odbiorcach danych

i ewentualnym zamiarze przekazania danych do państwa trzeciego. Należy ponadto poinformować o okresie przechowywania danych, prawie dostępu, sprostowania, usunięcia lub ograniczenia przetwarzania, a także prawie do wniesienia sprzeciwu i prawie do przenoszenia danych, prawie skargi do organu nadzorczego. W przypadku przetwarzania w oparciu o zgodę, należy poinformować o możliwości jej wycofania. Konieczna jest też informacja o tym, czy podanie danych jest obowiązkiem (i jaka jest podstawa tego obowiązku) i jakie są konsekwencje niepodania danych. Jeżeli administrator podejmuje decyzje w sposób zautomatyzowany (w tym profiluje), należy wyjaśnić zasady tego mechanizmu.

Z kolei obowiązki informacyjne w przypadku zbierania danych od różnego rodzaju pośredników (uzyskanie przez badaczy danych z gotowej bazy zebranej przez kogoś innego) precyzuje art. 14. Zgodnie jednak z ust. 5 lit. b tego przepisu udzielenie informacji nie jest w takim przypadku konieczne, gdy okazuje się niemożliwe lub wymagałoby niewspółmiernie dużego wysiłku, w szczególności w przypadku przetwarzania danych do celów naukowych, o ile obowiązki informacyjne mogłyby uniemożliwić lub poważnie utrudnić realizację tych celów. Zamiast tego, „administrator podejmuje odpowiednie środki, by chronić prawa i wolności oraz prawnie uzasadnione interesy osoby, której dane dotyczą, w tym udostępnia informacje publicznie”.

Prawo o szkolnictwie wyższym (art. 469b) wyłącza niektóre obowiązki administratora w odniesieniu do instytucji naukowych przetwarzających dane osobowe do celów naukowych, jednak

tylko wtedy, gdy zachodzi prawopodobienstwo, że prawa te uniemożliwią lub poważnie utrudnią badania i wyłączenie ich jest konieczne do realizacji tych celów. Administratorzy nie muszą zapewniać możliwości realizacji prawa dostępu do danych, sprostowania, ograniczenia przetwarzania oraz prawa sprzeciwu. W kontekście wskazanych wyżej obowiązków informacyjnych wydaje się właściwe, aby informować osoby, których dane dotyczą, że prawa te im nie przysługują i z jakiego powodu.

Zarządzanie danymi o wysokiej wrażliwości

RODO narzuca podwyższony poziom ochrony dla „danych wrażliwych", wymienionych w art. 9, czyli: „danych osobowych ujawniających pochodzenie rasowe lub etniczne, poglądy polityczne, przekonania religijne lub światopoglądowe, przynależność do związków zawodowych oraz przetwarzania danych genetycznych, danych biometrycznych w celu jednoznacznego zidentyfikowania osoby fizycznej lub danych dotyczących zdrowia, seksualności lub orientacji seksualnej tej osoby". Danych tych przetwarzać nie wolno, o ile administrator nie dysponuje jedną z podstaw prawnych wskazanych osobno w art. 9 ust. 2.

Spośród tych podstaw, w kontekście otwartego udostępniania danych badawczych znaczenie może mieć w zasadzie tylko wyraźna zgoda osoby, której dane dotyczą oraz oczywiste upublicznienie danych przez tę osobę. Z pozostałych podstaw krótkiego omówienia wymaga jeszcze niezbędność ze względu na interes publiczny uregulowany w przepisach prawa. W chwili

obecnej nie istnieją takie przepisy, a musiałaby to być konkretna regulacja, w tym zapewniająca ochronę danych nie tylko na poziomie ich wykorzystania do badań, lecz również późniejszego otwartego udostępniania. Natomiast w art. 469b ust. 2 Prawa o szkolnictwie wyższym dopuszczono przetwarzanie danych wrażliwych na potrzeby badań naukowych, jednak pod warunkiem, że publikowanie wyników tych badań następuje w sposób uniemożliwiający identyfikację danej osoby.

Wśród możliwych podstaw prawnych w art. 9 ust. 2 wymieniono ponadto niezbędność m.in. do celów badań naukowych. RODO wymaga przy tym zachowania zasady proporcjonalności, zachowania istoty prawa do ochrony danych oraz zapewnienia środków ochrony praw osób. Tu także trzeba odróżnić wykazanie tych okoliczności na potrzeby samego prowadzenia badań od późniejszego udostępniania danych badawczych. Administrator danych powinien zatem osobno zadbać o zgromadzenie argumentów dla obu tych przypadków i udokumentowanie ich spełnienia.

Zasady przetwarzania danych wrażliwych mogą w konkretnych przypadkach wynikać dodatkowo z przepisów szczególnych. Przykładem takiej sytuacji jest eksperyment medyczny uregulowany przede wszystkim w art. 21 i nast. Ustawy o zawodach lekarza i lekarza dentysty. Przepisy te wymagają m.in. przeprowadzenia oceny ryzyka (eksperyment badawczy może być przeprowadzony na ludziach, gdy uczestnictwo nie wiąże się z ryzykiem lub gdy ryzyko jest minimalne) oraz oceny korzyści i zasadności. Wprowadzają także szereg szczegółowych

zasad, jak np. to, że eksperymentem może kierować tylko lekarz odpowiedniej specjalizacji i kwalifikacji, z wyjątkiem wskazanych w ustawie przypadków. Art. 24 określa obowiązki informacyjne względem uczestników eksperymentu, w tym konieczność poinformowania o wszelkim przewidywanym dalszym użyciu wyników eksperymentu medycznego, danych oraz materiału biologicznego zgromadzonego w jego trakcie, w tym jego użycia dla celów komercyjnych. Z kolei art. 25 i 25a regulują zasady wyrażania zgody na eksperyment i szczególne sytuacje umożliwiające przeprowadzenie eksperymentu bez zgody uczestników. Natomiast zgodnie z art. 28 wskazanej ustawy informacja uzyskana w związku z eksperymentem medycznym lub badaniem przesiewowym może być wykorzystana do celów naukowych bez zgody uczestnika w sposób uniemożliwiający jego identyfikację. Oznacza to z jednej strony znaczne podwyższenie wymagań związanych z zarządzaniem danymi w odniesieniu do eksperymentów medycznych. Z drugiej natomiast strony zachowana została wynikająca już z RODO zasada, że dane zanonimizowane mogą być udostępniane swobodnie, podczas gdy udostępnianie danych osobowych (tu: wrażliwych danych medycznych) wymagać będzie bezwzględnie zgody osoby, której dane dotyczą.

Ramy prawne wynikające z przepisów o komercjalizacji wyników badań oraz regulaminów zarządzania własnością intelektualną jednostek naukowych

Komercjalizacja wyników badań to w największym skrócie czerpanie z nich korzyści finansowych. Może to polegać na udostępnianiu bezpośrednio takich danych badawczych, które same w sobie mają wartość rynkową. Związek między komercjalizacją a danymi badawczymi może być też pośredni, wtedy gdy dane stanowią podstawę dalszych badań, które prowadzą do opracowania nowego produktu lub metody, mających zastosowanie przemysłowe. Narzędziami prawnymi często wykorzystywanymi przy komercjalizacji są prawa wyłączne, zwłaszcza patenty i prawa własności przemysłowej, ale można też sobie wyobrazić komercjalizację za pomocą praw autorskich lub praw *sui generis* – zależy to bowiem od sytuacji i rodzaju komercjalizowanego produktu lub usługi. Tam, gdzie prawa wyłączne nie są dostępne lub przydatne, sięga się często po ochronę tajemnicy przedsiębiorstwa (np. wniesienie danych badawczych lub wyników analiz jako know-how do spółki celowej, która podejmuje działania zmierzające do zachowania ich poufności, sprzedając jednocześnie efekty ich zastosowania).

