**Zabezpiecz swoje dane badawcze**

**James Baker**

Przewodnik wskazuje, w jaki sposób historycy mogą dokumentować i strukturyzować swoje dane badawcze, aby zapewnić ichużyteczność w przyszłości.

**Redaktor:**

Adam Crymble

**Recenzenci:**

Jane Winters

Sharon Howard

William J. Turkel

**Tłumaczenie:**

Beata Starosta

Maja Starosta

#### opublikowano

2014-04-30

#### zmodyfikowano

2018-05-21

#### poziom trudności

Niski

**Spis treści**

* Wstęp
* Dokumentowanie danych badawczych
* Formaty plików
* Podsumowanie 1
* Strukturyzacja danych badawczych
* Podsumowanie 2
* Podsumowanie końcowe
* Bibliografia

**Wstęp**

Roy Rosenzweig, w swojej pracy "Niedobór czy Obfitość" (ang. “ScarcityorAbundance”) z 2003 roku, starał się ostrzec historyków przed tym, co nazwał "kruchością dowodu w erze cyfrowej" (Rosenzweig, 736). Pomimo że jego obawy dotyczyły głównie źródeł dostępnych w otwartej sieci, koncepcję tę można z powodzeniem rozszerzyć na materiały lub dane powstającew formie cyfrowej podczas badań prowadzonych przez historyków.

Dlaczego niniejszy przewodnik podejmuje tematzabezpieczenia cyfrowych danych badawczych?

Otóż, historycy skłaniają się ku wykorzystaniu komputera jako podstawowego narzędzia do przechowywania wszystkich danych i materiałów badawczych. Ich rękopisy już od jakiegoś czasu przybierają formę obiektów cyfrowych, a badania podążając w tym samym kierunku, rejestrowane są w formie sporządzanych komputerowo notatek, archiwów fotografii, czy też danych tabelarycznych. Ponadto dane badawcze przechowywane w wersji cyfrowej mają wyraźną przewagę nad swoimi fizycznymi poprzednikami: można je przeglądać, wyszukiwać i udostępniać z wielu miejsc, a także łączyć i porównywać z innymi danymi badawczymi.

Sam fakt przechowywania danych badawczych w formie cyfrowej nie gwarantuje ich przetrwania. Jednak nie mam tutaj na myśli przetrwania danych w sensie dosłownym, jak również tego, że będą one możliwe do odczytania w kolejnej wersji Microsoft Word, ale raczej że zachowają swoją użyteczność dla odbiorcy. Nawet, jeśli problem nie został jeszcze rozstrzygnięty, to potencjalne rozwiązania dotyczące praktycznych kwestii zabezpieczenia danych badawczych na przyszłość były już szeroko rozważane, z myślą o historykach i nie tylko. Również specjaliści, usługodawcy i inni zajmujący się zarządzaniem danymi dyskutowali na temat najlepszych naukowych praktyk dotyczących dokumentowania, strukturyzowania i organizowania danych badawczych. Tym niemniej, dane badawcze wytwarzane przez pojedynczego historyka narażone są na utratę użyteczności, jeżeli historyk nie będzie umiał wygenerować i zabezpieczyć tych danych w formie umożliwiającej zrozumienie ich znaczenia po kilku lub kilkudziesięciu latach tak, by odbiorca z przyszłości mógł bez większego wysiłku odkryć szczegóły procesu badawczego. Krótko mówiąc, ryzyko utraty użyteczności danych pojawia się w konsekwencji ich oderwania od kontekstu wytworzenia, od wiedzy ukrytej, która niegdyś czyniła te dane przydatnymi do przygotowania wykładu X lub referatu Y. Jak ujmuje to William Stafford Noble:

*Główna zasada przewodnia jest prosta: ktoś niezaznajomiony z twoim projektem powinen być w stanie przejrzeć twoje pliki komputerowe i zrozumieć ze szczegółami, co zrobiłeś i dlaczego [...] Jednakże najczęściej ten “ktoś” to właśnie ty. Za kilka miesięcy możesz nie pamiętać, co miałeś na myśli, gdy tworzyłeś ten konkretny zestaw plików, lub jakie wyciągnąłeś wnioski. Będziesz zmuszony albo poświęcić czas na rekonstrukcję swoich poprzednich eksperymentów, albo stracić wszelkie obserwacje, które zdobyłeś w wyniku tych eksperymentów*

