

<https://doi.org/10.59298/NIJEP/2026/614959>

Standards, Datasets, and Transparency Norms in Digital Humanities Reproducibility

Mugabo Kalisa G.

Faculty of Engineering Kampala International University Uganda

ABSTRACT

This paper examines standards, datasets, and transparency norms as foundational pillars of reproducibility in Digital Humanities (DH). It situates reproducibility within the broader computational turn in humanities scholarship, where large-scale datasets, algorithmic methods, and digital infrastructures increasingly shape research practices. While reproducibility has become a central ideal in DH, the paper argues that its implementation remains complex due to the interpretive, contextual, and diachronic nature of humanities inquiry. It critically explores the distinctions between reproducibility, replication, and reusability, highlighting how these concepts intersect yet serve different epistemic purposes. The study further investigates how standards, metadata frameworks, and interoperability protocols, such as FAIR principles, Dublin Core, RDF, and DOI systems attempt to structure data sharing and enhance transparency. Attention is also given to workflows, provenance tracking, ethical constraints, licensing regimes, and infrastructural limitations that influence dataset accessibility and long-term sustainability. Ultimately, the paper contends that while transparency norms and standardized datasets enhance accountability and research visibility, they also introduce tensions related to interpretive flexibility, intellectual property, and disciplinary diversity. It concludes that reproducibility in Digital Humanities should be understood not as a fixed technical benchmark but as a negotiated, context-sensitive practice embedded in scholarly ecosystems.

Keywords: Digital Humanities; reproducibility; datasets; transparency norms; and metadata standards.

INTRODUCTION

Historical scholarship and literary studies have increasingly adopted the computational turn characterizing digital humanities [1]. Sociologists Avilova et al. assert that “through large-scale data collection and automated analysis, digital humanities gather, preprocess, analyse, and visualize archival data texts, images, languages, and built environments.” In this new environment, automated methods become comparable to laboratory or experimental practices supporting reliability demands [1]. In particular, the notion of reproducibility has gained symbolic strength and is increasingly adopted by research institutions. However, the challenge lies in balancing strict reproducibility with the interpretive nature of humanities research, which involves complex diachronic qualitative corpora requiring hermeneutic approaches [2]. Research procedures in the digital era often rely on pre-existing datasets collection, curation, and access to these datasets remain important concerns [2]. The integration of existing datasets can lead to zero investment in these topics, and yet the presence of datasets does not signal adherence to scientific and technological solutions [3]. Inspection of the datasets themselves without datasets description cannot guarantee their suitability, and data cleaning is an almost ubiquitous pre-processing step. Datasets presumably prepared for research may still conceal systematic, structural biases and problematic representations that directly impede factual analysis without prior exhaustive steps. Collecting hundreds of datasets may further catalyse the abandonment of evaluation [2]. Nevertheless, datasets become rapidly available on every topic and continue to circulate massively.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

Conceptual Foundations of Reproducibility in Digital Humanities

Reproducibility has gained symbolic strength and is increasingly adopted within the digital humanities (DH). The publication of code and the availability of corpora are now required to ensure the reproducibility of DH methods [1]. Digital research is furthermore often viewed through the lens of a cartesian ideal [2]. Such an emphasis on reproducibility and explicability can however seem perplexing: a similar compulsion for explicability does not appear to extend to the presentation of non-computational, multimedia or multimodal approaches, nor to text that relies on explicative figures of speech meta-critical, narratorial and allegorical for that matter. Reinforcement of the ideal of reproducibility may nevertheless contribute to micro-political strife in DH by introducing a constraint that artists of other media can evade [3]. The aspiration to reproduction, replication and fungibility has historically been characteristic of the hard sciences and some of the social sciences, and it continues to feature prominently in the discourse of DH[4]. When DH methods and approaches are seen to infringe on a purported standard of rigour, in a milieu long contaminated by the presence of mathematical formalism and scepticism regarding humanities object-hood, the rhetorical means whereby their remaining reproducibility is asserted sometimes equate DH with a notion of (digital-)social science that remains contested[5]. Attempting to demonstrate the reproducibility of aspects of DH in such a context may render it vulnerable to mischaracterization as a sub-field of so-called difficult DH and might risk portraying it as a boundary condition on the field. Contaminated reproduction discourse in DH is therefore deserving of close attention [6].

Definitions and Scope

The Digital curation centre defines reproducible research as the process of publishing research materials, methods, data and computational related materials as a set of materials sufficient to enable an independent researcher to repeat a given experiment [2]. Reproducibility implies independence in recreating results by having the recipients of the research materials reproduce results obtained earlier with similar resource by the original researcher. Reproducibility of results in research and research methods are ideal to observe standardization and transparency in research. [1]. A completely transparent, reproducible research task across humanities and digital humanities using pre-registration, sharing data, processing data and sharing methodology is difficult in practice. Barzaghi et al states that research areas must be addressed even when using visualization methods and tools [3]. Transparency remains an issue where only a subset of all digital humanities research projects that describes fully transparent procedures is published [2]. Other challenges remain to ensure that the whole research task including digits publication and any digital artefacts linked to the research is reproducible [1].

