Exploring Data Provenance in Handwritten Text Recognition Infrastructure: Sharing and Re-Using Ground Truth-Data, Referencing Models, and Acknowledging Contributions.

Starting the Debate on How *We* Could Do This.[1]

C. Annemieke Romein [1) 2)], Tobias Hodel [3)], Femke Gordijn [1)], Joris J. van Zundert [1)], Alix Chagué [4) 5)], Milan van Lange [6)], Helle Strandgaard Jensen [7)], Andy Stauder [8)], Melissa M. Terras [9)], Pauline van den Heuvel [10)], Carlijn Keijzer [6)], Achim Rabus [11)], Chantal Sitaram [1)], Aakriti Bhatia[1)], Katrien Depuydt [12)], Mary Aderonke Afolabi [13)], Nastya Anikina [2)], Elisa Bastianello [14)], Lukas Vincent Benzinger [2)], Arno Bosse [1)], David Brown [15)], Ash Charlton [9) 16)], André Nilsson Dannevig [17)], Klaas Van Gelder [18) 19)], Sabine C.P.J. Go[2) 22)], Marcus J.C. Goh [2)], Silvia Gstrein [20) 21)], Sewa Hasan[2)], Stefan von der Heide [23)], Maximilian Hindermann [24)], Dorothee Huff [25)], Ineke Huysman[1)], Ali Idris [2)], Liesbeth Keijser [26)], Simon Kemper [26)], Sanne Koenders [2)], Erika Kuijpers [2)], Lisette Rønsig Larsen [27)], Sven Lepa [28)], Tommy O. Link [2)], Annelies van Nispen [6)], Joe Nockels [9) 16)], Laura M. van Noort [2)], Joost Johannes Oosterhuis [29)], Vivien Popken [30)], María Estrella Puertollano [2)], Joosep J. Puusaag [2)], Ahmed Sheta [31)], Lex Stoop[34)], Ebba Strutzenbladh[32)], Nicoline van der Sijs [12)], Jan Paul van der Spek [34)], Barry Benaissa Trouw [34)], Geertrui Van Synghel [1)], Vladimir Vučković [2)], Sonia Weiss [8)], David Joseph Wrisley[33)], Riet Zweistra[34)], and further anonymous citizen scientists/ volunteers of the Goetgevonden-project!

1) KNAW Humanities Cluster/ Huygens Institute; 2) Vrije Universiteit Amsterdam; 3) University Bern; 4) Université de Montréal; 5) ALMAnaCH, Inria, Paris; 6) NIOD Institute for War, Holocaust, and Genocide Studies; 7) University of Aarhus; 8) READ-COOP SCE; 9) University of Edinburgh; 10) Amsterdam City Archives; 11) University of Freiburg;12) Instituut voor de Nederlandse Taal; 13) Bonn Center for Dependency and Slavery Studies at the University of Bonn; 14) Bibliotheca Hertziana/ Max-Planck-Institute for Art History; 15) *Trinity College Dublin, 16) National Library of Scotland; 17) National Archives of Norway; 18) Vrije Universiteit Brussel; 19) State Archives Brussels; 20) University of Innsbruck 21) State Library of Tyrol; 22) University of Exeter; 23) CCS Content Conversion Specialists GmbH;* 24) *University of Basel; 25) University Library of Tübingen; 26) National Archives of the Netherlands; 27) Danish National Archives; 28) Rahvusarhiiv Estonia; 29) University of Amsterdam; 30) Research Centre for Hanse and Baltic History (FGHO); 31) Friedrich Alexander Universität Erlangen-Nürnberg; 32) University of Aberdeen; 33) NYU Abu Dhabi; 34) independent citizen scientist.*

[1] This article is the result of a writing sprint organised during a workshop at the Transkribus User Conference (TUC) 2022 on the Re-Use of Ground Truth and Acknowledging Contributions by: Annemieke Romein, Tobias Hodel, Femke Gordijn, Helle Strandgaard Jensen, Pauline van den Heuvel, Andy Stauder and Melissa Terras. Contributions have also been made by students from the Vrije Universiteit Amsterdam, who participated in the course *Introduction to Digital Humanities and Social Analytics* (2022) (which is part of the university minor *Digital Humanities*) taught by Annemieke Romein.

Corresponding author(s):
Dr. C. Annemieke Romein: Annemieke.Romein@Huygens.knaw.nl
Prof. Dr. Tobias Hodel: tobias.hodel@unibe.ch

Abstract: This paper discusses *best practices for sharing and re-using Ground Truth in Handwritten Text Recognition infrastructures, as well as ways to reference and acknowledge contributions to the creation and enrichment of data within these systems.*? We discuss how one can place Ground Truth data in a repository - and consequently notify others through HTR-United. Furthermore, we want to raise awareness about how to cite HTR-data, models and contributions made by volunteers. Moreover, when using digitised sources (digital facsimile) it becomes increasingly important to distinguish between the physical object and the digital collection. These topics all relate to the proper acknowledgment of work put into digitisation, transcribing and sharing Ground Truth HTR-data. This also points to broader issues about the use of machine-learning with archival and library contexts, and how the community should begin to acknowledge and record contribution and data provenance

Within the humanities, working with digital tools is no longer a novelty. However, there is a difficulty regarding how to cite digital resources and – in the case of transcriptions – Handwritten Text Recognition (HTR) models. When we discussed how to reference those above properly, we also began questioning whom we should acknowledge for their role in the creation process. HTR and – more generally – the latest engines for automatic text recognition processes depend on digitised sources and the transcriptions being made, to create and synthesise models via machine-learning. For general models, massive amounts of documents accompanied by correct and (ideally) uniform transcriptions are key. The production of these masses is therefore a challenge that falls in the category of big sciences.[2] Volunteers (citizen scientists) are often involved in data creation – in our case – Ground Truth transcriptions – but how do we properly acknowledge their contribution? Moreover, when talking about the digitisation efforts of the GLAM sector, we should acknowledge the production of digital facsimiles/digital images of documents. In this article we use the term *digital facsimiles* in a way that is a translation of the German *Digitalisat*, in other words: the resulting product of a digitisation. In our case we are talking about digital reproductions - either photos or scans - of physical objects containing text. At the same time, we should insist on getting insights into their digitisation strategies to determine what is digitally available and what has been left out.

The unease surrounding the topic of adequately referencing and acknowledgement and – the consequent – re-use of resources led to the organisation of a hybrid workshop at the Transkribus User Conference 2022. We aimed to discuss: *How to properly re-use, reference and acknowledge contributions: what are the best practices thus far*? The activities we channelled through a *silent discussion/ writing sprint* exceeded our expectations. This paper is the result of that exchange of ideas.

In this article, we start with an introduction to Ground Truth and an overview of how we propose to share and re-use it. In this context, we contextualise this with ethics and legal limitations of the sharing. Consequently, re-using Ground Truth requires proper referencing of the contributions *and* contributors discussed in the second part. In our conclusion, we combine these parts. This article is a first proposal that wants to start a discussion about ways to conduct and acknowledge work carried out to deal with training data for machine learning.

[2] For approaches to big science that depends on (worldwide) collaboration and strives for unified goals, see for example Chawla, Dalmeet. 2017. "A New 'Accelerator' Aims to Bring Big Science to Psychology." *Science*, November. https://doi.org/10.1126/science.aar4464.

1. Research Context: Ground Truth

In order to understand Ground Truth, we would like to make aware of the necessity to produce such labelled data. Only based on so-called Ground Truth it is possible to provide the means to train text recognition models and furthermore to judge the quality of said models. The term “Ground Truth” suggests that this is a form of data that adheres to specified standards and is (at least by a group of people) considered to be a correct representation of (in our case hand written material).

Initial transcriptions may still contain quite a few mistakes, thoroughly checking these by – most often – another person can lead to a *perfect* transcriptions. Therefore, Ground Truth should be understood as ‘gold standard’. As such, it can function as benchmarked data. The value of Ground Truth is already implicit in these assumption. Especially if we take the data-hunger of machine learning systems into account. From this perspective we learn that it is essential to have as much Ground Truth available as possible in order to provide large (or even general) models for certain scripts or types of handwriting. Also, we would like to make aware of the reduced amount of training data that is necessary, once large models are available that can be fine-tuned (in the sense of transfer learning).

Ground Truths can be drawn from many sources. A bespoke transcription can be produced from scratch for a specific HTR project, but it is often easier and more efficient to adapt a Ground Truth from a transcription or edition that already exists. This raises the issue of varying or conflicting transcription conventions that may not be easy to identify but can impact on the project that the Ground Truth, or combination of Ground Truths, is to be applied to. If the Ground Truth is to be shared and potentially bundled into multiple models, it is essential that these conventions are included in the description or metadata, or at least are made available in some form. This will help the ultimate user to select the GT that is most appropriate for their own project and help explain certain behaviours of a model.

Providing Ground Truth ánd *metadata*

Generally speaking, there are two main-stream ways of making transcriptions: diplomatic or semi-diplomatic. The former transcribes *as-is* taking the character-set into consideration; the latter allows for changes to improve readability, e.g. writing out abbreviations and reducing complexities with regards to special characters. There are also transcriptions that are hyperdiplomatic in the sense that ligatures ‘st’ are transcribed, or types of ‘s’ or ‘r’ are distinguished. All of them can be used as benchmarked data – but it should always be clear which choices have been made.

From a legal perspective and since the data is so valuable, Ground Truth should be understood as data (by)-product of a project and considered for publication (we come back to this below). In most legal systems Ground Truth is independent of image rights and can be made available by the creators/ producers of the data. Unfortunately, image rights might be in the way of (re)training Handwritten Text Models - as both Ground Truth and images are needed - storing whatever we can, in a - as good as possible - manner for further and future re-use.

2. Sharing Ground Truth

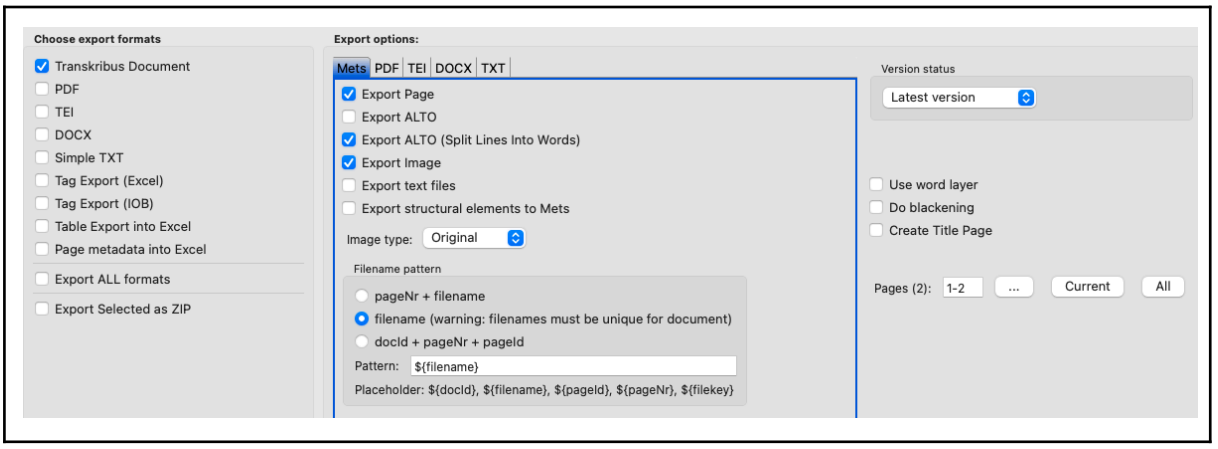
Much work and money are poured into producing manually and semi-manually Ground Truth-transcriptions. Re-using transcriptions – and the images they belong to – promises to greatly

support small(er) projects and speed up their processes. Furthermore, for advancing digital techniques, all available material could provide valuable training data for future projects and (new versions of) tools, like HTR engines or other downstream tasks like for example language models for Named Entity Recognition, to mention only one downstream task.[3] However, sharing transcriptions, e.g. in a repository, is, in our opinion, not enough and does not fully adhere to the FAIR principles:[4] it should be findable for others. Unfortunately, sharing data could have ethical and legal limitations.[5] These topics are discussed in the first section of this article. It should be stressed that we are explicitly talking about sharing Ground Truth data and not about sharing HTR models. The latter is tool-dependent and cannot be shared *outside* the tool; this topic will be addressed in this contribution's second section.

