Article

# FAIRness of Research Data in the European Humanities Landscape

Ljiljana Poljak Bilić and Kristina Posavec

## Special Issue

MDPI

*Article*

# FAIRness of Research Data in the European Humanities Landscape

Ljiljana Poljak Bilić [1,2,*] and Kristina Posavec [3]

1    Department of Information Sciences, University of Zadar, 23 000 Zadar, Croatia
2    University of Split Library, University of Split, 21 000 Split, Croatia
3    Data Management Department, University of Zagreb University Computing Centre SRCE,
10 000 Zagreb, Croatia; kristina.posavec@srce.hr
*    Correspondence: lpoljak@unizd.hr

**Abstract:** This paper explores the landscape of research data in the humanities in the European context, delving into their diversity and the challenges of defining and sharing them. It investigates three aspects: the types of data in the humanities, their representation in repositories, and their alignment with the FAIR principles (Findable, Accessible, Interoperable, Reusable). By reviewing datasets in repositories, this research determines the dominant data types, their openness, licensing, and compliance with the FAIR principles. This research provides important insight into the heterogeneous nature of humanities data, their representation in the repository, and their alignment with FAIR principles, highlighting the need for improved accessibility and reusability to improve the overall quality and utility of humanities research data.

## 1. Introduction

Research data play a crucial role in the development of the scientific communication landscape since sharing and reusing data are considered key activities in the transition to open science at the European level. Research data in the humanities are the most diverse of all scientific disciplines because almost any data on human activity can be considered research data, including newspapers, photographs, diaries, church records, court files, etc. [1]. Given the confusion surrounding the definition of data, Borgman believes that it is more relevant to question when something becomes data rather than what data are, because it "usually involves the process by which a scientist recognises that an observation, object, record, or other entity can be used as evidence of phenomena, and then collects, acquires, presents, analyses, and interprets these entities as data" [2].

Data sharing is essential in enabling the transparency, validation, and reproducibility of the research process, and further discovery based on reused data [3]. Therefore, it is possible to say that "the value of data lies in their use" [2]. To enable reuse, data must be shared in a way that is primarily FAIR (Findable, Accessible, Interoperable, Reusable), which ensures their finding, availability, interoperability, and reuse. The FAIR principles provide guidelines for the publication of digital resources in a way that makes them findable, accessible, interoperable, and reusable, and describe a continuum of characteristics, attributes, and behaviours that will bring digital resources closer to this goal [4].

The aim of this paper is to examine and answer three research questions related to datasets in the field of humanities: what types of research data are represented in the humanities? (1), to what extent are datasets in repositories in the field of the humanities represented by openness and under what licence? (2), and to what extent do research data align with the FAIR principles? (3).

In the literature, there is still no consensus on the definition of research data in the humanities, so they are often used as a broad term that includes different types of research

sources. This paper aims to determine the most prevalent data types in the humanities found in repositories.

To evaluate the FAIRness of datasets, this study used Wilkinson et al.'s [5] FAIR guidelines principles and Routledge Open Research data guidelines [6]. The F-UJI tool was used to assess the FAIRness level of datasets. This tool evaluates all FAIR elements using the URL or persistent identifier of a dataset [7]. Using tools to assess the FAIRness of data helps the future improvement of data FAIRness and, therefore, their reuse [8]. The use of the F-UJI tool is already acknowledged in scientific research [9,10]. Petrosyan et al. [9] used the F-UJI tool for the FAIR assessment of agriculture datasets and proposed a methodology that integrates the process of obtaining datasets and their subsequent evaluation, which can be used to investigate the degree of FAIRness of datasets in different disciplines.

## 2. Review of the Literature

The concept of research data in the humanities is still vague, especially in comparison to data-driven fields like STEM. This is due to the unique nature of humanities scholarly primitives [11], which refer to the basic functions that are common in scholarly activity across various disciplines. However, to participate effectively in current research practices, there is a need for a clearer understanding of what research data mean in the humanities.

Research data are often determined by definition as a set of symbols, that is, quantitative or qualitative, and by type, i.e., numerical, textual, and similar (definitions by Baretto, Burrel, Oppenheim) [12]. These characteristics describe research data, i.e., data that have been collected, recorded, or generated for analysis to obtain new, original research results. Each scientific field interprets and groups research data differently, most often according to the method of collection [13] and the level of processing [14].

According to DARIAH-DE [15], research data are all sources, materials, and results that have been collected, recorded, or evaluated in the context of research and answers to research questions in the field of humanities and culture, as well as computer-processed data for permanent storage, citation, and further processing. The definition is deliberately broadened to encompass the particular characteristics of humanities research, as well as the heterogeneous nature of research data [15]. Since there is no consensus on the definition, research data in the humanities field are often used as an umbrella term that includes different types of sources for research instead of the specific material that needs to be researched using quantitative tools [16]. Research data can also differ according to different perspectives, so we distinguish between artefacts intentionally created by people; texts, which are subject to interpretation; and information that is machine-processable and can be quantitatively analysed. However, such a division does not allow for the distinction of research data from research sources. Perhaps the most significant specificity of the humanities, in the context of scientific research data, are the unclear boundaries between data and publications. They are often used as equals, so primary sources are considered research data in the humanities [17]. Borgman describes this in the example of an old book, which can be considered not only a publication, but also a data source for retrieving names, places, and events [1]. According to the results of research by Allen and Hartland [18], in some fields, especially in history, researchers do not even recognise that they use data. It is often not considered a weakness, but a resistance to using the general term "data" when there is a more developed and accepted terminology, such as primary and secondary sources, bibliographies, and notes, and it is possible to describe research processes using them [18]. But quite often, they are not autonomous products of research, but are part of the cultural and social practices of the institutions that preserve, curate, and (co)produce them, such as libraries, museums, and archives [19]. Thus, another particularity of the humanities in this context is that they are not only devoted to physical entities, as in other disciplines, but also to abstract entities that include work as a physical and intellectual manifestation.

