



**Staatsbibliothek
zu Berlin**
Preußischer Kulturbesitz

STAATSBIBLIOTHEK ZU BERLIN – PK

INTERMEDIARIES, CRAFTED BY TRUSTEES

DATASHEETS FOR DIGITAL CULTURAL HERITAGE DATASETS

Dr. Jörg Lehmann, Research Project „Human.Machine.Culture“

- Cultural heritage institutions s enjoy a high level of trust per se
- This year only, the first survey on trust in museums was published in Germany (and Europe), conducted by the Institute for Museum Studies – Prussian Cultural Heritage Foundation (IFM-SPK)
- Museums are supported by the public, have reliably fulfilled expectations for centuries and represent consistency
- They are mission-driven & are not working towards profit maximisation (sigh)
- They provide valuable digital assets – objects are accompanied by high-quality metadata established by trained cultural heritage practitioners

- Machine learning models do not work well if the context in which they are used does not match the training or evaluation datasets
 - Machine learning models do not work well if these datasets reflect undesirable social biases
-
- ⇒ Datasheets and model cards are required („instruction leaflets“)
 - ⇒ Datasheets as “**Intermediaries**” between the domains of cultural heritage and machine learning
 - ⇒ There are templates available for both uses, which can be imagined as questionnaires
 - ⇒ Until last year, there was no template available for digital cultural heritage datasets
 - ⇒ „**Crafted by Trustees**”: Cultural heritage practitioners as both curators of datasets and authors of datasheets

„Datasheets for Digital Cultural Heritage Datasets“ (2023) Results from a combined Europeana Tech and Research Working Group

- Standardised form of documentation to facilitate communication between dataset creators and dataset consumers
- Structure:
 - Motivation
 - Composition
 - Collection process
 - Preprocessing / cleaning / labeling
 - Uses
 - Distribution
 - Maintenance
- Research article: Alkemade, H., Claeysens, S., Colavizza, G., Freire, N., Lehmann, J., Neudecker, C., Osti, G., & van Strien, D. (2023). *Datasheets for Digital Cultural Heritage Datasets*. Journal of Open Humanities Data, 9:17, pp. 1–11. DOI: <https://doi.org/10.5334/johd.124>
- Template: Alkemade, H., Claeysens, S., Colavizza, G., Freire, N., Irollo, A., Lehmann, J., Neudecker, C., Osti, G., & van Strien, D. (2023). *Template Datasheet for Digital Cultural Heritage Datasets*. Version 1. DOI: <https://doi.org/10.5281/ZENODO.8375033>

„Datasheets for Digital Cultural Heritage Datasets“ (2023) Some Specifics

- Collection Process
 - Digitisation Pipeline
 - Data Provenance
 - Use of Linked Open Data, Controlled Vocabulary, Multilingual Ontologies / Taxonomies
 - Version Information
 - Personal and Sensitive Information
- Uses
 - Social Impact of Dataset
 - Unanticipated Uses made of this Dataset
- Distribution
 - Dataset Curators

- Balancing act of integrating the expertise of employees of cultural heritage institutions and data science perspectives may complicate the creation of a datasheet
- Challenge of interdisciplinarity: Tension between the claim to objectivity of statistical methods and the unavoidable positionality in documentation
- So far no metrics for racism, misogyny, discrimination, exclusion or toxicity in historical data
- Power relations lead to the subjugated subjects being systematically silenced in the sources, but how should this be identified?
- Relational ethics: not only the dominant knowledge of the collectors, but also that of the collection "subjects" should be taken into account
- The dilemma of "mitigating harm" (= intervention into the dataset) vs. accepting the status quo
- Data collection means decontextualisation, but recontextualisation is laborious and expensive

▪ **Alignment with Data Catalog Vocabulary DCAT (ongoing)** => machine-readable metadata

[bold = actual fields; grey = clusters of fields; orange = possibly important gap; yellow = less important gap]

Source	Section	subsection / field	field	note	format/example	Source	Section	field	note	format/example
Datasheets for DCH	Motivation		Title [To be added]			DCAT-AP	Dataset	dct:title		
			Dataset summary [to be moved from its current 'independent' position]	how/why the dataset was created, language, domain, topic, genre...	free text		Dataset	dct:description		
		[Dataset description]	[higher-level description]	guidelines on how it differs from data set summary						
			Homepage	main page	URL		Dataset	foaf:page		
			Repository	if hosted on platform	URL		Distribution	dcat:accessURL		
			cf above - distributions can be related via repository (and others?), but there's no structured description for them	we will need to make a decision on whether to keep Distribution-level fields scattered or group them						
							Dataset	dcat:distribution	Points to a Distribution (see below)	URI
							Distribution	dcatap:availability	controlled voc, e.g. "available"	

- **Creation of a web app** presenting the template as well as supporting questions
 - => enables creation of datasheets in several formats according to the needs of their users
 - => various output formats like md, rdf, xml, pdf
 - => machine-readability to ensure interoperability and automatic ingest into existing catalogues and finding aids within CHIs, metasearch engines or dataset hubs
 - => webtool facilitates the preview of a given cultural heritage dataset and generates descriptive statistics