Чуканова Світлана Олександрівна, канд. педагог. наук,
завідувач сектору Наукової бібліотеки Національного університету
«Києво-Могилянська академія»

УДК 021.61:004.65-047.64

УПРАВЛІННЯ ДАНИМИ ДОСЛІДЖЕНЬ: ВАЖЛИВІСТЬ ЗАСТОСУВАННЯ РЕПОЗИТАРІЇВ ДЛЯ ЗБЕРЕЖЕННЯ ДАНИХ

С. О. Чуканова

Національний університет «Києво-Могилянська академія»
Наукова бібліотека

Анотація: У статті описано важливість застосування репозитаріїв для даних з метою забезпечення належного рівня управління даними досліджень. Показано принцип дії реєстру репозитаріїв для даних та можливості його використання.

Ключові слова: *Управління даними досліджень, репозитарій для даних, реєстр репозитаріїв для даних, куратор даних, пакети даних, цитування даних, життєвий цикл даних.*

З розвитком концепції відкритої науки та вільного доступу до знань набула поширення практика управління даними досліджень, тобто супровід даних певного дослідницького проекту упродовж життєвого циклу даних. Мережа даних спостереження за землею DataOne, як і багато інших наукових установ, має свій підхід до класифікації етапів життєвого циклу даних. Таким чином, ця організація пропонує визначати життєвий цикл даних як певну послідовність етапів, що забезпечують процес

управління даними досліджень і, що може відрізнятись в залежності від технічних можливостей та практик організації даних різних наукових установ [5].

На думку фахівців цієї мережі життєвий цикл даних включає в себе наступні етапи:

- планування дослідження, тобто опис того, які дані будуть зібрані і як керувати ними та надавати до них доступ протягом тривалості дослідження;
- збір даних відбувається або вручну, або посередництвом спеціальних приладів, після чого дані конвертують у цифровий формат;
- забезпечення якості даних відбувається шляхом перевірок та верифікацій;
- опис даних відбувається посередництвом якісних метаданих;
- збереження даних, особливо довготривале, відбувається шляхом розміщення даних у електронному архіві, репозитарії, центрі даних;
- відкриття нової інформації можливе лише за наявності якісного опису даних (метаданих, супровідної інформації);
- інтеграція даних з різних джерел здійснюється для формування однорідного набору даних, який можна легко проаналізувати;
- аналіз даних відбувається для забезпечення результатів дослідження [5] (*тут і надалі переклад з англійської наш – С.Ч.*).

Усі перераховані етапи графічно можуть бути відображені у вигляді кола (Схема 1)

Схема 1. Життєвий цикл даних – версія DataONE [5]

Отже, як бачимо, на етапі збереження виникає потреба у виборі надійного архіву, щоб можна було забезпечити захист даних і у той самий час доступ до пакетів даних. Кетрін МакНіл зазначає, що більшість досліджень може ґрунтуватись на повторному використанні раніше зібраних даних. Таке використання та обмін даними можливий лише за умов ефективного використання репозитаріїв [9, с.15].

Для того, щоб депонувати свої дані до певного репозитарію, науковцеві потрібно проаналізувати, які варіанти збереження будуть найбільш надійними. Зберігати свої дані можна в інституційному репозитарії, тобто архіві, що адмініструється кураторами даних наукової установи, де проводиться дослідження, або ж дані можна депонувати до тематичних репозитаріїв, якщо установа не має свого електронного сховища.

Підібрати тематичний репозитарій можна за допомогою реєстру re3data.org [11]. Цей інструмент є продуктом компанії DataCite [6]. Він використовується на безкоштовній основі і є зведеним реєстром усіх

репозитаріїв даних. Принцип пошуку досить простий: реєстр шукає за такими пунктами:

- предметом;
- типом контенту;
- країною.

Якщо ми шукаємо у реєстрі репозитарій за предметом, то ми можемо вибрати або графічний, або текстовий інтерфейс з такими категоріями, які у свою чергу розбиваються на дрібніші підкатегорії. Основні категорії предметів у реєстрі представлені наступним чином:

- соціо-гуманітарні науки;
- природничі науки (науки про життя);
- природничі науки (точні науки);
- інженерні науки.

Кожна з цих основних категорій поділена на менші складові, які у свою чергу поділені на ще дрібніші елементи. На прикладі соціо-гуманітарних наук простежимо поділ категорій на дрібніші елементи. На прикладі соціо-гуманітарних у схемі 2 відобразимо підкатегорію гуманітарних наук, в якій у свою чергу виокремимо історію стародавнього світу, в якій виокремимо єгиптологію. За подібним принципом відбувається поділ усіх основних категорій, таким чином, науковець може обрати репозитарій для депонування або ж повторного аналізу даних саме з тієї галузі, яка стосується безпосередньо його або її дослідження [11].

