Czy model „danych otwartych od narodzin” powinien stać się standardem udostępniania danych badawczych?

Bartosz Maćkiewicz

Instytu Filozofii, Uniwersytet Warszawski

14 grudnia 2018

Tło: kryzys replikacyjny w psychologii

Plan

1. Metodologiczne i praktyczne zalety otwartości danych
2. Modele udostępniania danych i ich problemy
3. Praktyczne strategie radzenia sobie z problemami otwierania danych

Otwarta nauka i otwarte dane

Otwarte dane są jednym z fundamentów otwartej nauki i mogą pomóc wyeliminować niedostatki metodologiczne na wielu etapach pracy badawczej (Spellman, Gilbert i Corker 2017).

Etap procesu badawczego	Problem
Zbieranie danych i raportowanie metod	▶ braki w raportowaniu wszystkich zbieranych danych
Analiza danych i raportowanie wyników	▶ HARKing ▶ p-hacking ▶ „elastyczny” proces czyszczenia danych i doboru narzędzi analitycznych
Przechowywanie i archiwizacja	▶ problem „szuflady” ▶ utrata informacji powoduje osłabienie rzetelności metaanaliz

Modele udostępniania danych

Udostępnianie danych na żądanie

- Najpopularniejszy model
- Dane udostępniane są zainteresowanym podmiotom po kontekcie z autorem i nieformalnej prośbie o udostępnienie danych

Udostępnianie danych wraz z publikacją

- Coraz popularniejszy model, obecnie standard dobrej praktyki naukowej
- Dane dołączane są do manuskryptu na wstępnym etapie procesu recenzyjnego
- Po ukazaniu się publikacji dane udostępnione są wszystkim podmiotom mającym do niej dostęp

Udostępnianie danych po okresie embarga

- Sposób radzenia sobie z efektem „szuflady”
- Badacz zawczasu decyduje się na okres embarga, po którym wszystkie zebrane dane będą zdeponowane w repozytorium
- Okres embarga ma umożliwić „wyciśnięcie” z danych maksymalnej liczby publikacji autorom danych

Udostępnianie danych w praktyce

- Model „na żądanie” nie zdaje egzaminu, ponieważ:
  - dane giną bezpowrotnie lub są źle zarchiwizowane;
  - przygotowanie danych wymaga pracy (anonimizacja, opisanie danych, itp.), a często te roboczogodziny nie są uwzględnione w projekcie.

- Model „wraz z publikacją” nie zdaje egzaminu, ponieważ:
  - generuje „efekt szuflady”;
  - pozwala wyselekcjonować tylko te dane, które wspierają tezy stawiane w publikacji.

- Model „po okresie embarga” nie zdaje egzaminu, ponieważ:
  - kiedy upływa okres embarga, to projekt badawczy jest już dawno skończony - osoby odpowiedzialne za przechowywanie i archiwizacje danych mogą nie móc znaleźć danych i dokumnetacji, pracować na innym uniwersytecie, umrzeć.

73%

Jelte M Wicherts i in. “The poor availability of psychological research data for reanalysis.”. W: *American Psychologist* 61.7 (2006), s. 726

Praktyczne strategie radzenia sobie z problemami otwierania danych

Born-open data (Rouder 2016)

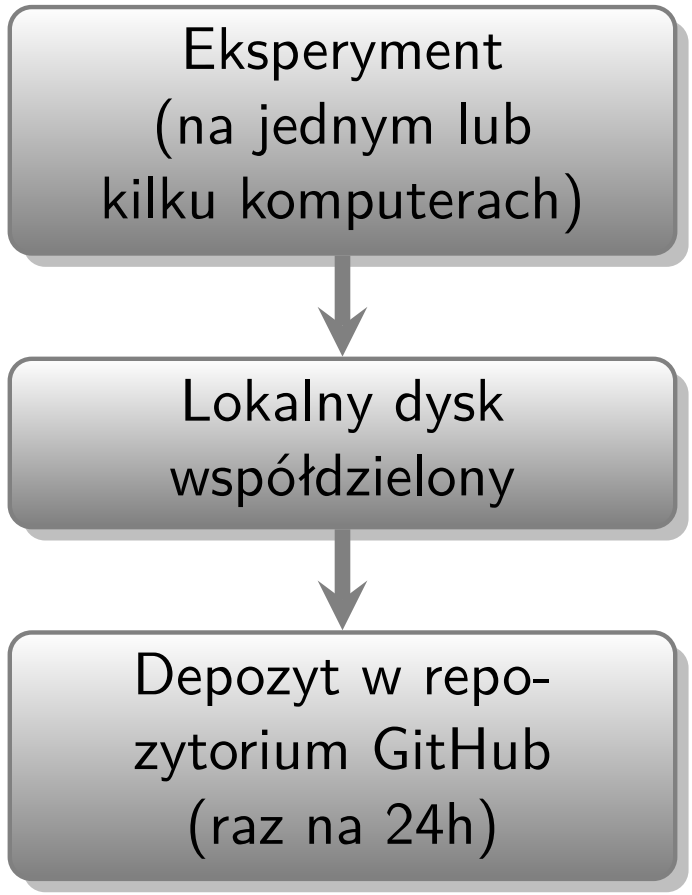
Radykalne rozwiązanie problemu udostępniania danych

