Research Article

AND DATA SCIENCES
BULLETIN OF COMPUTER

Are Research Data Really FAIR? A Metadata Quality Audit of Research Data Repositories at South African Universities

Siviwe Bangani

North-West University, ZA

Abstract

The FAIR Guiding Principles have become a global benchmark for evaluating the stewardship and reusability potential of research data. Yet, in many contexts, evidence about FAIR alignment is inferred from self-reported practices rather than from direct inspection of deposited datasets and their metadata. Building on prior survey-based work that documented limited data sharing and inconsistent documentation among South African researchers, this paper specifies an artefact-based audit protocol for evaluating the FAIRness of research datasets hosted in institutional and disciplinary repositories associated with South African universities. We develop an operational FAIRness assessment framework that translates the four FAIR dimensions—findability, accessibility, interoperability, and reusability—into a transparent set of observable indicators, including persistent identifiers, metadata completeness, access conditions, licensing, file formats, and documentation. We then fully specify a stratified sampling strategy and a coder-based assessment procedure, including decision rules and inter-rater reliability checks, so that the audit can be implemented consistently across heterogeneous repositories and disciplines. The analysis plan is designed to quantify overall FAIRness and to test for differences by discipline, repository type, and deposition year, while explicitly reporting uncertainty and conducting robustness checks against alternative aggregation choices. Because the contribution of this manuscript is methodological, we do not present repository-level numerical results; instead, we provide an explicitly illustrative example of the reporting and interpretation workflow that the audit enables. The resulting protocol offers a reusable basis for evidence-based evaluation of research data stewardship and for guiding improvements to research data management policy, infrastructure, and training in South African universities and comparable settings.

Keywords: FAIR Guiding Principles, research data management, FAIRness assessment framework, South African universities, institutional and disciplinary data repositories

1. Introduction

Over the past decade, research data management (RDM) and data sharing have moved from peripheral concerns to central components of the scientific enterprise. Across disciplines, there is now a broad recognition that research data are not merely by-products of scholarly activity, but primary research outputs that should be curated, preserved, and, where possible, made available for reuse. Large-scale survey work has repeatedly shown both the centrality of data to contemporary research

Corresponding author: Siviwe Bangani (siviwe.bangani@gmail.com)

BCDS 2025, Vol. 6(2), 1-19. DOI: 10.71448/bcds2562-1

and the persistence of barriers to systematic sharing and reuse across fields and regions [1–4]. Funders, governments, and universities increasingly require that data produced with public funds be stored securely, described adequately, and shared in ways that enable verification, reproducibility, and secondary analysis, often as part of broader open science agendas [5]. This policy pressure coincides with a cultural shift in many research communities, where open science, transparency, and accountability have become prominent values.

In this global context, the FAIR Guiding Principles have emerged as a widely adopted framework for evaluating the quality and reusability of research data [6]. FAIR emphasises that data and their associated metadata should be findable through persistent identifiers and rich descriptions, accessible through stable and well-documented mechanisms, interoperable through the use of standard formats and vocabularies, and reusable through clear licensing, provenance information, and sufficient documentation of context and methods. Subsequent work has elaborated and operationalised these principles in different domains, as well as highlighting challenges in interpreting and implementing FAIR consistently across infrastructures [7, 8]. The principles are intentionally high level and technology neutral, which allows them to be implemented in different ways across disciplines, repositories, and national systems, but they set a shared expectation against which data practices can be assessed.

South Africa has been an active participant in this broader shift towards open and FAIR-aligned data practices. The National Research Foundation (NRF) Open Access Statement of 2015 explicitly calls for NRF-funded research outputs, including underlying data, to be deposited in suitable repositories, preferably with open access, subject to legal and ethical constraints. In parallel, several South African universities have begun to develop RDM policies, integrate data management planning into grant and ethics processes, and invest in institutional repositories and library-led support services [9–11]. These initiatives mirror international trends in which academic libraries and related units take on expanding roles in RDM, from policy development and training to repository management and data stewardship [12, 13]. They aim to ensure that research data are not only preserved for the long term, but also curated in ways that increase their visibility, impact, and potential for reuse both within and beyond the country.

Empirical work has started to document how these policies and infrastructures are experienced by researchers in practice. Survey-based studies such as Bangani and Moyo [14] provide an important snapshot of data practices at South African universities. Their findings suggest that a large proportion of researchers make use of data created by others, but that comparatively few share their own data outside their immediate research teams or institutions. Many respondents report storing data on personal computers, external drives, or departmental servers rather than in dedicated repositories, and they often acknowledge that their documentation and labelling practices are rudimentary or inconsistent. Legal and ethical concerns, fear of data misuse or misinterpretation, uncertainty about intellectual property, and a perceived lack of time or resources emerge as recurrent barriers to wider data sharing, echoing patterns observed in global surveys of scientists' data practices [1–3].

A key limitation of this growing body of work, however, is that it relies primarily on self-reported practices. Researchers are asked whether they share data, which formats they use, how they back up and document their datasets, whether they are aware of institutional policies, and how they perceive the benefits and risks of sharing. These self-reports are informative, but they are vulnerable to several kinds of bias. Social desirability may encourage respondents to describe their behaviour in ways that align with perceived expectations rather than with day-to-day practice. Recall limitations may lead

to incomplete or inaccurate answers, especially when data management is distributed across multiple projects and storage locations. Differences in how respondents interpret terms such as "repository", "metadata", or "documentation" can also complicate comparisons and mask underlying heterogeneity in practice.

Most importantly for the present study, survey responses do not directly reveal the actual FAIRness of the datasets that have been deposited in repositories. It is one thing for a researcher to state that they "document" their data or that they have "shared" a dataset; it is quite another to examine the resulting dataset record and assess whether it has a persistent identifier, whether the metadata are sufficiently rich to enable discovery, whether the files are provided in interoperable formats, whether licences and access conditions are clearly specified, and whether there is enough contextual information for a third party to understand and reuse the data appropriately. At present, there is limited systematic evidence about whether datasets archived at South African institutions satisfy these FAIR-oriented expectations in practice, despite the growing international literature on FAIR assessments and metrics [8].

This paper addresses this gap by shifting attention from attitudes and self-reports to observable artefacts: the datasets and metadata records stored in research data repositories at South African universities. Instead of asking researchers what they do, the audit approach examines what is actually exposed at the point of deposit, treating the public dataset record as the relevant object of assessment. In this sense, the framework operationalises FAIRness at the level of the repository-facing record (metadata plus any accessible files and documentation), rather than as a claim about the intrinsic scientific quality of the underlying data or the internal practices of a research group.

