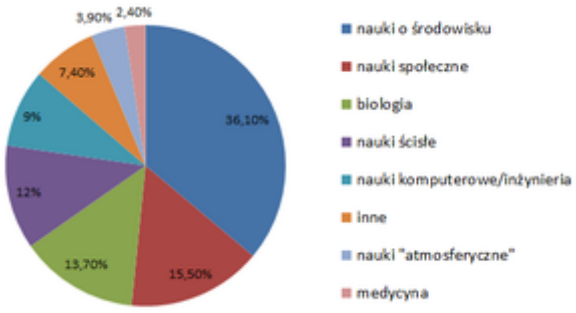
Daleka droga do Open Data

Przy okazji Międzynarodowego Tygodnia Wolnego Dostępu dużo mówiło się o tym, jak zmienia się naukowy przemysł wydawniczy; o tym, że instytucje przyznające fundusze na badania coraz częściej wymagają od naukowców publikacji wyników w trybie open access; o tym, jaki nacisk na wydawców wywierają autorzy, czytelnicy i biblioteki. Mniej jednak mówi się o open data – o potrzebie udostępniania publicznego nie tylko przetrawionych do formy publikacji wyników, ale i samych „surowych" rezultatów.



Pracując w niektórych dziedzinach nauki można odnieść wrażenie, że open data to jest coś, co już jest, co ma miejsce, nad czym nie trzeba dyskutować. Myśleć tak mogą genetycy przyzwyczajeni do korzystania z zasobów NCBI, z źródeł danych z takich projektów jak Projekt Sekwencjonowania Ludzkiego Genomu, czy ENCODE. Myśleć tak mogą biolodzy korzystający z banku danych białkowych, PDB, albo z danych na temat deforestacji puszczy amazońskiej. O tym, jak pomocne może być open data – zwłaszcza w połączeniu z crowsourcingiem badań – wiedzą astronomowie stojący za projektem Galaxy Zoo. Zaś wyniki badań w CERNie stają się także powoli otwarte – chociaż na razie tylko dla środowiska naukowego, a nie dla wszystkich chętnych.

Jakie korzyści płyną z dzielenia się danymi, z publicznego i dowolnego do nich dostępu? Po pierwsze, ponowna analiza naszych danych przez kogoś z zewnątrz może pomóc w weryfikacji naszych wniosków. Nie jest bowiem tak, że wystarczy zrobić eksperyment, a z surowych, nieobrobionych wyników wyskoczy na nas objawienie. O nie, nie. Dane trzeba najpierw przeanalizować, a błędów przy analizie można popełnić tyle samo, o ile nie więcej, jak przy samym doświadczeniu. Niezależny głos potwierdzający naszą analizę jest więc zawsze miłym dla ucha potwierdzeniem, że rzeczywiście mieliśmy rację. Po drugie, osoby z zewnątrz patrząc na nasze dane mogą w nich dojrzeć coś, co nam samym umknęło, zaproponować jakiś rodzaj analizy, o którym nie pomyśleliśmy, albo nawet alternatywną interpretację naszym wyników.

Po trzecie, żeby dzielić się danymi, muszą one być zdigitalizowane i przechowywane w sposób, który dostęp do nich ułatwi. A gdy są już w takiej formie, to duża szansa, że rozpełzną się szybko po świecie, co tylko pomoże w ich utrwaleniu i zachowaniu – bo wbrew pozorom największym dziedzictwem nauki nie jest stos publikacji, ale właśnie dane. Po czwarte, dostępność danych na jakiś temat pomaga w optymalizacji środków – bowiem fakt, że można sobie wynik jakiegoś doświadczenia po prostu ściągnąć, oznacza, że nie musimy go sami przeprowadzać (chyba że bardzo chcemy oczywiście), oszczędzając czas i pieniądze.

I po dwa ostatnie: dzielenie się danymi jest swego rodzaju bezpiecznikiem dla naukowych oszustw, gdyż dostępność surowych danych oznacza, że prędzej czy później ktoś się nimi może zainteresować i je przeanalizować i jeśli jest z nimi coś nie tak – prawdopodobnie roztrąbić to na cały świat. Ponadto zaś open data jest niesamowitym źródłem materiałów dydaktycznych: zarówno tylko jako dane, które można analizować potem na dziesiątą stronę, jak i jako inspiracja i tzw. benchmark dla prostych doświadczeń wykorzystywanych do szkolenia nowych pokoleń badaczy.

