Datasheets for Digital Cultural Heritage Datasets

HENK ALKEMADE (iD)

STEVEN CLAEYSSENS (iD)

GIOVANNI COLAVIZZA (iD)

NUNO FREIRE (iD)

JÖRG LEHMANN (iD)

CLEMENS NEUDECKER (iD)

GIULIA OSTI (iD)

DANIEL VAN STRIEN (iD)

*Author affiliations can be found in the back matter of this article

ABSTRACT

Sparked by issues of quality and lack of proper documentation for datasets, the machine learning community has begun developing standardised processes for establishing datasheets for machine learning datasets, with the intent to provide context and information on provenance, purposes, composition, the collection process, recommended uses or societal biases reflected in training datasets. This approach fits well with practices and procedures established in GLAM institutions, such as establishing collections' descriptions. However, digital cultural heritage datasets are marked by specific characteristics. They are often the product of multiple layers of selection; they may have been created for different purposes than establishing a statistical sample according to a specific research question; they change over time and are heterogeneous. Punctuated by a series of recommendations to create datasheets for digital cultural heritage, the paper addresses the scope and characteristics of digital cultural heritage datasets; possible metrics and measures; lessons from concepts similar to datasheets and/or established workflows in the cultural heritage sector. This paper includes a proposal for a datasheet template that has been adapted for use in cultural heritage institutions, and which proposes to incorporate information on the motivation and selection criteria, digitisation pipeline, data provenance, the use of linked open data, and version information.

(1) INTRODUCTION

The massive uptake of machine learning (ML) in both commercial and academic settings has led to increased interest in using ML with Digital Cultural Heritage (DCH) data. Fueled by this growth of various computational approaches to working with large amounts of data, the Collections as Data movement has gained considerable momentum. The Collections as Data movement encourages "computational use of digitised and born digital" cultural heritage collections by making these collections available as data, "amenable to computation" (Padilla et al., 2022, p. 20). Central to the first Collections as Data statement is the case for shared documentation, attesting "to the history of how the collection has been treated over time" (Padilla et al., 2022, p. 21). In a recent effort to create a checklist for publishing Collections as Data in GLAM (Galleries, Libraries, Archives, and Museums) institutions, this call for documentation was echoed, stating that "documentation is a key element to foster the re-use by the community" (Candela et al., 2023, p. 8).

While there was already significant demand for DCH data among scholars, nowadays also Big Tech engages in a fair amount of DCH data processing. It can only be applauded that in this way, the use of DCH transcends the domains of humanities and cultural heritage. However, it also means that the data is used in contexts that are barely familiar with the specifics and intricacies of DCH, making the plea for documentation even more urgent. DCH collections without proper information providing the required context needed for (re-)use, are prone to exploitation and misuse.

This paper elaborates on the use of datasheets, as introduced by (Gebru et al., 2021) to the ML community for the first time in 2018, for creating and disseminating documentation about DCH materials shared as "collections as data." The two main concepts here, DCH and datasheets need some clarification. In this paper, we comply with the UNESCO definition of digital cultural heritage: "Digital materials include texts, databases, still and moving images, audio, graphics, software, and web pages, among a wide and growing range of formats. They are frequently ephemeral and require purposeful production, maintenance, and management to be retained" (UNESCO, 2003, p. 1 of Annex I). DCH collections are composed of discrete digital heritage objects with associated metadata. They are extremely diverse by nature, biased by definition and hardly ever created or collected with computation in mind. This necessitates careful consideration when applying ML or other computational methods to DCH.

Then, why datasheets? Dataset documentation can take on a myriad of shapes and forms, ranging from highly structured data, for both humans and machines to read (for example, metadata description in the Data Catalog Vocabulary[1] (DCAT)), over semi-structured datasheets, organised around a standard list of questions, to unstructured, primarily narrative data papers. They all accompany the publication of a dataset to provide a basis for informed decisions about using the data. They may serve as a valuable starting point for source criticism, a crucial initial building block for reproducibility, and, in general, bring in much-needed transparency by creating a space for facilitating intersectoral communication. Datasheets, however, bring a structured approach to the description of datasets, which provide guidance to the data publisher in describing the datasets according to the information needs of data re-users, and they offer the advantage of allowing information to be collected in both a structured manner, whenever possible, and in a narrative form, whenever necessary. Considering the particularly diverse nature of DCH collections, that combination is invaluable.

CONTEXT: RELATED WORK

The adoption of datasheets by cultural heritage institutions is growing. A notable example are the datasheets prepared during the BigLAM[2] workshop, organised by BigScience, an open scientific collaboration of nearly 600 researchers from 50 countries and 250 institutions who collaborated on various projects within the Natural Language Processing (NLP) domain. BigLAM focused on making data from Libraries, Archives, and Museums potentially suitable for machine-learning applications more discoverable by publishing datasets and datasheets at Hugging Face Hub, an online platform where people can easily share and collaborate on ML assets, e.g., the Contentious Contexts Corpus (Brate et al., 2021). Another example available on the Hugging Face hub is the datasheet for the DEArt dataset (Reshetnikov et al., 2022), an object detection and pose classification dataset

---

1   https://www.w3.org/TR/vocab-dcat-2/.

