Data-Envelopes for Cultural Heritage: Going beyond Datasheets

Mrinalini Luthra, Maria Eskevich
Huygens Insitute, Koninklijke Nederlandse Akademie van Wetenschappen (KNAW)
Oudezijds Achterburgwal 185, 1012 DK Amsterdam
{mrinalini.luthra, maria.eskevich}@huygens.knaw.nl

Abstract

Cultural heritage data is a rich source of information about the history and culture development in the past. When used with due understanding of its intrinsic complexity it can both support research in social sciences and humanities, and become input for machine learning and artificial intelligence algorithms. In all cases ethical and contextual considerations can be encouraged when the relevant information is provided in a clear and well structured form to potential users before they begin to interact with the data. Proposed data-envelopes, basing on the existing documentation frameworks, address the particular needs and challenges of the cultural heritage field while combining machine-readability and user-friendliness. We develop and test data-envelopes usability on the data from the Huygens Institute for History and Culture of the Netherlands.

Keywords: machine-readable datasheets, cultural heritage, data ethics, transparency, auditability, FAIR

1. Introduction

The digitisation of historical collections presents opportunities for research and education, transforming how we understand and access the past, and define how the future for the past can be collectively shaped (Trouillot, 2015; McGillivray et al., 2020). However, this digital transformation is accompanied by significant ethical, legal, and practical challenges, especially as historical datasets become critical resources for not only academic scrutiny but also serve as fuel for advanced computational models. The complexity of these challenges necessitates a robust framework to guide the use of cultural heritage (CH) data, ensuring their accessibility, transparency, and ethical (re)use.

In response to these challenges, this paper presents the following contributions: i) we highlight the complexity of CH data, featuring the unique ethical and contextual considerations they entail on the example of materials that are offered by Huygens Institute; ii) we evaluate and compare existing dataset documentation frameworks, examining their suitability for CH datasets; iii) we introduce the "data-envelope"–a machine-readable adaptation of existing dataset documentation frameworks, to tackle the specificities of CH datasets. Its modular form is designed to serve not only the needs of machine learning (ML), but also and especially broader user groups varying from humanities scholars, governmental monitoring authorities to citizen scientists and the general public. Importantly, the data-envelope framework emphasises the legal and ethical dimensions of dataset documentation, facilitating compliance with evolving data protection regulations and enhancing the accountability of data stewardship in the cultural heritage sector. We discuss and invite the readers for further conversation on the topic of ethical considerations, and how the different audiences should be informed about the importance of datasets documentation management and their context.

2. Diversity in Cultural Heritage Data

In this section, we delve into the multifaceted nature of CH data, emphasising the specific ethical and contextual considerations that it necessitates. By examining data from the Huygens Institute for History and Culture of the Netherlands[1], part of the KNAW Humanities Cluster[2] we illustrate three key aspects of CH data: the extensive historical range of the collections, the unique contexts of their creation and aggregation, and the intricate data structures within these datasets. This institution is selected for its representative practices and data interaction types that are common within the CH sector in the Netherlands, suggesting that solutions identified here may be applicable more broadly.

The collections managed by the Huygens Institute showcase the evolution of CH data from physical to digital realms. Initially, data were selected and published in book form, starting in 1902 (Kooijmans and de Valk, 1985). This historical approach laid the groundwork for contemporary digital projects such as GLOBALISE[3], Oorlog voor de Rechter[4] [War in Court], and REPUBLIC[5]. These initiatives reflect the shift towards digital accessibil-

[1] https://www.huygens.knaw.nl/en/

[2] KNAW is an abbreviation of the Koninklijke Nederlandse Akademie van Wetenschappen [Royal Dutch Academy of Arts and Sciences]

[3] https://globalise.huygens.knaw.nl

[4] https://oorlogvoorderechter.nl

[5] https://republic.huygens.knaw.nl

LEGAL2024 @LREC-COLING-2024, pages 52–65
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

ity and the ongoing efforts to process and release data.

The nature of CH datasets, often spanning over centuries, is distinguished not only by their historical depth but also by their collection and selection processes. These processes, historically influenced by various biases, shape the datasets' structure and content. Digital historians and scholars, equipped with a deep understanding of the field's evolution and ongoing debates, approach these datasets critically, mitigating biases through careful analysis (Tasovac et al., 2020; Maryl et al., 2023). This scholarly perspective informs the data's structure, metadata quality, and its application, diverging from the requirements commonly associated with machine learning and artificial intelligence (AI) disciplines (Heger et al., 2022). Therefore, dataset documentation could benefit from integrating such rich contextual information, ensuring CH data is utilised responsibly and effectively in technological applications (Jo and Gebru, 2020).

Research in (digital) humanities utilises complex data structures and/or interconnected datasets to deepen historical understanding and introduce new insights into past events. On the one hand, scholars navigate numerous challenges, including handling low-resource languages, accommodating spelling variations, and correcting text recognition errors (Koolen et al., 2023). The diversity of document types and domains, coupled with language evolution and noisy inputs, further complicates analysis. On the other hand, the information about the same entity, such as a migrant person, can be scattered across different registries, archives, and other official documents as well as informal records collected by civil society organisations, churches, and other non-governmental organisations, thus varying in form, structure, and availability (Arthur et al., 2018). Moreover, the research might combine both analysis of the content of particular types of documents such as letters, and the way the communication was evolving through the network analysis (Hotson, 2019). To effectively utilise these rich historical resources, data brokers must provide comprehensive, accessible information on data limitations and considerations, ensuring users can fully engage with the historical context.