Distinctions among Replication, Reproduction, and Reusability

Replication, reproduction, and reusability refer to different scientific processes, but they are frequently confused. This ambiguity creates obstacles for researchers and staff when they seek to apply the ideal of “reproducible research,” particularly within a humanities context [6]. Replication, reproduction, and reusability are distinct but complementary concepts [7]. Replication entails the execution of the same experiment, with the goal not of obtaining the same result but of validating an original finding. In other words, replication serves as an independent check on a prior publication [8]. Reproduction involves the collection and analysis of new data to evaluate the credibility of an original claim, generally by applying the same computational steps. Finally, reusability refers to the degree to which data and scripts may be freely assessed, adapted, and combined in new work [2]. Reusability practices also contribute to transparency. The curation, annotation, and integration of materials to facilitate origination from paraphrased sources significantly enhance both transparency and reuse. Comprehensive inventories of datasets harvested during experimentation may be directly deposited into repositories, especially for diverse projects requiring elaborate data combinations [3]. Consequently, reusability extends the timeframe usually assumed in the “reproducibility” or “verifiability” paradigm to embrace a broad range of earlier and subsequent operations [1].

Standards for Reproducible Digital Humanities Research

Publications in the digital humanities seldom provide sufficient information to assess the reproducibility of analysis [5]. The definition of standards relevant to the field remains so broad that practitioners do not know if it applies to their work. According to L. Jones, wide-ranging guidelines for reproducibility in the digital humanities often include recommendations that fall outside the framework of research as traditionally defined by the discipline [6]. Standards that align closely with the narrower definition of research are not widely accepted. By contrast, R. Barzaghi et al., call for a specification of reproducibility that is less restrictive and can accommodate a broader range of investigations across the digital humanities [7]. In some quarters, standards are also considered undesirable. A. Joyeux-Prunel argues that the strict standards of reproducibility adopted by many research institutions and mandated by high-tier journals impose an excessively technical discourse and unjustly question the legitimacy of outcomes that do not satisfy such requirements. Standards must respect the corpus itself and remain sensitive to the historical, disciplinary, and societal context [8]. Shared semantics in a digital humanities

NIJEP Publications

context inform the provenance and citation of data and metadata as well as the documentation of how they have been transformed, each represented by a unique identifier across time and space [9]. Providing documentation of how data were assembled, cleaned, processed, and interpreted enhances the transparency of the workflow and the description of adjustments made to the publication, thereby facilitating further investigation [2].

Data Standards and Interoperability

The diversity of data representations across disciplines and databases presents a major challenge to analysis through interoperability [4]. Most datasets fall into heterogeneous categories of text, image, connectives, symbolic representations, transaction sequences, and structured tabular data [5]. Recognizing this diversity, many communities design frameworks for interoperable tools and data, such as the W3C Graph Data Model [6] for representation of ontologies, the Dublin Core Metadata Initiative (DCMI) and the resource Description Framework (RDF, [Link] for metadata and knowledge representation, and standards for highly structured data such as FITS (Flexible Image Transport Systems) for astrophysics data and the instrument Data Model of the International Virtual Observatory Alliance (IVOA) [1]. Metadata are the main means for describing and understanding datasets and providing the maximum number of interoperable links with other datasets [6]. The DCMI standard for administrating metadata has been proposed widely throughout the humanities, but can become cumbersome if a large number of concepts are involved [8]. Metadata usually take the form of a simple table (author, date, title, etc.), but as datasets and associated knowledge extend it is more concise to define specifications using a dedicated ontology [9]. Ontologies provide documentation for datasets and guide tools for building datasets, intersubjective contributions, browsing, and navigation both locally and remotely [10].

Provenance, Metadata, and Documentation

The digital humanities, as means for generating databases, corpora, edited texts, visualizations, and network models, should give more attention to reproducibility [7]. Different distributions of widely shared materials or models ensure cross-fertilization of methodologies. Documentation of digital research and publication of data should go beyond conventional itemization [3]. Research materials and protocols should be specified to permit meaningful secondary exploration [2]. Investigating reproducibility in the digital humanities requires definition of both scope and scale. Scope refers to the materials or concepts that the author aims to make available for further use, whereas scale identifies the degree of specifications or elaboration necessary to enable reproduction [1]. The research exquisitely developed by an art historian concerning the activities of a well-known gallery could be shared in a spreadsheet listing the artists represented and other significant characteristics, thus permitting systematic macro-evaluation of gallery commercialisation policies [4]. A historical study addressing the recovery of a specific document held by a French archive could usefully be coupled with a sample of archives from which recovery has been established, together with a description of the extensive regulations protecting humanitarian documents [5]. Research at a global scale by a historian ranking the partitions occurring throughout the tumultuous twentieth century and the frame of reference adopted by a historian pointing to the lists of partitions, consequently clustering texts with similar characteristics according to a variable approach adds an important dimension to the plausible dissemination of content [6].

Computational Workflows and Workflow Management

A computational workflow describes a structured sequence of tasks for processing data, wherein each task can be marked as completed when producing results consistent with expectations [6]. Typical elements of a workflow definition include data inputs, intermediate and final outputs, parameters, processes, and execution order. Workflows can be linear, branched, hierarchical, or nested, and may involve conditional branching and feedback loops [7]. Workflow management denotes the use of a specialized tool to define, execute, and store a computational workflow; typical functions include syntax checking, preparation of execution context, execution of tasks, concurrent task execution, and management of semaphores, lineage tracing, and logging [8]. Reproducibility is enhanced by toolbox-independent workflow descriptions, allowing identical semantics to be processed by different tools [1]. Curation standards and conventions for archiving workflows and specifying the required operating environment have been articulated [2].

