Exploring Data Provenance in Handwritten Text Recognition Infrastructure: Sharing and Reusing Ground Truth Data, Referencing Models, and Acknowledging Contributions

Starting the Conversation on How *We* Could Get It Done[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)*], Jake Purcell [35)], 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)], Anastasiia Anikina [29)], 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)], Heleen Wilbrink[36)], Sonia Weiss [8)], David Joseph Wrisley[33)], Riet Zweistra[34)], and further anonymous citizen scientists/volunteers of the Goetgevonden project![2]

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; 35) American Historical Association; 36) Utrechts Archief.*

---

[1] This article is the result of a writing sprint organised during a workshop at the Transkribus User Conference (TUC) 2022 on the Reuse 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 Digital Humanities minor) taught by Annemieke Romein.
Corresponding author(s):
Dr. C. Annemieke Romein: Annemieke.Romein@Huygens.knaw.nl
Prof. Dr. Tobias Hodel: Tobias.Hodel@unibe.ch

[2] This paper has multiple first-authors, all marked with *-sign; the list of authors/contributors is first based on relative contribution and second alphabetically.

Abstract: This paper discusses *best practices for sharing and reusing 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, subsequently, inform others through HTR-United. Furthermore, we want to we want to suggest appropriate citation methods for HTR data, models, and contributions made by volunteers. Moreover, when using digitised sources (digital facsimiles), it becomes increasingly important to distinguish between the physical object and the digital collection. These topics all relate to the proper acknowledgement of labour put into digitising, transcribing, and sharing Ground Truth HTR data. This also points to broader issues surrounding the use of machine learning in archival and library contexts, and how the community should begin to acknowledge and record both contributions and data provenance.

Within the humanities, working with digital tools is no longer a novelty. However, a difficulty remains regarding how to cite digital resources and, in the case of transcriptions, Handwritten Text Recognition (HTR) models. When we collaboratively began to discuss how to reference these properly, we also started to think about 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 the digitization of sources and the production of transcriptions to create and synthesise models via machine learning. For general models, massive number of documents, accompanied by correct and (ideally) uniform transcriptions, are fundamental. The production of these massive corpora is therefore a challenge that falls in the category of big science.[3] 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 Galleries, Libraries, Archives, and Museum sector (GLAM), we should acknowledge the production of digital facsimiles/digital images of documents.[4]

Unease over adequate citation and acknowledgement in the creation of digital resources – an issue for their reuse – led to the organisation of a hybrid workshop at the Transkribus User Conference 2022. We aimed to discuss: *How can we properly reuse, reference, and acknowledge contributions? What are the best practices thus far*? The activities we channeled 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 reuse it. We then contextualize these strategies within the ethical and legal limitations of the sharing. Because of these limitations, the reuse of Ground Truth requires that contributions *and* contributors be acknowledged, which is discussed in the second section. In our conclusion, we combine

[3] For approaches to big science that depend on (worldwide) collaboration and strive for unified goals, see, for example: Dalmeet Singh Chawla, 'A New 'accelerator' Aims to Bring Big Science to Psychology', Science, 8 November 2017, https://www.science.org/content/article/new-accelerator-aims-bring-big-science-psychology.

[4] In this article, we use the term *digital facsimile* essentially as a translation of the German *Digitalisat*, or, the resulting product of an instance of digitisation. In our case, we are talking about digital reproductions, either photos or scans, of physical objects containing text. In addition, we suggest that, alongside the reproduction itself, researchers should insist on getting information about the digitisation strategies used to create it, in order to determine what is available digitally and what has been left out.

these parts. This article is a preliminary proposal intended to start a discussion about how to conduct and acknowledge the work that goes into generating training data for machine learning.

1. Research Context: Ground Truth

To explain what Ground Truth is, we will show why it is necessary to create data that can be called 'Ground Truth'. Only with Ground Truth is it possible to provide the means of training text recognition models and, furthermore, to judge the quality of said models. As the term 'Ground Truth' suggests, it is a form of data that adheres to specified standards and is considered, at least by a group of people, to be an accurate representation of material, in our case handwritten.[5]

Initial transcriptions may contain quite a few mistakes, but thoroughly checking them – most often by a human – can lead to *perfect* transcriptions. As perfect transcriptions, Ground Truth should be understood as the 'gold standard'. Alternatively, as Mühlberger et al describes it: '[Ground Truth] is a term commonly used in machine learning to refer to accurate, objective information provided by empirical, direct processes, rather than that inferred from sources via the statistical calculation of uncertainty.'[6] As such, it can function as benchmarked data. If we take machine learning systems' hunger for data into account, the value of Ground Truth becomes even clearer. 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 specific scripts or types of handwriting. Also, we would like to note 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. Suppose the Ground Truth is to be shared and potentially bundled into multiple models. In that case, 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 end user select the Ground Truth that is most appropriate for their project and help explain certain model behaviours.

[5] Guenter Muehlberger et al., 'Transforming Scholarship in the Archives through Handwritten Text Recognition', *Journal of Documentation* 75, no. 5 (2019): 957.
[6] Muehlberger et al., 957.

Providing Ground Truth and *metadata*

Generally speaking, there are two mainstream ways of making transcriptions: diplomatic and semi-diplomatic. The former transcribes *as is*, taking the full character set into consideration; the latter allows for changes to improve readability, e.g. writing out abbreviations and simplifying some or all special characters. There are also transcriptions that are hyper-diplomatic, in the sense that ligatures, such as 'st' ligatures, are transcribed, or that types of 's' (e.g. as 'long s') or 'r' are distinguished. All types can be used as benchmarked data, but it should always be clear which choices have been made. Machine learning models will only be able to distinguish characters based on training material.

From a legal perspective and because of the data's value, Ground Truth should be understood as data (by)product of a project and considered for publication (we return 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 may be an obstacle to (re)training Handwritten Text Recognition models, because both Ground Truth and images are needed for training processes. In any case, we should store whatever data we can as well as possible for further and future reuse.

2. Sharing Ground Truth

Much labour and resources are poured into manually and semi-manually producing Ground Truth transcriptions. Reusing transcriptions – and their associated images – promises to support small(er) projects greatly and speed up their work. Furthermore, to advance 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, such as language models for Named Entity Recognition, to name just one example.[7] However, sharing transcriptions, e.g. in a repository, is, in our opinion, not enough and does not fully adhere to the FAIR (Findable, Accessible, Interoperable, Reusable) principles:[8] it should also be (easily) findable by others. Still, sharing data could have legal limitations, and sometimes should also have ethical limitations.[9] These topics are discussed in 2.2 of this article. It should be stressed that we are explicitly talking about sharing Ground Truth data and not about sharing HTR models in this section. The latter is tool-dependent and can only occasionally be used outside or independently of a specific tool or framework; this topic will be addressed in this article's third section.

[7] 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.

[8] 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.

[9] See section 2.5.

2.1 How to Export Data

The various programs that allow for the creation of Optical Character Recognition (OCR)/HTR have options to export the generated and/or corrected transcriptions. When possible, both the transcriptions and images should be exported, depending on any potential copyright/image rights.[10] If this is not possible, it is still helpful *to at minimum* sustainably store the transcriptions.

Within the *Transkribus* tool, provided by the READ-COOP SCE[11], the export appears as shown in figure 1 (below). The ALTO XML and the PAGE XML allow for an alignment between image and transcription – based on coordinates – which is required to connect transcribed text on a word or line basis with images and allows for the opportunity to (re)train machine learning based models.[12] These two formats are also supported by the eScriptorium application (see figure 2 below), which has been developed in the context of a variety of national (France) and European projects.[13]

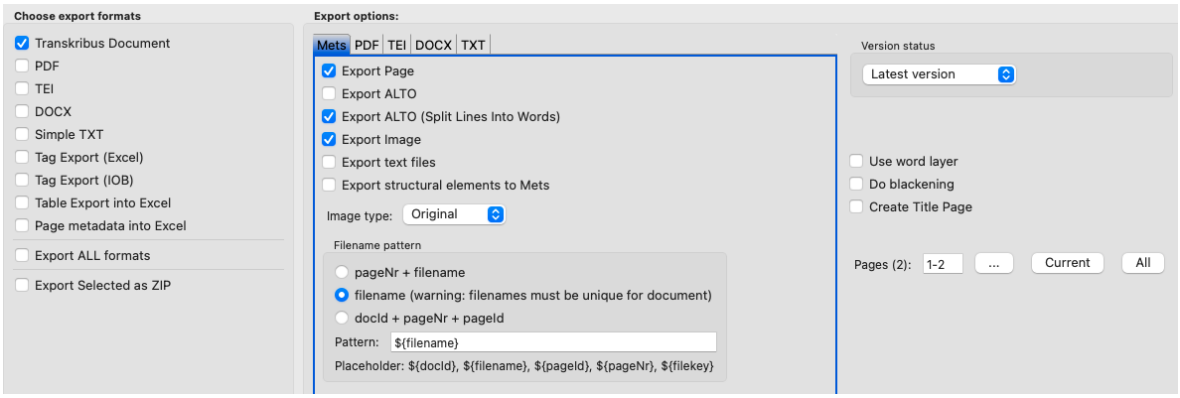


Figure 1. Screenshot Transkribus Export [version 1.22.0.1-SNAPSHOT]. [30 September 2022]

[10] 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 can be included in provided XML files.