2.1. Retrodigitised Editions

Building on the exploration of the complexities of CH data, this subsection explores the specific case of retrodigitised editions. Historians frequently engage with these editions, which are historical documents compiled and commented on in book form, later digitised for broader access (Kooijmans and Th.S. Bos, 1985; Tollebeek, 1994). This process exemplifies the transformation of CH data across formats (varying from the actual physical instance to a plethora of data representations), highlighting the necessity for clear documentation on annotation and content transformation decisions. Such detailed documentation is crucial for users to understand the historical context and the interpretative layers added through digitisation, further illustrating the challenges and limitations of existing documentation frameworks mentioned above.

2.2. GLOBALISE: Commodities Dataset

The GLOBALISE project, focused on leveraging AI to transcribe and extract data from the Dutch East India Company (VOC) archives (Petram and van Rossum, 2022), underscores the limitations of current dataset documentation standards in digital cultural heritage. For instance, the documentation of the commodities dataset (Pepping et al., 2023)—detailing classifications and a thesaurus of commodities traded in the early modern Indian Ocean World—highlights these gaps. Existing templates fail to adequately capture the complexity of the provenance inherent in such datasets, derived from primary sources and enriched through the multiple secondary sources. Furthermore, they fall short in addressing the linguistic diversity and temporal scopes, both being crucial aspects for accurately documenting digital cultural heritage data.

2.3. Potential Legal and Ethical Issues with Huygens Institute Resources

Institutions such as the Huygens Institute combine running projects with managing access to legacy datasets (more than 200 in this case) which brings a lot of potential legal and ethical issues along the way:

- *Copyright:* In principle, within this institute copyright is less of an open issue, even though different copyright regulations apply to the datasets. In a lot of cases the copyright stays with the institute, as it publishes or has published the materials (Kooijmans and Th.S. Bos, 1985).
- *Licenses:* As Large Language Models (LLMs) increasingly rely on structured datasets for training, it is crucial to consider the potential risks associated with using cultural heritage data. Given the historical intricacies and biases inherent in cultural heritage data, there is a danger that LLMs trained on such datasets may inherit these biases. When applied in contemporary contexts, these models may perpetuate discriminatory practices and reinforce historical prejudices. Moving forward, it is important to develop strategies for mitigating the potential risks of bias amplification when making cultural heritage datasets available for LLM

training (Hicks, 2017; Thylstrup, 2019; Noble, 2018). Additionally, cultural heritage institutions need to navigate and rethink the landscape of intellectual property rights and openness in the era of generative AI. These institutions may need to adopt a more nuanced approach, differentiating between private users, researchers, and commercial entities, while also renegotiating license agreements and addressing technical challenges related to copyright protection in the context of AI (Lehmann, 2024).

- *Privacy:* There is a number of projects that make public and digitally accessible personal information which requires special attention and contextualisation. For example, the project “Oorlog voor de Rechter" (War in Court) aims at disclosure of archival documents about collaboration during Second World War[6], and another project, “Child Separation", works with the information about extraction of children from their indigenous context and putting them into foster care (Mak et al., 2020).
- *Information security:* Historical documents reflect different aspects of the past within the country, and when accessed and processed without proper contextualisation, this information might provoke wrongful assumptions, statements, and even prosecutions.
- *Ethical and Emotional:* Such considerations are particularly poignant in cases involving information about living individuals, such as data related to wartime collaboration or child adoption, which require sensitive handling to mitigate potential harm or distress (Wood et al., 2014).

3. Harmonising Machine Learning, Cultural Heritage, and Legal Insights

In the rapidly evolving digital landscape, the documentation of CH datasets emerges as a critical juncture where machine learning practices, cultural heritage stewardship, and legal compliance intersect. This section delves into the existing documentation frameworks, underscoring the limitations within machine learning paradigms, the unique complexities of cultural heritage data, and the increasing importance of aligning with legal standards. Through this examination, we highlight the imperative for a nuanced, comprehensive approach to dataset documentation that is answered by the proposed data-envelope framework.

[6] https://oorlogvoorderechter.nl/

3.1. ML perspective

Dataset documentation, often referred to as "datasheets", first introduced by Gebru et al. (2021) advocates for the inclusion of comprehensive documentation alongside machine learning dataset publications. Such documentation is envisioned to serve multiple critical functions: facilitating informed decision-making regarding dataset application, enhancing transparency concerning the datasets’ composition and creation, and establishing clear guidelines for dataset development (Gebru et al., 2021; Pushkarna et al., 2022; Library of Congress, 2021; Roman et al., 2023).

3.2. CH perspective

The complexity inherent in (digital) cultural heritage data transcends the technical dimensions typically addressed by machine learning documentation standards. These datasets are situated within diverse social, cultural, and historical contexts, often encompassing multiple perspectives and interpretations (Cameron and Kenderdine, 2007) as demonstrated in Section 2. The temporal and spatial complexity of the data adds another layer of challenge, as does the presence of uncertainties and incompleteness. Furthermore, cultural heritage data is often subject to copyrights, traditional knowledge, and intellectual property considerations (Torsen and Anderson, 2010). The collaborative nature of knowledge production in this domain necessitates careful attribution and recognition of contributors (Srinivasan et al., 2010; Powell, 2016).These factors collectively underscore the need for documentation practices that can adequately capture and convey the nuances and complexities of cultural heritage data (Candela et al., 2023).

A recent paper by the Datasheets for digital cultural heritage Working Group, set up within the Europeana Research Community and EuropeanaTech Community, has made a first attempt to documenting datasets from the cultural heritage sector (Alkemade et al., 2023). However, these initial steps, while pioneering, reveal gaps in usability, machine-readability, and the depth of coverage on critical issues like provenance, ethical, and legal considerations.