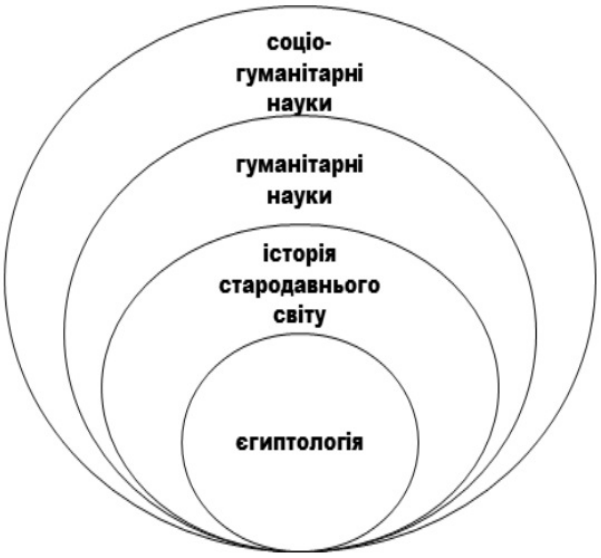

Схема 2. Приклад поділу категорій у реєстрі re3data.org

У разі пошуку за країнами потрібно обрати країну розміщення репозитарію. Наприклад, якщо ми хочемо проаналізувати кількість репозитаріїв даних в Україні, то на інтерактивній мапі платформи ми вибираємо Україну. Таким чином, система показує нам, що в Україні є два репозитарії для даних і належать вони таким установам:

- Відділ космічних інформаційних технологій і систем (Інститут космічних досліджень НАН України та ДКА України).
- Світовий центр даних з геоінформатики та сталого розвитку [1;2;11].

Якщо здійснювати пошук за типом контенту, то система запропонує всі можливі формати даних, які можуть зберігатись у репозитаріях.

Оскільки репозитарії забезпечують збереження даних та доступ до них, необхідно пам’ятати: по-перше, – репозитарій повинен мати ліцензію (Seal of Approval), що підтверджує його надійність, по-друге, – повторне використання даних потребує належного цитування (так само як і друковані матеріали, що мають автора), а це означає, що репозитарій повинен мати інструменти для генерування бібліографічних описів для пакетів даних. Ініціативна група Data Citation Synthesis Group розробила

узагальнення щодо цитування даних [4], які можна відобразити наступним чином, як це показано у таблиці1 [3].

Таблиця 1. Принципи цитування даних від Data Citation Synthesis Group

Importance – важливість	Дані є такими самими продуктами дослідження як і наукові статті і мають бути цитовані за тим же принципом.
Credit and Attribution -атрибуція, визнання	Інструменти цитування даних повинні забезпечувати належне цитування даних та згадування усіх розробників, причетних до створення пакетів даних.
Evidence - доказ	У разі, якщо гіпотеза залежить від даних, відповідні дані повинні бути процитовані.
Unique Identification - унікальний ідентифікатор	Цитування даних повинно відбуватись на основі постійного та незмінного методу для ідентифікації, машинорозпізнаваного та визнаного світовою науковою спільнотою.
Access – доступ	Цитування даних повинні забезпечити доступ до даних та відповідних метаданих, документації, кодів та інших матеріалів, що сприймаються як людьми, так і технікою з метою інформованого використання зазначених даних.
Persistence - постійність	Унікальні ідентифікатори, метадані та їх розташування мають бути постійними і перевищувати строк існування самих даних.
Specificity and Verifiability- специфікація та верифікація	Цитування даних повинні забезпечити ідентифікацію, доступ та верифікацію визначених даних, що підтверджують гіпотезу. Цитування або метадані цитування повинні містити інформацію про походження та закріплення, достатню для проведення верифікації того, що дата створення, версія, порція завантажених даних відповідає набору даних, процитованого у роботі.
Interoperability and flexibility	Метод цитування даних повинен бути достатньо гнучким, щоб забезпечити варіативність

інтероперабельність та гнучкість	підходів у різних наукових спільнотах, проте у той же час не відрізнятись докорінно, аби не порушити інтероперабельність цитування даних відповідно до загальноприйнятих норм.

Забезпечити цитування пакетів даних можна посередництвом он-лайн інструментів, як автономних, так і вбудованих у деякі репозитарії для даних. Наприклад, платформа Відкритої Науки Open Science Framework (OSF), до якої можна приєднатись посередництвом створення аккаунту або номеру ORCID, надає можливість своїм користувачам шукати необхідні дані, депонувати власні дані, створювати унікальні цифрові ідентифікатори DOI для пакетів даних, створювати бібліографічні описи у міжнародних форматах [10].

Прикладом автономного інструменту для створення посилань на дані є продукт від розробників реєстру репозитаріїв DataCite під назвою DOI Citation Formatter [7]. Принцип роботи цього інструменту полягає у тому, що маючи DOI пакету даних, який необхідно процитувати, науковець має ввести його у відповідне поле інструменту, вибрати необхідний стиль цитування, вибрати мову та країну і згенерувати посилання.

