EBIB Biuletyn EBIB, nr 7 (177)/2017,
Surowe dane badawcze
Artykuły

Anna Małgorzata Kamińska
Instytut Bibliotekoznawstwa i Informacji Naukowej
Uniwersytet Śląski w Katowicach
anna.kaminska@us.edu.pl

Dobre praktyki publikowania danych badawczych

Streszczenie: *Artykuł przedstawia spostrzeżenia autorki na temat dobrych praktyk publikowania danych badawczych w zakresie ich dostępności, spójności, identyfikowalności, integralności, niezaprzeczalności, wiarygodności oraz wersjonowania. Stanowi on wyraz subiektywnych opinii wyrobionych na podstawie doświadczeń nabytych w trakcie realizacji badań własnych, w których konieczne okazało się wdrożenie praktyk ułatwiających operowanie na licznych i dużych zbiorach danych różnych typów. Po omówieniu kolejnych aspektów składających się na dobre praktyki publikowania danych przedstawiono przykład publikacji w repozytorium Zenodo danych dokumentujących autorskie badania bibliometryczne opisane w poprzednim numerze „Biuletynu EBIB". Pozwala to więc nie tylko zobrazować poszczególne aspekty dobrej praktyki, lecz także ułatwić wszystkim zainteresowanym prowadzenie badań własnych podobnych do opisanych w tamtym artykule.*

Słowa kluczowe: *autentyczność, dane badawcze, dobre praktyki, dostępność, spójność, identyfikowalność, integralność, niezaprzeczalność, repozytoria, wiarygodność, wersjonowanie, Zenodo*

Wprowadzenie

Rozwój nauki warunkowany jest możliwościami prowadzenia badań nad obiektami leżącymi w sferze zainteresowania poszczególnych jej dziedzin. Badania te w najogólniejszej definicji polegają na weryfikacji hipotez badawczych dotyczących bądź to własności badanych podmiotów, bądź też metod na nich operujących. Jako że w dobie upowszechnienia się technologii informatycznych komputery na stałe zagościły w środowiskach badawczych, badania naukowe w dużym uproszczeniu opisać można jako transformowanie danych wejściowych (źródłowych), opisujących wspomniane właściwości przedmiotów, przy pomocy metod naukowych do wyników, które najczęściej opisuje się danymi wyjściowymi (wynikowymi). W tak przyjętym modelu badań naukowych rozwój wiedzy naukowej następuje przez opracowywanie nowych (szybszych, prostszych, dokładniejszych) metod lub poprzez stosowanie metod już znanych dla danych opisujących cechy przedmiotów dziedzin, do których nie były jeszcze stosowane, co pozwala na odkrywanie nowych prawideł. Warto tutaj zauważyć, że dane wynikowe stanowić mogą dane źródłowe dla badań prowadzonych w ramach tej samej dziedziny bądź też przekraczać granice dziedzin, a przez to przyczyniać się do rozwoju innych gałęzi nauki.

O ile dokumentowanie nowo powstających czy ulepszanych metod naukowych praktykowane było w zasadzie od początku rozwoju samej nauki, to dokumentowanie danych badawczych, czy to z powodu barier ilościowych, czy technologicznych, jest zjawiskiem, które zaczęło się upowszechniać ze sporym opóźnieniem. Jego obecnemu rozwojowi sprzyja niewątpliwie rozwój sieci internet, technologii przetwarzania danych w chmurze oraz rozwój koncepcji otwartej nauki.

W zależności od poszczególnych dyscyplin naukowych dane badawcze mogą przybierać różną postać i opisywać całkiem odmienne właściwości różnych podmiotów przy pomocy całego wachlarza typów i formatów danych. Mogą to być więc ilościowe opisy cech badanej populacji wyrażane w postaci danych tabelarycznych, ciągi liczbowe urządzeń pomiarowych zebrane w zadanych interwałach czasu, zdigitalizowane zapisy bodźców odbieranych przez różne ludzkie zmysły w postaci plików dźwiękowych, obrazów, filmów i wiele innych. Cechą wspólną danych badawczych jest natomiast fakt, że stanowią one jedynie pochodną opisującą w mniejszym bądź większym przybliżeniu zaistniałe w otaczającej nas rzeczywistości byty lub zjawiska.

Warto zwrócić uwagę na przedrostek „surowe", który często poprzedza termin „dane badawcze" w polskiej literaturze. Jako surowe, a więc nieprzetworzone (zarówno w języku polskim, jak i angielskim przymiotnik ten odnosi się również do żywności) dane, mogłyby uchodzić za dane będące danymi źródłowymi dla jakichś metod, zaś dane będące wynikiem przetworzenia danych źródłowych przez konkretne metody można by określić jako przetworzone. Jak zauważono jednak wcześniej, dane wynikowe mogą stanowić dane źródłowe dla innych badań, stąd nie mając pewności, w jaki sposób publikowane dane będą wykorzystywane, wydaje się, że nie ma potrzeby określania ich dodatkowo w taki czy inny sposób, a przynajmniej określania wszystkich danych badawczych jako surowe. Dodatkowym argumentem wydaje się spostrzeżenie, że sam sposób pozyskania danych mogących być określonymi jako surowe jest skutkiem zastosowania metody ekstrakcji tych danych z otoczenia. Tymczasem może się okazać, że ekstrakcja ta może być wykonana na wiele sposobów prowadzących do różnych wyników. Spostrzeżenie to zdaje się pozostawać w zgodzie z trendami zagranicznymi, gdzie przykładowo wydawnictwo Elsevier wśród danych badawczych wymienia siedem kategorii[1], a dane surowe są tylko jedną z nich. Uwaga ta nie odnosi się oczywiście do dziedzin zamkniętych, gdzie przymiotnik „surowe" przyjmuje konkretne znaczenie – na przykład w fotografii „surowe pliki" to pliki będące „pełnym zrzutem danych" zarejestrowanych przez element światłoczuły w chwili zrobienia zdjęcia i nieprzetworzonych do powszechnie znanych formatów opisu obrazu.