2.1 How to Export Data

The various programs that allow for the creation of OCR/ HTR have options to export the generated and/ or corrected transcriptions. When possible – depending on potential copyright/ image rights – both should be exported, the transcriptions as well as the images.[6] If this is not possible, it is already helpful *to at least* sustainably store the transcriptions.

Within the *Transkribus* tool, provided by the READCOOP, the export looks as shown in image 1 (below). The ALTO XML and the PAGE XML allow for an alignment between image and transcription – based on coordinates – which is a must to connect transcribed text on a word or a line basis with images and allows for the opportunity to (re)train machine learning based models.[7] These two formats are also supported by the eScriptorium application (see figure 2 below) that has been developed in the context of a variety of national (France) and European projects.[8]



[3] Phillip Benjamin Ströbel et al., 'Evaluation of HTR Models without Ground Truth Material' (arXiv, 29 April 2022), https://doi.org/10.48550/arXiv.2201.06170.

[4] Mark D. Wilkinson et al., 'The FAIR Guiding Principles for Scientific Data Management and Stewardship', *Scientific Data* 3, no. 1 (15 March 2016): 160018, https://doi.org/10.1038/sdata.2016.18.

[5] See section xx

[6] Some institutions make their images available through IIIF; in such cases one should not need to (re)share the images as the path-information to the images is included in the XML.

[7] In the Transkribus environment, depending on the amount of documents and pages, this might take a while and when chosen the server export one will receive an e-mail with a link to download the files when they are available.

[8] For more information see Kiessling, Benjamin, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. "EScriptorium: An Open Source Platform for Historical Document Analysis." In 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), 2:19–19. https://doi.org/10.1109/ICDARW.2019.10032.

Figure 1. Screenshot Transkribus Export [version 1.22.0.1-SNAPSHOT]. [30 September 2022]

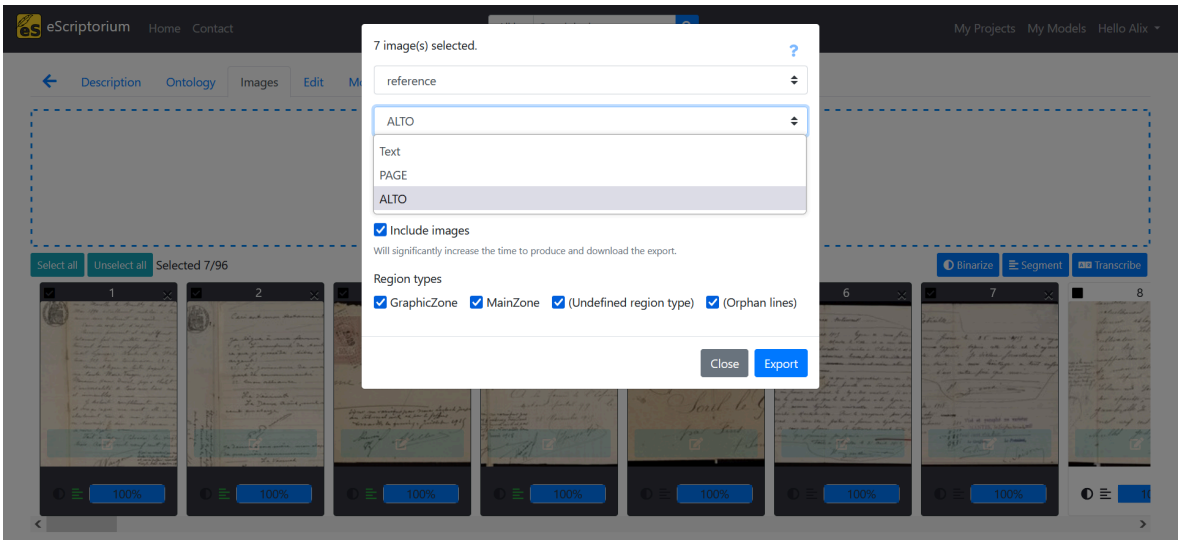


Figure 2. Screenshot eScriptorium Export [Version 0.12.5b]. [14 October, 2022]

ALTO and PAGE are the main formats used to store HTR output, while TEI, which is better known in the DH community, is primarily dedicated to further steps toward critical digital scholarly editions. Although it is hard to foresee future developments, we are optimistic that at least a future conversion to later formats from PAGE and ALTO XML will be possible. As a consequence, we encourage exports in these formats. Both PAGE and ALTO XML are open data formats defining an XML structure while leaving the option to add custom properties.

While some would call for a centralised Ground Truth repository, this could be both a costly affair (who could or should declare itself responsible for the longevity of such an environment) and result in doubling the work as funding agencies could have requirements to store output on specified (e.g. national) repositories. Consequently, a solution to the decentralised distribution of sources is discussed below.

2.2 Publishing Data in a Repository

There are various factors involved in choosing a repository to store your data: your institution or country might have its own repository that you are obliged to use, or you might have restrictions on what you can share, which are data-dependent. Generally, storing data in a FAIR-compliant, non-commercial repository with a persistent identifier, like Digital Object Identifiers (DOIs) is preferred. At the same time, it is highly encouraged to make the data output in a clearly structured form accessible. Images and XML files should reside in sub-folders, with indicating names for folders as well as images.

Repositories, like Zenodo, offer the possibility to add structured metadata that includes the name of contributors, licenses for re-use and (if applicable) URLs to external webpages. Furthermore, adding more detailed information in a README file is good practice and helps in the case of re-use

to navigate the data dump.[9] As an alternative, data can be provided using publicly available Git repositories such as GitHub (owned by Microsoft) or Gitlab – but these do not provide DOIs. In order to use both easy to use git-environments while receiving DOIs, a mixed solution is a possible pathway: version management can be done through GitHub, while Zenodo stores versioned and (in-)frequently updated datasets. Conveniently, some platforms like GitHub offer the possibility to link a repository with Zenodo semi-automatically. GitHub then handles the versioning (and creation of releases). At the same time, Zenodo provides the user with a DOI, on top of making the repository findable in the Zenodo search engine (see figure 3 below). If set in place, this allows for different versions of transcriptions and documents to become available online, based upon different parameters or (underlying) HTR models. Here, again the question of which types of transcriptions we are talking about comes up: manual Ground Truth or automatic transcriptions. Whichever version you are posting: it should be clear to potential other users what the type and status of the transcriptions are.

At the same time, one would like to have insights into the Ground Truth transcription rules of underlying documents could be considered, and consequently different understandings of 'text' could be made available as outputs.[10] For example, the textual output could include larger or smaller character sets (e.g. including/excluding abbreviated specific Unicode characters or silently expanded versions) or only parts of a 'document' (from a specific hand/time) would be published.

Using a repository to share data is – obviously – good academic practice and useful. However, to make the data not only available but furthermore findable – as is required by FAIR principles – at least a link to a sharing platform like HTR-United (see below) should be considered.[11]

[9] Miguel-Angel Sicilia, Elena García-Barriocanal, and Salvador Sánchez-Alonso, 'Community Curation in Open Dataset Repositories: Insights from Zenodo', *Procedia Computer Science*, 13th International Conference on Current Research Information Systems, CRIS2016, Communicating and Measuring Research Responsibly: Profiling, Metrics, Impact,Interoperability, 106 (1 January 2017): 54–60, https://doi.org/10.1016/j.procs.2017.03.009.

[10] For different understandings of text with regards to scholarly editions, see Patrick Sahle, 'What Is a Scholarly Digital Edition?', in *Digital Scholarly Editing : Theories and Practices*, ed. Matthew James Driscoll and Elena Pierazzo, Digital Humanities Series (Cambridge: Open Book Publishers, 2017), 19–40, http://books.openedition.org/obp/3397.

[11] An alternative to accessing Ground Truth, the IMPACT group offers its own Ground Truth repository (https://www.digitisation.eu/resources/impact-dataset/). In order to upload individual ground truth, contact to the IMPACT centres needs to be established.

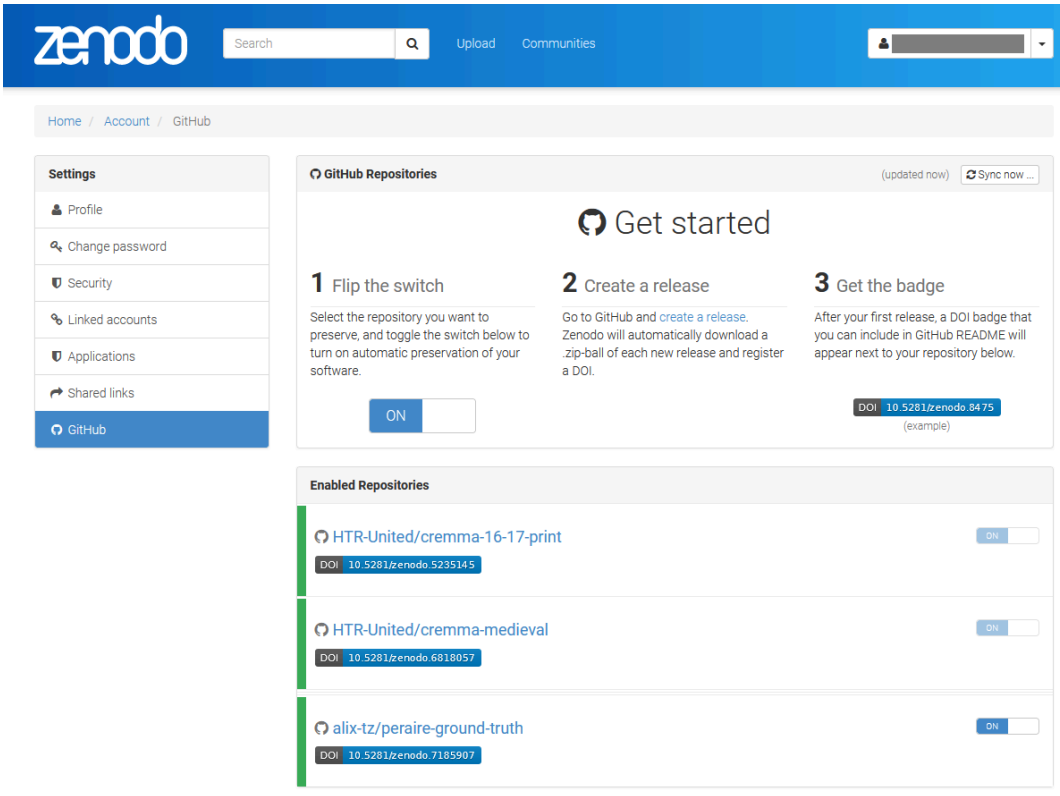


Figure 3. GitHub synchronisation and DOI generation on Zenodo. [17 October 2022]

2.3 HTR-United: Sharing Your Data

Several programs allow for the creation of OCR and/ or HTR. Regardless of the program used, it is up to the creators whether – or not – they want to share their work. Given the enormous diversity and the potential limitless repositories where work could be stored, there is an increasing need to have an overview of available Ground Truth datasets or – if possible – open-sourced models. Furthermore, the relative novelty of the output type requires new habits to publish them.

Supporting users and (small) projects, Alix Chagué and Thibault Clérice developed the HTR-United initiative (see figure 4).[12] HTR-United consists of three imperatives: 'a collaborative enterprise for the community; friendly to consumers and data producers; as low tech as possible (because $$).'[13] Furthermore, 'minimal computing connotes digital humanities work undertaken in the context of some set of constraints. These could include lack of access to hardware or software, network capacity, technical education, or even a reliable power grid.'[14]

[12] Alix Chagué and Thibault Clérice, 'HTR-United', 17 October 2022, https://htr-united.github.io/index.html..
[13] Alix Chagué and Thibault Clérice, 'Sharing HTR Datasets with Standardized Metadata: The HTR-United Initiative', in *Documents Anciens et Reconnaissance Automatique Des Écritures Manuscrites* (Paris, France: CREMMALab, 2022), sec. 13, https://hal.inria.fr/hal-03703989.
[14] Roopika Risam and Alex Gil, 'Introduction: The Questions of Minimal Computing', *Digital Humanities Quarterly* 2, no. 16 (2022): sec. 3.

