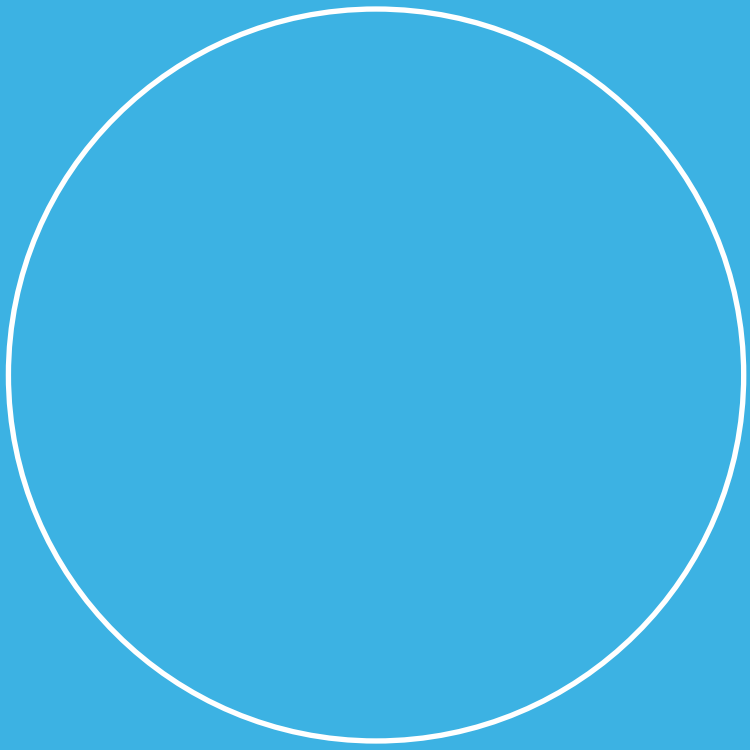
Selekcja i przygotowanie danych badawczych do udostępnienia

Wersja 1.0



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

Broszura powstała w ramach projektu Dziedzinowe Repozytoria Otwartych Danych Badawczych, finansowanego ze środków Programu Operacyjnego Polska Cyfrowa.

Strona internetowa projektu: drodb.icm.edu.pl

Uniwersytet Warszawski
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
ul. Tyniecka 15/17
02-630 Warszawa
https://icm.edu.pl/

Selekcja i przygotowanie danych badawczych do udostępnienia

Wersja 1.0

Spis treści

Słowniczek

Dane badawcze – dane zebrane lub wytworzone jako materiał do analizy w ramach badań naukowych.

Data access committee – składająca się z ekspertów grupa, do której należy decyzja o udostępnieniu zbioru danych.

Data journal – czasopismo naukowe, które publikuje artykuły opisujące zbiory danych badawczych, udostępnione w repozytoriach danych lub (rzadko) w formie suplementu do samego artykułu.

Data management plan – zob. plan zarządzania danymi.

DMP – zob. plan zarządzania danymi.

DOI – ang. „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.

Embargo – okres, przez który dane badawcze nie mogą zostać udostępnione publicznie. Jest on zwykle wykorzystywany po to, aby uzyskać związane z nimi 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.

Interoperacyjność – cecha tych danych, które można łączyć z innymi danymi, wykorzystywać w wielu różnych systemach komputerowych i analizować przy użyciu różnorodnego oprogramowania.

Licencja – upoważnienie do korzystania w określony sposób z utworu lub bazy danych. Przedmiotem licencji może być na przykład zbiór danych badawczych.

Licencje Creative Commons – popularne wzory licencji opracowane przez organizację Creative Commons.

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

Ograniczony dostęp – model dostępu, w którym dane udostępniane są jedynie określonym osobom (np. tym, które uzyskały zgodę dysponenta danych) lub kategoriom osób (np. prowadzącym badania naukowe lub zatrudnionym w konkretnej instytucji).

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

Plan zarządzania danymi (*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 – 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.

Repozytorium danych – serwis internetowy służący do deponowania (umieszczania), przechowywania i udostępniania za pośrednictwem internetu danych badawczych w formie cyfrowej.

Wstęp

Udostępnianie danych badawczych innym osobom i instytucjom staje się coraz istotniejszym elementem krajobrazu nauki w Polsce i na świecie. Mają na to wpływ naukowcy, którzy coraz częściej sami decydują się na udostępnienie swoich danych; redakcje czasopism naukowych, które wymagają od autorów udostępniania danych pozwalających na weryfikację twierdzeń zawartych w publikacji; a także instytucje finansujące, które coraz częściej zobowiązanie do udostępnienia danych badawczych czynią elementem umów grantowych z beneficjentami.

Dzięki udostępnianiu danych naukowcy zwiększają szanse na cytowanie swoich publikacji i zbiorów danych, czasopisma mają lepszą renomę i wyższą jakość ukazujących się w nich artykułów, a instytucje finansujące dbają o racjonalność wydatkowania środków publicznych.

Niniejsza broszura zawiera wskazówki dotyczące selekcji i przygotowania danych badawczych do udostępnienia. Największe korzyści z jej lektury odniosą sami naukowcy zaangażowani w realizację badań, w ramach których pozyskiwane lub wytwarzane są dane. Jej struktura wprost nawiązuje do wymogów związanych z udostępnianiem danych badawczych powstałych w wyniku badań finansowanych w ramach programu Horyzont 2020, którym podlegają również polscy badacze zaangażowani w ich realizację.

