УДК 378.

Л. Ф. Панченко,
доктор педагогічних наук, професор
lubov.felixovna@gmail.com
Н. О. Самовілова,
аспірант
(ДЗ "ЛНУ імені Тараса Шевченка", м. Старобільськ)
nata15101970@gmail.com

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ЯК ФАКТОР ПРОЗОРИХ ТА ВІДТВОРЮВАЛЬНИХ ДОСЛІДЖЕНЬ

Анотація

У статті обговорюються питання підготовки майбутніх науковців до планування і проведення прозорих та відтворювальних досліджень, що характеризуються достовірними результатами, надають можливості для подальших розвідок з цієї тематики, підвищують якість аналізу та інтерпретації отриманих даних. Аналізуються інструменти інформаційно-комунікаційних технологій щодо прозорості наукових досліджень, які стануть корисними студентам педагогічних спеціальностей у процесі виконання навчально-дослідницьких проектів у вишах та в майбутній науковій діяльності.

Ключові слова: інформаційні технології, відтворювані дослідження, прозорі дослідження, середовище R.

Summary

The article discusses the issues of preparing future scholars for planning and conducting transparent and reproducible research. Such studies are characterized by reliable results, provide opportunities for further research on this topic, improve the quality of data analysis and data interpretation. The tools of information and communication technologies for the transparency of scientific research, which will be useful for students of pedagogical specialties in the process of conducting research and development projects in higher education and in future scientific activities are analyzed.

Key words: informational technology, reproducible research, transparent research, R, data analysis, environment R

Постановка проблеми. Наука зіткнулася з кризою довіри до досліджень у зв'язку з труднощами їх відтворювання [6; 8; 11]. Так, за джерелом [11] в 2011р. дослідники Bayer повідомили про змогу відтворення лише 17 з 67 важливих фармацевтичних досліджень; у 2012 р. Amgen виявив, що лише 6 з 53 досліджень на рак мали результати, які вони могли відтворити; у серпні 2015 р. проект "Відтворюваність" у Вірджинії повідомив, що вони змогли відтворити лише 39 з 100 досліджень психології; у вересні 2015 р., за даними Федеральної резервної системи США, було відтворено лише 29 з 67 економічних досліджень.

Прозорі та відтворювальні дослідження збільшують підстави в достовірності результатів, розширюють можливості для подальших пошуків з цієї тематики, підвищують якість аналізу та інтерпретації отриманих даних [1-3; 6; 7]. Про актуальність означеної тематики свідчить ряд масових онлайн курсів та ініціатив [12-14; 16].

Аналіз останніх досліджень та публікацій. Науковцями розглядаються питання прозорості наукових досліджень: так, R. D. Peng вивчає питання відтворювальних досліджень з точку зору комп'ютерних наук та аналізу даних [8]; J. F.Robert, W.Johnson [15], G.Christensen [6] розглядають стандарти та кращі практики відтворювальних досліджень для соціальних наук: Lucas C. Coffman, Muriel Niederle [3] присвятили свої публікації складанню плану попереднього аналізу даних дослідження. Аналіз робіт науковців показує, що більш інтенсивно ця тематика розвивається науковцями в галузі психологічних, соціальних наук, галузі Data Science, а питання підготовки майбутніх педагогів-дослідників до проведення прозорих та відтворювальних досліджень залишилися, на жаль, поза увагою науковців.

Мета статті – проаналізувати інструменти інформаційно-комунікаційних технологій щодо прозорості наукових досліджень, які стануть корисними студентам педагогічних спеціальностей у процесі виконання навчально-дослідницьких проектів у вишах та в майбутній науковій діяльності.

Виклад основного матеріалу. У рамках цієї публікації розглянемо декілька інструментів: Open Science Framework (OSF), масові відкриті онлайн курси з цієї тематики, можливості середовища R.

