

ІНСТРУКЦІЯ

підготовки файлів звітів, дисертацій, публікацій і відповідних файлів даних для розміщення в Електронному архіві (репозиторії) і Архіві наукових даних (Dataverse)

1. Загальні положення

1.1. Ця Інструкція визначає вимоги до підготовки файлів наукових матеріалів, супровідних документів і пов'язаних із ними наукових даних перед розміщенням у:

- **Електронному архіві Інституту (Репозиторії)** – для зберігання та публікації наукових і службових документів;
- **Архіві наукових даних Інституту (Dataverse)** – для зберігання, опису, публікації та довгострокового використання наборів наукових даних.

1.2. Інструкція застосовується, зокрема, до:

- звітів про науково-дослідні роботи;
- дисертацій та пов'язаних із ними матеріалів;
- наукових публікацій;
- матеріалів конференцій;
- інших завершених результатів наукової діяльності;
- наборів даних, на яких ґрунтуються відповідні наукові результати.

1.3. Основним принципом розміщення є розділення:

науковий документ / результат дослідження → Репозиторій;

наукові дані та документація до них → Dataverse.

Між записами в Репозиторії та Dataverse створюються взаємні постійні посилання.

2. Загальна схема підготовки і розміщення

Для завершеного наукового результату формується два взаємопов'язані інформаційні об'єкти:

2.1. У Репозиторії розміщується основний науковий документ:

- звіт;
- дисертація;
- стаття;
- тези;
- монографія;
- інший науковий результат.

У метаданих запису зазначається посилання на відповідний набір даних у Dataverse.

2.2. У Dataverse розміщується набір даних, який може включати:

- таблиці первинних або підготовлених даних;
- результати вимірювань;
- зображення;
- результати моделювання;
- статистичні дані;
- програмний код;
- документацію набору даних;
- інші матеріали, необхідні для інтерпретації даних.

У метаданих набору даних зазначається посилання на відповідний документ у Репозиторії.

Таким чином формується зв'язок: **Репозиторій $\rightleftharpoons$ Dataverse**

3. Підготовка матеріалів для Репозиторію

3.1. Основний файл.

Основний науковий документ рекомендується подавати у форматі **PDF/A** – для довгострокового зберігання (додаток А).

Якщо створення PDF/A технічно неможливе, допускається PDF.

Документ повинен:

- відкриватися без помилок;
- не бути захищеним паролем;
- не містити технічних обмежень на читання;
- мати остаточну версію;
- містити коректні назву, авторів та афіліацію;
- відповідати версії документа, зазначеній у метаданих.

Для матеріалів, що пройшли редакційно-видавниче опрацювання та були опубліковані видавництвом, редакцією журналу або іншою видавничою організацією (зокрема наукових статей, монографій, розділів книг, матеріалів конференцій), під час розміщення в Репозиторії необхідно враховувати умови видавця щодо допустимої для відкритого доступу версії документа:

- опублікована версія (Version of Record, published version);
- прийнятий авторський рукопис (Accepted Manuscript);
- препринт (Preprint).

3.2. Найменування файлів документів

Імена файлів повинні бути короткими, однозначними та придатними для автоматизованої обробки.

Рекомендується:

- використовувати латиницю;
- не використовувати пробіли;
- використовувати символ **_** як роздільник;
- не використовувати спеціальні символи **/ \ : * ? " < > |**;
- не включати до імені файла персональні дані пацієнтів.

4. Підготовка наукових даних для Dataverse

Перед завантаженням даних необхідно визначити:

- склад набору даних;
- формат файлів;
- структуру даних;
- рівень доступу;
- наявність персональних або чутливих даних;
- документацію, необхідну для інтерпретації даних.

Кожен набір повинен бути достатньо документованим для розуміння його структури без звернення до автора.

Якщо дослідження включає суттєво різні типи даних, наприклад:

- клінічні дані;
 - біомеханічні дані;
 - медичні зображення;
 - результати математичного моделювання,
- вони оформлюються як окремі набори даних, кожен з яких має свій файл опису.

У записі Репозиторію в такому випадку зазначаються посилання на всі пов'язані набори даних.