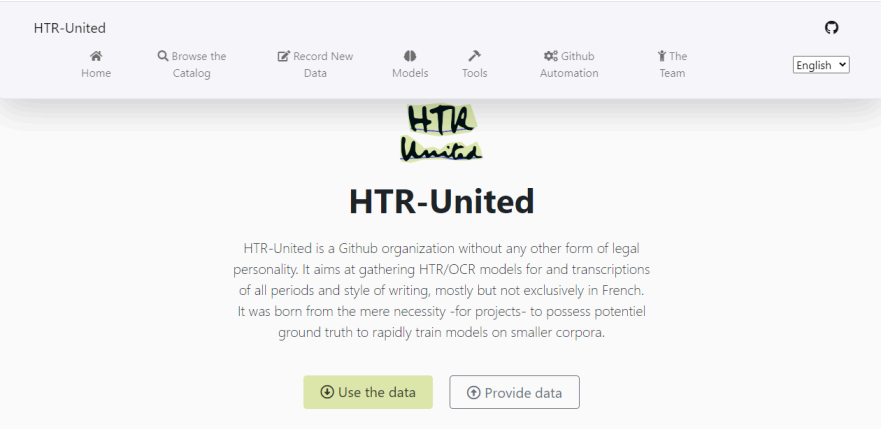


Figure 4. Website of HTR-United. https://htr-united.github.io/index.html [30 September 2022]

This much-needed initiative offers an easy to use and available solution, allowing contributors to store their dataset at any given location (preferably with a DOI). It also centralises an overview of those Ground Truth Datasets. The HTR-United interface allows users to filter Ground Truth on language, script/-type, and periodisation. Furthermore, the catalogue contains metadata (.yml), updated through *Continous Integration* through GitHub Actions. Chagué and Clérice developed a form that eases the process of creating .yml and badges and uploading metadata in the catalogue.[15] The developers (and at the same-time initiators) know that the schema with questions gains complexity. However, they think it is worth the effort as it provides a uniform overview of the digital environment.

[15] Chagué and Clérice, 'Sharing HTR Datasets with Standardized Metadata', sec. 15.

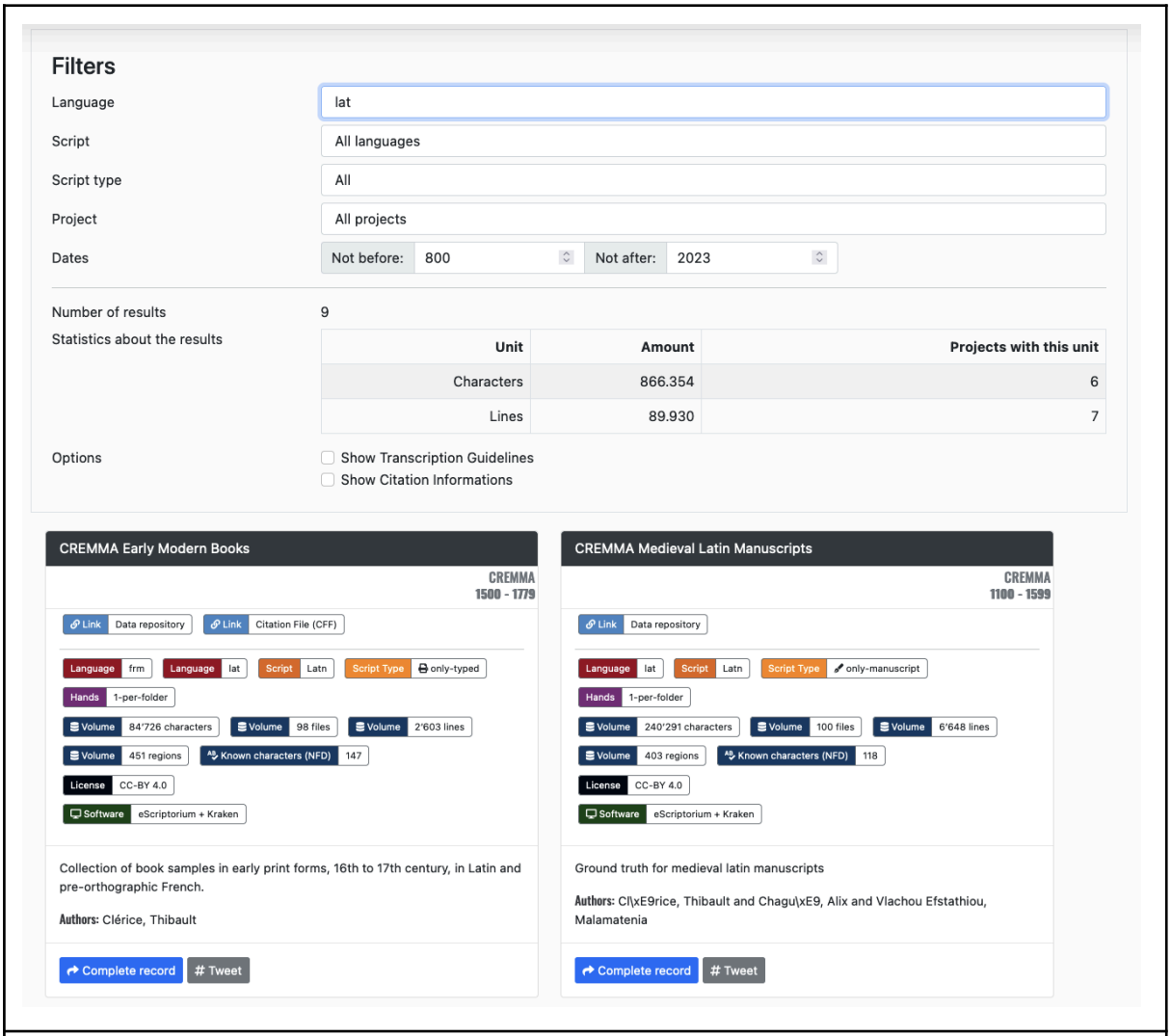
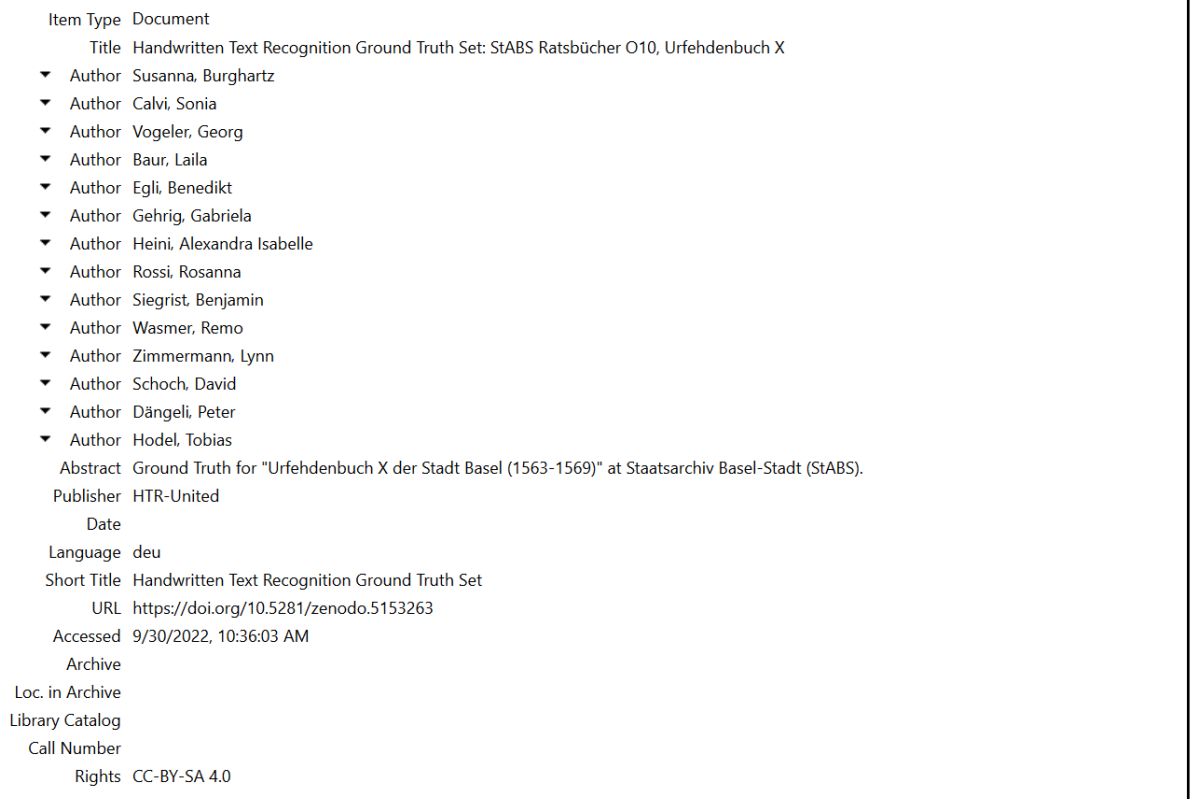


Figure 5. Catalogue of HTR-United. https://htr-united.github.io/catalog.html [30 September 2022]

While HTR-United limits itself – or the collaborators – to a pre-described way to share the Ground Truth, it does – for practical reasons – provide a relatively strict schema for the catalogue allowing for a machine-actionable way to check the conformity of the submissions and supports searches across the catalogue.[16]

From the catalogue on the HTR-United (see figure 5), it is possible to download the metadata into Zotero as an *Item Type:* (digital) document. This download-option eases the – future – referencing process (see figure 6).

[16] Ibidem, 10.



Item Type	Document
Title	Handwritten Text Recognition Ground Truth Set: StABS Ratsbücher O10, Urfehdenbuch X
Author	Susanna, Burghartz
Author	Calvi, Sonia
Author	Vogeler, Georg
Author	Baur, Laila
Author	Egli, Benedikt
Author	Gehrig, Gabriela
Author	Heini, Alexandra Isabelle
Author	Rossi, Rosanna
Author	Siegrist, Benjamin
Author	Wasmer, Remo
Author	Zimmermann, Lynn
Author	Schoch, David
Author	Dängeli, Peter
Author	Hodel, Tobias
Abstract	Ground Truth for "Urfehdenbuch X der Stadt Basel (1563-1569)" at Staatsarchiv Basel-Stadt (StABS).
Publisher	HTR-United
Date	
Language	deu
Short Title	Handwritten Text Recognition Ground Truth Set
URL	https://doi.org/10.5281/zenodo.5153263
Accessed	9/30/2022, 10:36:03 AM
Archive	
Loc. in Archive	
Library Catalog	
Call Number	
Rights	CC-BY-SA 4.0

Figure 6. Example of a GT Set. The example is of specific interest since it results from a multi-stage process. The transcription has been done within one project by several student assistants, following the structure of a DH expert and the project head (Burghartz, Susanna, Sonia Calvi, and Georg Vogeler. 2017. *Urfehdenbuch X Der Stadt Basel (1563-1569)*. Edited by Susanna Burghartz, Sonia Calvi, and Georg Vogeler. Graz: Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities Karl-Franzens-Universität Graz. http://edoc.unibas.ch/58852/.). Due to the open publication of the dataset (as TEI XML) alongside the images by the archives, an independent research group was able of running a text-to-image process that resulted in an annotated dataset suitable for training an HTR model. Only thanks to the publication as an open TEI XML data set in the first place, the further processing was possible.
https://htr-united.github.io/share.html?uri=https://doi.org/10.5281/zenodo.5153263 [30 September 2022]

To briefly conclude the section on sharing the data, we would like to emphasize four key approaches to processed textual data for future text recognition.

1) Export your data (including images, if possible);
2) Upload it online, using services compatible with versioning like GitHub or better in repositories;
3) Get a DOI, make it a publication;
4) Make others aware of it (through HTR-United or other possible means);

***

In the above, we focussed on sharing Ground Truth. In other words, the texts have been corrected manually. However, when models perform well, we may reach a point where sharing large datasets of raw HTR-produced transcriptions – while not perfect due to their errors – might also prove helpful. These, too, might be useful to share; however, they should be explicitly indicated as

machine-generated transcriptions, indicating the calculated or assumed Character Error Rate (CER) is necessary in that case.[17] It should thus be made evident that it is not human-corrected Ground Truth. In the case of such machine-generated transcriptions, it could be advisable to indicate not only the used model but also the CER of that model as a point of reference for the – potential – correctness of these machine-generated transcriptions.[18]

2.4 Referencing Digitised Resources and Digital Output

Several software solutions to create, collect, edit, and re-use bibliographic references for annotation purposes exist – to name but a few: EndNote, Citavi, Zotero, and Mendeley.[19] Zotero is a free, open-source referencing tool provided by the corporation for digital scholarship – that can adapt to various referencing styles.[20] As this is a free and open-source tool that has been programmed by and for humanities scholars,[21] we use this as a point of reference to come to suggestions on how to reference and acknowledge digital (re)sources and contributors.

