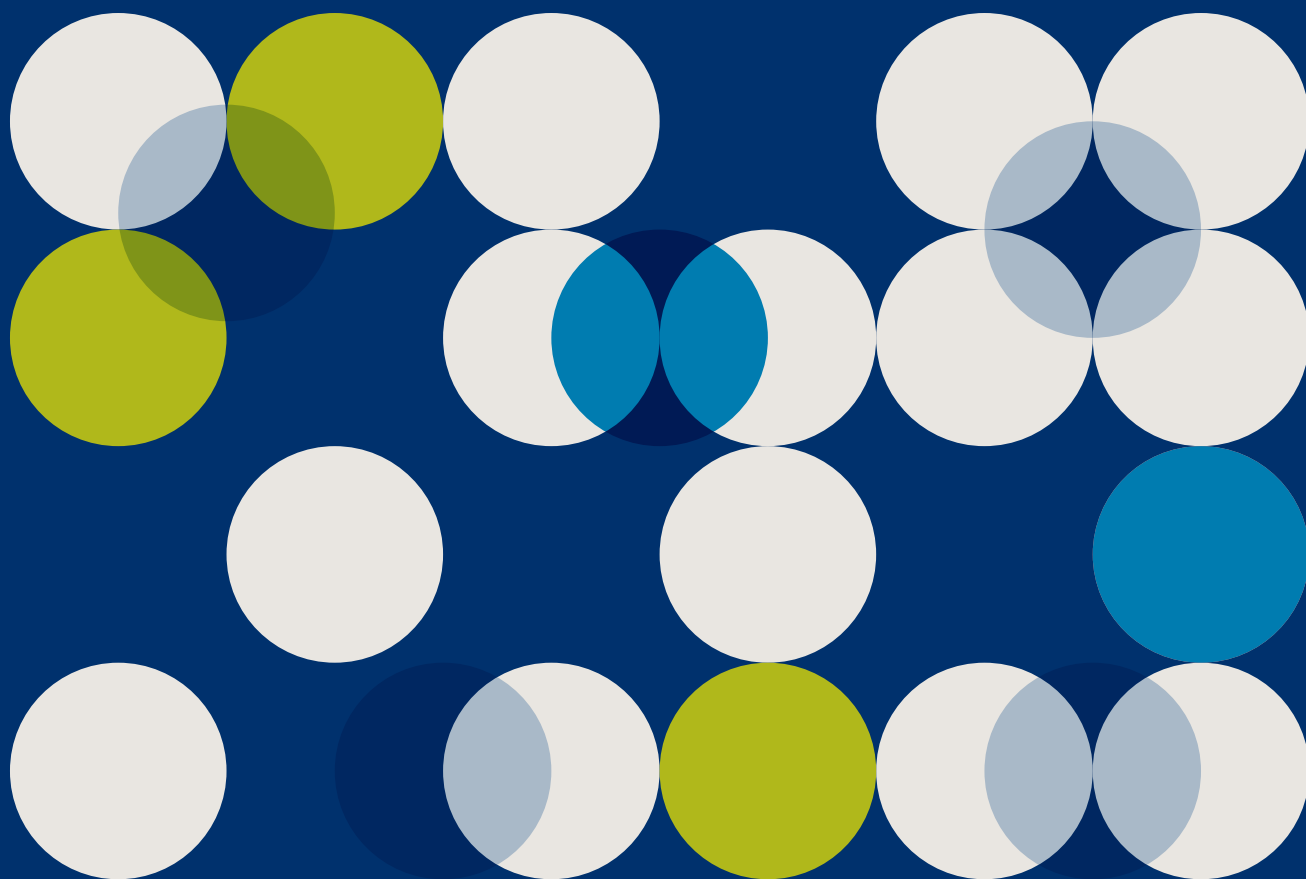


PORADNIK DLA NAUKOWCÓW I DATA STEWARDÓW

Zarządzanie danymi badawczymi



Zarządzanie danymi badawczymi

Poradnik dla naukowców i data stewardów

Wersja 2.0

Opracowanie wersji 1.0: Wojciech Fenrich, Natalia Gruenpeter, Krzysztof Siewicz, Jakub Szprot

Wersja 1.0 poradnika została opracowana w ramach kursów z zarządzania danymi badawczymi udostępnionych na platformie edukacyjnej Navoica (navoica.pl).

Opracowanie wersji 2.0: Magdalena Bielińska, Agnieszka Cybulska-Phelan, Wojciech Fenrich, Jakub Szprot

Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego

Publikacja dostępna na licencji Creative Commons – Uznanie Autorstwa 4.0. Postanowienia licencji dostępne są pod adresem creativecommons.org/licenses/by/4.0/pl/legalcode.



UNIwersYTET
WARSAWSKI



PLATFORMA
OTWARTEJ
NAUKI

Spis treści

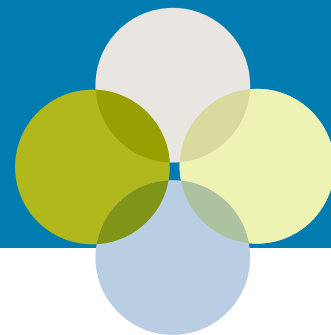


Wprowadzenie	5
Polityki otwartości	9
Komisja Europejska	9
Narodowe Centrum Nauki	11
Zasady FAIR	12
Narzędzia wspierające udostępnianie danych zgodnie z zasadami FAIR	15
Poradniki i wytyczne	16
Planowanie i organizacja	18
Plany zarządzania danymi	20
Plany zarządzania danymi w politykach otwartości Komisji Europejskiej i Narodowego Centrum Nauki	22
Repozytoria danych badawczych	26
Kryteria wyboru repozytoriów i bazy repozytoriów	29
Organizacja danych i zarządzanie ich wersjami	31
Dobre praktyki w zakresie organizacji folderów i plików	31
Przechowywanie danych badawczych	36
Kryteria wyboru miejsca przechowywania danych	36
Miejsca przechowywania danych	37
Przechowywanie danych podczas realizacji projektu	41
Bezpieczeństwo w bieżącej pracy z danymi	42
Przygotowanie danych do udostępnienia	45
Wybór danych do udostępnienia	45
Formaty danych	47
Dokumentacja i opis danych	48
Trwałe identyfikatory i zasady ich stosowania	48
Otwarte formaty	51
Metadane w kontekście standardów dyscyplinowych	56

Spis treści

Ponowne wykorzystanie danych badawczych	59
Czynniki wpływające na ponowne wykorzystanie danych	59
Rzetelne wykorzystanie istniejących danych	61
Replikacja i reprodukcja badań	63
Narzędzia wspierające kontrolę jakości udostępnionych danych	65
Prawne aspekty zarządzania danymi badawczymi	67
Dane badawcze jako przedmiot ochrony prawa własności intelektualnej	70
Prawa członków zespołu badawczego, prawa instytucji prowadzącej badania, prawa uczestników konsorcjum, prawa instytucji finansującej	73
Wykorzystywanie w badaniach danych objętych prawami osób trzecich	76
Zarządzanie danymi badawczymi stanowiącymi dane osobowe	79
Zarządzanie danymi o wysokiej wrażliwości	82
Ramy prawne wynikające z przepisów o komercjalizacji wyników badań oraz regulaminów zarządzania własnością intelektualną jednostek naukowych	83
Zarządzanie danymi a kwestie etyczne w projekcie badawczym	84
Odpowiedzialność z tytułu naruszenia przepisów prawa związanych z zarządzaniem danymi badawczymi	85
Zadania i zasoby w zarządzaniu danymi badawczymi	88
Jednostki uczelni i instytucji badawczych zaangażowane w zarządzanie danymi	89
Rozwój <i>data stewardship</i>	90
Rola i zakres obowiązków <i>data stewarda</i>	96
Słownik terminów z zakresu zarządzania danymi badawczymi	99

Wprowadzenie



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 (ang. *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 utrzymywania¹.

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 lub 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 badawczych szczególne znaczenie mają kwestie techniczne, a także aktywne kuratorowanie zasobów danych (ang. *data curation*) 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.

¹ Por. *Open Science Toolkit*, <https://www.unesco.org/en/open-science/toolkit> [data dostępu: 1.03.2025].

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 i wgląd w szczegóły procesu badawczego. To z kolei daje sposobność wykrycia błędów, nieścisłości czy przypadków nierzetelności naukowej, zarówno na etapie recenzji zapewnianej przez czasopismo naukowe, do którego zgłoszony został manuskrypt, jak i później, kiedy z artykułem zapoznają się czytelnicy.

2.

Większa efektywność badań naukowych

Otwarte udostępnianie danych zgodnie z dobrymi praktykami, zasadami FAIR i właściwymi standardami pozwala na uniknięcie powielania badań, umożliwiając w rezultacie zaoszczędzenie 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.

3.

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.

4.

Poprawa 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.

Wsparcie 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.

Wsparcie 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.



7.

Możliwość wykorzystania otwartych danych w dydaktyce akademickiej

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



8.

Znaczenie otwartych danych dla społeczeństwa

Otwarte udostępnianie danych jest korzystne nie tylko dla badaczy i instytucji naukowych. Może służyć całemu społeczeństwu i różnym grupom zawodowym, w zależności od dziedziny. Z otwartych danych w różnych kontekstach korzystać mogą m.in. organy administracji publicznej, organizacje pozarządowe, nauczyciele i edukatorzy, specjaliści z różnych dziedzin, przedsiębiorcy i wszyscy obywatele.



9.

Możliwość wykorzystania 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.

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 badaczy, w szczególności dotyczące odpowiedzialności za głoszone twierdzenia 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 odnośnych 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.



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.

Polityki otwartości



Ważnymi narzędziami kształtowania otwartej komunikacji naukowej są polityki otwartości, przyjmowane przez instytucje finansujące i prowadzące badania naukowe, a także państwa. Polityka otwartości to zbiór zasad określających sposób otwartego udostępniania rezultatów badań. Może zostać przyjęta na szczeblu krajowym lub instytucjonalnym oraz obejmować różne zagadnienia związane 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 stał się obowiązkowy. W kolejnym kroku polityką objęte zostały dane badawcze. Polityka ta określała zarówno wymogi związane z tworzeniem planu zarządzania danymi, jak i te dotyczące otwartego udostępniania danych. Aktualnie w programie Horyzont Europa obowiązuje kompleksowa polityka otwartości, obejmująca wszelkiego rodzaju 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.

W ramach programu 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: tak otwarte, jak to możliwe, tak zamknięte, jak to niezbędne (ang. *as open as possible, as closed as necessary*).

Wymogi dotyczące planu zarządzania danymi

- 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, w połowie realizacji projektu (jeżeli trwa on dłużej niż dwanaście miesięcy) i przed końcową oceną.

Wymogi dotyczące udostępniania danych badawczych

- Zapewnienie otwartego dostępu zgodnie z zasadą *as open as possible, as closed as necessary*.
- Zdeponowanie danych w zaufanym repozytorium tak szybko, jak to możliwe, bądź w terminach określonych w planie zarządzania danymi.
- Zdeponowanie danych powiązanych z publikacjami naukowymi najpóźniej w dniu 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, narzędziach i instrumentach potrzebnych do weryfikacji danych lub ich ponownego wykorzystania.
- Opracowanie metadanych zgodnie z zasadami FAIR i udostępnianie ich z wykorzystaniem licencji CC0 bądź równoważnej.

W rezultacie doświadczeń związanych z pandemią wirusa COVID-19, 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. Jeżeli zapewnienie takiego dostępu byłoby sprzeczne z uzasadnionymi interesami beneficjentów, należy 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 można znaleźć w dokumentach udostępnionych przez Komisję Europejską².

² 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: 1.03.2025].

Polityka otwartości Narodowego Centrum Nauki obejmuje recenzowane artykuły naukowe oraz dane badawcze³. W 2019 r. Narodowe Centrum Nauki wprowadziło wymóg sporządzania planów zarządzania danymi badawczymi w projektach.

Wymogi dotyczące planu zarządzania danymi

- Na etapie składania wniosku wymagane jest sporządzenie planu zarządzania danymi badawczymi, który podlega ocenie opisowej.
- Plan zarządzania danymi jest bezpośrednio związany z planem badawczym i 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 dotyczące udostępniania danych badawczych

- Dane powiązane z opublikowanymi artykułami (podstawowy zestaw danych) 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) lub na licencji CC BY.
- Dane powinny być udostępniane zgodnie ze standardami cytowania danych zawartych w *Joint declaration of data citation principles* oraz na zasadach zawartych w *Transparency and openness promotion guidelines* (TOP).
- 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.

³ Polityka Narodowego Centrum Nauki dot. otwartego dostępu do publikacji, https://www.ncn.gov.pl/sites/default/files/pliki/zarzadzania-dyrektora/zarzadzanieDyr-38_2020.pdf.
Por. Polityka NCN dot. otwartego dostępu do publikacji – instrukcja, https://www.ncn.gov.pl/sites/default/files/pliki/2021_10_instrukcja_open_access_NCN.pdf
oraz Pismo w sprawie zgłoszenia zapisów dotyczących polityki Narodowego Centrum Nauki dot. otwartego dostępu do publikacji, https://www.ncn.gov.pl/sites/default/files/pliki/pismo_ws_zlagodzenia_zapisow%20dotyczacych_polityki_na_w_ncn.pdf [data dostępu: 1.03.2025].

Zasady FAIR

Akronim FAIR określa wymagania, które powinny spełniać dane badawcze. Zasady te zostały opisane w artykule [FAIR guiding principles for scientific data management and stewardship](#) i stanowią ważny punkt odniesienia, przywoływany często w politykach otwartości. Zasady te rozwinąć można w sposób prezentowany poniżej.