## 2.1. Humanities Scientists on Data

In contrast to other disciplines, there is not much research on the perception of research data and generally on the management of research data among scientists in the humanities field. Thoegersen [20] conducts research on research data and its management among scientists in the humanities field, and the results indicate that respondents consider research materials as data and equate them with information. Some distinguish between information and data, in that information is considered everything they have, while data are numerical, that is, quantitative. They believe that when information is used in research, it becomes data, that is, research information. Although the research is interesting, limitations are apparent. The very small sample of nine respondents is perhaps the most significant limitation, so these results can be used as encouragement for further research on this topic. Hauf and coauthors [3], in their research on research data and their storage and use in the social sciences and humanities, intentionally use the term data instead of research data, because this phrase could refer too much to the STEM field and lead respondents (scientists in the social sciences and humanities) on the wrong track. The results indicate that the respondents most often use numerical data, including statistical and other quantitative data; scientific publications are in the second place, surveys and interviews in the third, and digital artefacts are in the fourth place. The correlation between the type of reused data and the discipline has been proven and the use of digital artefacts in the fields of archaeology, art, history, and literature is statistically significant. The fact that scientists from the social sciences and humanistic fields are involved in research partially explains the fact that numerical data are used most often, but it is necessary to investigate this in more detail. Gualandi, Pareschi, and Peroni [4] conducted research with the objective of defining research data in the humanities; that is, to define what that term includes and to provide an overview of current data practices. The research methodology was based on the previously mentioned research conducted by Thoegersen [20] and involved a sample of 19 scientists in the humanities field at the University of Bologna. The research resulted in a list of 13 types of research data, including publications, that is, primary and secondary sources. The respondents viewed the term "data" broadly, encompassing everything necessary for their research, and were not inclined to limit this term. Unlike the results of the research conducted by Thoegersen [20], the respondents did not consider the term "data" problematic and considered it relevant to their research context.

Limited research is available on how humanists conceptualise research data, and the studies that have been conducted are characterised by small sample sizes, making it hard to draw any conclusions. It should be noted that the term data itself is often problematic in the context of humanistic research.

## 2.2. Previous Research on FAIRness in Humanities

Quite often, the particularities regarding research data in the humanities make the implementation of the FAIR principles challenging. Kalinin and Skvortsov [21] identified four key characteristics of humanities data usage: 1. Research objects are collected, not created [22,23]; 2. Analogue data are critical; 3. Licence restrictions are imposed by data owners (libraries, galleries, and museums, not scientific institutes); 4. Context is of great importance [22]. Therefore, research data come in various forms, making it challenging to interpret and reuse them in other research once they are removed from their initial context.

The FAIR concept was the subject of research in different humanities areas. Calamai and Frontini [24] examined current data science and archiving practices in the area of speech and oral archives in the humanities. All FAIR principles were analysed and explained from the oral and speech archive perspective, i.e., the FAIR principles were applied to the CLARIN RI. Furthermore, de Jong et al. [25] further demonstrated the compliance of the CLARIN infrastructure with the FAIR concept through its data architecture. Strange, Gooch, and Collinson [26] investigated the question of data sustainability and existing repository services on the institutional level at the University of Oxford. Hiebel et al. [27] presented an approach to creating FAIR data for prehistoric mining archaeology based

on ontology and semantic web standards. They demonstrated that the creation of FAIR data concerns all the sub-principles and particularities of the field. One of the main implications for practice is that the demonstrated approach is generic and can be applied to any archaeological investigation carried out in Austria. Some cross-domain research studies aimed to incorporate research data from various fields, including the humanities [18,21]. Despite accounting for critical differences between disciplines, such as the diversity of data types, research infrastructures (RI), and attitudes towards sharing and perceived individual benefits [18], field particularities and lack of representation have resulted in positioning the humanities, understood as non-STEM areas such as the Social Sciences and Humanities (SSH) [3,28] and Arts and Humanities, as an area of science unrepresented in cross-disciplinary research [21].

Tóth-Czifra [22] found that the FAIRness of humanities research data is based on particularities of humanities compared to other fields: "First, scholarly data or metadata are digital by nature, second, scholarly data are always created and therefore owned by researchers, and third, there is a wide community-level agreement on what can be considered scholarly data. The problems around such assumptions in arts and humanities are cornerstones in reconciling disciplinary traditions with FAIR data management. By addressing them one by one, our aim is to contribute to a better understanding of discipline-specific needs and challenges in data production, discovery, and reuse".