Using this perspective, we develop a FAIRness auditing framework that translates the abstract FAIR Principles into a concrete set of operational indicators that can be applied consistently across repositories and disciplines. The indicators focus on observable features that are central to practical reuse, such as the presence and type of persistent identifiers, the completeness and structure of descriptive metadata, the explicitness of access and licensing information, the choice of file formats, and the availability of documentation on methods, provenance, and data structure. Importantly, we treat FAIR as conceptually distinct from openness: datasets may be legitimately restricted for ethical, legal, or commercial reasons, but they can still be FAIR if the conditions of access are explicit and the metadata remain discoverable and informative.

On this basis, the paper specifies a stratified audit design covering multiple institutional and disciplinary repositories, a range of disciplinary domains, and multiple deposition years. The analysis is organised around three guiding questions. First, to what extent do research datasets deposited in South African institutional and disciplinary repositories satisfy core FAIR criteria when assessed through observable indicators? Second, how does FAIRness vary across disciplines, repository types, and deposition years, and what patterns are robust to alternative scoring and aggregation choices? Third, to what extent are observed FAIRness patterns consistent with previously reported data management and sharing practices, including those documented by Bangani and Moyo [14], and where do they suggest gaps between stated intentions, policy frameworks, and the repository-visible reality of data and metadata?

Rather than presenting repository-level numerical outcomes, this paper emphasises measurement validity, methodological transparency, and analytical reproducibility. The goal is to provide a fully specified blueprint that can be implemented and extended in follow-up empirical work, and that supports evidence-based evaluation of research data stewardship in South Africa and comparable set-

tings. By articulating a practical approach to FAIRness assessment—including explicit assumptions, decision rules, and planned robustness checks—the study aims to support more informed discussions about how RDM policies, infrastructures, and support services can be strengthened to realise the promise of FAIR data in contexts where resources, capacities, and research cultures are diverse and unevenly developed.

2. Research Design and Methodology

2.1. Overall Approach

We adopt a cross-sectional, multi-repository audit of research datasets hosted in institutional and disciplinary repositories associated with South African universities. "Cross-sectional" here means that the audit evaluates each sampled dataset record as it is presented in the repository at the time of assessment, providing a snapshot of repository-visible FAIRness rather than tracking records longitudinally as they evolve. This design choice makes the audit feasible across heterogeneous repositories and enables clear interpretation of measured indicators as properties of the deposited record.

The unit of analysis in this study is the dataset record. For the purposes of the audit, a dataset record is defined as the combination of (i) the structured metadata describing a research dataset as displayed on the repository landing page or metadata view and (ii) any associated data files or documentation that are accessible through the repository interface. In practice, for each sampled record we inspect the landing page, record all relevant metadata fields, verify identifier and access information, and examine the list of files and formats. Where documentation files are available (for example readme documents, codebooks, or methodological appendices), we open them to assess whether they provide the minimum contextual information needed for reuse. We do not attempt to reanalyse the substantive content of the data, nor to evaluate the scientific quality of the underlying research; the audit focuses strictly on stewardship features that are observable at the repository interface and that plausibly condition discoverability and reuse.

The methodological pipeline consists of four interlinked stages. In the first stage, we identify suitable repositories and construct a sampling frame that enumerates eligible dataset records, along with core attributes such as repository affiliation, disciplinary domain, and deposition year. In the second stage, we develop and pilot a FAIRness assessment rubric that translates the high-level FAIR Principles into concrete indicators that can be applied consistently to the sampled records, with explicit decision rules for ambiguous cases. The third stage involves manual assessment of a stratified sample of dataset records by trained coders, using the rubric as a scoring instrument and recording structured notes that support later adjudication. The fourth stage analyses the resulting indicator-level data and FAIRness scores to characterise distributions, test pre-specified comparisons, and quantify uncertainty, while conducting planned robustness checks. Throughout these stages, procedures are refined only through documented rule updates, so that any changes in scoring practice are transparent and reproducible.

2.2. Repository Selection and Sampling Strategy

The first methodological task is to identify repositories that are relevant to the South African research data landscape and that provide sufficient information to support a FAIRness audit. We focus on two main categories of repositories. The first category comprises institutional repositories that are

maintained or endorsed by individual universities, often managed by academic libraries or research offices. These repositories typically host a mix of content types, including theses, journal articles, and research datasets, and they play an increasingly important role in institutional RDM strategies. The second category comprises discipline-specific or thematic repositories that have a clear connection to South African research institutions, either through formal partnerships, national infrastructure status, or a demonstrable concentration of South African datasets.

Repository inclusion is governed by operational criteria aimed at maximising comparability and reducing measurement error. Each repository must (i) be maintained or explicitly endorsed by a South African university or national research infrastructure, (ii) expose dataset records through a public web interface with stable landing pages, and (iii) provide a way to identify and filter research datasets (as distinct from articles, theses, or other non-data content). Repositories that host datasets but obscure key metadata fields behind authentication, or that provide only unstructured, non-persistent record pages, are excluded because indicators cannot be measured consistently. For each included repository, we document the repository platform (where identifiable), the metadata fields visible in the interface, and any repository-level deposit requirements that may affect observed FAIRness.

Once eligible repositories have been identified, we construct a sampling frame that lists all dataset records that meet a set of inclusion criteria. We include records that are explicitly tagged or categorised as research datasets, data collections, data packages, or equivalent labels that clearly denote underlying data rather than publications, software-only artefacts, or non-research content. We restrict attention to datasets deposited within a defined time window, from 2015 onwards, to align the audit with the contemporary policy environment following the NRF Open Access Statement and to reduce heterogeneity introduced by legacy records created under substantially different metadata practices. For each eligible record, we extract and store a minimal set of frame variables: repository identifier, persistent record URL, deposition year, and any disciplinary or subject tags provided by the repository.

Because the audit seeks to characterise variation across different parts of the research system, we employ a stratified sampling strategy. Strata are defined along three dimensions. The first dimension distinguishes repository type, separating institutional repositories from disciplinary or thematic repositories. The second dimension is broad disciplinary domain (for example natural sciences, health sciences, engineering and technology, social sciences, and humanities). The mapping from repository subject tags to these domains is specified in advance and then refined during piloting through explicit rules to minimise discretionary recoding. The third dimension is deposition year, discretised into a small number of periods to differentiate between early and later stages of FAIR-related policy and infrastructure development.

Within each stratum, dataset records are sampled at random. Target stratum sizes are chosen to balance two requirements: (i) sufficient observations to estimate stratum means and medians with useful precision and to support pairwise comparisons, and (ii) feasibility under manual coding constraints. To avoid conclusions being driven by arbitrary weighting choices, the primary analysis reports both overall (pooled) estimates and stratum-specific estimates, and it uses stratification variables explicitly in all comparative models. Any deviations from the planned sampling targets (for example due to the scarcity of eligible records in certain strata) are recorded and reported so that representativeness can be evaluated.