Findable – możliwe do znalezienia

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

Accessible – dostępne

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

Interoperable – interoperacyjne

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

Reusable – możliwe do ponownego użycia

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

Kluczowym 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 badawcze). 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 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.

Warto zauważyć, że wiele z opisanych poniżej aspektów jest związanych z wyborem odpowiedniego miejsca przechowywania i udostępnienia danych, najlepiej przystosowanego do tego repozytorium. Inne działania w pełni zależą natomiast od autorów i osób przygotowujących dane do udostępnienia.



Dane są możliwe do znalezienia, kiedy są dobrze opisane, posiadają trwały identyfikator i zamieszczone zostały w serwisach, których zasoby można przeszukiwać, a ich metadane są indeksowane przez zewnętrzne wyszukiwarki i agregatory.

Co zrobić, aby ułatwić innym znalezienie danych?

- Opatrzeć je trwałym identyfikatorem, najlepiej DOI (ang. *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.



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?

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



Interoperacyjność (meta)danych to możliwość łączenia ich z innymi (meta)danymi, wykorzystywania w wielu rozmaitych systemach komputerowych i analizowania przy użyciu różnego rodzaju 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.



Możliwość ponownego wykorzystania danych to najważniejszy cel zasad FAIR. Osiągnięcie go jest możliwe, kiedy dane są dobrze i rzetelnie opisane, w szczególności towarzyszy im dokumentacja wskazująca na to kto, kiedy i w jaki sposób dane wytworzył lub zebrał. Umożliwia to określenie wiarygodności i rzetelności danych, a także adekwatności 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 odpowiednią dokumentację i udostępnić ją wraz z danymi.
- Korzystać ze standardowych wolnych licencji, które mają postać możliwą do odczytu maszynowego, np. z licencji Creative Commons.

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 (ang. *self-assessment*) zaprojektowane tak, aby wspierać naukowców w podejmowaniu decyzji dotyczących sposobu przygotowania i udostępnienia danych badawczych, jak i narzędzia informatyczne, analizujące stopień zgodności konkretnego zbioru danych z zasadami FAIR.

FAIR-Aware

Narzędzie [FAIR-Aware](#) pomaga ocenić poziom wiedzy na temat zasad FAIR i lepiej zrozumieć, w jaki sposób zgodność z nimi może wpłynąć na oddziaływanie danych badawczych. 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 mogą wykorzystać także osoby prowadzące szkolenia lub przygotowujące materiały szkoleniowe. Narzędzie rozwijane jest przez DANS (Data Archiving and Networked Services), holenderskie centrum kompetencyjno-infrastrukturalne w zakresie przechowywania i udostępniania danych badawczych.

FAIR Data Self Assessment Tool

Umożliwiające przeprowadzenie samooceny [FAIR Data Self Assessment Tool](#) pomaga 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. Zostało stworzone 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

Narzędzie [F-UJI Automated FAIR Data Assessment Tool](#) służy 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 szesnaście parametrów przypisanych do poszczególnych zasad FAIR. Twórcy narzędzia zastrzegają, że jest ono wciąż rozwijane i oparte na ograniczonych 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 [How to FAIR](#) udostępnione zostały materiały edukacyjne, wywiady i quizy, które przybliżają różne aspekty zasad FAIR. Strona została stworzona przez Danish National Forum for Research Data Management przy wsparciu Danish e-Infrastructure Cooperation (DeiC).

FAIR Cookbook – przewodnik

Przewodnik [FAIR Cookbook](#) dotyczy wdrażania rozwiązań zgodnych z zasadami FAIR w naukach o życiu.

FAIR Software – rekomendacje

Zasady FAIR można odnieść do różnych typów obiektów cyfrowych, również do oprogramowania. Strona [FAIR Software](#) zawiera pięć rekomendacji w tym zakresie, a dodatkowo wskazówki i dodatkowe informacje opracowane przez Netherlands eScience Center oraz DANS.

How to make your data FAIR – strona informacyjna

Przewodnik przygotowany przez ekspertów OpenAIRE, [How to make your data FAIR](#), zawiera objaśnienia, materiały szkoleniowe, listę kontrolną dla badaczy przygotowujących się do realizacji projektów i odnośniki do dodatkowych źródeł informacji.

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 (ang. *sharing*) bywa określane jako publikowanie (ang. *publishing*), długoterminowe przechowywanie (ang. *preserving*) rozumiane bywa także jako archiwizacja (ang. *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 danym 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 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 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 i analizy wielu kwestii, 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 powinno uwzględniać 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 dzieć 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. Szablony takie zwykle składają się z pytań dotyczących zarządzania danymi, które 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, z których można skorzystać 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 planów zarządzania danymi ma swoje wady i zalety. Jej zaletą jest to, że 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 przez różne instytucje, z których każda określa własne zasady zarządzania danymi.

Wzorcowe rozwiązania stowarzyszenia Science Europe

Aby ukierunkować tę różnorodność, stowarzyszenie Science Europe opracowało zestaw sześciu kluczowych elementów (zwanymi wymaganiami), jakie zawierać powinien wzorcowy plan zarządzania danymi.

Opis danych oraz zbieranie lub ponowne wykorzystanie istniejących danych

- W jaki sposób nowe dane zostaną zebrane lub wytworzone, oraz/lub jak zostaną ponownie wykorzystane dane, które już istnieją?
- Jakie dane (np. pod względem rodzaju, formatów i wielkości) zostaną zebrane lub wytworzone?

Dokumentacja i jakość danych

- Jakie metadane i dokumentacja (np. dotycząca metodologii zbierania danych oraz sposobu ich organizacji) będą towarzyszyć danym?
- Jakie środki kontroli jakości danych zostaną wykorzystane?

Przechowywanie oraz kopie zapasowe danych w trakcie procesu badawczego

- W jaki sposób dane i metadane będą przechowywane w trakcie procesu badawczego i jak będą tworzone kopie zapasowe?
- W jaki sposób w trakcie badań zostanie uwzględniona kwestia zabezpieczeń oraz ochrony danych wrażliwych?

Wymogi prawne i etyczne, kodeksy postępowania

- Jeśli dojdzie do przetwarzania danych osobowych, w jaki sposób zostanie zapewniona zgodność z ustawodawstwem dotyczącym danych osobowych oraz zabezpieczania danych?
- W jaki sposób zarządzane będą inne kwestie prawne, takie jak prawa własności intelektualnej? Jakie przepisy będą mieć tu zastosowanie?
- W jaki sposób zostaną uwzględnione możliwe kwestie etyczne oraz zgodność z kodeksami postępowania?

Udostępnianie danych oraz ich długoterminowe przechowywanie

- W jaki sposób i kiedy dane zostaną udostępnione? Czy możliwe są ograniczenia w udostępnianiu danych lub istnieją powody do wprowadzenia embarga?
- W jaki sposób zostaną wybrane dane, które będą przechowywane? Gdzie dane będą przechowywane długoterminowo (np. w repozytorium danych lub archiwum)?
- Jakie metody lub narzędzia związane z oprogramowaniem będą konieczne do uzyskania dostępu do danych i ich wykorzystania?
- W jaki sposób zostanie zapewnione przypisanie unikalnego i trwałego identyfikatora (takiego jak DOI) do każdego zbioru danych?

Obowiązki oraz zasoby związane z zarządzaniem danymi

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

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

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

Plan zarządzania danymi badawczymi wymagany jest przez wiele instytucji finansujących badania naukowe, które mogą formułować 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.

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 zasadami określonymi w umowie grantowej i zgodnie z zasadą: tak otwarte, jak to możliwe, tak zamknięte, jak to niezbędne.

Plany zarządzania danymi:

- 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, lecz także publikacjami i innymi rezultatami badań, zarówno cyfrowymi, jak i niecyfrowymi; dotyczy to:
 - oprogramowania stworzonego podczas projektu,
 - procedur,
 - protokołów,
 - materiałów takich jak próbki, linie komórkowe, przeciwiata,
 - 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,
- powinny być udostępniane w sposób otwarty, jeżeli to możliwe,
- mogą być publikowane w odpowiednich czasopismach lub na służących do tego platformach (np. [RIO](#)) albo deponowane w repozytoriach (np. [DMPonline](#)),
- ich pełne wersje nie są wymagane na etapie składania wniosku o finansowanie,
- rekomendowane – ale nie obowiązkowe – jest korzystanie ze wzoru udostępnionego na stronie Komisji Europejskiej.

Elementy planu zarządzania danymi⁴:

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

Narodowe Centrum Nauki

Narodowe Centrum Nauki wymaga złożenia planu zarządzania danymi badawczymi jako elementu wniosku o finansowanie. Aby ułatwić pracę nad planami, NCN przygotowało [wytyczne dla wnioskodawców w języku polskim](#) i [wytyczne w języku angielskim](#).

Elementy planu zarządzania danymi:

- 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.

Zalety tworzenia planu zarządzania danymi

Wśród korzyści, jakie dają plany zarządzania danymi, twórcy przewodnika do zarządzania danymi badawczymi opracowanego przez organizację CESSDA ERIC wymieniają⁵:

- 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.

⁴ 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: 1.03.2025].

⁵ Opracowanie na podstawie CESSDA Data Management Expert Guide, <https://dmeq.cessda.eu/Data-Management-Expert-Guide/I-Plan/Benefits-of-data-management> [data dostępu: 1.03.2025].

- 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.
- Łatwiejsze określenie potrzebnego budżetu. Dzięki DMP łatwiej 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.
- 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.
- 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.

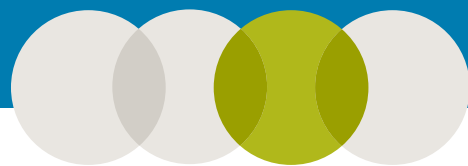
Narzędzia wspierające tworzenie planów zarządzania danymi

W sieci dostępne są serwisy wspomagające tworzenie planów zarządzania danymi. Warto zwrócić uwagę na dwa z nich: [Argos](#) oraz [DMPonline](#).

Zawierają one liczne szablony planów zarządzania danymi. Zalogowani użytkownicy mogą wybrać i wypełnić szablon oraz udostępnić 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.

Repozytoria danych badawczych



Zgodnie z definicją zawartą w *Annotated grant agreement* dla programu Horyzont Europa 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 różne 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

Gromadzą dane wąsko określonego rodzaju i często stawiają bardzo konkretne wymagania dotyczące typu deponowanych plików oraz sposobu ich dokumentacji. Taka specyfika pozwala repozytoriom specjalistycznym na opatrywanie danych metadanymi dobrze przystosowanymi do ich charakteru. Przykładem repozytorium specjalistycznego jest repozytorium PDB (Protein Data Bank), gromadzące wyłącznie dane dotyczące struktur białkowych.

Repozytoria dziedzinowe

Poświęcone danym o szerszym spektrum, służą badaczom z określonej dziedziny (maksymalnie kilku dziedzin). Przykładem jest repozytorium Pangaea, gromadzące informacje z zakresu nauk o Ziemi i środowisku oraz polskie repozytoria: CLARIN (gromadzące dane badawcze z zakresu językoznawstwa) oraz Repozytorium Danych Społecznych, w ramach którego działają Archiwum Danych Jakościowych oraz Polskie Archiwum Danych Społecznych (serwis dla danych ilościowych).

Repozytoria instytucjonalne

Gromadzą dane powiązane z konkretną instytucją. To, na czym w szczególności może polegać takie powiązanie, określa samo repozytorium. Repozytoria instytucjonalne są zwykle zainteresowane danymi 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. Przykładem może być repozytorium danych badawczych Uniwersytetu Warszawskiego – [Dane Badawcze UW](#).

Repozytoria ogólnego przeznaczenia (sieroce)

Są gotowe przyjąć każdy rodzaj danych badawczych. Nazwa *repozytoria sieroce* bierze się z tego, że winny do nich trafiać jedynie te zbiory danych, w wypadku których nie istnieją odpowiednie repozytoria specjalistyczne, dziedzinowe czy instytucjonalne. Przykładem takiego repozytorium jest prowadzone przez ICM UW [Repozytorium Otwartych Danych RepOD](#). Pełni ono jednak również funkcję repozytorium instytucjonalnego szeregu polskich instytucji naukowych. Jednym z największych repozytoriów tego rodzaju jest z kolei repozytorium [Zenodo](#), 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. Poziomy weryfikacji danych mogą być następujące⁶:

- Brak weryfikacji: zbiory danych publikowane są w takiej postaci, w jakiej zostały przygotowane i dostarczone do repozytorium przez użytkownika.
- Podstawowy poziom weryfikacji: repozytorium oferuje elementarne sprawdzenie poprawności elementów zbioru, a także uzupełnienie metadanych lub dokumentacji w podstawowym zakresie.
- Rozszerzony poziom weryfikacji: repozytorium oferuje konwersję do nowych formatów oraz rozszerzenie dokumentacji.
- 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 wypadku repozytoriów ogólnego przeznaczenia, które muszą być przygotowane na przyjęcie bardzo szerokiego spektrum danych.

⁶ CoreTrustSeal Standards and Certification Board, *CoreTrustSeal Requirements 2023-2025* (V01.00), Zenodo, 2022, <https://doi.org/10.5281/zenodo.7051012> [data dostępu: 1.03.2025].

Ograniczenia wielkości danych

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, choć może sprawiać nieco kłopotów po stronie osoby deponującej dane (w wypadku większych zbiorów wymusza bowiem utworzenie wielu mniejszych plików archiwum), to jednak ułatwia 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 wyznaczają 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 podjęcia przez użytkownika 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 – bywa 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, że 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 ułatwiają bazy gromadzące informacje o istniejących serwisach. Na szczególną uwagę zasługują:

Registry of Research Data Repositories

Baza [re3data](#) gromadzi informacje o repozytoriach danych badawczych z całego świata. Za jej prowadzenie odpowiada partnerstwo, w skład którego wchodzi Berlin School of Library and Information Science, Helmholtz Open Science Office, KIT Library, Purdue University Libraries oraz DataCite.