3.3. Legal Perspective

The legal landscape around data use and governance is undergoing significant transformation on both international and national levels. Legislations such as the EU Data Act[7], EU Data Governance

[7] https://digital-strategy.ec.europa.eu/en/policies/data-act

Parameter	Datasheets	Data Cards	Open Datasheets	Datasheets for DCH	Data-Envelope
Structure	Questionnaire format	Structured Summaries	JSON-based metadata	Tailored for DCH data	Modular with detailed sections
Machine Readability	Not primary focus		Yes, fully supported	Not primary focus	Yes, Designed for machine readability
Provenance	Not explicitly/sufficiently considered				Extensively covered
Target Audience	ML/AI researchers			ML/AI researchers, CH Institutions	CH Institutions, ML community, legal institutions, broader public
FAIR	Not directly addressed				Designed with FAIR in mind, with specific section devoted to datasets' adherence to FAIR principles
Positionality	Not emphasized, only mentioned for annotators				Explicit focus on creators, contributors, annotators' positionality

Act[8], and EU Artificial Intelligence (AI) Act[9] on the EU level and Archiefwet [Law about archiving in the Netherlands][10] introduce complex requirements for dataset documentation, transparency, and accountability[11]. However, in practice the current lack of standardised, machine-readable documentation frameworks complicates the actual compliance and auditing processes. Our contribution lies in the development of a comprehensive machine-readable documentation framework, which enables automated auditing of datasets, particularly in areas concerning data collection, sharing, and (re)use. By bridging the gap between legal requirements and technical documentation, the proposed data-envelope facilitates compliance with regulatory mandates, thereby enhancing transparency and accountability in data governance practices.

3.4. Advancing Documentation Practices

Table 3.1 outlines a comparison between different dataset documentation frameworks (Gebru et al., 2021; Pushkarna et al., 2022; Alkemade et al., 2023; Roman et al., 2023). The data-envelope offers a machine-readable structured data alongside qualitative narrative elements, thereby ensuring versatility. This approach not only supports the development of AI models but also addresses the educational and research necessities of cultural heritage, underpinning the importance of a well-rounded, accessible data documentation method. The data-envelope's particular emphasis on positionality (Harding, 2003; Haraway, 2016; Mignolo and Walsh, 2018) and adherence to FAIR principles (Wilkinson et al., 2016; Harrower et al., 2020) demonstrates its comprehensive approach to dataset documentation and accessibility.

4. Data-Envelopes for Datasets

We introduce the "data-envelope", intended to provide clear guidance for the creation and documentation of CH datasets, ensuring that their complexity and context are effectively communicated and preserved for current and future use.

4.1. Contextual Wrapper for Datasets

At its core, the data-envelope is conceptualised as a contextual wrapper for datasets. Going beyond existing documentation frameworks (Gebru et al., 2021; Pushkarna et al., 2022), the data-envelope encases the dataset within a comprehensive context that elucidates the cultural, historical, and social dimensions of the data. By situating data within this contextual framework, the data-envelope empowers users to comprehend not just the 'what' but also the 'why' behind the data they engage with. This method guarantees that any interpretations

[8] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020PC0767

[9] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206

[10] https://wetten.overheid.nl/BWBR0007376/2022-05-01

[11] https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729541/EPRS_STU(2022)729541_EN.pdf

and utilisations of the dataset are rooted in an appreciation of its origins and importance, thereby encouraging more informed and thoughtful applications (Mignolo and Walsh, 2018).

In the current data interaction model, depicted in Figure 1, the CH sector oversees the creation and population of datasets, metadata, and datasheets primarily within its own confines. Subsequently, AI/ML algorithms typically ingest only the data and some metadata to generate models and tools, often stripping away valuable context.

The proposed model, illustrated in Figure 2, introduces the data-envelope as a pivotal innovation. Here, it acts as a central hub, harmonising access to comprehensive information and documentation for both CH users and the AI/ML community. This new paradigm aims to enrich AI/ML algorithms with a fuller context, enhancing the quality and applicability of the resulting models and tools.

The axis in both figures represents the amount of contextual information that the users are provided with when having access to the materials: when confronted with the trained model or a working tool they usually have way less context and explanation than when looking at the data and metadata itself. Under the current data interaction model, end-users engaging with AI/ML outputs encounter a notable deficit in context and explanation. In contrast, the data-envelope model facilitates direct access to extensive background information on datasets for more informed use.

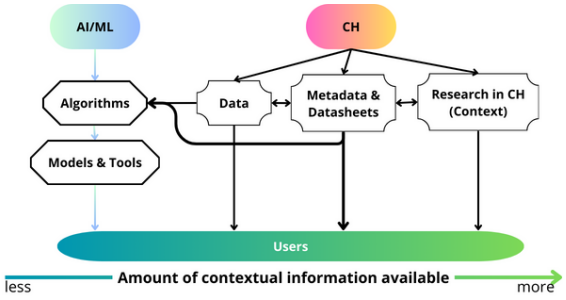

Figure 1: Current data interaction model with CH output (data, metadata, and research) in the context AI/ML development.

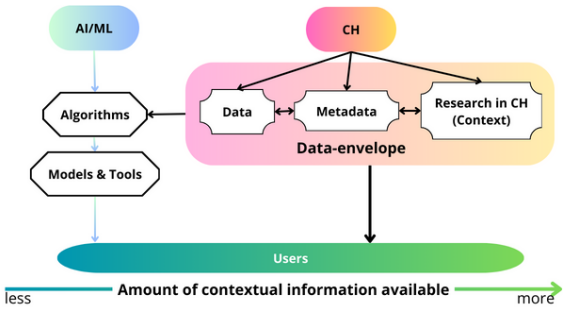

Figure 2: Proposed data interaction model with CH output (via data-envelope) in the context AI/ML development.