Jak przygotować dane badawcze do udostępnienia?

Opracowanie i udostępnienie danych badawczych wymaga ustalenia wielu kwestii, a potem – realizacji tych ustaleń w toku całego projektu badawczego. Ze względu na różnorodność danych badawczych niemożliwe jest określenie uniwersalnego zbioru zasad dotyczących tego, co i w jaki sposób należy zrobić, aby przygotować je do udostępnienia.

W przypadku danych badawczych nie dysponujemy więc – bo też nie możemy dysponować – czymś w rodzaju jednoznacznej listy kontrolnej, której elementy moglibyśmy „odhaczyć", by mieć

pewność, że jesteśmy dobrze przygotowani do ich udostępnienia. Rzetelne przygotowanie i udostępnienie danych badawczych jest procesem twórczym, wymagającym każdorazowego przeanalizowania kolejnych kwestii.

Dokument zawierający rezultaty takich analiz to plan zarządzania danymi (ang. *data management plan*, w skrócie DMP). Sporządzenia takiego planu coraz częściej wymagają instytucje akademickie, a przede wszystkim – instytucje finansujące badania naukowe. Co za tym idzie, istnieje wiele różnych wzorów planów zarządzania danymi (ang. *data management plan template*), opracowanych przez różne instytucje – część z nich znaleźć można na stronie internetowej https://dmptool.org/.

Na tej samej witrynie dostępnych jest wiele gotowych planów zarządzania danymi, udostępnionych przez ich autorów. Obok przykładów solidnie przygotowanych dokumentów znaleźć tam można również wiele innych, sporządzonych pobieżnie, więc ich lektura powinna mieć charakter krytyczny. Może być to jednak dobre ćwiczenie: po lekturze przykładowego planu możemy spróbować zastanowić się, czy rzeczywiście odpowiada on na wszystkie nasze pytania dotyczące zarówno samego badania (np. czy jest dla nas jasne, co jest jego celem, jakie dane zostaną w związku z nim wytworzone i kogo mogą one zainteresować), jak i danych (m.in. czy – a jeśli tak, to kiedy, gdzie i na jakich zasadach – zostaną one udostępnione, a także, czy możliwe będzie ich wykorzystanie np. do celów badawczych lub komercyjnych).

Choć wzory planów zarządzania danymi różnią się rozłożeniem akcentów i niekiedy opisują te same zagadnienia w różny sposób, wymieniają często podobne kwestie wymagające przeanalizowania w obrębie przygotowywanego dokumentu. Obejmują one zazwyczaj zagadnienia związane z wytworzeniem i/lub pozyskaniem danych, formatami plików i ich nazewnictwem, przechowywaniem danych, zasadami dostępu do nich oraz ich ponownego wykorzystania, aspektami prawnymi i etycznymi oraz niezbędnymi zasobami (zarówno finansowymi, jak i infrastrukturalnymi oraz kompetencyjnymi).

Data management plan w Horyzoncie 2020

Planem zarządzania danymi, z którym najczęściej mają do czynienia polscy naukowcy, jest dokument wymagany w projektach finansowanych w ramach programu Horyzont 2020. Zgodnie z zasadami FAIR, dokument ten powinien opisywać, jakie działania zostaną podjęte, aby wytworzone w ramach projektu dane były łatwe do odnalezienia, dostępne, możliwe do powiązania z innymi oraz do ponownego wykorzystania.

Tworząc plan zarządzania danymi na potrzeby programu Horyzont 2020, należy kierować się ogólną zasadą *as closed as necessary, as open as possible*. Podczas analizy kwestii danych w naszym projekcie powinniśmy więc zakładać ograniczenie dostępu do nich jedynie w takim stopniu i zakresie, w jakim jest to konieczne, zaś pojawiające się przeszkody starać się usuwać, zamiast czynić z nich pretekst dla ograniczenia tego dostępu.

Dane w ujęciu ogólnym

Punktem wyjścia dla planu zarządzania danymi powinno być ogólne określenie tego, jakie dane będą gromadzone lub wytwarzane w ramach projektu i w jaki sposób przyczynią się one do osiągnięcia jego celów.

W pierwszej kolejności trzeba rozważyć, czy wykorzystane zostaną już istniejące dane pochodzące ze źródeł zewnętrznych. Należy zadać pytanie, czy podstawa prawna, na jakiej będą one wykorzystane, umożliwia dalsze udostępnianie ich samych lub danych, które powstaną przy ich wykorzystaniu. Konieczne może okazać się zawarcie umów regulujących zakres możliwego wykorzystania i dalszego ich udostępniania, analiza i ujednolicenie ich dokumentacji oraz opisu czy przekonwertowanie ich na jednolity format, co ułatwi późniejsze ich powiązanie z nowo wytworzonymi danymi.

