BMC Medical Research Methodology

RESEARCH

Open Access

FAIR data management: a framework for fostering data literacy in biomedical sciences education

Rocio Gonzalez Soltero[1*], Debora Pino García[1], Alberto Bellido[1], Pablo Ryan[1,2,3,4,5] and Ana I. Rodríguez-Learte[1]

Abstract

Data literacy, the ability to understand and effectively communicate with data, is crucial for researchers to interpret and validate data. However, low reproducibility in biomedical research is nowadays a significant issue, with major implications for scientific progress and the reliability of findings. Recognizing this, funding bodies such as the European Commission emphasize the importance of regular data management practices to enhance reproducibility. Establishing a standardized framework for statistical methods and data analysis is essential to minimize biases and inaccuracies. The FAIR principles (Findable, Accessible, Interoperable, Reusable) aim to enhance data interoperability and reusability, promoting transparent and ethical data practices. The study presented here aimed to train postgraduate students at the Universidad Europea de Madrid in data literacy skills and FAIR principles, assessing their application in master thesis projects. A total of 46 participants, including students and mentors, were involved in the study during the 2022–2023 academic year. Students were trained to prioritize FAIR data sources and implement Data Management Plans (DMPs) during their master's thesis. An 11-item questionnaire was developed to evaluate the FAIRness of research data, showing strong internal consistency. The study found that integrating FAIR principles into educational curricula is crucial for enhancing research reproducibility and transparency. This approach equips future researchers with essential skills for navigating a data-driven scientific environment and contributes to advancing scientific knowledge.

Keywords Biomedical education, FAIR principles, Data literacy, Data stewardship, Master's thesis, Academic research

Introduction

Low reproducibility in biomedical research studies a widely debated topic [1]. Funding bodies, as the European Commission, are actively fostering the sharing of data as a regular practice and various efforts have been undertaken to increase reproducibility [2]. A fundamental and standardized framework for statistical methods, study design, and data analysis techniques will enable researchers to recognize potential sources of bias, confounding variables, and inaccuracies that could compromise reproducibility.

Data literacy, the ability to interpret, understand, and effectively communicate with data, plays a pivotal role

*Correspondence:
Rocio Gonzalez Soltero
mariadelrocio.gonzalez@universidadeuropea.es
[1]Facultad de Ciencias Biomédicas y de la Salud, Universidad Europea de Madrid, Villaviciosa de Odón, Spain
[2]Internal Medicine Department, Hospital Universitario Infanta Leonor, Madrid, Spain
[3]Faculty of Medicine, University Complutense of Madrid, Madrid, Spain
[4]Centro de Investigación Biomédica en Red de Enfermedades Infecciosas (CIBERINFEC), Instituto de Salud Carlos III (ISCIII), Madrid, Spain
[5]Postgrados Biomédicas, Facultad de Ciencias Biomédicas y de la Salud, Universidad Europea de Madrid, Villaviciosa de Odón, Spain

in mitigating low reproducibility in clinical research by empowering researchers with the skills necessary to critically assess, interpret, and validate data [3]. Ridsdale et al. considered the application of data literacy in educational contexts in 5 parts: conceptual framework, data collection, data evaluation, data management and data application [4].

In 2014, the group called FORCE 11 defined fundamental guiding principles called FAIR (for *Findable*, *Accessible*, *Interoperable*, and *Reusable*) to make scientific data and their metadata interoperable, persistent, and understandable for both humans and machines [5, 6]. These FAIR principles were introduced to tackle the challenges posed by the increasing volume of data generated in academic research in recent years. They serve as a comprehensive guide to optimize the appropriate reusability, collection, annotation, archiving, and management of data [6]. Assessing, managing, visualizing, and sharing data quality also requires an optimal balance between privacy and security provisions. Supporting FAIR principles compliance processes and increasing the human understanding of FAIRness criteria are critical steps in the data sharing process [7]. Following similar objectives, some initiatives as the RDA [8, 9], or the global initiative GO FAIR [10] have been highlighted for fostering open data sharing infrastructures. These two movements have profoundly reshaped the building blocks of data research methodology focused on research integrity and research fairness [11]. In the case of the GO FAIR initiative, they aim to create a structured framework that enhances the discoverability, accessibility, and usability of data, introducing the FAIR principles term and transmitting science researchers the urgent need for enhancing the infrastructure supporting the reuse of research data [6, 12]. The symbiotic relationship between data literacy and fair data strategies could facilitates the cultivation of a data-driven society grounded in integrity, trust, and social responsibility. Fostering data literacy among stakeholders, including researchers, policymakers, and the public, organizations can uphold fair data practices by enabling informed decision-making and promoting ethical data collection, analysis, and dissemination.