Ramy prawne komercjalizacji wyników badań wyznaczają przepisy Prawa o szkolnictwie wyższym oraz oczywiście odpowiednie przepisy regulujące korzystanie z odpowiednich praw wyłącznych lub tajemnicy przedsiębiorstwa. Prawo o szkolnictwie wyższym zobowiązuje uczelnie do przyjęcia

wewnętrznych regulaminów regulujących szczegółowo te kwestie, określając jednocześnie ogólne zasady. Wynika z nich obowiązek badaczy zatrudnionych na uczelni do zawiadamiania o wynikach badań przy jednoczesnym zachowaniu ich w poufności. Uczelnia może podjąć decyzję o komercjalizacji, a w przypadku rezygnacji taką możliwość uzyskuje autor. Regulaminy przyjęte przez uczelnie cechują się różnym stopniem szczegółowości, przy czym wiele z nich powtarza regulacje ustawowe.

Wybór przedmiotu komercjalizacji (dane czy przygotowane w wyniku ich analizy i przetworzenia produkty lub metody) oraz narzędzi prawnych (prawa wyłączne, umowy) wpływa na to, czy i w jakim zakresie możliwe będzie równoległe udostępnianie danych badawczych w sposób otwarty. Możliwość ta nie będzie istniała na pewno tam, gdzie warunkiem koniecznym komercjalizacji jest zachowanie danych w poufności (czyli przede wszystkim wtedy, gdy komercjalizowane są bezpośrednio same dane). We wszystkich pozostałych przypadkach możliwe jest natomiast takie skoordynowanie działań, aby otwartość danych mogła być realizowana równolegle do komercjalizacji. Zaczyna to niekiedy być też obowiązkiem, gdyż niektóre instytucje finansujące badania mające cel komercyjny wymagają jednocześnie otwartości samych badań i ich wyników (np. Fundacja Billa i Melindy Gatesów).

Przykładem praktycznej kwestii z tym związanej jest ustalenie zakresu możliwych do opublikowania informacji w sytuacji, gdy w wyniku badań powstało rozwiązanie kwalifikujące się do

ochrony patentowej jako wynalazek. Jedną z przesłanek patentowalności wynalazku jest nowość, czyli brak publicznej wiedzy o istocie wynalazku przed złożeniem wniosku do Urzędu Patentowego. Patent nie może zostać udzielony na rozwiązanie, które zostałoby przed złożeniem wniosku opisane w publikacji naukowej. W przypadku otwierania danych badawczych należy więc odpowiedzieć na pytanie, czy ich udostępnienie ujawni istotę wynalazku opracowanego na podstawie tych danych. Jeżeli tak, to z ich udostępnieniem należy powstrzymać się do czasu złożenia wniosku patentowego.

Podobnie w przypadku, gdy komercjalizacja realizowana jest co prawda bez pozyskiwania praw wyłącznych, ale w oparciu na poufności i „sprzedaży" know-how, których podstawę stanowią dane badawcze. Otwartość danych może być w takim przypadku realizowana w takim zakresie, w jakim ich opublikowanie nie ujawni istoty komercjalizowanego „know-how". Nie zawsze musi to oznaczać rezygnację z otwartości w ogóle, gdyż surowe dane mogą albo nie pozwalać w prosty sposób „domyślić" się co do zawartego w nich „know-how", albo można względnie łatwo się przed tym zabezpieczyć, odpowiednio przygotowując dane do udostępnienia (zachowując przy tym ich przydatność naukową).

Zarządzanie danymi a kwestie etyczne w projekcie badawczym

W poszczególnych obszarach badań naukowych funkcjonują różnego rodzaju sformalizowane kodeksy lub inne dokumenty wprowadzające zasady etyczne do stosunków prawnych

wiążących badaczy z zatrudniającymi ich instytucjami lub podmiotami zamawiającymi badania.

Przykładem takiego kodeksu jest przygotowany w ramach Europejskiego Stowarzyszenia Badaczy Opinii Publicznej i Rynku ESOMAR „Międzynarodowy kodeks badań rynku i opinii, badań społecznych oraz analityki danych" (wydany pod wspólną egidą ICC oraz ESOMAR[34]. Obowiązuje on członków ESOMAR, ale także wszystkich innych, którzy go przyjęli. „Przyjęcie" kodeksu może polegać na powołaniu się na niego w umowie zawartej pomiędzy instytucją a zamawiającym badanie, może on też być przyjęty przez daną instytucję jako dokument wiążący jej pracowników. Kodeks ten jest zbudowany wokół trzech podstawowych zasad: transparentności względem podmiotów danych, bezpieczeństwa danych i ochrony podmiotów danych. Zasady te są precyzowane w kolejnych postanowieniach kodeksu w sposób, który z jednej strony przypomina takie powszechnie obowiązujące przepisy jak RODO, a z drugiej strony mogą być postrzegane jako podwyższenie i uszczegółowienie standardu staranności w kwestiach takich jak np. informowanie podmiotów danych o tym, jakie dane są zbierane i w jakim celu będą przetwarzane. Kodeks określa też zasady postępowania w obszarach, które w RODO albo nie są uregulowane wprost, albo sposób tej regulacji mógłby rodzić różnice interpretacyjne – jak np. w odniesieniu do „pasywnego zbierania danych".

[34] The ICC/ESOMAR International Code, https://esomar.org/code-and-guidelines/icc-esomar-code [data dostępu: 17.07.2023].

Podobny charakter mają „WAPOR Code of Professional Ethics and Practices" czy opracowany przez Polskie Towarzystwo Socjologiczne Kodeks etyki socjologa[35].

Dokumenty takie mogą nie tylko podwyższać standard ochrony, ale również stymulować naukowców do podejmowania dodatkowych wysiłków w celu usuwania istniejących ograniczeń (prawnych i innych) na drodze do udostępniania danych (np. wtedy, gdy nie ma prawnego obowiązku zbierania danych tak, aby było możliwe ich późniejsze udostępnienie). Przykładem są postanowienia Kodeksu etyki socjologa, który wzywa do jak najszerszego udostępniania wyników badań.

Sformalizowane kodeksy etyczne mogą mieć znaczenie prawne nie tylko wtedy, gdy zostały wyraźnie przyjęte do stosowania w kontekście konkretnego zespołu badaczy lub projektu badawczego. Prawo niejednokrotnie odwołuje się do klauzul generalnych określanych jako „zasady współżycia społecznego", „należyta staranność", „rzetelność naukowa", „dobre praktyki" itp. W przypadku, gdy dana osoba podnosiłaby prawne roszczenia względem badaczy, bardzo często ustalenie zakresu ich ewentualnej odpowiedzialności byłoby uzależnione od tego, jakie dokładnie jest rozumienie tych pojęć w konkretnym przypadku. Z pomocą w ustaleniu tego rozumienia przychodzą funkcjonujące w danym środowisku standardy i zwyczaje, czego kodeksy takie jak wskazane powyżej mogą być bardzo dobrym przykładem.

---

[35] WAPOR Code of Professional Ethics and Practices, https://wapor.org/about-wapor/code-of-ethics/, Kodeks Etyki Socjologa, https://pts.org.pl/wp-content/uploads/2016/04/kodeks.pdf [data dostępu: 17.07.2023].

Oczywiście, im mniej formalny jest charakter takich dokumentów, tym trudniej wykazać, że mają one faktycznie charakter wpływający na treść pojęć prawnych.