2   https://github.com/bigscience-workshop/lam.

containing 15.000 annotated images of paintings from between the XIIth and the XVIIIth centuries. Other examples of datasheets from cultural heritage are the one for the dataset Unsilencing Colonial Archives via Automated Entity Recognition (Luthra et al., 2022a, 2022b), the 19th-century books dataset (British Library et al., 2021), and the De Boer Press Photography Datasheet (Wevers, 2022). Additional examples may be found in Fiorucci et al. (2020).

METHODOLOGICAL STEPS

To further deepen the concept and establish datasheets as a good practice for DCH, in 2022 a working group[3] was established by the Europeana Research and the EuropeanaTech Communities, joined with experts in the fields of DCH, data analysis and management, digital scholarship and ML. Drawing from their professional expertise and relevant literature (Gebru et al., 2021; Jo & Gebru, 2020; Lee, 2023; Pushkarna et al., 2022), the group members set to work with the primary goal of identifying the specific characteristics and needs of DCH, analysing and evaluating existing concepts of datasheets and datacards, and adapting as well as complementing relevant fields of datasheet templates. We discussed the example datasheets mentioned above and organised multiple (online) thematic sessions with regard to the scope and characteristics of digital cultural heritage datasets, possible metrics and measures, and lessons from similar concepts and/or established workflows in GLAMs.

Based on these discussions, a first attempt was made to create a datasheet template for DCH (Alkemade et al., 2023). Given the rapid development of the field, we explicitly understand this template as a proposal, a first working version for gaining experience and to collect feedback. We intend to continue working on it. In addition, it is important to stress that the template should be thought of as modular: it is up to those filling in the form to decide which questions should be answered and which questions could be ignored.

(2) DATASET DESCRIPTION

Object name: Template Datasheet for Digital Cultural Heritage Datasets

Format names and versions: PDF, Version 1

Creation dates: 2023-01-25 to 2023-09-25

Dataset creators: Henk Alkemade, Steven Clayessens, Giovanni Colavizza, Nuno Freire, Alba Irollo, Jörg Lehmann, Clemens Neudecker, Giulia Osti, Daniel van Strien

Language: English

License: Creative Commons Attribution 4.0 International

Repository name: Zenodo

Publication date: 2023-09-25

Repository location: https://doi.org/10.5281/zenodo.8375033

(3) RECOMMENDATIONS FOR CREATING RELEVANT AND RESPONSIBLE DCH DATASHEETS

SCOPE AND CHARACTERISTICS OF DIGITAL CULTURAL HERITAGE DATASETS

DCH datasets and industrial or research datasets share the complexities of data collection, digitisation and lack of knowledge over the data subjects contained therein. However, GLAM institutions have developed the language and procedures to document these issues and see it as part of their professional ethics to communicate them. DCH datasets are often built on top of existing collections, and may feature multiple layers of selection. Metaphorically, this can be imagined as a "ziggurat" structure, thus referring to the multiple layers and levels of terraced buildings in ancient Mesopotamia. The small fraction of our cultural heritage that is preserved in memory institutions could be imagined as the base of this ziggurat. Expertise, formal education, institutional frameworks and procedures established in cultural heritage

---

[3]   https://pro.europeana.eu/project/datasheets-for-digital-cultural-heritage-working-group.

institutions support these selection processes on an item-level basis. Not infrequently, GLAMs acquired pre-existing collections, eventually resulting from selection processes performed by people hundreds of years ago.

Digitisation presents another layer of selection, which does not necessarily correspond 1:1 to the collections held in a cultural heritage institution. Rather, the digitisation of a collection may be motivated by preservation purposes, to reproduce unique or rare items in digital form, or for a specific research project (Corrado & Moulaison Sandy, 2017). Therefore, provenance, source criticism and selection tool criticism are more important than in industrial datasets. This is supported by the findings of Holstein et al. (2019). Through interviews and surveys with ML practitioners in commercial product teams, they found that there is a general lack of guidance or requirements for conducting data collection, and the persistence of a significant communication gap between data collectors/curators and ML practitioners. In contrast to contemporary datasets, cultural heritage collections and their digital counterparts profit from the refined methodologies used in cultural heritage institutions for data collection and the expertise of cultural heritage practitioners, thus implementing foundational approaches such as consent, power, inclusivity, transparency, and ethics and privacy (Jo & Gebru, 2020).

Another feature of cultural heritage datasets is that many of them grow over time and are thus seen as dynamic, mutable objects (Conway, 2015). Digitised newspapers are a good example of this: initially, the impetus to digitise newspapers came from the intent to preserve deteriorating material, and therefore, such objects were amongst the first to be digitised in libraries (Beals & Bell, 2020; Beelen et al., 2023). In the meantime, this approach has been complemented by systematic digitisation policies, and the corpus of digitised newspapers is continually growing. This might produce an issue of its own with regard to digital datasets, because growth needs documentation, and introduces the need for versioning of the various dataset evolutions.