Upowszechnianiu i rozwojowi większości idei towarzyszą zwroty i kierunki, do których weryfikacji potrzebny jest czas na wypracowanie pozytywnej bądź negatywnej oceny poszczególnych doświadczeń. Podobnie metody upowszechniania czy publikowania danych badawczych ewoluowały od pewnego czasu, owocując powstaniem wielu dedykowanych platform dających możliwość ich upublicznienia zainteresowanym badaczom. Bez względu jednak na to, czy dane publikowane są z poziomu takich właśnie platform, czy z poziomu infrastruktury własnej (strony WWW, inne usługi dedykowane współdzieleniu plików), pewne spostrzeżenia dotyczące dobrych praktyk ich publikowania wydają się mieć charakter uniwersalny i zostały przedstawione poniżej.

Autorka nie uzurpuje sobie jednak prawa do traktowania niniejszego opracowania jako wyczerpującego, dzieli się natomiast własnymi spostrzeżeniami i wnioskami na podstawie własnych doświadczeń w publikowaniu danych badawczych oraz korzystaniu z danych publikowanych przez innych badaczy.

[1] *Research Data* [online]. Elsevier 2017. [Dostęp 2.10.2017]. Dostępny w: https://www.elsevier.com/about/our-business/policies/research-data.

Plan zarządzania danymi

W zależności od charakteru, celu i beneficjenta prowadzonych badań, jeszcze przed ich rozpoczęciem warto zastanowić się nad sposobem ich dokumentowania, a w szczególności dokumentowania danych badawczych. Niektóre instytucje wymagają wręcz formalnego dokumentu opisującego plan zarządzania danymi wykorzystywanymi podczas realizacji konkretnych badań bądź wręcz narzucają własne polityki postępowania. Mogą one dotyczyć zakresu dokumentowania wyników badań, formatów plików przechowujących te wyniki, przeglądu licencji danych wykorzystywanych w badaniach, a pozyskanych od zewnętrznych podmiotów, narzędzi oraz platform wykorzystywanych do składowania wyników badań, określenia osób odpowiedzialnych za poszczególne czynności związane z retencją danych czy nawet zdefiniowania procedur obowiązujących poszczególne osoby. Prowadząc badania własne, warto od początku zastanowić się nad sposobem i miejscem składowania danych, ich udostępniania w przypadku projektów zespołowych oraz zabezpieczania ich przez przypadkową utratą.

Chcąc natomiast podejść do zagadnienia w sposób bardziej formalny, warto skorzystać z dedykowanych platform, które umożliwiają opracowanie własnego planu bądź skorzystanie z planów opublikowanych przez inne podmioty. Przykładem takiej platformy może być DMPTool[2] (ang. Data Management Plan Tool), w której w zależności od potrzeb formalizacji udokumentować możemy na przykład sposoby implementacji poniżej przedstawionych pozostałych aspektów.

Dostępność

Celem publikowania danych badawczych jest ich udostępnienie innym badaczom. W zależności od potrzeb udostępnienie to może być zrealizowane dla wybranych osób, grup lub dla wszystkich zainteresowanych. Warto zawczasu zwrócić uwagę na cel tego udostępnienia, co wpływa na zasady, na jakich dane mogą zostać wykorzystane przez innych badaczy. Platformy współdzielenia danych badawczych często pozwalają wybrać spośród wielu konkretną licencję udzieloną potencjalnym użytkownikom danych. Spośród najbardziej upowszechnionych licencji znajdujących zastosowanie w odniesieniu do danych wymienić należy Creative Commons wraz z jej wariantami:

- BY (zezwalającą na kopiowanie, dystrybucję, wyświetlanie i użytkowanie danych pod warunkiem umieszczenia informacji o autorze),
- NC (zezwalającą kopiowanie, dystrybucję, wyświetlanie i użytkowanie danych tylko dla celów niekomercyjnych),
- ND (zezwalającą na kopiowanie, dystrybucję, wyświetlanie i użytkowanie tylko pierwotnych danych; niedozwolona jest ich zmiana i tworzenie zbiorów pochodnych) oraz
- SA (zezwalającą na kopiowanie, dystrybucję, wyświetlanie i użytkowanie pochodnych zbiorów danych, pod warunkiem, że będą one publikowane na tej samej licencji), przy czym możliwe jest łączenie powyższych niewykluczających się wariantów.

„Dostępność" jest to również termin dotyczący odporności systemu na awarie, czyli możliwości nieprzerwanego świadczenia usług. Warto zwrócić uwagę na warunki świadczenia

[2] *DMPTool* [online]. [Dostęp 2.10.2017]. Dostępny w: https://dmptool.org.

usług poszczególnych platform, choć trzeba zaznaczyć, że potencjalne możliwości ich nieprzerwanego świadczenia oferują platformy uruchamiane w centrach przetwarzania danych renomowanych dostawców, a dane tam przechowywane mogą być chronione przed utratą w sposób niewspółmiernie wyższy od danych składowanych na jakichkolwiek nośnikach prywatnych.

Spójność

Dane dokumentujące badania nierzadko składają się z kilku zbiorów zapisanych w niezależnych plikach. Jednak pomiędzy tymi danymi mogą występować zależności. Prosty przykład stanowić mogą dwa pliki, z których jeden zawiera informacje o użytkownikach biblioteki, a drugi o wypożyczanych przez nich książkach. Już na etapie publikacji danych należy zatroszczyć się o to, żeby obydwa pliki odzwierciedlały stan bazy danych sytemu bibliotecznego z tego samego momentu, w przeciwnym razie zdarzyć się może, że dane w pliku o wypożyczeniach mogą się odwoływać do użytkowników, którzy nie byli jeszcze zapisani do biblioteki. Pliki te dobrze jest więc publikować jako jeden nierozłączny byt, tak aby u innych potencjalnych badaczy nie wzbudzać niepotrzebnych wątpliwości, czy dane pobrane jako dwa autonomiczne byty zawierają spójne dane.