Насамкінець варто зазначити, що ефективність роботи з репозитаріями залежить від якісного курування даними і є прерогативою бібліотекарів по роботі з даними – новою спеціальністю, що виникає наразі на теренах Європи та США. Оскільки навчальні дисципліни та й сама спеціальність досить нові, багато інформаційно-бібліотечних фахівців займаються самоосвітою посередництвом он-лайн курсів та платформ. У мережі Інтернет розповсюджено низку рекомендацій та путівників, що стосуються різних аспектів роботи бібліотекаря-куратора даних. Особливої уваги заслуговує портал, розроблений Единбургським Університетом під назвою Research Data MANTRA [8]. Цей портал містить інформацію про: дослідницькі дані, формати файлів пакетів даних, забезпечення доступу та безпеки даних, створення планів управління дослідницькими даними, створення метаданих, ліцензування, упорядкування даних, збереження даних, проведення власних тренінгів тощо.

Таким чином, застосування репозитаріїв для даних є важливим аспектом здійснення практики управління даними досліджень як для бібліотечних фахівців, так і для дослідників. Репозитарії для даних забезпечують захист даних, можливість цитування даних і збереження авторського права, а також надають можливість забезпечити доступ до даних, на яких базуються результати досліджень, і, таким чином, розповсюдити інформацію серед зацікавлених аудиторії та ключових стейкхолдерів в контексті певного дослідження. Дані дослідження повинні відповідати вимогам концепції FAIR (Findable, Accessible, Interoperable, Reusable), тобто бути такими, які можна знайти, отримати доступ до них, відкрити доступними інструментами та повторно використати.

Література:

1. Відділ космічних інформаційних технологій і систем Інституту космічних досліджень НАН України та ДКА України : [Веб-сайт]. 2020. URL: http://inform.ikd.kiev.ua/?path=/en/index (дата звернення: 03.04.2020).
2. Світовий центр даних з геоінформатики та сталого розвитку : [Веб-сайт]. Київ, ІПСА, НТУУ «КПІ», 2015. URL: http://wdc.org.ua/ (дата звернення: 03.04.2020).
3. Чуканова С. О. Як уникнути плагіату при посиланні на пакети даних: [презентація] // eKMAIR. Наукова комунікація в цифрову епоху : 7-а Міжнар. наук.-практ. конф. Київ, 2019. URL: http://ekmair.ukma.edu.ua/handle/123456789/15359 (дата звернення: 04.04.2020).
4. Data Citation Synthesis Group: Joint Declaration of Data Citation Principles / за ред. Martone M. // FORCE11. San Diego CA: FORCE11, 2014. URL: https://www.force11.org/datacitationprinciples (дата звернення: 04.04.2020).
5. Data Life Cycle // DataONE : [Веб-сайт]. URL: https://www.dataone.org/data-life-cycle (дата звернення: 04.04.2020).

6. DataCite : [Веб-сайт]. URL: https://datacite.org/ (дата звернення: 04.04.2020).

7. DOI Citation Formatter : [Веб-сайт]. 2020. URL: https://citation.crosscite.org/ (дата звернення: 03.04.2020).

8. MANTRA: [Веб-сайт]. The University of Edinburgh, 2019. URL: https://mantra.edina.ac.uk/ (дата звернення: 04.04.2020).

9. Mcneill K. Repository Options for Research Data // Making Institutional Repositories Work / за ред. Burton B. Callicott, David Scherer, and Andrew Wesolek. West Lafayette, Indiana, 2016. Розд. 2. С. 15–17.

10. OSF : [Веб-сайт]. Center for Open Science, 2020. URL: https://osf.io/ (дата звернення: 02.04.2020).

11. Registry of Research Data Repositories : [Веб-сайт]. URL: http://re3data.org/ (дата звернення: 03.04.2020).

УПРАВЛЕНИЕ ДАННЫМИ ИССЛЕДОВАНИЙ: ВАЖНОСТЬ ИСПОЛЬЗОВАНИЯ РЕПОЗИТАРИЕВ ДЛЯ ХРАНЕНИЯ ДАННЫХ

С. А. Чуканова

Национальный университет «Киево-Могилянская академия»
Научная библиотека

Аннотация: В статье описана важность применения репозиториев для данных с целью обеспечения надлежащего уровня управления данными исследований. Показано принцип действия реестра репозиториев для данных и возможности его использования.

Ключевые слова: *Управление данными исследований, репозитарий для данных, реестр репозиториев для данных, куратор данных, пакеты данных, цитирование данных, жизненный цикл данных.*

RESEARCHDATA MANAGEMENT: THE IMPORTANCE OF DATA REPOSITORIES USAGE

S. Chukanova

The National University of Kyiv-Mohyla Academy
University Library

Annotation: The article describes the importance of using data repositories in order to ensure an appropriate level of Research Data Management. The operating principle of the data repository registry re3data.org and the possibility of its use.

Key words: *Research data management, data repository, data repository registry, data curator, data packages, data citation, research data life cycle.*