Digital datasets produced by cultural heritage institutions are marked by their *heterogeneity*– they are characterised by high dispersion in time, place, languages, social stratum of the people depicted or who contributed to them, and they come from a broad range of differing cultural contexts, thus making domain-specific knowledge necessary for a better understanding of what is incorporated in them. In contrast to industrial or research datasets that are assembled to create knowledge (often with the claim to create "objective science"), cultural heritage datasets may present knowledge as it was fabricated in earlier times, or community-based knowledge from lost local contexts. Cultural heritage datasets, therefore, often or even predominantly contain symbolic, historical, and aesthetic content, and the high degree of abstraction in the content they present is notable. Multilinguality is not only a characteristic of European datasets, but also of cultural heritage datasets in general. While some of the features described above are not exclusive to cultural heritage data, the complexity of cultural dimensions in such datasets is often underestimated.

Corresponding to the procedures and workflows established in cultural heritage institutions, rights, licences and other obligations derived from provenance are respected. In the tradition of classification systems developed in libraries, cultural heritage institutions use controlled vocabularies, ontologies and taxonomies and linked open data, an approach which is not commonly used within contemporary datasets. The practice of non-intervention into the dataset stands in the same tradition. DCH datasets are understood as a form of documentation into which should not be intervened. Seen from the standpoint of cultural heritage practitioners as well as historians and other humanists working with archival material, historical records must be left as they are, while an intervention into the dataset with the intent to "balance" or to "mitigate" a particular skewness is seen as inappropriate. Such specifics strongly require a contextualisation of each given dataset, and the provision of context is the task of the datasheet, or data card, as it may alternatively be called (Pushkarna et al., 2022).

Furthermore, it is imperative to maintain an ever-present awareness and adopt a critical perspective at every point of data handling. A key resource to guide libraries and other memory institutions in their approach to data science and ML is the OCLC report "Responsible Operations" (Padilla, 2019). The cornerstone of this report is the idea of "responsible operations," a concept originally introduced by Rumman Chowdhury, one of the leading women in the field of ML systems (O'Neil, 2023). As per her definition, responsible operations encompass the development of practices aimed at mitigating algorithmic bias arising from human input in computer programs,

and ensuring a comprehensive understanding of the data by all individuals involved in the process (Apte, 2017). While a variety of desirable attributes are associated with the domain of responsible operations, including fairness, transparency and privacy, to guide the implementation of any ML system or regulating the use of training data (Barredo Arrieta et al., 2020; Rakova et al., 2021), an unanimous consensus on this concept remains elusive. Nevertheless, the importance of examining different facets as exemplified by the critical examination of potential bias amplification in section 3.2, underscores the need for strong ethical commitments in the pursuit of responsible production, circulation, use and re-use of DCH datasets. We believe that these commitments could be furthered to some extent by carefully documenting as many relevant elements as possible from the datasheet template discussed here.

The intended *uses* to be made of digital cultural heritage datasets may differ from contemporary datasets. The purpose for which DCH datasets have been created was not necessarily the use in ML; mind the preservation example. In such cases, historical datasets may be very different from what the ML community needs, which recalls the gap between the creators of a particular dataset (cultural heritage practitioners and historians) and possible users (the ML community). It is not always straightforward to anticipate the use made of a dataset, because the use might be completely different from the "intention" with which the collection was compiled–or there never was an intended audience for the collection out of which the dataset was compiled. However, emphasis on the purpose of a cultural heritage dataset is rather on public accessibility and on research than on a deployment for profit. Another feature distinguishing the use of cultural heritage datasets from the use of industrial ones is that re-uses of the same dataset are common in the cultural heritage field. This may initially have resulted from the practices of historians who read and re-read the same sources over and over again according to new research questions, thus providing new interpretations of the same historical material and adding to the knowledge based on those same sources. However, processes such as optical character recognition (OCR), named entity recognition, disambiguation and linking, which are counted among standard tasks within libraries, also use the same datasets. Old datasets are often processed with lately developed models. Therefore, much research and development will be retroactive within the cultural heritage field. **Since the intended (and especially the unintended) use of a dataset may result in issues which are particular to a dataset, we recommend flagging them in the accompanying datasheet, indicating that historical data is mostly not suitable for modern-day applications.**

Datasheets in the cultural heritage sector often build on existing collection descriptions established with diligence and etiquette. However, this carefulness and the intent to provide saturated context to a collection brings forth a certain resistance to standardisation, because standardised description schemes such as the Encoded Archival Description (Library of Congress, 2019) or the standardised process offered by datasheet templates may be experienced as straitjackets into which complex circumstances have to be pressed. In this sense, "Tools and standards are *pharmaka*, giving much but taking as well" (Edmond & Lehmann, 2021, p. 100). However, datasheets are the place where the context of a particular dataset should be inserted, thus situating the dataset in its proper time, place and cultural context as well as inscribing the positionality of the datasheet creators and contemporary views of this dataset.