Data literacy is a key necessity for medical decision makers, and requires the ability to collect, manage, evaluate, and apply data in a critical manner [13]. Focusing on the reuse of scientific biomedical research data, this critical issue copes with the need of reserving patients' privacy. Patients' data presented in clinical records could be a good source for secondary data research purposes, however, they contain many sensitive personal information and the reuse is not always a possibility because of data privacy concerns [14]. A balance must be established to facilitate access to healthcare data must be accessible while keeping patients' privacy. De-identification and anonymization are the two most common terms used to refer to the technical approaches that protect privacy and facilitate the secondary use of health data and the significance of using precise terminology to describe the process of making those data less identifiable [15]. The increasing use of artificial intelligence urges also the application of strategies in the line of providing high-quality datasets for modelling [16, 17]. Data literacy and FAIR data strategies are intrinsically linked in ensuring ethical and equitable data practices in a clinical and biomedical research environment.

Concerning education, these practices start to permeate. Some health education journals now require or encourage practices like to storage data in a public repository, to guarantee data accessibility through statements and to ensure inclusion of the minimal dataset necessary for interpretation and replication [18]. Nevertheless, only a limited number of initiatives have endeavored to prompt graduate or postgraduate students and mentors to consider these references, aiming for the adoption of more effective research protocols [18, 19]. It seems crucial for the next generation of clinical and biomedical researchers to create a reliable framework rooted in the common good for data literacy and FAIR standards, aiming to generate data management skills to be cultivated within students [20]. To address this, future professionals within the biomedical domain, including academia, industry, management, and editorial roles, must receive training on the proper data literacy practices and data management, through fostering appropriate competencies for data management in accordance with the principles of research integrity and fairness.

Because of all these reasons, an educational innovation project was implemented for postgraduate students in the Biomedical Science and Health School at the Universidad Europea de Madrid. The objective of the study was to train students in data literacy skills, data management competences based on the FAIR principles and to develop a reliable tool to assess the level of FAIRness of research data used during their final master thesis work.

Materials and methods

Study design

A cross-sectional study was conducted to train and assess the students in the use of data literacy practices and the adherence to the FAIR principles during a postgraduate course in Bioinformatics at the European University in Madrid (Spain). The investigation was conducted during the academic year 2022–2023 comprising both students and academics mentors, for those students enrolled in master's theses. A total of 46 participants were recruited, consisting of 31 post-graduate students and 15 mentors with a PhD degree. The course was carried out entirely online. All participants were native Spanish speakers and

were notified about the study through the Large Management System Platform (LMS) Canvas, in the announcement section.

Implementation of data literacy and FAIR principles standards in the curricula

As included in the course syllabus, students were instructed to prioritize the use of data derived from experimental research, collections or repositories that adhered to FAIR data principles and data literacy practices along the curricula.

Students were encouraged to meticulously document the databases and repositories employed to control the source of data if they were for primary or secondary analysis. Student's guidelines include the necessity to report a protocol for collection, storage, and sharing within their research projects in the way of a data management plan (DMP). This DMP included [1] A description of the system(s) used, the data flow, the data management roles, and responsibilities, and [2] methods for back-ups, storage and archiving ensuring anonymization and privacy of data collected as explained in [11]. This research protocol including the DMP was submitted as a pre-task to a committee of 2 evaluators to check the quality of data to be used.

Creation of a questionnaire for assessing data FAIRness

A literature review was conducted to create a self-assessment questionnaire for evaluating the FAIR status of a dataset in a biomedical research study. Three tools were identified:

(1) The ARDC FAIR Data Self-Assessment Tool by the Australian Research Data Commons: This tool allows the determination of the "FAIRness" of a dataset through a series of questions and offers suggestions for improvement if necessary. It is a qualitative multi-response scale: https://ardc.edu.au/resource/fair-data-self-assessment-tool/.
(2) SATIFYD: Offers twelve questions to assess if your datasets comply with FAIR principles and provides advice. It uses a Yes/No response format: https://fairaware.dans.knaw.nl/.
(3) F-UJI: Uses a persistent identifier (such as a DOI) or a dataset URL to verify the extent to which FAIR criteria are met. It is not a scale but an application where a dataset can be analyzed: https://fairaware.dans.knaw.nl/.

None of these scales were validated at the time of the search either use for research purposes After analyzing the above scales, it we decided to adapt some questions from ARDC to a Likert-type scale to allow for the analysis of its internal consistency as part of the pilot following the recommendations from previous studies [21]. The scales vary depending on the number of response options. In all cases, 1 represents the ideal situation, while the highest number reflects being furthest from this ideal situation.