Problem: otwieranie danych jest trudnym, czasochłonnym i wymagającym zadaniem, które zawsze jest ostatnie na liście priorytetów.

Rozwiązanie: stworzenie automatycznego systemu, który archiwizuje, dokumentuje i udostępnia dane.

Born-open data (Rouder 2016)

Schemat działania



Born-open data (Rouder 2016)

Potencjalne problemy

- **Prywatność** uczestników badania i ich **zgoda** na udostępnianie danych
- **Niechęć** do udostępniania czegoś, co należy do nas
- **Strach** przed **„podkradaniem”**
- Podatność na **„zawodowe zranienie”**

Gwarancja integralności danych? (Bierman i Jolij 2018)

Problem: jednym z najważniejszych przyczyn Wątpliwych Praktyk Badawczych jest *selekcja danych* tak, aby wspierały hipotezę badawczą. Z tego względu **wszystkie dane w swojej oryginalnej formie** powinny być dostępne arbitrom w procesie recenzyjnym.

Rozwiązanie: zintegrowany z oprogramowaniem eksperymentalnym moduł wgrywający permanentnie wszystkie dane do bazy znajdującej się na niezależnym serwerze.

Gwarancja integralności danych? (Bierman i Jolij 2018)

1. Moduł przy każdym uruchomieniu programu sprawdza, czy badanie jest prerejestrowane (w przypadku brak prerejestracji oferuje odpowiedni formularz).
2. Podczas przebiegu eksperymentu moduł w czasie rzeczywistym wgrywa dane do bazy SQL (administrowanej przez Zaufaną Trzecią Stronę) zostawiając tym samym permanentny ślad działania programu. Serwer nie pozwala nikomu (w tym eksperymentatorowi) na edycję rekordów.
3. Formularz prerejestracji musi zostać wypełniony przed każdą formalną sesją. Wszystkie dane zebrane bez formularza prerejestracji oznaczone są w bazie jako eksploracyjne.

Gwarancja integralności danych? (Bierman i Jolij 2018)

- **Recenzent** ma możliwość przejrzenia **wszystkich danych zapisanych na serwerze**, dodatkowo udostępniane są mu odpowiednie skrypty pozwalające na podsumowanie zapisanych danych.
- Metoda ta pozwala na kontrolę liczby badanych rezygnujących z dalszego udziału w badaniu (*drop-outs*) i analizę tych przypadków.
- Testy autorów pokazały, że trudno taki moduł oszukać i jego użycie jest bezproblemowe.
- Możliwość wykorzystania technologii *blockchain*.

Desiderata

1. Standardowe modele udostępniania danych nie zdają egzaminu lub generują zbyt duże nakłady czasu i pracy.
2. Można wypracować praktyczne strategie radzenia sobie z problemami standardowych modeli.
3. Implementacja modelu *born-open data* w podstawowym kształcie nie wymaga specjalnej infrastruktury i dużych nakładów pracy i jest mniej problematyczna, niż by się zdawało.
   - W czasopismach metodologicznych z zakresu psychologii można znaleźć łagodne wprowadzenia do systemu kontroli wersji Git zaprojektowane dla nie-informatyków (np. Vuorre i Curley 2018).

Bibliografia

- Dick Bierman i Jacob Jolij. "Towards guaranteed data-integrity: A method of preventing Questionable Research Practices". 2018.
- Jeffrey N Rouder. "The what, why, and how of born-open data". W: *Behavior research methods* 48.3 (2016), s. 1062–1069.
- Bobbie Spellman, Elizabeth Gilbert i Katherine S Corker. "Open science: what, why, and how". 2017.
- Matti Vuorre i James P Curley. "Curating Research Assets: A Tutorial on the Git Version Control System". W: *Advances in Methods and Practices in Psychological Science* (2018), s. 219–236.
- Jelte M Wicherts i in. "The poor availability of psychological research data for reanalysis.". W: *American Psychologist* 61.7 (2006), s. 726.