We recommend establishing datasheets in an *interdisciplinary* manner in dialogue between domain and technical experts. The collaborative filling of datasheets helps to become aware of the particularities of the content and to reflect on possible issues from different perspectives. Often cultural heritage institution practitioners understand the specifics of a dataset better when researchers pose questions aimed at digital collections. They should therefore be asked to contribute to datasheets, especially in those cases where digital collections are established for research purposes. A closer collaboration between researchers and cultural heritage institutions is desirable, and could also help to integrate those cultural heritage practitioners into the dialogue when deciding what is being digitised. Their point of view is of particular interest wherever the possible use of such datasets is reflected. *Interdisciplinarity* is not only a noble (and expensive) goal, but also a means of creating transparency. As each discipline tends to create its own opaque technical language, which is "dense, and presumptive of a reader's background, making it difficult for non-technical stakeholders to interpret [...] transparency is attained when we establish a shared and socratic understanding of datasets" (Pushkarna et al., 2022, p. 1779).

POSSIBLE METRICS AND MEASURES

Metrics and measures are topics discussed controversially regarding cultural heritage datasets. Cultural heritage institutions' practices have a long pedigree which tightly connects them with qualitative item and collection descriptions containing only a minimum of numbers, such as the number of works being part of a multi-volume publication or the number of items of a specific class (e.g., photos) in a collection. Historians and humanists, in general, are sceptical vis-à-vis numbers because they are well aware of the length of the historical process necessary to establish faith and "Trust in Numbers" (Porter, 1996). Even though numbers, metrics and measures carry qualities like transferability, standardisation, interconvertibility, and generalizability, most people working in cultural heritage institutions and humanists do not view numbers as "objective," and view their use strongly as dependent on their social context (Urton, 1997). However, the ML community has a different background, where numbers, metrics and measures are paramount. This community, in general, has a strong interest in descriptive statistics of digital datasets, derived from data analysis, in data on how the content of such datasets has been influenced by digitisation (including metrics), and in statistics derived from the analysis of the annotations provided, if the latter are part of the dataset.

These diverging views on metrics and measures need not be seen as an irreconcilable conflict of interest. **We recommend choosing metrics based on whether they offer value, and whether the numbers and metrics provided can be of use depending on the intended purpose of a dataset.** However, no minimal viable set of metrics can be recommended for inclusion for each datasheet. Because of the heterogeneity of datasets produced in the cultural heritage sector, the decision on which numbers, metrics and measures to include has to be taken on a case-by-case basis, taking the possible use of such a dataset by the ML community into perspective. **We recommend establishing datasheets in a dialogue across different domains and letting domain experts, researchers and tech-savvy people collectively discover which metrics tell something valuable in each particular case and are, therefore, appropriate.**

All datasets are biased in one way or another. The description of biases can take several forms. Social biases can be well described narratively and may translate well into statistics, for example, about the sex or ethnicity of the people depicted in the dataset. By contrast, ethical biases must be understood as different from a statistical understanding of bias as *skewness*, and it is often inappropriate to describe ethical biases in numbers or metrics. Historians and other researchers working in cultural heritage institutions know historical biases well. From their point of view, it is a banality that throughout history most creators of written records were men, because they were literate and socially privileged. This argument can be turned around: the most challenging bias is not overrepresentation, but underrepresentation, which has also been described as the "Archive Gap" (Singh, 2019). Mitigating this bias by uncovering silenced and therefore underrepresented minorities has become a task in itself (Luthra et al., 2022a). While providing global information on social biases may contribute to a better understanding of a dataset, it may not directly be relevant to the use of the dataset as training data. This is because the relevant bias to be described is linked to the anticipated use made of the dataset, and the decision on what is regarded as important to include lies with the curators of the datasheet. From a contemporary point of view, including social and ethical considerations, many cultural heritage datasets are problematic with respect to their wording and their content. For these reasons, a few more questions regarding such biases have been added to the datasheet template (Alkemade et al., 2023).

Sensitive data are better not expressed in metrics but should be explained narratively. For example, a narrative description of sensitive content serves the purpose of datasheets well, even if it may be regarded as subjective, rather than e.g., quantitative information on how many contentious words are present per page. In this way, we contemporaries inscribe our positionality into the context of datasets provided by the datasheet. An example of cultural heritage datasets that might contain sensitive data that are still harmful today is a collection of sources from former colonies. Harmful effects for the "colonial subjects" are still conceivable, and cultural heritage institutions with colonial collections may also be interested in providing information on gender or ethnicity for this collection. However, this issue may probably not be regarded as pressing as with contemporary data collections, and if time and monetary resources allow for it, descendants of "colonial subjects" should be included in the process of establishing datasheets

(Kirk et al., 2022). Similar to sensitive data, there is the potential to describe toxic contents with metrics. Potentially, because everything that has an effect can be measured (Hubbard, 2010). However, it is not advisable to use metrics and measures which are not (yet) established or agreed upon by relevant communities. Some measures, like the carbon footprint, come with a long pedigree for their development–the discussion around this measure started in the 1990s.