4.1. Рекомендовані формати файлів даних

За можливості використовуються відкриті та широко підтримувані формати: Табличні та статистичні дані:

- `.csv` – універсальний текстовий табличний формат;
- `.xlsx` – Microsoft Excel;
- `.ods` – LibreOffice Calc / OpenDocument;
- `.sav` – IBM SPSS Statistics;
- `.dta` – Stata;
- `.RData`, `.rda` – R;
- `.rds` – R;
- `.mat` – MATLAB.

Перевага надається відкритим або широко підтримуваним форматам: `.csv` або `.xlsx`. Файли у форматах спеціалізованих статистичних пакетів доцільно, за можливості, супроводжувати експортованою копією даних у форматах `.csv/.xlsx`. Якщо дані зберігаються у власному форматі статистичного програмного забезпечення (SPSS, Stata або SAS), змінні повинні бути належним чином описані безпосередньо у відповідній програмі. Необхідно, наскільки це підтримується форматом файла, зазначати назви та мітки змінних, мітки категорій, одиниці вимірювання, коди пропущених значень та інші атрибути, необхідні для правильної інтерпретації даних. Наявність такого внутрішнього опису не замінює Codebook, який використовується як незалежна документація набору даних.

Текстові матеріали: `.txt`, `.md`, `.pdf`, `.pdf/a`.

Зображення: `.tif/.tiff`, `.png`, `.jpg`, `.dcm` для медичних зображень.

Статистичні та програмні матеріали: `.R`, `.Rmd`, `.qmd`, `.py`, `.m`, `.txt`;

Інші відкриті текстові формати.

Для спеціалізованих форматів допускається збереження оригінального файла разом із описом програмного забезпечення, необхідного для його відкриття.

4.2. Структура табличного набору даних

Для табличних даних рекомендується принцип:

один рядок = одне спостереження;

один стовпець = одна змінна.

Таблиця не повинна містити:

- об'єднаних комірок;
- декоративного форматування;
- кольору як єдиного носія інформації;
- проміжних заголовків усередині таблиці;
- підсумкових рядків між спостереженнями;
- прихованих значень, необхідних для аналізу.

Назви змінних повинні бути:

- унікальними;
- короткими;
- однозначними;
- стабільними в межах усього дослідження.

Наприклад: `patient_id, age, sex, bmi ...`

Кожне спостереження повинно мати унікальний ідентифікатор.

Наприклад: `id, sample_id, experiment_id`



Ідентифікатор не повинен містити:

- ПІБ;
- номер медичної картки;
- номер телефону;
- адресу;
- інші прямі персональні ідентифікатори.



Таблиця відповідності між внутрішнім кодом дослідження та особою, якщо вона існує, **не розміщується у Dataverse**.

Пропущені значення повинні кодуватися послідовно. Рекомендовано використовувати позначку `NA`. Не рекомендується використовувати для пропусків значення типу: `0`, `999`, та інші нерегламентовані позначки без їх чіткого документування.

4.3. Персональні та чутливі дані

До завантаження в Архів наукових даних (Dataverse) дані повинні пройти перевірку щодо:

- наявності персональних, медичних та інших чутливих даних;
- наявності прямих і непрямих ідентифікаторів;
- ризику повторної ідентифікації учасників дослідження;

- відповідності способу поширення даних умовам інформованої згоди учасників дослідження;
- відповідності рішенням та умовам, визначеним Комітетом з біоетики;
- наявності дозволу на вторинне використання або повторне використання даних, якщо це передбачено;
- відповідності вимогам законодавства, договорів, грантових угод та інших обмежень щодо поширення даних.

⚠ Для досліджень за участю людини перед публікацією або наданням доступу до даних необхідно перевірити, чи охоплює рішення Комітету з біоетики та інформована згода запланований спосіб зберігання, передачі, публікації або повторного використання даних.

Залежно від характеру даних та встановлених обмежень можуть застосовуватися:

- анонімізація;
- псевдонімізація;
- агрегування;
- вилучення окремих змінних або ідентифікаторів;
- контрольований доступ;
- обмеження доступу до окремих файлів;
- ембарго на доступ;
- відмова від публікації окремих даних, якщо безпечне поширення неможливе.