[11] https://readcoop.eu/, accessed 12 October 2022.

[12] In the Transkribus environment, depending on the number of documents and pages, this might take a while, and, when server export is chosen, one will receive an email with a link to download the files when they are available.

[13] For more information see Benjamin Kiessling et al., 'EScriptorium: An Open Source Platform for Historical Document Analysis', in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 2, 2019, 19–19, https://doi.org/10.1109/ICDARW.2019.10032.

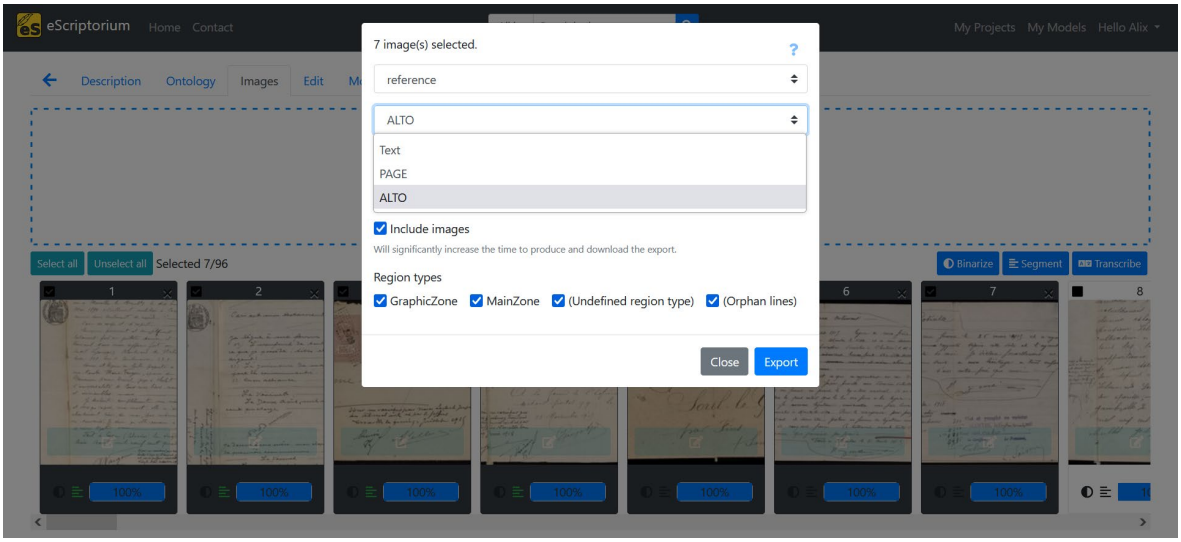


Figure 2. Screenshot eScriptorium Export [Version 0.12.5b]. [14 October, 2022]

ALTO and PAGE are the main formats used to store HTR output. At the same time, TEI, which is better known in the Digital Humanities community, is primarily dedicated to advancing critical digital scholarly editions. Although it is hard to predict future developments, we are optimistic that at least a future conversion to what is then the standard format, 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 keeping the option to add custom properties.

While some would call for a centralised Ground Truth repository, this could be a costly affair (who could or should declare itself responsible for the longevity of such an environment?), and furthermore result in double the work, as funding agencies could have requirements to store output in 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, depending on the data. Generally, storing data in a FAIR-compliant, noncommercial repository with a persistent identifier, like Digital Object Identifiers (DOIs), is preferred. At the same time, it is highly encouraged that data output be made accessible in a clearly structured format. Images and XML files should reside in subfolders, with descriptive names for folders as well as images.

Repositories, such as Zenodo, offer the possibility of adding structured metadata that includes the name of contributors, licenses for reuse and (if applicable) URLs to external webpages. Furthermore, adding more detailed information in a README file is good practice and helps to navigate the data dump in the case of reuse.[14] 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 both make use of user friendly git environments and receive a DOI, a mixed solution is a possible way forward: version management can be done through GitHub, while Zenodo stores versioned and (in)frequently updated datasets. Conveniently, some platforms like GitHub allow a repository to be linked 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 different versions of transcriptions and documents to become available online, based upon different parameters or (underlying) HTR models. Here, the question of which types of transcriptions we are talking about comes up again: manual Ground Truth or automatic transcriptions. Whichever version one posts, it should be clear to other potential users what the type and status of the transcriptions are.

At the same time, one would like to have information about the rules guiding and characteristics of the transcription of documents underlying a particular Ground Truth, this information would allows potential users to search for data sets for transcriptions that fit the criteria that they are interested in.[15] 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) could be published (see also the discussion above, section 1).

[14] 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.

[15] 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.

Using a repository to share data is obviously both useful and good academic practice. However, to make the data not only available but also findable – as is required by FAIR principles – at least a link to a sharing platform like HTR-United (see below) should be considered.[16]

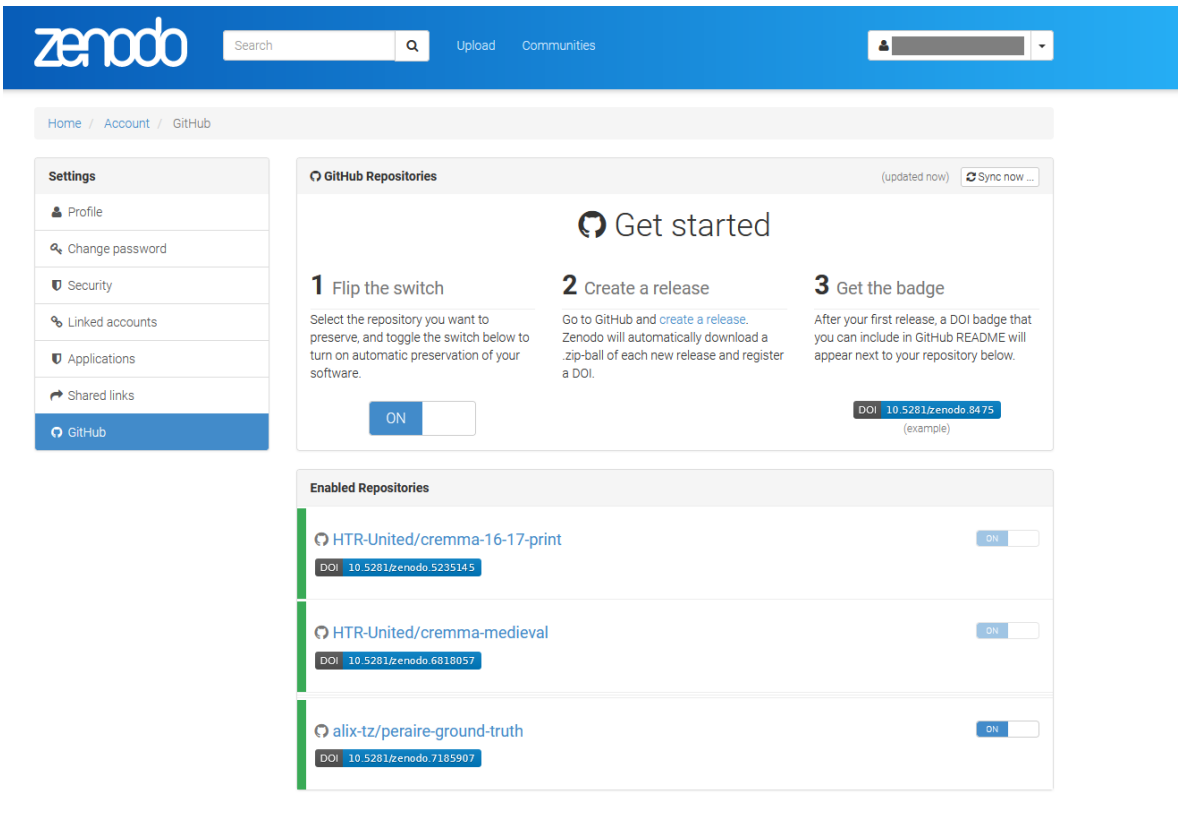


Figure 3. GitHub synchronisation and DOI generation on Zenodo. [17 October 2022]

2.3 HTR-United: Sharing Your Data

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

Supporting users and (small) projects, Alix Chagué and Thibault Clérice developed the HTR-United initiative (see figure 4).[17] HTR-United consists of three imperatives: 'a collaborative enterprise

[16] 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, one must contact the IMPACT centres.