Kwestie etyczne wymagają niekiedy od badaczy podjęcia rzeczowej dyskusji na temat udostępniania danych tam, gdzie nie ma takiego formalnego obowiązku bądź też ustalenia dobrego sposobu takiego udostępniania nawet wtedy, gdy z prawnego punktu widzenia każda forma udostępnienia byłaby dopuszczalna. Przykładem takiej sytuacji jest powstrzymywanie się od udostępniania niektórych kolekcji gromadzonych w koloniach przez badaczy pochodzących z dominiów, bez ich odpowiedniego opisania i wyjaśnienia braku obiektywizmu przy tworzeniu tych kolekcji.

Odpowiedzialność z tytułu naruszenia przepisów prawa związanych z zarządzaniem danymi badawczymi

Odpowiedzialność prawna polega na tym, że w związku z zaistnieniem negatywnie ocenianej sytuacji uregulowanej w przepisie prawa dochodzi do zastosowania sankcji względem określonego podmiotu. Rodzaj sankcji, zasady wskazywania odpowiedzialnego podmiotu oraz sposób ustalania i interpretowania przepisów zależy od tego, z jakim rodzajem odpowiedzialności prawnej mamy do czynienia. Wyróżniamy przede wszystkim odpowiedzialność cywilną, karną i administracyjną. Specyficznym rodzajem odpowiedzialności jest ponadto odpowiedzialność dyscyplinarna.

W ramach odpowiedzialności cywilnej wyróżnić można odpowiedzialność deliktową i kontraktową. W tym reżimie odpowiedzialności prawo chroni podmioty prawa prywatnego (osoby fizyczne i prawne) przed negatywnymi skutkami szkód wyrządzanych im działaniami innych takich podmiotów czy to w wyniku popełniania czynów niedozwolonych (działań zakazanych przez powszechnie obowiązujące prawo, zwanych deliktami), czy to w wyniku niewykonywania lub nienależytego wykonania zobowiązań (np. umów). W kontekście udostępniania danych badawczych odpowiedzialność cywilna będzie skutkiem naruszenia np. praw wyłącznych takich jak prawa autorskie, popełnienia deliktów takich jak bezprawne naruszenie dobra osobistego, czy wreszcie naruszenie umów (np. umowy o poufności, umowy przekazania prywatnych zbiorów do badań itd.). Podmiotem ponoszącym taką odpowiedzialność będzie ten, kto dokonał naruszenia, czyli w kontekście badań naukowych przede wszystkim prowadząca badania i udostępniająca ich wyniki instytucja naukowa. Nie jest jednak oczywiście całkowicie wykluczone przypisanie takiej odpowiedzialności indywidualnemu badaczowi, zwłaszcza gdy jego działanie nie odbywałoby się w ramach takiej instytucji. Odpowiedzialność cywilna to przede wszystkim odpowiedzialność majątkowa, odszkodowawcza (niekiedy istnieje możliwość zobowiązania przez sąd do określonego działania, np. zaniechania, złożenia oświadczenia). Zasądzenie odszkodowania wymaga jednak m.in. wykazania wysokości szkody, związku przyczynowego między czynem a szkodą oraz winy sprawcy. W szczególnych przypadkach, jak np. w prawie autorskim, istnieją rozwiązania ułatwiające dochodzenie odszkodowania, zwalniające z konieczności

udowadniania wysokości szkody – w zamiana za to uprawniony musi jednak wykazać, jaką kwotę uzyskałby, gdyby zawarto z nim umowę na korzystanie z utworu (i taką kwotę następnie mnoży się przez dwa).

Odpowiedzialność karna związana jest z popełnieniem przestępstw lub wykroczeń, czyli czynów zabronionych pod groźbą kary określonych w ustawie. W odróżnieniu od odpowiedzialności cywilnej, nie ponosi się jej wobec drugiej, równorzędnej strony – dochodzącym tej odpowiedzialności jest państwo. Odpowiedzialność ta może polegać na pozbawieniu lub ograniczeniu wolności bądź zapłacie grzywny (jest to kara majątkowa, nie będąca jednak „odszkodowaniem" za wyrządzoną szkodę). Odpowiedzialność karną może ponieść jedynie konkretna osoba fizyczna, której można udowodnić winę za popełniony czyn. Odpowiedzialność ta będzie zatem mogła być przypisana indywidualnym członkom zespołu badawczego lub innej osobie z instytucji prowadzącej badania. Jednak potrzebny jest do tego przede wszystkim konkretny przepis karny, którego naruszenie zostanie takiej osobie udowodnione wraz z dowodem winy takiego naruszenia. Udostępnianie danych badawczych jako takie nie stanowi przestępstwa, mogłoby nim być jedynie w dodatkowych okolicznościach wypełniających znamiona np. zniesławienia, przestępstw przeciwko bezpieczeństwu informacji itd.

Odpowiedzialność administracyjna jest podobna do odpowiedzialności karnej o tyle, że zachodzi ona również na linii jednostka (tu: badacz bądź instytucja) – państwo (organ władzy

państwowej). W odróżnieniu jednak od odpowiedzialności karnej nie jest tu potrzebne wykazanie winy sprawcy - wystarcza tylko wskazanie naruszenia obowiązku, nad którym nadzór sprawuje określony organ (np. organ ochrony danych osobowych w odniesieniu do obowiązków określonych w RODO). Źródłem takiego obowiązku może być ustawa, inny akt prawa powszechnie obowiązującego, ale także wydana na ich podstawie decyzja indywidualna. Ponadto o ile w przypadku odpowiedzialności karnej decyzję o sankcji podejmuje niezależny sąd, to w przypadku odpowiedzialności administracyjnej sankcja nakładana jest przez organ administracyjny, a dopiero potem może podlegać ewentualnej sądowej kontroli.

Odpowiedzialność dyscyplinarna z kolei to szczególny reżim odpowiedzialności, uregulowanej osobno w odniesieniu do przedstawicieli niektórych zawodów (np. odpowiedzialność zawodowa lekarzy). Ponadto odpowiedzialności dyscyplinarnej podlegają wszyscy pracownicy w ramach zasad określonych w prawie pracy, a pracownicy naukowi w ramach szczególnych zasad określonych w Prawie o szkolnictwie wyższym. Zatem udostępnianie danych badawczych realizowane z naruszeniem obowiązków pracownika naukowego mogłoby skutkować jego odpowiedzialnością na tych zasadach.

Zadania i zasoby w zarządzaniu danymi badawczymi

Kwestia zarządzania danymi badawczymi może pojawić się w projekcie już na wstępnym etapie przygotowań do jego realizacji, np. w trakcie prac nad wnioskiem grantowym. W konsekwencji już wtedy może wystąpić konieczność przeprowadzenia pierwszych konsultacji z data stewardem.

Najistotniejszymi punktami odniesienia powinny być na tym etapie dokumenty instytucji finansującej i/lub instytucji macierzystej, określające oczekiwania względem zarządzania danymi badawczymi oraz ich udostępniania. Oczekiwania te mogą dotyczyć tych elementów wniosku grantowego, które są

związane z zarządzaniem danymi badawczymi pośrednio (np. punktów formularza wnioskowego dotyczących działań informacyjno-promocyjnych lub spodziewanych rezultatów projektu) lub bezpośrednio (jeśli instytucja finansująca już na etapie składania wniosku oczekuje przygotowania planu zarządzania danymi, tak jak np. Narodowe Centrum Nauki).

W szczególności polityka instytucji macierzystej może przewidywać określone obowiązki związane z zarządzaniem danymi badawczymi również w przypadku tych projektów badawczych, w których nie mamy do czynienia z finansowaniem ze źródeł zewnętrznych, np. badań finansowanych ze środków instytucji lub badań stanowiących podstawę uzyskania stopnia doktora, w szczególności rozpraw doktorskich.