The 11-item questionnaire was implemented in Spanish, the language of instruction for the academic program. Three additional items were included to identify the source of the response (student/mentor) and to evaluate the usefulness of the tool and the video. The scale was established from 1 to 5, with 1 being "not useful" and 5 being "very useful." A comprehensive video was also created to understand the questionnaire and to facilitate the FAIR principles adherence. While the students were familiar with the term, they had not yet fully explored its practical application. The video served as their first in-depth introduction to the principles of data FAIRness, providing a more comprehensive understanding. The questionnaire (Data Dictionary Codebook with the questions and scales) and the video, hosted on the Vimeo platform, provides a step-by-step guide and can be accessed here in the Zenodo platform FAIR data literacy project (zenodo.org) [22].

Statistical analysis

For further analysis, the 11 questions were later grouped into the four attributes of the FAIR data principles (Findable (4 items), Accessible (2 items), Interoperable (3 items), and Reusable (2 items). For the implementation and to facilitate data collection we used the REDCap tool [23, 24]. REDCap is a secure and robust data collection tool created at Vanderbilt University and first conceived by clinical researchers to ensure secure data collection.

The data were exported from REDCap as a .csv file. Statistical analysis was performed with Jamovi statistical software (version 2.3.28.0). The internal reliability analysis of the questionnaire was performed using the 'Reliability Analysis' module within the Jamovi package. This module allows the computation of Cronbach's alpha (α) and McDonald's omega (ω) coefficients, providing insights into the reliability of the measurement instrument. The analysis involved a descriptive statistics item to item, and reliability coefficients. Both coefficients reinforce the reliability of the scale, ensuring that the questionnaire consistently measures the underlying constructs. The interpretation of reliability was based on established thresholds, considering higher values indicative of greater internal consistency [25]. Total score of the different categories were considered for correlation studies.

Table 1 Perception of the utility of the video

Item	Role	N	Median	IQR	Test	p-value
Video utility*	1	15	3	1.50	Mann-Whitney U test	0.082
	2	31	4	2.00		

H_a $\mu_1 \neq \mu_2$ 1: mentor; 2: students.
IQR: Inter Quartile Rank

Table 2 Repositories for collecting open data for student's biomedical projects

Website description	URL	Metadata availability	Type of data	Reference
PubMed/Medline	*pubmed.ncbi.nlm.nih.gov*	Not available	Biomedical literature	[26]
EudraVigilance	*ema.europa.eu*	Not available	Drug safety information in the European Union	[27]
SpainUDP	*spainudp.isciii.es*	Not available	Data about rare diseases studies in Spain	[28]
FEDRA	*notificaRAM.es*	Not available	Pharmacovigilance notifications data in Spain	[29]
The Genotype-Tissue Expression (GTEx) project	*gtexportal.org*	Yes	Gene expression and regulation data in various tissues	[30]
GEO and ArrayExpress	*ncbi.nlm.nih.gov/geo*	Yes	Gene expression and microarray experiment data	[31]
TCGABiolinks NCI Genomic Data Commons	*bioconductor.org*	Yes	Tool for accessing TCGA data through R	[32]

Results

Evaluation of the educational training in data literacy and data FAIRness

The integration of data literacy and FAIR data standards into master thesis projects represents a crucial step towards equipping future researchers with essential skills for navigating a data-driven approach. In our recent study, we introduced these principles to our students facilitating its application within their thesis projects. Following an initial presentation on the subject, a video was created to help the students to understand the data FAIRness concepts. The students and mentors were asked about the utility of the educational program. Although the median in the utility from reported data was different (3 in the case of mentors and 4 in the case of students), the U Mann-Whitney test applied showed no significant differences between groups (Table 1).

When starting their projects, students were encouraged to select a source of data for primary or secondary analysis. Remarkably, our data indicates that 55% of the students opted for primary data sources (as experimental data of retrospective clinical records). Additionally, the remaining 45% successfully located data from secondary open data sources. The list of open data sources is listed in Table 2. All of them share common features aligned with FAIR data principles and metadata availability is evident across platforms such as EudraVigilance, SpainUDP, GTEx, NCBI GEO, TCGABiolinks, Genomic Data Commons, BV-BRC, ENA Browser, Kaggle (Heart Attack Possibility), and PISA 2015 Results. These reported platforms prioritize data accessibility, offering users the capability to retrieve pertinent information. The types of accessible data encompass diverse categories, including drug safety details, disease-specific datasets, and gene expression data (Table 2).

Once the data were obtained, students may implement a DMP. One example of DMP implemented by the students can be found here: https://github.com/Tonibg2/TFMbioinformatica/tree/4f60c11b47eb2e62d7ce51da386745d0e24fff01.

Questionnaire reliability analysis

For performing the reliability analysis of the questionnaire, variables were coded according to the corresponding dimension of the FAIR data principles, supplementary S1: FAIR data literacy project (zenodo.org).