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

for the community; friendly to consumers and data producers; as low tech as possible (because $$).'[18] 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.'[19]

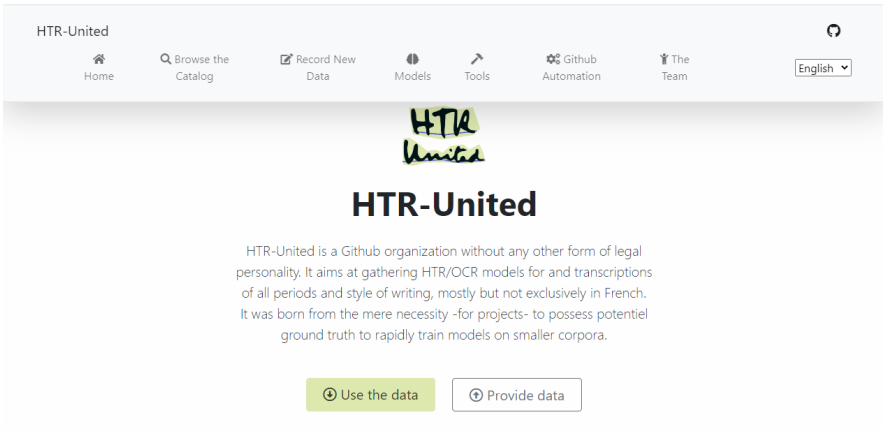


Figure 4. Website of HTR-United. https://htr-united.github.io/index.html [30 September 2022]

This much-needed initiative offers a solution that is easy to use and access, allowing contributors to store their dataset at any given location (preferably with a DOI). It also centralises an overview of those stored Ground Truth datasets. The HTR-United interface allows users to filter Ground Truth by language, script/type, and periodisation. Furthermore, the catalogue contains metadata (.yml), updated through *Continuous Integration* through GitHub Actions. Chagué and Clérice developed a form that simplifies the process of creating .yml and badges and uploading metadata in the catalogue.[20] The developers (and at the same time, initiators) know that the schema with questions gains complexity while trying to keep the overview of Ground Truth datasets simple. However, they think it is worth the effort as it provides a uniform overview of the digital environment.

[18] 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), https://hal.inria.fr/hal-03703989.

[19] Roopika Risam and Alex Gil, 'Introduction: The Questions of Minimal Computing', *Digital Humanities Quarterly* 16, no. 2 (2022): sec. 3.

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

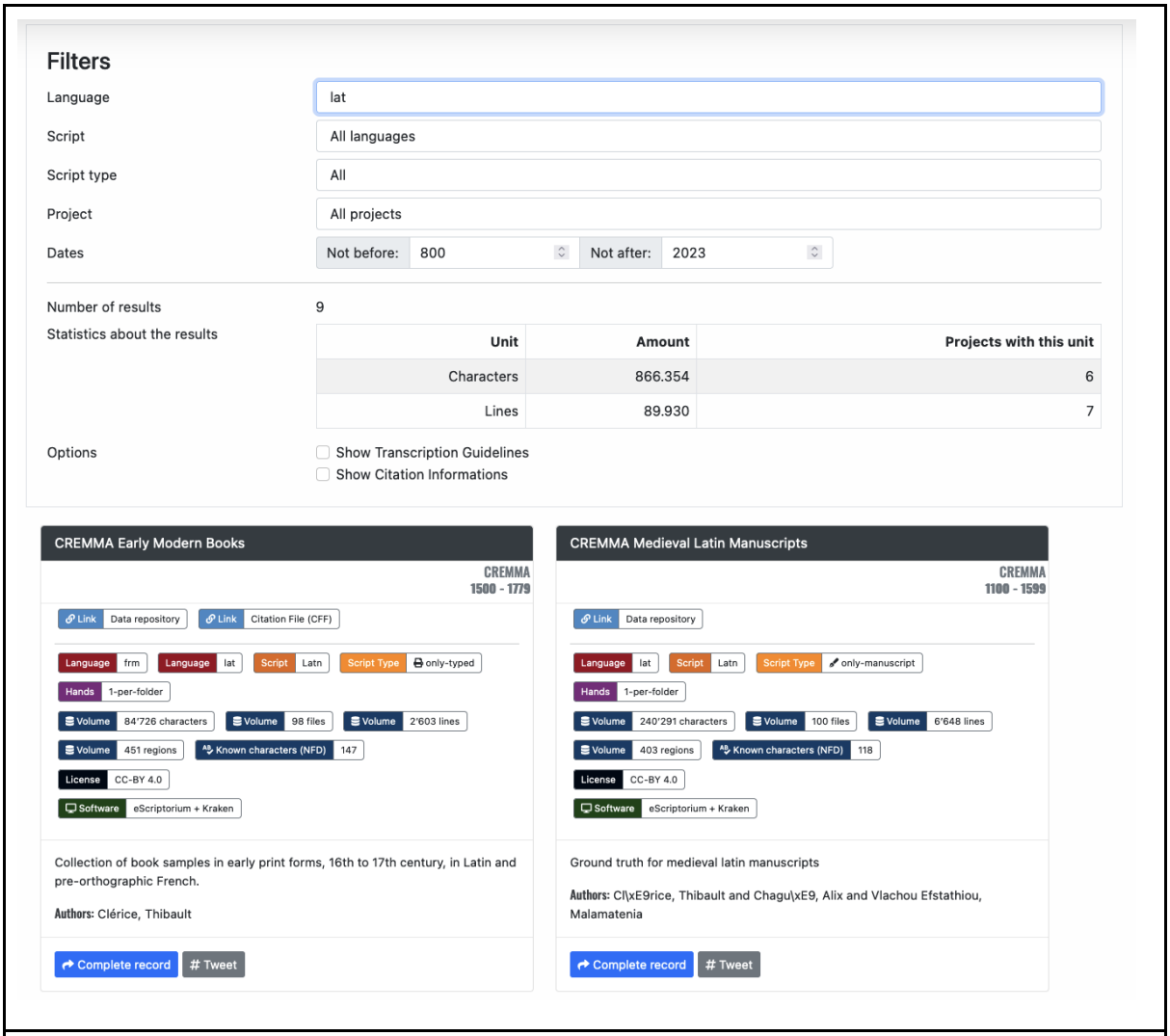
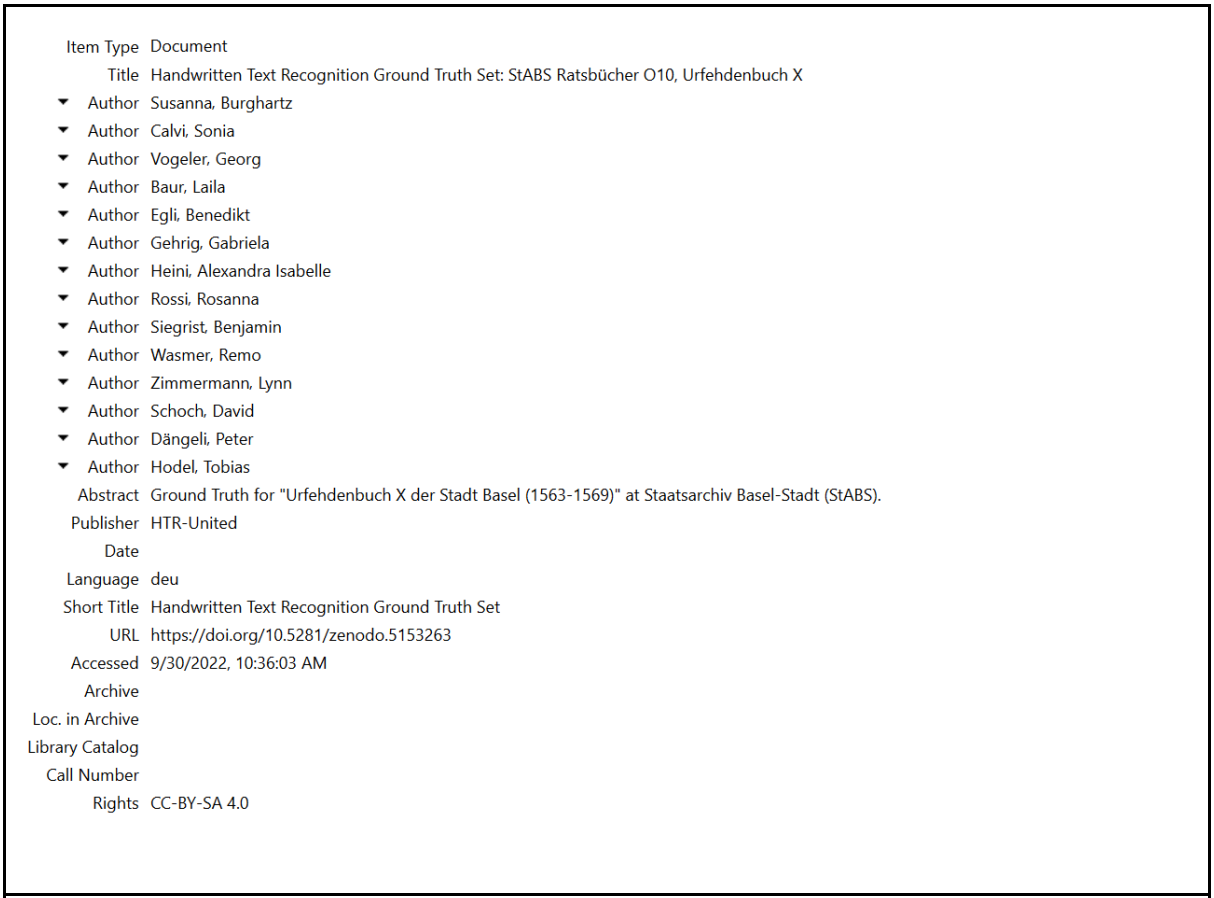