From previous research on the FAIR concept in the humanities, it can be observed that most of the research is focused or limited to Ris (e.g., CLARIN, DARIAH, SDS Oxford University services), with an emphasis on the infrastructure that supports FAIRification [29], not on the data themselves. But the importance of reaching a consensus around the notion of "data" [2,19,20] is a stepping stone for representable humanities research data in the open science landscape.

The question of determining what data are in the humanities is very current, reflecting challenges in distinguishing between data and publications. In addition to this problem, there are other specificities of research data in the humanities which need to be further investigated. The literature notes a lack of research on data in the humanities, including their definition and the adequacy of general concepts such as the FAIR principles.

## 3. Materials and Methods

The purpose of this paper is to gain insight into the humanities research data landscape, with a focus on the current data types, their representation in repositories, and alignment with the FAIR principles.

(1) Since consensus on the definition has not yet been reached, research data in the humanities are often used as an umbrella term that includes different types of sources for research. For this research, the DARIAH-DE definition of research data is used to determine the types of data represented in the humanities. An analysis of datasets published in institutional and thematic repositories in the field of humanities was performed using Digital Academic Archives and Repositories (DABAR), CROSSDA, and OpenAIRE. This research question aimed to determine the most common types of data found in open and/or closed access repositories in the humanities;

(2) In a study on the storage of research data in repositories, Buddenbohm et al. [30] found that a culture of sharing and reusing research data has not yet been established. Despite some data being stored in repositories, they are often difficult to find. The second research question is set up to examine the representation of humanities datasets published in open access sources and the licencing conditions under which they are published;

(3) Given the available number of open access datasets, this research question aims to investigate the extent to which the datasets align with the FAIR principles and, therefore, are appropriate for further use. The evaluation of FAIRness was conducted using F-UJI, a web service that enables the assessment of datasets. It is based on the metrics proposed by the RDA FAIR Data Maturity Model Working Group and developed by

the FAIRsFAIR project "Fostering FAIR Data Practices in Europe". The tool performs a FAIR assessment "based on aggregated metadata; this includes metadata embedded in the data (landing) page, metadata retrieved from a PID provider (e.g., Datacite content negotiation), and other services (e.g., re3data)" [31]. By inputting the URL or persistent identifier of a dataset into the tool interface, F-UJI assesses the dataset and categorises its FAIRness assessment into four levels: incomplete, initial, moderate, and advanced. Each FAIR element is evaluated on the metrics presented in Table S1. Each FAIR principle in the F-UJI tool has a maximum score that every dataset can reach. The maximum score for Findable is 7, for Accessible 3, for Interoperable 4, and for Reusable 10.

This research was conducted in humanities fields, including philosophy, theology, philology, history, art history, art science, archaeology, ethnology and anthropology, religious studies, and interdisciplinary humanities science [32]. The Croatian Social Science Archive (CROSSDA) [33], Digital Academic Archives and Repositories (DABAR) [34], and institutional repositories providing OpenAIRE Explore [35] were selected as data sources due to their search and filter functionalities that enable the breakdown of datasets by subject category. Other repositories, e.g., Zenodo, were not used due to the lack of search and filter functionalities as they could not provide the information needed for conducting this research.

Data were analysed using a deductive or "a priori" coding approach in which initial codes come from the literature or key variables from the conducted research [36]. For data analysis, attribute coding was used as the method in which information about sample data is recorded in the form of a framework matrix [37]. For each dataset in the sample, the following values were applied: name of the repository, scientific field, data type, openness, licences, file format, and FAIR assessment (Findable score (FS), FS FAIR level, Accessible score (AS), AS FAIR level, Interoperable score (IS), IS FAIR level, Reusable score (RS), RS FAIR level, FAIR level in general). This was performed using the F-UJI tool [34]. The aforementioned values were analysed to answer the research questions: the scientific field provided information about humanities fields, the data type and file format gave information about the specific types of datasets stored in repositories, licences gave information about the openness of the datasets, and the F-UJI tool conducted a FAIR assessment and provided information about the FAIR level of datasets.

## 4. Results

RQ1: What types of research data are represented in the humanities?

Although the question of defining research data in the humanities is still under discussion, for this research, we chose the DARIAH DE definition of research data, which includes various types of research data, such as text samples, tables, images, maps, databases, etc. However, based on this definition, research data are not a publication or research article.