Jeżeli zarówno instytucja finansująca projekt, jak i instytucje, przy których afiliowane są osoby go realizujące, posiadają polityki w zakresie zarządzania danymi badawczymi, należy zwrócić szczególną uwagę na ewentualne kolizje wymagań polityki grantodawcy i instytucji naukowych (np. dotyczące miejsca zdeponowania danych), a także na to, czy określony jest sposób postępowania w razie takich kolizji. Jednocześnie określając na etapie prac nad wnioskiem preferowane miejsca publikacji artykułów naukowych i/lub książek powstałych w ramach projektu, warto sprawdzić, czy te kanały dystrybucji (w szczególności czasopisma naukowe) posiadają własne polityki dotyczące zarządzania danymi. Można następnie przeanalizować możliwość ich uzgodnienia z politykami instytucji finansującej i naukowej, co pozwoli uniknąć trudności

w przyszłości i z wyprzedzeniem opracować odpowiedni sposób postępowania.

Również na etapie przygotowań wniosku grantowego należy zadbać o zabudżetowanie zakupu koniecznej infrastruktury i usług oraz kosztów zatrudnienia odpowiedniego personelu, np. data stewarda umiejscowionego na poziomie samego projektu, jeżeli zidentyfikowane zostaną potrzeby w tym zakresie i umożliwiają to warunki konkursu.

Jeśli instytucja naukowa realizuje własną politykę regulującą kwestie zarządzania danymi i ich udostępniania, to właśnie w takiej polityce należy poszukiwać bardziej szczegółowych regulacji dotyczących podziału odpowiedzialności za zarządzanie danymi pomiędzy poszczególne osoby i działy na różnych poziomach instytucji.

Podsumowując: zarządzanie danymi badawczymi w obrębie projektu powinno być zgodne z projektowym planem zarządzania danymi, który z kolei powinien być zgodny z odpowiednimi politykami instytucji finansującej oraz instytucji, przy której afiliowany jest projekt. Pamiętać też trzeba, że dokument taki powinien być stale aktualizowany w odpowiedzi na modyfikacje w zakresie celów i sposobu realizacji projektu, a także zmiany zachodzące w jego otoczeniu technicznym (np. pojawienie się nowych typów instrumentów badawczych), formalnym (zmiany wynikające z aneksów do umowy grantowej) czy prawnym (zmiany w zakresie obowiązującego ustawodawstwa).

Jednostki uczelni i instytucji badawczych zaangażowane w zarządzanie danymi

Centralną postacią łączącą różne działy instytucji zaangażowane w zarządzanie danymi w projekcie jest data steward. To z nim w pierwszej kolejności badacze powinni kontaktować się w kwestiach związanych z zarządzaniem danymi badawczymi. Data steward może rozwiązać przedstawiony mu problem samodzielnie, albo zasięgnąć w tym celu opinii innych działów instytucji lub skorzystać z dostępnej mu sieci ekspertów (wewnętrznych i/lub zewnętrznych).

Jeśli w instytucji funkcjonuje sieć data stewardów, część działań wspierających naukowców w zarządzaniu danymi badawczymi może być prowadzona w ramach jednostki koordynującej pracę takiej sieci. Na przykład działalność szkoleniowa dotycząca ogólnych aspektów zarządzania danymi badawczymi może być prowadzone przez tego rodzaju jednostkę umiejscowioną np. w bibliotece uczelnianej, podczas gdy działania szkoleniowe dostosowane do potrzeb badaczy reprezentujących określoną dyscyplinę lub specjalizujących się w określonym typie technik badawczych (np. badania jakościowe lub ilościowe) mogą być już prowadzone przez data stewardów umiejscowionych głębiej w strukturze instytucji, np. na poziomie wydziału.

Działania szkoleniowe w zakresie jeszcze ogólniejszych kompetencji cyfrowych mogą uzupełniać istniejące w niektórych instytucjach centra kompetencji cyfrowych. Dotyczy to w szczególności szkoleń z zakresu kompetencji, które mogą

pozwolić na bardziej efektywne wykorzystanie wyspecjalizowanych narzędzi cyfrowych wspomagających zarządzanie danymi badawczymi.

Obok wsparcia w sferze kompetencyjnej, badacze powinni też móc liczyć na pomoc ze strony działu IT na poziomie instytucji i/lub jej części. Pomoc ta dotyczyć może zarówno czysto sprzętowych kwestii związanych z infrastrukturą danych, jak i usług dostępnych pracownikom danej instytucji. W obu wypadkach – sprzętu oraz usług – pomoc może obejmować zarówno rozwiązania dostarczane przez samą uczelnię, jak i zapewniane przez podmioty zewnętrzne i udostępniane pracownikom i/lub studentom realizującym badania.

W obszarze prawnym realizację zadań związanych z zarządzaniem danymi wspierać powinni kompetentni prawnicy. W zależności od rozwiązań dostępnych w danej instytucji, wsparcie może zapewniać dział prawny funkcjonujący wewnątrz instytucji lub zewnętrzna kancelaria. Może ono polegać na opracowaniu wspólnie z data stewardem lub zespołem data stewardów instytucjonalnych poradników i szkoleń dotyczących prawnych aspektów zarządzania danymi badawczymi, a także obejmować indywidualne konsultacje w konkretnych sprawach wychodzących poza ogólne ramy zakreślone w poradnikach czy na szkoleniach.

Kolejny obszar wsparcia dotyczy formalnego rozliczenia projektów. Odpowiednie zarządzanie danymi to coraz częściej wymóg instytucji grantowych, którego spełnienie stanowi

warunek możliwości rozliczenia projektu. W tym zakresie powinni współpracować ze sobą kierownik projektu oraz dział instytucji zajmujący się wsparciem administracyjnym projektów, w razie potrzeby wspierani przez data stewarda oraz obsługę prawną instytucji.

Nieodzownym elementem zarządzania danymi jest również refleksja poświęcona zagadnieniom etycznym. W podstawowym zakresie pomocy dotyczącej tych zagadnień może udzielić data steward. W szczególności przydatna będzie tu pomoc data stewardów profilowanych dziedzinowo, znających kodeksy etyczne w danej dziedzinie oraz ich konsekwencje dla odpowiedniego zarządzania danymi. W odniesieniu do badań, w których uczestniczyć będą ludzie i/lub zwierzęta, konieczne może być również uzyskanie opinii komisji weryfikującej pod względem etycznym projekty badawcze w danej instytucji. Jej zalecenia mogą wprost dotyczyć odpowiedniego sposobu obchodzenia się z danymi.

W razie konieczności zakupu nowych elementów infrastruktury lub usług konieczne będzie wsparcie ze strony działu finansowego lub kwestury, a także działu zamówień publicznych. Jeśli zajdzie potrzeba wyspecyfikowania odpowiedniego sprzętu, wskazana może być współpraca z działem IT, zwłaszcza jeśli mowa jest o sprzęcie, nad którym w przyszłości pieczę sprawować będzie właśnie ten dział.

Na wyższym poziomie abstrakcji odpowiednie zarządzanie danymi badawczymi w projekcie wspierać powinny też władze

centralne instytucji (np. rektor i senat uczelni) oraz władze jej poszczególnych jednostek (dziekani wydziałów i dyrektorzy instytutów). Wsparcie takie powinno polegać przede wszystkim na tworzeniu odpowiednich warunków formalnych, technicznych, kadrowych i finansowych. Zarządzanie danymi badawczymi powinno bowiem wpisywać się w istniejące polityki i praktyki instytucji, w tym te dotyczące informacji i promocji oraz otwartego upowszechniania wyników badań naukowych.

