

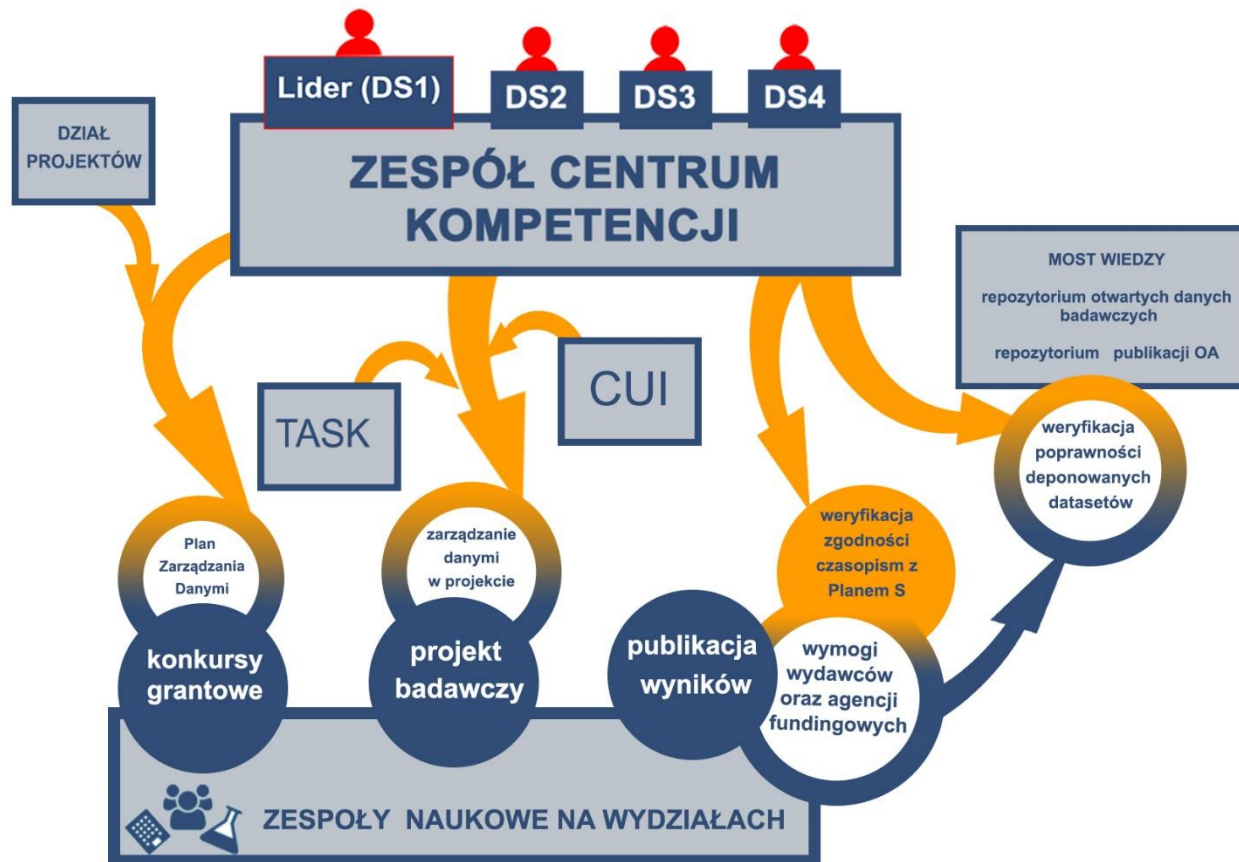


12. Seminarium Open Access Biblioteka Uniwersytetu Łódzkiego

Magdalena Szuflita-Żurawska
Biblioteka Politechniki Gdańskiej

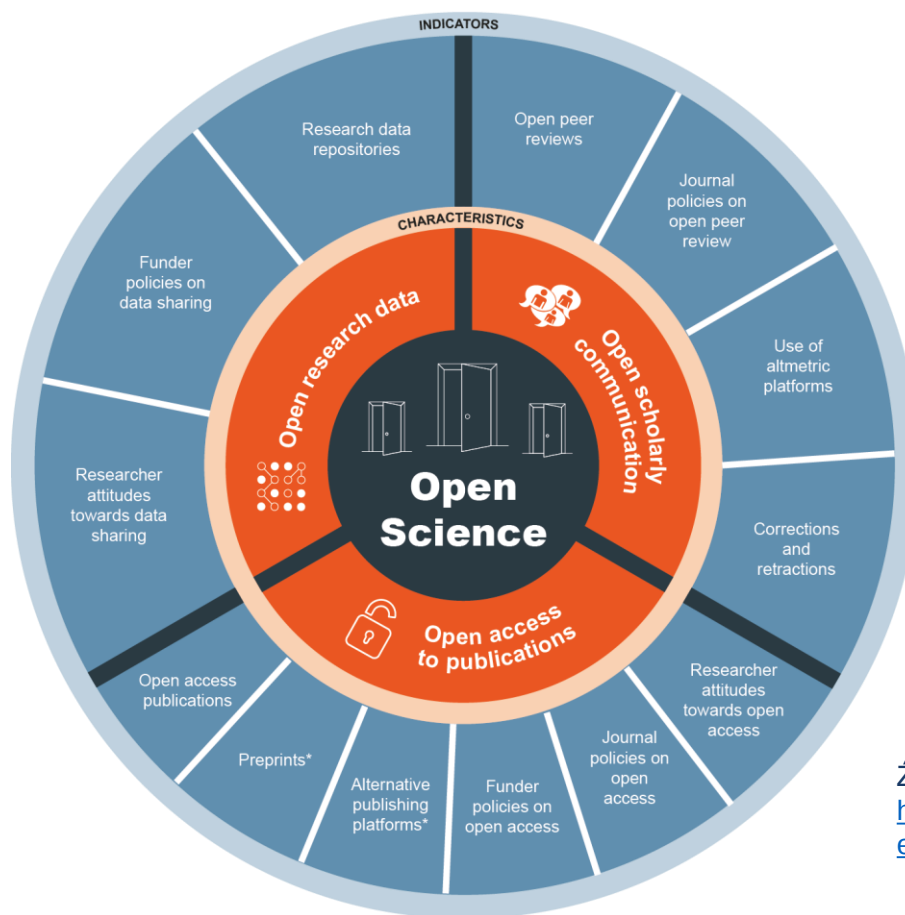


Centrum Kompetencji Otwartej Nauki PG





Open Science



Źródło:

<https://ec.europa.eu/research/openscienc e/index.cfm?pg=home§ion=monitor>



Co to są dane badawcze?

- Dane pozyskiwane w procesach badawczych
- Dane jakościowe i ilościowe
- „Primary and secondary” data
- Dane surowe - dane zebrane, ale jeszcze nieprzeanalizowane
- Dane potrzebne do walidacji wyników zaprezentowanych w publikacji naukowych („underlying data”)
- Uwaga! Jeśli dane potrzebują specjalnego oprogramowania do ich analizy, beneficjenci proszeni są o jego udostępnienie.

Typy danych badawczych

Dane badawcze można podzielić na cztery główne typy w oparciu o metody ich gromadzenia:

- **Dane obserwacyjne (Observational Data)**

Dane przechwytywane w czasie rzeczywistym, nie można ich „odzyskać”, często nazywane „unique data”. Przykłady: odczyty czujników, dane telemetryczne, wyniki ankiet, obrazy, obserwacje.

- **Dane eksperymentalne (Experimental Data)**

Dane ze sprzętu laboratoryjnego uzyskiwane w kontrolowanych warunkach, często powtarzalne, ale bardzo kosztowne. Przykłady: sekwencje genów, chromatogramy, odczyty pola magnetycznego, spektroskopia.

- **Dane symulacji (Simulation Data)**

Dane generowane modeli testowych badających rzeczywiste lub teoretyczne systemy. Modele i metadane wejściowe najczęściej są tu ważniejsze niż dane wyjściowe. Przykłady: modele klimatyczne, modele ekonomiczne, systemy inżynieryjne.



Typy danych badawczych

- **Dane pochodne/skompilowane (Derived / Compiled Data)**

Wyniki analiz danych bądź agregowane z różnych źródeł. Powtarzalne, ale mogą być bardzo kosztowne. Przykłady: eksploracja tekstu oraz danych, bazy danych, modele 3D.

- **Dane referencyjne (Reference or Canonical)**

Poprawione lub organiczne zbiory danych, zwykle recenzowane, często publikowane i selekcjonowane. Przykłady: bazy danych sekwencji genów, dane ze spisu powszechnego, dane GUS, struktury chemiczne.