From the research sample, the authors collected 123 objects (108 from OpenAire, 13 from DABAR, 2 from CROSSDA) stored as datasets, but only 20% were actual datasets, and the rest were research papers. Among the 26 actual datasets retrieved, the majority were mixed (55.6% were numeric and textual data). Figure 1 shows the types of repositories where it is evident that the largest types of dataset are mixed data that contain both numerical and textual data. The second dataset represented in the repository are textual data with 16.7% and audio data with 11.1%. With the same percentage of 5.6%, repositories represented databases, software, and applications.
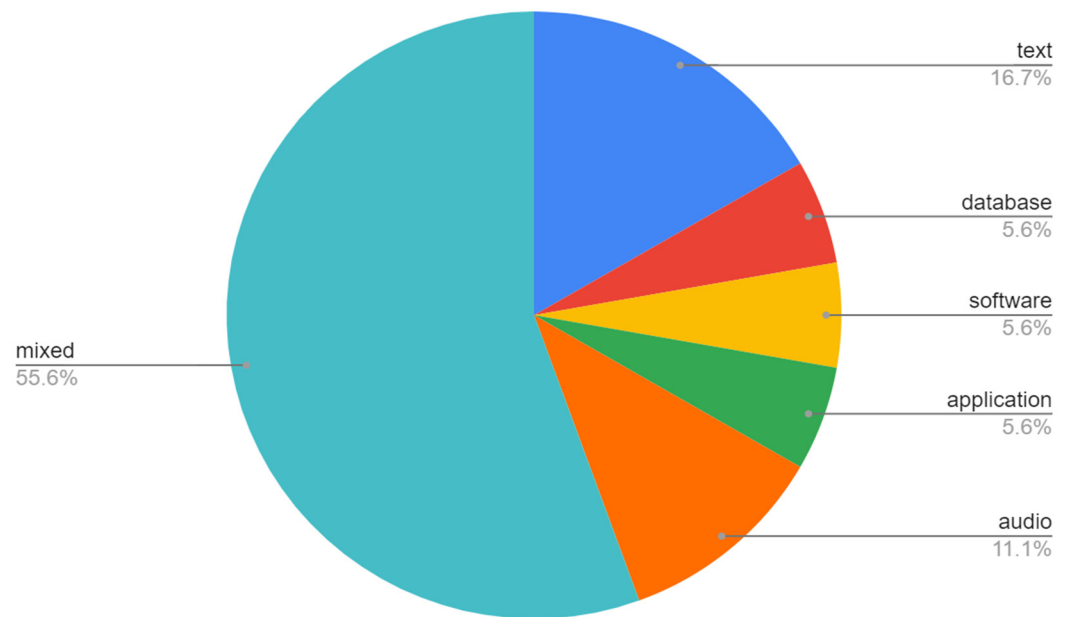
**Figure 1.** Representation of the type of dataset (*n* = 26).

The majority of datasets were from the interdisciplinary humanities (38.5%) and philology (30.8%); archaeology represented 23.1%, and art history and theology 3.8%.

Each dataset stored in the repositories consisted of several files with different data. The results indicate that 47.5% of the data are stored as spreadsheets or tabular files, e.g., .xlsx, .csv, .tab, .r, .tsv, .rda. The second category, representing 22% of the total, are text files, for example .docx, .txt, .pdf, .md. The third place, at 11.9%, is jointly shared by the categories of compressed files (.zip, .gz) and files for storing software code (.json, .mat, .loch, .es3). The research results showed that the least common formats for datasets stored in repositories are database formats (.sql), with 3.4%, and multimedia (.dcf), and audio files (.mp3) each with 1.7%.

RQ2: To what extent are datasets in repositories in the field of humanities represented by openness and under what licence?

Almost all data are openly accessible (88.5%); 3.8% of the data have limited access, and 7.7% of the data are not openly accessible. Although most of the datasets are open access, there are still some restrictions on reuse. Figure 2 shows a representation of the licences assigned to the repository datasets that were analysed. Most datasets have copyright licence with all rights reserved (34.62%). CC BY is one of the least restrictive licences, represented in 20% of the datasets and the second most popular licence in this sample, with 19.23%. Only 11.54% of the datasets are marked with CC0, which serves as a public domain dedication tool and is the most suitable for research data. The CC BY NC ND, CC BY NC, CC BY ND, and CC BY NC SA licences are represented in 7.69% of the datasets, while GPL-3.0 is assigned to 3.85% of the datasets.

RQ3: To what extent do research data align with the FAIR principles?

The analysis of the datasets included in this research showed that the average score for the Findable principle is 5.36 and the mode value is 5, which equals a moderate level of findability. None of the datasets has scores 1 and 3. One dataset had a score of 2, five datasets had a score of 4, ten datasets had a score 5, five datasets had a score of 6, and five datasets had a score of 7, representing an advanced level of findability. Based on this, most of the datasets have a moderate level of findability (*n* = 17), one dataset has a basic level, and eight datasets have an advanced level.
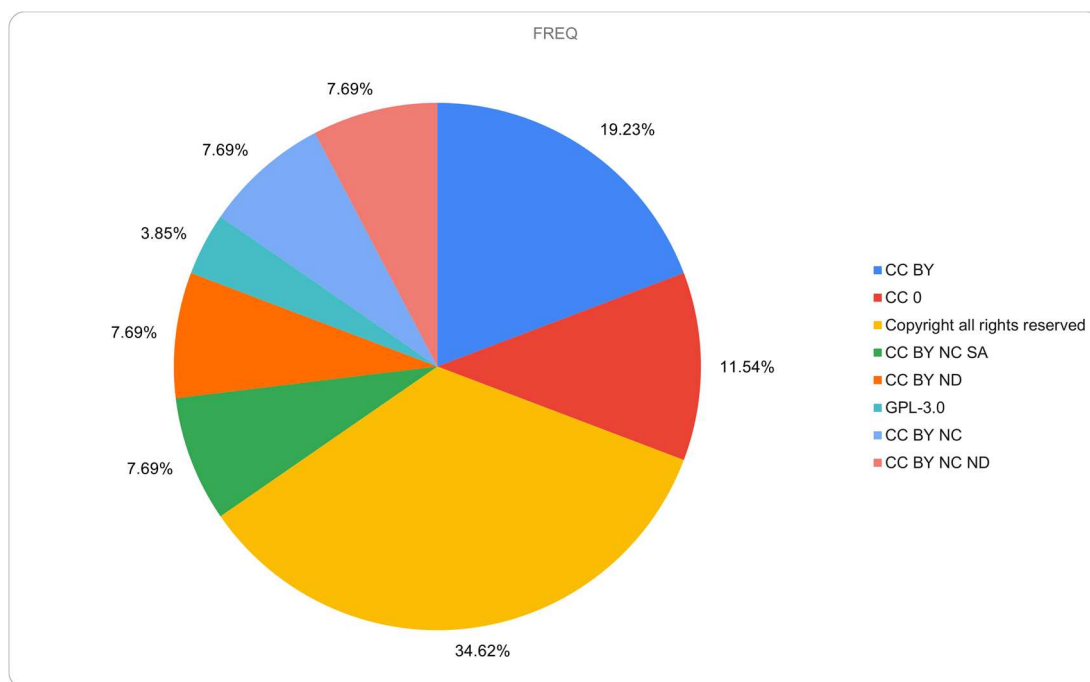
**Figure 2.** Licences assigned to the datasets (*n* = 26).