OpenDOAR

W bazie [OpenDOAR](#) można znaleźć informacje o repozytoriach danych badawczych, ale też o repozytoriach publikacji czy bibliotekach cyfrowych.

Open Access Directory

Katalog [Open Access Directory](#) jest prowadzony przez Simmons University.

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 serwisów – są to np. kryteria opracowane przez Science Europe, prezentowane poniżej.

1.

Trwały i unikalny identyfikator (ang. *persistent identifier*)

- umożliwienie odnalezienia i identyfikacji danych badawczych
- umożliwienie wyszukiwania, cytowania i pobierania danych
- umożliwienie wersjonowania danych

2.

Przechowywanie

- zapewnienie trwałości danych i metadanych
- transparentność misji, zakresu, polityk przechowywania oraz planów (w tym dotyczących zarządzania, długookresowej trwałości finansowania, okresu przechowywania danych oraz planów zapewnienia ciągłości działania)

3.

Metadane

- umożliwienie odszukania danych
- umożliwienie odniesienia się do odpowiednich powiązanych informacji, takich jak inne dane i publikacje
- zapewnienie i utrzymanie publicznie dostępnej informacji dotyczącej danych, nawet w wypadku danych nieopublikowanych, chronionych, wycofanych lub usuniętych
- wykorzystanie standardów metadanych, które są szeroko akceptowane przez społeczność naukową
- zapewnienie maszynowej dostępności metadanych

4.

Dostęp i licencje

- umożliwienie dostępu do danych pod dobrze określonymi warunkami
- zapewnienie autentyczności i integralności danych
- umożliwienie pobrania danych
- dostarczenie informacji dotyczących licencji i zgód (najlepiej w formie dostępnej maszynowo)
- zapewnienie poufności i poszanowania praw twórców i osób badanych

Organizacja danych i zarządzanie ich wersjami



Dobra organizacja folderów i plików z danymi badawczymi polega przede wszystkim na przyjęciu określonej konwencji i konsekwentnym jej stosowaniu. Celem jest tu przede wszystkim uniknięcie zamieszania i konfuzji u potencjalnych odbiorców danych (a więc również nas samych, np. jeśli 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, które znaleźć można na stronie Uniwersytetu w Cambridge⁷.

- Należy korzystać z folderów w taki sposób, aby informacje na określony temat znajdowały się w jednym miejscu.
- Warto trzymać się istniejących procedur i nie tworzyć rozwiązań *ad hoc*. Jeśli nasz zespół lub instytucja mają 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.

⁷ Opracowanie na podstawie University of Cambridge, *Data Management Guide*, <https://www.data.cam.ac.uk/data-management-guide/organising-your-data> [data dostępu: 10.3.2025].

- Nazewnictwo folderów powinno odpowiadać obszarowi zadań, którego dotyczy ich zawartość. Złym pomysłem jest natomiast nazywanie folderów w sposób związany 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.
- Warto dbać o spójność i trzymać się raz wybranej konwencji. Pomoże w tym jej jasne określenie na jak najwcześniejszym etapie procesu badawczego.
- 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.
- 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 edytowane.
- 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.
- 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.
- 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.

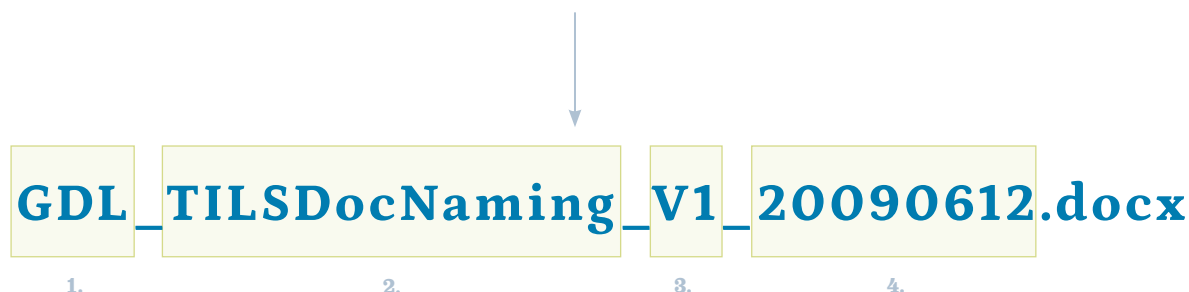
Konwencje nazewnnicze

Uzgodnienie konwencji nazewnniczej 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.

Przyjrzyjmy się przykładowej konwencji nazewnicznej przyjętej w Cambridge⁸. Weźmy nazwę pliku zawierającego wskazówki dotyczące konwencji nazewnicznej:

GDL_TILSDocNaming_V1_20090612.docx



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

1. Pierwszy, trzyliterowy element określa, z jakiego typu dokumentem mamy do czynienia. Litery GDL informują, że są to wskazówki (ang. *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*).
2. Kolejny element to skrócony tytuł dokumentu, którego pełna wersja to *TILS document naming convention*.
3. Następny element określa z kolei wersję dokumentu; w tym wypadku jest to wersja pierwsza.
4. 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.

^a Zob. TILS Document Naming Convention, <https://www.library.qut.edu.au/about/management/documents/OUTILSDocNamingConvention.pdf> [data dostępu: 1.03.2025].

Wersjonowanie

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, że tym, co powinno powodować inkrementację numeru wersji, jest pojawienie się nowej tzw. rewizji (ang. *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 zmiana nie ulega identyfikator DOI przypisany do zbioru danych.

Nieco inaczej podchodzi się do wersjonowania w repozytorium Zenodo. Tu każdy zbiór danych posiada dwa numery DOI: jeden przypisany do konkretnej wersji zbioru i drugi przypisany do abstrakcyjnej koncepcji (ang. *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 zawsze precyzyjnie wskazać wersję, do której się odnosimy, i w cytowaniu posłużyć się numerem DOI przypisanym właśnie do niej. Dopiero jeśli chcielibyśmy 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 na klawiaturze polskie znaki ani jak zostaną one 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 czy znaku minusa. Wariantów tych 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 wielkości poszczególnych plików i zbiorów czy formatów plików akceptowanych przez dane repozytorium.

W wypadku danych o rozmiarach wynoszących kilka terabajtów i więcej, dostęp do nich polegający na tym, że użytkownik pobiera takie dane i analizuje je na własnym komputerze, staje się bardzo problematyczny. Rozmiar taki najczęściej przekracza wielkość dostępnej przestrzeni dyskowej na lokalnej maszynie, co powoduje po stronie użytkownika konieczność zapewnienia macierzy dyskowej, na którą dane można by pobrać, lub odpowiednio dużego dysku przenośnego. 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 umożliwiających ich analizę.

Z kolei z perspektywy osoby zainteresowanej wykorzystaniem takich danych ważne jest, by odpowiednio wcześniej ustalić, na jakich zasadach finansowane są tego rodzaju obliczenia. Możliwe, że ich przeprowadzenie będzie np. wymagać odpowiedniego grantu obliczeniowego lub pokrycia kosztów obliczeń w inny sposób. Warto zadbać o ujęcie tych kwestii 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

Wybierając miejsce, w którym będziemy przechowywać swoje dane, należy rozważyć kilka kwestii.

- 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 długookresowego przechowywania danych badawczych, ponieważ wykazują podatność na utratę i zniszczenie. Co do zasady korzystanie z nich jest dobrą praktyką jedynie 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.

Zewnętrzne dyski twarde, dyski flash i płyty CD

Odpowiednie wyłącznie do tymczasowego, krótkotrwałego przechowywania danych lub do przenoszenia danych, gdy transmisja online nie jest możliwa. Urządzenia powinny być zabezpieczone silnym hasłem i szyfrowaniem, a ich sprawność podlegać regularnej kontroli.

Zalety

- Umożliwiają łatwe przenoszenie danych i plików bez konieczności przesyłania ich przez Internet. Może to być szczególnie pomocne podczas pracy w terenie.
- Stosunkowo niski koszt.

Wady

- 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.

Środki ostrożności dotyczące (wrażliwych) danych osobowych

Należy używać ich w połączeniu z szyfrowaniem i ochroną silnym hasłem.

Usługi chmurowe

Usługi przechowywania danych w chmurze polegają na korzystaniu z zewnętrznych serwerów, do których dostęp uzyskuje się przez Internet z dowolnego miejsca i urządzenia. Ułatwia to 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, lecz 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.

Google Drive, OneDrive, Dropbox, instytucjonalne ownCloud, Nextcloud, Open Science Framework, Tresorit itp.

Przydatne do zapewniania zdalnego i łatwego dostępu do danych wszystkim osobom zaangażowanym w projekt. Nie powinny być jedynym rozwiązaniem stosowanym do przechowywania i tworzenia kopii zapasowych ani służyć do przechowywania nieszyfrowanych danych osobowych. Warunki korzystania powinny zostać sprawdzone pod kątem praw usługodawcy do korzystania z treści. Pierwszeństwo przysługuje usługom europejskim, krajowym lub instytucjonalnym, które przechowują dane w Europie (np. SWITCHdrive – SWITCH, B2DROP – EUDAT).

Zalety

- Automatycznie tworzą kopie zapasowe.
- Często umożliwiają automatyczną kontrolę wersji.

Wady

- Nie wszystkie usługi są bezpieczne. Mogą nie być odpowiednie dla danych osobowych 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 udostępniania treści i wykorzystywania ich do własnych celów.

- Dane mogą zostać utracone, jeśli konto zostanie zawieszone lub przypadkowo usunięte, bądź jeśli dostawca zakończy działalność.

Środki ostrożności dotyczące (wrażliwych) danych osobowych

Należy zaszyfrować wszystkie (wrażliwe) dane osobowe przed przesłaniem ich do chmury. Jest to szczególnie ważne, jeśli nie ma pewności co do lokalizacji serwerów używanych 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.

Urządzenia lokalne

Służbowe komputery i laptopy są stosowane zarówno do bieżącej pracy z danymi, np. z wykorzystaniem odpowiedniego oprogramowania, jak i do przechowywania danych, w szczególności tymczasowych kopii roboczych. 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.

Komputery i laptopy

Odpowiednie tylko w wypadku projektów, w których bierze udział niewiele osób (najlepiej tylko jedna) i w których dane i pliki nie muszą być często przenoszone między komputerami osobistymi.

Plan pracy z wykorzystaniem różnych lokalnych komputerów, np. prywatnego laptopa i komputera stacjonarnego w miejscu zatrudnienia, powinien uwzględniać procedurę kontroli tworzenia kopii zapasowych oraz kontrolę wersji plików.

Zalety

- Umożliwiają sprawowanie pełnej kontroli nad plikami.
- Pozwalają na łatwą ochronę przed nieautoryzowanym dostępem.

Wady

- Jeśli dane są przechowywane tylko na jednym urządzeniu, są narażone na utratę, np. w wyniku awarii, kradzieży, nadpisania.
- Dostęp do danych ma tylko osoba dysponująca komputerem.

Środki ostrożności dotyczące (wrażliwych) danych osobowych

Należy zabezpieczyć komputer hasłem i zadbać o zaszyfrowanie dysku twardego.

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 współdzielone na serwerach instytucji naukowej lub serwerach NAS (Network Attached Storage)

Odpowiednie w projektach realizowanych we współpracy z wieloma osobami, które potrzebują dostępu do danych. Konieczna jest przy tym kontrola dostępu i uprawnień. Powinny być używane w połączeniu z odpowiednią strategią bezpieczeństwa i szczegółowymi zasadami wersjonowania, a także strategią długoterminowego archiwizowania.

Zalety

- Gwarancja centralnego przechowywania danych.
- Możliwość dostępu współdzielonego i zdalnego dla wszystkich osób zaangażowanych w projekt.
- Możliwość centralnego zarządzania kopiami zapasowymi i ich automatycznego tworzenia.

Wady

- Wymagane są wyższe środki bezpieczeństwa, aby zapobiec nieautoryzowanemu dostępowi, przypadkowemu usunięciu plików lub manipulacji danymi.
- Dostęp dla zewnętrznych partnerów projektu może być utrudniony lub niemożliwy.
- Wyższy koszt.

Środki ostrożności dotyczące (wrażliwych) danych osobowych

Należy korzystać w połączeniu z odpowiednią strategią bezpieczeństwa w celu ochrony danych przed nieautoryzowanym dostępem.

Tabele opracowane na podstawie przewodnika CESSDA: [Data management expert guide](#).

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 być związane z przedwczesnym lub nieplanowanym udostępnieniem danych, np. osobom pozbawionym 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 jej ukazaniem się. Zasada ta często znajduje zresztą wyraz w politykach instytucji finansujących i czasopism naukowych, które wymagają udostępnienia danych powiązanych z publikacją najpóźniej w momencie jej ukazania się.

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 kątem⁹.

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.

⁹ Por. University of Cambridge, <https://help.uis.cam.ac.uk/service/security/data-sec-classes>; Harvard University, <https://privsec.harvard.edu/data-security-levels-research-data-examples>; University of California, <https://security.berkeley.edu/resources/how-classify-research-data> [data dostępu: 1.03.2025].

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 jednym w innej lokalizacji geograficznej. Zarówno stworzenie planu zarządzania danymi badawczymi, jak i jego późniejsza realizacja wymagają odpowiedzi na poniższe pytania związane z bezpieczeństwem danych.

- W jaki sposób będą tworzone kopie zapasowe danych?
- Gdzie przechowywane będą kopie zapasowe?
- Jak często będą tworzone kopie zapasowe?
- Ile kopii powstanie?
- 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 prezentowane poniżej 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 dziesięć lat.