Software Licensing and Version Control

Reproducibility standards for software underpin hardware technologies, file formats, and coding languages. These standards are heavily influenced by the digital object identifier (DOI) system. Since its inception in the 1990s, the DOI has catered primarily to the needs of publishing scientific articles [6]. However, the growing diversity of digital objects routinely marked by DOI requires adaptations to the DOI scheme 1 for software objects, datasets, 3D models, audio files, and scientific reports (Allagui & Gagou; Gagou) [7]. Fully reproducible computer research overlooks the fact that the use of standard forms of coding language and the display of complete code, however desirable, are not sufficient for social sciences, personal sciences, and humanities [8]. Efforts must also be made to cite the processing and transformation of the dataset and the specific version of the transformation [9].

Datasets in Digital Humanities

Public-digital scholarship often draws upon datasets with long-standing provenance, yet the complexity of data curation and the need for formal consent frequently hamper their use [8]. Archives may contain materials added long after the original dataset was compiled, complicating provenance tracking, and conditions of use are often unclear. Such difficulties are repeatedly encountered in digitized cultural heritage datasets, resulting in watered-down or inconsistent access conditions that are not publicly communicated [9]. A wide range of publicly accessible datasets for network text analysis exist, yet many are subject to ambiguous or unclear access conditions [10]. Large-scale textual databases and cultural heritage corpora for text analysis represent several different types of cultural good, presenting numerous possibilities for assertion of rights and use. Multiple national institutions have offered careful reflection on data governance, yet correspondence with institutional personnel nonetheless reveals that continuity of funding and maintenance of clear position on access and conditions of use remain serious, unresolved issues [11].

Data Curation, Cleaning, and Stewardship

The curation, cleaning, and stewardship of datasets constitute essential steps for ensuring that textual datasets used in digital humanities research are at a level of quality, traceability, and documentation sufficient for reproducibility [7]. This framework draws on the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, which have rapidly become a widely accepted norm in the representation and dissemination of datasets across scientific disciplines. The framework also recognizes the extended definition of datasets presented in the UNESCO Text and Data Mining (TDM) guideline (UNESCO), which encompasses any “data collected, organized, and curated for the purposes of original research and analysis” and includes associated software [8]. The dataset must not only be accessible to other researchers, but it must also be sufficiently documented so that users can fully understand it and its context, as well as the manner in which it has been processed, in order to effectively apply the same analyses, re-use the content for other purposes, or build further upon it [9, 1].

Ethical Considerations and Consent

Publication ethics recognizes that privacy, anonymity, and confidentiality warrant particular attention in the humanities. The legislative framework is provided by the Framework for the Protection of Human Subjects (Common Rule) [5]. Researchers in the USA employing human subjects in research must address ethical considerations specific to these activities [1]. In the highly interdisciplinary landscape of the digital humanities, ethics related to human subjects has merited increasing attention in recent years; the decision by the National Endowment for the Humanities and the National Archives to enhance the ethical review of humanities grants incorporating such subjects exemplifies this [6]. Other funding procedures exist alongside the Common Rule and analogous regulations across the globe. When such considerations arise, obtaining approval from an institutional review board (IRB) of the National Institutes of Health or another entity is necessary prior to collecting data from human subjects or an exempt determination [7]. Accordingly, government and private entities, including the U.S. and Canadian federal funding agencies in 2023, provide dedicated resources to safeguard research subjects. Ninety-three per cent of humanities projects reported as publishing this information on the case management software Maelstrom and additional platforms acknowledge such ethical concerns, with some even submitting documentation to approvals and positions granting exemption from review [8]. As another funding opportunity promotes transparency concerning group readiness for undertaking potentially harmful research on sensitive materials while delineating diverse operative definitions of “ethical” [9].

Public Datasets, Access Policies, and Sustainability

Many present-day digital humanities scholarship practices are susceptible to distinct types of limitations on replicability that have withstood many postulation attempts [5]. A range of methodologies from different fields such as semiotics, ontology, and network science can be brought into play to securely demonstrate the significance of the digital humanities and their cornered landscape [10]. It is crucial to adopt measures that limit prevailing misconceptions of global digital humanities scholarship and articulate novel models to fraction open the meaning-making practices governing scholarship in the digital world [11]. Addressing questions about the meaning of the digital humanities in a networked world requires circumspection that a few texts have so far failed to deliver: namely, whether modelling the intellectual object as network might furnish a shared and unifying approach to the digital humanities and correlated disciplines [1].

Transparency Norms and Scholarly Communication

These projects employ computational techniques such as quantitative analysis, visualization, and artificial intelligence to address issues of reproducibility in human sciences [7]. A procedural definition of reproducibility is impractical due to the interconnectedness of corpus, method, and interpretation. Multiple types of reproducibility should be considered, with a focus on ‘post-computational reproducibility’ as a more rigorous standard. This approach aims to prevent digital humanities from isolating themselves and instead sharing discoveries broadly