The average score for the Accessible principle is 1.3 and the mode value is 1, which equals a basic level of accessibility. Most datasets had a score of 1 (*n* = 20), five datasets had a score of 2, and one dataset had a score of 3, which represents an advanced level of accessibility. Based on this, most of the datasets have a basic level of accessibility (*n* = 20), one dataset has an advanced level, and five datasets have a moderate level.

The average score for the Interoperable principle is 1.56 and the mode value is 0, which equals an incomplete level of interoperability. Most of the datasets did not score any points; they scored 0 (*n* = 12) out of 4. One dataset had a score of 1, three datasets had a score of 2, eight datasets had a score of 3, and two datasets had a score of 4, which represents an advanced level of interoperability. Based on this, most of the datasets have an incomplete level of interoperability (*n* = 12), one dataset has a basic level, ten datasets have a moderate level, and three datasets have an advanced level.

The average score for Reusable principle is 2.32 and the mode value is 1, which indicates a basic level of reusability. None of the datasets has scores of 10, 9, 8, 7, or 5. Most of the datasets had a score of 1 (*n* = 13), three datasets had a score of 2, four datasets had a score of 3 and 4, and two datasets had a score of 6. Based on this, the majority of datasets have a basic reusability level (*n* = 23), two datasets have a moderate level, and one dataset has an advanced level.

Based on these results, the FAIR principles with the lowest scores were reusability and accessibility. Most datasets (88.5%) are at a basic FAIR level regarding the reusability principle and the accessible principle with 76.9%. The Findable principle reached the best score, with one dataset at the moderate level and 30.8% at the advanced level. The lowest FAIR principle was Interoperable, because 46.2% of the datasets were at the incomplete level.

Table 1 presents the number of datasets classified by the FAIR levels that the datasets achieved during testing with the F-UJI tool.

**Table 1.** Number of datasets (*n* = 26) classified by FAIR levels.

| | Incomplete | Initial | Moderate | Advanced |
|---|---|---|---|---|
| Findable | 0 | 1 | 17 | 8 |
| Accessible | 0 | 20 | 5 | 1 |
| Interoperable | 12 | 1 | 10 | 3 |
| Reusable | 0 | 14 | 10 | 2 |
| Total FAIR level | - | 14 | 10 | 2 |

The assessment of each dataset provides information about the overall percentage of datasets on each FAIR level. The analysis of the results reveals that most of the datasets are at a basic level of FAIRness (*n* = 14), with an average score of 28.86%; ten datasets are at a moderate FAIR level, with an average of 55.55%; and two datasets are at an advanced FAIR level, with an average of 79%. Based on this, the average score of the overall FAIR level for the datasets is 43.72% and the mode value is 29, which is equal to a basic FAIR level, which means that the average percentage of datasets is at a moderate level and the most frequent FAIR level percentage is 29, which is equal to the basic level. Only one dataset had the highest FAIR level, at 83%, and the lowest FAIR level was 25%. Overall, 53.8% of the datasets are at a basic level of FAIRness, 38.5% of the datasets are at a moderate level of FAIRness, and only 7.7% of the datasets are at an advanced level of FAIRness.

Based on the analysis of FAIRness, datasets in the DABAR system have an average score of 28.38%, which equals a basic FAIR level, the datasets in CROSSDA have an average score of 59%, which equals a moderate FAIR level, and the datasets in OpenAIRE have an average score of 57.71%, which equals a moderate FAIR level.

## 5. Discussion

This article highlights several key findings regarding the state of research data in the humanities, emphasising their significance in the context of open science and the need for improvements in their accessibility and reusability. It confirms that humanities data are diverse, and the line between data and publications is still unclear, since most of the datasets stored in repositories are research articles rather than datasets. This implies that more education and training related to research data management and the principles of FAIRness is needed.