Sharing data and models are of great importance to the academic world and the data providers; we – the contributors to this article – since there are some challenges in correctly referencing this in output. There is a risk of negligence, because there is no consensus on referencing 'digitised texts' or their contributors. As a brief clarification, when we talk about 'digitised texts', we mean the 'digital facsimile' (basically the output of a scanner or digital camera).

Questions about referencing are quite apparent for typical objects of the humanities, such as physically published books. It needs to be stated who wrote the text, who contributed, and what was the source. Depending on the citation style, the exact structure needs to be followed. For his section, we focus on referencing digital objects, whether they are resources (digitised texts), datasets (recognized texts) or even HTR models (algorithms that allow the processing of digitised texts). In comparison to manuscripts, prints, and other forms of written documentation that have been referred to for centuries and even millenia, we are only in the infancy of dealing with digital (ephemeral) objects.[22] We have thus combined experiences, suggestions and guidelines in this section. Like above, with a focus on FAIR and CARE principles and while striving to use persistent identifiers. The dominant point of view will be on the application and the *how* to apply references.

---

[17] Tobias Hodel et al., 'General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example', *Journal of Open Humanities Data* 7, no. 0 (9 July 2021): 13, https://doi.org/10.5334/johd.46; Ströbel et al., 'Evaluation of HTR Models without Ground Truth Material'.

[18] Cordell, Ryan. 2017. "Q i-jtb the Raven": Taking Dirty OCR Seriously. Book History John Hopkins University Press, 20(1):188-225.
Cordell, R. 2020. Talking about Viral Texts Failures. Available at: https://ryancordell.org/research/VT-database-fail/

[19] https://www.mendeley.com/ [22 Sept. 2022]; https://endnote.com/product-details/compatibility [22 Sept. 2022]; Ann Chen, 'Library Guides: Mendeley: Home' [, https://aut.ac.nz.libguides.com/c.php?g=359376&p=2427744].

[20] Zotero version 6.0.15: https://www.zotero.org/ [22 Sept. 2022].

[21] Takats, Sean. 2010. "Facing Abundance: Zotero as an Enlightenment Tool." In American Society for Eighteenth-Century Studies. New Mexico.

[22] See as an introduction Föhr, Pascal. 2018. "Historische Quellenkritik Im Digitalen Zeitalter." Basel 2018. http://edoc.unibas.ch/diss/DissB_12621.pdf.

2.4.1 Referencing Datasets

Data models are only starting to be referenced at this point in time in research[23], which results in a lack of standards within the humanities domain.[24] However, within computer sciences and machine learning, guidelines on how to cite datasets – and software – do exist.[25]

Let us be clear about what we mean in the context of this article with datasets that could need referencing. In the first place, we think about transcriptions, which include the information on where on an image page to find the specific word or line. Thus, the manually created Ground Truth and machine-generated transcriptions and anything in between machine-generated but manually corrected Ground Truth. In the second place, in addition to either of these types of transcriptions, text enrichment or more generally semantic annotation is thought of: e.g. georeferencing place names, named entity recognition, and linking terms to authority data. While these activities may be integrated within one dataset, it should – under all circumstances – be stated clearly what has been done and/ or used, and by whom.

Standard literature management software, are only adopting the citation of datasets and software. Zotero, for example, is currently [22 September 2022] not supporting the output types of 'datasets' or 'data/ HTR models', but they state that the category 'datasets' will soon be added to their output types.[26] To a dataset would – according to the Zotero-forum – be added the following elements/metadata: *author(s); dataset title; publication date; version; data repository/ publisher; DOI; URL; License/ Rights;* and *Resource/ Medium*. Like this, the citability of such scholarly and scientific contributions will be even easier made aware of. And we can only hope that in future releases of large data sets such contributions get accordingly acknowledged.[27]

Reflections exports and clarifying documentation

One of the questions explored by the HTR-United initiative is to know how we can better document HTR Ground Truth datasets. Besides the information related to the source or the authorship of the datasets, elements such as which guidelines or choices were made during the transcription are essential. The majority of the available models (and, thus, the Ground Truth they were trained with) are not entirely diplomatic in the strict sense of the word. Some add word separation (no scriptura continua) or hyphens, some omit diacritical marks, modify punctuation, bring superscripts down the line, or ignore bold, italics, smallcaps and other typographical peculiarities. Some models (and the Ground Truth they are based upon) are designed to show intelligent capabilities such as resolving abbreviations, normalising orthography, and transcribing into another script.

Moreover, the (completeness) of the character set should be addressed too - e.g. missing numbers could cause challenges when re-using the dataset. It would be helpful to specify the features of the Ground Truth transcriptions and the capabilities of the models based on these transcriptions, respectively. It would be good to have a standardised way to describe such features and to make the Ground Truth-creators reflect upon the choices they made in the process of Ground Truth creation, possibly affecting data re-use.

It is challenging to imagine properly combining or re-using different datasets without such information. Of course, such a question is difficult to extend to existing datasets, but it is crucial as

---

[23] Above, we have shown that HTR-United currently uses the *Item Type: Document*.
[24] Only in the Natural Language Processing field we encounter references to hubs like huggingface (https://huggingface.co/) and language models, taggers, etc. stored on the platform.
[25] Timnit Gebru et al., 'Datasheets for Datasets' (arXiv, 1 December 2021), https://doi.org/10.48550/arXiv.1803.09010.
[26] 'DataSets'.
[27] The background here is that we see for example in Computer Vision a multitude of datasets that are only partially acknowledging the contributors. See e.g. the Cocos Dataset: https://cocodataset.org/#home.

soon as the dataset contains or is made of annotations. If it were possible to provide a set of parameters which need to be documented before allowing the export of data from the tool in which the Ground Truth was produced, that could create awareness of the importance of documenting choices.[28] Then the next step would be to have an integrated export option to a FAIR-abiding repository; such as Zenodo. Especially the technical details (e.g. not only name, used base model(s) but also the origin of training & validation set, number of epochs) should be included and ideally visible for all other users. Future users might want to trace the documents or the workflow when creating their own models/building new models with base.

Guidelines for metadata

Data[29], models[30], and even concrete objects[31] are seldom neutral. Over the last few years, the potentially egregious effects of using skewed or biased training data have been more coherently acknowledged in CS, ML, NLP, and other data intensive fields.[32] Some work has been done in these areas, particularly from data ethics and algorithmic bias perspectives. One approach is to apply bias mitigating algorithms or causal inference models as in-analysis mitigation strategies. Another approach is to ensure that sufficient pre-analysis documentation exists to warrant the responsible use of data. As Gebru et al. state, bias may be mitigated by 'careful reflection on the process of creating, distributing, and maintaining a dataset, including any underlying assumptions, potential risks or harms, and implications of use.'[33] Thus, responsible metadata does not just imply the application of FAIR principles[34] and the existence of sufficient provenance information. Responsible metadata also details why the data was gathered, for what research purposes and to what end the research was conducted. It adds a description of the data, how it was gathered, which relevant tools and technologies were used in the collection process, and if and how it underwent possible transformation processes (selection and 'cleaning')[35] and/or annotation ('labelling'). All this information is essential in determining if a dataset can be used or repurposed for specific research. Datasets that do not provide such information should probably be treated as suspects and with the highest reserve or at least tested in-depth.

Unfortunately, because of the incredible variety in formats and content of humanities digital data and resources, no single agreed-upon metadata schema exists that serves all purposes, needs, and contexts of researchers. The heterogeneity of humanities data is only matched by the prolifacy of metadata standards, of which at least three hundred exist.[36] However, the salient point is not that a particular data standard should be primary but that a trustworthy data source clearly states to which metadata schema its metadata is adhering.

Clear and comprehensive metadata allow for correct and comprehensive referencing and citation. As with digital data standards, there is no agreed-upon standard for referencing datasets.

---

[28] Information concerning the authority data to which reconciliation has taken place, should be mentioned too. To establish the source of information.

[29] Lisa Gitelman, ed., *Raw Data Is an Oxymoron* (MIT Press, 2013), https://doi.org/10.7551/mitpress/9302.001.0001.

[30] Robyn Speer, 'How to Make a Racist AI without Really Trying', ConceptNet blog, 13 July 2017, http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/.

[31] Steve Woolgar and Geoff Cooper, 'Do Artefacts Have Ambivalence? Moses' Bridges, Winner's Bridges and Other Urban Legends in S&TS', *Social Studies of Science* 29, no. 3 (1999): 433–49.

[32] Ninareh Mehrabi et al., 'A Survey on Bias and Fairness in Machine Learning' (arXiv, 25 January 2022), https://doi.org/10.48550/arXiv.1908.09635.

[33] Gebru et al., 'Datasheets for Datasets'.

[34] Wilkinson et al., 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'.

[35] With regards to 'cleaning' see also Rawson, Katie, and Trevor Muñoz. 2019. "Against Cleaning (Chapter 23)." In Debates in the Digital Humanities 2019, edited by Matthew K. Gold and Lauren F. Klein. University of Minnesota Press. https://doi.org/10.5749/j.ctvg251hk.

[36] Jenny Riley and Devin Becker, 'Seeing Standards: A Visualization of the Metadata Universe.', 2010, http://jennriley.com/metadatamap/seeingstandards.pdf.

However, like research software, datasets should best 'be cited on the same basis as any other research product such as a paper or a book'.[37] Proper citing of datasets facilitates research transparency and ensures credit and accountability on the part of the dataset producers. The Edinburgh University-based Digital Curation Center report on how to cite datasets provides excellent guidance.[38] Metadata fields that should be part of any data set citation, if known, comprise author, publication date, title, version, resource type, publisher, identifier, and location.

2.4.2 Referencing HTR Models

The *Item Type: Software* could be used for referencing an HTR-models – this is suggested on the Zotero forum.[39] It requests information such as *title, programmer, abstract, series, version* and *date, programming language, URL and rights*. Questions arise about whether such an *Item Type* is suitable for HTR models or whether other disciplines might offer more suitable approaches. Let us first make an inventory of useful elements for an HTR model.

One of the elements the authors of this paper would like to see in the annotations of the models – which could fall under the term URL but might be even more specific – is the possibility of having a DOI for their HTR model. If it were possible to generate these automatically – either when sharing publicly within a system, such as within the Transkribus infrastructure, or outside the system, such as with eScriptorium – that would be appreciated. Another possible desired integration is with ORCID to be unambiguous about the creator of the HTR model.[40] In order to even further complicate the issue, we would advise to also mention the programmer of the training and evaluation algorithms (the text recognition engines) in one way or another.

Another issue is adding another layer that keeps returning to the surface is that of the quality of a model, expressed in Character Error Rate (CER), and the number of tokens this has been based upon. Both the CER of the training- and the validation set and their respective sizes are of use here, as they tell potential other users about the quality of the HTR model, as well as tendencies to overfit.[41] That, however, brings us to the point that we would need to know more about the underlying sources, too: are they homo- or heterogeneous and what is the character set used. A sample of the text could be a helpful attribute here.

Then, to complicate matters, it is possible to create new models based on existing models (so called fine-tuning in machine learning terms). The existing model is then used as a so-called *base* model. Base models can also be stacked while creating the ideal model. By principle, the entire stack of base models preceding new base models should be referenced. However, from a technical point of view and regarding the HTR+ technique – one of the licensed techniques that the Transkribus platform uses, re-training of the whole dataset after three cycles (base model + training as a base model + training as a base model) would be advisable.[42] When referencing earlier base models, proper descriptions with attributions to the creator(s) are also very important.

---

[37] Stephan Druskat, 'Research Software Citation for Researchers', Research Software Citation, 17 October 2022, https://cite.research-software.org/researchers/.
[38] Alex Ball and Monica Duke, *How to Cite Datasets and Link to Publications*, A Digital Curation Centre 'Working Level' Guide (DCC How-to Guides. Edinburgh: Digital Curation Centre., 2015), https://doi.org/10.1007/1-4020-5340-1.
[39] 'Data Models', Zotero Forums, accessed 20 October 2022, https://forums.zotero.org/discussion/99896/data-models.
[40] 'ORCID', accessed 17 October 2022, https://orcid.org/.
[41] See with regards to interpreting HTR models Hodel, Tobias. 2020. "Best-Practices Zur Erkennung Alter Drucke Und Handschriften – Die Nutzung von Transkribus Large- Und Small-Scale." In DHd 2020. Spielräume Digital Humanities Zwischen Modellierung Und Interpretation, edited by Christof Schöch, 84–87. Paderborn. https://doi.org/10.5281/zenodo.3666689.
[42] Here we refrain from making remarks about the Pylaia-engine or Kraken.