4.2. Modular Structure

The data-envelope is structured into modular sections, each designed to encapsulate different facets of the dataset in a systematic manner. The philosophy behind this five-level structure is to provide a comprehensive yet organised representation of the dataset. By separating the information into distinct levels, users can quickly locate the specific details they need without being overwhelmed by a monolithic documentation. The five-level structure, visualised in Figure 3, is elaborated on below, highlighting the basic ideas, philosophy, and differentiation from other templates. Further details about each level of the data-envelope are provided in the Section 8 (Appendix A), offering a more granular view of the specific contents and considerations within each section.

4.2.1. Basic Information/What Goes on the Data-Envelope

This section is dedicated to outlining the core details of both the data-envelope and the dataset it encompasses. It goes beyond traditional documentation practices by introducing a dual-versioning system: one for the dataset and another for the data-envelope itself. Recognising the dynamic nature of dataset documentation, this approach allows for the data-envelope to evolve independently of the dataset, adapting to the changing needs and standards of data management over time.

Additionally, this segment includes comprehensive contact information for individuals involved in various stages of the project. From conceptualisation and technical implementation to administration and more, users are provided with direct avenues to connect with experts for specific inquiries. This not only enhances the accessibility and transparency of the dataset but also fosters a collaborative environment where users can seek guidance, clarification, or further information as needed.

4.2.2. Basic Dataset Metadata

The Basic Dataset Metadata section conforms to the Data Catalog Vocabulary (DCAT) standards to guarantee compatibility with machine-readable formats (World Wide Web Consortium, 2014).[12] It catalogues key dataset information such as title, identifier, version, and a detailed description, along with the genre and topic classification. This section also outlines the dataset's geographical and temporal scope, essential for situating cultural heritage data within specific contexts.

[12]We refer to the most recent version: DCAT-3.

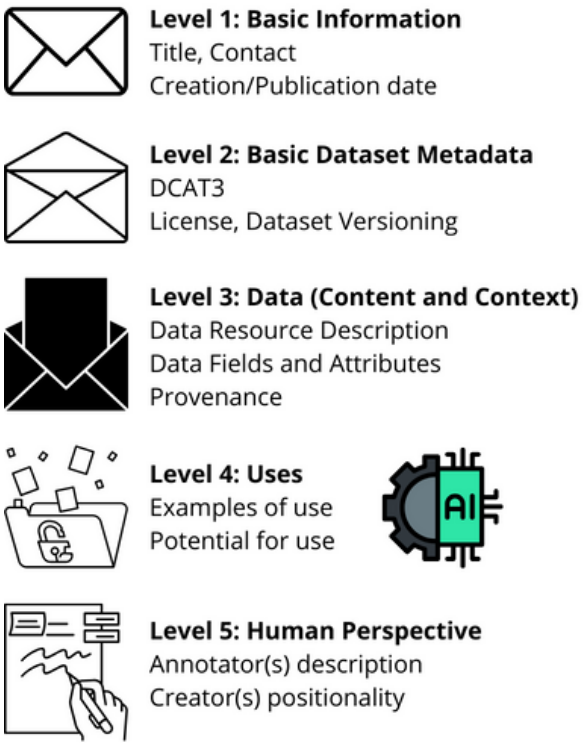

Figure 3: Data-envelope structure

Details about the dataset's inception and release provide insight into its relevance. Acknowledgement of contributors affirms transparency and credits those involved. Information on distribution, access, licensing, and maintenance is meticulously presented, equipping users with knowledge about usage conditions. Furthermore, a dedicated subsection ensures compliance with FAIR data principles, emphasising Findability, Accessibility, Interoperability, and Reusability (Wilkinson et al., 2016; Devaraju et al., 2021; Singh et al., in press). This commitment to FAIR principles ensures that the datasets are well-documented and suitable for broader use, aligning with global data management standards.

4.2.3. Data Content and Context

This section addresses the inclusion of diverse resources within the dataset, such as thesauri, reference data, and annotations. It is comprehensive, covering languages, encoding formats, resource creation dates, subjects of the data, modality, and descriptive statistics. It also describes data fields and attributes, presents sensitivity assessments, and provides examples to illustrate common errors and redundancies. Additionally, it details the annotation and labelling processes.

Furthermore, it has an extensive section on data provenance, connecting to additional documentation such as datasheets for sources and handwritten text recognition outputs, annotation instructions, and any other documentation, where available, providing users with supplementary information. Lastly, this section concludes with ethical reviews, social impact assessments, and bias considerations.

4.2.4. Uses

This section encourages dataset creators to introspect and articulate both recommended and discouraged uses of the dataset. It invites consideration of various application contexts, offering a platform for detailed descriptions and linkages to related datasets, publications, and models. This proactive reflection on the dataset's appropriate and inappropriate uses fosters responsible utilisation and helps users understand the boundaries within which the dataset is intended to operate.

4.2.5. Human Perspective

Positionality, rooted in Sandra Harding's (2003) standpoint theory, emphasises that personal backgrounds—encompassing gender, ethnicity, socioeconomic status, and more—influence an individual's knowledge and actions. This idea challenges the belief in objective, absolute truths within scientific research, instead suggesting that knowledge is created within a web of personal and social experiences (Haraway, 2016). Feminist epistemologists have thus argued that acknowledging and integrating positionality into the research can lead to more comprehensive and nuanced understandings (Mignolo and Walsh, 2018; Harding, 2013).