Taka perspektywa wiąże się z koniecznością uwzględnienia zachodzącego z czasem procesu degradacji danych (ang. *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 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 udostępnienia 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, że 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, że będą one bezpiecznie przechowywane przez długi czas (najlepiej wieczyście), ale też – w wypadku wybranych repozytoriów – że podejmowane będą dodatkowe działania mające zapewnić dostępność danych również po wielu latach, np. konwersja na nowe formaty.

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 udostępnienia

Jeżeli w ramach projektu badawczego powstały duże ilości danych, możemy stanąć przed dylematem, które z nich – w obliczu ograniczonego budżetu, jaki mamy do 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 udostępniemy.

Odpowiedzi na powyższe trzy pytania mogą pomóc w rozstrzygnięciu dylematu dotyczącego tego, które dane udostępniać, a których 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 pojawić się dodatkowe możliwości pozyskania środków na opracowanie danych. Sprawia to, że te dane, których 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, że 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 danych

Po ustaleniu, jakie dane zamierzamy udostępnić, należy zadbać o nadanie im określonego formatu. Należy sprawdzić, czy repozytorium, na które się zdecydowaliśmy, nie ma w tym zakresie określonych preferencji, oczekiwań lub nawet wymagań.

Wiele repozytoriów preferuje otwarte formaty danych, w wypadku których nie występują zwykłe trudności z 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 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 formaty zamknięte i dysponują odpowiednim oprogramowaniem umożliwiającym korzystanie z nich, jak i osobom, którym w przyszłości może być łatwiej prowadzić analizy z wykorzystaniem formatów otwartych.

Zatem 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 dać pierwszeństwo tym formatom, 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. O tej pierwszej można myśleć jako o swoistej próbce, która pozwoli zainteresowanym na wstępną selekcję danych, których bardziej szczegółowa analiza odbywałaby się już przy użyciu nieskompresowanych plików.

Na szczególną uwagę zasługuje format kompresji archiwów ZIP. Choć nie brak formatów przewyższających go pod pewnymi względami (np. 7-ZIP czy RAR charakteryzują się lepszą kompresją), 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, że 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 zawartość plików.

Dokumentacja i opis danych

Deponowanym danym powinny 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 (ang. *dark deposit*). Jest to rekord w repozytorium pozbawiony plików z danymi, ale posiadający metadane. Rekord taki ma zwykle przypisany trwały identyfikator, co umożliwia jego poprawne zacytowanie. Dzięki temu informacja o danych (choć nie one same) trafia do szerokiego obiegu.

Trwałe identyfikatory i zasady ich stosowania

Rodzaje trwałych identyfikatorów

Historia trwałych identyfikatorów jest długa. Nie potrzebują one 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. Jest 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 – 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 powiązanie z identyfikatorem 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 wprowadzanie jakieś zasoby czy treści, ale które po pewnym czasie przestawały działać np. z powodu przeniesienia plików w inne miejsce czy 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. Istnieje 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. data.cimmyt.org).

Na systemie identyfikatora handle nabudowany jest z kolei najpopularniejszy w świecie naukowym identyfikator, a mianowicie DOI (ang. *digital object identifier*).

Elementy identyfikatora DOI:

- 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 (odpowiadający głównie 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 – dodając informację o autorach zbiorów, możemy np. wskazać identyfikator ORCID każdego z nich. Opisując afiliację 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

Z perspektywy naukowca uzyskanie identyfikatora DOI dla zbioru danych 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 identyfikator 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 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*, czyli strony rozprowadzającej 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. Numer DOI jest jego stałym identyfikatorem – dostęp będzie 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, 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 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 gromadzi m.in. informacje o cytowaniach zawartych w publikacji i opatrzonych numerem DOI.

Otwarte formaty

Jednym z powodów wielozadaniowości współczesnych komputerów jest to, że bity używane do przechowywania informacji mogą być interpretowane na różne sposoby, w zależności od kontekstu, przyjętych standardów 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, natomiast w odniesieniu do całych plików – 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, wolne od ograniczeń licencyjnych i mogą być swobodnie używane, implementowane oraz modyfikowane przez każdego – wtedy mówimy o formacie otwartym. Wykorzystanie innych formatów podlega różnym ograniczeniom, np. ze względu na niedostępność specyfikacji (objęcie tajemnicą przedsiębiorstwa), ograniczoną interoperacyjność wynikającą z zamkniętego projektu, czy konieczność uzyskania licencji na prawa „własności intelektualnej” – wtedy mówimy o formacie zamkniętym.

Wybrane definicje formatów otwartych

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

Pojęcie otwarty format zostało zdefiniowane m.in. przez rząd Stanów Zjednoczonych w [Dyrektywie otwartego rządu](#) (*Open government directive*) przyjętej w 2009 r. Zgodnie z nią format otwarty jest:

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

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

Przyjęte w 2012 r. przez rząd Wielkiej Brytanii [Zasady otwartych standardów](#), 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 na podstawie przemyślanych decyzji,
- 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.

Definicja zaproponowana przez [Projekt informacyjny Linuxa](#) (*Linux information project*, 2004) rozróżnia z kolei formaty otwarte i wolne: format otwarty to *format, który został opublikowany w taki sposób, aby każdy mógł 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 funkcjonował początkowo 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 znaczy, 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

Interoperacyjność

Jak widać 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, wymagającej często dużych nakładów czasu i zasobów.

Długoterminowa archiwizacja danych

Różnorodność platform staje się jeszcze istotniejszą okolicznością, jeśli uwzględnimy nie tylko aktualną sytuację, lecz także jej zmianę w czasie. Ogólnodostępność specyfikacji formatów otwartych połączona z gwarancjami ich trwałości sprawia, że możemy być pewni, że w dającej się przewidzieć przyszłości otwarte formaty 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.

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 zadbane 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, PPTX) • RTF (chroniony prawem autorskim, dokumentacja dostępna publicznie, ale format pozostaje pod kontrolą Microsoftu)	• 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	• AVI	• MP4	• AMV
Archiwa	• 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 wspomaganych komputerowo)

Metadane w kontekście standardów dyscyplinowych

Definicja, rodzaje i funkcje metadanych

Metadane 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 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. Zatem 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 wypadkach 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 poszczególnych zmiennych. 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 metadanych są obowiązkowe, a które opcjonalne, a także jaka powinna być ich kardynalność (czyli minimalna i maksymalna liczba wystąpień danego elementu).

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 mieć tylko jeden identyfikator ORCID – a jednocześnie dopuszczać, by autor miał 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 ma nazwę oraz identyfikator ROR. Grant może mieć też – dokładnie jeden – tytuł oraz numer.

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 identyfikator 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 wymagania 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 wypadku 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. Prezentuje 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.

Serwis Research Data Alliance

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

Serwis FAIRsharing

[FAIRsharing](#) to serwis prowadzony przez Uniwersytet Oxfordzki. Gromadzi 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 do identyfikacji standardów, baz danych lub repozytoriów oraz ich zgodności z politykami instytucji finansujących lub wydawców naukowych.

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 inny¹⁰. Istnieją jednak inne możliwości, które komplikują proste rozróżnienie na wykorzystanie i ponowne wykorzystanie na podstawie wspomnianych kryteriów. Na przykład dane zdeponowane w repozytorium 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 całej 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 zaprezentowano w artykule [Seeing oneself as a data reuser. How subjectification activates the drivers of data reuse in science](https://doi.org/10.5334/dsj-2017-008). Stanowią one dobry punkt wyjścia do omówienia kluczowych czynników wpływających na podatność danych do ponownego wykorzystania.

¹⁰ 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: 1.03.2025].

Czynniki zależne od projektu

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.

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 i 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 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.

Adekwatność danych w stosunku do celów projektu

W zakres tego czynnika wchodzi 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.

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. Na przykład 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 wśród 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 różne instytucje¹¹.

¹¹ 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: 1.03.2025].

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 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)¹².

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](#), 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 powiązаныmi 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 i miejsce udostępnienia, np. nazwa repozytorium lub archiwum, wersja, trwały identyfikator.

¹² Jak korzystać z zasobów w repozytoriach danych, oprac. N. Gruenpeter, Warszawa 2022, <http://pon.edu.pl/dane-materialy> [data dostępu: 1.03.2025].

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](#). Przyjęte w nim definicje uwzględniają kontekst nauk obliczeniowych.

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

Reprodukcja Rozumiana jako przeprowadzenie tych samych analiz na tych samych zbiorach danych (ang. *reproduction*). Wynik uznać można za reprodukowalny (ang. *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 Stabilność wyników (ang. *robustness*) 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 Uogólnienie wyników (ang. *generalisation*) jest rezultatem połączenia stabilnych i zreplikowanych 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 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 sprawdzanie 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),
- 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 i metodami badań. 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. Często 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, 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 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 też zalecenia dotyczące sposobu przygotowania danych mających trafić do konkretnego repozytorium. Zalecenia tego rodzaju (a niekiedy nawet wymagania, które trzeba spełnić, by opublikować dane w określonym serwisie) 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](#), [Zenodo](#) czy na stronie informacyjnej [Polskiego Archiwum Danych Społecznych](#).

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 umożliwiających przygotowanie i/lub walidację danych znaleźć można na stronach specjalistycznego repozytorium [Protein Data Bank](#) czy repozytorium ogólnego przeznaczenia [Dryad](#).

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 przysyłająca 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 po oprogramowaniu Dataverse](#).

Fakt, że 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 do osoby deponującej 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.

Ograniczenia techniczne

Są związane 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 – zostały przyjęte i opublikowane przez organizację normalizacyjną w drodze formalnego procesu, który gwarantuje 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, 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 technicznych sposobów ich wykorzystania. Oczywiście nawet wtedy mogą istnieć inne ograniczenia, w tym techniczne, wynikające chociażby 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 wypadku – jedynie jako wewnętrzna dokumentacja producenta. Skorzystanie z takich specyfikacji jest albo niemożliwe, albo ograniczone do wąskiego kręgu (czasem wyłącznie do producenta jednej aplikacji). Użycie alternatywnego oprogramowania 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 (np. gdy nie jest w stanie zintegrować tego oprogramowania ze swoim systemem).

Do okoliczności technicznych można 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 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 działań w ramach procesu udostępniania – w tym usuwania ograniczeń wynikających z pozostałych okoliczności – oraz koordynacji współpracy różnych interesariuszy w ten proces zaangażowanych. 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. Niekiedy osoby uczestniczące w procesie udostępniania danych nie podlegają bezpośrednio udostępniającemu ani konkretnej polityce otwartości. Implikuje to konieczność dodatkowych uzgodnień i może skutkować specyficznymi ograniczeniami organizacyjnymi. Z drugiej strony nawet hierarchiczna podległość nie wystarczy, jeżeli osoby zaangażowane w proces udostępniania danych nie mają wystarczającej wiedzy, kompetencji czy motywacji do działania. Do wyzwań organizacyjnych należy wobec tego zaliczyć odpowiednie zarządzanie zespołem oraz zapewnienie dobrej komunikacji i współpracy pomiędzy zespołami.

Ograniczenia finansowe

Są związane z koniecznością pokrycia kosztów przygotowania danych do udostępniania oraz kosztów związanych z ich przechowywaniem i przesyłaniem. Kluczowe jest 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 dłuższej perspektywie, czy to przez pozyskanie dodatkowego finansowania, czy to przez deponowanie danych w repozytoriach, które dają gwarancję długoterminowej archiwizacji.

Odpowiednie zarządzanie procesem badawczym może pomóc zredukować koszty. Istotne jest wczesne identyfikowanie ewentualnych trudności i planowanie całego procesu tak, aby jak najdłużej zachować 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), lecz także ewentualne koszty jego zmiany (np. konwersji danych do innego formatu). W tym kontekście możliwość modyfikacji oprogramowania oraz możliwość sięgania po wsparcie w jego rozwoju (istniejące z definicji przy wolnym oprogramowaniu, a niedostępne lub trudno dostępne przy oprogramowaniu własnościowym) oraz swoboda wyboru różnych rozwiązań, związana z otwartymi standardami, nabierają istotnego znaczenia.

Ograniczenia o charakterze prawnym i etycznym

Udostępnienie danych lub określony sposób tego udostępnienia może naruszać czyjeś uprawnienia bądź powszechnie obowiązujące prawo. Szczególnym przypadkiem ograniczenia prawnego będzie sytuacja, w której 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. Skutkuje to koniecznością pozyskania zgody tych osób bądź ograniczenia lub zaniechania podejmowanych działań.

Na uwagę zasługują w pierwszym rzędzie prawa wyłączne, skuteczne 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 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 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 należy rozumieć przede wszystkim sformalizowane kodeksy etyczne obowiązujące badaczy. Mogą one wprowadzać takie ograniczenia udostępniania danych, których nie ma w systemie prawnym. Z drugiej jednak strony mogą 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 obowiązujących umów, np. zatrudnienia w instytucji naukowej czy dofinansowania.

W dalszej kolejności mieszczą się tu różnego rodzaju normy postępowania przyjmowane przez badaczy kierujących się wartościami takimi jak rzetelność naukowa, ochrona osób badanych itp. Normy te nie muszą być zawsze skodyfikowane w postaci kodeksów etycznych. Mogą jednak nabrać prawnego znaczenia, gdyż przepisy prawa często odwołują się do pojęć takich jak *rzetelność naukowa* przy ocenianiu zakresu praw i obowiązków naukowców. Wobec tego dane normy etyczne mogą przełożyć się na prawną odpowiedzialność.

Dane badawcze jako przedmiot ochrony prawa własności intelektualnej

„Własność intelektualna” to niezbyt precyzyjny termin używany zazwyczaj jako 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 osobny wobec „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 dane osobowe nie zaliczają się do „własności intelektualnej”, to podstawowe zasady ochrony tych danych wymagają omówienia.