W zależności od rodzaju prowadzonych badań do analiz danych wykorzystywane mogą być narzędzia, które wyniki badań składować mogą w plikach o różnych formatach. Jeśli analizy te mają charakter wielotorowy (tzn. prowadzone są równolegle na tych samych danych źródłowych – w odróżnieniu od charakteru etapowego, gdzie dane wynikowe jednego etapu mogą stanowić dane źródłowe kolejnego), a wyniki analiz każdego z tych torów mają wpływ na ostateczną konkluzję badań, to publikacja analizowanych danych, jako jednego nierozłącznego bytu, wydaje się przynosić więcej korzyści niż podział całego zbioru danych na podzbiory zawierające pliki w poszczególnych formatach.

Integralność

Publikowanie jakichkolwiek danych badawczych pozbawione byłoby sensu, gdyby dane pobierane przez innych badaczy nie stanowiły dokładnej kopi danych opublikowanych. Niestety zdarza się nierzadko, że dane pobierane z odległych platform, które są transmitowane przez sieć komputerową ulegają przekłamaniu. Również dane składowane na nośnikach lokalnych, a w szczególności zewnętrznych napędach dyskowych czy też kartach pamięci, są podatne na utratę integralności. Biorąc pod uwagę powyższe, zainteresowany badacz, korzystając z publikowanych danych badawczych, powinien mieć możliwość weryfikacji integralności tych danych nie tylko w fazie ich pobierania z dedykowanych platform, lecz także przed rozpoczęciem ich wykorzystania do prowadzenia własnych analiz. Najprostszym i najbardziej popularnym mechanizmem zapewniającym kontrolę integralności danych jest obliczanie sum kontrolnych i publikowanie ich razem z danymi badawczymi. Co prawda niektóre z bardziej zaawansowanych formatów danych mają te mechanizmy już wbudowane, jednak zwykle są to formaty obsługiwane przez węższą grupę aplikacji, co może znacząco ograniczać możliwości wykorzystania danych na innych platformach. Wśród najczęściej wykorzystywanych algorytmów obliczania sum kontrolnych wymienić należy rodziny MD oraz SHA, z czego największą popularność zyskały MD5 oraz SHA1. Istnieje wiele aplikacji (w tym również ogólnodostępnych) zarówno o prostych,

jak i rozbudowanych funkcjonalnościach umożliwiających obliczanie sum kontrolnych, które następnie mogą być publikowane razem z danymi badawczymi.

Inną metodą zapewnienia integralności danych może być zastosowanie aplikacji do kompresji i archiwizacji danych, które podczas odtwarzania archiwum automatycznie sprawdzają jego integralność. Jednak niektóre platformy udostępniania danych badawczych umożliwiają przeglądanie zawartości plików z danymi bez potrzeby ich pobierania, zaś publikowanie danych w formie skompresowanych archiwów znacznie tą możliwość ogranicza.

Wiarygodność

Aby dane mogły stanowić podstawę wiarygodnych prac badawczych, same muszą zostać uwiarygodnione. Dzieje się to zwykle poprzez rzetelny opis procesu badawczego przedstawionego w publikacji naukowej z dokładnością pozwalającą na odtworzenie tego procesu i weryfikację rezultatów otrzymanych przez danego naukowca. Czasem jednak publikacji danych nie towarzyszy dodatkowy artefakt naukowy dokumentujący ich pochodzenie. W takich przypadkach opis sposobu pozyskania danych powinien zostać zamieszczony wraz z samymi danymi. Dobrym przykładem może być tutaj upublicznienie badań ankietowych przeprowadzonych na pewnej populacji, które mogą dać początek analizom realizowanym przez wielu badaczy. Jednak aby oszacować poziom wiarygodności uzyskanych wyników, potrzebna jest wiedza o sposobie przeprowadzenia tych badań, tak aby na przykład każdy mógł samodzielnie oszacować reprezentatywność próby badawczej. Podobna sytuacja ma miejsce w przypadku gromadzenia danych pomiarowych, ekstrakcji danych ze składnic informacji i wielu innych. Jeśli publikowane dane zostały dodatkowo w jakiś sposób wstępnie przetworzone, powinno to również zostać opisane.

Identyfikowalność

W czasach ciągłego wzrostu ilości przetwarzanych informacji rośnie również liczba danych badawczych. O ile kiedyś w ramach poszczególnych dziedzin zbiorom danych wystarczyło nadać nazwy i w ten sposób identyfikować je w pracach naukowych, o tyle obecnie przestaje to być możliwe.

Doprowadziło to do przejściowej sytuacji, w której dane badawcze identyfikowane były miejscem ich publikacji w sieci internet przy pomocy opisów URL. Jednak połączenie funkcji identyfikacji z opisem lokalizacji nie należy do dobrych praktyk ze względu na ulotność nazw domen internetowych oraz struktur witryn WWW – po pewnym czasie od publikacji adresu URL danych badawczych może się okazać, że stał się on nieaktualny lub wskazuje na zupełnie inne zasoby. Spostrzeżenie to przyczyniło się do separacji koncepcji identyfikacji i umiejscowienia zasobów i skutkowało powstaniem kilku standardów identyfikacji, spośród których największą popularność zyskał standard Handle System opracowany przez CNRI (Corporation for National Research Initiatives), a rozwinięty i wdrożony przez IDF (International DOI Fundation) pod nazwą DOI (ang. *digital object identifier*). Stanowi on cyfrowy identyfikator dla dowolnego przedmiotu własności intelektualnej, a jego zadaniem jest stałe identyfikowanie dowolnych obiektów w sieciach cyfrowych w powiązaniu z aktualnymi danymi na jego temat. Dzięki niemu można nadać jeden identyfikator zbiorowi danych badawczych, opisać go metadanymi, a następnie opublikować go na

dowolnych platformach pod tym samym identyfikatorem. Pozwala to na jednoznaczną identyfikację danych badawczych w pracach naukowych je wykorzystujących, w sposób jednolity i niezależny od możliwych przyszłych zmian lokalizacji tych danych. Mimo że sposób nadawania DOI jest ściśle kontrolowany przez agencje powołane przez IDF, to pozyskanie identyfikatorów dla własnych potrzeb nie jest trudne i często wspierane wprost przez platformy służące udostępnianiu danych badawczych.