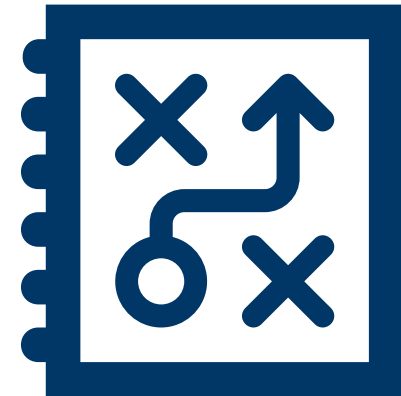
Inne formy danych: dane tekstowe, numeryczne, multimedialne, audio, kod, oprogramowanie, wzory, video, instrumentalne.



Otwarte Dane Badawcze?

Otwarte dane to dane, do których każdy ma dostęp. Można je ponownie wykorzystywać, modyfikować i redystrybuować.

- Dostęp do treści naukowych (np. w ramach licencji Creative Commons) pozwala na lepsze „zrozumienie” dołączonych do treści danych (kontekst, metodologia pozyskania).





Udostępnianie danych badawczych

Dane powinny być tak otwarte, jak to możliwe i na tyle zamknięte, na ile to jest konieczne.

- Dane badawcze najczęściej udostępniane są w postaci tzw. datasetów, czyli zbiorów stanowiących pewną odrębną całość i zawierających dane powiązane z jedną publikacją, projektem naukowym, bądź eksperymentem.
- Aby pomóc naukowcom w odpowiednim przygotowaniu oraz udostępnianiu danych badawczych sformułowano zasady FAIR.
- Nie są to sztywne zasady, a raczej wytyczne do właściwej pracy z danymi. Zasady FAIR zostały stworzone i są w dalszym ciągu rozwijane przede wszystkim po to, aby uczynić dane dostępnymi zarówno dla użytkowników jak i oprogramowania komputerowego przeszukującego bazy danych bez udziału człowieka.
- **UWAGA! Zasady FAIR odnoszą się do danych, metadanych, ale także do infrastruktury np. repozytorium.**



FAIR Data i metadata

Findable – Możliwe do odnalezienia

- Zbiór danych opatrzony jest metadanymi, które umożliwiają odnalezienie tego zbioru zarówno przez ludzi jak i programy komputerowe;
- Do zbioru przypisany jest unikalny identyfikator (np. DOI), który stanowi jednocześnie element metadanych opisujących ten zbiór;
- Metadane są indeksowane w ogólnodostępnych bazach danych umożliwiających ich przeszukiwanie.



FAIR Data i metadata

Accessible – Dostępne

- Dostęp do zbioru danych, a przynajmniej do metadanych, możliwy jest bezpośrednio poprzez unikalny identyfikator i nie wymaga dodatkowych narzędzi ani oprogramowania;
- Metadane są zawsze dostępne, nawet jeśli sam zbiór danych został już usunięty lub przeniesiony.



FAIR Data i metadata

Interoperable – Interoperacyjne

- Dane oraz metadane są dostarczone w formacie zapewniającym łatwy odczyt i przetwarzanie zarówno przez ludzi jak i komputery;
- Zbiory danych i metadane je opisujące zawierają odnośniki do innych powiązanych z nimi zbiorów.



FAIR Data i metadata

Reusable – Możliwe do ponownego użycia

- Metadane zawierają liczne atrybuty dokładnie opisujące zbiór danych i ułatwiające użytkownikom określenie ich przydatności dla ich własnych badań;
- Zbiór danych zawiera licencję określającą jednoznacznie warunki ponownego wykorzystania i przetwarzania danych;
- Metadane wyraźnie określają autora oraz miejsce powstania danych;
- Metadane są skonstruowane według ogólnie przyjętych standardów specyficznych dla danej dyscypliny oraz rodzaju danych.