According to Tóth-Czifra [38], the reusability of research data in the humanities is equal to their sustainability, and the problem is not openness but a lack of transparency and proper documentation of research data. This research confirms Tóth-Czifra's [38] findings, because most datasets were openly available, but there was a lack of metadata documentation. This research has also confirmed that the culture of sharing and reusing data has not yet been established [30]. While research data are stored in repositories, they are difficult to find due to the lack of functionalities such as subject categories for filtering datasets on platforms and repositories. The analysis of the dataset confirmed that there is a need to improve the existing platform and repository interfaces with more functional search and filter options, allowing users to obtain the information they need. Although Tóth-Czifra [38] equals reusability and sustainability, it does not mean that reusability is unimportant in the humanities. Based on our results, it is challenged by the fact that the majority of the research data are licenced with copyright or restrictive Creative Commons licences (CC BY NC ND, CC BY NC, CC BY ND, CC BY NC SA). These licenses pose requirements or restrictions that make future reuse challenging.

Within the scientific community, repositories have an important role in developing and disseminating standards such as the FAIR concept, as they "often bridge the data sharing gap between the individual researcher collecting data, funding agencies, publishers, and standards authorities" [39]. This research has shown that all datasets tested using the F-UJI tool adhere to the FAIR principles. However, most of these datasets are only at a basic FAIRness level. The reusability and accessibility principles need improvement, as the datasets scored the lowest on these FAIR principles. These findings are consistent

with previous research conducted by Petrosyan et al. [9], whose investigation showed that the principles of reusability and interoperability have the lowest FAIR assessment score. Another study by Dumontier [10] also found that the interoperability principle had the lowest score, which can be attributed to the focus of F-UJI on structure and not metadata and data measurements.

This study has provided several recommendations for improving the FAIR level of datasets. Comprehensive and standardised metadata for each dataset must be ensured to enhance findability. This can be achieved by providing detailed descriptions of the data, keywords, and relevant contextual information in the form of a Readme file that is stored in the repository along with the dataset. To improve accessibility, datasets should be clear and detailed, including information on data structure, methodology, and any specific considerations for usage. To enhance interoperability, standardised and widely accepted data formats are recommended, e.g., CSV, DOCX, JSON, etc. In addition, it would be useful to represent metadata using a formal knowledge representation language, using semantic resources, or establishing links between data and related entities [9]. To promote reusability, datasets need to adequately address aspects such as whether the metadata actually specify the content of the data, whether they include information on licences, or whether both the data and metadata use standards and formats recommended by the scientific discipline to which they belong [9], to ensure that they are not only accessible but also reliable and usable for subsequent research. In addition, a clear and permissive licensing practice needs to be cultivated regarding the reusability of datasets, such as the use of licenses like Creative Commons (CC) that promote reuse while respecting intellectual property rights. To improve the level of FAIRness, the integration of FAIR assessment tools such as F-UJI into the data publication workflow is recommended. Furthermore, researchers should be provided with support and resources to self-assess and improve the FAIRness of their datasets.

According to Wilkinson et al. [40], FAIRmetrics and FAIRness are focused on metrics that cover various aspects of the FAIR principles and apply to all types of digital objects. These metrics should be universal but flexible enough to adapt to a specific community's needs. They must adhere to the FAIR principles and be open standards to encourage the development of assessment tools. Various methods for FAIR assessment, such as self-assessment, task forces, crowd-sourcing, and automation, should be enabled. However, scalability for assessing billions or even trillions of diverse digital objects is crucial. FAIRness assessments must be continuously updated, revised, timestamped, and publicly accessible. A simple visualisation of a "FAIRness grade" can be a powerful way to inform users and guide resource producers. The assessment process should be designed to encourage digital resource providers to participate and improve based on these assessments. Furthermore, Wilkinson et al. [40] state that FAIRmetrics should have the following qualities: clear, so that anyone can understand the purpose metric; realistic, so it should not be complicated for a resource to comply with the metric; the discriminating metric should measure something important for FAIRness; measurable, which means that the assessment can be made in an objective, quantitative, machine-interpretable, scalable, and reproducible manner, ensuring transparency about what is being measured and how; and universal, so that metrics apply to digital resources. For an improvement in FAIR level, the integration of FAIR assessment tools such as F-UJI into the data publication workflow is recommended because it could provide support and resources for researchers to self-assess and improve the FAIRness of their datasets.

Community engagement is one of the ways to promote FAIRness, e.g., through the organisation of workshops and training sessions to educate researchers, repository managers, and other stakeholders about the FAIR principles. Establishing a feedback mechanism to allow users to provide input on dataset quality and FAIRness is useful and can drive continuous improvement.

## 6. Conclusions

This study is significant because of the ambiguity surrounding humanities research data and the lack of research in this field. The findings reveal a high level of uncertainty in the definition of research data, with most of the analysed datasets being research articles, underscoring a crucial distinction between data as evidence and publications as a source in the humanities. Although the datasets align with the FAIR principles, most are at a basic FAIR level, indicating the need for improvement in terms of compliance. Research data play a fundamental role in the European open science landscape, making it necessary to ensure openness and FAIRness for effective sharing and reuse. In general, this study emphasises the need for enhanced accessibility and reusability to improve the quality and utility of humanities research data.