The reliability analysis of the questionnaire resulted in high internal consistency, as indicated by the Cronbach's alpha (α) and McDonald's omega (ω) coefficients. Specifically, the overall scale achieved a Cronbach's alpha of 0.929 and a McDonald's omega of 0.946. These values suggest that the items on the questionnaire are highly correlated and consistently measure the intended constructs.

Data FAIRness analysis

The descriptive analysis of the 11 individual items related to the FAIR data principles is summarized in the Table 3. For assessing the normality of data, Shapiro-Wilks normality test was applied. All 11 items show non-normal distribution. Median and IQR were calculated for each item and described in Table 3. To assess possible differences between students and mentors, and considering the non-asymmetry of the data, the independent samples Mann-Whitney U test were conducted to assess potential differences in FAIR data principles scores between students (coded as 1) and mentors (coded as 2). No significant differences were identified between students and mentors replies.

In the Table 4, results for different categories are summarized. For the "Findable" category, responses to the items varied. In "Findable_1," most responses (45.7%) rated it 4, indicating high findability, with 30.4% rating it 3. For "Findable_2," a significant majority (58.7%) rated it 2, and 41.3% rated it 1, suggesting that this aspect may need improvement. In "Findable_3," most participants rated it 4 (47.8%) and 1 (26.1%), showing a mixed but generally positive assessment. "Findable_4" had varied ratings, with 30.4% at 5 and 32.6% at 2.

In the "Accessible" category, responses were also varied. For "Accessible_1," ratings were highest at 6 (39.1%) and 1 (37.0%), suggesting that while some found the data highly accessible, others did not. "Accessible_2" had a majority rating of 5 (47.8%), followed by 21.7% at 3.

The "Interoperable" category showed significant challenges. For "Interoperable_1," the majority of responses (56.5%) rated it 1, indicating low interoperability, while 41.3% rated it 3. "Interoperable_2" had responses spread across 4 (26.1%) and 1 (37.0%), highlighting the need for better standardization and integration.

In the "Reusable" category, responses were more positive. "Reusable_1" had the highest rating at 5 (54.3%), with 21.7% rating it 1, indicating strong reusability for many of the participants. "Reusable_2" showed varied responses, with 41.3% at 1 and 34.8% at 4.

Discussion

Data literacy facilitates transparent reporting practices, ensuring that methodologies and findings are comprehensively documented and accessible for scrutiny and replication. By equipping researchers with the tools to navigate complex datasets and evaluate the robustness of study results, data literacy acts as a safeguard against

Table 3 Descriptive statistics for the FAIR item's dimensions

Item	Role	N	Median	IQR	Statistic	p-value*
findable_1: does the data set have any identifier assigned?	1	15	4.00	1.00	178	0.175
	2	31	3.00	2.00		
findable_2: *Is the data set identifier included in all records/metadata files that describe the data?*	1	15	2.00	1.00	228	0.913
	2	31	2.00	1.00		
findable_3: *How is data described with metadata?*	1	15	3.00	2.50	227	0.890
	2	31	3.00	2.50		
findable_4: *What type of repository or registry is the metadata record located in?*	1	15	4.00	2.50	218	0.734
	2	31	3.00	3.00		
accesible_1: *To what extent is the data accessible?*	1	15	4.00	4.00	212	0.610
	2	31	5.00	5.00		
accesible_2: *Is the data available online without the need for specialized protocols or tools once access is approved?*	1	15	4.00	2.00	231	0.970
	2	31	5.00	2.00		
interoperable_1: *In what format is the data available?*	1	15	1.00	2.00	218	0.705
	2	31	1.00	2.00		
interoperable_2: *Is the data available online without the need for specialized protocols or tools once access is approved?*	1	15	3.00	2.00	154	0.056
	2	31	2.00	2.00		
interoperable_3: *What best describes the types of vocabularies/ontologies/labeling schemes used to define the data?*	1	15	3.00	0.00	165	0.023
	2	31	3.00	1.00		
reusable_1: *Which of the following best describes the rights of license/use associated with the data?*	1	15	5.00	2.00	210	0.562
	2	31	5.00	4.00		
reusable_2: *How much provenance information has been captured to facilitate data reuse?*	1	15	3.00	3.00	209	0.566
	2	31	5.00	4.00		