With regard to DCH datasets, claims for representativeness are usually problematic. This is the result of what has been described above as the "ziggurat" structure of such datasets: in most cases, we do not know about the "whole" of historical records that once existed and out of which the collections we are now dealing with have been taken out. Moreover, what has been digitised more often than not represents only a fraction of a collection as a whole. This observation applies especially to digitised archival records.

If a cultural heritage dataset is specifically prepared for ML tasks, we recommend providing–beyond basic descriptive statistics–a description of how the content has been influenced by digitisation, e.g., what part of a collection has been digitised and what not, information about the main features and on further processing of the dataset (like, e.g. OCR) including state-of-the-art metrics, as well as disclaimers or warnings about potential risks and hazards that may result from the training of a model on this dataset. Moreover, it is important to provide information on whether the dataset was annotated, about the quality of the annotation, some related statistics (like, e.g. the inter-annotator agreement), provide links to annotation guidelines or instructions given to the annotators and documentation of the annotation process if available. Finally, if a dataset has been annotated, it is likely that models have already been trained on it; a reference to these models and related publications, such as research papers and/or blog posts, are valuable.

For the ML community, benchmark datasets play an important role in the organisation of research. Benchmark datasets are used to coordinate research activities around shared research problems. They are a means to measure progress with respect to particular tasks, often in "yearly challenges where researchers compete to develop the best performing model" (Scheuerman et al., 2021, p. 317:2). In the cultural heritage sector, one may think of text recognition, named entity recognition, named entity disambiguation and linking, image classification, or other specific ML tasks. Again, benchmark datasets depend on the modality of data (e.g., image versus text) and on the tasks for which they were created. The composition of benchmark datasets specifically for such tasks is desirable since it supports the re-use of datasets and the advance of ML research. Even more complicated benchmark datasets for the cultural heritage sector are imaginable, such as a multilingual dataset on time period recognition or diachronic datasets. Moreover, benchmark datasets offer the possibility to compare the use made of datasets relevant to the cultural heritage field and to prepare case studies on this basis.

Finally, it is advisable to consider the FAIR data principles (Wilkinson et al., 2016) while establishing the datasheet. Findability, Accessibility, Interoperability, and Reusability (FAIR) are principles that support the composition and publication of datasets as well as the accompanying metadata. To calculate a score for the completion of each FAIR principle per dataset, several metrics and tests have been developed and applied (Devaraju, Anusuriya et al., 2020; Van Erp et al., 2018; Wilkinson et al., 2018) which may support and orient the creators of a datasheet.

LESSONS FROM SIMILAR CONCEPTS AND/OR ESTABLISHED WORKFLOWS IN GLAMS

Museums present digital representations of their holdings online. Archives have often digitised at least their finding aids and provide archival descriptions in digital form. Libraries provide access to their digitised collections. Sooner or later, datasets and ML models trained on them will be integrated as new assets into existing catalogues, finding aids, and databases. This prospect is a reminder that establishing datasheets should become a standard feature of the curatorial procedures and workflows for DCH already in place in cultural heritage institutions. This applies especially to the activities related to digitisation, starting with the selection process, adding metadata during digitisation, and further processing steps undertaken before assembling a dataset. **The creation of a datasheet should, at best, start alongside the creation of the dataset, and we recommend documenting explicitly what has been done in each antecedent step of the digitisation pipeline.**

Digitisation processes involve heads of digitisation divisions and/or project managers. Alongside those cultural heritage practitioners who already know the collections by heart, such digitisation professionals should be placed at the beginning of an interdisciplinary dialogue between domain experts and sociotechnical personnel, thus integrating the knowledge available in any case in the cultural heritage institution. Heads of digitisation divisions and project managers are also aware of the data management plans (DMPs) usually prescribed by donors or funders of digitisation projects. Datasheets fit well alongside data management plans, and it is expectable that they will soon become a standard item within them, even if both are meant for different purposes and audiences.

As a matter of principle, datasheets should feed back into existing catalogues, finding aids, databases, and data publications within cultural heritage institutions. **We recommend providing them in a machine-readable format, such as xml, rdf, comma-separated values (csv), and linking them with controlled vocabularies to enable ingestion into existing systems.** This opens the possibility for (meta-)search engines or dataset hubs to automatically ingest datasheets, perform search and filter operations over hundreds or thousands of datasheets and provide links between datasheets accompanying the same or related datasets. This might be the case if a dataset is being published in several places and with modified datasheets accompanying it, or, e.g. with 3D scans or models of objects, where it is possible to have parent models and "derivative" children models.

Dealing with multiple versions of a dataset will certainly be a challenge for traditional cultural heritage reference systems. With growing, changing or sub-set datasets, the idea of static and durable datasheets becomes obsolete. Rather, they should be conceived of as flexible and dynamic objects which change according to the version of a dataset documented by them, and/or with changing uses made of a particular dataset. **We recommend providing opportunities to respond to and comment on the datasheet, to indicate a reliable contact to the datasheet curators and information on maintenance or updating of the dataset, and to incorporate responses and comments into a new version of the datasheet where sensible.** From this requirement follows that there have to either exist many versions of a datasheet or that the reference systems will have to offer the possibility to record different versions of the same datasheet, like a version control system does.