Wersjonowanie

Wersjonowanie to koncepcja publikowania danych opisujących wybrane zagadnienie przy pomocy różnych wersji badawczych zbiorów danych. Potrzeba tworzenia nowych wersji wcześniej opublikowanych danych może wynikać ze zmienności w czasie atrybutów opisujących dany byt, ciągłości procesu obserwacji skutkującego potrzebą uaktualniania danych, zmiany/doskonalenia metod badawczych pozwalających na uzyskanie bardziej dokładnych wyników analiz i innych. Zwykle zakłada się, że kolejne wersje zbiorów danych badawczych posiadać będą podobną, jeśli nie identyczną strukturę i opisywać te same zagadnienia.

Aby stosowanie mechanizmów wersjonowania nie wprowadzało chaosu w miejsce oczekiwanych korzyści, musi być ono odpowiednio zaadresowane przez koncepcję identyfikacji zasobów. Mimo że system DOI nie przewiduje dedykowanych możliwości tworzenia identyfikatorów wersji, to zwykle w przypadku potrzeby wersjonowania zbiorów badawczych realizuje się to poprzez nadanie identyfikatora DOI dla wszystkich wersji danego zbioru (używanego do opisywania samej koncepcji – czyli najczęściej źródła danych, stosowanych metod i innych) oraz dla niezależnych identyfikatorów DOI dla każdej z wersji z osobna. Daje to możliwość cytowania przez innych badaczy zarówno konkretnych wersji (np. w przypadku realizacji analiz na danych), jak i jedynie koncepcji (np. w przypadku luźniejszego odnoszenia się do stosowanych metod badawczych, sposobu pozyskiwania danych).

Autentyczność i niezaprzeczalność

„Autentyczność” i „niezaprzeczalność” to terminy zapożyczone z dziedziny koncepcji podpisu elektronicznego. W tradycyjnym modelu publikowania prac naukowych to na wydawcy ciąży obowiązek dbałości o weryfikację tożsamości autorów i zapewnienia czytelników, że publikowane artykuły zgłoszone były przez podpisujących się pod nimi badaczy. Upowszechnianie się koncepcji publikowania prac naukowych w modelach otwartych przenosi odpowiedzialność zapewnienia niezaprzeczalności publikowanych artykułów na platformy otwartej nauki, które wywiązują się z niej zwykle słabo (poprzez weryfikację domeny zadeklarowanego konta pocztowego) lub wcale. Wzrost presji i konkurencja wśród dostawców usług opartych na zasadach otwartości w nauce może prowadzić do działań sabotujących i prób podszywania się pod tożsamość innych osób lub zaprzeczania autorstwa danej pracy. Problem ten został już rozwiązany w wielu innych dziedzinach życia poprzez stosowanie mechanizmów podpisu elektronicznego, które zapewniają przy okazji także integralność dokumentów. Prawdopodobnie jest tylko kwestią czasu, kiedy podobne mechanizmy zostaną wprowadzone w obszarze działań otwartej nauki.

Metadane

Metadane zbiorów badawczych to atrybuty opisujące ich zawartość, pochodzenie, stosowane metody badawcze i inne. Mimo że zwykle powyższe aspekty zostały opisane w danej publikacji naukowej lub też umieszczone wewnątrz samego zbioru badawczego (patrz **Wiarygodność**), to opisanie ich dodatkowo za pomocą ustrukturyzowanych metadanych daje dodatkowe korzyści w postaci lepszego indeksowania przez wyszukiwarki internetowe bądź mechanizmy wyszukująco/podpowiadające konkretnych platform. Warto zauważyć, że oprócz możliwości opisywania zbiorów metadanymi często istnieje również możliwość zamieszczenia opisów związanych z identyfikatorem konkretnego zbioru. Możliwość taką oferują oczywiście bazy danych systemu DOI, gdzie dodatkowo zamieścić możemy adresy URL, pod którymi znaleźć można publikacje danych.

Oprócz wyczerpującego opisania zbiorów metadanymi warto nie zapominać o przemyślanej strukturze katalogów, jeśli publikowane dane składają się z wielu plików oraz o nadawaniu wymownych nazw konkretnym plikom, co będzie ułatwiać identyfikację ich zawartości innym badaczom w trakcie prowadzenia analiz.

Formaty danych

W zależności od opisywanych informacji dane badawcze gromadzone są w plikach o różnych formatach danych. Formaty te są często narzucone przez wybrane narzędzia wchodzące w skład infrastruktury badawczej, ale często też są skutkiem wyboru jednego z wielu dostępnych, z którymi dane narzędzie współpracuje. Aby nie ograniczać możliwości innym potencjalnym użytkownikom, dane powinny być zapisane w standardach umożliwiających jak największą interoperacyjność. Wybór formatu często nie jest kwestią oczywistą i powinien on uwzględniać następujące czynniki:

- otwartość formatu (jeśli to możliwe to należy unikać publikowania danych w formatach zamkniętych, których użycie zmuszałoby innych do zakupu licencji na oprogramowanie),
- wierność odwzorowania danych (prowadząc badania przy pomocy platformy obsługującej specyficzny format pliku, należy rozważyć, na ile przekodowanie danych badawczych do potencjalnie lepszego formatu może przyczynić się do jakościowej /dokładność/ bądź ilościowej /utrata opisu pewnych atrybutów/ degradacji danych),
- cel badań (mimo że poszczególne profile badań mogą operować na podobnych klasach danych, to wybór konkretnego specyficznego dla nich formatu może mieć jednak istotne znaczenie – trywialnym przykładem może być przetwarzanie danych tekstowych w ramach dziedzin rozpoznawania tekstu (ang. OCR) oraz przetwarzania języka naturalnego (ang. NLP),
- popularność formatu (jeśli inne kryteria nie wskazują inaczej, to należy sugerować się popularnością danego formatu plików, co może przyczynić się do zwiększenia liczby potencjalnych odbiorców danych badawczych).