*Mann-Whitney U test

Table 4 Distribution of frequencies for FAIR data dimensions

	Metric	Frequency	%From Total
findable_1: *Does the data set have any identifier assigned?*	1	8	17.4%
	2	3	6.5%
	3	14	30.4%
	4	21	45.7%
findable_2: *Is the data set identifier included in all records/metadata files that describe the data?*	1	19	41.3%
	2	27	58.7%
findable_3: *How is data described with metadata?*	1	12	26.1%
	2	2	4.3%
	3	10	21.7%
	4	22	47.8%
findable_4: *What type of repository or registry is the metadata record located in?*	1	3	6.5%
	2	15	32.6%
	3	5	10.9%
	4	9	19.6%
	5	14	30.4%
accesible_1: *To what extent is the data accessible?*	1	17	37.0%
	2	3	6.5%
	4	6	13.0%
	5	2	4.3%
	6	18	39.1%
accesible_2: *Is the data available online without the need for specialized protocols or tools once access is approved?*	1	2	4.3%
	2	6	13.0%
	3	10	21.7%
	4	6	13.0%
	5	22	47.8%
interoperable_1: *In what format is the data available?*	1	26	56.5%
	2	1	2.2%
	3	19	41.3%
interoperable_2: *What best describes the types of vocabularies/ontologies/labeling schemes used to define the data?*	1	17	37.0%
	2	7	15.2%
	3	10	21.7%
	4	12	26.1%
reusable_1: *Which of the following best describes the rights of license/use associated with the data?*	1	10	21.7%
	2	5	10.9%
	3	2	4.3%
	4	4	8.7%
	5	25	54.3%
reusable_2: *How much provenance information has been captured to facilitate data reuse?*	1	19	41.3%
	2	3	6.5%
	3	8	17.4%
	4	16	34.8%

erroneous conclusions and enhances the reliability and reproducibility of clinical research findings. In the same line, fostering FAIR among clinicians and researchers fosters a culture of rigor, transparency, and accountability essential for advancing evidence-based medicine and improving patient outcomes. In an era marked by a growing emphasis on interdisciplinary collaboration, prioritizing FAIR principles becomes crucial for ensuring that data, a fundamental element of scientific inquiry, can be readily located and accessed across diverse academic disciplines.

The cross-sectional study presented here involved 46 participants, comprising 31 post-graduate students and 15 mentors. 55% of students opted for primary data sources and the remaining 45% effectively utilized open data sources, showcasing the value of publicly available datasets in research. Noteworthy repositories like EudraVigilance, SpainUDP, and NCBI GEO align with FAIR

principles, emphasizing data accessibility and robust metadata provision (*European data strategy - European Commission*, s. f.); Table 1). Following data acquisition, students implemented DMPs, ensuring systematic data handling throughout their projects [33].

Regarding the reliability analysis of questionnaire items aligning with FAIR principles, strong internal consistency was observed across dimensions, indicating a reliable assessment tool, showing no significant differences in FAIR principles adherence between students and mentors, suggesting a uniform understanding and application of FAIR principles across the cohort. Internal validation showed a too high α coefficient (close to 0.95) can be a sign of redundancy in the scale items [34]. These findings collectively affirm the internal consistency and reliability of the questionnaire in effectively measuring the targeted FAIR principles.

This study underscores the efficacy of integrating FAIR data principles into educational curricula, fostering proficient data stewardship, and enhancing research reproducibility and transparency. By equipping students with the skills to navigate primary and open data sources while adhering to FAIR principles, educational institutions can empower the next generation of researchers to contribute meaningfully to the advancement of scientific knowledge. Furthermore, the utilization of DMPs ensures the systematic management of research data, promoting data integrity and facilitating data sharing within the scientific community (Kratz & Strasser, 2015). Overall, the findings highlight the importance of incorporating FAIR data principles into academic training to prepare researchers for the data-driven landscape of modern science [35].

These results are in accordance to the data Science and Professionalization Work Package (WP7) from the EU commission project who reported a handbook on good practices in FAIR competence for higher education institutions, providing practical support for universities to integrate research data management (RDM) and FAIR data skills at the bachelor, master, and doctoral levels [36].

The current project endeavors to advocate for best practices in data collection and management aligned with the FAIR principles within the academic workflow. To achieve this goal, a scalable adaptation of the FAIR list has been introduced for the benefit of both students and mentors. The joint Research Data Alliance/World Data System Data Publishing Bibliometrics Working Group aims to 'conceptualize data metrics and corresponding services by investigating current and potential applications for data metrics [37]. The National Information Standards Organization (NISO) Alternative Assessment Metrics Initiative is working to define standards, and best practices for applying altimetric to non-traditional products like software and data [38].

One aspect to highlight is the use of metadata. This practice, although not very extended, provides information about the dataset's content, structure, and context, and hinders its accessibility. Metadata is essential for understanding how to access and use the data. Without clear documentation, users may struggle to interpret the dataset, limiting its accessibility. In this regard, most data in open databases currently lack metadata [39]. The lack of adequate metadata is cited as a barrier to accessibility, making it difficult for other researchers to reuse the data (Tedersoo et al., 2021). Metadata is crucial for interpreting the context, structure, and content of the data. The varied access methods, such as web access, online downloads, and formal agreements, highlight the need for standardization in data access procedures, a point discussed by [40].