Guidelines for metadata

Zotero supports an item type called *software*. However, in disciplines such as computational sciences and machine learning, such a generic designation falls short of describing the diverse digital objects that may be produced in any scientific domain currently, and it is, in any case, not sufficient to cover HTR models. Congruent to what has been said about metadata concerning datasets, we need quite a granular schema for describing HTR models. Mitchell et al. propose a 'model card' to inscribe sufficient metadata and context about a model.[43] Such model cards have been implemented in the Hugging Face[44] repository, the current go-to repository to publish data sets, models, and documentation for language models used in AI technologies. Metadata fields include model description, intended use, a how-to for application, limitations and bias, a description of the training data and procedure, evaluation methods and results, and a suggestion for how to cite the model.[45] HTR models being in a sense character-based language models combined with computer vision approaches, are a close relative of the language models made available in Hugging Face. The same model card metadata scheme would therefore be a good fit and a solution to make aware of bias and editorial decisions. This would also have communities strive for better understanding what different practices of preparing and curating datasets exist.

As for citing models, we suggest the same approach as suggested for data sets in the previous section. HTR models should be cited on the same basis as any other research product, a practice for which the report of the Digital Curation Center previously mentioned provides good practices.[46] Consequently, the metadata fields to include for HTR model citing are congruent with those to use for data set citation: author, publication date, title, version, resource type, publisher, identifier, and location. Right now, information is put on the Web without any of this information being present which makes it hard to know any of this information.

2.5 Ethics and Limitations of Sharing

Those sharing data must be aware of ethical implications and how to handle them. These can be regarding economic or societal aspects or related to personality rights, among other things. Questions include — but are not limited to: Does the sharing contribute to the sharer's subsistence? Who can contribute most to society by having (some control) over the data – e.g. by improving a HTR-platform? For how long should the data of people in the documents be protected? In this section, we will briefly venture into these aspects of sharing Ground Truth and HTR models to indicate various points of view without siding with either.

For a business, regardless of its governance mode – e.g., an ltd. company or a cooperative – maintaining a business model for survival is essential. As such, e.g. READ-COOP SCE has as a philosophy: *share as much as possible, and retain as much control as necessary.* As money flows to upholding the infrastructure and the servers, this is the part that needs to be controlled, but other than that – e.g. the Ground Truth or transcriptions that are produced – can be shared. With this, understanding and respect for all the hard work being poured into creating datasets are shown, and consequently, proper referencing is essential. This proper referencing – be it using eScriptorium/Kraken, Transkribus or any other tool – is important as any of these tools benefit from

[43] Margaret Mitchell et al., 'Model Cards for Model Reporting', in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19 (New York, NY, USA: Association for Computing Machinery, 2019), 220–29, https://doi.org/10.1145/3287560.3287596.

[44] 'Hugging Face – The AI Community Building the Future.', accessed 20 October 2022, https://huggingface.co/.

[45] Cf. for instance a concrete example on the GPT-2 model at 'Gpt2 · Hugging Face', accessed 20 October 2022, https://huggingface.co/gpt2.

[46] Ball and Duke, *How to Cite Datasets and Link to Publications*.

economies of scale, improving the AI's ability to correctly recognise texts, which is a shared, joint-cooperative-goal. While the joint, rather abstract goal is to improve recognition, the business model to maintain the service is not that surprising. Creating and maintaining a sustainable infrastructure (e.g. servers) would provide challenges for many users. Thus, sharing and exporting Ground Truth and machine-generated output data is allowed. However, sharing HTR models is only allowed within the platform, without the ability to export the models, into whose training a lot of computing power is invested free of cost for the users. Other tools, such as eScriptorium allow the sharing of models but require the user to set up their own infrastructure and maintain it.[47]

Not with regards to a legal but an ethical point of view, it is crucial to think about creators, curators and descendants of the treated material. Especially when working with historical materials originating from colonial contexts, it is crucial to take the biography of a document into consideration and to describe how it became part of an institution; and what implications there might be connected to making documents or sources as data publicly available.[48] There are also other considerations from non-Western communities who may have very different models and understanding of ownership and what is means to respect the content of historical documents. Thus, consequences of working and sharing data must be kept in mind. For this reason, besides FAIR principles, CARE principles need to be considered, since they cover a multitude of aspects and have been proposed by the Global Indigenous Alliance. CARE does not have the same standing as FAIR for the moment but brings ethics, as one key aspect, into the discussion, asks for the collective benefit of data production and sharing, plus demands authority to control and responsibility of the actors.[49]

An example from the NIOD Institute for War, Holocaust, and Genocide Studies illustrates the challenges sources can bring to the surface. In the HTR-based digitisation project 'First-Hand Accounts of War: War letters (1935-1950) from NIOD digitised' issues arise – due to traceable personal information that these letters sometimes contain (GDPR) and ethical considerations related to and caused by implications of the disclosure of information could have on next-of-kin or third parties involved, (past) agreements with restrictions by those donating their archives and, last but not least, author's rights that might apply to the original texts.[50]

Dutch legislation has not specified in detail how to deal with these issues. The community of archival professionals has provided additional but informal guidelines. 'Werkgroep AVG' (*Workgroup GDPR*) of the Royal Society of Archivists in the Netherlands (KVAN) illustrates how a data controller can comply with legal and ethical restrictions. [51] Some strategies relevant to the case at hand:

- *Anonymisation*: by editing personal data in such a way that they cannot be traced back to the actual person (neither directly nor indirectly);
- *Pseudonymisation*: by editing personal data in such a way that they can be traced back to the actual person (neither directly nor indirectly) without using additional data. These preconditions could be a 'key' that only authorised individuals have access to, which should be stored separately;

---

[47] 'eScriptorium Tutorial (en)', *LECTAUREP* (blog), 17 October 2022, https://lectaurep.hypotheses.org/documentation/escriptorium-tutorial-en.

[48] See as an example Ortolja-Baird, Alexandra, and Julianne Nyhan. 2021. "Encoding the Haunting of an Object Catalogue: On the Potential of Digital Technologies to Perpetuate or Subvert the Silence and Bias of the Early-Modern Archive." Digital Scholarship in the Humanities, October, fqab065. https://doi.org/10.1093/llc/fqab065.

[49] 'CARE Principles of Indigenous Data Governance', Global Indigenous Data Alliance, 17 October 2022, https://www.gida-global.org/care.

[50] Carlijn Keijzer, Milan van Lange, and Annelies van Nispen, 'First-Hand Accounts of War', accessed 20 October 2022, https://www.niod.nl/en/projects/first-hand-accounts-war.

[51] Working Group GDPR (Werkgroep AVG) of Information and Archive Knowledge Network (Kennisnetwerk Informatie en Archief – KIA), "Weten of vergeten? Handreiking voor het toepassen van de Algemene verordening gegevensbescherming in samenhang met de Archiefwet in de dagelijkse praktijk van het informatiebeheer bij de overheid" (2020) p.33-34. See: https://kia.pleio.nl/attachment/entity/a8e1caa5-0d59-4267-bbc0-4cd288b2a56c

- *Data minimisation:* storing no more personal data than is strictly necessary for the prescribed goal. This restriction is not a safeguard in itself, but the procedure can reduce the need for other safeguards;
- *Retention period and timely deletion;*
- *Privacy 'by default':* by implementing checks and balances into the system, such as authorised access, logging and monitoring;
- *Honouring the rights of whom the data concerns:* providing civilians with the right to view, rectify and/ or delete the data that concern them. Exceptions to this rule may apply, but always involve a careful weighing procedure of the rights of various stakeholders;
- *Information security:* by implementing risk analysis, data classification and auditing.

Legal and/ or ethical restrictions do not necessarily imply the impossibility of Ground Truth transcriptions or machine-generated transcriptions with a larger public. The strategies mentioned above show how customised approach solutions and technical and organisational measures can offer a solution to dealing with these restrictions.

3. Acknowledging Contributions

When we consider the proper acknowledgment of datasets and HTR-models – *ergo*, all the hard work that was put into the creation of Ground Truth and the human desire to be open about this (credit where credit is due!) when referencing the work of others in footnotes; we should not forget that the creation was a joint effort. As said, lacking consensus or even the knowledge of referencing digital contributions could result in negligence. As Ground Truth and transcriptions are often aided by 'the crowd', volunteers, or citizen scientists and digitisation is often the result of institutional activities; we would like to address arising issues with acknowledging these contributions in this part of our paper.

3.1 Acknowledging the Crowd/ Citizen Scientists

In an increasing number of digitisation projects, '*the crowd*' is essential in generating Ground Truth data by transcribing or correcting transcriptions. Consequently, these people often bring new historical facts to light. In that sense, these volunteers or 'crowd' frequently take up the role of being 'citizens in science'. In this section, we focus on the recognition and reward of the labour that has been poured into projects through the many hands of volunteers, and we look at the *best practices* of various projects and make new recommendations. Acknowledging the crowd is important because of their hard work and to provide insight into *how* and with *what* resources Ground Truth Data was produced. Proper referencing the crowd thus contributes to a more transparent data policy. However, there are no clear standards yet for how this should be done. The following section thus deals with the question of how to acknowledge the crowd sustainably and fairly.

Figure 7. Part of the REPUBLIC team-website: here volunteers are mentioned as a group. https://republic.huygens.knaw.nl/index.php/en/about-republic/team-2/

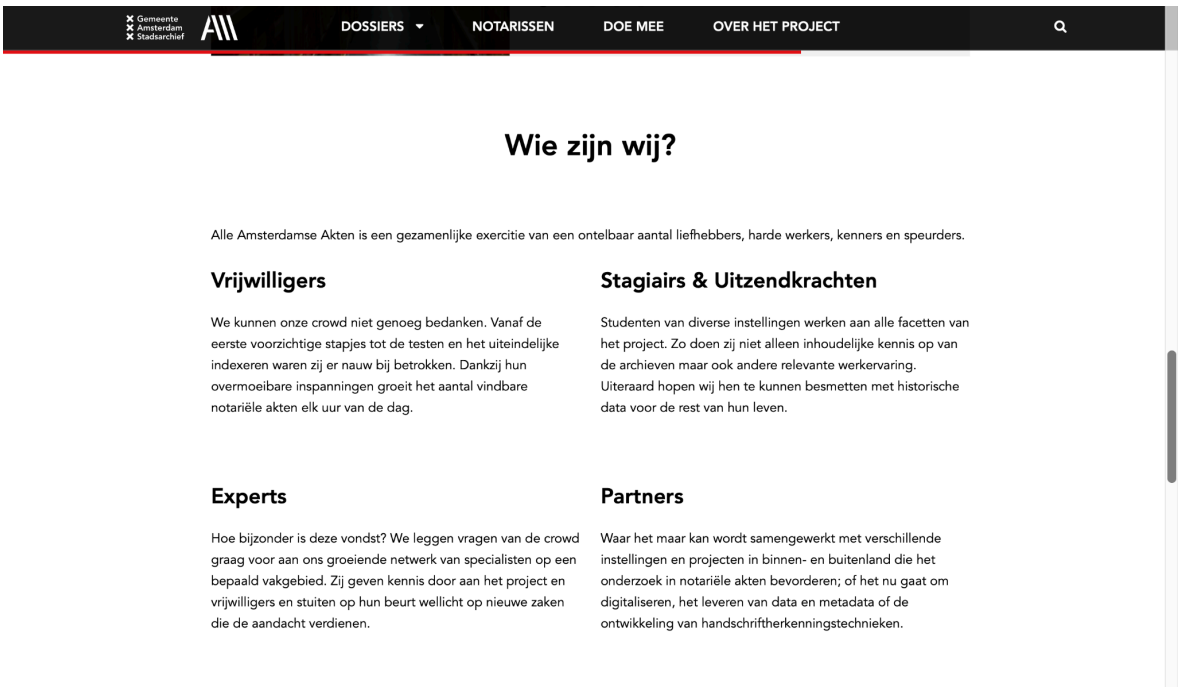


Figure 8. Part of *Alle Amsterdamse Akten* (All Amsterdam Notarial Deeds) Website - acknowledging the volunteers (crowd) as one group.