2.3. FAIRness Assessment Framework

2.3.1. Conceptual Foundations. The core of the methodology is the FAIRness assessment framework, which translates the abstract FAIR Principles into a set of concrete indicators that can be applied consistently across a diverse set of repositories and disciplines. We adopt three explicit assumptions to make this translation testable. First, we assume that the repository landing page and its linked files constitute the authoritative public representation of the dataset for potential reusers; the audit therefore treats the record as the object of evaluation. Second, we assume that FAIRness is graded rather than binary: datasets can be more or less findable, accessible, interoperable, and reusable, and they may perform well on some dimensions while falling short on others. Third, we assume that, for an audit intended to support policy and service improvement, indicators must be observable, rule-based, and comparable across repositories, even if this requires prioritising widely applicable stewardship features over highly discipline-specific best practices.

Within this framing, the rubric is designed to be transparent, practical, and comparable. Transparency means that each indicator has an explicit definition and an auditable decision rule, so that a reader can reconstruct why a dataset received a particular score. Practicality means that the rubric can be applied by trained coders who are not domain experts in every represented field; indicators therefore rely on information that is visible in the repository interface or in accessible documentation, rather than on specialised disciplinary interpretation. Comparability requires that the same indicator set and scoring rules apply across repositories with different platforms and metadata schemas; where repositories expose different field names for similar concepts, coders map fields to the indicator definitions using a standardised crosswalk.

The framework is also careful to distinguish FAIRness from adjacent but non-equivalent constructs. In particular, FAIRness is not treated as a proxy for scientific validity or for the substantive quality of the data, and it is not equated with openness. Datasets may be restricted for legitimate reasons; in such cases, accessibility is operationalised in terms of whether access conditions and procedures are explicit and whether metadata remain discoverable and informative. These distinctions are necessary to ensure that indicator scores can be interpreted as stewardship features that can be improved through repository workflows, policy requirements, and training.

The rubric development process is grounded in three sources. First, we draw on existing FAIR metrics and assessment tools to ensure that indicator choices correspond to widely discussed interpretations of the FAIR dimensions. Second, we examine the metadata schemas and interfaces of the selected repositories to identify which fields are available, how they are populated, and what kinds of documentation are typically provided, thereby reducing the risk of defining indicators that cannot be observed reliably. Third, we incorporate feedback from librarians and RDM practitioners to ensure feasibility and relevance in the local context, and to calibrate decision rules for ambiguous cases that occur frequently in practice.

2.3.2. Operational Indicators and Scoring. On the basis of these conceptual foundations, we define a set of operational indicators grouped under the four FAIR dimensions. Indicators are chosen to satisfy two criteria simultaneously: they must be closely connected to the FAIR principles as commonly interpreted in repository practice, and they must be measurable from the dataset record without requiring privileged access or specialised disciplinary knowledge. For each indicator, the coding manual specifies the exact evidence sources within the record (metadata fields, landing-page elements, and accessible documentation) and applies a conservative "visible-evidence" rule: coders

score only what is explicitly present in the record, and they do not infer information that might exist elsewhere.

For the findable dimension, indicators capture whether the dataset is associated with a globally unique and persistent identifier (for example a DOI or Handle), whether the metadata include a minimally sufficient descriptive bundle (a descriptive title and summary/abstract, creator/contributor information, keywords or subject terms, and a deposition date), and whether the record is discoverable within the repository catalogue through basic search and browsing mechanisms. Where repositories support richer discovery features (for example faceted filtering), these are recorded as supporting evidence but do not substitute for the presence of record-level metadata.

For the accessible dimension, indicators assess whether the record provides stable, protocol-based access to the data or, where access is restricted, to a clear access pathway. Specifically, coders verify whether the landing page provides a stable URL and whether the record specifies access status (open, embargoed, or restricted) together with an explicit mechanism for obtaining the data (direct download, mediated request, or controlled access). Because FAIR does not require data to be open, restricted records are not penalised for restriction per se; rather, they are scored on the clarity, persistence, and transparency of access conditions and on whether metadata remain visible even when files are inaccessible.

For the interoperable dimension, indicators focus on syntactic and semantic interoperability as observable in the record. Syntactic interoperability is assessed by whether primary data files are provided in non-proprietary, widely used formats where feasible, or whether proprietary formats are accompanied by open equivalents or clear conversion guidance. Semantic interoperability is assessed by whether metadata use standardised elements (for example structured fields rather than free text alone) and whether controlled vocabularies, standard subject headings, or persistent identifiers for related entities (such as associated publications or authors) are used when available. The audit records the presence of explicit links between the dataset and related outputs, recognising that such linkages materially increase integrability and reuse.

For the reusable dimension, indicators assess whether the record provides a clear legal and contextual basis for reuse. Coders record whether an explicit licence or terms-of-use statement is present and unambiguous, and they evaluate the availability of documentation that enables interpretation by a third party (for example readme files, codebooks, data dictionaries, methodological descriptions, or well-specified links to methods in associated publications). Reusability scoring is anchored in the principle that a dataset can be technically accessible yet practically unusable if core contextual information (provenance, collection methods, variable definitions, or processing steps) is missing.

Each indicator is scored on an ordinal scale reflecting the degree to which the dataset record meets the relevant expectation. A three-point scale is used for most indicators: zero indicates that the feature is absent or clearly inadequate, one indicates partial fulfilment (present but incomplete, ambiguous, or minimally informative), and two indicates that the record meets the criterion in a complete and usable way. The coding manual provides decision rules and examples for each score point to minimise coder discretion. For instance, licensing is scored as zero if no licence or terms of use are mentioned, one if rights information is present but ambiguous or non-standard, and two if a clear, standardised licence or equivalent explicit permission statement is provided.

For each dataset record, dimension subscores are computed by summing the indicator scores within a dimension and normalising by the maximum attainable score for that dimension under the indicators assessed. The primary overall FAIRness score is then defined as the simple (unweighted) average of

the four dimension subscores. An unweighted aggregation is used as the primary specification to avoid embedding a normative weighting of the FAIR dimensions into the headline score; instead, the dimension subscores are reported alongside the overall score for interpretation. As a planned robustness check, we also compute alternative overall scores under plausible reweighting schemes and verify whether substantive comparisons (for example between repository types or over time) are sensitive to these choices.