According to the FAIR principles, to ensure data accessibility, data should use a standardized protocol, without necessarily having to be open, since sometimes public access is not possible for reasons of privacy, national security, or commercial interests, without detriment that in these exceptional cases the conditions of access are transparent and clear [10]. Data and metadata should be obtainable through their identifiers using a standardized communication protocol. This protocol should be open, freely accessible, and universally implementable, with the option for authentication if necessary. Additionally, metadata should remain accessible even in cases where the original data is unavailable [6]. In relation to data accessibility, users were asked to check in the cross-check questionnaire the type of access to the data and metadata used.

Data interoperability is a concept related to the ways in which data is formatted so that it can be interpreted by a computer to automatically combine with other datasets in a meaningful way. It is a key aspect of the FAIR Data Principles and constitutes the "I" in FAIR. Thus, for data and metadata to use community-agreed formats, languages and vocabularies and contain links to related information through identifiers. Interoperability relies on standardized formats and structures. In our case, for interoperability, the prevalence of structured and open data formats (59.5%) aligns with the recommendations of the FAIR Data Principles (Wilkinson et al., 2016) and the guidelines provided by the Research Data Alliance [37]. The low use of ontologies or global identifiers (11.9%) is consistent with the findings of [41], who emphasized the importance of ontologies in enhancing data interoperability. As for Reusability of the data used, this is compromised when a dataset lacks proper documentation. Users need metadata to comprehend the dataset's purpose, variables, and any specific considerations for analysis. Incomplete or absent metadata reduces the likelihood that others can effectively reuse the data for

different purposes. The limited response of 15 participants highlights the ongoing challenges in establishing clear licensing terms for data reuse, as discussed by [42], encouraging the use of standardized, machine-readable licenses. Licenses, such as Creative Commons, has been proposed by several authors to enhance data reusability [41, 43]. Our results suggest that numerous vital datasets originating from traditional, low-throughput bench science face challenges in fitting into various general-purpose data repositories. These repositories, spanning from institutional (e.g., a single university) to open repositories, accept a wide range of data types in diverse formats. Typically, they do not strive to integrate or harmonize deposited data and impose minimal restrictions (or requirements) on data deposition descriptors. As a result, the evolving data ecosystem is becoming more diverse yet less integrated, heightening the difficulties associated with discovery and reusability for both human and computational stakeholders.

Limitations

It is of relevance that the tool suggested in this project is designed exclusively for self-assessment, so it has some concerns and restrictions. While the self-assessment tool offers valuable insights into students' understanding of the FAIR principles, it should be complemented by external evaluations and feedback mechanisms to provide a more comprehensive assessment of their learning and progress.

Future directions

The tool has been used again in the current academic year, and we plan to compare the results from this academic cycle with those of the previous year. We will evaluate and implement adjustments to refine the tool and improve its performance in future iterations of the questionnaire.

Additionally, the statistical analysis revealed that some questions performed better than others, particularly those related to the interpretation of metadata, which posed challenges for students. To address this, we plan to reinforce the teaching of metadata concepts within the curriculum to ensure a clearer understanding moving forward. Furthermore, to expand the use of this tool to other audiences, it will be important to offer training, particularly for faculty members less familiar with data usage, to ensure that both students and faculty are adequately prepared.

Conclusions

Our findings collectively underscore the importance of promoting FAIR principles in the context of research data [5, 43]. Nevertheless, implementation of the use of a checklist for adherence to FAIR data has resulted a valuable experience, as this introduction sets the stage for a transformative educational experience, wherein students not only engage with the theoretical underpinnings of the FAIR principles but also apply them directly to their Master's Thesis projects, thereby bridging the gap between theory and real-world application in a research. This tool will enable students to assess the quality of their research data against FAIR standards at the initial stages of their future research tasks works when their considered.

In summary, this project marks a pivotal step towards equipping students with a focus in the understanding data literacy and the FAIR principles application. By integrating these principles into the academic workflow, students are poised to enhance the quality and robustness of their research endeavors, aligning with contemporary standards of data stewardship, and promoting a culture of responsible and effective data management.

Acknowledgements

The authors acknowledge Miguel Campoy Ederra its help for technical support in data preparation.

Author contributions

Conceptualization: Rocío González-Soltero; Methodology: Rocío González-Soltero, Débora Pino, Pablo Ryan; Formal analysis and investigation: Rocío González-Soltero, Pablo Ryan, Ana I. Rodríguez Learte; Writing - original draft preparation: Rocío González-Soltero, Anabel R. Learte; Writing - review and editing: Rocío González-Soltero, Pablo Ryan; Resources: Pablo Ryan, Rocío González-Soltero, Débora Pino; Supervision: Rocío González-Soltero.

Funding

This work was conducted without any sources of funding.