NIJEP Publications

[8]. Digital humanists have endeavored to establish credibility by disclosing datasets, algorithms, and publishing data papers, often mirroring computer science publication formats [8]. However, these efforts are often insufficient to communicate findings effectively to those outside the field, especially the broader public and traditional humanists. The increasing technicality in reporting practices can make digital humanities research less accessible to colleagues lacking computational expertise [9]. The tremendous potential of the Internet, networks, and digitization has encouraged different sectors and domains to share and communicate information and knowledge by opening up access, especially in research. There is an unprecedented challenge and promise for the sustainability of knowledge capital in the long run, and the “openness” of knowledge capital has become a crucial means to ensure its sustainability [10]. These ideas are particularly important to digital humanists, who seek to share and disseminate knowledge in a time when the sustainability of knowledge capital is at stake. Academic publishing in the traditional manner still limits circulation within the “invisible college [11].” For digital humanists, sharing and disseminating knowledge via academic publications, as traditionally defined, is insufficient. In the domain of human sciences, some digital projects employ computational techniques, from quantitative analysis and visualization to artificial intelligence (AI), that raise the question of reproducibility [10]. Yet a procedural definition of reproducibility, in which corpus, method, and interpretation can be disentangled, remains impracticable [2]. Multiple types of reproducibility should be envisaged, among which “post-computational reproducibility” emerges as a notional ideal deserving further elaboration [1]. Centrality, or, rather, sacrality, of the corpus precludes indulging in method agnosticism. To avoid falling into a retrograde isolation, broader dissemination of findings is preferable [11]. Digital humanists have therefore endeavored to assure credibility by the disclosure of datasets and algorithms companioning significant publications, sometimes escalating to the provision of data papers mirroring the conventions of computer science [7]. Despite continuing efforts to accomplish reproducibility through conventional formats and repositories, communication is seldom effective outside the digital sphere, within the larger collectivity of the scholarly enterprise, among the educated general public, and especially among humanists trained in disciplines other than those now termed “digital humanities [8].” The increasing technicality of documentation further compounds the situation and may hinder exchanges even among digital humanists lacking computational expertise [9].

Open Access and Preregistration Practices

The adoption of open-access publishing has led to the widespread dissemination of scholarly works, encouraging reuse and facilitating the development of interconnected knowledge [6]. These shared works facilitate greater scrutiny and support the integration of additional information within materials, such as extended empirical evidence or theoretical formalizations. The Digital Humanities community has emphasized the importance of openness within its scholarship [7]. In parallel, preprint servers like arXiv have reinforced this movement by encouraging the dissemination of works prior to formal peer review [1]. Digital Humanities researchers have also embraced supplementary pre-registration, specifying research questions and hypotheses prior to data analysis on a dedicated internet platform.

Reproducibility in Publications and Supplementary Materials

Publication practices, including the form and contents of articles, their peer review, and the material accompanying them, signal a research community’s commitment to reproducibility [2]. The degree of openness of a publication, the extent to which information is made freely available to the public, is a central aspect of this transparency. Two approaches have recently gained attention: “open access” and “pre-registration [3].” Both allow the community to assess a study’s reproducibility without requiring access to its software, data, or detailed methodology. Under both approaches, the publisher offers a supplementary material facility for authors wishing to share these additional items [8]. Open-access journals or repositories provide free access to the full text of articles on the publisher’s website. Such arrangements were once limited to publications housed in specialized repositories or low-impact journals [8]. The term now encompasses a wider range of venues, including respected high-impact journals and books in various disciplines [3]. Pre-registration punctuates freedom at the research-design stage and limits it to subsequent phases. Authors specify and document their intended research approach in advance method, data, and criteria for interpreting results, and subsequently publish their outcomes in a journal not ex ante assigned to them [1]. In comparison to archives or repositories, pre-registration therefore permits wider dissemination of a study’s objectives within a community not necessarily attuned to its methods [3].

Peer Review and Reproducibility Audits

Scholarly publications often undergo peer review prior to acceptance, with reviewers evaluating the approach, methodology, and interpretations of the published research [4]. In many cases, however, the reproducibility of the research is not assessed. Reproducibility approaches are even rarely a consideration in the review process. Although researchers apply standards to promote reproducibility throughout the research, little is done regarding considerations in peer review except for encouraging establishing broader and peer-reviewed communities [2].

NIJEP Publications

The issue, therefore, often involves identifying information that may provide additional consideration for reproducibility at different stages of research. It also involves dedicating time comparatively limited focus on peer review, according to the length of the research [1]. At various stages of the research process, supplementary materials and pre-registration offer potential avenues for improving consideration of reproducibility [7]. Published supplementary materials receive more consistent attention from proposed standards than other sorts of supplementary material, such as data availability statements [7]. Furthermore, the proposals also imply potential consideration attached to supplementary material, such as the formal assurance of access during and after the peer review process [8]. An alternative type of supplementary material, outlining adherence to open science or reproducibility practices, has also gained traction in the community. Pre-registration may also constitute a practice for improving the consideration of reproducibility during peer review [9]. Pre-registration defines submission of documented description of the research project to a registry prior to implementation of any data collection or analysis, with guaranteed access during the peer-review process [10].

Page | 54

Infrastructure and Tools Supporting Reproducibility

Reproducibility in research relies on effective infrastructure and tools. Best practices include documenting and sharing research data, software, and provenance [11]. Tools such as Docker-based packaging systems facilitate platform-agnostic implementations. Reproducibility efforts also involve imputation methods, data sharing policies, and community standards that promote transparency and data reusability. Platforms like virtual appliances and cloud computing environments support consistent and replicable research workflows [4]. To ensure that others can reanalyze research results, original datasets, and the tools necessary to process them need to be made accessible [3]. Assigning persistent identifiers (digital object identifiers, or DOIs) can encourage deposition by making datasets easier to cite, facilitate data discovery, and support data reuse across research projects [5]. Data citation enables greater emphasis on an essential part of the research process, often left undocumented, while also enhancing discoverability if authors cite the dataset in their own published work. Supporting persistent identifiers for individual datasets contributes to the long-term sustainability and traceability of richer scholarly outputs [3]. Alongside dataset publication, documenting the technical steps involved in processing datasets can improve the timeliness of analyses, clarify complex datasets and analyses, ensure that datasets remain up to date, and drive collaboration by enabling others to rerun analyses [4]. Comprehensive provenance capture that enables reproducibility, alongside already-archived datasets, instead addresses a different but equally important aspect. Even when datasets are unchanged, revisiting the question posed at the time of analysis can gain clarity or urgency; yet recreating an entire analysis is frequently impractical, time-consuming, or impossible [5]. Providing provenance information to accompany archived datasets can thus facilitate iterative collaboration [1].