Acknowledging the crowd: current situation and room for improvement

While using the existing situation of crowdsourcing projects as examples, we find roughly two different methods of acknowledging volunteers. First, some projects refer to their volunteers in general, as if they were a homogeneous group (see figures 7 and 8).[52] Sometimes out of practical reasons, sometimes intentional, wanting to emphasise the collaborative effort instead of the individual. Second, there are also projects, especially smaller ones, that acknowledge their volunteers by listing them with their full credentials as a recognition for their work (see figures 9 and 10). In our view and in line with the previous sections of this article, these acknowledgements should be

[52] ivdnt.org, 'AI-Trainingset – Tag de Tekst voor Named Entity Recognition (NER)', *INT Taalmaterialen* (blog), accessed 20 October 2022, https://taalmaterialen.ivdnt.org/download/aitrainingset1-0/.

incorporated in the publication of the actual resulting datasets, too. How should that be done?

Whereas it is understandable that, due to administrative tasks, especially larger projects tend to acknowledge their volunteers in a more generalised manner, there are also arguments in favour of listing members of the crowd as individuals in the case of Ground Truth publication. We would like to provide three of such arguments.

First, choosing to name individuals is a more personal acknowledgement of their pivotal role in the data production process. Some volunteers appreciate being named for their efforts, and listing specific names gives credit to those deserving.

Second, a named acknowledgement in the case of a published dataset can also serve as a certificate of participation for members of the crowd. Participants can then list the dataset as a publication on their CVs, which allows them to demonstrate their knowledge of digital skills. These skills are especially important considering that humanities students, interns, and young programmers make up part of the crowd in many projects.

Third, acknowledging individuals as contributors to a dataset provides transparency to (future) users on how and by whom it was created (see also above, 2.4.1).

Experience learns that in many crowdsourcing projects, a small group of individuals contributes the majority of the work. Additionally, there often is a somewhat larger group of which the members contribute on a regular basis. Many of the volunteers however, only make a limited contribution after which they quit or they never actually start the work at all. In these cases, it could be considered to only name the volunteers who have exceeded a specific threshold of work. A personalised recognition could also provide room for listing the people who delivered most of the transcriptions first, whereas those who made smaller contributions are placed last on the list. However, instead of ranking members of the crowd for their contributions, names could be attached to the individual documents – or even pages – they transcribed. As such, not only credit is given to the person who produced the data but can also provide insight into the quality of individual transcriber's contributions.

Some caution is in place. While the above certainly provides future users with more transparency in the data curation process, it is essential to keep in mind that the idea from which crowdsourcing projects departed is that every contribution is welcome and valued. Many volunteers who start a new project are insecure about their palaeography skills, and not every participant can contribute substantial work due to personal situations. One should thus be cautious about ranking as this could be considered a (dis)qualification of their efforts. If at all, ranking volunteers or attaching individual transcribers' names to their specific contributions should be done in a motivating and engaging way. If a positive outcome of ranking is unsure, it is advisable to list the names alphabetically.

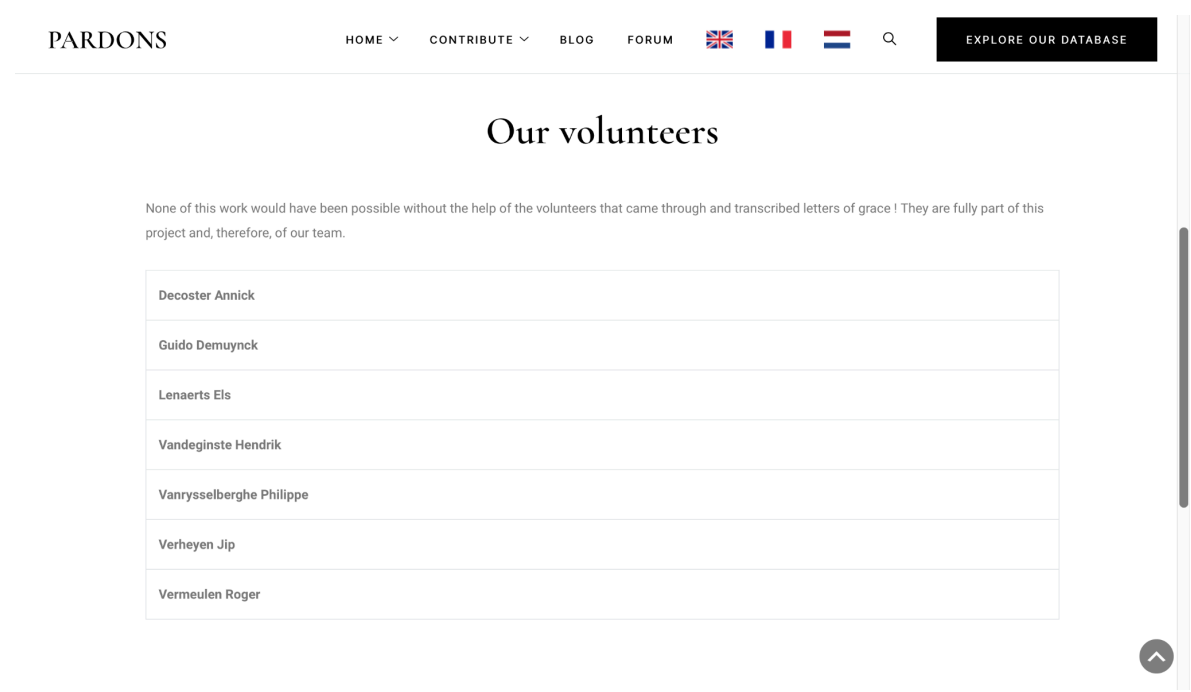


Figure 9. Part of the project-website of Pardons, here the names of all the volunteers are listed (https://pardons.eu/the-team/).

Figure 10. Volunteers listed on the website National archives of Estonia, including the files they transcribed (https://www.ra.ee/vallakohtud/index.php/site/top).

GDPR issues: opt-in or opt-out?

While listing individual citizen scientists is something to consider, there are some hurdles to take into account when publishing such a list. According to the European Union's General Data Protection

Regulations (GDPR), a person's name is a form of personal data. In this case, when listing the names of individual contributors, these people should be informed, and consent for using their names needs to be sought. Additionally, the possibility should remain that people – or their heirs, can redraw their names in the future.

Future complications could be solved by presenting the citizen scientists with a digital form asking them to check a box if they agree to being named in a publication before they apply to the project. Thus, they can knowingly opt-in. It is crucial that such a form clearly states how *exactly* their name would be used as part of expectation management *if* the participant allows for their name to be used at all. Under what conditions are names listed? Should a certain threshold have been met before a person is acknowledged? Are the names in alphabetical order, ranked, and/ or even connected to the individual output? The form should also provide information on how personal information is stored and kept safe.

However, one can imagine that, especially for larger projects which have already started, asking every individual member of the crowd for their consent can result in an administrative nightmare. There is an 'opt-out' method for these cases to deal with the GDPR regulations. Opt-out refers to a situation in which people are presented with the statement that data will be published with their names unless *they themselves* reach out and express their demand to be excluded to a specified person within a specific, reasonable timeframe. It is sufficient for projects to send the option to opt-out once, as this serves as proof for the initiative. One should be aware, though, that this method is riskier than using an opt-in, especially when many participants in a project are no longer active. If people miss the opportunity to opt out (due to changed contact details, for example) and specifically do not want to be mentioned by name, this could lead to discontent.

For both the opt-in and the opt-out option, the possibility should remain for volunteers and their heirs to withdraw their name at a later point in time. Information as to how this can be achieved should be available. In the case when someone requests *withdrawal* of their name being used, the name can no longer be used for future publications. However, the GDPR also allows for an appeal to *data erasure*. In these cases, the name should, if reasonably possible, also be removed from past publications. Thereby, the question should be asked if deleting the name prevents the achievement of the goals of the publication and/or research.

The practice of acknowledgement

Mentioning citizen scientists/volunteers on a project's website either by full name or a pseudonym is, at this point, the most common practice.[53] Nonetheless, at some moment, 'slow science' results in publications: either the publication of Ground Truth data, as previously mentioned or in an analysis of that same data. While within, e.g. the field of history, single-authored articles and monographs are still the norm, working with citizen scientists/volunteers may require adaptations to this approach, similar to the approach of publishing and citing the used datasets or HTR models.

As such, we would like to bring the CRediT-taxonomy forward.[54] Some journals, such as

[53] Although this section predominantly dealt with acknowledging volunteers by listing their names in the case of publication of Ground Truth data, there are of course a myriad of other ways to let the crowd know that their work is appreciated. Material gifts are often offered to members of the crowd but experience learns that volunteers' prime motive to get involved in projects is because they want to contribute to research. For this reason, a more appreciated form of acknowledgement is to involve the crowd in the field of research they contributed to for instance by inviting them to write a blog; or attend a (online) meeting during which results are presented. This also creates network opportunities from which, especially participating students and interns could benefit.

[54] Liz Allen et al., 'Publishing: Credit Where Credit Is Due', *Nature* 508, no. 7496 (April 2014): 312–13, https://doi.org/10.1038/508312a.

Science, already work with this model and add an acknowledgement section to their article.[55] The CRediT website states that it:

> '[...] grew from a practical realisation that bibliographic conventions for describing and listing authors on scholarly outputs are increasingly outdated and fail to represent the range of contributions that researchers make to published output. Furthermore, there is growing interest among researchers, funding agencies, academic institutions, editors, and publishers in increasing both the transparency and accessibility of research contributions.'[56]

The taxonomy lists – at this moment – fourteen different roles contributors could have, as indicated on the screenshot in figure 11 below:

Figure 11. CRediT-taxonomy. https://credit.niso.org/

While this overview might look complicated, work on Ground Truth, datasets or databases generally fits within the frame of *data curation* or *resources*.[57] Being explicit about a person's role will not only help avoid confusion about someone's role but also make aware of the different categories of contribution. Providing citizen scientists/ volunteers with a specific task – e.g., transcribing, correcting or tagging texts – could immediately be connected to such a specific role or task. Independent of such an initial indication, when the citizen scientists come across an exciting find, which leads to specific research, an additional role could be assigned in consultation with this individual.

From a legal perspective, one's role relates to potential author's rights – depending on one's legal system – as information is processed/ converted to a machine-readable format in the previously described cases. This activity implies that no original work is created, and therefore it is not covered by the author's rights.[58] However, courtesy could and should require a proper acknowledgement of work put into creating files.

[55] Mike Kestemont et al., 'Forgotten Books: The Application of Unseen Species Models to the Survival of Culture', *Science* 375, no. 6582 (18 February 2022): 765–69, https://doi.org/10.1126/science.abl7655.
[56] 'Background', *CRediT* (blog), 14 April 2020, https://credit.niso.org/background/.
[57] 'Data Curation', *CRediT* (blog), 12 June 2020, https://credit.niso.org/contributor-roles/data-curation/; 'Resources', *CRediT* (blog), 12 June 2020, https://credit.niso.org/contributor-roles/resources/.
[58] ECLI:NL:RBDHA:2022:8828, Rechtbank Den Haag, C/09/586380 / HA ZA 20-36, No. ECLI:NL:RBDHA:2022:8828 (Rb. Den Haag 3 August 2022).

3.2 Acknowledging Institutional Activities: Digitisation-Actions and Contextualisation

The GLAM sector – but certainly also private institutions – digitises their collections. Digitisation is a time-consuming process that is not magic even if the best technical solutions are sometimes seen as those that provide the most seamless (magical?) user experience! From the perspective of institutions, digitisation is a costly, time-consuming practice which belongs – by now – to their core business.[59] It will take time, and this steadily-paced process is not something that is often the topic of communication to the outside world. From the researcher's perspective, communicating the relationship between the – current version of the – online collection and the offline archive is of great use, as it will support critical reflection on the possible methodological implications of the choices made in the digitisation process. Alternatively, a document or video explaining how subject categories, search fields or filtering options were made/conceptualised can help understand the (in)complete online collection. This document or video could provide crucial details contributing to the researchers' understanding of data-provenance and archive structure and design.

Reflections exports and clarifying documentation
Researchers active with digital resources have developed a different understanding of provenance; for them, it covers questions, as shown in figure 12 below.

- Where does your data come from? - Who created it and why? - Has the data been selected from a more extensive set? If so, what were the criteria? - What does the data represent? - Is the (meta)data reliable? - Is there a bias of some sort we should be aware of? - What has been done to the data by the publisher/creator?	- What tools were used to for datafication and what is their (expected) quality? - Cleaned, modelled, altered, annotated, enriched? On what purpose? - Is the data well documented/described by the publisher/creator? - What physical aspects have gotten lost in the process of digitisation? - Can it be published/shared (license), or are there any restrictions? - What metadata system has been used to describe the data?
Figure 12. Selection of questions regarding *provenance* in the understanding of the digital humanities.[60]	

[59] Although not explicitly covered in this article, acknowledging institution or funder is polite. Their efforts and/or financial support allowed for the creation of Ground Truth. If it concerns earlier published texts (scholarly editions), their contribution toward state-of-the-art research in AI space is – often – considered rewarding.