Plan zarządzania danymi - Wymagania agencji fundingowych

- **Data Management Plan (DMP) czyli Plan Zarządzania Danymi** to dokument opisujący czynności wykonywane na każdym etapie pracy z danymi badawczymi. DMP powinien powstać już na wstępnym etapie badań naukowych. Jest on coraz częściej wymagany przez instytucje i agencje przyznające środki finansowe na badania naukowe (Narodowe Centrum Nauki, Komisja Europejska, Economic and Social Research Council, Natural Environmental Research Council).
- **DMP ułatwia planowanie procedur związanych z pozyskiwaniem, przetwarzaniem i dzieleniem się danymi badawczymi.**
- Plan zarządzania danym - „living document”
- Utworzony przed rozpoczęciem badań lub na wczesnych etapach, może być aktualizowany podczas trwania projektu
- Nie wszystkie odpowiedzi są znane na początku!
- NCN nie przewiduje kontroli planów w czasie trwania projektów



Plan zarządzania danymi NCN

1. Opis danych oraz pozyskiwanie lub ponowne wykorzystanie dostępnych danych

- 1.1 W jaki sposób będą pozyskiwane lub wytwarzane nowe dane lub ponownie wykorzystywane dane już istniejące?
- 1.2 Jakie dane (tj. rodzaje, formaty, objętości) będą pozyskiwane lub wytwarzane w projekcie?

2. Dokumentacja i jakość danych

- 2.1 Jakie metadane i dokumentacja (np. metodologia oraz sposoby pozyskiwania i organizacji danych) będą towarzyszyć danym w projekcie?
- 2.2 Jakie planują Państwo zastosować środki kontroli jakości?

3. Przechowywanie i tworzenie kopii zapasowych podczas badań

- 3.1 W jaki sposób w trakcie projektu będą przechowywane dane i metadane? W jaki sposób będą tworzone ich kopie zapasowe?
- 3.2 W jaki sposób zostanie zapewnione bezpieczeństwo i ochrona danych wrażliwych w okresie trwania projektu?



Plan zarządzania danymi NCN

4. Wymogi prawne, kodeksy postępowania

- 4.1 Jeżeli będzie miało miejsce przetwarzanie danych osobowych, w jaki sposób zostanie zapewniona zgodność z przepisami dotyczącymi danych osobowych oraz ich ochrony?
- 4.2 W jaki sposób planują Państwo zapewnić zgodność z innymi przepisami, takimi jak prawa własności intelektualnej i prawa własności? Jakie przepisy znajdują w tym przypadku zastosowanie?

5. Udostępnianie i długoterwale przechowywanie danych

- 5.1 Kiedy i w jaki sposób będą udostępniane dane z projektu? Czy istnieją ewentualne ograniczenia i zakazy dotyczące ich udostępniania?
- 5.2 Jak będzie wyglądać selekcja danych przeznaczonych do utrwalenia i gdzie będą one długoterminowo przechowywane (np. w repozytorium danych, archiwum)?



Plan zarządzania danymi NCN

- 5.3 Jakie metody lub oprogramowanie umożliwiają dostęp do danych i korzystanie z danych?
- 5.4 W jaki sposób zagwarantują Państwo stosowanie unikalnego i trwale przypisanego identyfikatora (takiego jak cyfrowy identyfikator dokumentu elektronicznego (DOI)) dla każdego zbioru danych?

6. Zadania związane z zarządzaniem danymi oraz zasoby

- 6.1 Kto będzie odpowiadał za zarządzanie danymi (tj. kto będzie ich opiekunem)?
- 6.2 Jakie zasoby zostaną przeznaczone na cele zarządzania danymi i zagwarantowanie przestrzegania zasad FAIR?



Przygotowanie danych do udostępnienia

1. Selekcja – nie wszystkie dane muszą zostać udostępnione. Dobierając zbiory danych do archiwizacji warto kierować się takimi czynnikami jak:

- Wymagania agencji finansujących badania naukowe;
- Wartość naukowa danych badawczych;
- Wyjątkowość – warto sprawdzić czy dane nie duplikują się z innymi istniejącymi już zbiorami danych;
- Możliwość replikacji wyników badań – czy dane zawierają wszystkie parametry umożliwiające powtórzenie eksperymentu;
- Kwestie ekonomiczne – jakie koszty wiążą się z zarządzaniem i przechowywaniem danych i czy są one uzasadnione.



Przygotowanie danych do udostępnienia

2. Usunięcie danych wrażliwych umożliwiających identyfikację badanych osób:

- Anonimizacja polega na przekształceniu danych osobowych w sposób uniemożliwiający przyporządkowanie poszczególnych informacji do określonej lub możliwej do zidentyfikowania osoby.
- Pseudonimizacja – to przetworzenie danych tak, by nie można ich było przypisać osobie, której te dane dotyczą, bez użycia dodatkowych informacji.
- Podstawową cechą odróżniającą pseudonimizację i anonimizację jest odwracalność. Anonimizacja jest procesem nieodwracalnym, pseudonimizacja zaś jest odwracalna.