Таблиці відповідності між кодом учасника та його персональними даними, ключі псевдонімізації та інші матеріали, що дозволяють встановити особу учасника, в Архіві наукових даних не розміщуються.

5. Документування даних у DataBook Wizard

Перед завантаженням табличного набору даних його структура описується за допомогою програми DataBook Wizard.

Програма CodeBook Wizard призначена для опису файлу даних у форматах (.xlsx і .csv) для подальшого його розміщення в Dataverse.

Перед початком роботи всі файли даних, які необхідно розмістити в Dataverse розмістить в одній папці.

Програма CodeBookWizard знаходиться за посиланням
https://m0ygzx-0-0.shinyapps.io/DataBook_Wizard/

Запустить програму.

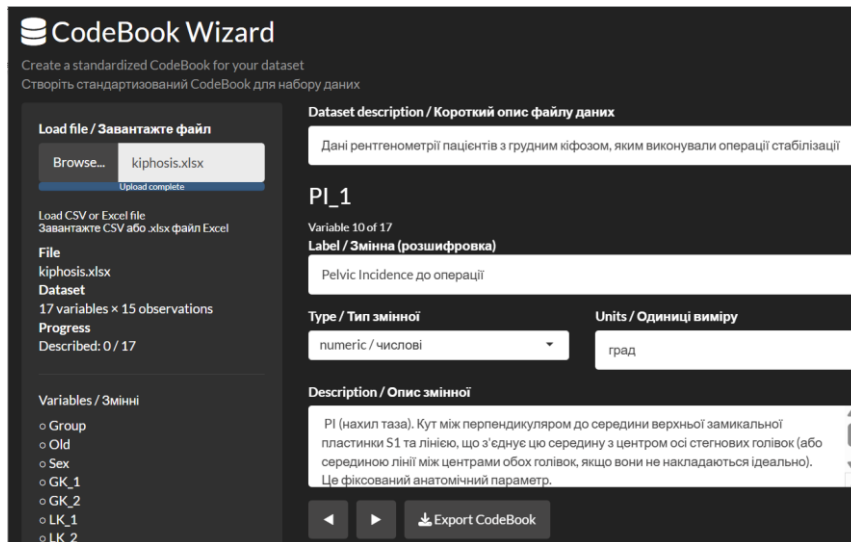


Рис. Базовий вигляд інтерфейсу програми

1.Виберіть файл, дані якого треба описати.

Натисніть кнопку *Browse*, виберіть необхідний файл Excel або .csv

2. Програма виведе основні характеристики файлу даних:

Назва файлу (*File*)

Склад набору даних (*Dataset*)

Перелік змінних (*Variables*)

Лічильник описаних даних (*Described*)

3. У вікні *Dataset description / Короткий опис файлу даних* дати короткий опис набору даних (10-15 слів)

4. В основному вікні програми висвічується назва змінної (на скані - PI_1) для неї у полі *Label / Змінна (розшифровка)* дати більш повну розшифровку даних

6. У спадному меню вибрати тип змінної.

Доступні варіанти:

"numeric / числові",

"nominal / номінальні",

"ordinal / порядкові",

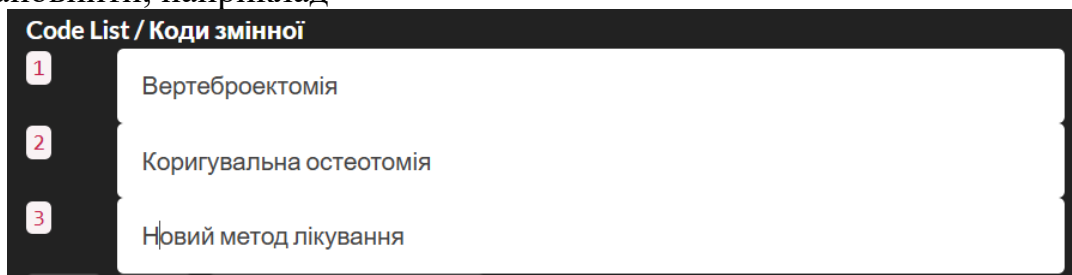
"text / текст",

"date / дата",

"identifier / ідентифікатор".