*William Stafford Noble (2009) A Quick Guide to Organizing Computational Biology Projects.PLoSComputBiol 5(7): e1000424. doi:10.1371/journal.pcbi.1000424*

Opierając się na doświadczeniach i wiedzy ekspertów w zakresie danych badawczych, niniejszy przewodnik wskazuje, w jaki sposób historycy mogą dokumentować i strukturyzować swoje dane badawcze, aby zapewnić ich użyteczność w przyszłości. Przewodnik ten w zamyśle nie ma mieć charakteru nakazowego, zakłada raczej, że ​​czytelnicy będą naśladować, zmieniać i dostosowywać prezentowane pomysły tak, by jak najlepiej pasowały do ​​ich własnych badań.

**Dokumentowanie danych badawczych**

*Birkwood, Katie (girlinthe). “Zwycięstwo należy do mnie: jakiś czas temu opracowałam  parę Mądrych Rzeczy™ w Excelu. I ZROBIŁAM Z TEGO NOTATKI. I te notatki POZWOLIŁY MI TO POWTÓRZYĆ.” 7 października 2013, 3:46, Tweet.*

[*https://twitter.com/Girlinthe/status/387166944094199809*](https://twitter.com/Girlinthe/status/387166944094199809)

Celem dokumentowania jest uchwycenie procesu tworzenia danych, zmian wprowadzanych do danych i związanej z danymi wiedzy ukrytej. Metody zarządzania projektami, takie jak PRINCE2, kładą duży nacisk na precyzyjną, uporządkowaną i pełną dokumentację. Wspomnianametoda może przynosić korzystne efekty, zwłaszcza w przypadku dużych, złożonych projektów obejmujących wielu partnerów, jednak przeciętny historyk będzie raczej wybierał elastyczne, dostosowane do indywidualnych potrzeb podejście do dokumentacji, które wywodzi się z zasad zarządzania projektami, ale nie jest nimi ograniczone. W przypadku badań historycznych dokumentacjaumożliwiająca zachowanie użyteczności danych badawczych obejmuje:

* dokumentację opisującą notatki sporządzone podczas badania dokumentu w archiwum, tj. archiwalne odniesienia do oryginalnego dokumentu, reprezentatywność notatek (np. czy są to pełne transkrypcje, częściowe transkrypcje,czy streszczenia), wskazanie, jaka część dokumentu została zbadana lub zapis decyzji podjętych w celu wyłączenia fragmentów dokumentu z procesu badawczego.
* dokumentację opisującą dane w formie tabelarycznej, tj. sposób ich wygenerowania (np. ręczny lub zautomatyzowany), archiwalne odniesienia do oryginalnych źródeł, z których pochodzą dane lub wskazanie, które cechy oryginalnych źródeł zostały zachowane (i dlaczego).
* dokumentację opisującą katalog obrazów cyfrowych, tj. sposób utworzenia każdego obrazu, miejsca, z którego obrazy zostały pobrane lub notatki badawcze, które się do nich odnoszą.

Zpowyższych przykładów wynika, że jednym z kluczowych celów dokumentacji jest opisanie mających znaczenie powiązań, które istnieją pomiędzy danymi badawczymi, a które wraz z upływem czasu mogą nie wydawać się oczywiste.

Moment tworzenia dokumentacji jest w dużej mierze uzależniony od preferencji i rytmu pracy badacza. Główną zasadą jest opanowanie zwyczaju pisania i aktualizowania dokumentacji w regularnych odstępach czasu, najlepiej każdorazowo po zakończeniu porannej, popołudniowej lub całodziennej partii pracy. Jednocześnie nie należy martwić się o perfekcję dokumentacji, ale raczej dążyć do tego, by była ona napisana w sposób spójny i efektywny, a wskutek tego użyteczny dla dokumentującego oraz, miejmy nadzieję, innej osoby, która użyje tych danych badawczych wiele lat później.