Data availability

Gonzalez-Soltero, R. (2024). FAIR data literacy project [Data set]. Zenodo. https://doi.org/10.5281/zenodo.11109044.

Declarations

Ethical approval

The information presented in this paper was gathered as part of an educational innovation project conducted at the Universidad Europea de Madrid (Spain). The methodology employed for data collection underwent ethical review and approval by the UEM Ethical Committee (code: 2023/410). No formal ethics consent was provided for this study. At the beginning of the questionnaire, participants were advised that data will be handled in accordance with European and Spanish data protection guidelines, in compliance with the General Data Protection Regulation (GDPR) of the European Union and Organic Law 3/2018, of December 5, on the Protection of Personal Data and guarantee of digital rights in Spain.

Competing interests

The authors declare no competing interests.

Received: 30 July 2024 / Accepted: 6 November 2024
Published online: 16 November 2024

References

1. Lee RS, Hanage WP. Reproducibility in science: important or incremental? Lancet Microbe 1 de junio de. 2020;1(2):e59–60.

2. European data strategy. - European Commission [Internet]. [citado 4 de mayo de 2024]. Disponible en: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en
3. Griffin GW, Holcomb D. Data Literacy and Skills Development. En: Griffin GW, Holcomb D, editores. Building a Data Culture: The Usage and Flow Data Culture Model [Internet]. Berkeley, CA: Apress; 2023 [citado 4 de mayo de 2024]. pp. 109 – 22. Disponible en: https://doi.org/10.1007/978-1-4842-9966-1_6
4. Ridsdale C, Rothwell J, Smit M, Bliemel M, Irvine D, Kelley D et al. Strategies and Best Practices for Data Literacy Education Knowledge Synthesis Report. En [object Object]; 2015 [citado 3 de mayo de 2024]. Disponible en: http://rgdoi.net/10.13140/RG.2.1.1922.5044.
5. Guiding Principles for Findable. Accessible, Interoperable and Re-usable Data Publishing version b1.0 – FORCE11 [Internet]. [citado 4 de mayo de 2024]. Disponible en: https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/
6. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding principles for scientific data management and stewardship. Sci Data 15 de marzo de. 2016;3(1):160018.
7. David R, Mabile L, Specht A, Stryeck S, Thomsen M, Yahia M, et al. FAIRness literacy: the Achilles' heel of applying FAIR principles. Data Sci J 11 de agosto de. 2020;19:32.
8. RDA | Research. Data Sharing without barriers [Internet]. [citado 15 de enero de 2024]. Disponible en: https://rd-alliance.org/
9. Treloar A. The Research Data Alliance: globally co-ordinated action against barriers to data publishing and sharing. Learn Publish. 2014;27(5):S9–13.
10. GO FAIR [Internet]. [citado 13 de diciembre de 2023]. GO FAIR Initiative. Disponible en: https://www.go-fair.org/go-fair-initiative/
11. Alba S, Lenglet A, Verdonck K, Roth J, Patil R, Mendoza W, et al. Bridging research integrity and global health epidemiology (BRIDGE) guidelines: explanation and elaboration. BMJ Glob Health Octubre De. 2020;5(10):e003237.
12. Mons B, Neylon C, Velterop J, Dumontier M, Da Silva Santos LOB, Wilkinson MD, Cloudy FAIR. Revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services and Use. 1 de enero de 2017;37(1):49–56.
13. Hoffmann I, Behrends M, Consortium H, Marschollek M. Data literacy in Medical Education – An Expedition into the World of Medical Data. Advances in Informatics, Management and Technology in Healthcare [Internet]. En: IOS; 2022. pp. 257–60. https://ebooks.iospress.nl/doi/. https://doi.org/10.3233/SHTI220711. [citado 3 de mayo de 2024].
14. Durneva P, Cousins K, Chen M. The current state of Research, challenges, and future research directions of Blockchain Technology in Patient Care: systematic review. J Med Internet Res 20 de Julio De. 2020;22(7):e18619.
15. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and understanding of anonymization and de-identification in the Biomedical Literature: scoping review. J Med Internet Res 31 de mayo de. 2019;21(5):e13484.
16. Parimbelli E, Soldati F, Duchoud L, Armas GL, de Almeida J, Broglie M et al. Cost-utility of two minimally-invasive surgical techniques for operable oropharyngeal cancer: transoral robotic surgery versus transoral laser microsurgery. BMC Health Serv Res. 29 de octubre de 2021;21:1173.