2.4. Coder Training and Reliability

Because the FAIRness assessment relies on human judgement in applying the rubric to heterogeneous dataset records, it is crucial to ensure that coders apply indicators and scoring rules consistently and that measured variation reflects real differences between records rather than coder idiosyncrasies. The reliability procedure therefore has two roles: to validate that indicators are sufficiently well-defined to be applied consistently, and to identify indicators or decision rules that require refinement.

The first step is the preparation of a detailed coding manual that defines each indicator, specifies the evidence sources coders must consult, and describes the criteria for each point on the scoring scale. The manual includes worked examples of common record patterns (for example open datasets with full documentation, records with restricted files but visible metadata, and records with minimal metadata) and provides explicit guidance on how to handle ambiguous cases such as partial documentation, broken links, or inconsistent repository interfaces.

Before formal coding begins, coders participate in a structured training workshop in which the manual is reviewed and applied to a pilot set of records spanning multiple repositories and disciplines. Training emphasises consistent application of the "visible-evidence" rule and requires coders to record brief justifications for non-trivial scores (especially zeros and ones) so that disagreements can be diagnosed. During piloting, disagreements are discussed and resolved through documented rule clarifications; the manual is then updated so that subsequent coding reflects stable decision rules rather than ad hoc consensus.

To quantify reliability, at least two coders independently score an overlapping subset of records sampled to cover all major repository types and disciplinary domains. Inter-rater agreement is evaluated at the indicator level and at the dimension-subscore level using appropriate statistics (for example Cohen's kappa for categorical indicators and an intraclass correlation coefficient for aggregated subscores). Reliability thresholds are defined in advance: indicators that do not reach an acceptable agreement level trigger targeted refinement of definitions and examples, followed by recoding of the pilot subset. This process continues until agreement is stable enough to justify scaling the assessment to the full sample, ensuring that subsequent analyses can be interpreted as reflecting record-level stewardship features rather than measurement noise.

2.5. Data Analysis

Once coding is complete and FAIRness scores have been computed for all sampled dataset records, the analysis proceeds in three layers: (i) descriptive characterisation of indicator frequencies and score distributions, (ii) pre-specified comparisons across repository types, disciplines, and deposition periods, and (iii) multivariable modelling to examine associations while controlling for potential confounders.

Descriptive analysis reports the distribution of the overall FAIRness score and each dimension subscore using means and medians together with uncertainty summaries. Because scores are bounded

and derived from ordinal indicators, we report bootstrap confidence intervals for means and medians and present distributional visualisations (for example histograms and boxplots) to avoid over-reliance on single summary statistics. Indicator-level reporting is also essential for interpretability: alongside aggregate scores, we tabulate the prevalence of key stewardship features (for example persistent identifiers, explicit licences, and documentation files) so that improvements can be linked to specific workflow components.

For group comparisons, we test differences in overall scores and dimension subscores between institutional and disciplinary repositories, across disciplinary domains, and across deposition periods. Given the bounded nature of the scores and the possibility of non-normal distributions, the primary comparisons use non-parametric tests where appropriate and emphasise effect sizes (for example differences in medians and rank-based measures) rather than p-values alone. Where multiple pairwise comparisons are conducted, we control the family-wise error rate through a documented correction procedure. Uncertainty is communicated through confidence intervals for group differences and, where relevant, through permutation-based significance assessment.

To examine determinants of FAIRness more systematically, we estimate regression models with the overall score and dimension subscores as outcomes and repository type, disciplinary domain, and deposition year (or period) as predictors, with additional repository-level covariates included when they are observable and consistently defined. Model choice is tied to the score properties: linear models are used only when residual diagnostics support their use; otherwise, bounded-outcome or ordinal modelling strategies are adopted to respect the scale. All models are reported with uncertainty intervals for coefficients and with sensitivity analyses that assess whether conclusions are robust to alternative score aggregation rules and to the exclusion of records with restricted files.

Finally, we conduct planned robustness checks that address the most common threats to inference in repository audits. These include repeating key comparisons under alternative weighting schemes for the overall score, verifying that conclusions do not hinge on a small number of repositories or disciplines (leave-one-repository-out checks), and assessing the impact of potentially ambiguous indicators by re-running analyses with those indicators removed. Together, these procedures ensure that any substantive conclusions drawn from a completed audit are explicitly supported by the observed evidence and are not artefacts of arbitrary parameter settings.

3. Illustrative Findings and Analysis

The figures and narrative in this section are explicitly illustrative: they are included to make the analysis plan concrete and to show how results from a completed audit should be reported, interpreted, and stress-tested. They should not be read as empirical claims about any specific repository or institution. The aim is to clarify the reporting logic that links indicator-level observations to dimension subscores, to overall FAIRness summaries, and to statistically supported comparisons. In a completed implementation, the illustrative statements below would be replaced or supplemented with numerical estimates, uncertainty intervals, and documented robustness checks as specified in the methodology.

3.1. Overall FAIRness Levels

A first layer of analysis summarises the distribution of FAIRness scores across all sampled dataset records. For each record, we compute an overall score as described in Section 2.3, together with

subscores for findability, accessibility, interoperability, and reusability. Reporting begins with the empirical distribution of the overall score (for example via a histogram or density plot) and the corresponding summary statistics (mean, median, and interquartile range), accompanied by bootstrap confidence intervals to convey uncertainty in central tendencies. Figure 1 is included as an example of the type of visualisation used for this purpose.

A key interpretive requirement is to avoid treating the overall score as a black box. The four dimension subscores are therefore examined separately and reported side-by-side. Figure 2 illustrates the standard presentation: boxplots (or equivalent) that show medians, interquartile ranges, and outliers for each dimension. In a completed audit, this representation would be accompanied by indicator-level prevalence tables so that differences between dimensions can be traced back to concrete stewardship features (for example whether low reusability is driven primarily by absent licences, weak documentation, or missing provenance information).

Because the subscores are computed from ordinal indicators and are bounded by construction, interpretation focuses on magnitude and practical meaning rather than on small differences in means. In addition, the completed analysis would report sensitivity of the overall score to the aggregation rule by recomputing the overall score under alternative weighting schemes and verifying whether qualitative conclusions about the distribution (for example whether scores cluster toward the middle or polarise) remain stable. These checks ensure that any narrative about "typical" FAIRness levels is directly supported by indicator evidence and is not an artefact of a particular scoring convention.