6.1. Для числових даних заповнити поле *Units / Одиниці виміру*.

6.2. Для номінальних і порядкових даних відкривається спадне меню, яке треба заповнити, наприклад

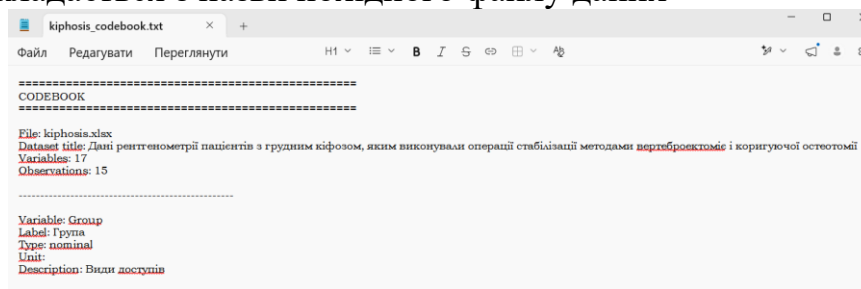


7. Після завершення опису всіх даних натиснути кнопку *Export CodeBook*.



8. Програма сформує звіт.

Назва складається з назви похідного файлу даних



В Dataverse файл *.codebook.txt розміщується разом з файлами даних.

6. Підготовка Readme файла

Файли набору/наборів даних супроводжуються файлом опису Readme.txt.

У файлі вказується

- Назва дослідження
- Автори та їх контактні дані
- Тип дослідження (когортне, РКД тощо)
- Короткий опис, для чого збирались дані

Приклад Readme.txt наведено в додатку Б.

7. Метадані матеріалів для розміщення в Репозиторіях та Dataverse

Матеріали в Репозитарії та Архіві наукових даних (Dataverse) розміщує адміністратор відповідного ресурсу або уповноважена відповідальна особа.

Для передачі матеріалів на розміщення автор або відповідальний виконавець подає файли та супровідну інформацію, необхідну для формування метаданих відповідного ресурсу.

Оскільки записи Репозитарію передаються до Національного репозитарію академічних текстів (НРАТ), супровідна інформація повинна містити:

- назву матеріалу українською та англійською мовами;
- прізвище, ім'я та по батькові авторів українською та англійською мовами;
- ORCID авторів (обов'язково);
- афіліацію авторів;
- опис (анотацію) українською та англійською мовами;
- ключові слова українською та англійською мовами;
- рік публікації або створення матеріалу;
- тип ресурсу;
- бібліографічні відомості або DOI, якщо наявні;
- посилання на пов'язаний набір наукових даних у Dataverse, якщо такий набір створено.

8. Послідовність створення крос-посилань

Якщо постійні посилання формуються лише після створення запису, рекомендується така послідовність:

- підготувати науковий документ;
- підготувати та перевірити набір даних;
- сформувати Codebook у DataBook Wizard;
- створити запис набору в Dataverse;
- отримати DOI або постійний ідентифікатор набору;
- внести посилання на Dataverse до запису в Репозиторії;
- отримати постійне посилання або ідентифікатор запису Репозиторію;
- внести його до метаданих Dataverse;
- перевірити працездатність обох посилань.

9. Особливості різних типів наукових результатів

9.1 Звіт про НДР

В Репозиторій розміщується затверджений звіт та, за необхідності, додатки, які є частиною звіту.

В Dataverse можуть розміщуватися:

- дані, отримані під час НДР;
- результати експериментів;
- таблиці;
- зображення;
- моделі;
- програмний код;
- документація даних.

Одна НДР може бути пов'язана з декількома наборами даних.

9.2 Дисертація

В Репозиторій розміщується остаточний текст дисертації та, за необхідності, автореферат або інші передбачені документи.

В Dataverse можуть розміщуватися

- дані дисертаційного дослідження;
- код статистичного аналізу;
- результати моделювання;
- додаткові матеріали.

Якщо дисертація використовує декілька незалежних наборів, допускається створення декількох записів Dataverse.

9.3 Наукова стаття, книга

В Репозиторій розміщується версія статті, дозволена умовами видавця.