In dataset documentation, embracing positionality is vital for various reasons. Firstly, it illuminates the biases and assumptions that may influence data collection and analysis. Secondly, it provides transparency, allowing users to understand the context in which the dataset was created and to consider how this context may affect their use of the data. Thirdly, it promotes inclusivity by recognising the diverse standpoints of dataset creators and subjects, encouraging a multiplicity of perspectives in data interpretation. While positionality of annotators is becoming common practice (Geva et al., 2019), it is yet uncommon to see mention of positionality of the curators of datasets. The data-envelope will have a dedicated section on positionality of the institutions, projects, and persons involved in dataset creation.

An illustrative example is the work of Dutch linguist Jo Daan. In her seminal 1963 study at the Meertens Institute, Daan did not merely catalog dialects; she contextualised the data within the social dynamics of the speakers (Daan and Meertens, 1963). Her approach to documenting language patterns was inherently tied to the positionality of the communities she studied, pioneering a path in

linguistic research that considered the complex interplay of language with social identity and culture. This historical example underscores the depth and richness that positionality can bring to dataset documentation, and why it is increasingly becoming a best practice in the field.

4.3. Machine-Readable Implementation

The development of the data-envelope template is underway, aiming to transform it into a user-friendly, fillable form accessible on a static website. This innovative approach is designed to streamline the process of documenting datasets by allowing users to input detailed information directly into the form. Once completed, the form will enable the download of documentation in formats that are both human-readable and machine-readable.

Inspiration for this model comes from successful implementations such as CFFINIT[13], developed by the Netherlands e-Science Center, which facilitates the creation of citations for software and datasets. Similarly, Microsoft's introduction of the 'Open Datasheet' form, which outputs information in JSON format, exemplifies the potential of such tools in promoting standardised, machine-readable dataset documentation (Roman et al., 2023).

The ultimate goal is for these machine-readable documents to seamlessly integrate with Open Science repositories, like Zenodo[14], facilitating the automatic population of metadata fields. This integration would significantly advance the FAIRness of datasets, making them more discoverable and usable across the scientific community (Wilkinson et al., 2016). Although the practice of automatically integrating machine-readable datasheets into repositories is not yet commonplace, it embodies a progressive strategy to ensure that datasets are not only easily accessible but also thoroughly documented.

Balancing Metrics and Narratives in Cultural Heritage Datasets

The use of metrics and measures in cultural heritage datasets is a topic of ongoing debate. Cultural heritage institutions have a long history of qualitative item and collection descriptions, with minimal reliance on numbers. Historians and humanists are often skeptical of quantitative measures, recognizing their dependence on social context (Urton and Llanos, 1997). In contrast, the machine learning community places great value on descriptive statistics, digitization metrics, and annotation analysis (Alkemade et al., 2023). Resolving this divergence requires a case-by-case approach, selecting metrics based on their value and relevance to the dataset's intended purpose. Dialogue between domain experts, researchers, and tech-savvy individuals is crucial in determining appropriate metrics.

Moving forward, as the authors further develop the data-envelope template, they will consider incorporating controlled vocabularies for sensitive content categories and mitigation measures. This approach aims to facilitate the communication of crucial information in a standardized, machine-readable format while allowing for the inclusion of both quantitative and qualitative information as deemed appropriate for each specific dataset. By striking a balance between metrics and narratives, the data-envelope template seeks to promote transparency, accountability, and ethical considerations in the documentation of cultural heritage datasets.

5. Conclusion, Future Work and Challenges

This paper advocates for a paradigm shift in how we document, use, and understand cultural heritage datasets through the introduction of the data-envelope framework. By addressing the limitations of existing documentation practices and proposing a solution that caters to both technical requirements and broader societal needs, we invite the academic community and stakeholders in the cultural heritage sector to engage in a critical dialogue about the future of dataset documentation. Our work underscores the importance of a multidisciplinary approach to data governance, one that recognises the intricate web of legal, ethical, and practical considerations surrounding the stewardship of cultural heritage in the digital age.

While initially conceived to address the specific challenges of CH data, we argue that the data-envelope framework holds potential for broader applicability across diverse datasets. As many contemporary datasets are inherently socially and historically constructed, our documentation template serves as a valuable tool for enhancing transparency and understanding across various data domains.

The data-envelope template, as presented in the appendix, is a comprehensive framework designed to capture the intricacies of (Digital) Cultural Heritage datasets. As we continue to refine the template through collaborative iterations with diverse research groups within the Huygens Institute, we are actively engaged in a bottom-up approach to finalise the template to fit the needs of diverse projects, datasets, creators, and users. This iterative process involves gathering feedback, identifying common challenges, and adapting the template to ensure its flexibility and applicability across a

[13] https://citation-file-format.github.io/cff-initializer-javascript

[14] https://zenodo.org

wide range of cultural heritage contexts.

5.1. Ethical Considerations and Novelty

The ethical dimensions of this work are twofold. Firstly, the data-envelopes incorporate explicit statements about data bias and (re)use policies, addressing critical ethical concerns in the (re)use of historical datasets. Secondly, by harmonising the differing perspectives of data scientists and legal experts, proposed data-envelopes serve as a bridge between technical and legal frameworks, facilitating a more ethical and legally compliant use of historical data.

5.2. Technical Implementation and Embedding Data-Envelopes into the Data Life-Cycle

The scientific novelty of our approach lies in its emphasis on machine-readability, which not only enhances transparency and trust but also allows for the data-envelopes to be easily harvested and utilised as by the institutions internally, as well as by data marketplaces and repositories on the (inter)national level. We envisage that filling in and updating data-envelopes can become part of the standard research procedures, as they complement already established practice of creating data management plans.