Przygotowanie danych do udostępnienia

■ 3. Wybór formatów plików

Dane powinny zostać opublikowane w ogólnodostępnym formacie, który nie wymaga komercyjnego oprogramowania i wykorzystuje standardowe kodowanie (ASCII, UTF-8). Warto również zastanowić się, jakie formaty plików funkcjonują w naszej dyscyplinie, tak aby użytkownicy po pobraniu zbioru danych nie byli zmuszeni do ich dodatkowej konwersji, która może prowadzić do utraty jakości danych.

Formaty plików

- Text, Documentation, Scripts: XML, PDF/A, HTML, Plain Text.
- Still Image: TIFF, JPEG 2000, PNG, JPEG/JFIF, DNG (digital negative), BMP, GIF.
- Geospatial: Shapefile (SHP, DBF, SHX), GeoTIFF, NetCDF.
- Graphic Image:
 - raster formats: TIFF, JPEG2000, PNG, JPEG/JFIF, DNG, BMP, GIF.
 - vector formats: Scalable vector graphics, AutoCAD Drawing Interchange Format, Encapsulated Postscripts, Shape files.
- cartographic: Most complete data, GeoTIFF, GeoPDF, GeoJPEG2000, Shapefile.
- Audio: WAVE, AIFF, MP3, MXF, FLAC.
- Video: MOV, MPEG-4, AVI, MXF.
- Database: XML, CSV, TAB.



Metadane danych badawczych

Zgodnie z zasadami [FAIR](#) istotny element każdego zbioru danych stanowią metadane. Metadane to dane o danych. Są kluczem do uzyskania dostępu do danych badawczych, ich zrozumienia i ponownego wykorzystania. Istnieją trzy główne typy metadanych:

- **Metadane opisowe** – dostarczają informacji niezbędnych do odszukania czy też identyfikacji zbioru danych. Mogą zawierać elementy, takie jak tytuł, streszczenie, autor i słowa kluczowe.
- **Metadane strukturalne** – opisują relacje i zależności pomiędzy poszczególnymi zbiorami oraz elementami tych zbiorów w celu np. ułatwienia nawigacji.



Metadane danych badawczych

Metadane administracyjne – zawierają informacje pomocne w zarządzaniu danym zasobem. Zawierają informacje takie jak sposób i datę jego utworzenia, typ pliku i informacje dotyczące dostępu. Istnieje kilka podzbiorów danych administracyjnych. Dwa z nich często wymieniane jako oddzielne typy metadanych to:

- **Metadane zarządzania prawami**, które dotyczą praw własności intelektualnej,
- **Metadane konserwacji**, które zawierają informacje potrzebne do archiwizacji i utrzymania zasobu.

Metadane powinny informować m.in. o: strukturze danych, ograniczeniach ich dotyczących (jeśli takie istnieją), o tym co dane oznaczają i w jaki sposób je cytować.



Metadane danych badawczych

- Wszystkie dane badawcze muszą być udostępnione wraz z ich metadanymi
- Metadane: podstawowe informacje np. tytuł, autor, terminy, prawa dostępu...
- Dokumentacja: kod, metody, słownik danych, kontekst...
- Standardy metadanych - <http://www.dcc.ac.uk/resources/metadata-standards>



Archiwizacja i udostępnianie danych

- Należy dokładnie zapoznać się z warunkami korzystania z serwisu i sprawdzić czy spełnia on nasze wymagania;
- Warto również dowiedzieć się gdzie i na jakich zasadach będą przechowywane nasze dane oraz w jaki sposób będą zabezpieczone;
- Aby zbiór danych spełniał wymogi zasad FAIR musi być możliwy do odnalezienia, dlatego należy upewnić się, że dane repozytorium zapewnia przypisanie naszym zbiorom identyfikatora DOI. Ważne jest również odpowiednie powiązanie zbioru z jego autorami, dlatego istotna jest opcja wprowadzenia do opisu identyfikatora naukowca (np. numeru ORCID);
- Przy wyborze warto też sprawdzić czy inni naukowcy z naszej dyscypliny korzystają z danego repozytorium;
- Należy dowiedzieć się, czy repozytorium wspiera używany w naszej dyscyplinie standard metadanych;
- Trzeba mieć na uwadze, że niektóre repozytoria mogą pobierać opłatę za archiwizację danych – tzw. Data Processing Charge;
- Ważnym aspektem aktywności naukowej jest popularyzacja dorobku naukowego, dlatego warto upewnić się, że zasoby wybranego przez nas repozytorium znajdują się w bazach indeksujących repozytoria danych badawczych. Przykładami takich baz są Data Citation Index, Mendeley Data czy Google Dataset Search.