Rozwój data stewardship

Wraz ze wzrostem rangi danych badawczych, jaki dokonuje się w ostatnim czasie, pojawiło się zapotrzebowanie na osoby stanowiące rolę swoistego zwornika szeregu działań i procesów koniecznych do tego, by w cyklu życia danych odpowiednio nimi zarządzać. Osobą taką jest właśnie data steward.

Chyba najpopularniejsza definicja wyrażenia data steward głosi, iż jest to:

> Osoba odpowiedzialna za utrzymanie jakości, integralności i ustaleń dotyczących dostępu do danych oraz metadanych, w sposób spójny z obowiązującym prawem, politykami instytucjonalnymi oraz indywidualnymi zgodami. Data stewardship obejmuje profesjonalne i staranne obchodzenie się z danymi w trakcie wszystkich stadiów procesu badawczego. Data steward dąży do tego, by zagwarantować, że dane są traktowane w odpowiedni sposób we

> wszystkich stadiach cyklu badawczego (tj. projektowania, gromadzenia, przetwarzania, analizy, długoterminowego przechowywania i zabezpieczania (*long term preservation)*, udostępniania danych i ich ponownego wykorzystania [oraz reprodukowalności/ odtwarzalności]”[36].

Inna definicja głosi z kolei, iż data stewardship to:

> Proces i postawa, która sprawia, że postępujemy odpowiedzialnie z naszymi własnymi oraz cudzymi danymi, w trakcie i po zakończeniu początkowego cyklu naukowego wytworzenia i odkrycia[37].

W tym kontekście osobą odpowiedzialną za kreowanie odpowiednich procesów i postaw jest właśnie data steward. Co ważne, rola data stewarda nie ogranicza się tu do okresu realizacji projektu badawczego, ale wykracza poza jego ramy. Projekty badawcze mają swój początek i koniec, ale wytworzone dane trwają znacznie dłużej, a zachowanie ich użyteczności przez lata również może wymagać podjęcia odpowiednich działań (np. konwersji danych do nowych formatów, które przez

---

[36] Jetten M., Grootveld M, Mordant A. i in., *Professionalising data stewardship in the Netherlands. Competences, training and education. Dutch roadmap towards national implementation of FAIR data stewardship* (1.0), Zenodo, 2021, https://doi.org/10.5281/zenodo.4486423 [data dostępu: 17.07.2023].

[37] Mons B., *Data Stewardship for Open Science: Implementing FAIR Principles*, Chapman and Hall/CRC, 2018, https://doi.org/10.1201/9781315380711 [data dostępu: 17.07.2023].

lata zdążyły wyprzeć formaty oryginalne). Polskie badaczki zauważają z kolei, że:

> W środowiskach z wysoce rozproszonymi systemami i projektami, czy też w jednostkach naukowych prowadzących liczne i multidyscyplinarne projekty, zarządca danych (tj. data steward – przyp. autorzy kursu) staje się centralnym punktem kontaktowym dla licznej grupy pracowników naukowych ze względu na coraz bardziej złożone systemy archiwizacji i rosnące ilości danych oraz koszty ich przechowywania[38].

Znaczenie data stewarda polegałoby więc na tym, by zapewnić odpowiednie – zgodne z zasadami FAIR – zarządzanie danymi w organizacji. Kwestią otwartą pozostaje natomiast, w jaki konkretnie sposób odpowiednie procesy zostaną zorganizowane oraz kto powinien przejawiać odpowiednie postawy. Jako że zasady FAIR są właśnie zasadami, a nie zestawem ściśle określonych standardów, zadania data stewarda mają w sobie istotny element twórczy.

Kwestią otwartą pozostaje również to, kto odpowiadać będzie za obsługę tych procesów. W niektórych wypadkach może to być sama osoba pełniąca rolę data stewarda (i w mniejszych jednostkach prawdopodobnie będzie to właśnie ona), w innych jednak takich osób będzie więcej, a wówczas sam data steward skoncentruje się na zadaniach związanych z projektowaniem

[38] Pawłowska M., Wachowicz M., *Wprowadzenie do zarządzania danymi naukowymi*, Difin, 2020.

procesów, szkoleniem zaangażowanych w nie osób czy opracowywaniem zaleceń i wewnętrznych dokumentów (np. regulaminów, polityk) związanych z zarządzaniem danymi.

Ważna jest tu ponadto rola data stewarda jako centralnego punktu kontaktowego. Szczególnie w większych jednostkach, które z jednej strony prowadzą zróżnicowaną działalność badawczą, z drugiej zaś dają swoim pracownikom dostęp do licznych, a zarazem złożonych usług IT – do których dodać trzeba również usługi otwarte, niewymagające dostępu za pośrednictwem organizacji – data steward jest przede wszystkim osobą, która bardzo dobrze zna ekosystem danych badawczych i potrafi się w nim poruszać na tyle biegle, że jest w stanie doradzić pracownikowi (w najlepszym przypadku) konkretny sposób – a najlepiej standard – postępowania lub przynajmniej wskazać kierunek poszukiwań.

Jednocześnie osoba pełniąca tę funkcję sama powinna móc liczyć na współpracę ze strony innych osób i jednostek funkcjonujących w obrębie instytucji. Trzeba tu uwzględnić przede wszystkim:

- Współpracę z badaczkami i badaczami realizującymi badania w instytucji, której dotyczą aktywności podejmowane przez data stewarda. Są oni często pierwszymi odbiorcami jego działań, zarówno tych dotyczących takich aspektów zarządzania danymi, które można określić mianem „twardych” (np. dostępu do infrastruktury oraz usług służących do krótko- i długoterminowego przechowywania danych), jak i tych, które można określić

mianem „miękkich” (np. identyfikacja luk kompetencyjnych i realizacja odpowiednich szkoleń).

- Współpracę i wsparcie ze strony działu odpowiedzialnego za obsługę IT. Działalność ta, w zależności od przyjętego sposobu organizacji wewnętrz instytucji, może obejmować zarówno techniczną obsługę usług IT, jak i tzw. *support* dla innych osób zaangażowanych w proces wytwarzania danych i zainteresowanych wykorzystaniem wewnętrznej infrastruktury instytucjonalnej.
- Współpracę i wsparcie ze strony działu promocji w zakresie działań informacyjnych oraz organizacji wydarzeń szkoleniowych i promocyjnych związanych z zarządzaniem danymi badawczymi.
- Współpracę i wsparcie ze strony działu zajmującego się obsługą prawną instytucji. Wsparcie to powinno dotyczyć zarówno formalnych aspektów dokumentów opracowywanych przez data stewarda (polityk, regulaminów), jak i pojawiających się *ad hoc* wątpliwości prawnych dotyczących konkretnych sytuacji.
- Współpracę i wsparcie ze strony instytucjonalnego inspektora ochrony danych osobowych, który może pomóc w rozstrzygnięciu wątpliwości dotyczących zbiorów zawierających dane osobowe lub tych, które takie dane zawierały, ale zostały zanonimizowane lub poddane procedurze pseudonimizacji.
- Współpracę i wsparcie ze strony działu odpowiedzialnego za komercjalizację wyników badań naukowych – szczególnie ważne w przypadku tych projektów badawczych, w których istnieje rzeczywisty potencjał komercjalizacji.

- Wsparcie ze strony jednostek wspomagających przygotowanie i obsługę wniosków grantowych, które winny współdziałać z data stewardem w zakresie przygotowania, modyfikowania i realizacji planów zarządzania danymi w zakresie zgodności z wymogami instytucji finansujących.
- Współpracę i wsparcie ze strony bibliotek instytucji naukowych. Biblioteki takie często prowadzą już działania w zakresie otwartej nauki i otwartych publikacji, a stąd mogą też być dobrym miejscem rozwijania kompetencji dotyczących otwartych danych badawczych. W przypadku bardziej rozproszonych modeli data stewardship, w przypadku których data stewardzi ulokowani są na poziomie wydziału, instytutu, a nawet pojedynczego projektu, biblioteka akademicka może np. pełnić rolę koordynującą ich działania.