Figure 5. Catalogue of HTR-United. https://htr-united.github.io/catalog.html [30 September 2022]

HTR-United limits itself, and so its collaborators, to a predetermined way of sharing Ground Truth, and does so for practical reasons. Providing a relatively strict schema for the catalogue allows for a machine actionable method of checking the conformity of the submissions, and also supports searches across the catalogue.[21]

From the catalogue on the HTR-United webiste (see figure 5), it is possible to download the metadata into Zotero as an 'Item Type: (Digital) Document'. This download option simplifies the future referencing process (see figure 6).

[21] Chagué and Clérice, sec. 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 Ground Truth Set. The example is of particular interest because it results from a multi-stage process. The transcription was done within one project by several student assistants, under the direction of a Digital Humanities 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 to run a text-to-image process that resulted in an annotated dataset suitable for training an HTR model. This further processing was only possible because of the initial publication of the open TEI XML data set.
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 focused on sharing Ground Truth, or texts that have been corrected manually. However, when models perform well, we may reach a point where sharing large datasets of raw HTR-produced transcriptions would also prove helpful, even though they are not perfect due to errors. These, too, might be useful to share; however, they should be explicitly designated as machine generated transcriptions, in which case it is necessary to note the calculated or assumed Character Error Rate (CER).[22] 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 also to indicate the model used.

2.4 Referencing Digitised Resources and Digital Output

Several software solutions exist for creating, collecting, editing, and reusing bibliographic references for annotation purposes. These include, to name only a few, EndNote, Citavi, Zotero, and Mendeley.[23] Zotero is a free, open-source referencing tool provided by the Corporation for Digital Scholarship that can adapt to various referencing styles.[24] As it is a free and open-source tool that has been programmed by and for humanities scholars,[25] we use Zotero as a point of reference for suggesting how to reference and acknowledge digital (re)sources and contributors.

Sharing data and models are of great importance to the academic world, as well as to data providers (including us, the contributors to this article, since there are some challenges in correctly referencing our work output). There is a risk of neglect, because there is no consensus on standards for referencing either '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).

For certain objects in the humanities, such as physically published books, it is quite clear what questions need to be answered in a citation. It needs to state who wrote the text, who contributed, and what was the source. An exact structure needs to be followed, which will depend on the citation style. 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 referenced for centuries and even millennia, approaches to dealing with digital (ephemeral) objects are

[22] 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'. See also Ryan Cordell, '"Q i-Jtb the Raven": Taking Dirty OCR Seriously', *Book History* 20, no. 1 (2017): 188–225, https://doi.org/10.1353/bh.2017.0006; Ryan C. Cordell, 'Talking about Viral Texts Failures', 25 June 2020, https://ryancordell.org/research/VT-database-fail/.

[23] https://www.mendeley.com/ [22 Sept. 2022]; https://endnote.com/product-details/compatibility [22 Sept. 2022]; Ann Chen, 'Library Guides: Mendeley: Home', 13 October 2022, https://aut.ac.nz.libguides.com/c.php?g=359376&p=2427744.

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

[25] Sean Takats, 'Facing Abundance: Zotero as an Enlightenment Tool' (American Society for Eighteenth-Century Studies, Albuquerque, New Mexico, March 2010), https://orbilu.uni.lu/handle/10993/50339.

in their infancy.[26] We have thus combined experiences, suggestions, and guidelines in this section. As above, we maintain a focus on FAIR and also CARE (CARE = Collective Benefit, Authority to Control, Responsibility, and Ethics) principles (see below 2.5), while striving to use persistent identifiers. The principal point of view will be on when digital resources should be cited, what elements should be included in the citation, what aspects of a digital resource should be acknowledged.

2.4.1 Referencing Datasets

Data models are only starting to be cited at this point in time in research[27], which results in a lack of standards within the humanities.[28] However, in the fields of computer sciences and machine learning, guidelines on how to cite datasets and software do exist.[29]

Let us be clear about what we mean, in the context of this article, when we consider datasets that could need to be cited. In the first place, there are transcriptions, which include information about where on an image-page a specific word or line can be found. These transcriptions encompass manually created Ground Truth and machine-generated transcriptions, as well as anything in between, such as machine-generated but manually corrected Ground Truth. Second, in addition transcriptions, there is text enrichment or, more generally, semantic annotation, e.g. georeferencing place names, named entity recognition, and linking terms to authority data. While these activities may be integrated within one dataset, what has been done and/or used and by whom should be clearly stated in all circumstances.

Standard literature management software is only beginning to incorporate citation of datasets and software. Zotero, for example, is, as of 22 September 2022, not supporting output types 'datasets' or 'data/HTR models', though they state that the category 'datasets' will soon be added.[30] According to the Zotero forum, the following elements/metadata will be added to a dataset: *author(s); dataset title; publication date; version; data repository/publisher; DOI; URL; license/rights;* and *resource/medium*. In this way, the ability to cite these kinds of scholarly and scientific contributions will be even easier. We can only hope that, in future releases of large data sets, such contributions will be accordingly acknowledged.[31]

---

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

[27] Above, we have shown that HTR-United currently uses 'Item Type: Document'.

[28] 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.

[29] Timnit Gebru et al., 'Datasheets for Datasets' (arXiv, 1 December 2021), https://doi.org/10.48550/arXiv.1803.09010.

[30] 'DataSets', Zotero Forums, accessed 20 October 2022, https://forums.zotero.org/discussion/77019/datasets.

[31] 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.

Reflections exports and clarifying documentation

One of the questions explored by the HTR-United initiative is how we can better document HTR Ground Truth datasets. In addition to 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, while others omit diacritical marks; modify punctuation; bring superscripts down the line; or ignore bold, italics, small caps, 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 character set and its completeness should be addressed, too; for example, missing numbers could cause issues when reusing 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 reuse.

It is challenging to imagine properly combining or reusing different datasets without such information. Of course, it is difficult to extend such considerations to existing datasets, but it is crucial as soon as the dataset contains or is made of annotations. If it were possible to provide a set of parameters that need to be documented before allowing the export of data from the tool in which the Ground Truth was produced, it could raise awareness of the importance of documenting choices.[32] Then the next step would be to have an integrated export option to a FAIR-abiding repository, such as Zenodo. In particular, technical details (e.g. not only name and base model(s) used, but also the origin of the training and validation set, and the 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 the base model.

Guidelines for metadata

Data[33], models[34], and even concrete objects[35] 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

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

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

[34] 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/.

[35] 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.

computer science, machine learning, Natural Language Processing, and other data-intensive fields.[36] 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 allow for 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.'[37] Thus, responsible metadata does not just encompass the application of FAIR principles[38] and the existence of sufficient provenance information. Responsible metadata also details why the data was gathered, and 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')[39] 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 suspect and with the greatest of reservations, or at least tested in depth.