Open Science Framework [7] – це вільний ресурс, який підтримує повний цикл дослідження: планування, виконання, створення звітів, архівування документів і т. і. За технологією прозорих досліджень створюється план управління даними, який пояснює, що дослідник збирається робити з власними даними підчас проекту та після нього. У плані описується джерело для кожного набору даних, форма датасетів (текст, числові дані, моделі, комп'ютерний код, аудіовізуальні дані і. т. д.). Далі створюється словник, призначення якого зробити дослідження відтворюваним, який допомагає іншим дослідникам зрозуміти дані цього дослідження. У словникові наводяться імена змінних, одиниці вимірювання, допустимі значення, визначення змінних тощо.

Дані досліджень можна завантажувати безпосередньо на цей ресурс або користуватися іншими. Так, Open Science Framework підтримує для зберігання даних такі сервіси, як Amazon S3 (5 Гб), Box (10 Гб), Dataverse (1Тб), Dropbox (2 Гб), Figshare (20 Гб), Owncloud, Github, Google Drive (15 Гб), для управління цитуваннями – Mendeley та Zotero.

Студенти можуть проводити пошук вже розшарованих досліджень, наприклад, за темою "емоційний інтелект", на 19.07.2017 р. можна знайти 308 проектів, 259 файлів, 155 реєстрацій, 84 препринти, 34 компоненти (рис.1). В одному з винайдених досліджень йдеться про академічні емоції, які притаманні студентам у процесі навчання в вишах; в дослідженні використовується фінський опитувальник щодо академічних емоцій, мережевий аналіз та факторний аналіз для опрацьовування даних (https://osf.io/x29gh/files/).

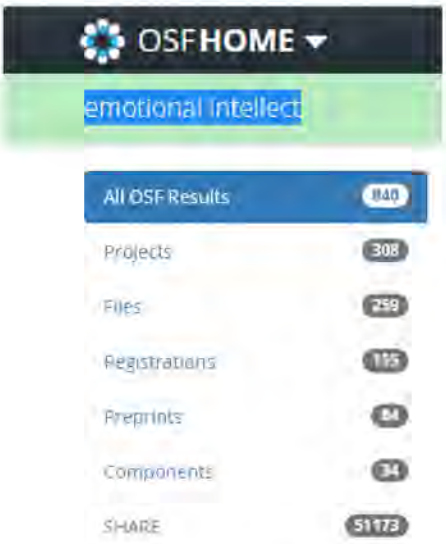

Рис.1. Результати пошуку досліджень за темою "емоційний інтелект" в OSF

Center for Open Science [2], який є партнером ініціативи Берклі в проекті прозорості досліджень [1] надає також вільні консультації для науковців щодо використання статистичних методів, які включають проведення аналізу потужності, проведення метааналізу, використання вільного середовища R, використання OSF, створення плану попереднього аналізу, створення робочих матеріалів для підвищення прозорості.

На допомогу науковцям пропонується шаблон плану попереднього аналізу [9]. Наведемо фрагмент шаблону, який стосується польового етапу збору даних і включає питання щодо інструментів, збору та обробки даних.

Інструменти. Які інструменти збору даних використовуватимете? Які (групи) показників будуть охоплювати кожен інструмент? Як розроблявся кожен інструмент? Чи використовувався кожен інструмент раніше? Якщо так, хто? Якщо ні, то чи проводився пілотаж цього інструмента? Які основні переваги / недоліки кожного інструмента?

Збір даних. Як довго триватиме весь процес збору даних від початку до кінця? Що спричиняє збір даних? Які кроки потрібно вжити, щоб зберігати зібрані дані на цьому етапі?

Обробка даних. Скільки часу займає обробка даних від початку до кінця? Що таке обробка даних? Які кроки потрібно вжити, щоб зберігати конфіденційність оброблених даних? Хто володіє обробленими даними? Як дані будуть використані / збережені після вивчення на цьому етапі?

Розглянемо масові відкриті он-лайн курси з тематики прозорості досліджень, фрагменти яких можна застосовувати в рамках змішаного навчання студентів вишів з метою підготовки останніх до реалізації концепцій прозорості та відтворюваності у власних дослідженнях. Ми узагальнили деякі курси, що пропонуються на платформах Coursera та Future Learn у таблиці.