Reproducible Research Environments

The concept of reproducibility has been widely discussed in digital humanities (DH) since the emergence of computing in the 1950s, and several approaches for its operationalization have been proposed [5]. Considering reproducibility as a broad umbrella term encompassing different degrees of freedom, a framework enables independent and original researchers to engage productively with prior computational work [1]. A specific case study investigates the reproduction of some results obtained with the Textalyser software applied to a small but well-known 19th-century literary corpus [2]. A salient issue is whether humanness approaches or, more broadly, the digital humanities must comply with an agenda of strict reproducibility inspired by machine learning, the goal being control, verification, generalization, and assurance of portability [6]. The software used for the study, written mainly in R, was published along with the unaltered raw corpus and an extensive documentation workflow. All files are hosted in the GitHub repository to ensure persistent access [7]. The study was neither preregistered nor deposited in a formal archive, title pages, or dedicated repositories, thus raising the task of framing an appropriate approach to reproducibility [8]. A publication is needed that describes the subject of inquiry, the goals of the study, the rationale justifying the selected method, the description of the material used, and selection criteria that allow further access to the corpus, which has been published offline [9].

Data Repositories, Registries, and DOIs

Building on the contributions of other disciplines, the digital humanities have developed a variety of approaches to reproducibility [1]. These approaches are supported by an infrastructure of tools and practices that facilitate transparent research throughout the research cycle, fostering reproducibility and enabling “post-reproducibility” by allowing others to repurpose or build on initial work [2]. Data repositories, registries, and persistent identifiers such as Digital Object Identifiers (DOIs) promote the management, storage, and sharing of datasets [5]. Elevating the metadata of data repositories to the global infrastructure further enhances the discoverability and interoperability of research data [6]. In the United States, the Office of Science and Technology Policy (OSTP) seeks to increase access to the results of federally funded research, including not only articles but also the underlying data and codes. The Inter-university Consortium for Political and Social Research (ICPSR) and the

NIJEP Publications

National Archive of Historical Data (NAHDAP) are examples of suitable deposit locations that ensure data longevity and continued accessibility [7]. Acceptable data repositories should provide the required characteristics of effective data management and sharing. Metadata standards, including Dublin Core, together with the FAIR principles (findable, accessible, interoperable, and reusable), improve the discoverability of datasets [8]. Initiatives such as the Open Archives Initiative Protocol for Metadata Harvesting facilitate the harvesting of metadata from data repositories [9]. The DataCite schema and associated DOI services support the persistent identification of datasets and research outputs, while the Research Organization Registry (ROR) facilitates the persistent identification of institutions [10].

Provenance Tracking, Logging, and Audit Trails

Provenance tracking has been proposed as a promising approach to address the reproducibility crisis in data-oriented disciplines such as computational biology, botany, and particle physics [6]. The term provenance refers to information that describes the origins and history of data and the processes that create and transform it [7]. Contemporaneously, the lectures of LBO at the EAO yielded further epistemological insights into the reproduction of DH artefacts; thus connects institutional and epistemological aspects of reproducibility, provenance, and interpretability [8]. Provenance information records the context and history of data assets, detailing their origins, derivatives, and their transformations. It encompasses two distinct yet interconnected subfields: provenance metadata and provenance models [9]. Provenance metadata constitute the information that describes these aspects; provenance models specify the structure, semantics, and behaviours (e.g. the transformation mechanics) of such information [10]. Substantial efforts exist to develop automated methods for capturing provenance data associated with operations executed within software environments such as Jupyter notebooks and R Markdown [1]. Audit trails serve to assist the human-driven activities involved in the creation of artefacts such as DH publications, which frequently require complex workflows where, unlike in trials, no single section may constitute an unambiguous answer [8]. Such audit trails elicit questions related to provenance, the 'where do I come from?' of science, and interpretability, the 'why should I believe you?' aspect of computational artefacts in the humanities [9, 10].

Case Studies Illustrating Reproducibility in Digital Humanities

Reproducibility in the digital humanities can improve the credibility, reliability, visibility and long-term impact of research outcomes [2]. Three case studies illustrate how different types of reproducibility are being implemented in textual analysis, cultural heritage digitization, and collective scholarly datasets [3]. In a project on the literary and linguistic use of indefinite articles in British English, the model, the data and all scripts required to fully replicate the results were published as open-access materials [4]. The priority was to allow immediate reuse of the experimental set-up, to explore how the resources comply with existing standards and to demonstrate that use of tiers of abstraction; conversion routines and provenance recording facilitate reproducibility [2]. The Diachronic Corpus of Swedish also aimed at full reproducibility. The project addressed collection, wrangling, enrichment, modelling and rendering of language data [3]. It recognised the distinction between reproducing formalised, standardised research procedures and producing replicative results where output, experimental procedures and intermediate transformations are documented, even flexibly [1]. A further initiative, reSynth, set out to collect reliable datasets for the representation of syntactic, semantic and ontological knowledge. Only freely available data with formal or semi-formal annotations were sought [2]. Existing repository systems, though, were neither uniform nor sufficiently well-known to accommodate a clear, adequate articulation of the Linköping University Centre for Language Technology's data-enhancement strategy [3].