Unfortunately, because of the incredible variety in format 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 prolificacy of metadata standards, of which at least three hundred exist.[40] However, the salient point is not that a particular data standard should be primary, but that a trustworthy data source will clearly state 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. 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'.[41] 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.[42] Metadata fields

[36] 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.

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

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

[39] With regards to 'cleaning' see also: Katie Rawson and Trevor Muñoz, 'Against Cleaning', in *Debates in the Digital Humanities 2019*, ed. Matthew K. Gold and Lauren F. Klein (University of Minnesota Press, 2019), 279–92, https://doi.org/10.5749/j.ctvg251hk.26.

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

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

[42] 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.

that should be part of any data set citation, if known, include 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 HTR models, as is suggested on the Zotero forum.[43] This 'Item Type' 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 fitting approaches. Let us first make an inventory of elements useful for citing 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 option of having a DOI for their HTR model. One appreciated possibility would be 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 (through upload to Zenodo). Another possible desired integration would be with ORCID, to be unambiguous about the creator of the HTR model.[44] In order to even further complicate the issue, we would also advise mentioning the programmer of the training and evaluation algorithms (the text recognition engines) in one way or another.

An added layer that keeps coming up 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 set and the validation set, as well as their respective sizes, are of use here, as they tell other potential users about the quality of the HTR model and tendencies to overfit.[45] 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 (called 'fine-tuning' in machine learning terms). The existing model is then used as a 'base model'. Base models can also be stacked while creating the ideal model. By principle, the entire stack of base models preceding any 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), retraining of the whole dataset after three cycles (base model + training as a base model + training as a

[43] 'Data Models', Zotero Forums, accessed 20 October 2022, https://forums.zotero.org/discussion/99896/data-models.

[44] 'ORCID', accessed 17 October 2022, https://orcid.org/.

[45] See with regards to interpreting HTR models: Tobias Hodel, '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. Konferenzabstracts*, ed. Christof Schöch (Paderborn: Christof Schöch, 2020), https://doi.org/10.5281/zenodo.3666690.

base model) would be advisable.[46] When referencing earlier base models, proper descriptions with attributions to the creator(s) are also very important.

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 currently be produced in any scientific domain, and it is, in any case, insufficient to cover HTR models. Congruent with what has been said about metadata and datasets, we need a quite granular schema for describing HTR models. Mitchell et al. propose a 'model card' to inscribe sufficient metadata and context about a model.[47] Such model cards have been implemented in the Hugging Face[48] repository, the current go-to repository for publishing data sets, models, and documentation for NLP 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.[49] 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 inform users of bias and editorial decisions. This would also allow communities to strive for a better understanding of 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 standards.[50] Consequently, the metadata fields to include for HTR model citing are congruent with those to use for dataset citation: author, publication date, title, version, resource type, publisher, identifier, and location. Right now, data is put on the Web without any of this information being present, which makes it hard to know.

[46] Here we refrain from making remarks about the Pylaia engine or Kraken.

[47] 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.

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

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

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

2.5 Ethics and Limitations of Sharing

Those sharing data must be aware of ethical implications of doing so 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 more 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, maintaining a business model for is essential for survival, regardless of its governance mode, such as a ltd. company or a cooperative. As one such example, READ-COOP SCE has as a philosophy: 'share as much as possible, and retain as much control as necessary'. Because money flows to upholding the infrastructure and the servers, that is the part that needs to be controlled, but other things can be shared, for example, the Ground Truth or transcriptions that are produced. Sharing as much as possible an 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 economies of scale, improving the AI's ability to correctly recognise texts, which is a shared, 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 be challenging for many users. As such, sharing and exporting Ground Truth and machine-generated output data is allowed; however, sharing HTR models is only allowed *within* the platform. These HTR-models cannot be exported outside the READ-COOP as it invested a lot of computing powerfree of cost for users. Other tools, such as eScriptorium, allow the sharing of models but require the user to set up their own infrastructure and maintain it.[51]

From an ethical rather than legal point of view, it is crucial to think about creators, curators, and descendants of the material in question - which is the focus of the third section of this article. Especially when working with historical materials originating from colonial contexts, one must take the biography of a document into consideration and describe how it became part of an institution, as well as consider what implications there might be of making documents or sources publicly available data.[52] There are also other considerations from non-Western communities that may have very different models and understanding of ownership and what it means to respect the content of historical documents. Thus, consequences of working with and sharing data must be kept in mind. For this reason, in addition to

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

[52] See as an example: Alexandra Ortolja-Baird and Julianne Nyhan, '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 1', *Digital Scholarship in the Humanities* 37, no. 3 (1 September 2022): 844–67, https://doi.org/10.1093/llc/fqab065.

FAIR principles, CARE principles need to be considered, since they cover a multitude of aspects and have been proposed by the Global Indigenous Data Alliance. CARE does not have the same standing as FAIR for the moment, but it brings ethics into the discussion as a key aspect, it asks for the collective benefit of data production and sharing, and it demands that communities keep the authority to control "their" data, while all players act in a responsible manner.[53]

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, publishing this would be a violation of the the 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.[54]

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. [55] 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 (either directly or indirectly);
- *Pseudonymisation*: by editing personal data in such a way that they cannot be traced back to the actual person (either directly or indirectly) without using additional data. These preconditions could be a 'key' that only authorised individuals have access to, which should be stored separately;
- *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;

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

[54] 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.

[55] 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

- *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 procedure weighing 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 sharing Ground Truth transcriptions or machine-generated transcriptions with a larger public. The strategies mentioned above show how customised approaches 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 – all the hard work that was put into the creation of Ground Truth and the human desire to be open about it (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 neglect. 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 issues that come up when 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 essential because of their hard work and to provide insight into *how* and *with what resources* Ground Truth data was produced. Properly citing 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/

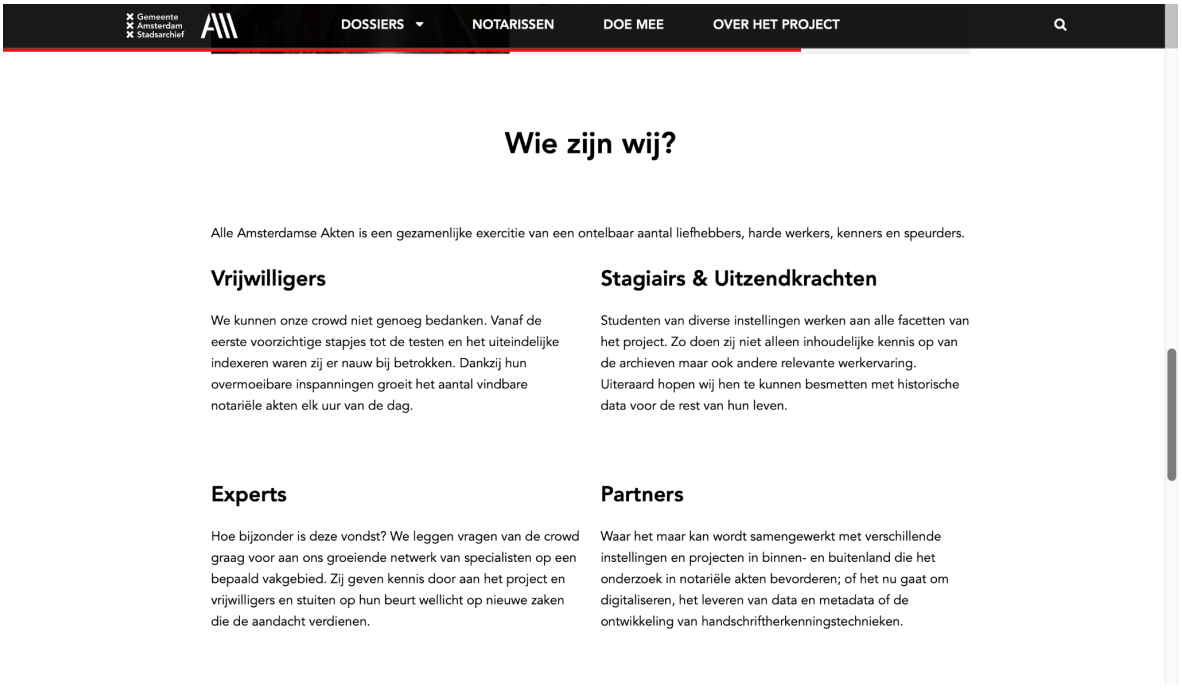


Figure 8. Part of *Alle Amsterdamse Akten* (All Amsterdam Notarial Deeds) Website, acknowledging the volunteers (crowd) as one group.