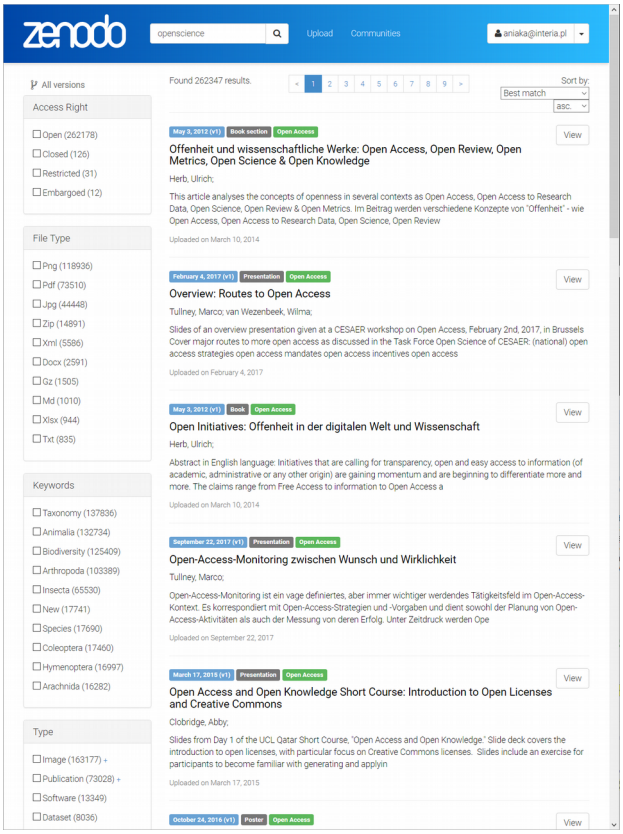
Dziedzina badawcza

Przed powstaniem dedykowanych i uniwersalnych platform udostępniania danych badawczych niektóre z ośrodków naukowych, odczuwając potrzebę ich składowania i udostęp-

niania, budowały rozwiązania własne, które w sposób naturalny gromadziły dane opisujące specjalizacje tych ośrodków. W ten sposób powstawały dziedzinowe platformy skupiające badaczy konkretnych gałęzi nauki. W chwili obecnej każdy badacz przed podjęciem decyzji o wyborze konkretnej platformy powinien samodzielnie odpowiedzieć sobie na pytanie, czy ważniejsze jest dla niego publikowanie danych na platformie dziedzinowej, czy też więcej korzyści odniesie z korzystania z platform niesprofilowanych, które często oferują bogatsze możliwości funkcjonalne, a i tak również umożliwiają budowanie społeczności związanych z dziedzinami zadeklarowanymi przez samych ich użytkowników.

Platforma Zenodo

Jako przykład repozytorium „ogólnego przeznaczenia" do publikowania danych badawczych przedstawiona zostanie platforma Zenodo[3]. Powstała ona w roku 2013 wspólnym wysiłkiem organizacji CERN oraz konsorcjum OpenAIRE i pozwala naukowcom na publikowanie danych badawczych o maksymalnej objętości 50GB. Jej użytkowanie jest proste i intuicyjne, a podstawowe funkcjonalności skupiają się wokół wyszukiwania zbiorów danych według zadanej frazy wyszukiwania, słów kluczowych, formatów plików, zasad udostępniania oraz typu danych i zostały zaprezentowane na ilustracji 1.



II. 1. Przykładowe okno panelu wyszukiwania danych platformy Zenodo
Źródło Zenodo. [Dostęp 2.10.2017]. Dostępne w: https://zenodo.org/search?page=1&size=20&q=openscience.

[3] *Zenodo* [online]. [Dostęp 2.10.2017]. Dostępny w: https://zenodo.org/.

Opis sposobu i procesu publikacji dotyczył będzie danych badawczych[4] wypracowanych w artykule[5] autorki opublikowanym w poprzednim numerze „Biuletynu EBIB", który dotyczył zagadnień ewaluacji nauki. Artykuł przedstawiał sposób ekstrakcji danych dotyczących artykułów publikowanych przez wydawnictwo PLOS ONE, a pozyskanych z otwartego indeksu cytowań OpenCitations, a następnie analizy tych danych pod kątem ilościowym w zakresie cytowalności. Dane pozyskane zostały z grafowej bazy danych, której zawartość publikowana jest w miesięcznych odstępach w repozytorium Figshare w miejscach wskazanych pod adresem OpenCitations: http://opencitations.net/download[6].

Mimo zamieszczonego w artykule szczegółowego opisu pozyskania tych danych ich duży wolumen oraz zastosowane technologie mogą stać na przeszkodzie niektórym zainteresowanym w samodzielnym pozyskaniu tych danych do analiz własnych. Autorka postanowiła więc opublikować dane wyekstrahowane z bazy danych i opublikować je jako dane badawcze w formie gotowej do użycia przez większość narzędzi analitycznych używanych w przedmiotowej dziedzinie. Mimo że sama operowała jedynie na plikach formatu CSV, postanowiła dodatkowo dokonać ich konwersji na natywny format (gexf) wykorzystywany do analiz narzędzia, tak aby inni w jednym prostym kroku mogli nimi zasilić tę platformę. Dodatkowo te same dane udostępniła również w formacie (net) stworzonym dla jednego z najczęściej wykorzystywanych programów do analizy sieci cytowań[7]. Warto jednak zaznaczyć, że format ten ma ograniczone możliwości, przez co utracone zostały informacje o latach publikacji poszczególnych artykułów.