Jeśli wszystkie dane zostaną wytworzone w ramach projektu badawczego, odpowiednio wcześnie podjęte ustalenia odnośnie do tego, kto będzie nimi dysponować i w jaki sposób będą one wykorzystywane, mogą znacząco ułatwić ich udostępnienie i ponowne wykorzystanie. Jest to tym istotniejsze, im więcej osób i instytucji zaangażowanych jest w realizację projektu. Zawarcie stosownych umów

pozwoli uniknąć sporów i nieporozumień oraz wynikających z nich opóźnień na późniejszych etapach realizacji projektu.

Kolejną kwestią, którą należy uwzględnić w ramach DMP, jest spodziewany rozmiar wytwarzanych i udostępnianych danych. Pomoże to oszacować szeroko rozumiane nakłady potrzebne do ich krótko- i długoterminowego przechowywania. Jeśli dane są niewielkich rozmiarów, zapewne istniejąca i dostępna bez dodatkowych opłat infrastruktura repozytoryjna okaże się w pełni wystarczająca do ich przechowywania i udostępniania. Jeśli jednak wielkość danych liczona będzie np. w petabajtach, konieczne może okazać się zapewnienie lub stworzenie przeznaczonej specjalnie dla nich infrastruktury (np. macierzy dyskowych) oraz oprogramowania (umożliwiającego np. odfiltrowanie i pobranie jedynie potrzebnej części danych albo zdalne prowadzenie analiz przez inne osoby zainteresowane ich wykorzystaniem).

Rozważyć też należy, dla kogo wytwarzane dane mogą być użyteczne. Im więcej osób i/lub instytucji może chcieć je wykorzystać, tym więcej powodów, by podjąć wysiłek zmierzający do ich udostępnienia i opracowania w sposób przystępny dla wszystkich potencjalnych użytkowników. Identyfikacja danych na wczesnym etapie projektu badawczego będzie w tym bardzo pomocna.

W całym procesie udostępniania danych jedną z najważniejszych zasad jest stosowanie szeroko przyjętych standardów wszędzie tam, gdzie to możliwe. Jeśli do opisu naszego zbioru danych wykorzystamy słowa kluczowe powszechnie stosowane w danej dyscyplinie, osoba zainteresowana ich wykorzystaniem nie powinna mieć trudności z odnalezieniem tego zbioru oraz ustaleniem znaczenia użytych do jego opisu terminów. Chociaż z prawnego punktu widzenia nic nie stoi na przeszkodzie, byśmy swój zbiór danych opatrzyli licencją napisaną samodzielnie lub przez zaprzyjaźnionego prawnika, nie ułatwi to jednak ustalenia, czy konkretny sposób jego użycia jest dopuszczalny na mocy takiej licencji oraz czy i jak można łączyć go ze zbiorami danych pochodzącymi z innych źródeł. Odstępstwo od ustalonych reguł zawsze powinno być sytuacją rzadką, poprzedzoną dogłębną refleksją nad tym, czy rzeczywiście jest ono uzasadnione i niezbędne.

Findable data, czyli jak ułatwić innym znalezienie naszych danych

Podstawowym elementem, ułatwiającym wszystkim zainteresowanym odnalezienie naszego zbioru danych, jest odpowiedni ich opis za pomocą metadanych. W pierwszej kolejności powinniśmy do jego przygotowania wykorzystać istniejące, standardowe schematy stosowane w naszej dziedzinie. Jeśli takie schematy nie istnieją, warto wykorzystać istniejące schematy ogólnego zastosowania, uwzględniające podstawowe parametry zbioru danych, i ewentualnie rozszerzyć je o dodatkowe informacje.

W odnalezieniu i jednoznacznej identyfikacji zbioru pomagają również trwałe i unikalne identyfikatory, takie jak popularne DOI (Digital Object Identifier) lub Handle (często spotykany również w repozytoriach publikacji). Umożliwiają one dotarcie do zbioru nawet wtedy, gdy zmieni się jego internetowy adres URL (ponieważ np. zostanie on przeniesiony do innego repozytorium danych lub zmieni się oprogramowanie wykorzystywane przez repozytorium obecne).

Do opisu danych warto stosować przyjęte w danej dziedzinie konwencje nazewnicze – ułatwi to odnalezienie zbiorów dotyczących tych samych przedmiotów badań. Należy również zadbać o to, by dane opatrzone były odpowiednimi słowami kluczowymi. Jeśli w danej dyscyplinie istnieją ich standardowe słowniki, to właśnie nimi w pierwszej kolejności powinniśmy się posłużyć.

Dane, które zamierzamy udostępnić, mogą zmieniać się w czasie. Mogą one mieć charakter przyrostowy, co znaczy, że w miarę upływu czasu ich ilość będzie się zwiększać. Możliwe jest też, że na dalszych etapach badań będziemy chcieli udostępniać kolejne, coraz lepiej opracowane wersje zbioru, mające zastąpić poprzednie. Nie znaczy to jednak, że starsze wersje mogą w takim wypadku zostać usunięte jako bezużyteczne, w międzyczasie mogły bowiem zostać zacytowane lub wykorzystane przez innych badaczy. Przewidując, że nasz zbiór danych będzie się zmieniać, powinniśmy wybrać dla niego jasny sposób identyfikacji kolejnych wersji oraz takie miejsce udostępniania, które pozwala na jego implementację.