Nie wszystkie rodzaje praw „własności intelektualnej” mają znaczenie w odniesieniu do każdego zbioru danych. Za najczęściej spotykane ograniczenia można uznać te wynikające z praw autorskich, praw *sui generis* oraz z ochrony danych osobowych (ta ostatnia, jak widzieliśmy, 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 równoległej prowadzonej komercjalizacji wyników badań.

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. 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 utwór nie jest efektem działań odtwórczych, prowadzących do powtarzalnych rezultatów).

Ochrona prawa autorskiego przysługuje automatycznie – nie wymaga zgłaszania utworu do żadnej instytucji ani badania stopnia twórczości (w odróżnieniu od np. ochrony patentowej). Jest zatem dość prawdopodobne, że dane dobro intelektualne jest utworem. Nie istnieje jednak możliwość uzyskania jednoznacznego potwierdzenia takiego stanu rzeczy – oczywiście poza ewentualnym rozstrzygnięciem sądowym. Z kolei ocena ekspercka w tej sytuacji będzie tylko prywatnym zdaniem danego eksperta. Własna ocena musi więc w praktyce wystarczyć.

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 (nie ma jednak wyraźnej definicji pozwalającej ustalić granicę między tymi pojęciami). Na pewno nie są chronione suche fakty, informacje ani odkrycia naukowe. Nie będą więc chronione dane polegające na mechanicznej, odtwórczej rejestracji rzeczywistości. Jednocześnie jednak twórcza aranżacja czy 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 przy tym pamiętać, 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 samo 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. Ochronie może podlegać dowolny zbiór, o ile tylko wykazany zostanie nakład inwestycyjny. Pojęcie nakładu inwestycyjnego może nie nastroczać szczególnych trudności, natomiast ustalenie, czy jest on *istotny*, może już 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 czy urzędowego poświadczania ochrony *sui generis*, podobnie jak w wypadku prawa autorskiego może pojawić się trudność w ustaleniu, czy dany przedmiot jest chroniony, czy nie. Zgodność z kryteriami nakładu inwestycyjnego i ponoszenia ryzyka inwestycyjnego należy 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 może działać na wielu poziomach zbioru danych. Poszczególne dane w zbiorze mogą np. 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. Można wobec tego mówić o ochronie kumulatywnej i wielopoziomowej. Kolejne reżimy prawne mogą dodatkowo komplikować tę kwestię.

Powyższe zasady ochrony wynikają z przepisów ustawowych. Choć są one w dużym stopniu zharmonizowane przez prawo Unii Europejskiej oraz prawo międzynarodowe, to szczegółowe rozwiązania mogą różnić się w zależności od kraju, którego prawu podlega 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 w istotnym zakresie zastąpiło różne krajowe ustawy. W związku z tym, że rozporządzenia unijne są bezpośrednio skuteczne (bez konieczności implementacji do prawa krajowego), występuje tu znacznie wyższy stopień harmonizacji.

Jak wspomniano, ochrona danych osobowych nie jest zaliczana do prawa „własności intelektualnej”, inna jest też jej konstrukcja – skupia się przede wszystkim na określeniu obowiązków administratora danych, których podstawowa egzekucja zapewniana jest drogą administracyjną.

Podstawowe zasady ochrony danych osobowych, wyznaczające m.in. zakres obowiązków administratorów danych, zostały ujęte w art. 5 RODO. Są to:

1. wymaganie zapewnienia przetwarzania zgodnie z prawem, rzetelnie i przejrzystość dla osoby, której dane dotyczą;
2. zasada ograniczonego celu: dane osobowe muszą być zbierane w konkretnych, wyraźnych i prawnie uzasadnionych celach i nieprzetwarzane 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 danych gromadzonych w ramach badań.

W skład praw autorskich wchodzi 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 oryginału. 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. W polskim prawie nie ma wyjątków od tej zasady.

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 wcześniej nie zostały przez twórcę przeniesione np. na wydawcę.

Wyjątkiem od zasady powstawania praw po stronie twórców utworu jest sytuacja pracowniczych programów komputerowych. 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. W wypadku lakonicznych umów o pracę mogą powstać wątpliwości co do zakresu nabycia, więc mimo wszystko wskazane jest rozwinięcie tych kwestii w postanowieniach umowy. 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.

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

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 – pracownicy często są wręcz 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, czy dopuszczalne jest udostępnianie przez instytucję danych badawczych zawartych w pracownicznych utworach naukowych, nawet jeżeli pracownicy zachowają prawa do tych utworów i do doprecyzowania zakresu tego udostępniania, np. w regulaminach zarządzania własnością intelektualną uczelni.

W wypadku danych badawczych, o ile uznamy ochronę danego zbioru z uwagi na wkład 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 także 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*). 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, dlatego 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. Zależać to będzie oczywiście od relacji pomiędzy tymi dwoma podmiotami, rodzaju badań, sposobu zapewniania finansowania i towarzyszących temu warunków. Status producenta nabywa się z mocy prawa, nie można więc co do tego poczynić własnych ustaleń, jednak umowa z grantodawcą może precyzować kwestie pomagające ustalić, kto ponosi ryzyko inwestycyjne.

Jeżeli w projekt badawczy zaangażowanych jest kilka instytucji, pojawia się dodatkowe pytanie, której z nich przysługują prawa związane z danymi lub czy dzielą się one między te instytucje. Zastosowanie mają tutaj zasady omówione powyżej, przy czym postanowienia umów mogą określać wymagania konieczne do nabycia praw lub wprost regulować przeniesienie tych praw na konkretną instytucję. Określone prawa mogą przysługiwać kilku instytucjom wspólnie, podobnie jak nie jest wykluczone, że mogą być 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 pracownicznych lub umów zawartych z jednym ze współtwórców, osoba prawna (instytucja naukowa) może wejść w posiadanie udziału tego współtwórcy w prawach. Prawa *sui generis* z kolei mogą powstać jednocześnie na rzecz kilku podmiotów, jeżeli wszystkie one w jakimś stopniu ponosiły (współdzieliły) ryzyko inwestycji. Przysługiwanie praw wspólnie oznacza, że zarówno udostępnienie, jak i inne sposoby korzystania z danych wymagają jednomyślnej decyzji wszystkich uprawnionych.

Wykorzystywanie w badaniach danych objętych prawami osób trzecich

Osoby trzecie należy tu rozumieć jako wszystkie podmioty inne niż badacze oraz zatrudniający ich instytucja naukowa. Osobom takim mogą przysługiwać prawa do danych badawczych – takie jak prawa autorskie i prawa *sui generis* – a wykorzystanie danych może w te prawa wkraczać.

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 żadne z tych działań nie może naruszać tej więzi. Poza oczywistymi kwestiami, takimi jak zachowanie autorstwa utworu, niewprowadzanie w błąd co do tego autorstwa czy oznaczanie autorstwa zgodnie z wolą twórcy (nazwiskiem, pseudonimem, anonimowo) należy też dopilnować, aby nie doszło do innego naruszenia osobistej więzi twórcy z utworem. Nie można naruszać integralności utworu, np. udostępniać modyfikacji w sposób wprowadzający w błąd, sugerując odbiorcy, że ma nadal do czynienia z oryginałem. Twórca ma także prawo do nadzoru nad sposobem korzystania z utworu, a dzięki temu może przeciwstawiać się jego prezentacji w niepożądanym kontekście. Innym przykładem autorskiego prawa osobistego jest prawo do pierwszego publicznego udostępnienia utworu – musi być ono uzgodnione z twórcą. 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 utworu i jego przeznaczenia oraz stosowanych w danym obszarze zwyczajów.

Autorskie prawa majątkowe obowiązują za życia twórcy, a wygasają po upływie siedemdziesięciu lat od jego śmierci. 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) oraz 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, to znaczy nie umieszczono zastrzeżenia praw na odbitkach. Takie fotografie nie są chronione prawem autorskim, nawet jeżeli nie minęło jeszcze siedemdziesiąt lat od śmierci ich autorów.

Prawa *sui generis* podlegają podobnej zasadzie – wygasają po piętnastu 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 braku innych praw osób trzecich). Jeżeli natomiast są one jeszcze chronione, to możliwy zakres 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 jest 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 ramy dozwolonego użytku. Ramy te zaś obejmują m.in. cytowanie, dozwolony użytek naukowy, edukacyjny czy 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. Obowiązek wykazania, że dane działanie mieści się w zakresie dozwolonego użytku, leży po stronie tego, kto z utworu korzysta. 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 określa art. 27 Ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych. Mowa tu jednak o badaniach, ale 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 na podstawie umów 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. Twórca może odstąpić od umowy z uwagi na swoje *istotne interesy twórcze* (wiąże się to z koniecznością wykazania istotności, a dodatkowo jest obwarowane ograniczeniami wskazanymi 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 po stronie twórcy jest ciężar dowodu tych okoliczności. 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 powyżej 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. W praktyce prowadzi to do podobnego rezultatu – tam, gdzie brakuje 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. naruszałoby 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 nie ma dobrych argumentów, że interes społeczny przemawia za otwartym udostępnieniem danych (tzn. że nie byłoby wystarczające np. przetworzenie danych wewnątrznie 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. Trudno też byłoby argumentować, że takie udostępnianie jest działaniem w ramach własnego prawa przeważającego zawsze nad ochroną dóbr osobistych.

Istotne w kontekście danych badawczych prawa osób trzecich to ponadto prawa z patentów na wynalazki oraz prawa do wzorów użytkowych. Jednak są one zwykle powiązane z danymi pośrednio, 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, opierając się na zebranych danych, prowadząc analizy i eksperymenty.

Bardziej bezpośrednia relacja może zachodzić pomiędzy danymi badawczymi a innymi rodzajami własności przemysłowej, takimi jak znaki towarowe czy wzory przemysłowe. Ich przedstawienia mogą znajdować się w bazach tworzonych przez badaczy, np. w ramach badań marketingowych dotyczących rozpoznawalności marek. Jednak 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 przypadki wykorzystania cudzej tajemnicy 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.

Ograniczenia wynikające z umów podlegają 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 zawiera on 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 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 (RODO wymaga ponadto zawarcia umowy regulującej wskazane w przepisach kwestie w razie powierzenia osobom trzecim przetwarzania danych w imieniu administratora).

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:

1. osoba, której dane dotyczą, wyraziła zgodę,
2. przetwarzanie jest niezbędne do wykonania (lub zawarcia) umowy z osobą, której dane dotyczą,
3. przetwarzanie jest niezbędne do wypełnienia obowiązku prawnego administratora,
4. przetwarzanie jest niezbędne do ochrony żywotnych interesów osoby fizycznej,
5. przetwarzanie jest niezbędne do wykonania zadania realizowanego w interesie publicznym lub w ramach władzy publicznej administratora,
6. przetwarzanie jest niezbędne do celów wynikających z prawnie uzasadnionych interesów realizowanych przez administratora lub stronę trzecią.

W kontekście otwartego udostępniania danych badawczych zawierających dane osobowe na pewno może wystąpić pierwsza sytuacja, polegająca na zgodzie na takie udostępnianie. RODO wymaga, aby było to *dobrowolne, konkretne, świadome i jednoznaczne okazanie woli*. Oznacza to, że proces pozyskiwania zgód należy zorganizować tak, aby następnie można było bez przeszkód udowodnić, że zgody mają wszystkie te cechy. Dodatkowe wymagania zawiera art. 7 RODO, w którym np. w wypadku oświadczeń pisemnych wskazano obowiązek przedstawienia zapytania o zgodę *w sposób pozwalający wyraźnie odróżnić je od pozostałych kwestii, w zrozumiałej i łatwo dostępnej formie, jasnym i prostym językiem*. Osoba, której dane dotyczą, musi zostać powiadomiona o możliwości wycofania zgody, zanim ją wyrazi, a wycofanie zgody 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.

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. Nie istnieją natomiast obecnie 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 pozyskiwania zgody ani konieczności udowadniania, że jest to niezbędne w konkretnych przypadkach.

Pozostałe wymienione powyżej sytuacje raczej nie będą miały miejsca w kontekście prowadzenia badań.

Istotną 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 wypadku przetwarzania na podstawie zgody 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 wypadku 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 wypadku konieczne, gdy okazuje się niemożliwe lub wymagałoby niewspółmiernie dużego wysiłku, w szczególności w sytuacji 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 prawdopodobieństwo, 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, a także 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.

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. Obecnie nie istnieją takie przepisy, a musiałyby 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. Należy 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 znaczne podwyższenie wymagań związanych z zarządzaniem danymi w odniesieniu do eksperymentów medycznych.

Z drugiej 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, 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 przemysłowe. 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 są wyznaczane przez przepisy Prawa o szkolnictwie wyższym oraz 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. Zobowiązują one badaczy zatrudnionych na uczelni do zawiadamiania o wynikach badań przy jednoczesnym zachowaniu ich w poufności. Uczelnia może zdecydować się na komercjalizację wyników badań. Jeśli tego nie zrobi, 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 produkty lub metody przygotowane w wyniku ich analizy i przetworzenia) oraz narzędzi prawnych (prawa wyłączne, umowy) wpływa na to, czy możliwe będzie równoległe udostępnianie danych badawczych w sposób otwarty i w jakim zakresie. 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 wypadkach możliwe jest natomiast takie skoordynowanie działań, aby otwartość danych mogła być realizowana równoległe do komercjalizacji. Coraz częściej jest to nie tylko możliwość, lecz także obowiązek – 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 wypadku otwierania danych badawczych należy więc odpowiedzieć na pytanie, czy udostępnienie ujawni istotę wynalazku opracowanego na podstawie tych danych. Jeżeli tak, to z udostępnieniem danych należy wstrzymać się do czasu złożenia wniosku patentowego.