Opracowane w trakcie badań wizualizacje zostały również udostępnione jako dane badawcze w rozdzielczościach umożliwiających wielkoformatowy wydruk lub dalsze szczegółowe analizy w przeglądarkach obrazów z wykorzystaniem funkcji przybliżania/powiększania interesujących w danym momencie regionów. Jako format obrazów wykorzystano PNG (ang. Portable Network Graphics), gdyż w takim właśnie formacie obrazy generuje wykorzystane narzędzie, a przy okazji jest on formatem wystarczająco popularnym i otwartym. Trzeba zauważyć, że konwersje obrazów (grafiki rastrowej) zarówno w zakresie rozdzielczości, jak i głębi kolorów czy zmiany formatów zwykle powodują obiektywny spadek jakości, więc powinno się je stosować z pewną dawką ostrożności, a kiedy tylko jest to możliwe generować obrazy o z góry potrzebnych parametrach, tak aby nie trzeba już ich było później dodatkowo opracowywać.

Przyjęta konwencja nazw plików z danymi źródłowymi pozwala po przeczytaniu artykułu źródłowego (zarówno odnośnik, jak i jego DOI zostały wskazane w opublikowanych danych) bez czytania dalszych opisów domyślić się, jakie dane w jakich formatach zostały w nich zawarte. Nazwy plików z obrazami zawierają w sobie natomiast numerację ilustracji zaczerpniętą bezpośrednio z artykułu. Wyjątek stanowi jeden dodatkowy obraz, który nie

[4] KAMIŃSKA, A.M. PLOS ONE – a case study of citation analysis of research papers based on the data in an open citation index (The OpenCitations Corpus). *Zenodo* [online]. [Dostęp 13.10.2017]. Dostępny w DOI: 10.5281/zenodo.1010450. DOI: 10.5281/zenodo.1010450.

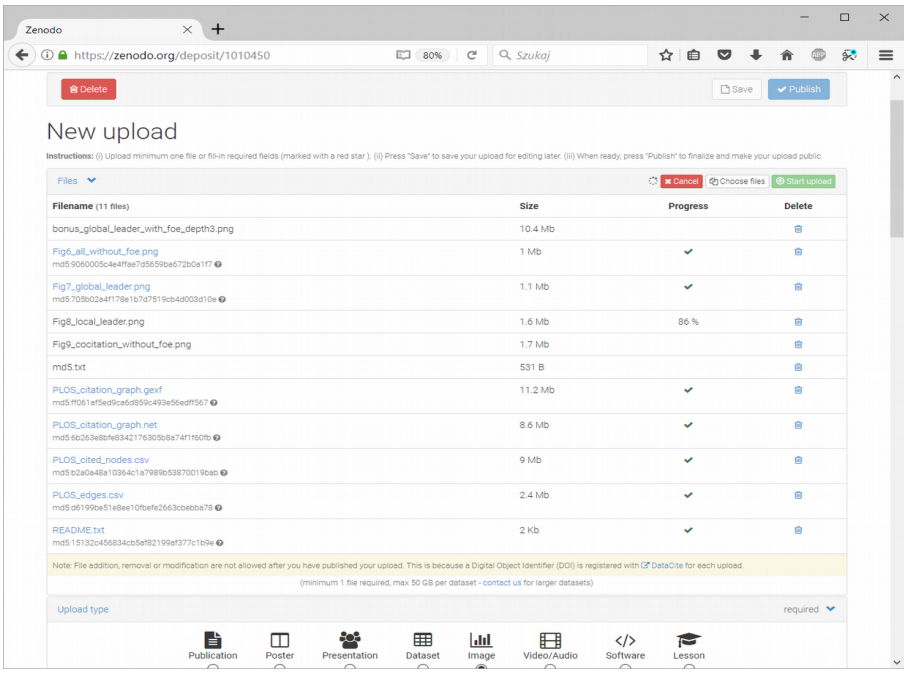
[5] KAMIŃSKA, A.M. PLOS ONE – studium przypadku analizy cytowań prac naukowych na podstawie danych otwartego indeksu cytowań (OpenCitations Corpus). *Biuletyn EBIB* [online]. 2017, no 176. [Dostęp 12.12.2017]. Dostępny w: http://open.ebib.pl/ojs/index.php/ebib/article/view/564.

[6] Wszystkie odesłania do stron internetowych przedstawiają wersję aktualną w dn. 20.10.2017.

[7] Najnowszą wersję programu wraz z listą opisów zmian zrealizowanych w ramach kolejnych wersji znaleźć można na stronie Pajek. *Pajek: analysis and visualization of large networks* [online]. [Dostęp 2.10.2017]. Dostępny w: http://mrvar.fdv.uni-lj.si/pajek/.

został opublikowany w źródłowym artykule, ale stanowił jeden z rezultatów przedstawionych tam badań. Dodano także plik md5.txt z sumami kontrolnymi wyliczonymi dla wszystkich plików danych badawczych oraz plik README.txt zawierający zwięzły anglojęzyczny opis zawartości plików pozwalający zrozumieć ich przeznaczenie nawet osobom, które nie przeczytały artykułu opisowego.

Repozytorium Zenodo umożliwia publikowanie zbioru plików pod jednym identyfikatorem DOI. Poniżej na ilustracji 2 przedstawiono przykładowe okno do przesyłania plików w ramach wybranej publikacji.



