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

C. Annemieke Romein, Tobias Hodel, Femke Gordijn, Joris Zundert, Alix Chagué, Milan Van Lange, Helle Strandgaard Jensen, Andy Stauder, Jake Purcell, Melissa Terras, et al.

Journal of Data Mining and Digital Humanities

EPIsciences overlay journals

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[1,2]

***C. Annemieke Romein[1,2,3], *Tobias Hodel[3], *Femke Gordijn[1], *Joris van Zundert[1], *Alix Chagué[4,5], *Milan van Lange[6], *Helle Strandgaard Jensen[7], *Andy Stauder[8], Jake Purcell[35], Melissa 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[2], Elisa Bastianello[14], Lukas Vincent Benzinger[2], Arno Bosse[1], David Brown[15], Ash Charlton[9,16], André Nilsson Dannevig[17], Klaas van Gelder[18,19], Sabine C.P.J. Go[2,22], Marcus J.C. Goh[2], Silvia Gstrein[20,21], Sewa Hasan[2], Stefan von der Heide[23], Maximilian Hindermann[24], Dorothee Huff[25], Ineke Huysman[1], Ali Idris[2], Liesbeth Keijser[26], Simon Kemper[26], Sanne Koenders[2], Erika Kuijpers[2], Lisette Rønsig Larsen[27], Sven Lepa[28], Tommy O. Link[2], Annelies van Nispen[6], Joe Nockels[10,16], Laura M. van Noort[2], Joost Johannes Oosterhuis[29], Vivien Popken[31], María Estrella Puertollano[2], Joosep J. Puusaag[2], Ahmed Sheta[32], Lex Stoop[36], Ebba Strutzenbladh[33], Nicoline van der Sijs[12], Jan Paul van der Spek[36], Barry Benaissa Trouw[36], Geertrui Van Synghel[1], Vladimir Vučković[2], Heleen Wilbrink[30], Sonia Weiss[8], David Joseph Wrisley[34], Riet Zweistra[36], and further anonymous citizen scientists/volunteers of the "Goetgevonden!" project[2]**

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

Corresponding authors: C. Annemieke Romein, annemieke.romein@huygens.knaw.nl , Tobias Hodel, tobias.hodel@unibe.ch

ISSN 2416-5999, an open-access journal

Abstract

This paper discusses best practices for sharing and reusing Ground Truth in Handwritten Text Recognition infrastructures, and ways to reference and acknowledge contributions to the creation and enrichment of data within these Machine Learning systems. We discuss how one can publish Ground Truth data in a repository and, subsequently, inform others. Furthermore, we 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.

Keywords

Automatic Text Recognition, Handwritten Text Recognition, Data Publication, Open Data, Data Curation, Ground Truth, Sharing.

I INTRODUCTION

Within the humanities, working with digitised (primary) source material is no longer a novelty. Due to both large and small projects over recentyears an increasing number of digital sources have become available. Increasingly, these projects have been realised and enriched with Automatic Text Recognition (ATR, machine learning-based text recognition for print and handwriting) techniques. Although the resulting datasets of machine readable texts are immensely promising for the humanities, these developments also inevitably challenge existing disciplinary practices.

This paper revolves around several challenges tied to preparing and publishing ATR results: No clear practices have been established on how digital resources like ATR recognition models and training material should be properly stored and cited. We lack a clear guideline of how we should make aware of the several layers of contributions in publishable products. These are the perspectives that require in-depth elaboration.

ATR and, more generally, the latest engines for automatic text recognition processes depend on the digitisation of sources and the production of transcriptions to create and synthesise models via machine learning. For general models, massive numbers of documents, accompanied by correct and (ideally) uniform transcriptions, understood as Ground Truth in machine learning, are fundamental; the production of these corpora is, therefore, a challenge that falls in the category of big science (e.g. Chawla [2017]). Groups of volunteers (citizen scientists) are frequently involved in this data creation process, which raises the question of how we should properly acknowledge their contribution. Moreover, when talking about the digitisation efforts

[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's Digital Humanities minor) taught by Annemieke Romein.

[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.

of the Galleries, Libraries, Archives, and Museum sector (GLAM), we should acknowledge the production of digital images of documents.[3]

To discuss this problem fully, we organised a hybrid workshop at the Transkribus User Conference 2022 (Innsbruck, Austria). In the context of Ground Truth creation, we aimed to discuss: *how can we properly reuse, reference, and acknowledge contributions. What are the best practices thus far?* Many participants shared our sense of urgency for these questions and proposed fruitful ideas. This paper is the result of that exchange, via a resulting writing sprint with the community. This paper contributes to ever more important workflow processes for data generation based on shared and highly practical experiences.