5.3. Standardisation

To ensure the interoperability and widespread adoption of the data-envelope framework, we recognise the importance of aligning our template with existing standards and best practices in the cultural heritage sector. This includes considering the compatibility of the data-envelope with metadata standards such as Dublin Core (Weibel et al., 1998) and CIDOC-CRM (Doerr, 2005), as well as ensuring compliance with the FAIR principles (Findable, Accessible, Interoperable, and Reusable) for research data management (Wilkinson et al., 2016).

By actively engaging with the cultural heritage community and relevant standardisation bodies, we aim to develop a data-envelope template that aligns with existing standards while still addressing the unique challenges of cultural heritage datasets. This standardisation effort will not only facilitate the integration of the data-envelope into existing data management workflows but also promote its adoption across various cultural heritage institutions and projects.

5.4. User-friendliness and Collaborative Documentation

As we continue to engage with the research community and refine the data-envelope template, our primary goal is to achieve a balance between comprehensive documentation and practical implementation. To ensure the template's accessibility and ease of use, we will present the data-envelopes in the form of user-friendly, fillable forms accompanied by clear explanations for each section and field. These explanations will include illustrative examples and outline the purpose of each section, empowering dataset creators to provide accurate and relevant information.

Recognising the collaborative nature of dataset creation and documentation within the cultural heritage domain, we have designed the data-envelope template to facilitate teamwork and collective input. The template will allow multiple team members to contribute to the forms simultaneously, with features such as real-time collaboration, version control, and the ability to save progress as they work through the various sections. This collaborative approach not only streamlines the documentation process but also ensures that the final data-envelope benefits from the diverse expertise and perspectives of the entire research team.

To maximize the benefits of the data-envelope framework, we strongly advise implementing this documentation process at the outset of a research project. By conducting a thorough structural analysis of the dataset during the planning phase, researchers can effectively define their work plans, allocate resources, and identify potential data ethics issues early on. This proactive approach not only saves time and effort in the long run but also promotes a culture of responsible data stewardship from the very beginning of the research lifecycle.

6. Acknowledgements

This work resulted from a joined initiative in the context of two projects carried out at the Huygens Institute: Werk aan Uitvoering (WaU) and GLOBALISE, which is financed by the Dutch Research Council (NWO) in the research programme Research Infrastructure (project number 175.2019.003). We thank the GLOBALISE team, particularly the historical contextualisation team, for their valuable discussions during the initial development of data-envelopes. We also benefitted greatly from Leon's insights on machine readability, Kay's work on the previous datasheet version, and the inspiration drawn from existing datasheets. The Data Management department at Humanities Cluster KNAW, especially Lotte, Douwe, Harm, and Lodewijk, provided crucial feedback in refining the data-envelope.

7. Bibliographical References

Henk Alkemade, Steven Claeyssens, Giovanni Colavizza, Nuno Freire, Jörg Lehmann, Clemens Neudeker, Giulia Osti, Daniel van Strien, et al. 2023. Datasheets for digital cultural heritage datasets. *Journal of Open Humanities Data*, 9(17):1–11.

Paul Longley Arthur, Jason Ensor, Marijke van Faassen, Rik Hoekstra, and Nonja Peters. 2018. Migrating people, migrating data: Digital approaches to migrant heritage. *Journal of the Japanese Association for Digital Humanities*, 3(1):98–113.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Fiona Cameron and Sarah Kenderdine. 2007. *Theorizing digital cultural heritage: A critical discourse*.

Gustavo Candela, Nele Gabriëls, Sally Chambers, Milena Dobreva, Sarah Ames, Meghan Ferriter, Neil Fitzgerald, Victor Harbo, Katrine Hofmann, Olga Holownia, et al. 2023. A checklist to publish collections as data in GLAM institutions. *Global Knowledge, Memory and Communication*.

Nicole Contaxis, Jason Clark, Anthony Dellureficio, Sara Gonzales, Sara Mannheimer, Peter R. Oxley, Melissa A. Ratajeski, Alisa Surkis, Amy M. Yarnell, Michelle Yee, and Kristi Holmes. 2022. Ten simple rules for improving research data discovery. *PLOS Computational Biology*, 18(2):1–11.

Jo Daan and Pieter Jacobus Meertens. 1963. Toelichting bij de taalatlas van noord- en zuinederland. Technical report, Bijdragen en Medelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappente Amsterdam.

Anusuriya Devaraju, Mustapha Mokrane, Linas Cepinskas, Robert Huber, Patricia Herterich, Jerry de Vries, Vesa Akerman, Hervé L'Hours, Joy Davidson, and Michael Diepenbroek. 2021. From conceptualization to implementation: Fair assessment of research data objects. *Data Science Journal*, 20(1):1–14.

Martin Doerr. 2005. The cidoc crm, an ontological approach to schema heterogeneity. Schloss-Dagstuhl-Leibniz Zentrum für Informatik.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*.

Donna Haraway. 2016. 'situated knowledges: The science question in feminism and the privilege of partial perspective'. In *Space, gender, knowledge: Feminist readings*, pages 53–72. Routledge.

Sandra Harding. 2003. How standpoint methodology informs philosophy of social science. *The Blackwell guide to the philosophy of the social sciences*, pages 291–310.

Sandra Harding. 2013. Rethinking standpoint epistemology: What is "strong objectivity"? In *Feminist epistemologies*, pages 49–82. Routledge.

Natalie Harrower, Maciej Maryl, Timea Biro, Beat Immenhauser, and ALLEA Working Group E-Humanities. 2020. Sustainable and fair data sharing in the humanities: : Recommendations of the allea working group e-humanities. Technical report, ALLEA.

Amy K Heger, Liz B Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–29.

Mar Hicks. 2017. *Programmed inequality: How Britain discarded women technologists and lost its edge in computing*. MIT press.