To, jak intensywne powinno być omawiane wsparcie, zależy z jednej strony od możliwości danej instytucji, z drugiej zaś od zapotrzebowania samego data stewarda oraz od posiadanych przez nią lub niego kompetencji. Przykładowo, innego wsparcia ze strony działu prawnego wymagać i oczekiwać będzie data steward wywodzący się ze środowiska technicznego, a innego taki, który posiada wykształcenie prawnicze.

Nieostrość roli data stewarda

Jednocześnie w tekstach poświęconych zarządzaniu danymi badawczymi można napotkać konstatacje mówiące o wielości

treści, jakie mogą kryć się za tym terminem[39], a także o wielości terminów mających oddawać zbliżone treści. Funkcja określana przez jednych jako „data steward", przez innych może być określana jako „data concierge", „data curator", „data manager", „data librarian", „data coordinator", „data handler", a nawet „data scientist". Niekiedy terminy te traktowane są jak synonimy, kiedy indziej z kolei służą do bardziej zniuansowanego opisu ról związanych z zarządzaniem danymi badawczymi (a wtedy powyższe wyrażenia używane byłyby w różnych znaczeniach). Jeszcze inni próbują zniuansować zadania osób określanych jako data stewardzi, wyróżniając np. „dziedzinowych" czy „zakorzenionych" (*embedded*) w obrębie konkretnego projektu data stewardów.

Ponadto obok data stewardów związanych z instytucjami akademickimi terminu tego używa się również w odniesieniu do ról pełnionych w obrębie organizacji innego typu, np. przedsiębiorstw. W tym przypadku również można pokusić się o dalsze zniuansowanie funkcji i wyróżnić procesowych, systemowych, czy biznesowych data stewardów[40].

---

[39] Por. Helling P., Rau F., Linne M. i in., *The Importance of Demand and Environment for Defining and Establishing the Role of Data Stewards. International FAIR Convergence Symposium*, Zenodo, 2022, https://doi.org/10.5281/zenodo.6511185 [data dostępu: 17.07.2023].

[40] Plotkin D., *Data Stewardship. An Actionable Guide to Effective Data Management and Data Governance*, Academic Press, 2021.

Wielość kontekstów, wielość zadań

Powody takiego stanu rzeczy są złożone. Przez długi czas dane nie były traktowane na równi z innymi rezultatami działalności naukowej, takimi jak artykuły naukowe, książki czy patenty. Były one postrzegane jako wytwór czy też – szerzej – element procesu badawczego, który pełni służebną rolę wobec wymienionych wyżej narracyjnych form komunikacji naukowej, które bywają niekiedy nazywane podejściem Artykuł(+)[41]. Z perspektywy wytworów działalności badawczej, dane stanowiły tu suplement do tekstu, wiedza kryła się zaś w samym tekście. Cień tego podejścia do dziś bywa obecny w tych politykach instytucji finansujących i czasopism, w których udostępnienie danych zostaje powiązane z momentem udostępnienia tekstu. Takie podejście od lat jest również (nieintencjonalnie) wspierane przez istniejący system oceny pracowników i instytucji naukowych. Koncentruje się on na tradycyjnie uznanych wytworach pracy naukowej i wykorzystuje mierniki niedostrzegające produktów innego rodzaju:

> [...] wiemy, że cytowanie naszego artykułu niekoniecznie jest skorelowane z wpływem, jaki nasze badania mają na naukę lub innowacje społeczne, ale, co ważniejsze, te klasyczne metody ignorują wartość innych obiektów badawczych, które są ponownie wykorzystywane przez innych. [...] Miernik ten naraża na szwank możliwości kariery młodych naukowców

[41] Por. Mons B., dz. cyt.

w ogólności, w szczególności zaś kariery data stewardów[42].

Owo narażenie na szwank bierze się z jednej strony z tego, iż młodzi naukowcy, ze względu na zbyt krótką obecność w świecie akademickim, zwyczajnie nie zdążyli jeszcze wypracować sobie wysokich wskaźników bibliometrycznych. Z drugiej strony, wielu młodych naukowców na wczesnym etapie swoich karier wytwarza duże ilości danych (np. w postaci kodu), które – choć konieczne do uzyskania finalnych wyników w postaci publikacji – w dużej mierze pozostają niezauważone przez system oceny dorobku. Podobnie sytuacja ma się w przypadku tych naukowców, którzy zdecydowali się na pełnienie roli data stewarda.

Co za tym idzie, również refleksja nad szczegółowym podziałem zadań związanych z obsługą cyklu życia danych jest czymś, co wciąż jest przedmiotem ustaleń. Ustalenia takie nie mają przy tym charakteru odgórnie narzuconej definicji. Częściej są wynikiem „docierania się" osoby pełniącej funkcję określaną jako „data steward" z jej otoczeniem. Rezultat tego procesu będzie zależał w dużej mierze od tego, jak ukształtowane jest to otoczenie. Inna będzie rola stewarda danych w zetknięciu z badaczem posiadającym niskie kompetencje informatyczne, a inna w zetknięciu z badaczką posiadającą duże doświadczenie programistyczne.

[42] Tamże.

Data stewardzi pracują w różnych instytucjach akademickich, które w różny sposób definiują swoje potrzeby w odniesieniu do danych. Różne instytucje mają też różną wielkość i charakteryzują się różnym stopniem specjalizacji. Inne potrzeby związane z danymi będzie mieć niewielki instytut badawczy o bardzo wąskiej specjalizacji, a inne uniwersytet bezprzymiotnikowy, w skład którego wchodzą liczne i bardzo zróżnicowane wydziały.

Nawet w przypadku instytucji zbliżonych wielkością mogą one być różnie zorganizowane i posiadać odmienną wewnętrzną strukturę. Sytuacja instytucji, która już posiada prężnie działający dział IT, dysponujący dobrze zorganizowanym systemem wsparcia użytkowników oraz skuteczny dział promocji, będzie więc różna od sytuacji takiej instytucji, która ich nie posiada. Ta pierwsza we wspomnianych dwóch zakresach będzie mogła wesprzeć swojego data stewarda, podczas gdy w przypadku tej drugiej steward danych może być zmuszony wykorzystywać jedynie własne kompetencje.

Różne osoby określane mianem data stewarda pracują z różnymi danymi, a różne dane wiążą się z różnymi szczegółowymi potrzebami. Stąd też rola data stewarda mającego zajmować się ilościowymi danymi społecznymi na szczegółowym poziomie będzie wymagać innego zdefiniowania niż rola data stewarda zajmującego się danymi krystalograficznymi czy nawet jakościowymi danymi społecznymi.

Wzrost znaczenia danych badawczych i jego konsekwencje

W ostatnich latach znaczenie danych badawczych zmienia się i coraz częściej traktowane są one jako samodzielny atut znajdujący się w posiadaniu instytucji naukowych. Otwartość rozumiana zarówno jako transparentność procedur badawczych i reprodukowalność ich wyników, jak i najszersza możliwa dostępność wytworów badań naukowych (zarówno dla osób spoza świata akademickiego, jak i dla innych badaczy), możliwa jest jedynie o tyle, o ile w odpowiedni sposób uwzględni się w niej rolę danych badawczych.