This article departs from the concept of Ground Truth. This concept stems from computer science, claiming that an object can be described as it is. From a philosophical and epistemological point of view, this is highly problematic. Supervised machine learning algorithms require such information to imitate the result in the form of a model. 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 the material, in our case, handwritten or printed material [Muehlberger et al., 2019, 957]. This form of representation then informs about the accuracy of algorithms since Ground Truth is partially used to measure errors

Initial transcriptions may contain quite a few mistakes, but thoroughly checking them – most often by a human – can lead to accurate transcriptions according to defined standards. Ground Truth should thus be understood as the 'gold standard' ideally being reached. Alternatively, as Muehlberger et al. [2019, 957] 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.' As such, it can function as benchmarked data. Having as much Ground Truth available as possible is essential to provide large (or even general) models for specific scripts or types of handwriting. However, once large models are available that can be fine-tuned (in the sense of transfer learning) a reduced amount of training data is needed.

Ground Truth can be drawn from many sources. A bespoke transcription can be produced from scratch for a specific ATR project, but it is often more efficient to adapt 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 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, such conventions must be included in the description or metadata, or at least are made available in some form. This will help potential future users select the Ground Truth that is most appropriate for their project and help explain certain behaviours of a model.

Generally and roughly speaking, there are various ways of producing transcriptions. Two frequently used approaches are diplomatic and semi-diplomatic transcriptions. The former transcribes as much as possible as is, taking a large character set into consideration; the latter allows for adaptations to improve readability, e.g. writing out abbreviations and simplifying

[3] 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 to determine what is available digitally and what has been left out.

some characters. Some transcriptions 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. Machine learning-based models do not care about character sets, but of course, trained models can only reproduce what they have been trained on.

From a legal perspective and because of the data's value due to its laborious production, Ground Truth should at least 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. Since both Ground Truth and images are needed for training processes, unfortunately, image rights may be an obstacle to (re)training Handwritten Text Recognition models.In any case, we should store different stages for future reuse.

In the first part of this article, we contextualise strategies within the ethical and legal limitations of sharing Ground Truth. Because of these limitations and the urge to make aware of labour, the reuse of Ground Truth requires that contributions and contributors be acknowledged, which is discussed in the second section. In our conclusion, we combine and synthesize the two parts. This article is a proposal intended to start a discussion about how to conduct and acknowledge the work that goes into generating training data for machine learning. It must be mentioned that the proposed solutions are thus not meant as definite or provide a complete overview of all thinkable options. Additionally, one should remember that this article is the result of a large group of people with varying backgrounds. Consequently, we want to make you aware that definitions may vary according to different fields, and we won't elaborate on all perspectives.

II 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 and institutions with various material 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 ATR engines, or other downstream tasks, such as language models for Named Entity Recognition [Ströbel et al., 2022]. 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 [Wilkinson et al., 2016]: it should also be (easily) findable by others. Still, sharing data can have legal and/or ethical limitations. It should be stressed that we are explicitly talking about sharing Ground Truth data and not about sharing ATR models in this section.

2.1 How to Export Data

The various programs that allow for the creation of ATR material 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[4]. If this is not possible, it is helpful *to at minimum* sustainably store the "pure" transcriptions.

Within the *Transkribus* tool, provided by the READ-COOP SCE, the export appears as shown in 1. Widely used standards, like ALTO XML, PAGE XML, and hOCR allow for an alignment between image and transcription – based on coordinates – which is required to connect transcribed text on a character, word or line basis with images and allows for the opportunity to

[4]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 or via metadata.

(re)train machine learning based models.[5] These two formats are also supported by the eScriptorium application (see 2), which has been developed in the context of a variety of national and European projects [Kiessling et al., 2019].

