

DATA STEWARD

Дослідницькі дані: від первинної інформації до опублікованого набору

Що таке дослідницькі дані?

Дослідницькі дані — це не лише таблиці з результатами експериментів. Це можуть бути вимірювання, спектри, зображення, анкети, транскрипти інтерв'ю, аудіо й відео, код, моделі, польові нотатки, бази даних та інші матеріали, на яких ґрунтуються результати дослідження.

Від сирих даних до опублікованого набору: що відбувається з даними під час дослідження?

Дослідницькі дані рідко існують в одному незмінному вигляді. У процесі дослідження вони проходять кілька етапів: від первинної інформації, отриманої безпосередньо з експерименту, опитування чи джерела, до очищеного та проаналізованого набору, а згодом — до версії, яку можна архівувати або поширювати для повторного використання.

У спрощеному вигляді цей шлях можна представити так:



Це не завжди строго лінійний процес. Дослідник може кілька разів повертатися до попередніх етапів, змінювати процедуру обробки, доповнювати набір новими даними або створювати декілька похідних наборів. Проте розрізнення цих етапів допомагає правильно організувати файли, документувати зміни та вирішити, які саме дані потрібно зберігати й поширювати.

1. Raw data — первинні, або «сирі», дані

Raw data — це дані у максимально близькому до первинного отримання вигляді, до змістовної обробки, очищення чи інтерпретації.

Для експериментального дослідження це можуть бути вихідні файли приладу, спектри, мікрофотографії, результати вимірювань або журнали експерименту. Для опитування — первинний експорт відповідей із LimeSurvey, Google Forms чи іншої системи. Для якісного дослідження — аудіозаписи інтерв'ю, польові нотатки або первинні транскрипти. Для гуманітарного дослідження — цифрові копії джерел, корпус текстів, фотографії архівних документів чи первинна бібліографічна база.

Важливо: первинні дані бажано зберігати окремо й не перезаписувати під час подальшої обробки. Якщо ви виправляєте, очищуєте або трансформуете дані, краще створювати нову версію.

2. Processed data — оброблені дані

Processed data — це первинні дані після технічного очищення, структурування, кодування або інших перетворень, необхідних для подальшого аналізу.

На цьому етапі можуть виправлятися очевидні помилки, усуватися дублікати, стандартизуватися назви змінних, перетворюватися одиниці вимірювання, кодуються категоріальні відповіді, створюються псевдоніми, об'єднуються файли або формуються єдина таблиця.

Критично важливо, щоб такі зміни були **відтворюваними**. Інший дослідник — або ви самі через рік — повинен мати можливість зрозуміти, як із первинних даних була отримана оброблена версія. Для цього варто зберігати скрипти обробки, журнал змін, правила перекодування або короткий опис у README-файлі.

3. Analysed data — дані, підготовлені або отримані в результаті аналізу

Analysed data — це дані, на основі яких безпосередньо отримують наукові результати, а також похідні результати обчислень.

Це можуть бути розраховані показники, нормалізовані значення, результати статистичних моделей, тематичні коди інтерв'ю, класифікації, результати моделювання, таблиці для побудови рисунків або набори, з яких отримано конкретні значення, представлені у статті.

Саме на цьому етапі особливо важливо забезпечити зв'язок між **даними** → **методом аналізу** → **результатом у публікації**. Якщо у статті наведено графік, таблицю або статистичний результат, бажано мати можливість встановити, з якого набору та за допомогою якої процедури він був отриманий.

4. Published dataset — набір даних, підготовлений до збереження або поширення

Published dataset — це не просто копія робочої таблиці, завантажена в Zenodo.

Це спеціально підготовлена версія даних, достатньо зрозуміла для іншого користувача й безпечна для поширення. Перед депонуванням необхідно перевірити структуру файлів, метадані, права, персональні та конфіденційні дані, документацію, ліцензію та умови доступу.

Опублікований набір зазвичай супроводжується README-файлом, описом методів отримання та обробки даних, словником змінних або іншою документацією, інформацією про авторів, версію та ліцензію.

При цьому **published не обов'язково означає open**. Якщо первинні дані містять інформацію, яку не можна безпечно відкрити, у репозитарії може бути розміщено анонімізовану або агреговану версію, лише метадані, або запис із контрольованим доступом.

Як це виглядає у різних науках?

Етап	Природничі / технічні науки	Соціальні науки	Гуманітарні науки
Raw data	Оригінальні спектри, SEM-зображення, показники сенсорів, файли вимірювального обладнання	Експорт анкети, аудіозаписи інтерв'ю, первинні транскрипти	Скан-копії архівних матеріалів, корпус первинних текстів, фотографії документів
Processed data	Калібровані вимірювання, видалення артефактів, конвертовані формати	Очищена таблиця, кодування відповідей, псевдонімізовані транскрипти	OCR/транскрибований текст, нормалізований корпус, каталогізовані джерела
Analysed data	Розраховані параметри, апроксимації, результати моделей	Статистичні показники, регресійні моделі, тематичні коди	Частотний аналіз, категоризація текстів, анотований корпус
Published dataset	CSV + вихідні/оброблені файли + README + методологія	Анонімізований dataset + codebook + анкета + README	Дозволений до поширення корпус + metadata + опис принципів відбору

Які дані потрібно зберігати?

Не обов'язково зберігати кожен тимчасовий файл, створений під час роботи. Але потрібно зберегти достатньо інформації, щоб забезпечити наукову прозорість, перевірюваність і, наскільки це можливо, відтворюваність дослідження.

Практичне запитання, яке варто поставити собі:

Якби інший дослідник отримав лише фінальну статтю та ті файли, які я вирішив/ла зберегти, чи зміг би він зрозуміти, звідки взяли результати і як були отримані основні висновки?

Якщо відповідь «ні», ймовірно, потрібно зберегти більше даних або краще їх задокументувати.

Що робити далі?

Якщо ви вже визначили, які дані створюватимете, наступний крок — вирішити, де вони зберігатимуться, як називатимуться файли, хто матиме доступ і що станеться з даними після завершення дослідження.