Acknowledging the crowd: current situation and room for improvement

Using the existing landscape 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).[56] Sometimes they do so for practical reasons, at other times they do so to intentionally emphasise the collective effort instead of the individual. Second, there

[56] 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/.

are also projects, especially smaller ones, that acknowledge their volunteers by listing them with their full credentials in recognition of their work (see figures 9 and 10). In our view, and in line with the previous sections of this article, these acknowledgements should be incorporated into the publication of the actual resulting datasets, too. How should that be done?

It is understandable that, due to administrative labour, larger projects in particular tend to acknowledge their volunteers in a more generalised manner, but there are also arguments in favour of listing members of the crowd as individuals in the case of Ground Truth publication. We want 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, acknowledgement by name 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 teaches 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 individuals who 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, once could consider only naming the volunteers who have exceeded a specific threshold of work. A personalised recognition could also provide the space to list the people who delivered most of the transcriptions first, whereas those who made smaller contributions are placed last on the list. Alternatively, 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 insight is also provided into the quality of individual transcribers' contributions.

Some caution is required. 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 uncertain, it is advisable to list the names alphabetically.

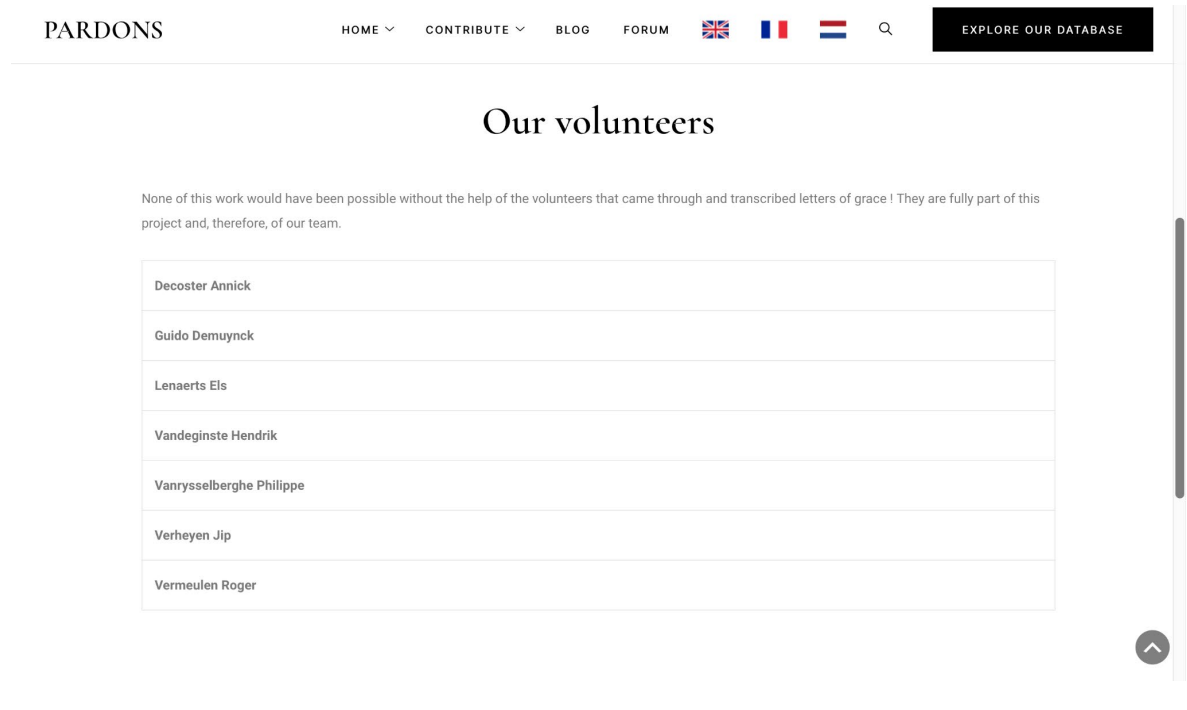
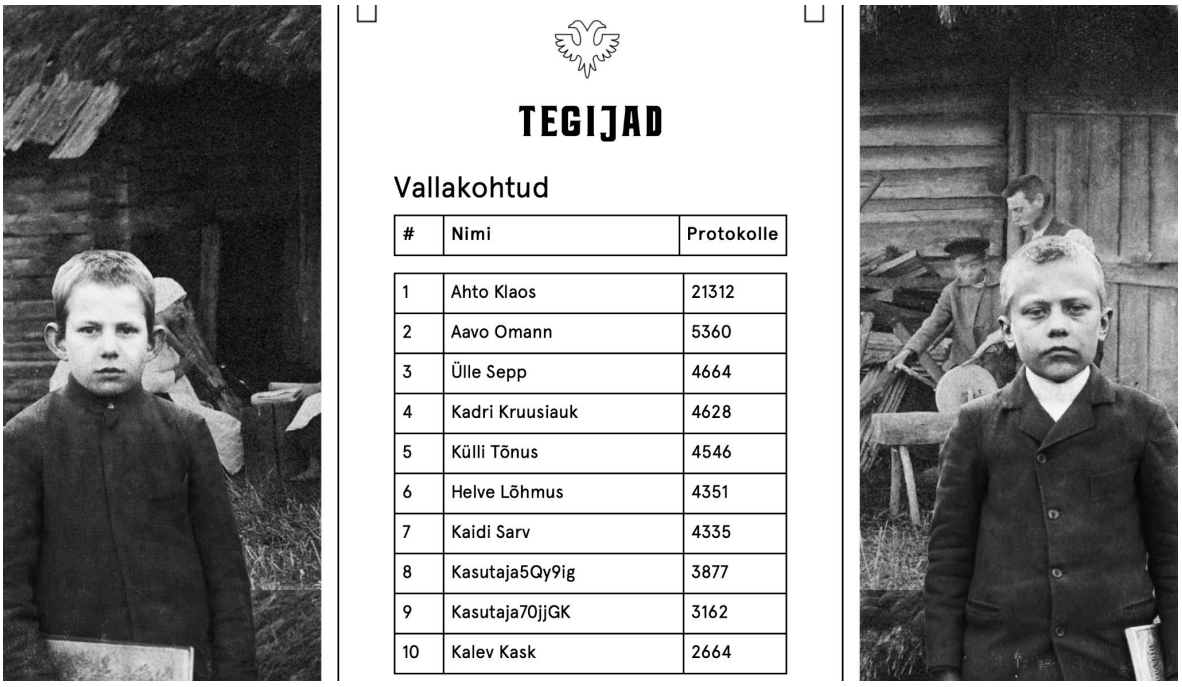


Figure 9. Part of the project website of Pardons; here, the names of all volunteers are listed (https://pardons.eu/the-team/).



TEGIJAD

Vallakohtud

#	Nimi	Protokolle
1	Ahto Klaos	21312
2	Aavo Omann	5360
3	Ülle Sepp	4664
4	Kadri Kruusiauk	4628
5	Külli Tõnus	4546
6	Helve Lõhmus	4351
7	Kaidi Sarv	4335
8	Kasutaja5Qy9ig	3877
9	Kasutaja70jjGK	3162
10	Kalev Kask	2664

Figure 10. Volunteers listed on the website of the National archives of Estonia, including the number of 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, those people should be informed, and consent for using their names needs to be sought. Additionally, the option should remain that people or their heirs can withdraw their names in the future.

Future complications could be avoided 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 state 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 option should remain for volunteers and their heirs to withdraw their name at a later point in time. Information about how they can do so should be available. In cases when someone requests *withdrawal* of their name from use, the name can no longer be used for future publications. However, the GDPR also allows for a request for *data erasure*. In these cases, the name should, if reasonably possible, also be removed from past publications. When doing so, it 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.[57] 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, for example, the field of history, single-authored articles and monographs are still the norm, working with citizen scientists/volunteers may require changes to this approach, as with publishing and citing datasets used or HTR models.

As such, we would suggest the CRediT taxonomy.[58] Some journals, such as *Science,* already work with this model and add an acknowledgement section to their article.[59] 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.'[60]

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

[57] Although this section predominantly deals 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 teaches that volunteers' primary motivation for getting involved in projects is that 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 post, or to attend a (online) meeting during which results are presented. This also creates networking opportunities from which, participating students and interns could benefit in particular.

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

[59] 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.

[60] 'Background', *CRediT* (blog), 14 April 2020, https://credit.niso.org/background/.