Regarding existing datasheets and datasets, the demand for documentation of multiple versions prompts questions regarding whether it is possible to retrospectively keep record of existing versions, whether these versions should all be documented in a single datasheet, or whether this will lead to multiple datasheets, where each pertains to a different version of the dataset. The ML community is interested in comprehensible documentation of the relations between datasets and models in order to enable reproducibility (Kapoor & Narayanan, 2022). This is a current issue and an emergent practice that aims at validating previous results and making the progress of research visible (Padilla et al., 2022, p. 120). Such questions should be discussed for each case individually and decided by the cultural heritage institution that provided the dataset according to archival or curatorial principles.

Finally, this discussion points to issues surrounding the re-use of existing datasets. Processing datasets assembled many years ago with models recently developed by researchers usually lead to better results, e.g. with respect to OCR or named-entity recognition. However, updating existing data with better results is not the task of research projects. Rather, the responsibility for ingesting improved data lies–at least in part–with the cultural heritage institution hosting the initial dataset. Such institutions should adapt existing workflows in order to incorporate the updated datasets.

(4) CONCLUSION

Given the availability of massive amounts of digitised material on the side of cultural heritage institutions and the strong interest in the ML community in receiving large datasets, the benefits to standardise the process of establishing datasheets (as well as model cards) for cultural heritage datasets are obvious for supporting the responsible re-use of datasets. The datasheet template (Alkemade et al., 2023) pushes this intention forward, providing both a structure and explanations on what kind of information should be provided where, as well as recommendations and hints supporting cultural heritage practitioners in the process of documentation. Given the broad range and heterogeneity of cultural heritage datasets, it is

certainly a challenge to provide information on curated data from cultural heritage institutions in a consistent and standardised way. However, the datasheet template provided here should be thought of as modular, i.e. with some sections fitting for the dataset to be described while others have to be discarded (or be appropriate in other cases).

Cultural heritage institutions certainly need to be motivated to adopt the practice of furnishing a comprehensive datasheet alongside the datasets prepared by them. Meanwhile, providing descriptions of their holdings in a standardised way fits very well with current workflows and procedures established in libraries, archives and museums if it becomes clear to cultural heritage practitioners that datasheets provide essential information on the provenance of datasets, an activity that is part of GLAM institution's daily business. Even though the development of ML applications is currently rather the exception than the norm in cultural heritage institutions, it is clear that the provision of datasets and accompanying datasheets does not only serve the demand of the ML domain. It also empowers practitioners to play an active role in circulating digital content, strongly contributes to the profile of the providing institutions, serves as evidence of their capacity for conducting research projects in the field of ML and artificial intelligence, and thus may indirectly contribute to receiving funds.

To stimulate the adoption of datasheets by cultural heritage institutions and researchers, we aim to create a web-based tool for the establishment of datasheets (and model cards) for cultural heritage datasets to be used by a broad range of cultural heritage institutions, thus improving transparency, reusability and reproducibility and fostering the adoption of the standard provided by the datasheet template. Such a tool, which can be imagined as a structured form with some sections to be filled mandatorily and others optionally, may be used to generate machine-readable datasheets and initial metadata, keep track of different versions of datasets and datasheets and thus to continually update growing datasets, and provide flexibility in incorporating feedback. It should be enhanced by a tool for previewing of the dataset, and, if applicable, to extract information on the annotations, like e.g. inter-annotator agreement or changes in the annotations. Such a tool will certainly support the adoption of datasheets in cultural heritage institutions and their daily workflows.

ACKNOWLEDGEMENTS

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Conceptualization – all authors; Writing – original draft: Henk Alkemade, Steven Clayssens, Jörg Lehmann; Writing – review and editing – all authors.

AUTHOR AFFILIATIONS

Henk Alkemade orcid.org/0000-0002-4413-5463
EuropeanaTech Community, Europeana Network Association, The Hague, Netherlands; CARARE, Dublin, Ireland

Steven Claeyssens orcid.org/0000-0003-1110-5935
KB, National Library of the Netherlands, The Hague, Netherlands; Europeana Research Community, Europeana Network Association, The Hague, Netherlands

Giovanni Colavizza ⓘ orcid.org/0000-0002-9806-084X
Department of Classical and Italian Philology, University of Bologna, Italy

Nuno Freire ⓘ orcid.org/0000-0002-3632-8046
School of Social Sciences and Humanities, NOVA University of Lisbon, Lisbon, Portugal; Europeana
Foundation, The Hague, Netherlands

Jörg Lehmann ⓘ orcid.org/0000-0003-1334-9693
Staatsbibliothek zu Berlin – Berlin State Library, Berlin, Germany