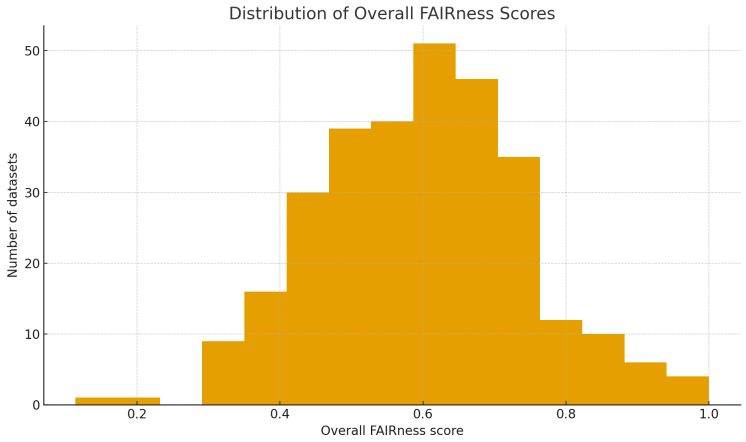

Fig. 1. Illustrative distribution of overall FAIRness scores across all sampled dataset records. Bars indicate the proportion of datasets falling into each score interval; the dashed line marks the sample mean.

3.2. Differences Across Disciplines and Repository Types

The next analytical step compares FAIRness scores across disciplinary domains and repository types using the stratification variables defined in the sampling design. The primary outputs are stratum-specific distributions and effect-size summaries: for each discipline and repository type, we report the median and interquartile range of the overall score and each dimension subscore, together with confidence intervals for between-group differences. Figure 3 provides an example of a visual summary that separates institutional from disciplinary repositories within each domain.

Group comparisons are conducted in a way that respects the bounded, potentially non-normal nature of the scores. In a completed audit, the primary inferential statements would be based on

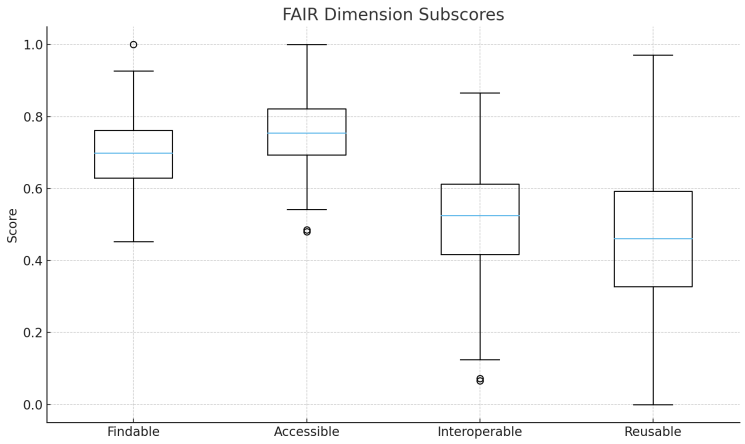

Fig. 2. Illustrative boxplots of FAIR dimension subscores (Findable, Accessible, Interoperable, Reusable). Each box summarises the median, interquartile range, and outliers for the corresponding dimension across all sampled datasets.

rank-based or permutation procedures where appropriate, accompanied by multiple-comparison corrections when many domains are compared. Importantly, comparisons are interpreted as descriptive of observable record-level stewardship in the sampled repositories, not as claims about disciplinary "quality" or researcher intent.

To strengthen causal interpretation where relevant, multivariable models are used to estimate associations while controlling for potentially confounding factors, such as deposition period and repository type. Robustness checks include verifying that key differences are not driven by a single repository (leave-one-repository-out analysis) and that conclusions are stable when alternative score aggregations are used. These checks are crucial because repository platforms and deposit workflows can cluster within disciplines, creating the appearance of disciplinary effects that are in fact infrastructural.

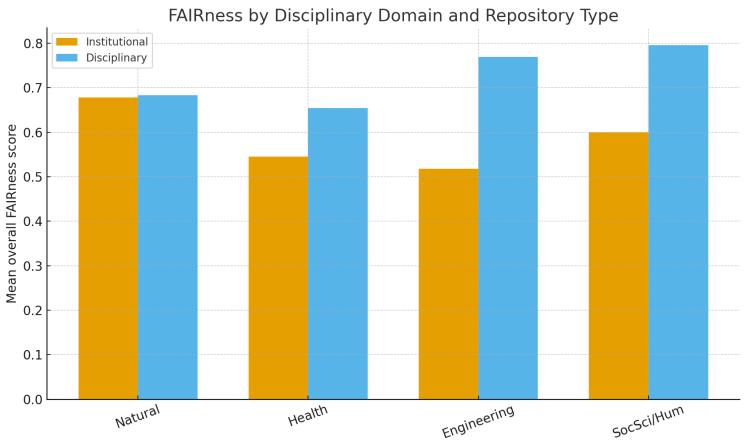

Fig. 3. Illustrative FAIRness profiles by disciplinary domain and repository type. The figure shows, for each domain, the distribution of overall FAIRness scores in institutional repositories (light shading) and disciplinary repositories (dark shading).

3.3. Temporal Trends and Policy Effects

By incorporating deposition year (or deposition period) into the sampling frame, the audit enables analysis of temporal variation in FAIRness. The descriptive starting point is a plot of mean or

median overall FAIRness by year, separated by repository type, together with uncertainty intervals that reflect sampling variability. Figure 4 illustrates this reporting style. In a completed audit, such plots are interpreted cautiously: apparent year-to-year fluctuations may reflect changing repository coverage or sampling error, so smoothing and uncertainty presentation are essential.

To move beyond visual inspection, temporal patterns are evaluated using regression models in which FAIRness outcomes are regressed on deposition year (or period) and repository type, optionally interacting these predictors to test whether temporal change differs between institutional and disciplinary repositories. Where the analysis aims to relate changes to policy milestones, such claims require explicit modelling assumptions: the completed audit would therefore treat policy dates as contextual reference points rather than as automatically causal breakpoints, and it would report sensitivity to alternative period definitions. Robustness checks include repeating the analysis after excluding very recent deposits (which may have incomplete curation) and after excluding records with restricted files to ensure that conclusions about temporal improvement are not driven by changes in access policy alone.

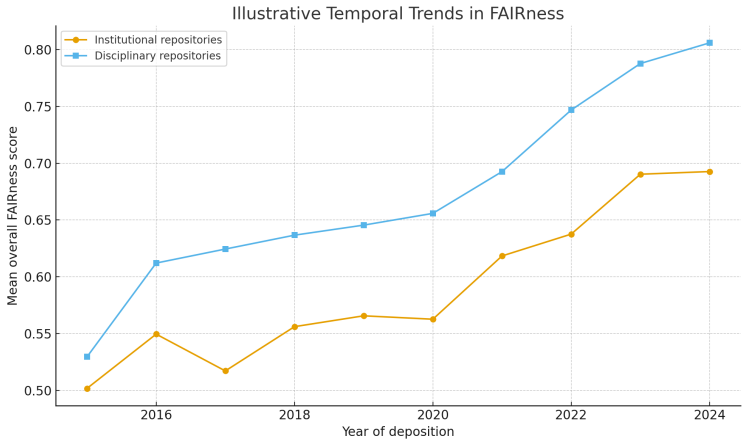

Fig. 4. Illustrative temporal trends in mean overall FAIRness scores by year of deposition. Separate lines show institutional repositories and disciplinary repositories; vertical reference lines can be used to indicate key policy milestones such as the 2015 NRF Open Access Statement.