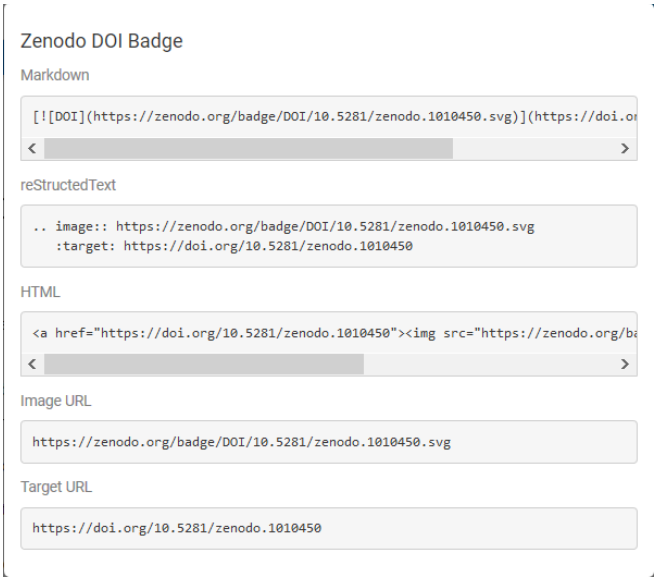
II. 2. Przykładowe okno do przesyłania plików w ramach jednej publikacji Źródło Zenodo. [Dostęp 2.10.2017]. Dostępne w: https://zenodo.org.

Mamy tutaj możliwość wskazania (przycisk *Choose files*) wielu plików składowanych w zasobach lokalnych komputera, które po naciśnięciu przycisku *Start upload* zostaną kolejno przesłane do repozytorium. Dla każdego przesłanego pliku wyliczona i zaprezentowana zostanie suma kontrolna, którą należy porównać z wyliczoną wcześniej w oparciu o plik składowany lokalnie, aby uzyskać pewność, że na skutek błędów transmisji nie nastąpiło żadne przekłamanie danych. Mimo że platforma wyświetla sumy kontrolne, jednym z publikowanych plików jest w tym przypadku również md5.txt zawierający sumy kontrolne wszystkich pozostałych. Może to być szczególnie przydatne wówczas, gdy dane badawcze zostały pobrane z repozytorium jeden raz, a później są kolejno kopiowane pomiędzy różnymi lokalizacjami środowiska badawczego bez dostępu do internetu. Wtedy na każdym etapie korzystania z danych badawczych istnieje możliwość weryfikacji ich integralności, bez potrzeby odwoływania się do informacji udostępnianych z poziomu repozytorium.

Dla całego zbioru publikowanych plików ustalić należy najbardziej odpowiedni typ zbioru (artykuł, plakat, prezentacja, zbiór danych, obraz, audio/video, oprogramowanie, lekcja),

tytuł, autorów, język, słowa kluczowe, licencję udostępniania i wiele innych opcjonalnych metadanych. Warto zauważyć, że w przyjętym modelu w razie potrzeby publikowania plików w ramach różnych licencji należy je opublikować jako oddzielne zbiory badawcze. Mimo że w podanym przykładzie publikuje się pliki różnych formatów, przyjęto, że dominującym jest „obraz", jako że przedstawiane wizualizacje stanowią końcowe etapy badania, zaś dane źródłowe umieszczono głównie celem uwiarygodnienia wyników wizualizacji. Po przesłaniu wszystkich plików i opisaniu publikowanego zbioru metadanymi możemy go zapisać (przycisk *Save*) w celu dokonywania późniejszych poprawek czy zmian plików składających się na publikację, ale nie będzie on jeszcze widoczny dla innych. Chcąc upublicznić zbiór danych, należy skorzystać z opcji *Publish*, która spowoduje nadanie identyfikatora DOI[8] i zamknie drogę do dalszych poprawek – od tego momentu jakiekolwiek zmiany w plikach składających się na publikację możliwe będą jedynie z wykorzystaniem mechanizmu wersjonowania publikacji, który na platformie Zenodo udostępniony został zaledwie kilka miesięcy temu[9].

Dane badawcze opublikowane w repozytorium mogą być wyszukiwane wprost z poziomu platformy Zenodo, ale są również dobrze indeksowane przez wyszukiwarkę Google. Dodatkowo opisy każdego opublikowanego zbioru przesyłane są do europejskiego systemu indeksującego OpenAIRE. Ciekawą opcją jest również możliwość wyświetlenia okna z gotowymi do użycia „plakietkami" w różnych formatach umożliwiającymi zamieszczenie odnośnika do publikacji konkretnego zbioru w różnych zasobach elektronicznych, w tym naukowych platformach społecznościowych. Mechanizm ten został zaprezentowany na ilustracji 3.