Rik Hoekstra and Marijn Koolen. 2019. Data scopes for digital history research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 52(2):79–94.

Howard Hotson, editor. 2019. *Reassembling the Republic of Letters in the Digital Age*. Universitätsverlag Göttingen, Göttingen.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 306–316.

Kees Kooijmans and Johannes Petrus de Valk. 1985. “Eene dienende onderneming". De Rijkscommissie voor Vaderlandse geschiedenis en haar Bureau 1902-1968. pages 203–271.

Kees Kooijmans and C.E. Keij J.G. Smit Th.S. Bos, A.E. Kersten, editors. 1985. *Bron en publikatie. Voordrachten en opstellen over de ontsluiting van geschiedkundige bronnen*. Bureau der Rijkscommissie voor Vaderlandse Geschiedenis, Den Haag.

Marijn Koolen, Rik Hoekstra, Joris Oddens, Ronald Sluijter, Rutger Van Koert, Gijsjan Brouwer, and Hennie Brugman. 2023. The value of preexisting structures for digital access: Modelling the resolutions of the dutch states general. *J. Comput. Cult. Herit.*, 16(1).

Jörg Lehmann. 2024. Orientation in turbulent times.

Library of Congress. 2021. labs-ai-framework/Experiment/Data-Processing-Plan-template-2021-12-01-draft.docx at main · LibraryOfCongress/labs-ai-framework.

Geertje Mak, Marit Monteiro, and Elisabeth Wesseling. 2020. Child separation: (post-)colonial policies and practices in the netherlands and belgium. *Bijdragen en Mededelingen Betreffende de Geschiedenis der Nederlanden*, 135(3-4):4–28. Introduction to special issue.

Maciej Maryl, Marta Błaszczyńska, Ilaria Bonincontro, Beat Immenhauser, Szilvia Maróthy, Eveline Wandl-Vogt, and Joris J. van Zundert. 2023. Recognising digital scholarly outputs in the humanities. Technical report, ALLEA.

Barbara McGillivray, Beatrice Alex, Sarah Ames, Guyda Armstrong, David Beavan, Arianna Ciula, Giovanni Colavizza, James Cummings, David De Roure, Adam Farquhar, Simon Hengchen, Anouk Lang, James Loxley, Eirini Goudarouli, Federico Nanni, Andrea Nini, Julianne Nyhan, Nicola Osborne, Thierry Poibeau, Mia Ridge, Sonia Ranade, James Smithies, Melissa Terras, Andreas Vlachidis, and Pip Willcox. 2020. The challenges and prospects of the intersection of humanities and data science: A White Paper from The Alan Turing Institute.

Walter D Mignolo and Catherine E Walsh. 2018. *On decoloniality: Concepts, analytics, praxis*. Duke University Press.

Safiya Umoja Noble. 2018. Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press.

Kay Pepping, Henrike Vellinga, Manjusha Kuruppath, Leon Van Wissen, and Matthias Van Rossum. 2023. GLOBALISE Thesaurus - Commodities.

Lodewijk Petram and Matthias van Rossum. 2022. Transforming historical research practices–a digital infrastructure for the voc archives (globalise). *International journal of maritime history*, 34(3):494–502.

Timothy B Powell. 2016. Digital knowledge sharing: forging partnerships between scholars, archives, and indigenous communities. *Museum Anthropology Review*, 10(2):66–90.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826.

Anthony Cintron Roman, Jennifer Wortman Vaughan, Valerie See, Steph Ballard, Nicolas Schifano, Jehu Torres, Caleb Robinson, and Juan M. Lavista Ferres. 2023. Open datasheets: Machine-readable documentation for open datasets and responsible ai assessments.

Navroop K Singh, Shuai Wang, Angelica Maineri, and Tycho Hofstra. in press. Aligning data management plans with community standards using fair implementation profiles.

Ramesh Srinivasan, Katherine M Becvar, Robin Boast, and Jim Enote. 2010. Diverse knowledges and contact zones within the digital museum. *Science, technology, & human values*, 35(5):735–768.

Toma Tasovac, Sally Chambers, and Erzsébet Tóth-Czifra. 2020. Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper. DARIAH’s response to European Commission’s evaluation and possible revision of the Commission Recommendation of 27 October 2011 on Digitisation and Online Accessibility of Cultural Material and Digital Preservation (REC 2011/711/EU).

Nanna Bonde Thylstrup. 2019. *The politics of mass digitization*. MIT Press.

Jo Tollebeek. 1994. *De ijkmeesters : opstellen over de geschiedschrijving in Nederland en België*. Bakker, Amsterdam.

Molly Torsen and Jane Anderson. 2010. Intellectual property and the safeguarding of traditional cultures.

Michel-Rolph Trouillot. 2015. *Silencing the past: Power and the production of history*. Beacon Press.

Gary Urton and Primitivo Nina Llanos. 1997. *The social life of numbers: A Quechua ontology of numbers and philosophy of arithmetic*. University of Texas Press.

Stuart Weibel, John Kunze, Carl Lagoze, and Misha Wolf. 1998. Dublin core metadata for resource discovery. Technical report.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Stacy Wood, Kathy Carbone, Marika Cifor, Anne Gilliland, and Ricardo Punzalan. 2014. Mobilizing records: re-framing archival description to support human rights. *Archival Science*, 14:397–419.

World Wide Web Consortium. 2014. Data Catalog Vocabulary (DCAT).