Rośnie wolumen wytwarzanych danych (zarówno tych badawczych, jak i danych innego rodzaju, które wszakże również mogą stać się przedmiotem namysłu naukowego), ale i możliwości ich cyfrowej analizy. Rośnie też wielkość rynku otwartych danych (184,45 miliarda euro w 2019 r.), liczba związanych z nimi miejsc pracy (1,09 miliona w 2019 r.) oraz potencjał sektora otwartych danych[43].

W rezultacie coraz więcej aspektów życia społecznego może funkcjonować w oparciu na danych. Na dane badawcze – zwłaszcza te, których wytworzenie lub zebranie sfinansowane zostało ze środków publicznych – coraz częściej patrzy się jak na określone dobro czy też wartość. W konsekwencji niewłaściwe zarządzanie takimi danymi lub jego brak może oznaczać

[43] Huyer, E., Knippenberg, L. Van, *The economic impact of open data. Opportunities for value creation in Europe*, European Data Portal, 2020, https://doi.org/10.2830/63132 [data dostępu: 17.07.2023].

zmarnotrawienie jakiegoś fragmentu tego dobra. Dane, którymi nikt nie zarządza i ich nie udostępnia, nie przyniosą nikomu pożytku. Z kolei niewłaściwe zarządzanie danymi może nawet takie marnotrawstwo pogłębić, oznacza bowiem, iż zainteresowane nimi osoby – często młodzi badacze otrzymujący stypendia lub pensje również finansowane ze środków publicznych – muszą poświęcać dużą część swego czasu na wstępne czynności, które mają dopiero uczynić dane zdatnymi do użytku. Z tych samych powodów, choć rosną możliwości automatycznego, maszynowego analizowania danych, jedynie ich niewielka część może być wykorzystana „od ręki", bez żadnych dodatkowych czynności przygotowawczych, które po angielsku określa się terminem *data munging*.

> *Data munging* to proces czyszczenia, agregowania i wzbogacania surowych danych do postaci nadającej się do analizy. Może obejmować wiele zadań, takich jak obsługa brakujących lub nieprawidłowych danych, przekształcenie danych do innych formatów, scalanie danych pochodzących z różnych źródeł i tworzenie nowych zmiennych wspierających analizę. Choć *data munging* może być nużący i zabierać dużo czasu, jest to krytyczny krok w organizacji zadań w obrębie analizy danych, ponieważ jakość naszych wyników będzie zależeć od jakości naszych danych[44].

[44] Jarmul K., Kazil J., *Data Wrangling or 'Munging' with Python*, [w:] *Python for Data Science Handbook*, red. J. VanderPlas, O'Reilly Media, 2017, s. 431.

Role i zakres obowiązków data stewarda na uczelni, na wydziale i w instytucie

Szeroką analizę ról pełnionych przez data stewarda zawierają dwa raporty dotyczące sytuacji w – odpowiednio – Danii i Holandii.

Rola i zakres obowiązków – ujęcie duńskie

Przypadek duński opracowany został w ramach kompleksowego raportu zatytułowanego „National Coordination of Data Steward Education in Denmark. Final report to the National Forum for Research Data Management (DM Forum)”[45], opublikowanego w styczniu 2020 r. Celem raportu było zebranie materiału wspomagającego optymalizację oferty edukacyjnej skierowanej do osób mających pełnić funkcję data stewarda. W badaniach zrealizowanych na potrzeby tego raportu posłużono się bardzo szerokim spektrum technik. Wykorzystane zostały bowiem:

- analiza istniejących programów edukacyjnych przeznaczonych dla data stewardów, zarówno w Danii, jak i poza jej granicami;
- analiza publicznych profili data stewardów w serwisie LinkedIn;
- analiza ogłoszeń o pracę na stanowisku data stewarda;
- badanie kwestionariuszowe, którym objęto osoby pracujące w środowiskach, w których dane pełnią istotną rolę, na wszystkich poziomach organizacji;

[45] Wildgaard L., Vlachos E., Nondal L. i in., *National Coordination of Data Steward Education in Denmark: Final report to the National Forum for Research Data Management (DM Forum)*, Zenodo, 2020, https://doi.org/10.5281/zenodo.3609516 [data dostępu: 17.07.2023].

- wywiady mające na celu określenie różnic pomiędzy oczekiwaniami wobec data stewardów w świecie akademickim oraz w biznesie.

Tak szerokie spektrum technik pozwoliło na wyróżnienie czterech ról pełnionych przez data stewardów:

- administratora,
- agenta zmiany,
- analityka,
- dewelopera.

Administrator odpowiada za opracowanie dobrych praktych w zakresie *compliance* oraz ochrony prywatności. Musi to być osoba zdolna do szybkiej nauki, o analitycznym usposobieniu, skupiona na realizacji zadań i poszukująca wyzwań w zakresie strategicznego rozwoju. Implementuje rozwiązania i szkoli z nich użytkowników końcowych. Osoba pełniąca tę rolę odpowiada za tworzenie polityk oraz zabezpieczeń IT, z uwzględnieniem rozwiązań chmurowych. Jej ważną cechę jest umiejętność pracy zespołowej.

Analityk odpowiada z kolei za zapewnienie jakości danych. Cechuje go znajomość dostępnych rozwiązań chmurowych. Szybko się uczy i potrafi w nowatorski sposób wykorzystywać customowe oprogramowanie oraz bazy danych. Dobrze radzi sobie z wieloma zadaniami; posiada też umiejętności programistyczne związane ze statystyką i analizą danych. Poszukuje wyzwań i pozytywnie podchodzi do obowiązków związanych z raportowaniem.

Deweloper to rola, która związana jest z doradztwem w zakresie zasad FAIR. Pełniąca ją osoba powinna posiadać umiejętności planistyczne i zarządcze w odniesieniu do danych. Powinna skupiać się na współpracy i dzieleniu się wiedzą, być innowacyjna i refleksyjna, a także posiadać umiejętność opracowywania procedur i wskazówek. Deweloper dobrze zna się również na zarządzaniu projektami. Współpracuje w jednym zespole z osobami odpowiedzialnymi za compliance oraz ekspertami zajmującymi się ochroną prywatności, próbując wspólnie z nimi opracować dobre praktyki.

Agenta zmiany charakteryzuje z kolei elastyczne nastawienie. Jest to rola zorientowana na klienta, wymagająca zrozumienia zarówno użytkowników, jak i procesów. Agent zmiany potrafi opracowywać przyjazne dla użytkownika procedury i wytyczne. Szkoli użytkowników w kwestiach związanych z etyką czy odpowiedzialną realizacją badań. Potrafi też pracować w formule projektowej oraz posługiwać się technikami zarządzania zmianą.

Rola i zakres obowiązków – ujęcie holenderskie

Podobne co do celu badania zrealizowano w Holandii, czego efektem stał się raport zatytułowany „Towards FAIR Data Steward as profession for the Lifesciences”[46]. Raport ten również oparty został na zróżnicowanym spektrum technik badawczych, na które składały się:

[46] Scholtens S., Jetten M., Böhmer J. i in., *Towards FAIR data steward as profession for the lifesciences. Report of a ZonMw funded collaborative approach built on existing expertise*, Zenodo, 2019, http://doi.org/10.5281/zenodo.3471708 [data dostępu: 17.07.2023].

- analiza 40 opisów stanowisk z ogłoszeń o pracę, opublikowanych w latach 2016–2019,
- analiza i mapowanie istniejących ram kompetencji dla data stewardów,
- konsultacje eksperckie,
- konsultacje z osobami pełniącymi funkcje data stewardów.

Autorzy raportu zidentyfikowali następujące role data stewardów:

- data steward odpowiedzialny za obszar polityki,
- data steward odpowiedzialny za obszar badań,
- data steward odpowiedzialny za obszar infrastruktury.