The limitations of this research study relate to a few main issues. Since we only used the F-UJI tool and considered three repositories, additional data collection and the use of different tools could have provided different results. The identified limitations signal a continuing commitment to advancing the understanding of humanities data to foster openness, accessibility, and reusability in the field.

Future research could be conducted on a broader scale, including more databases and repositories in the analysis. To gain a complete understanding of FAIRness, future research could benefit from using several similar tools on a sample.

This research covers the FAIR principles in digital objects and questions the fundamental principles of open science in humanities. This study can be useful to open science service providers in the humanities field, policymakers, various scholarly or professional interest groups, and researchers. The quest for functional, FAIR, and reusable research data in the humanities depends on their shared understanding, and this research contributes to it.

## References

1. Borgman, C.L. Data, Disciplines, and Scholarly Publishing. *Learn. Publ.* **2008**, *21*, 29–38. [CrossRef]
2. Borgman, C.L. *Big Data, Little Data, No Data: Scholarship in the Networked World*; MIT Press: Cambridge, MA, USA, 2015; ISBN 978-0-262-02856-1.
3. Hauf, N.; Fürholz, A.; Klaas, V.C.; Morger, J.; Šimukovič, E.; Jaekel, M. *Data Reuse in the Social Sciences and Humanities: Project Report of the SWITCH Innovation Lab "Repositories & Data Quality"*; ZHAW Zürcher Hochschule für Angewandte Wissenschaften: Winterthur, Switzerland, 2021. [CrossRef]
4. Gualandi, B.; Pareschi, L.; Peroni, S. What Do We Mean by "Data"? A Proposed Classification of Data Types in the Arts and Humanities. *J. Doc.* **2022**, *79*, 51–71. [CrossRef]
5. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef]

6. Open Data and Accessible Source Materials | Routledge Open Research. Available online: https://routledgeopenresearch. org/for-authors/data-guidelines?utm_source=google&utm_medium=sem&utm_campaign=JRE28264&gad_source=1&gclid= CjwKCAiAvJarBhA1EiwAGgZl0Jp_SEnDs-9Tf7yH8tnZkwn11k8qZy_AxbVbmFsg9btZRiMq5y40GRoCd0IQAvD_BwE (accessed on 28 November 2023).

7. Devaraju, A.; Huber, R.; Mokrane, M.; Cepinskas, L.; Davidson, J.; Herterich, P.; L'Hours, H.; de Vries, J.; White, A. *FAIRsFAIR Data Object Assessment Metrics*; European Commission: Brussels, Belgium, 2020. [CrossRef]

8. Krans, N.A.; Ammar, A.; Nymark, P.; Willighagen, E.L.; Bakker, M.I.; Quik, J.T.K. FAIR Assessment Tools: Evaluating Use and Performance. *NanoImpact* **2022**, *27*, 100402. [CrossRef]

9. Petrosyan, L.; Aleixandre-Benavent, R.; Peset, F.; Valderrama-Zurián, J.C.; Ferrer-Sapena, A.; Sixto-Costoya, A. FAIR Degree Assessment in Agriculture Datasets Using the F-UJI Tool. *Ecol. Inform.* **2023**, *76*, 102126. [CrossRef]

10. Sun, C.; Emonet, V.; Dumontier, M. A Comprehensive Comparison of Automated FAIRness Evaluation Tools: 13th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences. *Semant. Web Appl. Tools Health Care Life Sci.* **2022**, *3127*, 44–53.

11. Unsworth, J. Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this. In Proceedings of the Symposium on Humanities Computing: Formal Methods, Experimental Practice, London, UK, 13 May 2000; King's College: London, UK; p. 5-00.

12. Zins, C. Conceptual Approaches for Defining Data, Information, and Knowledge. *J. Am. Soc. Inf. Sci. Technol.* **2007**, *58*, 479–493. [CrossRef]

13. National Science Board. *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century*; Technical Report NSB-05-40; National Science Foundation: Arlington, VA, USA, 2005.

14. Celjak, D.; Dorotić Malič, I.; Matijević, M.; Poljak, L.J.; Posavec, K.; Turk, I. *Istraživački Podaci—Što s Njima? Priručnik o Upravljanju Istraživačkim Podacima*; Sveučilište u Zagrebu, Sveučilišni Računski Centar: Zagreb, Croatia, 2020.

15. Forschungsdaten—DARIAH-DE. Available online: https://de.dariah.eu/en/forschungsdaten (accessed on 28 November 2023).

16. Pawlicka, U. Data, Collaboration, Laboratory: Bringing Concepts from Science into Humanities Practice. *Engl. Stud.* **2017**, *98*, 526–541. [CrossRef]

17. Owens, T. Defining Data for Humanists: Text, Artifact, Information or Evidence? *J. Digit. Humanit.* **2011**, *1*, 325–341.