3.4. Comparing Observed FAIRness with Self-Reported Practices

A final strand of analysis links the FAIR audit to prior survey-based evidence on researcher attitudes and self-reported practices, such as the findings reported by Bangani and Moyo [14]. The empirical objective is not to "validate" surveys against the audit (the two measure different constructs), but to quantify the extent to which self-reported practices align with repository-visible stewardship features at the level of institutions or broad disciplinary groupings.

In a completed implementation, this comparison requires careful operational alignment. Survey items are mapped to the closest observable indicators (for example documentation practices to reusability indicators, and sharing practices to accessibility status), and comparisons are conducted at an aggregation level for which both data sources are available. Figure 5 illustrates a typical visual output: a scatter plot comparing an institution-level self-report measure (for example the share of respondents reporting documentation) with an institution-level observed reusability subscore. Inference focuses on the direction and magnitude of association (with uncertainty), and interpretation explicitly acknowledges limitations such as survey response bias, differences in disciplinary compo-

sition, and the possibility that surveyed researchers deposit data in repositories not covered by the audit.

Where systematic gaps are observed (for example self-reports suggesting widespread documentation but low observed documentation evidence in records), the audit provides a concrete basis for diagnosing where the gap arises: it may reflect weak deposit requirements, insufficient training on documentation standards, or a mismatch between what researchers consider "documentation" and what is reusable by third parties. The completed analysis would therefore accompany any gap claims with indicator-level evidence and with sensitivity checks that examine whether the gap persists under alternative mappings and aggregation rules.

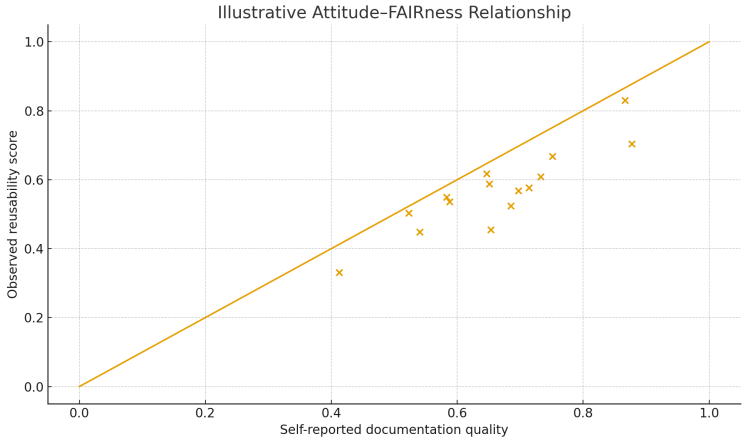

Fig. 5. Illustrative relationship between self-reported documentation practices and observed reusability scores at the institutional level. Each point represents an institution; the diagonal line indicates equality between reported and observed practices, so that points below the line suggest a documentation gap.

Overall, the illustrative analyses and figures in this section clarify how the proposed FAIRness audit links indicator-level coding to interpretable summaries and statistically supported comparisons. Once populated with empirical data from the specified sampling and coding procedure, the same reporting workflow would provide a transparent basis for describing the distribution, determinants, and evolution of repository-visible FAIR-aligned practices, with conclusions explicitly tied to uncertainty estimates and robustness checks.

4. Discussion

4.1. Implications for RDM Policy and Practice

A systematic FAIRness audit of research datasets in South African repositories has several important implications for research data management policy and practice. At the most basic level, it provides an evidence-based diagnosis of where current RDM efforts are succeeding and where they are falling short. Rather than relying on anecdotal impressions or the perceptions of particular stakeholder groups, the audit offers a direct view of the quality and reusability of datasets as they are actually preserved and exposed in repositories, in line with broader arguments that data stewardship should be evaluated on the basis of concrete artefacts and workflows rather than aspirations alone [15]. If, for example, the findings indicate that findability and accessibility scores are consistently high while reusability remains weak, this pattern would suggest that repositories and institutional policies have been relatively successful in addressing the core tasks of preservation and discovery, but have not yet

fostered a culture of rich documentation, explicit licensing, and careful articulation of provenance. Such a diagnosis not only identifies weaknesses but also acknowledges areas where substantial progress has already been made, echoing international experiences where basic FAIR compliance was achieved earlier than deep interoperability and reusability [8, 16].

The availability of this kind of detailed profile enables more targeted interventions than would be possible on the basis of general policy aspirations alone. If the audit reveals that datasets frequently lack clear licences and that documentation is sparse or unstructured, institutions can respond by introducing or strengthening requirements for minimal documentation bundles at the point of deposit, for example by mandating the inclusion of a readme file, a codebook, or a data dictionary and by insisting on the selection of a standard licence or terms-of-use statement. Repository workflows can be adjusted to make these steps as seamless as possible, for instance by integrating licence selection into submission interfaces and by providing templates or semi-structured forms for documentation. These kinds of workflow and policy adjustments are consistent with the trajectories reported in case studies of library-led RDM programmes, where service design, template development, and policy refinement have been used to nudge practice toward better data stewardship [17, 18]. In turn, repository managers and librarians can use the audit results to prioritise enhancements to repository software, such as adding fields that support richer metadata or improving the visibility of licensing information on dataset landing pages, responding to the kinds of gaps repeatedly identified in reviews of RDM services and training [19, 20].

Beyond immediate corrective action, a FAIRness audit also creates the possibility of monitoring progress toward policy goals over time. If the audit is repeated periodically, even at modest intervals, universities and funders can track whether investments in RDM training, infrastructure, and policy refinement are yielding measurable improvements in data quality, interoperability, and reusability. FAIR-oriented maturity models already emphasise the value of repeated assessment and benchmarking as a way to harmonise expectations and demonstrate progress [16]. Such longitudinal evidence is particularly valuable in an environment where resources are limited and competing priorities are numerous, because it allows decision-makers to assess which forms of intervention are most effective and where efforts may need to be redirected. Over time, a series of FAIRness audits can function as a form of performance indicator for the research system's data stewardship, complementing traditional bibliometric measures of publication output with a richer account of data outputs and their readiness for reuse.