Każdej z tych ról przyporządkowane zostały odpowiednie grupy interesariuszy.

I tak, data steward odpowiedzialny za obszar polityki skupia się na tych grupach interesariuszy, które biorą udział w tworzeniu ustaleń dotyczących tego, jak należy obchodzić się z danymi: twórcy polityk, fundatorzy, osoby zarządzające, dziekani czy członkowie kolegialnych ciał uczelni.

Data steward odpowiedzialny za sprawy badawcze styka się natomiast przede wszystkim z naukowcami i data scientistami, którzy wytwarzają dane i pracują z nimi na co dzień, głównie w związku z badaniami naukowymi.

Data steward odpowiedzialny za infrastrukturę kontaktuje się przede wszystkim z dostawcami infrastruktury IT oraz danych (personelem IT, technikami i menedżerami aplikacji).

Jest to grupa interesariuszy dostarczających narzędzia umożliwiające implementację polityk dotyczących danych i ułatwiających zgodne z nimi zarządzanie danymi.

Dla wszystkich tych ról autorzy raportu wyszczególnili osiem obszarów odpowiedzialności:

1. Obszar polityk/strategii – związany z opracowywaniem, wdrażaniem i monitorowaniem polityk i strategii dot. zarządzania danymi.
2. Obszar *compliance* – związany z monitorowaniem zgodności z kodeksami praktyk akademickich czy kodeksami etycznymi.
3. Obszar zapewnienia zgodności z zasadami FAIR, obejmujący również zapewnienie zgodności z zasadami, jakimi kieruje się otwarta nauka.
4. Obszar usług – związany z adekwatnym wsparciem dla zarządzania danymi, obejmujący zarówno dostępność odpowiedniego personelu, jak i usług.
5. Obszar infrastruktury – związany z zapewnieniem dostępności do odpowiedniej infrastruktury dla zarządzania danymi.
6. Obszar zarządzania wiedzą – związany z odpowiednim poziomem wiedzy i umiejętności dotyczących zarządzania danymi na poziomie instytutu, wydziału czy projektu.
7. Obszar sieciowania – związany ze stworzeniem i utrzymaniem sieci kontaktów z ekspertami oraz zewnętrznymi i wewnętrznymi organizacjami.
8. Obszar archiwizacji danych – związany z odpowiednim wsparciem dla długoterminowej archiwizacji danych.

Naturalnie jedna osoba pełniąca rolę data stewarda nie musi posiadać kompetencji we wszystkich tych obszarach. W rezultacie zapewnienie kompleksowych usług związanych z zarządzaniem danymi badawczymi w organizacji tak naprawdę zawsze jest efektem współpracy między wieloma osobami i działami. „Jednorożec danych" – mityczne stworzenie łączące wszystkie potrzebne do tego kompetencje – zwyczajnie nie istnieje[47].

Rodzaje i specjalizacje data stewardów

Zarysowane powyżej modele ról data stewarda umożliwiają ich przełożenie na rodzaje i specjalizacje data stewardów. Pamiętać przy tym należy, iż są to właśnie modele czy też typy idealne, umożliwiające nam koncepcyjne ujęcie analizowanych zagadnień. Nie jest jednak konieczne, aby w zarysowanym kształcie znajdowały one pełne odzwierciedlenie w rzeczywistości – ani tym bardziej, aby jedna osoba łączyła wszystkie wymienione umiejętności i kompetencje.

Powiemy więc np., że jakaś osoba jest data stewardem – administratorem czy analitykiem – jeśli charakterystyka jej zadań w dużej części odpowiada charakterystyce zadań „modelowej" roli administratora lub analityka. Możemy mieć również do czynienia z takimi rodzajami data stewardów, którzy łączą zadania analityka i dewelopera w zarysowanym powyżej sensie.

---

[47] Kvale L.H., *Using Personas to Visualize the Need for Data Stewardship*, „College & Research Libraries" 82(3), 2021, https://doi.org/10.5860/crl.82.3.332 [data dostępu: 17.07.2023].

Nie należy spodziewać się jednak, że będą oni w stanie w pełni realizować obie role w modelowym ujęciu.

Podobnie rzecz ma się z drugą wyszczególnioną wcześniej typologią ról, obejmującą data stewardów odpowiedzialnych za obszar polityki, badań oraz infrastruktury. Również w tym wypadku nie należy się spodziewać, iż będziemy w stanie łatwo odnaleźć data stewarda posiadającego wszystkie przedstawione kompetencje we wszystkich zarysowanych obszarach odpowiedzialności. W praktyce będziemy mieli raczej do czynienia z pewną kombinacją odpowiadających tym rolom zadań i kompetencji.

Poziomy działań data stewarda

Data steward może realizować swoje zadania na różnych poziomach instytucji. Na podstawie analizy dostępnej literatury[48] wyróżnić można w szczególności data stewardów, w wypadku których centrum działań jest biblioteka oraz zakorzenionych (*embedded*) data stewardów, operujących na poziomie wydziału czy instytutu, czy nawet pojedynczej grupy badawczej. Data stewardzi drugiego rodzaju są silnie sprofilowani dziedzinowo. Koncentrują się na dobrych praktykach oraz standardach specyficznych dla danego obszaru badań. W konsekwencji, data stewardzi tego rodzaju muszą nie tylko rozumieć np. kwestie długoterminowego przechowywania danych oraz ogólną rolę metadanych, ale także metody oraz dane, z którymi pracują.

[48] Por. Tamże.

Drugie podejście zidentyfikowane w literaturze przedmiotu dotyczy nowych wyzwań stojących przed bibliotekarzami zajmującymi się danymi (*data librarians*). Rosnące znaczenie zarządzania danymi badawczymi sprawia, że powinni oni rozwijać swoje kompetencje w tym obszarze. W odróżnieniu od zakorzenionych data stewardów nie posiadają oni jednak szczegółowych kompetencji dziedzinowych, pozwalających na bardziej intensywne zaangażowanie w konkretny projekt lub prace konkretnej grupy badawczej.

Podobną typologią posługują się autorzy raportu „Data Stewardship on the map: A study of tasks and roles in Dutch research institutes". I tu spotykamy zakorzenionego (*embedded*) data stewarda, który może bezpośrednio wspierać same badania.

> On lub ona zna odpowiedni obszar badań oraz specyficzne potrzeby koleżanek i kolegów badaczy pracujących w jednostce realizującej badania, a także przekłada ogólną politykę danych na jej praktyczną implementację. Zakorzeniony data steward posiada ekspercką wiedzą dotyczącą sposobów pracy związanych z badaniami i charakterystycznych dla określonej dziedziny. Zakorzeniony data steward może na przykład pomóc przy tworzeniu kodu oprogramowania, skryptów oraz algorytmów do analizy danych[49].

[49] Verheul I., Imming M., Ringerma J. i in., *Data Stewardship on the map: A study of tasks and roles in Dutch research institutes*, Zenodo, 2019, https://doi.org/10.5281/zenodo.2669150 [data dostępu: 17.07.2023].

Na przeciwnym biegunie autorzy lokują z kolei ogólnego (*generic*) data stewarda, który

> pomaga badaczom w wypadku pytań związanych z wszelkimi rodzajami danych lub odsyła ich dalej. On lub ona dostarcza informacji oraz prowadzi szkolenia w odniesieniu do wymogów zawartych w politykach oraz przewodnikach dotyczących danych, a także pomaga w tworzeniu planów zarządzania danymi[50].

Data steward tego rodzaju jest więc z perspektywy badaczy swego rodzaju centralnym węzłem komunikacyjno–informacyjnym w obszarze zarządzania danymi badawczymi.

[50] Tamże.