Podejmując decyzję o wyborze miejsca, w którym zamieścimy nasze dane, warto wziąć pod uwagę to, czy będzie można odnaleźć je w istniejących wyszukiwarkach danych badawczych (np. prowadzonej przez Google https://toolbox.google.com/datasetsearch). Znalezienie się w takich wyszukiwarkach zdecydowanie ułatwi odnalezienie naszych danych osobom zainteresowanym.

Accessible data, czyli jak zapewnić otwarty dostęp do danych

Jeśli jakieś zbiory nie mogą zostać udostępnione lub dostęp do nich musi zostać ograniczony, w DMP należy wyjaśnić, dlaczego jest to konieczne. Wyróżnić możemy ograniczenia o charakterze prawnym, umownym i woluntarystycznym. W pierwszym przypadku na drodze otwartego udostępnienia danych stają obowiązujące przepisy prawa, w drugim – umowne zobowiązania nasze lub naszej instytucji, w trzecim – to, w jaki sposób planujemy wykorzystać dane w trakcie projektu i po jego zakończeniu.

O tym, w jakim stopniu dane będą dostępne dla innych zainteresowanych, decyduje również miejsce, w którym zostaną one udostępnione. Najbardziej oczywistym – a co za tym idzie, zalecanym w pierwszej kolejności – są repozytoria danych, umożliwiające ich udostępnianie w modelu otwartym. Również w tym wypadku, jeśli to tylko możliwe, powinniśmy starać się korzystać z rozwiązań istniejących. Może w tym pomóc witryna https://www.re3data.org/, zawierająca informacje o ponad 2300 repozytoriach danych, dzięki której dużo łatwiej będzie nam znaleźć serwis, w którym możemy zarchiwizować i udostępnić wyniki badań.

Repozytoria danych mogą – choć nie muszą – mieć określone wymogi o charakterze prawnym czy technicznym (np. dotyczące formatów akceptowanych w repozytorium dziedzinowym lub maksymalnej wielkości zbioru), których spełnienie wymaga odpowiednich przygotowań. Wymogi te mogą rzutować na sposób bieżącego opracowywania danych, więc warto zapoznać się z nimi przed przystąpieniem do niego i uwzględnić je na możliwie wczesnym etapie. Na przykład specjalistyczne repozytorium

PDB (Protein Data Bank, http://www.wwpdb.org) wymaga deponowania danych dotyczących części struktur białkowych jedynie w formacie PDBx/mmCIF, a maksymalna wielkość zbioru danych, który można zdeponować w repozytorium Zenodo, wynosi 50GB.

Udostępniając dane w repozytorium, warto również skorzystać z możliwości, jakie dają *data journals*, czyli czasopisma naukowe publikujące recenzowane artykuły opisujące dostępne zbiory danych. Czasopisma te stanowią odpowiedź na fakt, że z punktu widzenia systemu nauki w wielu krajach opracowanie i udostępnienie zbioru danych nie stanowi osiągnięcia mogącego wpływać na rozwój kariery naukowej badaczy. Jeśli udostępnieniu zbioru towarzyszy publikacja artykułu w takim czasopiśmie, osoby wykorzystujące ten zbiór otrzymają możliwość równoczesnego zacytowania zbioru i opisującego go artykułu. Cytowanie tego drugiego rodzaju może pozytywnie wpłynąć na rozwój kariery naukowej już teraz, bez konieczności czekania na jakiekolwiek zmiany systemowe.

Aby inne zainteresowane osoby mogły skorzystać z danych, udostępniony zbiór powinien zawierać również dokumentację dotyczącą potrzebnego do tego oprogramowania. Jeśli to możliwe, można jego elementem uczynić samo oprogramowanie; nie ma z tym zwykle problemu, jeśli ma ono status wolnego i otwartego.

Nawet jeśli dane z jakichś powodów nie mogą być udostępnione w modelu otwartym, wciąż możemy i powinniśmy zastanowić się nad innymi formami ich udostępnienia. Postępując zgodnie z zasadą *as closed as necessary, as open as possible*, możemy spróbować udostępnić je w modelu ograniczonego dostępu, np. wyłącznie do celów naukowych. Niektóre repozytoria pozwalają na udzielenie dostępu do danych konkretnemu, pojedynczemu użytkownikowi (bywają to te same repozytoria, w których dominującym modelem udostępniania danych jest model otwarty). Gdy dane będą umieszczone w takim miejscu, jego użytkownikom bardzo łatwo będzie wystąpić o konieczną zgodę, a dysponent danych w równie prosty sposób będzie mógł jej udzielić.

Aby zobiektywizować proces podejmowania decyzji odnośnie do udostępnienia danych konkretnej osobie, można powołać tzw. *data*

access committee. To składające się z ekspertów gremium, które decyduje o tym, czy w konkretnym przypadku zgoda taka powinna zostać udzielona, czy nie.

O dostępności danych decyduje również to, jaką licencją zostaną one objęte oraz w jaki sposób ich potencjalni użytkownicy zostaną o tym fakcie poinformowani. Optymalna jest sytuacja, w której są one opatrzone licencją w postaci nadającej się do odczytu zarówno przez ludzi, jak i przez maszyny, które również mogą być odbiorcami naszych danych.