Open data to jednak dla większości badaczy na razie tylko mit. Bo też, jeśli się dobrze zastanowić, dzielenie się swoimi danymi dla wielu osób nie ma po prostu sensu. Powiedzmy sobie wprost: dopóki badacze na uczelniach rozliczani są z publikacji, nie będzie miało znaczenia, jak niesamowite robią badania, jeśli nie będę publikować wyników. Co więcej, jak długo do dorobku bardziej będzie się liczyła publikacja w Nature a nie w specjalistycznych periodykach, tak długo wielu naukowców będzie te swoje oryginalne trzymać w ukryciu i je kisić, aż im się uzbiera dość na bang warty podboju tegoż Nature (czy jakiegokolwiek innego pisma o wysokim profilu). Innymi słowy,

naukowcom po prostu często brak odpowiedniej motywacji, żeby się swoją krwawicą dzielić.

Prawda jest zresztą taka, że w wielu dziedzinach, które obecnie można podawać jako piękny przykład wdrożenia open data, dzielenie się swoimi danymi nie oznacza jakiejś nieokreślonej moralnej wyższości badaczy w tej dziedzinie pracujących. Z prostego powodu: gdyż zazwyczaj jest wynikiem nacisku instytucji, z których płyną pieniądze. Jak inaczej w końcu przekonać wszystkich badaczy z danej dziedziny, że warto to robić, jak inaczej zmusić ich do robienia czegoś, co jest wbrew ich jestestwu?

W zeszłym roku grupa amerykańskich badaczy opublikowała w – a jakże – otwartodostępowym piśmie PLoS ONE wyniki swoich badań nad tym, jak naukowcy dzielą się swoimi danymi i co stoi na przeszkodzie takiemu dzieleniu się. Przepytano ponad 1300 naukowców z różnych dziedzin – pomiędzy dyscypliny rozkłada się to mniej więcej tak:

80% respondentów było czynnymi akademikami, około 1/8 było zatrudnione na etatach rządowych, po ok. 2.5% pracowało w przemyśle lub instytucjach non-profit. Blisko połowa badanych pracuje na stanowisku lub posiada tytuł profesora.

Zapytani, z jakich źródeł danych korzystają, prawie 40% respondentów odpowiedziała, że korzystają z repozytoriów instytucjonalnych, zaś 27% z „innych źródeł". Warte uwagi jest tutaj to, że niemal wszystkie pozostałe wymienione źródła danych to różnego rodzaju repozytoria gromadzące dane dotyczące ekologii, bioróżnorodności, środowiska. Badacze odpowiadali też na długą listę pytań typu „zgadzam się, trochę się zgadzam, ani się zgadzam ani nie zgadzam…", dotyczących tego, czy projekt badawczy lub organizacja, w której pracują, posiada protokoły pozwalające zarządzać i przechowywać dane eksperymentalne, poziomu satysfakcji z tego, jak tego rodzaju protokoły funkcjonują na każdym etapie badań, czy dane są dostępne dla innych badaczy, jak brak dostępu do danych innych badaczy wpływa na badania, jeśli mamy dostęp do danych, jakiej jakości są to dane itd. Rzućmy okiem na kilka ciekawych wyników.

Aby dzielić się danymi efektywnie, nie wystarczy wrzucić je do publicznego repozytorium – muszą one dodatkowo być dobrze opisane: z jakiego eksperymentu pochodzą, jakie były warunki doświadczalne, kto badanie przeprowadził i według której wersji protokołu. I tak dalej, i tym podobne. Innymi słowy do zestawu danych muszą być dodane tzw. metadane (czyli dane o danych). Na pytanie jednak, jakiego rodzaju standard metadanych jest stosowany w grupach badawczych respondentów, odpowiedź jest druzgocąca:

W prawie połowie laboratoriów nie stosuje się bowiem żadnych metadanych. Co oznacza nie tylko, że korzystać z nich nie będą mogli badacze z innych instytucji, ale także, że jest spora szansa, że za dziesięć lat, gdy studia pokończą obecni doktoranci, a postdocy ruszą na podbój innych instytucji, w labie nie będzie nikogo, kto byłby w stanie coś z nich zrozumieć…

Bardzo ciekawie wyglądają odpowiedzi respondentów na pytania o to, czy dzielą się danymi z innymi badaczami, a także o to, czy inni badacze mają łatwy do nich dostęp:

Do dzielenia się danymi przyznaje się ponad 70% osób biorących udział w badaniu, jednocześnie jednak zaledwie jedna trzecia przyznaje, że dostęp jest dla innych badaczy łatwy – czyli że są one w jakimś repozytorium i posiadają zrozumiałe metadane. Co to oznacza? Uwzględniając tę sporą chęć do dzielenia się wynikami, najprawdopodobniej winę można zrzucić właśnie na brak odpowiednich repozytoriów i standardów – co powoduje, że dostępność danych jest tak znacznie

niższa od chęci ich udostępnienia.