**Formaty plików**

Najlepiej, gdy dane badawcze i dokumentacja zapisywane są w formatach uniwersalnych(niezależnych od platformy), takich jak .txt dla notatek i .csv (ang. comma-separated values - wartości rozdzielone przecinkiem) lub .tsv (ang. tab-separated values - wartości rozdzielane tabulatorami) dla danych tabelarycznych. Te proste formaty tekstowe mają przewagę nad formatami zastrzeżonymi, używanymi jako domyślne w pakiecie Microsoft Office lub iWork, ze względu na fakt, że mogą być otwierane za pomocą różnych pakietów oprogramowania, co zwiększa szanse na ogląd i edycję plików w przyszłości. Większość standardowych pakietów biurowych zawiera opcję zapisywania plików w formatach .txt, .csv i .tsv, co oznacza, że ​​możliwa jest kontynuacja pracy ze znanym oprogramowaniem i jednocześnie podejmowanie działań w celu udostępnienia jej wyników. Z punktu widzenia zachowania danych, te formaty mają tę dodatkową zaletę, że, w przeciwieństwie do formatów .doc lub .xls, zawierają jedynie elementy do odczytu maszynowego.Pogrubienie, kursywa i kolorowaniesą powszechnie praktykowane w celu wyróżnienia nagłówków lub stworzenia wizualnej ciągłości pomiędzy elementami danych.Jednak te dodatki ukierunkowane na obrazowanie nie nadają się do odczytu maszynowego,w związku z czym nie można ich znaleźć za pomocą wyszukiwarki, co ma szczególne znaczenie w przypadku dużej ilości informacji. Zalecane są zatem proste schematy zapisu, takie jak użycie podwójnej gwiazdki (\*) lub trzech symboli hash (**#)** do określenia właściwości danych. Przykładowo: w moich notatkach trzy znaki zapytania wskazują na coś, co muszę sprawdzić. Wybrałem taki sposób zapisu, ponieważ “???” mogę z łatwością znaleźć przy użyciu wyszukiwania za pomocą klawiszy CTRL + F.

Najprawdopodobniej w wielu przypadkach schematy zapisu będą wynikiem indywidualnej praktyki (i w konsekwencji będą musiały zostać opisane), aczkolwiek dostępne są także gotowe schematy, np. Markdown (pliki [Markdown](http://en.wikipedia.org/wiki/Markdown) zapisywane są jako .md). Doskonała ściągawka z Markdown jest dostępna na GitHub (<https://github.com/adam-p/markdown-the>) dla tych, którzy chcieliby naśladować - lub zaadaptować - ten istniejący już schemat. Dla użytkowników Windows jest zalecany edytor Notepad ++ <http://notepad-plus-plus.org/>, choć nie jest on niezbędny do pracy z plikami .md. Użytkownicy systemów operacyjnych Mac lub Unix za pomocne mogą uznać programy [Komodo Edit](http://komodoide.com/komodo-edit/) lub [TextWrangler](https://itunes.apple.com/gb/app/id404010395?mt=12).

**Podsumowanie 1**

Najważniejsze wskazówki dotyczące dokumentacji i formatów plików:

* Dąż do tego, aby dokumentacja mogła w precyzyjny i spójny sposób odzwierciedlić wiedzę ukrytą związaną z procesem badawczym. Czyń to w odniesieniu do prowadzenia notatek, tworzenia danych tabelarycznych, gromadzenia materiałów wizualnych.
* Zachowaj prostotę dokumentacji poprzez korzystanie z formatów plików i schematów zapisu, które są uniwersalne i mogą być odczytywane maszynowo.
* Włącz tworzenie i aktualizację dokumentacji w normalny przebieg pracy, ale nie pozwól, by dokumentowanie stało się zbyt dużym obciążeniem.
* Zainwestuj czas w prowadzenie aktualnej dokumentacji, tak aby zaoszczędzić go później, podczas prób rekonstrukcji twoich eksperymentów w przyszłości.

**Strukturyzacja danych badawczych**

Dokumentowanie twoich badań będzie łatwiejsze, jeśli zorganizujesz swoje dane badawcze w logiczny i możliwy do przewidzenia sposób.

Dlaczego?