Jeśli z jakichś powodów nie możemy udostępnić samych danych lub musimy wycofać uprzednio udostępniony zbiór, powinniśmy zadbać o dostępność opisujących je metadanych. W pierwszym przypadku będą one sygnałem, iż dane w ogóle istnieją, w drugim zaś – oznaką tego, że w określonym czasie były one udostępnione i mogły zostać zacytowane lub poddane wtórnym analizom.

Interoperable data, czyli jak umożliwić powiązanie naszych danych z innymi danymi

Mówiąc o interoperacyjności danych, mamy na myśli to, czy jest możliwa ich wymiana i wykorzystanie przez inne osoby i podmioty (w tym maszyny), w szczególności zaś przez innych naukowców. Interoperacyjność można zapewnić, trzymając się tak bardzo, jak to tylko możliwe, istniejących standardów dotyczących formatów metadanych i samych danych, a także oprogramowania (ze szczególnym wskazaniem na wolne oprogramowanie, które w przyszłości będzie mogło być modyfikowane). Jeśli wykorzystanie zbioru wymaga opracowania translatora, sposobu mapowania czy konwersji formatów, świadczy to o jego ograniczonej interoperacyjności.

Standaryzacja dotyczy również licencji, na jakiej udostępnione zostały dane. Posługując się jedną z popularnych, znanych w środowisku licencji (takich jak licencje Creative Commons w wersji 4.0) lub zwolnieniem ze zobowiązań (tzw. *waiver*) Creative Commons Zero ułatwimy innym użytkownikom danych zrozumienie, jaki sposób wykorzystania danych jest dopuszczalny i czy możliwe jest ich połączenie z innymi istniejącymi zbiorami.

Jeśli to tylko możliwe, powinniśmy korzystać z istniejących standardów w zakresie metadanych (np. stosować posiadające dobrą dokumentację i opatrzone trwałym identyfikatorem słowniki kontrolowane). Uzyskana dzięki nim jednoznaczność ułatwi zadanie innym osobom chcącym połączyć wiele zbiorów danych.

Jeśli z jakichś powodów konieczne jest utworzenie na potrzeby projektu nowego słownika, zbiór danych powinien zawierać również mapowanie pól tego słownika na inne istniejące słowniki, co także ułatwi jego powiązanie z innymi, już istniejącymi danymi.

Z kolei jeśli nasze dane bazują na innych zbiorach danych, wymagają użycia określonego oprogramowania lub w jakikolwiek inny sposób zależą od zasobów zewnętrznych, ich zbiór powinien zawierać odpowiednie odniesienie do tych zasobów, co ułatwi innym skorzystanie z niego.

Należy pamiętać, że odbiorcami danych i metadanych mogą być nie tylko ludzie, lecz także maszyny, czyli zewnętrzne systemy komputerowe – istotne jest, aby zarówno metadane, jak i dane dostępne były również dla nich. O pełnej interoperacyjności możemy więc mówić wtedy, gdy zdeponowany przez nas zbiór dostępny jest za pośrednictwem API, a jego metadane – dodatkowo za pośrednictwem protokołu OAI-PMH.

Reusable data, czyli jak umożliwić ponowne wykorzystanie danych

Ponownemu wykorzystaniu danych sprzyja bogaty zestaw opisujących je metadanych. O ile podstawowe informacje będą pomocne w samym odnalezieniu zbioru, o tyle ustalenie, czy dane będą przydatne w określonym kontekście, może wymagać informacji o wykorzystanej metodologii, próbkach i materiałach, instrumentach badawczych czy ich konkretnych parametrach. Zakres tych informacji powinien być możliwie szeroki, nie jesteśmy bowiem w stanie z góry przewidzieć, czego potrzebować będą wszystkie osoby zainteresowane wykorzystaniem naszego zbioru teraz i w przyszłości.

Pomocne będą również informacje o pochodzeniu danych i kolejnych etapach ich przetwarzania – najlepiej w postaci umożliwiającej odczyt maszynowy.

Jeśli w naszej dziedzinie istnieją standardy określające to, jakiego rodzaju dane powinny znaleźć się w zbiorze, jak powinny być zorganizowane, udokumentowane i sformatowane, to trzymanie się ich również ułatwi ponowne wykorzystanie naszych danych.

To, czy możliwe będzie ponowne wykorzystanie udostępnionych danych – a jeśli tak, to w jakim zakresie – zależy też od tego, na jakiej licencji zostaną one udostępnione. Warto przy tym wykorzystać istniejące licencje, takie jak Creative Commons Zero czy inne licencje Creative Commons w wersji 4.0. Z każdym rokiem stają się one coraz popularniejsze – nie tylko w środowisku akademickim – więc, korzystając z nich znacząco ułatwiamy odbiorcom uzyskanie informacji o tym, jaki sposób wykorzystania naszych danych jest dopuszczalny. Dla coraz większej liczby osób, które znają licencje Creative Commons, sama ich nazwa lub graficzne oznaczenie niosą ze sobą informację o warunkach licencji. Natomiast jeśli zdecydujemy się na napisanie własnej licencji lub zlecimy to zadanie prawnikowi, każdy użytkownik będzie zmuszony uważnie przestudiować jej unikalną treść, która dodatkowo nie będzie zapewne dostępna w postaci zrozumiałej dla maszyn.