Textual Analysis and Linguistics Corpora

Research in textual analysis and computational linguistics would benefit greatly from shared corpora. The variation in spoken and written language, coupled with the diverse range of annotated corpora generally unattainable for a single research group, motivates this need [6]. Language datasets and preprocessing, along with the environments governing their use, have been described with the intent of facilitating their acquisition and adoption [1]. The spaCy library is primarily utilized, which includes a pipeline for standard analysis [7]. Preprocessing, however, is seldom reported, although it usually comprises basic cleansing, sentence splitting, tokenization, and text categorization [8]. Collaborative and transparent initiatives in language modelling have emerged [9]. The ongoing push for exposition from model designers has already encouraged supplementary materials ranging from supplementary text to complementary notebooks, which, for some additional datasets, serve as insightful publication substitutes [1].

Cultural Heritage Digitization Projects

Digitization of cultural heritage constitutes an essential yet challenging area of digital humanities practice. The breadth of archived materials, diversity of scientific domains, and variety of underlying research questions mean that the conditions necessary for reproducibility differ widely both between projects and within a single

NIJEP Publications

information source [2]. From a practical viewpoint, cultural heritage digitization projects establish practices aimed at ensuring transparency, reliability, and equitable access to research outcomes [9]. Defining clear data management practices is critical. Incremental documentation and retention of raw data constitute best practices for preventing data loss and facilitating reproducibility. Adopting automated, periodic backup workflows assures long-term retention of operational versions [10]. Recording extensive details of processing steps employed on original data contributes toward dimensionality reduction, while sharing this information enables peers to comprehend and replicate the work. Utilization of repositories issuing persistent identifiers establishes traceability and guarantees dataset availability [11].

Networked Scholarly Datasets and Semantic Interoperability

A growing number of Digital Humanities (DH) projects focus on developing networked datasets to enhance so-called Linked Data principles [8]. This approach aligns with the notion that information is more valuable when published, linked, and semantically unambiguous [7]. For example, within a FAIR-compliant system, a primary step ensures that domain datasets link to shared authorities like IdRef, ORCID, Wikidata, or Geonames [8]. This feature supports data linking across heterogeneous datasets and systems without requiring a dominant vocabulary standard, significantly broadening the visibility of the linked data [9]. Such data enhancement facilitates automatic knowledge-graph-building processes capable of reasoning over the data through semantic-web-compliant engines. Supplementary open datasets expand the interconnections available during the linking process, promoting the integration of solid, porous, low-frequency primary datasets with transient or event-based datasets [3]. Representing datasets in aligned formats further promotes integration and semantic understanding. DH efforts have addressed Linked Data in domains such as history and music but remain hampered by a lack of suitable datasets and explicit semantic links [9]. Current DH paradigms emphasize publishing, documenting, and archiving datasets according to provenance standards, stressing that data remains indisputably valid only when the original subjects and contexts are independently communicable across time and space [10].

Challenges, Limitations, and Ethical Considerations

Adopting standards, sharing datasets, and endorsing transparency norms encourage the reproducibility of computational analyses [5]. The potential biases introduced through choice of standards, datasets, and transparency norms also warrant scrutiny [6]. The imposition of standards and sharing of datasets can amplify systemic biases. Textual analyses relying on corpora drawn from online sources face challenges of cultural representation, relevance, and online multilingualism [1]. Unavailability of provenance information further conceals these choices. Providing access to datasets enables others to critique standards and datasets employed, mitigating the risk of opaque choices [7]. Transparency regarding licensing and access conditions for shared datasets is essential to advance serendipitous, unrestricted research endeavors and remove uncertainty regarding permissible analyses. By ensuring that users are aware of ethical limits to reproducibility, whether legal, institutional, social, or personal, researchers can avert contravention of their fundamental objectives [8].

Intellectual Property and Licensing Tensions

Transparency practices essential for reproducibility in the social sciences and the natural sciences affect progress in the computational arts and humanities [6]. Datasets, documentation, and computational workflows must be made accessible to establish temporal, causal, and correlative relationships between stimuli and responses. Preregistration clarifies conditions specifying the specific variables studied and the types of evidence sought prior to data collection. Well-documented, sharable, and executable research workflows permit re-use of the procedures that generate an author's contribution [7]. In the historical and literary studies, however, dominated by concerns for assessment of originality, trends toward transparency and disclosure face severe objections, as do similar tendencies within artistic, architectural, and musical domains that rely primarily on the interaction of prior works. The interaction of intellectual property (IP) with practices that enhance transparency and reproducibility constitute another serious difficulty [8]. Adhering to copyright restrictions, principal investigators and scholars must consider a complex web of institutional requirements, donor wishes, team contracts, agreements with collaborators, co-author concerns, and discipline norms before divulging materials [8]. Many repositories that share graphical and textual materials insist on explicit permission and, even then, recommend licenses, such as the Creative Commons Attribution-Noncommercial-No Derivative Works license, that remains highly constraining. These barriers hamper curatorial efforts and duration of engagements with some datasets [9].