17. Yang J, Soltan AAS, Eyre DW, Yang Y, Clifton DA. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. Npj Digit Med 29 de marzo de. 2023;6(1):1–10.
18. McLaughlin JE, Tropsha A, Nicolazzo JA, Crescenzi A, Brouwer KL. Moving towards FAIR data practices in Pharmacy Education. Am J Pharm Educ marzo de. 2022;86(3):8670.
19. McMahon C, Houghton J, Wallis K. A brief introduction to creating and sharing FAIR data at UCL [Internet]. Presentation presentado en; University College London; 2021 may 18 [citado 15 de enero de 2024]. Disponible en: https://rdr.ucl.ac.uk/articles/presentation/A_brief_introduction_to_creating_and_sharing_FAIR_data_at_UCL/14612310/1
20. Adjekum A, Blasimme A, Vayena E. Elements of Trust in Digital Health Systems: scoping review. J Med Internet Res 13 de diciembre de. 2018;20(12):e11254.
21. Devaraju A, Mokrane M, Cepinskas L, Huber R, Herterich P, de Vries J et al. From Conceptualization to Implementation: FAIR Assessment of Research Data Objects. Data Science Journal [Internet]. 3 de febrero de 2021 [citado 2 de agosto de 2024];20(1). Disponible en: https://doi.org/10.5334/dsj-2021-004
22. Gonzalez-Soltero R. FAIR data literacy project [Internet]. Zenodo; 2024 [citado 2 de agosto de 2024]. Disponible en: https://zenodo.org/records/11109044
23. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inf Abril De. 2009;42(2):377–81.
24. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inf Julio De. 2019;95:103208.
25. Deng L, Chan W. Testing the difference between reliability coefficients alpha and omega. Educational Psychol Meas 18 de Julio De. 2016;77(2):185.
26. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 7 de enero de. 2022;50(D1):D20-6.
27. Postigo R, Brosch S, Slattery J, van Haren A, Dogné JM, Kurz X, et al. EudraVigilance Medicines Safety Database: Publicly Accessible Data for Research and Public Health Protection. Drug Saf Julio De. 2018;41(7):665–75.
28. López-Martín E, Martínez-Delgado B, Bermejo-Sánchez E, Alonso J, SpainUDP Network PM. SpainUDP: the Spanish Undiagnosed Rare diseases Program. Int J Environ Res Public Health. 14 de agosto de 2018;15(8):1746.
29. Información sobre el acceso a los datos de FEDRA | Agencia. Española de Medicamentos y Productos Sanitarios [Internet]. [citado 22 de enero de 2024]. Disponible en: https://www.aemps.gob.es/medicamentos-de-uso-humano/farmacovigilancia-de-medicamentos-de-uso-humano/informacion-sobre-el-acceso-a-los-datos-de-fedra/
30. GTEx Consortium. The genotype-tissue expression (GTEx) project. Nat Genet junio de. 2013;45(6):580–5.
31. Edgar R, Domrachev M, Lash AE. Gene expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 1 de enero de. 2002;30(1):207–10.
32. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res 5 de mayo de. 2016;44(8):e71.
33. Hart SA. Precision Education Initiative: moving toward Personalized Education. Mind. Brain Educ. 2016;10(4):209–11.
34. Taber KS. The Use of Cronbach's alpha when developing and Reporting Research Instruments in Science Education. Res Sci Educ 1 de diciembre de. 2018;48(6):1273–96.
35. Berman F, Cerf V. Who will pay for Public Access to Research Data? Science. 9 de agosto de. 2013;341(6146):616–7.
36. D7.4. How to be FAIR with your data. A teaching and training handbook for higher education institutions [Internet]. [citado 19 de enero de 2024]. Disponible en: https://zenodo.org/records/5787046
37. RDA | Research. Data Sharing without barriers [Internet]. [citado 15 de diciembre de 2023]. Disponible en: https://www.rd-alliance.org/
38. NISO Alternative Assessment Metrics. (Altmetrics) Initiative | NISO website [Internet]. [citado 19 de enero de 2024]. Disponible en: https://www.niso.org/standards-committees/altmetrics
39. Brown AD. Identity work and organizational identification. Int J Manage Reviews. 2017;19(3):296–317.
40. Smith M, Miller S. A principled approach to cross-sector genomic data access. Bioethics. 2021;35(8):779–86.
41. Rocca-Serra P, Gu W, Ioannidis V, Abbassi-Daloii T, Capella-Gutierrez S, Chandramouliswaran I, et al. The FAIR cookbook - the essential resource for and by FAIR doers. Sci Data 19 de mayo de. 2023;10(1):292.
42. Labastida I, Margoni T. Licensing FAIR data for reuse. Data Intell enero de. 2020;2(1–2):199–207.
43. Holub P, Kohlmayer F, Prasser F, Mayrhofer MT, Schlünder I, Martin GM, et al. Enhancing reuse of Data and Biological Material in Medical Research: from FAIR to FAIR-Health. Biopreserv Biobank Abril De. 2018;16(2):97–105.

Publisher's note