Таблиця 1

Масові відкриті он-лайн курси з тематики прозорості досліджень

Назва курсу / платформа	Частина спеціалізації	Програмне забезпечення, сервіси	Теми
Transparent and Open Social Science Research / Future Learn, 2017	–	OSR, Stata, R.	Введення до прозорих та відкритих досліджень: план попереднього аналізу та реєстрації; стан та майбутнє відкритої науки
Reproducible research / Coursera, 2017	Data Science	R, RStudio	Концепції, ідеї та структура. Markdown та knitr як засоби для розвитку відтворювальних звітів про дослідження. Checklist щодо доказового аналізу даних.
Communicating Data Science Results / Coursera, 2017	Data Science at Scale	Elastic Map Reduce, Pig, Amazon Cloud Service	Візуалізація, приватність та етика, відтворюваність та хмарні обчислення

Курс "Reproducible research" [12] є частиною спеціалізації Data Science, яка включає 10 курсів, запропонована Johns Hopkins University на платформі Coursera. Курс фокусується на концепціях та інструментах, що дозволяють відтворювати сучасний аналіз даних. У відтворюваних дослідженнях наукові твердження публікуються разом із зібраними даними та програмним кодом, таким чином, інші дослідники. можуть перевіряти отримані дані та спиратися на них. Потреба у відтворюваності різко зростає, оскільки аналіз даних стає все більш складним, включаючи великі набори даних та більш складні обчислення.

Автори курсу "Communicating Data Science Results" [4] зазначають, що відтворюваність важлива не тільки для вчених: аналітики, повинні вміти обмінюватися даними, пояснювати їх, захищати свої методи та дані від несанкціонованого доступу. У цьому курсі вивчається важливість відтворюваних досліджень та того, як хмарні обчислення пропонують нові механізми для спільного кодування даних, середовищ і навіть витрат, що є критичними для практичної відтворюваності. У курсі використовується хмарні сервіси Amazon, Elastic MapReduce та мова Pig для виконання аналізу графа датасету обсягом біля 600 Гб. Хмарні обчислення пропонуються як дружня зрозуміла платформа для відтворювальних досліджень із такими властивостями:

1. Незважаючи на масштаб та складність даних, для відтворювальних досліджень можна застосовувати формулу:

Data + Code + Environment + Platform ("Моя лабораторія в коробці").

2. Віртуальні машини (і контейнери) можуть бути збережені, розшарені, процитовані, вони є виконуваним середовищем для ваших експериментів.

3. Для дата-інтенсивних експериментів немає іншого вибору, ніж розшарити ресурси даних.

Зауважимо, що курс "Communicating Data Science Results" є кінцевим курсом спеціалізації, попередні два курси присвячені маніпуляції

з даними, методам та моделям передбачення: 1) Data Manipulation at Scale: Systems and Algorithms; 2) Practical Predictive Analytics: Models and Methods.

Курс "Transparent and Open Social Science Research" [16], який пропонується на платформі Future Learn, підтримує ініціативу Берклі [1] щодо подолання кризи відтворюваності. У ньому обговорюються основні проблеми прозорості та відтворюваності в сучасних соціальних науках, включаючи проблеми шахрайства, збігу публікацій та виявлення нечесних даних. Пропонуються варіанти вирішення цих проблем, як-от: попередня реєстрація дослідження та складання плану попереднього аналізу; виконання реплікацій, проведення мета-аналізів, відкритий доступ до даних, чесні та ефективні візуалізації.

У рамках ініціативи Берклі також пропонуються такі програмні засоби інформаційних технологій для використання у реалізації прозорих досліджень: Swirl, Dataverse, Git Version Control, Zotero. Розглянемо їх призначення [1].

Swirl – це програмний пакет для мови програмування R, який перетворює консоль R в інтерактивне навчальне середовище. Користувачі отримують негайний зворотній зв'язок у рамках самостійних уроків у галузі інформатики та програмування на R.