Podobnie w wypadku, 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 wówczas realizowana w zakresie, który pozwoli na nieujawnienie istoty komercjalizowanego *know-how*. Nie zawsze musi to oznaczać rezygnację z otwartości w ogóle. Surowe dane mogą nie pozwalać na wywnioskowanie w prosty sposób zawartego w nich *know-how*. Mogą też dać się względnie łatwo przed tym zabezpieczyć dzięki odpowiedniemu przygotowaniu do udostępnienia, zachowując przy tym przydatność naukową.

Zarządzanie danymi a kwestie etyczne w projekcie badawczym

W poszczególnych obszarach badań naukowych funkcjonują różnego rodzaju sformalizowane kodeksy lub inne dokumenty, które wprowadzają 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 ESOMAR¹³). Obowiązuje on członków ESOMAR oraz wszystkich tych, 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 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 przypomina powszechnie obowiązujące przepisy, takie jak RODO. Z drugiej strony postanowienia kodeksu mogą być postrzegane jako podniesienie 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, których RODO albo nie reguluje wprost, albo sposób tej regulacji mógłby rodzić różnice interpretacyjne – jak np. w odniesieniu do *pasywnego zbierania danych*.

Podobny charakter mają *WAPOR code of professional ethics and practices* czy opracowany przez Polskie Towarzystwo Socjologiczne *Kodeks etyki socjologa*¹⁴.

¹³ The ICC/ESOMAR International Code, <https://esomar.org/code-and-guidelines/icc-esomar-code> [data dostępu: 17.07.2023].

¹⁴ WAPOR Code of Professional Ethics and Practices, <https://wapor.org/aboutwapor/code-of-ethics/>, Kodeks Etyki Socjologa, <https://pts.org.pl/wp-content/uploads/2016/04/kodeks.pdf> [data dostępu: 17.07.2023].

Takie dokumenty mogą nie tylko podwyższać standard ochrony, lecz także 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 konkretnym zespole badaczy lub projekcie badawczym. Prawo niejednokrotnie odwołuje się do klauzul generalnych określanych jako *zasady współżycia społecznego, należąta staranność, rzetelność naukowa, dobre praktyki* itp. W sytuacji zgłaszania roszczeń prawnych pod adresem badaczy bardzo często ustalenie zakresu ewentualnej odpowiedzialności tychże badaczy jest uzależnione od tego, jakie dokładnie jest rozumienie wspomnianych pojęć w konkretnym wypadku. 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.

Oczywiście, im mniej formalny jest charakter takich dokumentów, tym trudniej wykazać, że mają one faktycznie wpływ 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 byłaby dopuszczalna.

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 zależą 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.

Odpowiedzialność cywilna

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). 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 całkowicie wykluczone przypisanie takiej odpowiedzialności indywidualnemu badaczowi, zwłaszcza gdy jego działanie nie odbywało 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 wypadkach, np. w prawie autorskim, istnieją rozwiązania ułatwiające dochodzenie odszkodowania, zwalniające z konieczności udowadniania wysokości szkody. Uprawniony musi jednak wykazać, jaką kwotę uzyskałby, gdyby zawarto z nim umowę na korzystanie z utworu (następnie tę kwotę mnoży się przez dwa).

Odpowiedzialność karna

Odpowiedzialność karna związana jest z popełnieniem przestępstw lub wykroczeń, czyli czynów zabronionych pod groźbą kary, określonych w ustawie. W odróżnieniu od odpowiedzialności cywilnej, odpowiedzialności karnej nie ponosi się 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, niebędąca jednak odszkodowaniem za wyrządzoną szkodę). Odpowiedzialność karną może ponieść jedynie konkretna osoba fizyczna, której można udowodnić winę za popełniony czyn. Będzie ona mogła być przypisana indywidualnemu członkowi 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 naruszenia. Udostępnianie danych badawczych jako takie nie stanowi przestępstwa. Mogłoby nim być jedynie w dodatkowych okolicznościach, wypełniających znamiona zniesławienia, przestępstw przeciwko bezpieczeństwu informacji itd.

Odpowiedzialność administracyjna

Odpowiedzialność administracyjna jest podobna do odpowiedzialności karnej o tyle, że zachodzi ona również na linii jednostka (tu: badacz bądź instytucja) – państwo (organ władzy państwowej). Odmienne niż w wypadku odpowiedzialności karnej, nie jest tu potrzebne wykazanie winy sprawcy. Wystarczy 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

Odpowiedzialność dyscyplinarna z kolei to szczególny reżim odpowiedzialności, uregulowanej osobno w odniesieniu do przedstawicieli niektórych zawodów (np. odpowiedzialność zawodowa lekarzy). Odpowiedzialności dyscyplinarnej podlegają też wszyscy pracownicy w ramach zasad określonych w prawie pracy, a pracownicy naukowcy 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, jak np. Narodowe Centrum Nauki).

Polityka instytucji macierzystej może przewidywać określone obowiązki związane z zarządzaniem danymi badawczymi również w wypadku 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, przyjęły 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. Wskazując 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) mają własne polityki dotyczące zarządzania danymi. Można następnie przeanalizować możliwość ich uzgodnienia z politykami instytucji finansującej i naukowej, co pozwoli 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 (takie jak 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, zasięgnąć w tym celu opinii innych działów instytucji lub skorzystać z dostępnej mu sieci ekspertów (wewnętrznych i 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 prowadzona przez tego rodzaju jednostkę umiejscowioną 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 obecnych 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 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 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 wewnętrzny dział prawny 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 określone 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 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 sprofilowanych dziedzinowo, znających kodeksy etyczne danej dziedziny oraz ich konsekwencje dla odpowiedniego zarządzania danymi. W odniesieniu do badań, w których uczestniczyć będą ludzie lub zwierzęta, konieczne może być również uzyskanie opinii komisji etycznej weryfikującej projekty badawcze w danej instytucji. Jej zalecenia mogą wprost dotyczyć odpowiedniego sposobu postępowania z danymi.

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

Na wyższym poziomie ogólności 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 pełniące rolę swego rodzaju 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, że 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 [ang. *long term preservation*], udostępniania danych i ich ponownego wykorzystania [oraz reprodukowalności i/lub odtwarzalności]¹⁵.

Inna definicja głosi, że *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 odkrycia¹⁶.

W tym kontekście *data steward* jest osobą odpowiedzialną za kreowanie odpowiednich procesów i postaw. Co ważne, rola ta 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 może wymagać podjęcia (a niekiedy ciągłego podejmowania) odpowiednich działań (np. konwersji danych do nowych formatów, które przez lata zdążyły wypierać 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, zarządzanie danymi [tj. *data steward*] 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 przechowywania¹⁷.

Zadanie *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 i reguł, zadania *data stewarda* mają w sobie istotny element twórczy.

¹⁵ 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: 1.03.2025].

¹⁶ Mons B., *Data Stewardship for Open Science: Implementing FAIR Principles*, Chapman and Hall/CRC, 2018, <https://doi.org/10.1201/9781315380711> [data dostępu: 1.03.2025].

¹⁷ Pawłowska M., Wachowicz M., *Wprowadzenie do zarządzania danymi naukowymi*, Difin, 2020.

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 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. Jednocześnie potrafi się w nim poruszać na tyle biegle, że jest w stanie doradzić pracownikowi konkretny sposób lub standard postępowania bądź przynajmniej wskazać kierunek poszukiwań.

Osoba pełniąca tę funkcję sama powinna móc liczyć na współpracę ze strony innych osób i jednostek funkcjonujących w obrębie instytucji. Trzeba tu uwzględnić kilka elementów.

- 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 wewnątrz instytucji, może obejmować zarówno techniczną obsługę usług IT, jak i bieżące wsparcie 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 wypadku tych projektów badawczych, w których istnieje rzeczywisty potencjał komercjalizacji.
- Współpraca z jednostkami wspomagającymi 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 dotyczące otwartej nauki i otwartych publikacji, stąd mogą też być dobrym miejscem rozwijania kompetencji związanych z otwartymi danymi badawczymi. W bardziej rozproszonych modelach *data stewardship*, w wypadku 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 ma 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 terminem¹⁸, a także o wielości terminów mających oddawać treści zbliżone. 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 (ang. *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 wypadku również można pokusić się o dalsze zniuansowanie funkcji i wyróżnić procesowych, systemowych czy biznesowych data stewardów¹⁹.

¹⁸ 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: 1.03.2025].

¹⁹ 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(+)*²⁰. 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 czasopiśmie, 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 w ogólności, w szczególności zaś kariery data stewardów²¹.

Owo narażenie na szwank bierze się z jednej strony z tego, że 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 wypadku tych naukowców, którzy zdecydowali się na pełnienie funkcji 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.

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 wchodzi wiele i bardzo zróżnicowane wydziały.

²⁰ Por. Mons B., dz. cyt.

²¹ Tamże.

Nawet instytucje zbliżone wielkością mogą być inaczej zorganizowane i mieć odmienną wewnętrzną strukturę. Sytuacja instytucji, która już posiada skuteczny dział promocji oraz prężnie działający dział IT, dysponujący dobrze zorganizowanym systemem wsparcia użytkowników, 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 wypadku 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 szczególnymi 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 – coraz częściej dane traktowane są 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 uwzględną 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 danych²².

W rezultacie coraz więcej aspektów życia społecznego może opierać się 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 wartość. W konsekwencji niewłaściwe zarządzanie danymi lub jego brak może oznaczać zmarnotrawienie jakiegoś fragmentu tego dobra. Dane, którymi nikt nie zarządza i których nikt nie udostępnia, nie przyniosą nikomu pożytku. Z kolei niewłaściwe zarządzanie danymi może nawet takie marnotrawstwo pogłębić, oznacza bowiem, że 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 natychmiast, bez żadnych dodatkowych czynności przygotowawczych, które po angielsku określa się terminem *data munging*.

²² 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: 1.03.2025].



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 danych²³.

Rola i zakres obowiązków data stewarda

Data steward może realizować swoje zadania na różnych poziomach instytucji. Na podstawie analizy dostępnej literatury²⁴ wyróżnić można w szczególności data stewardów, w wypadku których centrum działań jest biblioteka, oraz zakorzenionych (ang. *embedded*) data stewardów, operujących na poziomie wydziału czy instytutu, a 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, lecz także metody oraz dane, z którymi pracują.

Drugie podejście zidentyfikowane w literaturze przedmiotu dotyczy nowych wyzwań stojących przed bibliotekarzami zajmującymi się danymi (ang. *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 (ang. *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 wiedzę 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.

²³ 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.

²⁴ Por. Tamże.

Na przeciwnym biegunie autorzy lokują z kolei ogólnego (ang. *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.

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

Przykład – Uniwersytet Gandawski, Belgia

W Bibliotece Uniwersytetu w Gandawie działa [zespół ds. otwartej nauki](#), w skład którego wchodzi koordynatorka, pięcioro data stewardów i dwoje kuratorów danych. Członkowie zespołu udzielają naukowcom ogólnego i dziedzinowego wsparcia w zakresie zarządzania danymi badawczymi.

Do głównych zadań członków zespołu należą:

- promowanie dobrych praktyk w zakresie zarządzania danymi badawczymi,
- wspieranie zmian zachodzących w kulturze naukowej wśród badaczy i szerszej społeczności uniwersytetu, w tym zmian dotyczących danych otwartych, FAIR i zdalnych do ponownego wykorzystania,
- pomoc w podnoszeniu jakości danych i oprogramowania oraz ich zgodności z zasadami FAIR.

Kuratorzy danych (ang. *data curators*) oceniają poprawność danych i metadanych wszystkich zbiorów danych przeznaczonych do archiwizacji, publikacji i katalogowania. Pracują z (meta)danymi, dążąc do poprawienia ich jakości.

Data stewardzi stanowią uniwersytecki punkt kontaktowy dla osób poszukujących informacji na temat zarządzania danymi badawczymi. Są doradcami, którzy zapewniają wsparcie doktorantom i naukowcom prowadzącym badania. Każdy z pięciorga data stewardów odpowiada za jeden klaster wydziałowy: klaster prawa, sztuki i nauk humanistycznych; klaster nauk społecznych i behawioralnych; klaster nauk przyrodniczych; klaster (bio)inżynierii; klaster nauk o życiu i medycyny.

Przykłady działań data stewardów:

- udzielanie wsparcia w procesie przygotowania planów zarządzania danymi,
- prowadzenie centrum informacyjnego oferującego konsultacje indywidualne i grupowe w zakresie zarządzania danymi badawczymi oraz korzystania z narzędzi takich jak DMPonline czy OSF,
- organizowanie szkoleń, spotkań informacyjnych i innych wydarzeń związanych z tematyką danych badawczych.

Przykład – Uniwersytet Wiedeński, Austria

[Biblioteka Uniwersytetu Wiedeńskiego](#) zatrudnia czworo data stewardów oraz ich koordynatorkę. Data stewardzi są odpowiedzialni za konkretne wydziały i centra badawcze, w obrębie których współpracują z naukowcami i studentami. Ich rolą jest usprawnianie procesów zarządzania danymi w projektach badawczych i bezpośrednia praca z danymi (kuratorowanie), prowadzenie szkoleń z zarządzania danymi badawczymi, udzielanie konsultacji dotyczących planów zarządzania danymi, usprawnianie komunikacji i przepływu informacji między jednostkami uniwersyteckimi uczestniczącymi w procesie zarządzania danymi. Osoby zatrudniane na stanowisko data stewarda biorą udział w podyplomowym kursie specjalizacyjnym organizowanym przez uczelnię.

Przykład – Polska

Polskie instytucje planują i realizują wewnętrzne działania związane z zarządzaniem danymi badawczymi na rozmaite sposoby, a role i zadania z tym związane różnią się w zależności od rodzaju jednostki, jej struktury, przyjętych polityk i dostępnych zasobów.