В Dataverse можуть бути розміщені:

- dataset статті;
- додаткові дані, які не увійшли в текст статті;
- код аналізу;
- додаткові таблиці;
- матеріали, необхідні для відтворення результатів.

За наявності DOI статті він також зазначається в метаданих Dataverse.

 Якщо один dataset використовується у декількох роботах, в Dataverse зберігається один основний набір, а в його метаданих зазначаються всі пов'язані публікації.

У відповідних записах Репозиторію наводиться посилання на цей самий dataset.

10. Контроль перед завантаженням

10.1. Перед передачею матеріалів до Репозиторію або Dataverse автор або відповідальна особа перевіряє:

- остаточність версії документа;
- правильність назв файлів;
- можливість відкриття файлів;
- відсутність пошкоджених файлів;
- відповідність метаданих;
- наявність Codebook;
- наявність README за необхідності;
- відсутність прямих персональних ідентифікаторів;
- правильність рівня доступу;
- наявність дозволу на поширення матеріалів;
- відповідність умовам інформованої згоди;
- відповідність ліцензійним та авторським обмеженням;
- коректність посилань між Репозиторієм і Dataverse.

10.2. Мінімальний комплект матеріалів

Для Репозиторію:

- основний документ;
- метадані документа;
- посилання на dataset у Dataverse – якщо пов'язані дані опубліковані.

Для Dataverse

- файл або файли даних;
- Codebook, сформований у DataBook Wizard;
- метадані Dataverse;
- опис умов доступу;
- посилання на пов'язаний документ у Репозиторії.

Для складних наборів додатково:

- README;
- програмний код;
- опис методів підготовки даних;
- додаткова документація.

11. Відповідальність

Автори та відповідальні виконавці забезпечують достовірність переданих матеріалів, правильність структури даних і відсутність інформації, поширення якої суперечить законодавству, етичним вимогам, договорам або умовам інформованої згоди.

Адміністратори Репозиторію та Архіву наукових даних здійснюють перевірку технічної відповідності матеріалів установленим вимогам.

Порядок доступу, зберігання, ліцензування, захисту персональних даних, внесення змін та довгострокового збереження визначається відповідними положеннями і регламентами Інституту.

PDF/A – це спеціалізований варіант PDF, призначений для довгострокового архівного зберігання документів.

Головна відмінність від звичайного PDF полягає в тому, що файл PDF/A є максимально самодостатнім, в нього вбудовані шрифти та інші необхідні компоненти.

PDF/A стандартизований серією **ISO 19005**.

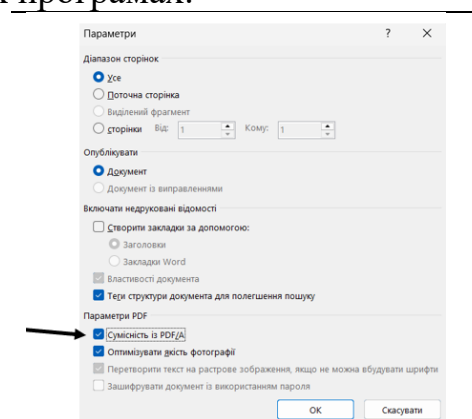
PDF/A краще підходить для довготривалого зберігання, зменшує ризик того, що документ через роки відображатиметься інакше, формат часто використовується в електронних архівах, бібліотеках і репозитаріях.

PDF/A можна створити у звичайних офісних програмах.

– Microsoft Word:

Файл → Зберегти як → PDF → Параметри → Сумісність із PDF/A

В деяких версіях Word можливий пункт **ISO 19005-1 compliant (PDF/A)**.



– **Adobe Acrobat Pro**: можна як створити PDF/A, так і перетворити звичайний PDF у PDF/A та перевірити відповідність стандарту.

– Для пакетної або автоматизованої конвертації можна використовувати **Ghostscript** або спеціальні архівні конвертери.

⚠ Важливо: звичайна команда «Зберегти як PDF» не означає автоматично PDF/A. Треба окремо вибрати режим PDF/A.

Файл *Readme.txt* рекомендується створювати для набору даних у цілому. Він містить загальний опис набору, перелік файлів, їх призначення та взаємозв'язок, особливості організації й використання даних та іншу інформацію, необхідну для правильного розуміння набору.