18. Allen, R.; Hartland, D. *FAIR in Practice—Jisc Report on the Findable Accessible Interoperable and Reuseable Data Principles*; Zenodo: Geneva, Switzerland, 2018. [CrossRef]

19. Edmond, J.; Tóth-Czifra, E. *Open Data for Humanists, A Pragmatic Guide*; Digital Research Infrastructure for the Arts and Humanities (DARIAH): Dublin, Germany, 2018. [CrossRef]

20. Thoegersen, J.L. "Yeah, I Guess That's Data": Data Practices and Conceptions among Humanities Faculty. *Portal* **2018**, *18*, 491–504. [CrossRef]

21. Kalinin, N.A.; Skvortsov, N.A. Difficulties of FAIR Principles Implementation in Cross-Domain Research Infrastructures. *Lobachevskii J. Math.* **2023**, *44*, 147–156. [CrossRef]

22. Tóth-Czifra, E. *The Risk of Losing Thick. Description: Data Management Challenges Arts and Humanities Face in the Evolving FAIR Data Ecosystem*; Open Book Publishers: Cambridge, UK, 2019; p. 29.

23. Tóth-Czifra, E.; Błaszczyńska, M.; Gelati, F.; Admiraal, F.; Blümm, M.; Buelinckx, E.; Chiquet, V.; Gautschy, R.; Gietz, P.; Király, P.; et al. *Research Data Management for Arts and Humanities: Integrating Voices of the Community*; Fachhochschule Potsdam: Potsdam, Germany, 2023.

24. Calamai, S.; Frontini, F. FAIR Data Principles and Their Application to Speech and Oral Archives. *J. New Music. Res.* **2018**, *47*, 339–354. [CrossRef]

25. de Jong, F.; Maegaard, B.; De Smedt, K.; Fišer, D.; Van Uytvanck, D. CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; European Language Resources Association (ELRA): Miyazaki, Japan, 2018.

26. Strange, D.; Gooch, M.; Collinson, A. Equality, Findability, Sustainability: The Challenges and Rewards of Open Digital Humanities Data. *Int. J. Perform. Arts Digit. Media* **2023**, 1–21. [CrossRef]

27. Hiebel, G.; Goldenberg, G.; Grutsch, C.; Hanke, K.; Staudt, M. FAIR Data for Prehistoric Mining Archaeology. *Int. J. Digit. Libr.* **2021**, *22*, 267–277. [CrossRef]

28. Quan-Hoang, V.; Anh-Tuan, B.; Viet-Phuong, L.; Minh-Hoang, N.; Hung-Hiep, P.; Thanh-Hang, P.; Thi-Hanh, V.; Thu-Trang, V.; Manh-Toan, H. Mirror, Mirror on the Wall: Is Economics the Fairest of Them All? An Investigation into the Social Sciences and Humanities in Vietnam. *Res. Evaluat.* **2021**, *30*, 57–72. [CrossRef]

29. FAIRification Process. Available online: https://www.go-fair.org/fair-principles/fairification-process/ (accessed on 4 November 2023).

30. Buddenbohm, S.; Cretin, N.; Dijk, E.; Gaiffe, B. State of the Art Report on Open Access Publishing of Research Data in the Humanities. *HAL SHS* **2018**, *49*. Available online: https://shs.hal.science/halshs-01357208 (accessed on 4 November 2023).

31. Devaraju, A.; Huber, R. F-uji-an Automated Fair Data Assessment Tool. 2020.

32. Nacionalno Vijeće za Znanost. Pravilnik o Znanstvenim i Umjetničkim Područjima, Poljima i Granama. [National Council for Science. Regulations on Scientific and Artistic Areas, Fields and Branches]. Narodne Novine 2009. p. 118. Available online: https://narodne-novine.nn.hr/clanci/sluzbeni/2009_09_118_2929.html (accessed on 4 November 2023).

33. CROSSDA. Available online: https://www.crossda.hr (accessed on 4 November 2023).
34. DABAR. Available online: https://dabar.srce.hr/ (accessed on 4 November 2023).
35. OpenAIRE | Find Research Compatible Data Sources. Available online: https://explore.openaire.eu/search/content-providers (accessed on 17 November 2023).
36. Connaway, L.S.; Radford, M.L. *Research Methods in Library and Information Science*, 7th ed.; Holtzbrinck: Santa Barbara, CA, USA, 2021; ISBN 978-1-4408-7857-2.
37. Saldaña, J. *The Coding Manual for Qualitative Researchers*, 2nd ed.; SAGE: Los Angeles, CA, USA, 2013; ISBN 978-1-4462-4736-5.
38. Tóth-Czifra, E. FAIR from the Perspective of the Humanities V2.pptx. In Proceedings of the Slides Presented at the Applying the FAIR Data Principles in Day-to-Day Library Practice Workshop at the LIBER2019 Conference, Dublin, Ireland, 26 June 2019.
39. Khodiyar, V.; Laine, H.; O'Brien, D.; Rodriguez-Esteban, R.; TÃ¼rkyilmaz-van der Velden, Y.; Baynes, G.; Brack, M.; Cambon-Thomsen, A.; Carr, D.; Carrus, E.; et al. *Research Data: The Future of FAIR White Paper*; Springer: Berlin/Heidelberg, Germany, 2021. [CrossRef]
40. Wilkinson, M.D.; Sansone, S.A.; Schultes, E.; Doorn, P.; Bonino da Silva Santos, L.O.; Dumontier, M. A design framework and exemplar metrics for FAIRness. *Sci. Data* **2018**, *5*, 180118. [CrossRef]