Otóż, za każdym razem, gdy korzystamy z katalogu biblioteki lub archiwum, opieramy się na zbiorzeuporządkowanych informacji, które pozwalają nam poruszać się po zasobachbibliotecznych lub archiwalnych (istniejącychw formie fizycznej lub cyfrowej). Bez tak uporządkowanych informacji nasze badania byłyby znacznie mniej efektywne.

Badanie adresów URL może być dobrym przyczynkiem do zastanowienia się, dlaczego strukturyzacja danych badawczych w logiczny i łatwy do przewidzenia sposób może być przydatna podczas badań. Nieprawidłowe adresy URL są nieodtwarzalne, zatem nie można ich zacytować w kontekście naukowym. Z drugiej strony, poprawne adresy URL jasno wskazują na stronę, którą identyfikują, albo dzięki elementom semantycznym, albo poprzez użycie pojedynczego elementu danych w całym zestawie lub na większości stron.

Typowymi przykładami pierwszych z wymienionych adresów URL są te używane przez dzienniki internetowe lub serwisy blogerskie. WordPress posługuje się następującym formatem adresów URL:

* *nazwa strony/rok(4 cyfry)/miesiąc(2 cyfry)/dzień(2 cyfry)/słowa-z-tytułu-rozdzielone-dywizem*
* <http://cradledincaricature.com/2014/02/06/comic-art-beyond-the-print-shop/>

Podobna forma używana jest przez agencje prasowe, np. The Guardian:

* *nazwa strony/zakładka strony/rok (4 cyfry)/miesiąc (3 litery)/dzień (2 cyfry)/słowa-opisujące-treść-rozdzielone-dywizem*
* <http://www.theguardian.com/uk-news/2014/feb/20/rebekah-brooks-rupert-murdoch-phone-hacking-trial>

W katalogach archiwów często używa się adresów URL uporządkowanych za pomocą pojedynczego elementu danych. The British Cartoon Archiveporządkuje swoje archiwum internetowe wedługnastępującego formatu:

* *nazwa strony/record/numer referencyjny*
* <http://www.cartoons.ac.uk/record/SBD0931>

Również OldBailey Online używa tego typu formatu:

* *nazwa strony/browse.jsp?ref=numer referencyjny*
* <http://www.oldbaileyonline.org/browse.jsp?ref=OA16780417>

Powyższe przykłady pokazują, że połączenie semantycznego opisu z elementami danych tworzy logiczne i możliwe do przewidzenia struktury odczytywane zarówno przez człowieka, jak i przez maszynę. Przełożenie tej metody na dane cyfrowe zgromadzone podczas badań historycznych może ułatwić przeglądanie i wyszukiwanie danych za pomocą standardowych narzędzi systemów operacyjnych (a także, co zostanie wyjaśnione w dalszej części przewodnika, poprzez zastosowanie bardziej zaawansowanych narzędzi).

W praktyce (użytkownicy systemów OS X i Linux powinni zastąpić wszystkie ukośniki lewe ukośnikiem prawym) struktura dobrego archiwum danych badawczych może wyglądać mniej więcej tak:

Katalog podstawowy lub główny, nazwany na przykład "praca".

\praca\

Szereg podkatalogów:

\praca\wydarzenia\

\badania\

\nauczanie\

\teksty\

W ramach tych katalogów znajdują się zestawy katalogów przyporządkowane do każdego wydarzenia, projektu badawczego, modułu lub utworu. Metoda nazewnictwa zawierająca elementy danych pozwala zorganizować informacje bez potrzeby stosowania podkatalogów według, powiedzmy, określania roku lub miesiąca.

\praca\badania\2012-01\_czasopismo\_artykuły

            \2014-02\_infrastuktura

Wreszcie, dalsze podkatalogi mogą być używane do rozdzielania informacji, których ilość rośnie wraz z rozwojem projektu.

\praca\badania\2014\_czasopismo\_artykuły\analizy

                                        \dane

                                        \notatki

Oczywiście nie wszystkie informacje będą perfekcyjnie dopasowane do danej struktury, a wraz z pojawieniem się nowych projektów system będzie musiał zostać zrewidowany.