The audit further has implications for the design and delivery of RDM training and support services. Indicators that consistently receive low scores, such as those related to interoperability, standard vocabularies, or licence specification, highlight topics that should be prioritised in workshops, online courses, and one-to-one consultations. Evidence from evaluations of RDM instruction suggests that focused, practice-oriented training can measurably improve researchers' skills and behaviours, especially when it is integrated into existing research workflows and curricula [19, 20]. Training materials can be tailored to the specific weaknesses identified in the audit, using real examples from local repositories to illustrate both common pitfalls and exemplary practice. Conversely, indicators that show strong performance can be used to identify departments, institutes, or individual projects that have developed effective approaches to data management. These can be showcased as case studies or champions, helping to normalise good practice and fostering peer learning, in line with findings that peer norms and institutional expectations play an important role in shaping data sharing and curation behaviours [2, 3]. In this way, the FAIRness audit not only diagnoses problems

but also supports a learning-oriented approach to improving RDM across the institution and the wider research community.

4.2. FAIRness in Resource-Constrained Contexts

The South African context adds a crucial dimension to global discussions of FAIR data. Universities and researchers operate under resource constraints that shape the feasibility and pace of change, with varying levels of technical support, repository infrastructure, and dedicated RDM staff. Studies of RDM service development in African and other resource-constrained higher education systems emphasise challenges such as limited staffing, uneven network infrastructure, and competing institutional priorities [9, 10]. A FAIRness audit in such a setting must therefore be sensitive to the fact that some aspects of FAIRness are relatively low-cost to implement, while others require sustained investment, community coordination, and often external infrastructure. For instance, assigning persistent identifiers and ensuring that basic descriptive metadata fields are populated can often be achieved through modest enhancements to repository workflows and staff training. By contrast, implementing rich semantic interoperability through domain ontologies or machine-actionable metadata frequently depends on the availability of domain standards, software tooling, and specialised expertise that may not yet be well established locally [15].

The framework proposed in this study acknowledges these constraints by emphasising pragmatic, observable indicators that can be improved incrementally. It focuses on what can actually be seen and acted upon in the current repository environment, such as the presence or absence of licences, the completeness of basic metadata fields, the choice of file formats, and the availability of human-readable documentation. This does not imply that more ambitious aspects of FAIRness, such as fully machine-actionable metadata or tightly integrated domain ontologies, are unimportant. Rather, it recognises that progress toward these ideals may need to proceed in stages, with initial efforts directed toward solidifying baseline practices and building institutional capacity. The audit can thus help institutions set realistic priorities and sequence their interventions in a way that matches their resource realities, much as phased approaches to RDM service development have been advocated in other low- and middle-income contexts [9, 10].

At the same time, the audit highlights the importance of aligning local practices with international standards and expectations to the greatest extent possible. South African researchers increasingly participate in global research collaborations and contribute to datasets and infrastructures that transcend national boundaries. If locally stored datasets are not easily discoverable, accessible, interoperable, or reusable, they risk being marginalised within global data ecosystems, even when the underlying research is scientifically robust and socially relevant. By identifying gaps in FAIRness, the audit draws attention to where alignment with international norms is weakest and where enhancing compatibility would yield the greatest benefits in terms of visibility, reuse, and impact. This alignment is not a matter of uncritically importing external models but of adapting and negotiating standards in ways that respect local constraints and priorities while ensuring that South African data are not isolated from broader developments [15, 16].

4.3. Bridging the Gap Between Policy, Attitudes, and Artifacts

By focusing on actual datasets rather than self-reported practices, the FAIR audit complements survey-based research such as the work of Bangani and Moyo [14]. Taken together, these perspectives illuminate the complex relationships between formal policies, researcher attitudes, and the concrete

artefacts of research, namely datasets and metadata records. Policy documents and institutional strategies articulate expectations about how data should be managed and shared; surveys capture how researchers understand, value, and claim to implement these expectations; and audits of repository content reveal how these expectations and attitudes translate into tangible outcomes. Only by considering all three components can we develop a realistic picture of where the research system stands and what kinds of interventions are likely to be effective.

If the audit reveals that FAIRness remains low in institutions where surveys report positive attitudes and strong willingness to share, this would point toward structural obstacles that prevent researchers from acting on their intentions. International studies repeatedly document such attitude–behaviour gaps, where researchers endorse data sharing in principle but share relatively little in practice, often citing lack of time, support, or clarity about procedures and responsibilities [2, 21–24]. Such obstacles might include a lack of clear, practical guidance on how to prepare datasets for sharing; insufficient time and recognition for the labour involved in documentation and curation; repository interfaces that are difficult to use or that do not support the kinds of metadata researchers need; or concerns about legal and ethical liability that are not adequately addressed by institutional policies. In such cases, the combined evidence suggests that interventions should focus on clarifying procedures, providing hands-on support, and adjusting incentive structures, rather than simply exhorting researchers to be more open [3].

Conversely, if FAIRness scores are higher than might be expected from survey responses, this may indicate that repositories and support units are able to impose or facilitate good practices even when individual researchers are sceptical or uncertain. For example, librarians might apply metadata standards consistently during the deposit process, or repository workflows might require the selection of a licence and the provision of a minimal documentation bundle, thereby raising the baseline level of FAIRness regardless of individual attitudes [17, 18]. This would highlight the potential leverage of institutional infrastructure and centralised support in shaping data practices, suggesting that investments in these areas can compensate, at least to some extent, for uneven cultural buy-in.

In either case, the combination of policy analysis, survey evidence, and FAIRness auditing can guide more nuanced and effective interventions than any single perspective alone. Rather than assuming that poor data practices are simply the result of ignorance or resistance on the part of researchers, institutions can use the audit to identify specific points in the data lifecycle where practices break down and to design interventions that address these bottlenecks. This might involve revising policy language to make expectations clearer, embedding RDM considerations in ethics and funding review processes, integrating data management planning into graduate training, or redesigning repository interfaces to better support FAIR-compliant deposits [19, 20]. The ultimate aim is to move from a situation in which FAIRness is primarily a rhetorical aspiration to one in which it is manifest in the everyday artefacts and workflows of research, thereby bridging the gap between policy, attitudes, and the realities of data and metadata as they appear in repositories.

5. Conclusion

This paper has proposed a systematic framework for auditing the FAIRness of research datasets deposited in institutional and disciplinary repositories at South African universities. Motivated by the recognition that existing evidence on research data management practices is dominated by self-reported surveys and policy documents, we have argued for the complementary value of examining

observable artefacts of data stewardship: the dataset records that are actually preserved and exposed in repositories. By translating the FAIR Guiding Principles into a set of operational, record-level indicators and by specifying a multi-repository, stratified sampling design, the study provides a concrete protocol for assessing the extent to which FAIR-aligned stewardship features are present in practice.