W przypadku programu Horyzont 2020 istotne jest dołożenie wszelkich starań, aby ponowne wykorzystanie danych było możliwe w jak najszerszym zakresie. Ewentualne ograniczenia powinny występować jedynie w tych przypadkach, w których pojawią się niedające się usunąć przeszkody (przykład: w naszych badaniach wykorzystaliśmy dane zewnętrzne, które stanowią integralny element zbioru danych, a których szerokie udostępnienie nie jest możliwe np. z powodów etycznych lub dlatego, że osoba, w gestii której się one znajdują, nie godzi się na to). W pierwszej kolejności powinniśmy postarać się, aby – w takim stopniu, w jakim to możliwe – usunąć przeszkody (np. w drodze negocjacji dotyczących problematycznego zbioru). Dzięki temu do ograniczenia dostępu do danych dochodzić będzie jedynie w ostateczności.

Należy również rozważyć kwestię, kiedy – w którym momencie realizacji projektu czy po jego zakończeniu – chcielibyśmy lub możemy udostępnić nasze dane. Pierwszym powodem, dla którego możemy zdecydować się na opóźnienie ich udostępnienia, może być chęć

uzyskania patentu na opracowany w ramach projektu wynalazek. Upublicznienie jego istoty przed złożeniem wniosku w urzędzie patentowym powodować będzie niemożność uzyskania patentu, w związku z czym wstrzymanie się z udostępnieniem zbioru danych, które niosłoby za sobą takie ryzyko, może być uzasadnione.

Udostępnienie danych można odroczyć również w związku z planami publikacyjnymi, które zależą od opracowanych w ramach projektu zbiorów danych. W tym przypadku embargo wynika z obawy przed utratą pierwszeństwa odkrycia na rzecz dysponujących większymi zasobami finansowymi i/lub kadrowymi naukowców z innych ośrodków, którzy byliby w stanie wykorzystać udostępnione dane szybciej. Okres embarga daje zespołowi, który wytworzył dane, czas niezbędny na ich opracowanie i opublikowanie wyników – a jednocześnie gwarantuje, że po jego upływie dane będą publicznie dostępne.

W takiej sytuacji dobrym rozwiązaniem jest zdeponowanie danych tak szybko, jak to tylko możliwe, w jednym z repozytoriów oferujących funkcję embarga, z jednoczesnym określeniem daty, od której dane będą publicznie dostępne. Dzięki temu możemy oddzielić moment umieszczenia ich w repozytorium od momentu ich udostępnienia, a jednocześnie nie musimy zwlekać ze zdeponowaniem danych – ich przygotowanie i umieszczenie w repozytorium po upływie dłuższego czasu od ich opracowania (np. po kilku latach od zakończenia projektu) mogłoby być kłopotliwe.

W obu przypadkach ważne jest, aby – mimo istniejących ograniczeń – udostępnienie danych nastąpiło tak szybko, jak to tylko możliwe. Jeśli wniosek patentowy został już złożony, a wszystkie planowane przez nas publikacje zdążyły się ukazać, to ustały powody, dla których dane pozostawały zamknięte. W takiej sytuacji dane można udostępnić, nawet jeśli zakładany okres embarga jeszcze nie upłynął.

W tym miejscu należy również rozważyć to, przez jak długi czas po zakończeniu projektu dane będą pozostawały dostępne i możliwe do ponownego wykorzystania. Jeśli mamy do czynienia z niewielkim zbiorem, który po udostępnieniu wymaga jedynie niewielkiego zakresu działań (np. regularnego tworzenia i weryfikacji kopii bezpieczeństwa przez

technicznego administratora repozytorium) i który może być atrakcyjny jeszcze przez wiele lat, to nie ma powodu aby zakładać, że pochodzące z niego dane będą udostępniane jedynie przez określony czas.

Jeśli jednak mówimy o danych, które wymagają stworzenia osobnej infrastruktury (na przykład ze względu na ich wielkość) i/lub opracowania specjalistycznego oprogramowania służącego do ich udostępniania czy regularnej konwersji na nowe formaty – co zawsze wiąże się z koniecznością przewidzenia i poniesienia konkretnych kosztów – a zarazem takich, które ze względu na dokonujący się bardzo szybko postęp w danej dziedzinie mogą pozostać atrakcyjne dla potencjalnych użytkowników jedynie przez kilka lub kilkanaście lat, pytanie o horyzont czasowy ich udostępniania staje się zasadne.

W okresie, w którym dane te są rzeczywiście wykorzystywane przez inne osoby i instytucje, ponoszenie kosztów ich udostępniania ma sens. W miarę tego, jak zainteresowanie danymi będzie słabnąć, coraz trudniej będzie je jednak rozsądnie uzasadnić. W takiej sytuacji konieczne może być określenie, jak długo dane będą udostępniane, a co za tym idzie – przez jaki okres należy zapewnić konieczne do tego finansowanie.

Trzeba też pamiętać, że dane zarówno przed udostępnieniem, jak i po nim mogą wymagać licznych zabiegów mających na celu umożliwienie ich ponownego wykorzystania. Mogą one dotyczyć na przykład konwersji samych danych lub opisujących je metadanych na nowe formaty, które nie były jeszcze znane w momencie opracowywania i udostępniania zbioru.