Tak czy inaczej, odstępstwa od metody są dopuszczalne, o ile ogólna struktura katalogów pozostaje logiczna i możliwa do przewidzenia oraz o ile wszystkie materiały, które nie posiadają tych cech, są udokumentowane w zrozumiały sposób: przykładowo, podkatalog “utwór” w zaprezentowanej powyżej strukturze może zawierać plik .txt ustalający, co ten podkatalog zawierał (np. wersje robocze i końcowe tekstu), a czego nie (np. badania dotyczące tej pracy).

Nazwa takiego pliku .txt, a właściwie jakiejkolwiek dokumentacji oraz danych badawczych, powinna zapewniać łatwą identyfikację zawartości pliku. Nazwa "Uwagi dotyczące folderu.docx" nie spełnia tego warunku, podczas gdy zapis "2014-01-31\_Utwór\_readme.txt" odzwierciedla tytuł katalogu i zawiera informacje o dacie (czytelnicy z Ameryki Północnej powinni zwrócić uwagę na wybrany przeze mnie format, tj. rok-miesiąc-dzień). Plik [“readme”, który stworzyłem dla najnowszego projektu](https://programminghistorian.org/assets/preserving-your-research-data/network_analysis_of_Isaac_Cruikshank_and_his_publishers_readme.txt), zawiera informacje, które mogą okazać się przydatne.

Niech krótka opowieść ku przestrodze podkreśli wartość tej metody.W trakcie poprzedniego projektu badawczego zebrałem około 2000 cyfrowych obrazów z różnych źródeł internetowych, przedstawiających georgiańskie rysunki satyryczne.Dla każdego z obrazów zachowałem nazwę pobieranego pliku. Gdybym od samego początku zastosował do nich metodę nazewniczą (np. "ROK PUBLIKACJI\_NAZWISKO ARTYSTY\_TYTUŁ PRACY.FORMAT), mógłbym z łatwością przeszukiwać te dane.Rozpoczynanie każdej nazwy pliku formatem RRRRMMDD pozwoliłoby na ułożenie plików w porządku chronologicznym w systemach operacyjnych Windows, OS X i Linux. Natomiast usunięcie wszystkich spacji i znaków interpunkcyjnych (z wyjątkiem myślnika, kropki i podkreślenia) z nazw plików umożliwiłoby pracę za pomocą wiersza poleceń. Jednakże, ponieważ nie zastosowałem się do tych zasad, musiałem spędzić bardzo dużo czasu na poprawieniunazwy każdego pliku z osobna, aby doprowadzić cały zbiór do stanu użyteczności.

Ponadto, stosowanie takich metod nazewnictwa do wszystkich danych badawczych w logiczny i możliwy do przewidzenia sposób ułatwia zrozumienie i zwiększa czytelność struktury danych. Przykładowo, dla projektu dotyczącego artykułów w czasopiśmie możemy wybrać katalog...

\praca\badania\2014-01\_czasopismo\_artykuły\

...w którym data (rok-miesiąc) oznacza, kiedy projekt się rozpoczął. Do tego katalogu włączamy katalog \dane\, w którym przechowujemy oryginalne dane związane z projektem.

2014-01-31\_czasopismo\_artykuły.tsv

Razem z danymi przechowujemy dokumentację, która je opisuje.

2014-01-31\_czasopismo\_artykuły.tsv

2014-01-31\_czasopismo\_artykuły\_notatki.txt

Wracając do poziomu katalogu \2014-01\_czasopismo\_artykuły\, tworzymy katalog \analiza\, w którym umieszczamy pliki:

2014-02-02\_czasopismo\_artykuły\_analiza.txt

2014-02-15\_czasopismo\_artykuły\_analiza.txt

Zwróć uwagę na różnice w atrybutach dat. Określają one daty przeprowadzenia analiz, według metody opisanej krótko w 2014-02-02\_czasopismo\_artykuł\_analiza\_readme.txt.

Wreszcie, w obrębie \danych\ tworzymy katalog \dane\_pochodne\ zawierający dane uzyskane z oryginalnego pliku 2014-01-31\_czasopismo\_artykuły.tsv. W tym przypadku nazwy plikówpochodnych zawierają słowa kluczowe: "afryka", "ameryka", "sztuka" itd.