W większości instytucji tematyką otwartych danych zajmują się specjaliści pracujący w bibliotekach i działach informacji naukowej, którzy rozszerzają w ten sposób zakres swoich obowiązków związanych z otwartą nauką. Dominuje model centralnie działającego opiekuna danych, którego funkcję pełni jeden z członków istniejącego zespołu.

Ważnymi elementami wciąż kształtującego się krajobrazu *data stewardship* w Polsce są powstające centra kompetencji oraz grupy robocze, np. Sieć Kompetencji Otwartej Nauki – Data Steward PL.

Słownik terminów z zakresu zarządzania danymi badawczymi



A

anonimizacja | *anonymisation*

Proces usuwania z danych informacji pozwalających na ustalenie tożsamości poszczególnych osób, a także odniesień do innych danych, które umożliwiają ustalenie ich tożsamości (por. Cornell University, [Glossary of data management terms](#)).

autorskie prawa osobiste | *author's moral rights**

Prawa chroniące więzi osobiste twórcy z utworem. Obejmują m.in. prawo do oznaczenia utworu swoim nazwiskiem, prawo do nienaruszalności treści i formy utworu oraz jego rzetelnego wykorzystania, a także prawo do nadzoru nad sposobem korzystania z utworu. Autorskie prawa osobiste są nieograniczone w czasie, nie można się ich zrzec ani przenieść ich na inne osoby.

autorskie prawa majątkowe | *copyright (economic rights)**

Wyłączne prawo do korzystania z utworu i rozpowszechniania go oraz do wynagrodzenia za korzystanie z niego. Uprawniony może upoważnić inne osoby do korzystania z utworu w określonym zakresie przez udzielenie im licencji. Może także przenieść autorskie prawa majątkowe na inne osoby, odpłatnie albo za darmo. Autorskie prawa majątkowe co do zasady wygasają po siedemdziesięciu latach od końca roku kalendarzowego, w którym zmarł twórca.

B

baza danych | *database*

Dane zorganizowane zgodnie z określonym modelem charakteryzującym te dane oraz zachodzące pomiędzy nimi relacje. Zorganizowanie danych w bazę znacznie ułatwia dostęp do nich, ich aktualizację oraz zarządzanie nimi.

***born digital* | (materiały) natywnie cyfrowe**

Termin odnoszący się do różnego typu materiałów, w tym danych badawczych, które zostały wytworzone wyłącznie w postaci elektronicznej; powstał w celu odróżnienia ich od materiałów, które w pierwotnej formie nie miały charakteru cyfrowego i zostały zdigitalizowane (stanowią elektroniczne odwzorowanie dokumentów w postaci pierwotnej).

C

cyfryzacja | *digitization****

Proces tworzenia plików cyfrowych poprzez skanowanie lub inny sposób konwersji materiałów analogowych.

D

dane badawcze | *research data**

Różnorodne materiały gromadzone w ramach prowadzenia badań naukowych (m.in. dane liczbowe, dokumenty tekstowe, notatki, kwestionariusze, wyniki ankiet, nagrania audio i wideo, fotografie, zawartość baz danych, oprogramowanie, wyniki symulacji komputerowych, protokoły metodologiczne, obserwacje laboratoryjne), definiowane między innymi na poniższe sposoby: 1) zarejestrowane materiały o charakterze faktograficznym, powszechnie uznawane przez społeczność naukową za niezbędne do oceny wyników badań naukowych; 2) dane zebrane, zaobserwowane lub wytworzone jako materiał do analizy w celu uzyskania oryginalnych wyników naukowych; 3) wszystko, co zostało wyprodukowane lub wytworzone w ramach prowadzonych badań.

dane dynamiczne | *dynamic data*

W Ustawie o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego zdefiniowane jako informacje sektora publicznego utrwalone w postaci elektronicznej podlegające częstym aktualizacjom lub aktualizacjom w czasie rzeczywistym, w tym ze względu na ich zmienność lub szybką dezaktualizację, w szczególności dane wygenerowane przez czujniki (zob. Dz. U. 2021 poz. 1641, [Ustawa z dnia 11 sierpnia 2021 r. o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego](#)).

dane osobowe | *personal data**

Informacje dotyczące żyjącego człowieka, które pozwalają go zidentyfikować (np. imię i nazwisko, PESEL, cechy wyglądu, stan zdrowia). Wykorzystywanie danych osobowych (np. na potrzeby badań naukowych) jest regulowane przez prawo Unii Europejskiej (rozporządzenie o ochronie danych osobowych, tzw. RODO) i dozwolone wyłącznie na przewidzianych tam zasadach. W szczególności zbieranie i analizowanie danych osobowych na potrzeby badań naukowych oraz ich udostępnianie może wymagać zgody osób badanych oraz udzielenia im informacji o określonej formie i treści.

dane wrażliwe | *sensitive data*

Szczególne kategoria danych osobowych, podlegających zgodnie z RODO dodatkowym wymogom związanym z ich ochroną. Są to dane ujawniające pochodzenie rasowe lub etniczne, poglądy polityczne, przekonania religijne lub światopoglądowe, przynależność do związków zawodowych, dane genetyczne, dane biometryczne, dane dotyczące zdrowia, seksualności lub orientacji seksualnej.

***data availability statement* | oświadczenie o dostępności danych**

Opis sposobu udostępnienia danych badawczych powiązanych z opublikowanym artykułem; stanowi element publikacji, który czasami jest wymagany przez czasopisma naukowe; może wskazywać konkretne repozytoria lub serwisy, w których przechowywane są dane, precyzować warunki dostępu do nich lub wskazywać ewentualne ograniczenia w dostępie.

data journal

Czasopismo naukowe publikujące artykuły opisujące zbiory danych badawczych. Celem *data journali* jest zwiększenie widoczności oraz ponownego wykorzystania danych opisywanych w artykułach.

data stewardship

Odpowiedzialne planowanie i wykonywanie wszystkich działań dotyczących danych cyfrowych przed, w trakcie oraz po zakończeniu projektu badawczego, mające na celu optymalizację użyteczności, ponownego wykorzystania i reprodukowalności powstałych danych (zob. Dutch Techcentre for Life Sciences, [Research Data Management](#)).

dobra osobiste | personal rights*

Niektóre cechy człowieka oraz przejawy jego aktywności (np. nazwisko, dobre imię, tajemnica korespondencji, godność osobista, twórczość naukowa i artystyczna), podlegające ochronie na podstawie prawa cywilnego. Ochrona ta wymaga aktywności samego zainteresowanego, który może wnieść pozew przeciwko osobie naruszającej jego dobra osobiste i zażądać np. przeprosin albo zadośćuczynienia.

DOI | Digital Object Identifier**

Jeden z trwałych identyfikatorów obiektów cyfrowych, pozwalający na ich odnalezienie w internecie niezależnie od wiodącego do nich adresu URL. Posiadający DOI zbiór danych można za jego pomocą zidentyfikować nawet wtedy, gdy zostanie on przeniesiony na inny serwer czy do innego repozytorium.

dokumentacja danych | data documentation

Ogół informacji niezbędnych do zrozumienia i poprawnej interpretacji danych. Informacje te mogą być zawarte w metadanych zbioru danych, słownikach, opisach metod oraz instrumentów badawczych wykorzystanych do zebrania i przetworzenia danych, a niekiedy również w samych plikach z danymi.

dozwolony użytek | copyright exceptions and limita- tions / fair use / fair dealing*

Korzystanie i rozpowszechnianie utworu chronionego prawem autorskim, które nie wymagają zgody tego, komu przysługują prawa do tego utworu. Jest ograniczone do wskazanych w ustawie o prawie autorskim celów i sposobów wykorzystywania informacji (np. na użytek osobisty albo naukowy).



**European Open Science Cloud,
EOSC (Europejska Chmura
Otwartej Nauki)**

Europejska inicjatywa, której celem jest stworzenie wirtualnego środowiska służącego m.in. do przechowywania, udostępniania i przetwarzania danych badawczych. W ramach EOSC tworzone są połączenia między istniejącymi i rozwijanymi infrastrukturami, usługami, narzędziami i zasobami naukowymi z różnych dziedzin, które wspólnie tworzyć będą sieć danych i usług opartych na zasadach FAIR (ang. *web of FAIR data and services*). Więcej informacji na temat EOSC znaleźć można [na stronie internetowej](#).

embargo**

Okres, przez który dane badawcze nie są udostępnione publicznie. Jest on zwykle wykorzystywany po to, aby uzyskać związane z danymi patenty i/lub inne prawa własności intelektualnej oraz przygotować oparte na nich publikacje naukowe. Po jego upływie opublikowanie danych badawczych staje się możliwe.



FAIR**

Akronim słów *findable* (możliwy do znalezienia), *accessible* (dostępny), *interoperable* (interoperacyjny) i *reusable* (możliwy do ponownego wykorzystania), określający wymogi, jakie powinny spełniać udostępnione dane badawcze.

**format dostępny maszynowo |
machine-readable format**

Format danych lub metadanych umożliwiający ich maszynową dostępność.



integralność danych | ***data integrity***

Właściwość danych, która gwarantuje ich kompletność, spójność, dokładność i wiarygodność. W praktyce zapewnienie integralności danych wymaga wdrożenia odpowiednich procedur na etapie zbierania, przetwarzania i przechowywania danych, które zapobiegają błędnym i nieuprawnionym zmianom, uszkodzeniu lub usunięciu danych. Wdrożenie odpowiednich procedur w tym zakresie jest szczególnie istotne w kontekście zarządzania danymi w formie cyfrowej ze względu na łatwość dokonywania modyfikacji plików.

interoperacyjność | ***interoperability***

Zdolność urzędów, systemów, aplikacji oraz treści do współdziałania z innymi urządzeniami, systemami, aplikacjami oraz treściami.



jakość danych | *data quality****

Jakość danych badawczych można rozpatrywać m.in. w aspekcie ich trafności, kompletności, aktualności, znaczenia, spójności źródeł, wiarygodności, odpowiedniości sposobu prezentacji oraz dostępności.



komercjalizacja wyników badań | *commercialization of research results**

Czerpanie korzyści finansowych z wyników badań naukowych. Najczęściej kojarzy się z opatentowaniem wynalazku stworzonego na podstawie uzyskanych wcześniej wyników badań, a następnie odpłatnym udzielaniem licencji zezwalających na korzystanie z niego. Komercjalizacja jest przedmiotem regulacji prawnych, które mogą mieć znaczenie dla otwartego udostępniania danych badawczych.

komisja ds. dostępu do danych | *data access committee*

Grupa osób odpowiedzialnych za weryfikację i ocenę próśb o dostęp do danych.

kontrola wersji | *version control****

Kontrolowanie w czasie zmian w danych, kodzie komputerowym, oprogramowaniu i dokumentach w sposób umożliwiający powrót do poprzedniej wersji, co ma kluczowe znaczenie dla możliwości śledzenia historii danych, czynionych w nich zmian czy poprawiania błędów. Kontrola wersji generuje (zmienioną) kopię danych, która zostaje oznaczona unikatowym numerem wersji.

kuratorowanie danych | *data curation****

Ogół procesów w toku cyklu życia danych, za sprawą których dane są czyszczone, dokumentowane, standaryzowane i wzajemnie ze sobą wiązane. Może obejmować między innymi wersjonowanie danych, tworzenie nowych zestawów na podstawie kilku źródeł danych, opatrywanie danych metadanymi czy kodowanie danych surowych.



licencje Creative Commons | *Creative Commons licenses**

Wzory licencji opracowane przez założoną w USA organizację pozarządową [Creative Commons](https://creativecommons.org/), współpracującą obecnie z partnerami w wielu państwach, również w Polsce. Osoba, której przysługują prawa własności intelektualnej do określonych treści (np. zbioru danych), może oznaczyć je symbolem jednej z licencji. W ten sposób określa zakres przysługujących użytkownikom uprawnień. Wolnymi licencjami są te licencje Creative Commons, które uprawniają zarówno do sporządzania opracowań chronionej twórczości, jak i do korzystania z niej w celach komercyjnych. Warunki te spełniają licencje CC BY (wymagająca od użytkownika podania informacji o autorstwie oraz warunkach licencji) oraz CC BY-SA (wymagająca ponadto stosowania takiej samej licencji w wypadku udostępniania opracowania danej twórczości), a także oświadczenie CC0 (będące na gruncie prawa polskiego bardzo szeroką licencją, nienakładającą na licencjobiorcę praktycznie żadnych zobowiązań).



metadane | *metadata***

Ustrukturyzowane informacje opisujące zasoby informacji, takie jak zbiory danych badawczych. Metadane zawierają informacje o formie i treści zasobów, dzięki czemu pozwalają na ich wyszukiwanie i identyfikację oraz zarządzanie nimi.

możliwość odczytu maszynowego | *machine readability*

Dostępność treści (w szczególności danych oraz metadanych) w postaci umożliwiającej ich wykorzystanie i interpretację przez maszyny (komputery).



obiekt cyfrowy | *digital object*

Sekwencja bitów tworząca pod względem operacyjnym całość analogiczną do obiektu fizycznego. Ta sama sekwencja bitów może mieć różne interpretacje, co znaczy, że różne obiekty cyfrowe mogą być tworzone za pomocą tych samych sekwencji bitów. W związku z tym założona interpretacja powinna być znana i przekazywana innym użytkownikom w formie metadanych o kodowaniu znaków i/lub formacie pliku.

ontologia | *ontology*

Sformalizowany model rzeczywistości, składający się z określonego zbioru pojęć (skategoryzowanych i zhierarchizowanych nazw) oraz relacji między nimi. Całość tworzy schemat pojęciowy opisujący daną dziedzinę wiedzy oraz podstawowe zasady wnioskowania w jej ramach na temat cech i własności przedmiotów nią objętych.