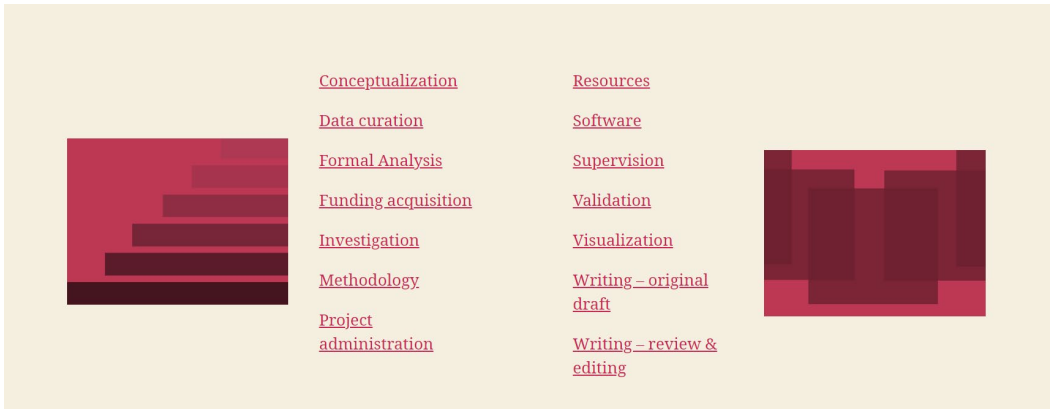


Figure 11. CRediT taxonomy. https://credit.niso.org/

While this overview might look complicated, work on Ground Truth, datasets, or databases generally fits within the frame of *data curation* or *resources*.[61] Being explicit about a person's role will not only help avoid confusion about their contribution, but also demonstrate the different kinds of contribution. When citizen scientists/volunteers are provided with a specific task (e.g. transcribing, correcting, or tagging texts), it could immediately be connected to one of the CRediT roles or tasks like *data curation* or *resources*. Regardless of their initial role, if the citizen scientists come across an exciting find that leads to specific research, an additional role could be assigned in consultation with the individual.

From a legal perspective, one's role relates to one's potential author's rights within a given jurisdiction. When information is processed and converted to a machine-readable format, as in the previously described cases involving transcription, it is implied that no original work is created, and therefore the processed information is not covered by the author's rights.[62] However, courtesy could and should require a proper acknowledgement of work put into creating files.

3.2 Acknowledging Institutional Activities: Digitisation Activity and Contextualisation

GLAM sector institutions, but of course also private institutions, digitise 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,

[61] '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/.
[62] 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).

digitisation is a costly, time-consuming practice that is, by now, part of their core business.[63] It takes time, and this steadily paced process is not something that is often communicated 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 clarify 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 from archives; for them, it covers questions such as those 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? For 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 conceptualisation of the digital humanities.[64]

[63] Although not explicitly covered in this article, it is polite to acknowledge institutions and funders. Their efforts and/or financial support allowed for the creation of Ground Truth. If the citation concerns previously published texts (scholarly editions), institutions/funders contribution toward state-of-the-art research in AI space is often considered rewarding.

[64] See: Rik Hoekstra and Marijn Koolen, 'Data Scopes for Digital History Research', *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 52, no. 2 (3 April 2019): 79–94, https://doi.org/10.1080/01615440.2018.1484676; Claudia Engelhardt et al., *D7.4 How to Be FAIR with Your Data. A Teaching and Training Handbook for Higher Education Institutions*, version V1.2 DRAFT (Zenodo, 2022), https://doi.org/10.5281/zenodo.5905866; Tessa Hauswedell 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', *Archival Science* 20, no. 2 (1 June 2020): 139–65, https://doi.org/10.1007/s10502-020-09332-1.

The questions above are essential for researchers to perform 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 new digital world requires technical skills and resources to set up an infrastructure that integrates characteristics archives are intended to guarantee: authenticity, reliability, integrity, and usability.[65] Here, a lack of a clear and distinctive overview of competing standards – handles, (P)URLs, DOIs, URIs – can cloud the understanding, which can lead to mere digitisation without guarantees of authenticity and reliability.

This also brings us to the question of *what* has been digitised by a particular institution 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 that went into choosing items for digitisation should be made clear to users.[66] Jensen suggests that digital archives could be encouraged to demonstrate the extent and content of their digitisation efforts.[67] Here, she implicitly refers to the reliability of the found digitised document – how much of the inventory has been digitised (as a percentage; 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 opportunity to be 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.[68] Though archivists are rarely credited for this work as individuals, the question is whether 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 in an increasingly precarious labour market a way to highlight their skills and experience (and be cited for it), make digital labour more visible, and let people who use the resources know who to contact if they have any questions related to the resources.

[65] 'What Are Archives? | International Council on Archives', accessed 2 October 2022, https://www.ica.org/en/what-archive.
[66] Hauswedell et al., 'Of Global Reach yet of Situated Contexts'.
[67] Helle Strandgaard Jensen, 'Digital Archival Literacy for (All) Historians', *Media History* 27, no. 2 (3 April 2021): 256, https://doi.org/10.1080/13688804.2020.1779047.
[68] 'Blog', Georgian Papers Programme, accessed 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.

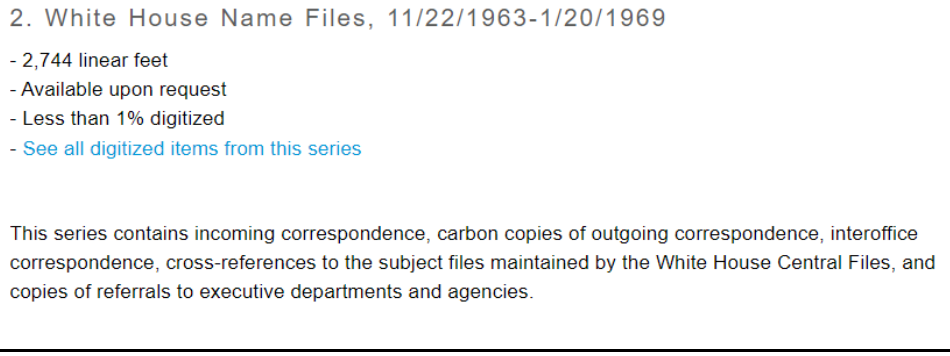


Figure 13. Section from the White House Central Files (WHCF) created or collected by President Lyndon B. Johnson and his staff, with an indication how much of the collection has been digitised. https://www.discoverlbj.org/exhibits/show/loh/pres/whcf

Combining the additional data with descriptions based on predefined categories and structures could allow for different search methods and so extend users' freedom. It would create multiple entries that allow for differences and similarities between conceptual models found in the archive and researchers' (changing) conceptual models.[69] Such room to manoeuvre is an asset to open and different interpretations without the apparent influence of the creators of such conceptual models. According to Jensen, this would/could result in different searches, including one targeting a range of related topics or production contexts.[70] 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.'[71]

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 depends on (additional) external funding, which means that they are likely to be subject to policies that emphasise popularity, marketisation, or current research trends.'[72] 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,

[69] Jensen, 'Digital Archival Literacy for (All) Historians', 257.
[70] Ibidem, 258.
[71] Ibidem, 259.
[72] Ibidem, 254.

and the ones paying for them will necessarily have an influence on the archive's contents. One could spin this thought out further and ask when the intentional omission of information starts (and where it will end).

On the other hand, it has been argued that the digitisation of archives reduces selection bias. Based on experience from small- and large-scale digitisation projects and from the literature, we cannot agree with that stance, noting in particular political and infrastructural decisions.[73] 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 also the availability of equipment and (financial) means to carry out such work and make it accessible, which favours the global north.

Whether digitisation really leads to increased information transparency is thus still up for discussion. For researchers with broad knowledge about an institution's collections, we nonetheless assume that educated conclusions about selection bias can be derived. Furthermore, based on the existence of certain materials online, it can also lead to more interest in certain documents or objects among the general public. 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, it's possible to select 'Item Type: Document and Books'. This includes fields to enter archives, location in the archive, and URLs, preferably a permanent URL (PURL).[74] Their availability certainly opens up opportunities; however, the author(s) should be aware that not every publisher is receptive to having indications of digitised sources in references. In these cases, 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 accidentially or intentionally not digitised, and questions of colouring and lighting, all leading to inaccurate or more broadly problematic machine-readable texts, ones that require critical approaches.[75] To properly differentiate between digital facsimiles and their physical objects, digitising institutions should provide explicit guidelines for how they want their digitised facsimiles to be referenced.[76]

---

[73] See for example for newspaper digitisation: Hauswedell et al., 'Of Global Reach yet of Situated Contexts'. And more broadly: Gerben Zaagsma, 'Digital History and the Politics of Digitization', *Digital Scholarship in the Humanities*, 16 September 2022, fqac050, https://doi.org/10.1093/llc/fqac050.

[74] 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.

[75] 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.