[60] See: Hoekstra, Rik, en Marijn Koolen. 'Data scopes for digital history research'. Historical Methods: A Journal of Quantitative and Interdisciplinary History 52, nr. 2 (3 april 2019): 79–94. https://doi.org/10.1080/01615440.2018.1484676; Engelhardt, Claudia, Katarzyna Biernacka, Aoife Coffey, Ronald Cornet, Alina Danciu, Yuri Demchenko, Stephen Downes, e.a. D7.4 How to be FAIR with your data. A teaching and training handbook for higher education institutions (versie V1.2 DRAFT). Zenodo, 2022. https://doi.org/10.5281/zenodo.5905866; Hauswedell, T., Nyhan, J., Beals, M.H. *et al.* Of global reach yet of situated contexts: an examination of the implicit and explicit selection criteria that shape digital archives of historical newspapers. *Arch Sci* 20, 139–165 (2020). https://doi.org/10.1007/s10502-020-09332-1

The questions above are essential to researchers to do a conceptual translation from the physical object to the digital collection, which is more than the inventory number in its context of origin, the archivists' concept of the word provenance. Adjusting to the digital new world requires technical skills and resources to set up an infrastructure that integrates archival guarantees' characteristics: authenticity, reliability, integrity, and usability.[61] Here a clear and distinctive overview of competing standards – handles, (P)URLs, DOIs, URIs – can cloud the understanding – leading to mere digitisation without guarantees of authenticity and reliability.

This also brings us to the question of *what* has been digitised so far. So far, an overview of what has been digitised should be available on the websites of GLAM institutions that do digitise. Hauswedell et al suggest that the institutional choices which went into choosing items for digitisation should be made clear to users.[62] Jensen suggests that digital archives could be encouraged to demonstrate the extent and content of their digitisation efforts.[63] Here she implicitly refers to the reliability of the found resources – how much of the inventory has been digitised (% – see e.g. figure 13) – but also, what type of *datafication* has been applied: has the entire text been described, or merely names and places? Is transcription ongoing (meaning that searches could give a different result if happening days, weeks or months later.) If additional data has been created, those involved in that process should have the possibility of being acknowledged, even if this is 'just' part of their job. Such tasks could be considered the modern equivalent of assembling or describing an archive, which is the traditional role of archivists.[64]. Though archivists are rarely credited for this work as individuals, the question is if it would be helpful for both archivists and scholars to be named when part of digital projects in a similar way to people who work on digital projects in academia? Having a credits list or page would give workers on an increasingly precarious labour market a way to highlight their skills and experience (and be cited for it); it would make digital labour more visible; and people who use the resources would know who to contact if they have any questions related to the resources.

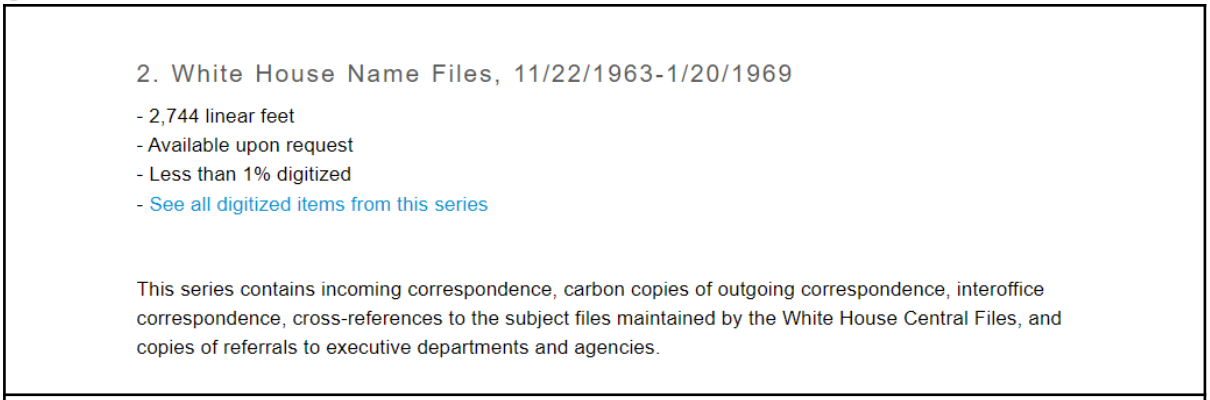


Figure 13. Section from the the White House Central Files (WHCF) created or collected by President Lyndon B. Johnson and his staff - with an indication how much of the files
hahttps://www.discoverlbj.org/exhibits/show/loh/pres/whcf

Combining the additional data – which could allow for different search methods – with the description of predefined categories and structures would extend users' freedom. It creates multiple entries that allow for differences and similarities between the archives and researchers' (changing) conceptual models.[65] Such

[61] 'What Are Archives? | International Council on Archives'.
[62] Hauswedell, T., Nyhan, J., Beals, M.H. *et al.* Of global reach yet of situated contexts: an examination of the implicit and explicit selection criteria that shape digital archives of historical newspapers. *Arch Sci* 20, 139–165 (2020). https://doi.org/10.1007/s10502-020-09332-1
[63] Jensen, 'Digital Archival Literacy for (All) Historians', 256.
[64] 'Blog', Georgian Papers Programme, 17 October 2022, https://georgianpapers.com/about/blog/; 'Stacks', Stacks, 17 October 2022, https://stacks.wellcomecollection.org/; Jensen, 'Digital Archival Literacy for (All) Historians', 258.
[65] Jensen, 'Digital Archival Literacy for (All) Historians', 257.

room for manoeuvre is an asset to open and different interpretations without an apparent influence of the contributors. According to Jensen, this would/ could result in different searches, including one targeting a range of related topics or production contexts.[66] She goes even further by saying:

> 'As with predefined subject categories, metadata reflects the conceptual model of the archival system's design. In the cases where archives use international ISO or similar standards, it is the biases of these that are reflected. This can easily cause problems for historians, both because of the differences in language historically and because our conceptual model does not comply with the metadata systems in use.'[67]

Given this situation, it would be interesting to adopt an (ISO-)standard for all institutions to follow to make data accessible and organised to whomever wants to access them.

A final concern voiced by Jensen is that: '[d]igitisation of archives depend on (additional) external funding, which means that they are likely to be subject to policies that emphasise popularity, marketisation, or current research trends.'[68] This concern could go two ways. On the one hand: one could argue that a selection bias based on the interests of funding individuals/institutions has been – and still is – also a problem of analogue archives. In other words, traditional archives require funding too, and the ones paying for them will necessarily have an influence on the archive's contents. One could spin this thought further and ask when the intended omission of information starts (and where it will end).

On the other hand, it has been argued that the digitisation of archives reduces the selection bias. Based on experience gained from small- and large-scale digitisation projects and gained from the literature, we cannot agree with that stance but point especially at political and infrastructural decisions.[69] Digitisation is thus often a combination of a selection made by institutions and requests made by users (scanning on demand or asking for better searchability of a digitised source) but furthermore also the availability of equipment and (financial) means to carry out such work and make it accessible, favouring the global north.

Whether digitisation really leads to increased information-transparency is thus still up for discussion. Still, from a researcher perspective with broad knowledge about an institution, educated conclusions can be derived. Furthermore, based on the online existence of certain materials online, can also for a general public lead to more interest in certain documents or objects. Referencing resources would lead to the GLAM sector's accountability for their work and, thus, hopefully, for (more) money to digitise other resources.

Guidelines for metadata

In Zotero, selecting the Item Type: Document and Books is possible. This includes fields to enter archives, location in the archive, and URLs, preferably a permanent URL (PURL).[70] Their availability certainly opens up possibilities; however, the author(s) should be aware that not every publisher is

---

[66] Ibidem, 258.
[67] Ibidem, 259.
[68] Ibidem, 254.
[69] See for example for newspaper digitisation: Hauswedell, Tessa, Julianne Nyhan, M. H. Beals, Melissa Terras, und Emily Bell. 2020. «Of Global Reach yet of Situated Contexts: An Examination of the Implicit and Explicit Selection Criteria That Shape Digital Archives of Historical Newspapers». *Archival Science* 20 (2): 139–65. https://doi.org/10.1007/s10502-020-09332-1. And more broadly Zaagsma, Gerben. 2022. «Digital History and the Politics of Digitization». *Digital Scholarship in the Humanities*, September, fqac050. https://doi.org/10.1093/llc/fqac050.
[70] Laura Rueda, Martin Fenner, and Patricia Cruse, 'DataCite: Lessons Learned on Persistent Identifiers for Research Data', *International Journal of Digital Curation* 11, no. 2 (4 July 2017): 39–47, https://doi.org/10.2218/ijdc.v11i2.421.

receptive to having the indicators of digitised sources in references. If that is the case, the author(s) should explain the importance of differentiating between physical and digital objects.

While digitised copies are distinct intellectual products from analogue materials, one should also be aware of possible discrepancies between digital and analogue versions, e.g. pages forgotten to be digitised or on purpose not digitised, questions of colouring and lighting, all leading to inaccurate or more broadly problematic machine-readable texts – that require critical approaches.[71] To properly differentiate, digitising institutions should provide explicit guidelines about how they want their digitised facsimile to be referenced - to distinguish it from their physical objects.[72]

Independent of the scale of document digitisation, issues arise while trying to distinguish between the physical and the digital object. In most cases, non-persistent identifiers are used, referring to an URL that is tied to the used technology or database system. This causes the risk of providing a link that is dead or – potentially worse – referring in the future to another object. Helle Strandgaard Jensen remarks in her thought-provoking piece about digital archival literacy that historians rarely disclose whether they accessed a physical or digitised version of their sources. She thus makes us aware that:

> '[c]hanging how we cite our digitised sources and what information we ask of others about their uses of digital archives when we peer-review, would be a good place to start a discussion about digital archives and how we can make better use of them in the future.'[73]

While the *idea* of the text might still be the same – clouding the understanding as to *why* a different way of citing is needed – their *form* is definitely not. This could have consequences with regards to research focusing on the materiality, as specific information (e.g. watermarks) can only be seen in the physical version and supported by specific infrastructure and not or only in a suboptimal/skewed way in the digitised version. But also the pro's of a digital version need to be brought forward, enrichment of the data (e.g. in the form of Linked Open Data) can only be provided in a data-fied version and not adequately in the physical object.

The digital turn in the humanities requires thus, that researchers need to be more aware of their data sources and – interestingly – its *materiality* than ever before. A methods' documentation, including digital paths (proper PURL citations), is the reasonable and only future-oriented course of action.[74] While the International Image Interoperability Framework is of immense help to re-use images, the manifests used for this purpose are in itself not enough to provide longevity, since they can be changed anytime and is, as such, not providing the stability academic users seek.[75]

Several GLAM institutions do furthermore offer references to exact locations even within their digitised resources.[76] They do so through page coordinates, which makes the research process highly transparent and easier to check/criticise as this is a feature that is exclusive to digital and digitised resources and could be a way of making historical research multilayered, transparent and accessible to readers.

In the above given example of the Dutch National Archives, it should be noted that the images are linked directly to their catalogue. While this is very useful for the user to see whether

---

[71] Ryan C. Cordell, 'How Not to Teach Digital Humanities', Debates in the Digital Humanities, 18 October 2022, https://dhdebates.gc.cuny.edu/read/untitled/section/31326090-9c70-4c0a-b2b7-74361582977e#ch36.

[72] See for example: https://data.nls.uk/data/digitised-collections/marjory-fleming/

[73] Helle Strandgaard Jensen, 'Digital Archival Literacy for (All) Historians', *Media History* 27, no. 2 (3 April 2021): 260, https://doi.org/10.1080/13688804.2020.1779047.

[74] The use of proper PURLs should then also result in not having to put a date between brackets after the weblink, which is now the case for all non-PURLs.

[75] See e.g.: https://zenodo.org/record/6884885#.Y15rvy-l2X1

[76] E.g. The Dutch National Archive, Journaal van Constantijn Rumpf, 33, https://www.nationaalarchief.nl/onderzoeken/archief/1.11.01.01/invnr/124/file/NL-HaNA_1.11.01.01_124_0033?tab=download [14 October 2022].

pieces have been digitised, the distinction between the physical and the digital is obscured unless the weblink is copied and applied.