Przed udostępnieniem danych można również zaplanować działania związane z kontrolą ich jakości. Jej wyniki również powinny stać się elementem zbioru danych. Jeśli na jej skutek jakieś dane zostały zmodyfikowane lub usunięte, informacja o tym fakcie powinna znaleźć się w opisie zbioru.

Etyczne aspekty udostępniania danych

Z udostępnianiem danych badawczych, w szczególności tych dotyczących ludzi, wiążą się również zagadnienia natury etycznej. Najczęściej dotyczą one możliwości ujawnienia tożsamości badanych (która to

kwestia ma również swój aspekt prawny) i wiążą się z pytaniem o to, czy udostępnienie danych może w jakiś sposób zaszkodzić innym osobom.

Sposobem na radzenie sobie z tą trudnością jest anonimizacja danych. Należy jednak pamiętać, że nie polega ona wyłącznie na usunięciu imienia i nazwiska osoby badanej, jej miejsca zamieszkania czy nazwy zakładu pracy. W szczególności w badaniach społecznych, w przypadku materiałów będących rezultatem wywiadów pogłębionych, możemy mieć do czynienia z tzw. efektem mozaiki: choć każdy fakt dotyczący rozmówcy wzięty z osobna nie wystarcza, by poznać jego tożsamość, to wzięte razem mogą już jednak umożliwiać jego identyfikację. Co więcej, zestawienie informacji, które z perspektywy osoby nieznającej badanego nie pozwala na identyfikację, komuś z grona jego znajomych, współpracowników czy rodziny może w zupełności do tego wystarczyć. Ludzie często zdają sobie sprawę z tego, że w danym miejscu prowadzone było badanie, w którym uczestniczyły określone osoby, co stanowi kolejny czynnik ułatwiający powiązanie ich tożsamości z udzielonymi wypowiedziami. Znajomi, współpracownicy i rodzina to jednocześnie często te osoby, przed którymi badacz powinien w pierwszej kolejności chronić swoich rozmówców: z faktu, że ich tożsamość uda się ustalić obcej osobie, mieszkającej w odległej miejscowości, może niewiele wynikać – co innego jednak, jeśli uda się to ich sąsiadowi lub przełożonemu.

Rozwiązaniem może być tu udostępnienie danych w modelu dostępu ograniczonego jedynie do celów naukowych, uzależnionego od akceptacji *data access committee* lub realizowanego w określonym miejscu (np. na wyznaczonym stanowisku w bibliotece lub w siedzibie instytucji sprawującej pieczę nad danymi). Jeśli występujących trudności nie daje się w żaden sposób usunąć, może być konieczne zamknięcie dostępu do danych. Rozwiązanie to stanowi jednak ostateczność.

Nawet jeśli decydujemy się nie udostępniać samych danych, możemy udostępnić ich metadane. Najłatwiej to osiągnąć umieszczając dane w repozytorium, które umożliwia zamknięcie dostępu do nich dla zewnętrznych użytkowników przy jednoczesnej publicznej prezentacji metadanych. Dzięki temu użytkownicy serwisu mogą odnaleźć opis

badania i zapoznać się z jego treścią, a w konsekwencji – skontaktować się z odpowiednią osobą i starać się uzyskać dostęp do danych na specjalnych zasadach. Tak zdeponowany zbiór przynosi również korzyści osobom biorącym udział w jego opracowaniu, jako że daje im możliwość łatwego powrotu do danych za pewien czas. Jest to dużo łatwiejsze, jeśli dane – nawet zamknięte – zostały zawczasu dobrze opracowane, opisane i zabezpieczone. Najważniejsze jednak, że wiadomo, gdzie ich szukać.

Warto pamiętać o tym, że choć z udostępnianiem danych mogą się wiązać wątpliwości natury etycznej, istnieją też jednak etyczne racje przemawiające za ich udostępnianiem. Jeśli dane, którymi dysponujemy, mogą w przyszłości umożliwić rozwiązanie jakiegoś istotnego problemu społecznego (np. opracowanie nowego leku lub sposobu na utylizację odpadów), to z pewnością jest to bardzo dobry powód do tego, by je udostępnić – zwłaszcza jeśli sami nie mamy w planach przeprowadzenia analiz mogących przynieść taki skutek. Jeśli badania konieczne do wytworzenia danych wiążą się dla badanych z jakimś ryzykiem (również o charakterze psychologicznym, jak miałoby to miejsce np. w przypadku badań, w których uczestniczyłyby ofiary przemocy), za ich udostępnieniem również stoją racje etyczne, gdyż udostępnienie danych pozwala ograniczyć – a nawet zupełnie wyeliminować – konieczność ponownego narażania tych osób na to same ryzyko w wypadku, gdyby ktoś chciał zrealizować podobne badania w przyszłości.

Koszty i zasoby

Wytworzenie, opracowanie i udostępnienie danych zgodnie z zasadami FAIR często wiąże się z koniecznością poniesienia wymiernych kosztów. Planując sposób zarządzania danymi, już na wczesnym etapie powinniśmy więc zadać sobie pytanie, jakimi zasobami dysponujemy, a jakie dopiero musimy pozyskać.

Rozważenie tej kwestii odpowiednio wcześnie ułatwi pozyskanie niezbędnych środków, które niekiedy – tak jak w przypadku projektów finansowanych w ramach Horyzontu 2020 – można uwzględnić w budżecie projektu.