Для окремого файла даних може створюватися додатковий *Readme*, якщо файл має складну структуру, специфічні умови використання або потребує окремих пояснень.

1. Назва набору

Зазначається повна назва набору даних, що відображає зміст дослідження або його частини.

Якщо README створюється для окремого файла, додатково зазначається назва відповідного файла.

Якщо набір містить дані, пов'язані зі звітом, дисертацією, публікацією або іншою науковою роботою, назва набору може відповідати назві роботи або стисло відображати зміст представлених даних.

2. Короткий опис

Наводиться коротка характеристика даних, зокрема:

- які типи даних містить набір: клінічні, демографічні, лабораторні, біохімічні, біомеханічні, інструментальні, дані моделювання тощо;
- до якого дослідження або його етапу вони належать;
- скільки файлів даних входить до набору;
- за необхідності – кількість спостережень і змінних у кожному табличному файлі.

Детальний опис змінних, їх типів, одиниць вимірювання, кодів категорій, допустимих і пропущених значень наводиться у відповідному файлі **Codebook**, сформованому за допомогою DataBook Wizard.

3. Структура набору

Наводиться перелік файлів, що входять до набору, із коротким поясненням їх призначення.

Наприклад:

- *harris.xlsx* – функціональна оцінка за HHS;
- *harris_codebook.txt* – Codebook для файла *harris.xlsx* ;
- *muscles_force.xlsx* – результати динамометрії;
- *muscles_force_codebook.txt* – Codebook для файла *muscles_force.xlsx*;
- *analysis.R* – програмний код статистичного аналізу;
- *README.txt* – загальний опис набору.

Якщо файли пов'язані між собою за допомогою ідентифікаторів або мають визначену послідовність використання, цей взаємозв'язок необхідно описати.

4. Організація даних

Описується спосіб організації даних у кожному файлі.

Для табличних даних, як правило:

- один рядок відповідає одному спостереженню;
- один стовпець – одній змінній;
- перший рядок містить назви змінних;
- зазначається змінна, що містить унікальний ідентифікатор спостереження,

наприклад ID.

Якщо використовується інша структура даних, вона повинна бути описана окремо.

Зазначається спосіб позначення пропущених значень, наприклад NA.

5. Програмне забезпечення

Зазначається програмне забезпечення, необхідне або рекомендоване для відкриття та використання файлів набору.

Для кожного спеціалізованого формату за необхідності вказуються:

- назва програмного забезпечення;
- версія;
- формат файла.

Якщо набір містить програмний код для аналізу даних або моделювання, зазначаються:

- програмне середовище, наприклад R, Python, MATLAB, SPSS, Stata, SAS, OpenSim тощо;
- версія програмного середовища;
- основні пакети, бібліотеки або модулі;
- за необхідності – послідовність запуску програмних файлів.

6. Етичні та правові обмеження

Для даних, отриманих у дослідженнях за участю людини, зазначається, що дані підготовлено до поширення відповідно до умов дослідження, рішення Комітету з біоетики та умов інформованої згоди учасників.

Рівень доступу до файлів набору визначається з урахуванням характеру даних, етичних і правових обмежень та ризику повторної ідентифікації учасників дослідження.

Для окремих файлів або всього набору можуть встановлюватися:

- відкритий доступ;
- обмежений доступ;
- доступ за запитом після погодження відповідальною особою;
- закритий доступ, якщо поширення даних не допускається;
- ембарго – тимчасове обмеження доступу до визначеної дати;
- обмежений строк доступності, якщо умови дослідження, договору, інформованої згоди або іншого документа визначають період, протягом якого дані можуть бути доступними.

Конкретні умови доступу зазначаються у метаданих Dataverse та, за необхідності, додатково пояснюються у README.

7. Додаткові примітки

Зазначаються особливості набору, які необхідно враховувати під час його аналізу або повторного використання.

До них можуть належати:

- причини систематичної відсутності частини даних;
- зміни методики збору даних;
- особливості окремих груп спостережень;
- виключені спостереження;
- перетворення або попередня обробка даних;
- відмінності між версіями набору;
- інші обставини, важливі для правильної інтерпретації даних.