It is thus strongly recommended that the entire GLAM sector becomes more aware of its crucial role in providing proper provenance data for digitised objects. While their core business towards physical objects is to store and preserve[77], the preservation of digital derivatives should – in our opinion – follow the same principles: *authenticity*, *reliability*, *integrity* and *usability*.[78] Through persistent identifiers, the GLAM sector could already guarantee *authenticity* and *usability*. At the same time, the reliability factor is partly met – but depends on the *integrity,* which relies on the 'coherent picture'.

For clarity, the International Standard Identifier for Libraries and Related Organisations (ISIL) could – and perhaps should – be integrated with the persistent identifier, adding additional information concerning the responsible institutions.[79] This information could function as an 'authority label', guaranteeing authority and reliability. If that were to be used, the structure of the filenames would be as follows (see figure 14):

Isilcode institute_id collection_id object_secquencenumber of the scan + extension
Figure 14. Suggested filename structure.

Available transcriptions could follow the same structure but with a different extension and possibly be followed by a number indicating a version. Under extreme circumstances, the above could also indicate if volunteers or researchers made a (less perfect) digital facsimile, as opposed to the official digitisation – this could potentially be helpful for GLAM institutions under threat or damaged. If and where possible, such a structure could be used to provide such versionized images within a IIIF manifest.[80]

4. Conclusions and recommendations

We started our contribution with the export and sharing of Ground Truth. However, with sharing comes caring: properly acknowledging who provided the data or models *and* who contributed to the creation. We have discussed and shown the initiative of HTR-United and how one can register available datasets on this platform. This platform functions as a 'umbrella' solution allowing contributors to use decentralised storage of their sources. At HTR-United, creators can be listed, and the metadata can be imported into Zotero for proper referencing.

Furthermore, we discussed the issues that consequently arise: acknowledging which digitised sources have been used – which, at this point – seems dependent on the accurate annotation of an author. Referring to a website, however, is not enough; we have indicated the need for persistent identifiers. Not only to distinguish the digitised collection from the physical objects; but – first and foremost – for the sake of the main characteristics of archival guarantees: authenticity, reliability,

[77] Mike Featherstone, 'Archive', *Theory, Culture & Society* 23, no. 2–3 (1 May 2006): 591–96, https://doi.org/10.1177/0263276406023002106.
[78] 'What Are Archives? | International Council on Archives', accessed 2 October 2022, https://www.ica.org/en/what-archive.
[79] 'ISO 15511:2019', ISO, 17 October 2022, https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/78/77849.html.
[80] Especially the archival specification for IIIF is to be mentioned in this regard. See online: https://archival-iiif.github.io/.

integrity, and (re)usability.[81] For clarity reasons, the structure of the filenames could contain the institutional ISIL codes.

Referencing datasets and HTR models requires an overview of not only underlying sources but also of the contributors and – in the case of HTR models – the quality and the processing of both the training and validation set. As this additional data is of great importance to future users, we propose working with a 'model card' implemented for example in Hugging Face to provide sufficient metadata about and contextualization of a model. To describe the role of contributors – and distinguish the various roles they could have – this contribution has put forward the CRediT (Contributor Roles Taxonomy), which allows to properly reference the work of volunteers/ citizens scientists, if they agree or desire to get mentioned.

Although this is one example of how machine-learning is being rolled out in the library and archival community, the ongoing discussions demonstrate that we are only beginning to understand how best to share data, and recognise contributions to shared datasets, that underpin artificial intelligence systems that are being used within heritage contexts. We hope that this provides an example that can encourage others to consider these aspects, within their own infrastructures.

References

Allen, Liz, Jo Scott, Amy Brand, Marjorie Hlava, and Micah Altman. 'Publishing: Credit Where Credit Is Due'. *Nature* 508, no. 7496 (April 2014): 312–13. https://doi.org/10.1038/508312a.

CRediT. 'Background', 14 April 2020. https://credit.niso.org/background/.

Ball, Alex, and Monica Duke. *How to Cite Datasets and Link to Publications*. A Digital Curation Centre 'Working Level' Guide. DCC How-to Guides. Edinburgh: Digital Curation Centre., 2015. https://doi.org/10.1007/1-4020-5340-1.

Georgian Papers Programme. 'Blog', 17 October 2022. https://georgianpapers.com/about/blog/.

Global Indigenous Data Alliance. 'CARE Principles of Indigenous Data Governance', 17 October 2022. https://www.gida-global.org/care.

Chagué, Alix, and Thibault Clérice. 'HTR-United', 17 October 2022. https://htr-united.github.io/index.html.

———. 'Sharing HTR Datasets with Standardized Metadata: The HTR-United Initiative'. In *Documents Anciens et Reconnaissance Automatique Des Écritures Manuscrites*. Paris, France: CREMMALab, 2022. https://hal.inria.fr/hal-03703989.

Cordell, Ryan C. 'How Not to Teach Digital Humanities'. Debates in the Digital Humanities, 18 October 2022. https://dhdebates.gc.cuny.edu/read/untitled/section/31326090-9c70-4c0a-b2b7-74361582977e#ch36.

CRediT. 'Data Curation', 12 June 2020. https://credit.niso.org/contributor-roles/data-curation/.

Zotero Forums. 'Data Models'. Accessed 20 October 2022. https://forums.zotero.org/discussion/99896/data-models.

Zotero Forums. 'DataSets'. Accessed 20 October 2022. https://forums.zotero.org/discussion/77019/datasets.

Druskat, Stephan. 'Research Software Citation for Researchers'. Research Software Citation, 17 October 2022. https://cite.research-software.org/researchers/.

ECLI:NL:RBDHA:2022:8828, Rechtbank Den Haag, C/09/586380 / HA ZA 20-36, No. ECLI:NL:RBDHA:2022:8828 (Rb. Den Haag 3 August 2022).

LECTAUREP. 'eScriptorium Tutorial (en)', 17 October 2022. https://lectaurep.hypotheses.org/documentation/escriptorium-tutorial-en.

Featherstone, Mike. 'Archive'. *Theory, Culture & Society* 23, no. 2–3 (1 May 2006): 591–96. https://doi.org/10.1177/0263276406023002106.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 'Datasheets for Datasets'. arXiv, 1 December 2021. https://doi.org/10.48550/arXiv.1803.09010.

Gitelman, Lisa, ed. *Raw Data Is an Oxymoron*. MIT Press, 2013. https://doi.org/10.7551/mitpress/9302.001.0001.

'Gpt2 · Hugging Face'. Accessed 20 October 2022. https://huggingface.co/gpt2.

[81] 'What Are Archives? | International Council on Archives'.

Hodel, Tobias, David Schoch, Christa Schneider, and Jake Purcell. ‘General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example’. *Journal of Open Humanities Data* 7, no. 0 (9 July 2021): 13. https://doi.org/10.5334/johd.46.

‘Hugging Face – The AI Community Building the Future.’ Accessed 20 October 2022. https://huggingface.co/.

ISO. ‘ISO 15511:2019’, 17 October 2022. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/78/77849.html.

ivdnt.org. ‘AI-Trainingset – Tag de Tekst voor Named Entity Recognition (NER)’. *INT Taalmaterialen* (blog). Accessed 20 October 2022. https://taalmaterialen.ivdnt.org/download/aitrainingset1-0/.

Jensen, Helle Strandgaard. ‘Digital Archival Literacy for (All) Historians’. *Media History* 27, no. 2 (3 April 2021): 251–65. https://doi.org/10.1080/13688804.2020.1779047.

Keijzer, Carlijn, Milan van Lange, and Annelies van Nispen. ‘First-Hand Accounts of War’. Accessed 20 October 2022. https://www.niod.nl/en/projects/first-hand-accounts-war.

Kestemont, Mike, Folgert Karsdorp, Elisabeth de Bruijn, Matthew Driscoll, Katarzyna A. Kapitan, Pádraig Ó Macháin, Daniel Sawyer, Remco Sleiderink, and Anne Chao. ‘Forgotten Books: The Application of Unseen Species Models to the Survival of Culture’. *Science* 375, no. 6582 (18 February 2022): 765–69. https://doi.org/10.1126/science.abl7655.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. ‘A Survey on Bias and Fairness in Machine Learning’. arXiv, 25 January 2022. https://doi.org/10.48550/arXiv.1908.09635.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. ‘Model Cards for Model Reporting’. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–29. FAT* ’19. New York, NY, USA: Association for Computing Machinery, 2019. https://doi.org/10.1145/3287560.3287596.

‘ORCID’. Accessed 17 October 2022. https://orcid.org/.

CRediT. ‘Resources’, 12 June 2020. https://credit.niso.org/contributor-roles/resources/.

Riley, Jenny, and Devin Becker. ‘Seeing Standards: A Visualization of the Metadata Universe.’, 2010. http://jennriley.com/metadatamap/seeingstandards.pdf.

Risam, Roopika, and Alex Gil. ‘Introduction: The Questions of Minimal Computing’. *Digital Humanities Quarterly* 2, no. 16 (2022).

Rueda, Laura, Martin Fenner, and Patricia Cruse. ‘DataCite: Lessons Learned on Persistent Identifiers for Research Data’. *International Journal of Digital Curation* 11, no. 2 (4 July 2017): 39–47. https://doi.org/10.2218/ijdc.v11i2.421.

Sahle, Patrick. ‘What Is a Scholarly Digital Edition?’ In *Digital Scholarly Editing : Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo, 19–40. Digital Humanities Series. Cambridge: Open Book Publishers, 2017. http://books.openedition.org/obp/3397.

Sicilia, Miguel-Angel, Elena García-Barriocanal, and Salvador Sánchez-Alonso. ‘Community Curation in Open Dataset Repositories: Insights from Zenodo’. *Procedia Computer Science*, 13th International Conference on Current Research Information Systems, CRIS2016, Communicating and Measuring Research Responsibly: Profiling, Metrics, Impact,Interoperability, 106 (1 January 2017): 54–60. https://doi.org/10.1016/j.procs.2017.03.009.

Speer, Robyn. ‘How to Make a Racist AI without Really Trying’. ConceptNet blog, 13 July 2017. http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/.

Stacks. ‘Stacks’, 17 October 2022. https://stacks.wellcomecollection.org/.

Ströbel, Phillip Benjamin, Simon Clematide, Martin Volk, Raphael Schwitter, Tobias Hodel, and David Schoch. ‘Evaluation of HTR Models without Ground Truth Material’. arXiv, 29 April 2022. https://doi.org/10.48550/arXiv.2201.06170.

‘What Are Archives? | International Council on Archives’. Accessed 2 October 2022. https://www.ica.org/en/what-archive.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. ‘The FAIR Guiding Principles for Scientific Data Management and Stewardship’. *Scientific Data* 3, no. 1 (15 March 2016): 160018. https://doi.org/10.1038/sdata.2016.18.

Woolgar, Steve, and Geoff Cooper. ‘Do Artefacts Have Ambivalence? Moses’ Bridges, Winner’s Bridges and Other Urban Legends in S&TS’. *Social Studies of Science* 29, no. 3 (1999): 433–49.

Acknowledgements

We thank the participants of the Transkribus User Conference 2022 and the organisers of this event for the possibility to discuss this topic there.

Funding: CAR was funded by a postdoctoral fellowship from the Dutch Research Council / Nederlandse Organisatie voor Wetenschappelijk Onderzoek [VI.Veni.191H.035];

Author contributions: **Conceptualisation**: C.A.R., T.H.; **Formal analysis**: C.A.R., F.G., A.C., J.v.Z.; **Resources**: C.A.R., T.H., F.G., A.C., J.v.Z., A.S., MT, H.S.J., P.v.d.H., M.v.L, CK, AR, MAA, NA, EB, LVB, AB, D.B., A.Ch., A.N.D., K.v.G., R.G., S.G., S.C.P.J. G., M.J.C. G., S. H., S. v.d. H., M. H., D. H., I. H., A. I., L.K., S. K., L.R. L. T.O.L, A.v.N., J.N., L.M.v. N., JJO, VP, MEP, JJ P., AS, E.S., N.v.d.S., G.V.S., VV, SW, DJW; **Methodology**: CAR TH, FG, A.C., J.v.Z., A.S., MT, H.S.J., P.v.d.H., M.v.L, CK, AR; **Writing – original draft**: AR, F.G., J.v.Z., A.C., M.v.L., C.K.; **Writing – review and editing**: C.A.R., T.H., F.G., A.C., A.S., M.T., A.R., J.N., C.S., A.B., K.D., S.G..

Competing interests: the authors declare no competing interests.