Role w opisie datasetów (na przykładzie Data Cite)

Przy rejestracji datasetu w repozytorium, warto pamiętać o zaznaczeniu roli w opisie datasetu np.

- Członek zespołu
- Dystrybutor danych
- Edytor
- Fundator
- Instytucja rejestrująca zasób
- Kierownik projektu
- Osoba gromadząca dane
- Producent
- Twórca
- Powiązana osoba



Wersjonowanie

- Kontrola wersji to sposób a śledzenie zmian w procesie bądź w zestawach danych.
- Szczególnie ważne jest zapisywanie wersji jeśli w projekcie pracuje więcej niż jedna osoba.
- Zaznaczaj każdą zmianę.
- Śledź zmiany w plikach.
- Dobrą praktyką jest używanie konwencji w nazewnictwie plików, np. nagłówki plików.
- Można używać oprogramowania do kontroli wersji (SVN, Git)



Zarządzanie prawami własności intelektualnej w odniesieniu do danych badawczych

Należy pamiętać, że istnieje **wiele sytuacji**, w których dane **nie są** chronione prawem autorskim i prawami pokrewnymi.

Takie dane mogą obejmować **fakty, nazwiska, liczby** - rzeczy, które są uważane za „nieoryginalne” i należą do domeny publicznej a zatem **nie podlegają** ochronie praw autorskich – **dane surowe**.

Podobnie baza danych (która jest uporządkowanym zbiorem danych) **może zostać** uznana za „nieoryginalną”, a zatem nie kwalifikującą się do ochrony praw autorskich, a dodatkowo może **zostać wyłączona z innych form ochrony** (takich jak unijne prawo do bazy danych sui generis, znane również jako „SGDR”, dla nieoryginalnych baz danych).



Zarządzanie prawami własności intelektualnej w odniesieniu do danych badawczych

- Licencje Creative Commons - <https://creativecommons.org/choose/>
- Open Data Commons - <https://opendatacommons.org/>

W ramach projektu Open Data Commons powstały licencje, które zostały zaprojektowane specjalnie w celu licencjonowania baz danych. Przykładami są:

- Public Domain Dedication and License (PDDL) – domena publiczna dla baz danych. Zakłada nieograniczoną możliwość pobierania, udostępniania i modyfikowania baz danych.
- Open Data Commons Attribution License (ODC-By) – licencja, w której jedynym warunkiem kopiowania i modyfikowania bazy danych jest uznanie autorstwa.
- Open Data Commons Open Database License (ODC – ODbL) – otwarta licencja zezwalająca na kopiowanie, przetwarzanie oraz rozpowszechnianie bazy danych pod warunkiem uznania jej autorstwa oraz upowszechniania na takich samych warunkach.
- Domena publiczna



Bezpieczeństwo danych

- Bezpieczeństwo danych to ochrona danych przed nieautoryzowanym dostępem, wykorzystaniem, zmianą, ujawnieniem i zniszczeniem.
- Dane mogą być sklasyfikowane jako wysoce wrażliwe, średnio wrażliwe lub niewrażliwe (publiczne). Wymagania dotyczące każdego z tych typów danych są różne.

Przykłady:

- Dane poufne nie powinny być przechowywane na komputerach z dostępem do sieci.
- Dane poufne nie powinny być wysyłane poprzez e-mail lub FTP (w razie potrzeby trzeba użyć szyfrowania).
- Pliki i komputery powinny być chronione silnym hasłem (zmienianym np. raz w miesiącu).
- Dobre praktyki archiwizacji danych wymagają stosowania reguły 3-2-1 – tworzenie trzech kopii zapasowych, na dwóch odrębnych nośnikach, w tym jednej kopii w innej lokalizacji fizycznej np. inny budynek lub "chmura".