2014-01-31\_czasopismo\_artykuł\_SK\_afryka.tsv

2014-01-31\_czasopismo\_artykuł\_SK\_ameryka.tsv

2014-02-01\_czasopismo\_artykuł\_SK\_sztuka.tsv

2014-02-02\_czasopismo\_artykuł\_SK\_brytania.tsv

[SK – Słowo Kluczowe – przyp. tłumacza]

**Podsumowanie 2**

Najważniejsze wskazówki dotyczące strukturyzacji danych badawczych:

* Struktury danych powinny być logiczne i możliwe do przewidzenia
* Rozważ użycie elementów semantycznych lub identyfikatorów danych, aby uporządkować katalogi danych badawczych
* Dostosuj strukturę danych badawczych do konkretnych badań
* Stosuj takie metody nazewnictwa dla katalogówi plików, które umożliwią ich identyfikację, tworzenie powiązań między danymi oraz zapewnią czytelność i zrozumiałość struktury danych w dalszej perspektywie

**Podsumowanie końcowe**

W niniejszym przewodniku zasugerowano sposoby dokumentowania i organizowania danych badawczych, których celem jest zapewnienie zachowania danych dzięki uchwyceniu wiedzy ukrytej zdobytej podczas procesu badawczego, a tym samym ułatwienia korzystania z tych informacji w przyszłości. Dla dokumentacji i danych badawczych zalecono stosowanie formatów uniwersalnych (niezależnych od platformy)i odczytywalnych komputerowo. Wskazuje się na adresy URL stanowiące przykłady zarówno dobrych jak i złych praktyk w zakresie struktury danych, które mogą być powielane w stosunku do danych badawczych historyków.

Przedstawione sugestie mają służyć jedynie jako wskazówki; oczekuje się, że naukowcy dostosują je we właściwy sposób do swoich celów. Przy tym, zaleca się, aby badacze mieli na uwadze stosowanie jak najlepszych strategii ochrony zasobów cyfrowych oraz praktyk zarządzania projektami, które jednocześnie nie sprawią, że dokumentacja i strukturyzowanie badań staną się uciążliwe czasowo. Ostatecznie, celem tego przewodnika jest sprawienie, by badania historyczne, które generują dane były jeszcze bardziej efektywne. Tutaj chodzi o twoje badania.

**Zalecana literatura**

Ashton, Neil, ‘Seven deadly sins of data publication’, School of Data blog (17 October 2013) http://schoolofdata.org/2013/10/17/seven-deadly-sins-of-data-publication/

Hitchcock, Tim, ‘Judging a book by its URLs’, Historyonics blog (3 January 2014) http://historyonics.blogspot.co.uk/2014/01/judging-book-by-its-url.html

Howard, Sharon, ‘Unclean, unclean! What historians can do about sharing our messy research data’, Early Modern Notes blog (18 May 2013) http://earlymodernnotes.wordpress.com/2013/05/18/unclean-unclean-what-historians-can-do-about-sharing-our-messy-research-data/

Noble, William Stafford, A Quick Guide to Organizing Computational Biology Projects.PLoSComputBiol 5(7): e1000424 (2009) https://doi.org/10.1371/journal.pcbi.1000424

Oxford University Computing Services, ‘Sudamih Project. Research Information Management: Organising Humanities Material’ (2011) https://zenodo.org/record/28329

Pennock, Maureen, ‘The Twelve Principles of Digital Preservation (and a cartridge in a repository…)’, British Library Collection Care blog (3 September 2013) http://britishlibrary.typepad.co.uk/collectioncare/2013/09/the-twelve-principles-of-digital-preservation.html

Pritchard, Adam, ‘Markdown Cheatsheet’ (2013) https://github.com/adam-p/markdown-here

Rosenzweig, Roy, ‘Scarcity or Abundance? Preserving the Past in a Digital Era’, The American Historical Review 108:3 (2003), 735-762.

UK Data Archive, ‘Documenting your Data’ http://data-archive.ac.uk/create-manage/document

**O AUTORZE**

James Baker jest Starszym Wykładowcą Historii Cyfrowej i Archiwów na Uniwersytecie Sussex oraz badaczem historycznychinterakcji pomiędzy ludźmi iobiektami.

**SUGEROWANY ZAPIS CYTOWANIA**

James Baker, "Preserving Your Research Data," The Programming Historian 3 (2014), <https://programminghistorian.org/en/lessons/preserving-your-research-data>.