Należy pamiętać o tym, że do udostępniania danych niezbędne są odpowiednie kompetencje i infrastruktura. Kwestia ta wiąże się bezpośrednio z pytaniami o to, jakie dane, w jaki sposób i jak długo będą udostępniane. Dopiero znając odpowiedzi na te pytania będziemy w stanie określić, jakich kompetencji i infrastruktury potrzebujemy.

Bezpieczeństwo danych

Bardzo istotne jest zadbanie o bezpieczeństwo danych. Zdeponowane i opublikowane dane nie mogą ulec nawet najmniejszym zmianom, nie mogą zostać zniszczone lub zdekompletowane; także dostęp do nich nie powinien podlegać technologicznym ograniczeniom.

Należy starannie wybrać sprzęt, dzięki któremu dane będa przetwarzane – dotyczy to zarówno serwerów, jak i przestrzeni dyskowej. Przy jego wyborze trzeba wziąć pod uwagę ilość danych i ich możliwy przyrost w czasie eksploatacji, a także kwestie utrzymania i serwisowania sprzętu (również w okresie utrzymania trwałości projektu). W dłuższej perspektywie konieczne jest planowanie modernizacji systemu.

Chcąc zapewnić nieprzerwany dostęp do danych, warto pomyśleć o lustrzanych lub zapasowych systemach serwerów i repozytoriów.

Bezpieczeństwo danych obejmuje także możliwość ich odzyskania, jeśli zostaną utracone. W tym kontekście istotne są odpowiednie systemy do tworzenia kopii zapasowych oraz zapasowe lub lustrzane systemy serwerów. Miejsca, w których składowane są kopie zapasowe, powinny być rozproszone geograficznie.

Dodatkowym wspomaganiem zapewnienia bezpieczeństwa danych są systemy archiwizacji, zbudowane na macierzach dyskowych lub na taśmach magnetycznych.

Oprogramowanie, na którym zbudowane są repozytoria, powinno spełniać powszechnie stosowane standardy i normy, które zapewnią bezpieczne, długoterminowe utrzymanie systemów i przechowywanie danych. Oprogramowanie powinno być monitorowane pod kątem rozwoju i aktualizacji, w szczególności – aktualizacji zabezpieczeń.

Cześć danych z ograniczonym dostępem powinna dodatkowo zostać zabezpieczona przed nieuprawnionym dostępem. Można to

zrealizować poprzez system rejestracji i logowania przy pomocy hasła. Bezpieczeństwo danych można także wzmocnić stosując technologię VPN lub szyfrowanie samych zbiorów. Narzędzia zabezpieczające powinny być jednak dobrane w sposób zgodny z powszechnie stosowanymi najlepszymi praktykami i standardami. Wdrożone zabezpieczenia powinny być monitorowane, aktualizowanie oraz modernizowane.

Uwzględnienie powyższych kwestii przez badaczkę zainteresowaną udostępnieniem danych nie musi, rzecz jasna, polegać na tym, by każdorazowo starać się samodzielnie zapewnić w danym aspekcie bezpieczeństwo danych. Warto jednak mieć je na uwadze podejmując decyzję o wyborze repozytorium, do którego trafią nasze dane – lub o potrzebach sprzętowych i kadrowych, jeśli decydujemy się na budowę własnej infrastruktury.

Inne kwestie

Dane badawcze, wzięte ogólnie, charakteryzują się bardzo dużą różnorodnością, więc nie sposób we wzorze planu zarządzania nimi ująć wszystkich istotnych czynników, które mogą wpływać na ich opracowanie i udostępnienie. DMP jest jednak właściwym miejscem na ich uwzględnienie, a ich wczesne rozpoznanie, przemyślenie i opisanie pozwoli uwzględnić je w toku realizacji projektu badawczego.

Podsumowanie

Opracowując plan zarządzania danymi, należy liczyć się z tym, że nie uda nam się już w jego pierwszej wersji uwzględnić wszystkich kwestii, które ostatecznie wpłyną na sposób opracowywanych i udostępnianych danych. Kształt badania często zmienia się w jego trakcie, pojawiają się nowe techniki badawcze, rozwija się również wykorzystywana w badaniach aparatura, przez co nierzadko rośnie (w stosunku do przewidywanej) wielkość danych uzyskiwanych w ramach projektu badawczego.

Rozwiązań przewidzianych w planie zarządzania danymi nie trzeba i nie należy trzymać się kurczowo, jeżeli w pewnym momencie stają one w sprzeczności z nowymi celami samego badania. Plan

zarządzania danymi to raczej żywy dokument, którego kolejne wersje powinny odzwierciedlać zmiany zachodzące w kształtującym go otoczeniu poznawczym, technicznym i kulturowym, zmienna jest bowiem zarówno nasza wiedza, jak i narzędzia, jakimi dysponujemy, oraz nasze oczekiwania względem tego, czym powinny być dane badawcze i kto powinien mieć prawo do ich wykorzystania. Na sam plan zarządzania można zresztą spojrzeć jako na jedno z ułatwiających rzetelne poznanie narzędzi, których racją istnienia jest to, czy są nam przydatne.