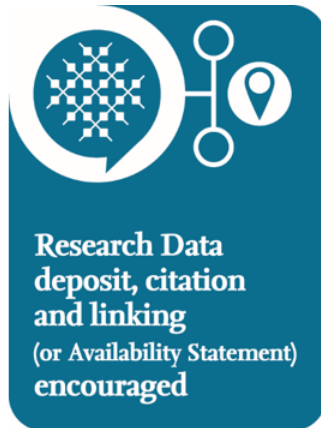
Dane badawcze i wymagania czasopism/wydawnictw

- **Wydawcy coraz częściej wymagają udostępnienia danych! Sprawdź politykę wydawniczą**
- Np. „A condition of publication in a Nature Research journal is that authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications”. - <https://www.nature.com/nature-research/editorial-policies/reporting-standards>
- Taylor & Francis - <https://authorservices.taylorandfrancis.com/wp-content/uploads/2019/11/Basic.pdf>



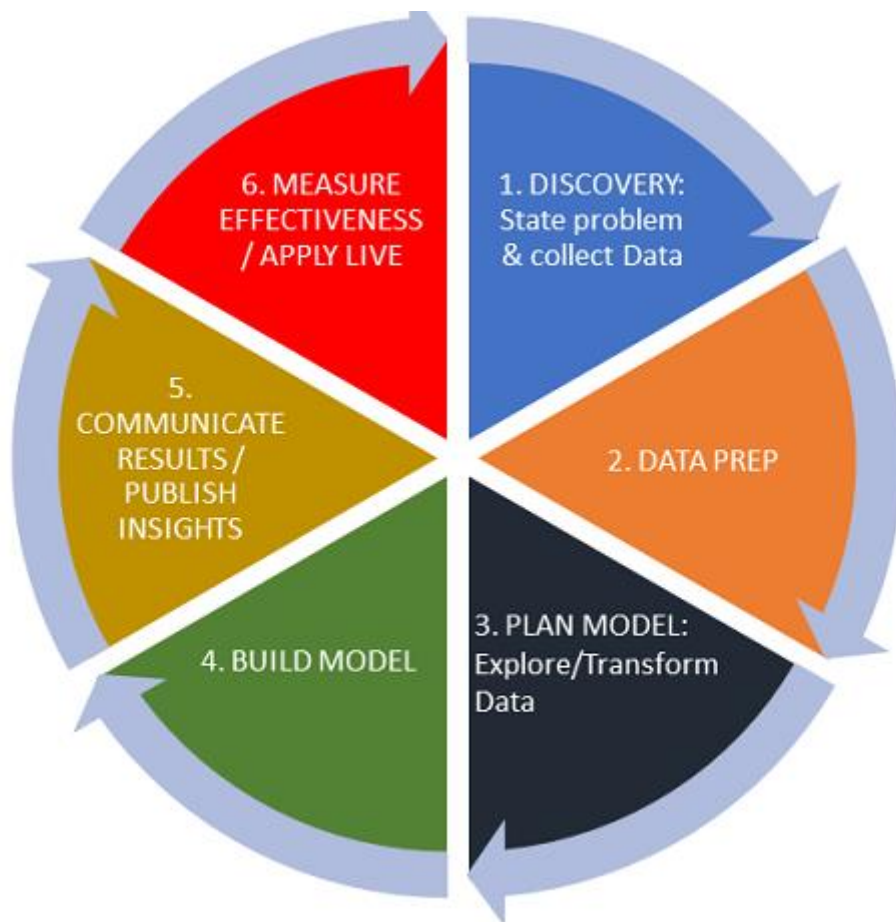


Dane badawcze i wymagania czasopism/wydawnictw



Elsevier - <https://www.elsevier.com/authors/author-resources/research-data/data-guidelines>

Cykl życia danych – The Lifecycle of Data



Modele zarządzania danymi na etapie realizacji projektów

„Cykl życia danych” jest również przedmiotem debaty, ponieważ dane tak naprawdę nie ewoluują i nie rosną tak jak ziarno. Niektórzy autorzy dodają różne etapy lub mogą je inaczej nazywać.

Źródło:
<https://www.datasciencecentral.com/profiles/blogs/the-lifecycle-of-data>



POLITECHNIKA
GDAŃSKA

12. Seminarium Open Access
Biblioteka Uniwersytetu Łódzkiego

Pytania?

Dziękuję za uwagę!
Zapraszam do kontaktu
magdalena.szuflita@pg.edu.pl