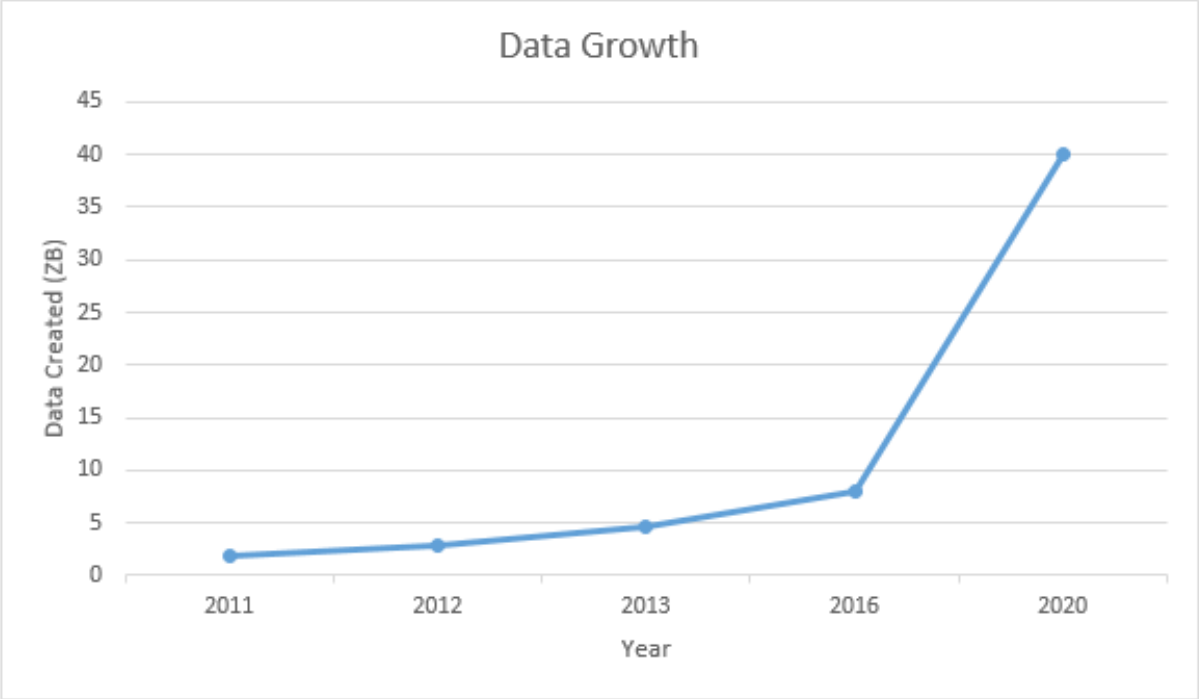
II. 3. Przykładowe okno platformy Zenodo z „plakietkami" opisującymi wybrany zbiór w różnych formatach w formie gotowej do użycia za pomocą „kopiuj-wklej"
Źródło Zenodo. [Dostęp 2.10.2017]. Dostępne w: https://zenodo.org.

[8] Warto zwrócić uwagę, że raz nadany identyfikator nie może zostać użyty ponownie w przypadku rezygnacji z publikacji, dlatego dla celów testowania funkcjonalności repozytorium Zenodo jego twórcy przygotowali równoległą wersję testową dostępną pod adresem: https://sandbox.zenodo.org.

[9] Więcej informacji o proponowanej przez twórców platformy Zenodo koncepcji wersjonowania danych z wykorzystaniem systemu identyfikacji DOI znaleźć można pod adresem: http://help.zenodo.org/#versioning.

Podsumowanie

Będąc świadkami odbywającej się właśnie na naszych oczach rewolucji informacyjnej, jak prognozuje amerykańskie przedsiębiorstwo analityczo-doradcze Gartner (il. 4), w nieodległej przyszłości będziemy musieli zmierzyć się z wyzwaniem gromadzenia i przetwarzania informacji osiągających wolumen 40ZB (czyli wielkości rzędu 1000^7 bajtów).



Il. 4. Prognozowany przyrost światowego wolumenu danych
Źródło Infragistics. [Dostęp 2.10.2017]. Dostępne w: https://www.infragistics.com/community/blogs/mobileman/archive/2015/12/15/big-data-and-the-internet-of-things.aspx.

Gromadzenie informacji rzadko bywa celem samym w sobie, a najczęściej jest jedynie warunkiem koniecznym do dalszego ich przetwarzania i przeszukiwania. Przyglądając się zawartości repozytorium Zenodo zarówno pod kątem liczby publikacji, jak i tematyki, której one dotyczą, można z całą pewnością stwierdzić, że zapowiadane jeszcze kilka lat temu przez krajowych badaczy[10] kierunki rozwoju nauki stają się faktem nie tylko w wymiarze legislacyjnym, ale również w wymiarze praktycznym. Dlatego biorąc pod uwagę tempo wzrostu gromadzonych danych, a dotyczy ono również danych badawczych, należy mieć na uwadze właściwy sposób ich publikowania, który ułatwi zapanowanie nad chaosem informacyjnym i nie pozostawi wątpliwości innym badaczom z nich korzystających co do ich wiarygodności, sposobu pozyskania, spójności i innych wymiarów przedstawionych w niniejszym opracowaniu.

[10] BEDNAREK-MICHALSKA, B. Komisja Europejska idzie do przodu. Otwartość nauki staje się w Europie faktem. *Biuletyn EBIB* [online]. 2014, no 147. [Dostęp 2.10.2017]. ISSN 1507-7187. Dostępny w: http://open.ebib.pl/ojs/index.php/ebib/article/viewFile/148/356.

Bibliografia:

1. BEDNAREK-MICHALSKA, B. Komisja Europejska idzie do przodu. Otwartość nauki staje się w Europie faktem. *Biuletyn EBIB* [online]. 2014, no 147. [Dostęp 2.10.2017]. ISSN 1507-7187. Dostępny w: http://open.ebib.pl/ojs/index.php/ebib/article/viewFile/148/356.
2. *DMPTool,* University of California Curation Center of the California Digital Library 2017. [online]. [Dostęp 2.10.2017]. Dostępny w: https://dmptool.org.
3. KAMIŃSKA, A.M. PLOS ONE – a case study of citation analysis of research papers based on the data in an open citation index (The OpenCitations Corpus). *Zenodo* [online]. [Dostęp 13.10.2017]. Dostępny w DOI: 10.5281/zenodo.1010450. DOI: 10.5281/zenodo.1010450.
4. KAMIŃSKA, A.M. PLOS ONE – studium przypadku analizy cytowań prac naukowych na podstawie danych otwartego indeksu cytowań (OpenCitations Corpus). *Biuletyn EBIB* [online]. 2017, no 176. Dostępny w: http://open.ebib.pl/ojs/index.php/ebib/article/view/564. ISSN 1507-7187.
5. *Pajek: analysis and visualization of large networks* [online]. [Dostęp 2.10.2017]. Dostępny w: http://mrvar.fdv.uni-lj.si/pajek/.
6. *Research Data* [online]. Elsevier 2017. [Dostęp 2.10.2017]. Dostępny w: https://www.elsevier.com/about/our-business/policies/research-data.
7. *Zenodo* [online]. [Dostęp 02.10.2017]. Dostępny w: https://zenodo.org/.