Żeby jednak nie winić tylko badaczy i okropnych instytucji, które uniemożliwiają dzielenie się danymi poprzez niezapewnienie odpowiednich do tego środków, spójrzmy na to, jakich odpowiedzi udzielano na pytanie o powody, dla których dane nie były udostępniane elektronicznie:

Z badań wyraźnie wybija się, że najczęściej powodem nie udostępniania danych jest brak czasu na dokonanie tego (tu kłania się np. brak standardów meta – bo stworzenie metadanych od podstaw po to, żeby móc dane udostępnić, jest bardzo, ale to bardzo czasożerne), oraz brak pieniędzy. Pytanie oczywiście: pieniędzy na co dokładnie – bo jeśli tylko na serwer, to tutaj w niektórych dziedzinach z pomocą przychodzą publiczne repozytoria. W innych jednak nie jest już tak łatwo. Dwie podane przyczyny, które powinny też przykuć uwagę, to brak uprawnień, aby dane uczynić publicznymi oraz to, że dane nie powinny być udostępniane publicznie. W tym pierwszym przypadku może być mowa o tym, że instytucje badawcze często przywłaszczają sobie prawa autorskie do wyników badań – nie bez powodu, ale też i często ten pęd za własnością intelektualną rozciągany bywa do granic absurdu. Dlaczego dane nie powinny być w ogóle udostępniane? Coraz częściej mówi się o tym, że prawie wszystko powinno być – włączając w to nawet do tej pory skrywane przez korporacje farmaceutyczne dane dotyczące prób klinicznych. Nawet bezpieczeństwo narodowe przestaje być dobrym argumentem.

Wszystkich wyników prezentował i omawiał tutaj nie będę – dość powiedzieć, że publikacja ma jakieś 30 tabel podsumowujących rezultaty i warto sobie na niektóre z nich zerknąć – różne ciekawe trendy wyłażą na przykład, gdy się odpowiedzi respondentów poukłada według ich wieku oraz dyscypliny, którą się zajmują.

Tutaj dość jednak powiedzieć, że ogólne wnioski są następujące: badacze chcą się swoimi danymi dzielić, nawet pomimo tego, o czym pisałem na początku tego wpisu – że dane to wielki skarb naukowca, którego należy strzec. Najczęstszymi przyczynami nie udostępniania danych nie są zatem niechęć czy rywalizacja, ale raczej bardzo przyziemne powody takie jak brak możliwości technicznej, brak czasu, czy też wreszcie odwieczna bolączka nauki – brak pieniędzy. Badacze często podkreślają, że niemożność dostępu do danych innych naukowców wpływa (negatywnie) na ich zdolność dokonania prawidłowej analizy własnych wyników.

Ważnym spostrzeżeniem tutaj, które każda osoba prowadząca jakiekolwiek badania naukowe powinna sobie wyryć złotymi zgłoskami na tabliczce nad biurkiem i spoglądać na nią co najmniej raz dziennie, jest to, że wyniki eksperymentów muszą posiadać metadane. Jest wiele standardów, które można do tego wykorzystać, warto więc spędzić chwilę na zastanowienie, który jest najbardziej odpowiedni dla danego typu badań, dla danej dziedziny, i zacząć go wdrażać u siebie jak najszybciej.

Autor: Rafal Marszalek
Źródło: http://nicprostszego.pl/, Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions PLoS ONE, 6 (6) DOI: 10.1371/journal.pone.0021101

p.s. Podejrzewam, że większość z Was nie potrzebuje wyjaśnienia, czym jest open access. Niemniej jednak szkoda by było zignorować tak niesamowity wysiłek Jorge Chama:

http://www.youtube.com/watch?feature=player_embedded&v=L5rVH1KGBCY

Informacje dnia: Mity na temat epilepsji Marzec był drugim najcieplejszym miesiącem w Europie Sporadyczne picie dużych ilości alkoholu W nagłych przypadkach ChatGPT Health często uspokaja Dieta bogata w warzywa i owoce zmniejsza ryzyko demencji nawet u seniorów Nie kompromitujcie nas, czyli jak chronić dane biometryczne Mity na temat epilepsji Marzec był drugim najcieplejszym miesiącem w Europie Sporadyczne picie dużych ilości alkoholu W nagłych przypadkach ChatGPT Health często uspokaja Dieta bogata w warzywa i owoce zmniejsza ryzyko demencji nawet u seniorów Nie kompromitujcie nas, czyli jak chronić dane biometryczne Mity na temat epilepsji Marzec był drugim najcieplejszym miesiącem w Europie Sporadyczne picie dużych ilości alkoholu W nagłych przypadkach ChatGPT Health często uspokaja Dieta bogata w warzywa i owoce zmniejsza ryzyko demencji nawet u seniorów Nie kompromitujcie nas, czyli jak chronić dane biometryczne

Partnerzy