Clemens Neudecker ⓘ orcid.org/0000-0001-5293-8322
EuropeanaTech Community, Europeana Network Association, The Hague, Netherlands; Staatsbibliothek zu
Berlin – Berlin State Library, Berlin, Germany

Giulia Osti ⓘ orcid.org/0000-0003-3179-6980
School of Information and Communication Studies, University College Dublin, Dublin, Ireland

Daniel van Strien ⓘ orcid.org/0000-0003-1684-6556
Hugging Face, Glasgow, Scotland, UK

REFERENCES

Alkemade, H., Claeyssens, S., Colavizza, G., Freire, N., Irollo, A., Lehmann, J., Neudecker, C., Osti, G.,
& **van Strien, D.** (2023). *Datasheets for Digital Cultural Heritage Datasets*. Version 1. (last accessed:
October 4th, 2023). DOI: https://doi.org/10.5281/ZENODO.8375033

Apte, P. (2017, September 27). The Data Scientist Putting Ethics Into AI. (last accessed: October 4th,
2023). https://web.archive.org/web/20170930075045/http://www.ozy.com/rising-stars/rumman-
chowdhury-the-human-centric-thinker/81044

**Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S.,
Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R.,** & **Herrera, F.** (2020). Explainable Artificial
Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.
Information Fusion, 58, 82–115. DOI: https://doi.org/10.1016/j.inffus.2019.12.012

Beals, M., & **Bell, E.** (2020). *The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges*.
(last accessed: October 4th, 2023). DOI: https://doi.org/10.6084/M9.FIGSHARE.11560059

Beelen, K., Lawrence, J., Wilson, D. C. S., & **Beavan, D.** (2023). Bias and representativeness in digitized
newspaper collections: Introducing the environmental scan. *Digital Scholarship in the Humanities,
38*(1), 1–22. DOI: https://doi.org/10.1093/llc/fqac037

Brate, R., Nesterov, A., Vogelmann, V., van Ossenbruggen, J., Hollink, L., & **van Erp, M.** (2021). Capturing
Contentiousness: Constructing the Contentious Terms in Context Corpus. *Proceedings of the 11th on
Knowledge Capture Conference*, 17–24. DOI: https://doi.org/10.1145/3460210.3493553

British Library, Morris, V., van Strien, D., Tolfo, G., Afric, L., Robertson, S., Tiney, P., Dogterom, A., &
Wollner, I. (2021). *19th Century Books—Metadata with additional crowdsourced annotations*. (last
accessed: October 4th, 2023). DOI: https://doi.org/10.23636/BKHQ-0312

**Candela, G., Gabriëls, N., Chambers, S., Pham, T.-A., Ames, S., Fitzgerald, N., Hofmann, K., Harbo, V.,
Potter, A., Ferriter, M., Manchester, E., Irollo, A., Van Keer, E., Mahey, M., Holownia, O.,** & **Dobreva,
M.** (2023). *A Checklist to Publish Collections as Data in GLAM Institutions*. (last accessed: October 4th,
2023). DOI: https://doi.org/10.48550/ARXIV.2304.02603

Conway, P. (2015). Digital transformations and the archival nature of surrogates. *Archival Science, 15*(1),
51–69. DOI: https://doi.org/10.1007/s10502-014-9219-z

Corrado, E. M., & **Moulaison Sandy, H. L.** (2017). *Digital preservation for libraries, archives, and museums*
(Second Edition). Rowman & Littlefield.

Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., de Vries, J., L'Hours, H., Davidson, J., &
White, A. (2020). *FAIRsFAIR Data Object Assessment Metrics*. (last accessed: October 4th, 2023). DOI:
https://doi.org/10.5281/ZENODO.4081213

Edmond, J., & **Lehmann, J.** (2021). Digital humanities, knowledge complexity, and the five 'aporias' of
digital research. *Digital Scholarship in the Humanities, 36*(Supplement_2), ii95–ii108. DOI: https://doi.
org/10.1093/llc/fqab031

Fiorucci, M., Khoroshiltseva, M., Pontil, M., Traviglia, A., Del Bue, A., & **James, S.** (2020). Machine
Learning for Cultural Heritage: A Survey. *Pattern Recognition Letters, 133*, 102–108. DOI: https://doi.
org/10.1016/j.patrec.2020.02.017

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & **Crawford,
K.** (2021). Datasheets for Datasets. *Communications of the ACM, 64*(12), 86–92. DOI: https://doi.
org/10.1145/3458723

Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & **Wallach, H.** (2019). Improving Fairness in
Machine Learning Systems: What Do Industry Practitioners Need? *Proceedings of the 2019 CHI Conference
on Human Factors in Computing Systems*, 1–16. DOI: https://doi.org/10.1145/3290605.3300830

Hubbard, D. W. (2010). *How to measure anything: Finding the value of 'intangibles' in business* (Second
Edition). Hoboken, NJ: Wiley. DOI: https://doi.org/10.1002/9781118983836

Jo, E. S., & **Gebru, T.** (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 306–316. DOI: https://doi.org/10.1145/3351095.3372829