8. Appendix A. Data-Envelope Structure

Currently, we envisage that the information in Levels 1 and 2 should constitute the default minimum requirements, while still allowing for flexibility in certain fields, such as the role of the project creators for legacy data. Level 3 should be completed to the best of the dataset creators' knowledge, providing essential context and provenance information. Levels 4 and 5 demand the most introspection from the dataset creators, challenging them to step back from the dataset and consider a variety of potential uses and issues. Consequently, these levels would have the least mandatory nature, but we strongly emphasise the importance of maximising their completeness to ensure responsible and ethical use of the datasets.

- **Level 1: Basic Information on Data-Envelope**
  - Title
  - Contact details for each relevant contact person (Name, ORCID, Role in Project, Email)
  - Data-envelope Creation Dates
  - Data-envelope Publication Date
- **Level 2: Basic Dataset Metadata**
  - Snapshot
    - Dataset Title
    - Version of dataset
    - Dataset URL
    - Description
    - Genre
    - Topic Classification
    - Geographical Coverage
    - Temporal Coverage
  - Dates
    - Dataset Creation Dates
    - Dataset Publication Date
  - Contributors
    - Publishing Organisation (Name, ROR ID, Organization Type)
    - Contributor (Name, ORCID, Organization Name, ROR ID, Role)
    - Funding Sources for each funding source: Institution(s) (Name of Institution, ROR ID, Funding or Grant Summary(ies), Relevant Links)
  - Distribution
    - Dataset Link: own dedicated website or if hosted on sites such as Zenodo, Dataverse
    - DOI
    - Repository
    - Download (URL, File Type(s) and Size)
    - Citation Information
  - Access/Licenses
    - Licensing Information for every license (Identifier, URL)
    - Access Level (Description, URL, Contact Information)
    - If Access level: restricted, then (Purpose of access controls, Highlight any restrictions or limitations, Access Prerequisites)
  - Dataset Version and Maintenance
    - Version Details (Current Version, Last Updated, Release Date)
    - Maintenance Status (Regularly Updated, Actively Maintained, Limited Maintenance, Deprecated)
    - Maintenance Plan (Versioning, Updates, Errors, Feedback)
    - Next Planned Update(s), if known (Version Affected, Next data update, Next Version)

- **Level 3: Data (Content and Context)**
  - Data Resource Description
    - Name of Resource
    - Description
    - Path, Format, Size, Date
    - Data Subject(s) (Sensitive data about people, Non-sensitive data about people, Data about natural phenomena, Data about places and objects, Synthetically generated data, Data about systems or products and their behaviour)
    - Language(s)
    - Encoding
    - Data Modality (Image Data, Text Data, Tabular Data, Audio Data, Video Data, Time Series, Graph Data, Geospatial Data, Multimodal)
    - Descriptive Statistics (Size of Dataset, Number of Fields, Labelled Classes, Number of Labels, Average labels per instance, Algorithmic labels, Human Labels)
  - Data Fields and Attributes
    - Data Fields Summary
    - Use of Linked Open Data, Controlled Vocabulary, Multilingual Ontologies/Taxonomies
    - Description of every data field in the resource (Data Field Name, Data Field Type, Description of the Field, Sensitivity, Notable Feature(s), Attributes)
    - Data Point Example
    - Atypical Datapoint
    - Any errors, sources of noise, or redundancies in this resource
    - Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, other datasets)
  - Annotation & Labeling
    - Annotation Workforce Type (machine vs human (from experts to non-experts, crowdsourcing, etc)
    - Annotation Characteristics (Description, Number of unique annotations, Total number of annotations, Average number of tokens/annotation, Total tokens annotated, Inter Annotator Agreement (or other relevant metric)
  - Social Impact, Sensitivity, and Biases
    - Does the resource contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
    - Does the resource contain data that might be considered confidential?
    - Known Biases in the resource
    - Sensitive Human Attributes
    - Unintentionally Collected Attributes
    - Any ethical review processes conducted (e.g., by an institutional review board)?
  - Data Provenance for each source used
    - Name
    - Path
    - Description
    - Creators of Source (name, affiliation, organization and contact if available)
    - Year of publication
    - Language
    - Temporal Scope
    - Geographical Scope
    - Notable Features
    - Datasheet/data-envelope
    - Data Selection Criteria:
  - Digitisation Pipeline

- **Level 4: Uses:** purpose of potential use, domain(s), motivating factors and problem space(s)
  - Uses
    - Dataset Use(s): safe for production use or for research use; conditional use-some unsafe applications; only approved use
    - Links to Related Datasets, Publications, and Models
    - Suitable Use Case(s)
    - Unsuitable Use Case(s)
    - Is there a repository that links to any or all papers or systems that use the dataset?
  - Use with other data
    - Safety Level: safe to use with other data, conditionally safe to user with other data, should not be used with other data
    - Known safe dataset(s) or data type(s)
    - Best Practices
    - Known unsafe dataset(s) or Data Type(s)
    - Limitation(s) and Recommendation(s)
  - Use in ML or AI Systems
    - Dataset Use(s): training, testing, validation, development or production use, fine tuning
    - Notable Feature(s)
    - Known Correlation(s)
    - Data splits
  - Sampling
    - Safety Level: safe to sample, conditionally safe to sample, should not be sampled
    - Acceptable Sampling Method(s)
    - Best Practice(s)
    - Risk(s) and Mitigation(s)
- **Level 5: Human Perspective**
  - Annotator Description(s) per each annotation type
    - Task type, e.g. survey, video annotation, text annotation, image annotation
    - Number of unique annotators
    - Expertise of Annotators
    - Description of annotators
    - Compensation
    - Language distribution of annotators
    - Age distribution of annotators
    - Geographic distribution of annotators
    - Gender distribution of annotators
    - Socio-economic distribution of annotators
    - Summary of annotation instructions
    - Summary of gold questions
    - Annotation platforms
  - Creators