Dataverse є веб-додатком для обміну, збереження, цитування, вивчення та аналізу даних досліджень. Його використання полегшує доступ до даних іншим науковцям і дозволяє відтворювати роботу попередників. У кожному сховищі Dataverse розміщується декілька версій даних; кожен набір даних містить описові метадані та файли даних (включаючи документацію та код, що супроводжує дані).

Open Science Framework (OSF), про яку ми вже згадували, є частиною системи контролю версій і програмного забезпечення, що дозволяє дослідникам переміщувати навчальні матеріали в хмару, ділитися та знаходити матеріали, зробити дизайн дослідження більш прозорим, зареєструвати матеріали для сертифікації дизайну дослідження. Для збільшення гнучкості робочого процесу OSF пропонує структуровану відкриту систему, в якій дослідники можуть опублікувати опис свого дослідження та його цілі. OSF поєднує універсальність з широким діапазоном інструментів та функцій, включаючи оголошення з інших пов'язаних сайтів: datavlesh і github.

Git – це безкоштовна та широко використовувана система керування версіями файлів. Система дозволяє дослідникам зберігати, відслідковувати та відновлювати різні версії своїх файлів у так званих сховищах Git. Програмне забезпечення Carpentry пропонує корисні підручники для навчання управління версіями за допомогою Git. Github – добре розроблений і популярний хост для сховищ Git, який також пропонує графічну оболонку для керування сховищами. Він використовується для обміну файлами проекту та співпраці в його рамках. Для вивчення способу використання Github пропонуються детальні підручники – Github Guides.

Zotero (https://www.zotero.org/) відноситься до бібліографічних менеджерів. Програмне забезпечення дозволяє користувачеві збирати,

упорядковувати та організовувати інформацію для власної дослідницької роботи з різних типів джерел: статті в форматі PDF, текстові файли, веб-сторінки і.т.і.). Zotero може зберігати бібліографічну інформацію з таких ресурсів, як Google Scholar, Google Books, Amazon.com, Wikipedia, тощо. Локальну копію джерел можна зберігати та додавати до них власні теги, нотатки, мета дані.

Вважаємо за необхідне знайомити з методикою відтворювальних досліджень майбутніх педагогів-науковців, зокрема в рамках курсів з методології та методів дослідження, з інформаційних технологій. Корисним інструментом для цього може виступити вільне середовище статистичного моделювання R, яке широко застосовується наразі для підготовки соціологів [10]. Зазначимо, що курси спеціалізацій з аналізу даних проекту Coursera та проведення відтворювальних досліджень практично всі побудовані із використанням R.

Важливою рисою середовища R для дослідників є можливість створювати динамічні звіти на основі даних, які збираються та опрацьовуються. Пакети середовища R (knitr, rmarkdown) дозволяють комбінувати статистичний аналіз та презентацію результатів в одному документі формату pdf або html. Для побудови звіту необхідні знання основ мов розмітки HTML, Markdown, LaTeX. Оболонка RStudio використовується як дружнє інтегроване середовище, що об'єднує R, різноманітні пакети та мови розмітки. У хмарних сховищах та Github можуть зберігатися дані, коди для їх обробки, презентаційні файли, їх попередні версії, тобто все, що робить дослідження прозорими та відтворювальними.

Наш досвід викладання вільного середовища R та R Studio студентам та магістрантам різного фаху (інформатика, соціологія, хімія,) свідчить про можливість опанування його основами в рамках курсів з інформаційних технологій та методів дослідження [17-20].

Висновки. Вважаємо, що програми курсів з наукових досліджень для магістрів та майбутніх PhD українських університетів, програми науково-дослідної практики доцільно доповнити навчальним модулем з розробки відтворювальних та прозорих досліджень, який буде фокусовано на сутності та принципах відтворювальних досліджень, розробці планів попереднього аналізу даних, реєстрації дослідження, управління його матеріалами та даними і кодами для їх обробки за допомогою засобу контролю версій Github, ресурсів Open Science Framework, Dataverse та інших, використання середовищ R та R Studio. Шляхи подальшого дослідження – розробка навчально-методичного забезпечення такого модуля.