Kapoor, S., & **Narayanan, A.** (2022). *Leakage and the Reproducibility Crisis in ML-based Science.* (arXiv:2207.07048). (last accessed: October 4th, 2023). DOI: https://doi.org/10.48550/ARXIV.2207.07048; https://doi.org/10.1016/j.patter.2023.100804

Kirk, H. R., Birhane, A., Vidgen, B., & **Derczynski, L.** (2022). *Handling and Presenting Harmful Text in NLP Research.* (arXiv:2204.14256). (last accessed: October 4th, 2023). DOI: https://doi.org/10.48550/ARXIV.2204.14256; https://doi.org/10.18653/v1/2022.findings-emnlp.35

Lee, B. C. G. (2023). The "Collections as ML Data" Checklist for Machine Learning & Cultural Heritage. *Journal of the Association for Information Science and Technology*, 1–22. DOI: https://doi.org/10.1002/asi.24765

Library of Congress. (2019, December). *Encoded Archival Description Tag Library Version EAD3 1.1.1.* (last accessed: October 4th, 2023). https://www.loc.gov/ead/EAD3taglib/EAD3.html

Luthra, M., Todorov, K., Jeurgens, C., & **Colavizza, G.** (2022a). *Unsilencing Colonial Archives via Automated Entity Recognition* (arXiv:2210.02194). (last accessed: October 4th, 2023). DOI: https://doi.org/10.48550/ARXIV.2210.02194

Luthra, M., Todorov, K., Wissen, L. van, Jeurgens, C., & **Colavizza, G.** (2022b). *Unsilencing Colonial Archives via Automated Entity Recognition.* (last accessed: October 4th, 2023). DOI: https://doi.org/10.5281/zenodo.7129316; https://doi.org/10.1108/JD-02-2022-0038

O'Neil, L. (2023, August 12). These Women Tried to Warn Us About AI. *Rolling Stone.* (last accessed: October 4th, 2023). Retrieved from https://www.rollingstone.com/culture/culture-features/women-warnings-ai-danger-risk-before-chatgpt-1234804367/

Padilla, T. (2019). *Responsible Operations: Data Science, Machine Learning, and AI in Libraries.* Dublin, OH: OCLC Research. DOI: https://doi.org/10.25333/XK7Z-9G97

Padilla, T., Allen, L., Frost, H., Potvin, S., Roke, E., & **Varner, S.** (2022). *Always Already Computational: Collections as Data.* (last accessed: October 4th, 2023). DOI: https://doi.org/10.17605/OSF.IO/MX6UK; https://doi.org/10.1108/JD-02-2022-0038

Porter, T. M. (1996). *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life*. Princeton: Princeton University Press. DOI: https://doi.org/10.1515/9780691210544

Pushkarna, M., Zaldivar, A., & **Kjartansson, O.** (2022). Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1776–1826. DOI: https://doi.org/10.1145/3531146.3533231

Rakova, B., Yang, J., Cramer, H., & **Chowdhury, R.** (2021). Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 7: 1–7: 23. DOI: https://doi.org/10.1145/3449081

Reshetnikov, A., Marinescu, M.-C., & **Lopez, J. M.** (2022). *DEArt: Dataset of European Art* (arXiv:2211.01226). (last accessed: October 4th, 2023). DOI: https://doi.org/10.48550/arXiv.2211.01226

Scheuerman, M. K., Hanna, A., & **Denton, E.** (2021). Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 1–37. DOI: https://doi.org/10.1145/3476058

Singh, A. (2019). Beyond the Archive Gap: The Kiplings and the Famines of British Colonial India. *South Asian Review*, 40(3), 237–251. DOI: https://doi.org/10.1080/02759527.2019.1599562

UNESCO. (2003, March). *UNESCO Charter on the Preservation of the Digital Heritage—UNESCO Digital Library.* (last accessed: October 4th, 2023). Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000229034.locale=en. DOI: https://doi.org/10.1007/978-3-031-25056-9_15

Urton, G. (1997). *The Social Life of Numbers. A Quechua Ontology of Numbers and Philosophy of Arithmetic* (First Edition). Austin: University of Texas Press.

Van Erp, J. A. A., Langen, C. D., Boon, A., & **Van Bochove, K.** (2018). Testing the FAIR metrics on data catalogs. *PeerJ Preprints*, 6, e27151v2. DOI: https://doi.org/10.7287/peerj.preprints.27151v2

Wevers, M. (2022). *Fotopersbureau De Boer Training Set on Scene Detection* (0.2). (last accessed: October 4th, 2023). DOI: https://doi.org/10.5281/zenodo.7118409

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. DOI: https://doi.org/10.1038/sdata.2016.18

Wilkinson, M. D., Sansone, S.-A., Schultes, E., Doorn, P., Bonino Da Silva Santos, L. O., & **Dumontier, M.** (2018). A design framework and exemplar metrics for FAIRness. *Scientific Data*, 5(1), 180118. DOI: https://doi.org/10.1038/sdata.2018.118