[76] See for example: 'Diary, Letters and Poems of Marjory Fleming – Data Foundry', accessed 31 October 2022, https://data.nls.uk/data/digitised-collections/marjory-fleming/.

Independent of the scale of document digitisation, issues arise when indicating differences between the physical and the digital object. In most cases, non-persistent identifiers are used, referring to an URL that is tied to the technology used or the database system. This causes the risk of providing a link that is dead or, potentially worse, refers 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.'[77]

While the *idea* of the text might still be the same, clouding the understanding as to *why* a different way of citing is needed, the *form* is definitely not. This could have consequences for research focusing on materiality, as specific information (e.g. watermarks) can only be seen in the physical version and supported by specific infrastructure, and cannot be seen at all or only seen in a suboptimal/skewed way in the digitised version. Nevertheless, the pros of a digital version need to be brought forward, and 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 be more aware of their data's source and, interestingly, its *materiality* than ever before. A methods' documentation, including digital paths (proper PURL citations), is the reasonable course of action, and the only future-oriented one.[78] While the International Image Interoperability Framework is of immense help for reusing images, the manifests used for this purpose are in themselves not enough to provide longevity, since they can be changed at any time, and so do not provide the stability academic users seek.[79]

Furthermore, several GLAM institutions even offer references to the exact locations of words within their digitised resources.[80] They do so through page coordinates, which make the research process highly transparent and easier to verify/critique, 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 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, who can see whether pieces

---

[77] Jensen, 'Digital Archival Literacy for (All) Historians', 260.

[78] 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.

[79] See e.g.: Joseph Padfield et al., 'Practical Applications of IIIF as a Building Block towards a Digital National Collection' (Zenodo, 22 July 2022), https://doi.org/10.5281/zenodo.6884885.

[80] 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].

have been digitised, it obscures the distinction between the physical and the digital object 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[81], the preservation of digital derivatives should – in our opinion – follow the same principles: *authenticity*, *reliability*, *integrity* and *usability*.[82] 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 *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 a persistent identifier, adding additional information concerning the responsible institutions.[83] 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 perhaps 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, which could potentially be helpful for GLAM institutions that are under threat or suffer damage. If and where possible, such a structure could be used to provide such versionized images within an IIIF manifest.[84]

4. Conclusions and Recommendations

We started our contribution by discussing the export and sharing of Ground Truth. However, with sharing comes caring: properly acknowledging who provided the data or models *and* who contributed to their creation. We have discussed the HTR-United initiative and shown how one can register available datasets on this platform. This platform functions as a 'umbrella' solution allowing

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

contributors to use decentralised storage of their sources. At HTR-United, creators can be listed and metadata can be imported into Zotero for proper referencing.

Furthermore, we discussed issues that arose consequently: how best to acknowledge what digitised sources have been used, which seems dependent, at this point, on an author providing accurate annotation. Referring to a website, however, is not enough; we have indicated the need for persistent identifiers, as well. A persistent identifier distinguishes the digitised collection from the physical objects, and, more importantly, preserves the main characteristics of archival guarantees: authenticity, reliability, integrity, and (re)usability.[85] For clarity, the structure of the filenames could contain the institutional ISIL codes, as well.

Proper referencing of datasets and HTR models requires an overview of not only the underlying sources, but also adequate acknowledgement of contributors. In addition, in the case of HTR models, information about the quality and the processing of both the training and validation sets should be provided. 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 for and contextualization of a model. To describe the role of contributors and distinguish the various roles they could have, this article has suggested CRediT (Contributor Roles Taxonomy), which allows researchers and projects to reference the work of volunteers/citizens scientists properly, if they agree to be mentioned.

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

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

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'. Accessed 17 October 2022. https://georgianpapers.com/about/blog/.

Global Indigenous Data Alliance. 'CARE Principles of Indigenous Data Governance'. Accessed 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.

Chawla, Dalmeet Singh. 'A New 'Accelerator' Aims to Bring Big Science to Psychology'. Science, 8 November 2017. https://www.science.org/content/article/new-accelerator-aims-bring-big-science-psychology.

Chen, Ann. 'Library Guides: Mendeley: Home', 13 October 2022. https://aut.ac.nz.libguides.com/c.php?g=359376&p=2427744.

Cordell, Ryan. '"Q i-Jtb the Raven": Taking Dirty OCR Seriously'. *Book History* 20, no. 1 (2017): 188–225. https://doi.org/10.1353/bh.2017.0006.

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.

———. 'Talking about Viral Texts Failures', 25 June 2020. https://ryancordell.org/research/VT-database-fail/.

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.

'Diary, Letters and Poems of Marjory Fleming – Data Foundry'. Accessed 31 October 2022. https://data.nls.uk/data/digitised-collections/marjory-fleming/.

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).

Engelhardt, Claudia, Katarzyna Biernacka, Aoife Coffey, Ronald Cornet, Alina Danciu, Yuri Demchenko, Stephen Downes, et al. *D7.4 How to Be FAIR with Your Data. A Teaching and Training Handbook for Higher Education Institutions* (version V1.2 DRAFT). Zenodo, 2022. https://doi.org/10.5281/zenodo.5905866.

LECTAUREP. 'eScriptorium Tutorial (en)'. Accessed 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.

Föhr, Pascal. 'Historische Quellenkritik Im Digitalen Zeitalter.' Basel, 2018. http://edoc.unibas.ch/diss/DissB_12621.pdf.

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.

Hauswedell, Tessa, Julianne Nyhan, M. H. Beals, Melissa Terras, and Emily Bell. '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, no. 2 (1 June 2020): 139–65. https://doi.org/10.1007/s10502-020-09332-1.

Hodel, Tobias. '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. Konferenzabstracts*, edited by Christof Schöch. Paderborn: Christof Schöch, 2020. https://doi.org/10.5281/zenodo.3666690.

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.

Hoekstra, Rik, and Marijn Koolen. 'Data Scopes for Digital History Research'. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 52, no. 2 (3 April 2019): 79–94. https://doi.org/10.1080/01615440.2018.1484676.

'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.

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

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.

Muehlberger, Guenter, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, Hervé Déjean, Markus Diem, and Stefan Fiel. 'Transforming Scholarship in the Archives through Handwritten Text Recognition'. *Journal of Documentation* 75, no. 5 (2019): 954–76.

Ortolja-Baird, Alexandra, and Julianne Nyhan. '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 Archive1'. *Digital Scholarship in the Humanities* 37, no. 3 (1 September 2022): 844–67. https://doi.org/10.1093/llc/fqab065.

Padfield, Joseph, Charlotte Bolland, Neil Fitzgerald, Anne McLaughlin, Glen Robson, and Melissa Terras. 'Practical Applications of IIIF as a Building Block towards a Digital National Collection'. Zenodo, 22 July 2022. https://doi.org/10.5281/zenodo.6884885.

Rawson, Katie, and Trevor Muñoz. 'Against Cleaning'. In *Debates in the Digital Humanities 2019*, edited by Matthew K. Gold and Lauren F. Klein, 279–92. University of Minnesota Press, 2019. https://doi.org/10.5749/j.ctvg251hk.26.

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* 16, no. 2 (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.

Takats, Sean. 'Facing Abundance: Zotero as an Enlightenment Tool'. Presented at the American Society for Eighteenth-Century Studies, Albuquerque, New Mexico, March 2010., March 2010. https://orbilu.uni.lu/handle/10993/50339.

'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.

Zaagsma, Gerben. 'Digital History and the Politics of Digitization'. *Digital Scholarship in the Humanities*, 16 September 2022, fqac050. https://doi.org/10.1093/llc/fqac050.

Acknowledgements

We thank the participants of the Transkribus User Conference 2022 and the organisers of this event for the opportunity 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., J.v.Z., A.C., A.S., M.T., H.S.J., P.v.d.H., M.v.L, C.K., A.R., C.S., A.B., K.D., M.A.A., A.A., E.B., L.V.B., A.B., D.B., A.Ch., A.N.D., K.V.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., E.K., L.R. L., S.L., T.O.L, A.v.N., J.N., L.M.v. N., J.J.O., V.P., M.E.P., J.J. P., L.S., A.S., E.S., N.v.d.S., J.P. v.d. Sp, B.B.T., G.V.S., V.V., H.W., S.W, D.J.W., R.Z.; **Methodology**: 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; **Writing – original draft**: C.A.R., T.H., F.G., J.v.Z., A.C., A.S., M.T., M.v.L., A.R., C.K., J.N., J.P., C.S., A.B., K.D., S.G.; **Writing – review and editing**: C.A.R., T.H., J.P.

Competing interests: the authors declare no competing interests.