The proposed research design combines a transparent FAIRness rubric, explicit decision rules, structured coder training with reliability evaluation, and a fully specified analysis plan that emphasises uncertainty quantification and robustness checks. The illustrative figures and narratives included in the manuscript are intended to clarify the reporting workflow that a completed audit would support—from indicator prevalence to dimension subscores, overall summaries, and statistically defensible comparisons—and they are not presented as empirical findings. This methodological emphasis is deliberate: it ensures that when the audit is implemented, conclusions can be tied directly to observable evidence and replicated across institutions and time periods.

The framework has three immediate forms of value to the academic community and to research systems. First, it enables an evidence-based diagnosis of specific, actionable stewardship gaps (for example in licensing, documentation, or identifier practices) that can be addressed through targeted changes to repository workflows, deposit requirements, and support services. Second, by prioritising observable and comparable indicators, it is feasible in settings where repository capacities and disciplinary practices are heterogeneous, allowing incremental but meaningful improvements to be identified and monitored. Third, by complementing survey-based studies and policy analyses, the audit helps to connect mandates and attitudes to the concrete state of repository-visible records, supporting more nuanced strategies for promoting FAIR data.

While the framework is tailored to the South African context, it is designed to be adaptable. Other national systems, consortia, or disciplinary communities can adopt and extend the rubric, adjust the sampling strategy to their repository landscapes, and incorporate additional indicators that reflect local standards and priorities. Future work should implement the audit at scale, report empirical results with uncertainty and robustness analyses as specified here, and explore staged integration of semi-automated FAIRness assessment where repository metadata practices are sufficiently mature. Taken together, such efforts can help ensure that FAIR principles are realised as verifiable stewardship outcomes rather than as aspirational policy language.

References

[1] Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., ... & Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PloS one, 6*(6), e21101.

[2] Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., ... & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PloS one, 10*(8), e0134826.

[3] Tenopir, C., Rice, N. M., Allard, S., Baird, L., Borycz, J., Christian, L., ... & Sandusky, R. J. (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PloS one, 15*(3), e0229003.

[4] Borgman, C. L. (2017). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT press.

[5] Andrikopoulou, A., Rowley, J., & Walton, G. (2022). Research data management (RDM) and the evolving identity of academic libraries and librarians: A literature review. *New Review of Academic Librarianship, 28*(4), 349-365.

[6] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data, 3*(1), 1-9.

[7] Boeckhout, M., Zielhuis, G. A., & Bredenoord, A. L. (2018). The FAIR guiding principles for data stewardship: fair enough?. *European Journal of Human Genetics, 26*(7), 931-936.

[8] Jacobsen, A., Kaliyaperumal, R., da Silva Santos, L. O. B., Mons, B., Schultes, E., Roos, M., & Thompson, M. (2020). A generic workflow for the data FAIRification process. *Data Intelligence, 2*(1-2), 56-65.

[9] Masenya, T. M. (2021). Research data management practices and services in South African academic libraries. *Library Philosophy and Practice*, Article 6311.

[10] Chiware, E., & Mathe, Z. (2015). Academic libraries' role in research data management services: a South African perspective. *South African Journal of Libraries and Information Science, 81*(2), 1-10.

[11] Patterton, L., Bothma, T. J., & Van Deventer, M. J. (2018). From planning to practice: An action plan for the implementation of research data management services in resource-constrained institutions. *South African Journal of Libraries and Information Science, 84*(2), 14-26.

[12] Cox, A. M., & Pinfield, S. (2014). Research data management and libraries: Current activities and future priorities. *Journal of Librarianship and Information Science, 46*(4), 299-316.

[13] Pinfield, S., Cox, A. M., & Smith, J. (2014). Research data management and libraries: Relationships, activities, drivers and influences. *PLoS one, 9*(12), e114734.

[14] Bangani, S., & Moyo, M. (2019). Data sharing practices among researchers at South African Universities. *Data Science Journal, 18*, 28-28.

[15] Mons, B. (2018). *Data Stewardship for Open Science: Implementing Fair Principles*. Chapman and Hall/CRC.

[16] Bahim, C., Casorrán-Amilburu, C., Dekkers, M., Herczog, E., Loozen, N., Repanas, K., ... & Stall, S. (2020). The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments. *Data Science Journal*, *19*(1), 41.

[17] Henderson, M. E., & Knott, T. L. (2015). Starting a research data management program based in a university library. *Medical Reference Services Quarterly, 34*(1), 47-59.

[18] Tang, R., & Hu, Z. (2019). Providing research data management (RDM) services in libraries: Preparedness, roles, challenges, and training for RDM practice. *Data and Information Management, 3*(2), 84-101.

[19] Xu, Z., Zhou, X., Kogut, A., & Watts, J. (2022). A scoping review: Synthesizing evidence on data management instruction in academic libraries. *The Journal of Academic Librarianship, 48*(3), 102508.

[20] Rod, A. B., Hervieux, S., & Lee, N. (2024). Evaluating an instructional intervention for research data management training. *Evidence Based Library and Information Practice, 19*(1), 114-131.

[21] Federer, L. M., Lu, Y. L., Joubert, D. J., Welsh, J., & Brandys, B. (2015). Biomedical data sharing and reuse: attitudes and practices of clinical and scientific research staff. *PloS one, 10*(6), e0129506.

[22] Kim, J., Hwang, H., Jung, Y., Cho, S. N., & Seo, T. S. (2023). Data sharing attitudes and practices of researchers in Korean government research institutes: a survey-based descriptive study. *Science Editing, 10*(1), 71-77.

[23] Thoegersen, J. L., & Borlund, P. (2022). Researcher attitudes toward data sharing in public data repositories: a meta-evaluation of studies on researcher data sharing. *Journal of Documentation, 78*(7), 1-17.

[24] Zabijakin-Chatleska, V., & Cekikj, A. (2020). Attitudes and practices of data sharing and data preservation among social science researchers in the Republic of North Macedonia. *Balkan Social Science Review, 15*(15), 251-275.

How to cite this article: Siviwe Bangani (2025). Are Research Data Really FAIR? A Metadata Quality Audit of Research Data Repositories at South African Universities. *Bulletin of Computer and Data Sciences*, 6(2), 1-19. DOI: 10.71448/bcds2562-1

Received: 21/12/2024 **Revised:** 21/04/2025 **Accepted:** 19/05/2025 **Publish:** 30/06/2025