ЛІТЕРАТУРА

1. Berkeley Initiative for Transparency in the Social Sciences [Електронний ресурс]. – Режим доступу: http://www.bitss.org/

2. Center for Open Science [Електронний ресурс]. – Режим доступу: https://cos.io.

3. Coffman, Lucas C., and Muriel Niederle. Pre-Analysis Plans Have Limited Upside, Especially Where Replications Are Feasible // Journal of Economic Perspectives. – 2015. – 29 (3): 81–98.

4. Communicating Data Science Results [Електронний ресурс]. – Режим доступу: https://www.coursera.org/learn/data-results.

5. Dal-Ré, Rafael, John P. Ioannidis, Michael B. Bracken, Patricia A. Buffler, An-Wen

Chan, Eduardo L. Franco, Carlo La Vecchia, and Elisabete Weiderpass. 2014. "Making Prospective Registration of Observational Research a Reality." Science Translational Medicine 6 (224): 224cm1-224cm1. doi:10.1126/scitranslmed.3007513.

6. Garret Christensen. Manual of Best Practices in Transparent Social Science Research. November 14, 2016 [Електронний ресурс]. – Режим доступу: http://www.bitss.org/education/manual-of-best-practices.

7. Open Science Framework [Електронний ресурс]. – Режим доступу: https://osf.io.

8. Peng R. D. Reproducible Research in Computational Science [Електронний ресурс]. – Режим доступу: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3383002.

9. Pre-analysis plan template [Електронний ресурс]. – Режим доступу: http://scholar.harvard.edu/files/alejandro_ganimian/files/pre-analysis_plan_template_0.pdf.

10. R for Reproducible Scientific Analysis: Producing Reports With knitr http://swcarpentry.github.io/r-novice-gapminder/15-knitr-markdown/

11. Reproducibility gold standard [Електронний ресурс]. – Режим доступу: https://www.coursera.org/learn/data-results/lecture/KOraT/reproducibility-gold-standard

12. Reproducible research [Електронний ресурс]. – Режим доступу: https://www.coursera.org/learn/reproducible-research

13. Research Transparency and Reproducibility Training (RT2) [Електронний ресурс]. – Режим доступу: http://www.bitss.org/wp-content/uploads/2015/12/2017-RT2-Report.pdf

14. Research Transparency Methods in the Social Sciences: Ph.D. course [Електронний ресурс]. – Режим доступу: http://emiguel.econ.berkeley.edu/assets/miguel_courses/12/Econ-270D-Syllabus_S15-01-20.pdf

15. Robert J. F, Johnson W. Replication standards for quantitative social science: why not sociology? [Електронний ресурс]. – Режим доступу: http://boydetective.net/docs/freese-reproducibility-webdraft.pdf

16. Transparent and Open Social Science Research [Електронний ресурс]. – Режим доступу: www.futurelearn.com/courses/open-social-science-research

17. Панченко Л. Ф. Практикум по анализу данных : учебное пособие для студентов высших учебных заведений / Л.Ф.Панченко // Луганск, Изд-во ГУ "ЛНУ имени Тараса Шевченко", 2013. – 269 с.

18. Панченко Л.Ф. R як інструмент аналізу соціологічних даних / Л.Ф.Панченко, І.В.Левітан // Інформаційні технології в освіті. – Мелітополь : МДПУ, 2014. – С.235-241.

19. Панченко Л.Ф. З досвіду навчання студентів аналізу даних в середовищі R / Л.Ф.Панченко, І.В.Левітан // Матеріали четвертої міжнародної конференції FOSS Lviv 2014. 26-28 квітня 2014. – Львів, 2014. – С.75-76. http://elartu.tntu.edu.ua/bitstream/123456789/16866/1/Matters14_v0.3.pdf

20. Панченко Л. Ф. До питання інструментів відтворення досліджень / Л. Ф. Панченко, Н. О. Самовілова // Держава та глобальні соціальні зміни: Матеріали VII міжнародної науково-практичної конференції. Київ; Одеса, Айпринт, 2016. – С.65–66.

Стаття надійшла до редакції 01.09.2017