oprogramowanie o otwartym kodzie źródłowym | *open source software*

Oprogramowanie, którego kod źródłowy jest dostępny pod względem technicznym, i objęte licencją umożliwiającą jego szerokie ponowne wykorzystanie. Zgodnie z formalną definicją (zamieszczoną na stronie [Open Source Initiative](#)), licencja taka musi zezwalać każdemu na redystrybucję oprogramowania bez ograniczeń w ramach większej całości, zarówno w postaci kodu źródłowego, jak i wynikowego (z odpowiednimi gwarancjami dostępu do kodu źródłowego), w postaci oryginalnej i zmodyfikowanej, w dowolnym celu. Większość oprogramowania

open source jest jednocześnie wolnym oprogramowaniem, którego definicja (zawarta na stronie [GNU](#)) wymaga zapewnienia użytkownikom wszystkich czterech wolności: uruchamiania programu w dowolnym celu, modyfikacji, rozpowszechniania kopii, publicznego udostępniania ulepszeń (które także wymagają dostępu do kodu źródłowego).

otwarte dane badawcze | open research data

Zgodnie z rekomendacjami UNESCO 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 (ang. *actionable*) formacie, zgodnie z zasadami dobrego zarządzania i opieki nad danymi, w szczególności zasadami FAIR (ang. *findable, accessible, interoperable, reusable*) oraz podlegają regularnym działaniom w zakresie ich kuratorowania i utrzymywania (por. UNESCO, [Recommendation on open science](#)).

otwarta nauka | open science

Zbiór zasad i praktyk w zakresie sposobu prowadzenia badań i udostępniania ich rezultatów, których celem jest zapewnienie jak najszerzego dostępu do wyników badań oraz możliwości ich ponownego wykorzystania. Otwarta nauka obejmuje m.in. otwarty dostęp do publikacji, otwarte dane badawcze, otwarte oprogramowanie, otwarte zasoby edukacyjne, a także inicjatywy z zakresu nauki obywatelskiej. W rekomendacjach UNESCO przyjętych w 2021 r. otwarta nauka zdefiniowana została jako 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 oraz otwarcie procesów tworzenia i oceny wiedzy naukowej czy komunikację z podmiotami społecznymi spoza tradycyjnie rozumianej społeczności naukowej (por. UNESCO, [Recommendation on open science](#)).

otwarty dostęp | *open access**

Udostępnienie określonych informacji (np. danych badawczych) w Internecie, w sposób umożliwiający swobodny dostęp do nich. Otwarty dostęp zakłada brak konieczności wnoszenia jakichkolwiek opłat przez odbiorców oraz brak ograniczeń technicznych (takich jak np. obowiązek rejestracji w systemie czy konieczność instalacji specjalistycznego oprogramowania). Nie jest to równoznaczne z brakiem prawnej ochrony udostępnionych informacji. Korzystanie z nich oraz dalsze udostępnianie może być ograniczone, przede wszystkim przepisami prawa własności intelektualnej. Jednak w odniesieniu do danych badawczych otwarty dostęp oznacza, że zostały one udostępnione na wolnej licencji lub w domenie publicznej. W odniesieniu do publikacji naukowych różni się natomiast otwarty dostęp *gratis* i *libre* (*gratis* oznacza dostępność nieodpłatną i bez ograniczeń technicznych, a dopiero *libre* – dodatkowe udzielenie wolnej licencji do publikacji).

otwarty format pliku | *open file format*

Sposób przechowywania danych na nośniku cyfrowym określony w opublikowanej i ogólnodostępnej specyfikacji, którego wdrożenie jest możliwe bez ograniczeń technicznych i prawnych. Otwarte formaty poddane procedurze standaryzacji (normalizacji) w ramach organizacji dbającej o pluralizm, inkluzywność oraz transparentność tej procedury, a także o minimalizację ograniczeń technicznych i prawnych we wdrażaniu specyfikacji, są nazywane otwartymi standardami. Nie ma jednak jednego, powszechnie przyjętego rozumienia otwartości w kontekście formatów i standardów. Samo pojęcie standardu bywa też często używane nieściśle na określenie popularnych formatów, niekoniecznie poddanych procedurze formalnej standaryzacji.



patent | *patent**

Prawo własności intelektualnej chroniące wynalazek – rozwiązanie pewnego problemu technicznego, jeżeli jest nowe, posiada poziom wynalazczy i nadaje się do przemysłowego stosowania. Uprawniony ma wyłączne prawo do korzystania z wynalazku w celach zarobkowych lub zawodowych. Patent jest przyznawany przez Urząd Patentowy.

**plan zarządzania danymi
badawczymi | data
management plan (DMP)****

Dokument opisujący to, co będzie działo się z danymi w trakcie projektu badawczego i po jego zakończeniu. Ma on charakter żywego dokumentu, który może i powinien zmieniać się wraz ze zmianami pojawiającymi się w innych obszarach projektu badawczego.

**ponowne wykorzystanie |
reuse****

Ogólny termin odnoszący się do technicznych, prawnych i metodologicznych uwarunkowań użycia danych przez dowolne osoby i/lub instytucje, w szczególności te, które nie były zaangażowane w ich wytworzenie.

**prawo autorskie |
copyright law***

Prawo wyłączne obejmujące utwory – przejawy twórczej działalności człowieka. Ochrona prawa autorskiego powstaje bez udziału organów państwa, w wyniku stworzenia utworu ustalonego w jakiegokolwiek – zewnętrznej wobec twórcy – postaci. Co do zasady uprawnionym jest sam twórca, chociaż przepisy przewidują odstępstwa od tej zasady, dotyczące np. utworów stworzonych przez pracowników w ramach ich obowiązków, a także programów komputerowych.

**prawo *sui generis* do bazy
danych | *sui generis*
database right***

Prawo chroniące uporządkowany zbiór danych jako całość. Nie obejmuje poszczególnych elementów tego zbioru, które mogą jednak podlegać odrębnej ochronie (jak np. zbiór nagrań wideo, chronionych prawem autorskim). Prawo do bazy danych przysługuje tylko wtedy, gdy sporządzenie, weryfikacja lub prezentacja zawartości zbioru wymagały istotnego nakładu inwestycyjnego. Uprawnionym jest ten, kto poniósł ryzyko wspomnianego nakładu (przepisy taką osobę nazywają producentem bazy). Może on zbyć prawo do bazy danych lub udzielić licencji, upoważniając tym samym do korzystania z niej w określonym zakresie.

**prawo „własności
intelektualnej” |
intellectual property***

Ogół przepisów lub wynikających z nich niekiedy praw wyłącznych regulujących korzystanie z dóbr niematerialnych. Zalicza się tu zazwyczaj prawo autorskie i prawa pokrewne, prawa *sui generis* do baz danych oraz prawa własności przemysłowej takie jak prawa z patentu, prawa do znaków towarowych czy prawa do wzorów przemysłowych. Pojęcie praw „własności intelektualnej” obejmuje często także zakazy i obowiązki wynikające z przepisów o tajemnicy przedsiębiorcy czy przepisów o zwalczaniu nieuczciwej konkurencji. Własność w ścisłym prawniczym rozumieniu odnosi się tylko do rzeczy materialnych, pojęcie „własności intelektualnej” natomiast nie ma ścisłej prawniczej definicji i jest raczej figurą retoryczną, akcentującą przyznanie wyłączności na korzystanie z dóbr niematerialnych określonym podmiotom na wzór prawa własności.

**pseudonimizacja |
pseudonymisation ******

Przetwarzanie danych osobowych w taki sposób, by bez użycia dodatkowych informacji dane nie mogły zostać powiązane z osobami, których dotyczą. Dodatkowe informacje muszą być przechowywane osobno i podlegać technicznemu oraz organizacyjnemu środkom ochrony, tak by zapewnić, że identyfikacja osób pozostanie niemożliwa. W odróżnieniu od anonimizacji pseudonimizacja jest procesem odwracalnym, co znaczy, że osoby, których dotyczą dane, będzie można zidentyfikować, o ile umożliwiony zostanie dostęp do dodatkowych informacji.

R

repozytorium | repository

System informatyczny umożliwiający przechowywanie publikacji naukowych lub danych badawczych oraz ich udostępnianie przez Internet.

**replikacja badań |
research replication**

Przeprowadzenie takich samych analiz na różnych zestawach danych (zob. poradnik [The Turing way](#)).

**reprodukcja badań |
research reproduction**

Przeprowadzenie tych samych analiz na tych samych zbiorach danych (zob. poradnik [The Turing way](#)).

S

**słownik kontrolowany |
controlled vocabulary**

Zestaw znormalizowanych terminów i wyrażen stanowiących metadane, które używane są w celu opisu, klasyfikacji i wyszukiwania treści w określonym kontekście (np. w wybranym serwisie, katalogu czy repozytorium); charakteryzuje się spójnym i jednoznacznym stosowaniem określonych terminów i wyrażen, wykluczającym ich różną pisownię.

**standard metadanych |
metadata standard**

Sformalizowany i szeroko stosowany sposób opisu danych. Standard metadanych określa strukturę opisu metadanowego oraz znaczenie stosowanych w nim pojęć. Przykładami takich standardów są DataCite (ogólny standard, którego spełnienie umożliwia zdobycie identyfikatora DOI) czy DDI (standard metadanych dla danych społecznych).

storage | sposób przechowywania danych****

Technologia informatyczna umożliwiająca zachowanie danych w formacie cyfrowym przy użyciu komponentów komputerowych i nośników zapisywalnych, w tym usług chmurowych. Przykładami takich technologii mogą być dyski przenośne lub pendrive'y, uniwersyteckie dyski sieciowe czy usługi przechowywania danych w chmurze.



trwały identyfikator | persistent identifier

Unikalna w skali globalnej sekwencja znaków umożliwiająca jednoznaczne i trwałe identyfikowanie obiektu, z którym jest powiązana; może identyfikować różnego typu materiały, w tym publikacje, dane badawcze i inne materiały (np. DOI, Digital Object Identifier; ISSN, International Standard Serial Number), instytucje (ROR, Research Organization Registry), a także osoby (ORCID, Open Researcher and Contributor ID); powszechnie stosowanym trwałym identyfikatorem danych badawczych jest DOI.

tworzenie kopii zapasowych | backup****

Proces kopiowania danych w postaci cyfrowej i przechowywania ich na innym urządzeniu, mający na celu zachowanie bezpieczeństwa danych i zapobieżenie ich utracie. Kopia zapasowa może mieć charakter pełny (przy okazji każdego tworzenia kopii zapasowej zachowywane są wszystkie pliki) lub częściowy (każdorazowo zachowywana jest jedynie część plików, np. nowe pliki).



wolne licencje | free licenses*

Licencje, na podstawie których każdy zainteresowany uzyskuje zezwolenie na korzystanie z licencjonowanego przedmiotu w dowolnym zakresie (na wszystkich polach eksploatacji, również w celu komercyjnym, w oryginale oraz w postaci opracowania lub modyfikacji), a także na jego swobodne rozpowszechnianie i modyfikowanie. Wolne licencje są nieodpłatne i nie nakładają na licencjobiorców zobowiązań idących dalej niż obowiązki przekazywania odbiorcom określonych informacji (np. o autorze, źródle, licencji – tzw. klauzule uznania autorstwa) lub obowiązki stosowania takiej samej licencji w wypadku rozpowszechniania modyfikacji (tzw. klauzule *copyleft* lub *sharealike*).



zabezpieczenie danych | *data preservation*

Ogół czynności i procesów mających zapewnić dostęp do cyfrowych danych badawczych w średnim i długim okresie. Odpowiednie zabezpieczenie danych powinno uwzględniać zarówno kwestie związane z ich przechowywaniem, jak i stale zachodzącymi zmianami o charakterze technologicznym – takimi jak pojawianie się nowych formatów danych i nowych form dostępu do nich – czy zachowaniem możliwości uruchomienia oprogramowania niezbędnego do ich analizy.

zarządzanie danymi badawczymi | *research data management*

Zbiór procedur, zasad i dobrych praktyk w zakresie gromadzenia, przechowywania, opracowywania, udostępniania i archiwizowania danych badawczych, którego celem jest zagwarantowanie bezpieczeństwa i integralności danych, a także określenie zasad ich udostępniania oraz ponownego wykorzystania. Narzędziem pozwalającym zaplanować właściwe zarządzanie danymi badawczymi jest plan zarządzania danymi.

zbiór danych | *dataset (data set)*

Zestaw plików zawierających dane badawcze, powstałych w toku badań w ramach realizacji projektu badawczego i/lub przygotowywania artykułu naukowego, wraz z towarzyszącymi im metadanymi. Metadane te opisują dane oraz wskazują, kto je wytworzył i kto może uzyskać do nich dostęp (zob. U.S. Department of Agriculture, [Data management glossary](#)).

Źródła opracowań haseł:

* K. Gaczyńska, N. Rycko, K. Siewicz, *Prawne aspekty otwierania danych badawczych – poradnik*, Warszawa 2022. Broszura dostępna na stronie Platformy Otwartej Nauki, <https://pon.edu.pl/dane-materialy> (na licencji Creative Commons – Uznanie Autorstwa 4.0).

** W. Fenrich, *Selekcja i przygotowanie danych badawczych do udostępnienia*, Warszawa 2019. Broszura dostępna na stronie Platformy Otwartej Nauki, <https://pon.edu.pl/dane-materialy> (na licencji Creative Commons – Uznanie Autorstwa 4.0).

*** Committee on Data International Science Council, *Research Data Management Terminology*, <https://codata.org/initiatives/data-science-and-stewardship/rdm-terminology-wg/rdm-terminology/> (na licencji Creative Commons – Uznanie Autorstwa 4.0).

**** OpenAIRE, *Research Data Management Glossary*, <https://www.openaire.eu/rdm-glossary> (na licencji Creative Commons – Uznanie Autorstwa 4.0).