Data Bias, Representation, and Inclusivity

Digital Humanities projects leverage tools such as 3D modelling and spatial humanities, drawing on the knowledge and methods developed in a vast range of interdisciplinary fields [5]. Unfortunately, rapid technological and methodological developments frequently compromise the sustainability of research, datasets, and scholar-led practices [4]. Throughout the entire lifecycle of a research project, transparency regarding the structure of datasets and the production process is essential for effective transmission and communication [6]. In

NIJEP Publications

parallel with the emergence of the socio-technical concept of FAIR (Findable, Accessible, Interoperable, Reusable) datasets, and with the substantial untapped corpus of carefully crafted records already produced, a focus on clearly articulating provenance and lineage can decrease the dissemination of incomplete and rapidly obsolete datasets [1].

Long-Term Sustainability and Funding Models

Digital research increasingly relies on sustainable funding models. Individual projects can rarely support the necessary infrastructure for long-term sustainability [7]. General disciplinary funding such as the Consolidator Grant from the European Research Council, especially when opened up to networks of smaller projects has helped many researchers [8, 9]. The King's Digital Lab, which manages over a hundred projects, devotes time at the outset to project sustainability, facilitating follow-on funding requests and providing a reliable guideline about the degree to which projects are likely to remain maintained [10]. Informatics-driven publications in cultural analytics or computational humanities still struggle to achieve sufficient traction whereas libraries, computer science departments, and similar locations can provide a much more robust framework to address them. The flexibility to adapt resourcing at various stages of a project is crucial [11]. Proposals that identify products without tools that remain active after the project finishes often are rejected. A clear financial scheme remains a work in progress, despite working with a number of different institutions, languages, and publishers [4]. The French "Évaluation et Valorisation de l'Information" program funds thematic networks to provide mutual assistance across projects in many disciplines, such as the "CulturAl-Med" network, which is gathering further momentum. For digital archives, which require formal definitions for statements assuring sustainability, a contributor-pays model integrating valorisation and sustainability might be worth considering [11]. The Centre for Digital Humanities at the University of Zurich has adopted best-practice recommendations for long-term sustainability proposed by the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Major themes are that the low-level costs of preservation should be part of research budgets from the outset; that long-term, low-variable-cost storage is necessary; and that the long-term and, if possible, permanent domain of data or resources should be determined, including appropriation for parallel or potential language shifts [1, 8]. The King's Digital Lab promotes and implements an integrated maintenance framework to safeguard the enduring usability of an extensive array of multi-language displays while minimising repository and storage development, in response to the increasing vulnerability of cultural analytics [1, 7].

Recommendations and Best Practices for Future Research

Well-defined, shared standards and norms for reproducibility in digital humanities, adapted to the community's specific conditions and needs, might support the development of reproducible scholarly outputs [5]. Several professionals in the domain suggest formulating an official charter on the subject [2]. Such a charter could establish a shared, community-transmitted understanding of reproducibility, specify a clear statement of principles, and clarify the framework of requirements for accompanying, endorsing, or developing practices in the humanities. Integrating reproducibility considerations into the design of computational pipelines and workflows encourages the preparation of materials and outputs suitable for dissemination inside and outside the scientific community [1, 6]. Reproducibility remains difficult in some instances (e.g., on account of the proprietary nature of materials; specific dependencies), yet when attainable, the incorporation of reproducibility-supporting packaging conventions enables the comprehensive reporting of required software, datasets, and documentation [7]. Several frameworks exist in the sector, addressing reproducibility at various stages of the research lifecycle (e.g., arrangements for version control and continuous integration), and may be applied to the production of preparation [8]. Alignment with widely recognised standards facilitates the sustainable preparation of valuable datasets worthy of community contribution [9]. The formulation of datasets should likewise encompass meaningful provenance and supplementary documentation, ensuring that data and associated authorship can be securely cited, that the intention and form of information remain usable, and that the policy of the dataset matches the context within which activities occur [10].

Developing Shared Standards and Communities

Adoption of formal standards in all materials that foster reproducibility in the arts and humanities threatens to generate overly technical assertions [9]. Reproducibility scholarship should not exercise undue technological determinism by assuming that these standards would necessarily be widely implemented, because their relevance is subject to historical, disciplinary, and societal contextualization [10]. In digital humanities, the drive for formal standards is complicated further by the typology of informational objects produced: humanist meanings can be diachronic, semantically multi-faceted, qualitative, and interpretive. The notion of operations, transformation, or manipulation reflexively escapes strict standards of reproducibility [1].

Designing Reproducible Computational Pipelines

Reproducibility of research findings lies at the heart of the scientific endeavour. However, the Digital Humanities, unlike traditional organisations of knowledge, face unique issues concerning the reproducibility of computational pipelines [9]. Within the framework of the Digital Humanities, computational pipelines refer to the set of procedures and operations applied to a dataset to generate the expected output, often consisting of new discoveries or means of accessing already-acquired knowledge [8]. Many constructionist approaches embracing computational pipelines scope five types of data: [7] corpus raw-data; data processing raw-code; final outputs; software specification; (v) pipeline documentation [2]. Constructionists extend this to remain aware of datasets originally collected for different objectives and to reuse at least one archival scientific footprint within the desired convention, such as equivalent identifications of projects, procedures, or output materials [1]. In the Digital Humanities, challenges such as copyright, data protection, value communication, and provenance transparency slow sharing efforts, hindering collective knowledge evolution and limiting the scope of exploratory research [7]. Overcoming such obstacles and designing reproducible datasets and pipelines would accelerate exploration cycles, unlock pattern emergence, sustain attributions across researchers and milestones, and maximise funding investments [8]. Moreover, there remain swathes of intellectual and artistic outputs awaiting discovery. Well-aided input-output pipelines operate beyond copyright restrictions, proposing an intermediary position balancing restricted access with creative dynamics. Tools exist to enhance accessibility and traceable augmentation; however, they often fall short of the required reality [4, 9].

Stewardship of Datasets and Transparent Reporting

Several new guidelines recommend a fair data ecosystem that supports machine and human-readable metadata facilitates the integration and connection of datasets, retains ownership of datasets, and ensures durability and accessibility over the long term [11, 1]. The desire to connect datasets, to build upon the scholarly work and corpus of others and to facilitate the organization of rich corpus collections facilitates reproducibility and inspires various initiatives for structured digital humanities datasets, such as the Text Encoding Initiative, an ongoing collaborative standard, Integrated Authority and Vocabulary Services (TAPoR), virtual research environments and Digital Research Infrastructure for the Arts and Humanities (DARIAH)[10].

CONCLUSION

This study has shown that reproducibility in Digital Humanities is deeply dependent on the quality, accessibility, and governance of datasets, as well as the transparency norms that guide their use. While standards such as FAIR principles, metadata schemas, and persistent identifiers have significantly improved the structure and accessibility of digital research materials, they do not eliminate the inherent complexities of humanities scholarship. The interpretive nature of humanistic inquiry means that reproducibility cannot be reduced to mere replication of computational outputs but must also account for context, methodology, and meaning-making processes. The analysis further demonstrates that datasets are not neutral entities; they are shaped by choices in curation, cleaning, representation, and institutional control. These choices can introduce bias, limit interoperability, or obscure provenance, thereby affecting the reliability of research outcomes. At the same time, transparency mechanisms such as workflow documentation, open access publishing, preregistration, and provenance tracking offer important pathways toward enhancing accountability and reusability in DH research. However, persistent challenges remain, including intellectual property restrictions, uneven access to data repositories, sustainability constraints, and the absence of universally accepted reproducibility standards tailored to humanities research. These limitations suggest that reproducibility should be understood as a flexible and evolving framework rather than a rigid requirement. In conclusion, strengthening reproducibility in Digital Humanities requires a balanced approach that integrates technical standards with disciplinary sensitivity, ethical responsibility, and long-term infrastructural support, ensuring that transparency enhances rather than constrains scholarly interpretation and innovation.

REFERENCES

1. Joyeux-Prunel B. Digital humanities in the era of digital reproducibility: Towards a fairest and post-computational framework. *Int J Digit Humanit.* 2024;6(1):23-43. doi:10.1007/s42803-023-00079-6.
2. Barzaghi S, Bordignon A, Gualandi B, Peroni S. Thinking outside the black box: Insights from a digital exhibition in the humanities [preprint]. *arXiv*; 2024. Available from: <https://arxiv.org/abs/2402.12000>
3. Stagge JH, Rosenberg DE, Abdallah AM, Akbar H, Attallah NA, James R. Assessing data availability and research reproducibility in hydrology and water resources. *Sci Data.* 2019;6:190030. doi:10.1038/sdata.2019.30.
4. Liu DM, Salganik MJ. Successes and struggles with computational reproducibility: Lessons from the Fragile Families Challenge. *Socius.* 2019;5:1-10. doi:10.1177/2378023119849803.

NIJEP Publications

5. Johnston LR, Hofelich Mohr A, Herndon J, Taylor S, Carlson JR, Ge L, et al. Seek and you may (not) find: A multi-institutional analysis of where research data are shared. PLoS One. 2024;19(4):e0302426. doi:10.1371/journal.pone.0302426.
6. Missier P. The lifecycle of provenance metadata and its associated challenges and opportunities. In: Lemieux VL, editor. *Building Trust in Information*. Cham: Springer; 2016. p. 127-137.
7. Beretta F. Semantic data for humanities and social sciences (SDHSS): An ecosystem of CIDOC CRM extensions for research data production and reuse [preprint]. arXiv; 2024. Available from: <https://arxiv.org/abs/2402.07531>
8. Chamanara J, Kraft A, Auer S, Koepler O. Towards semantic integration of federated research data. *Datenbank-Spektrum*. 2019;19(2):87-94. doi:10.1007/s13222-019-00315-w.
9. Meroño-Peñuela A. Digital humanities on the semantic web: Accessing historical and musical linked data. *J Catalan Intellect Hist*. 2017;1(11). doi:10.1515/jocih-2016-0013.
10. Smithies J, Westling C, Sichani AM, Mellen P, Ciula A. Managing 100 digital humanities projects: Digital scholarship and archiving in King's Digital Lab. *Digit Humanit Q*. 2019;13(1):000411.
11. Downs RR, Chen RS. Towards sustainable stewardship of digital collections of scientific data. In: *Proceedings of the Global Geospatial Conference 2012 (GSDI 13)*; 2012; Québec City, Canada. Québec City: GSDI Association; 2012.

CITE AS: Mugabo Kalisa G. (2026). Standards, Datasets, and Transparency Norms in Digital Humanities Reproducibility. NEWPORT INTERNATIONAL JOURNAL OF ENGINEERING AND PHYSICAL SCIENCES,6(1): 49-59. <https://doi.org/10.59298/NIJEP/2026/614959>