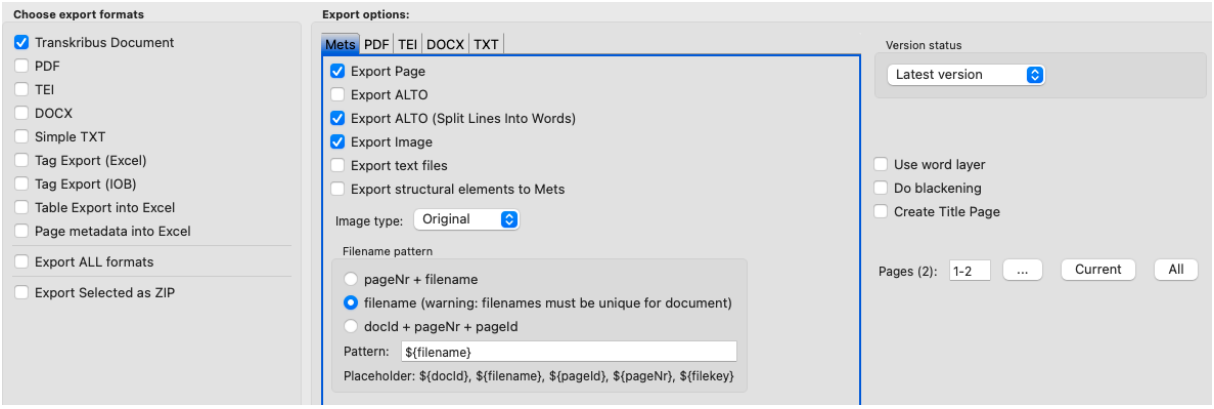

Figure 1: Screenshot Transkribus Export [version 1.22.0.1-SNAPSHOT]. [30 September 2022]

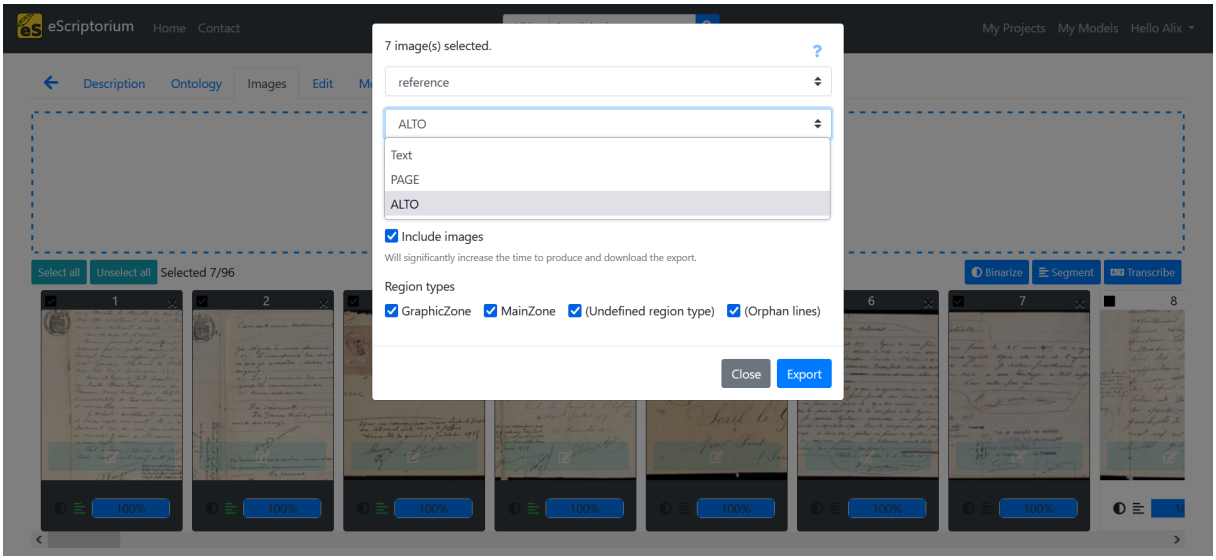

Figure 2: Screenshot eScriptorium Export [Version 0.12.5b]. [14 October, 2022]

ALTO and PAGE are the main formats used to store ATR output. TEI, better known in the Digital Humanities community, is primarily dedicated to producing critical digital scholarly editions but could also serve as a long-term storage format due to the wide user-base. The Gallicorpa project follows this approach and proposes TEI as an exchange format [Pinche et al., 2022]. 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 more or less open to adding custom properties. Exporting valid TEI XML as a third option is also sensible to us.

While some would call for a centralised Ground Truth repository, this could be a costly affair[6], and result in double the work, as funding agencies have requirements to store the output in

[5] 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.

[6] Including the uncertainty, who should declare itself responsible for the sustainability of such an environment.

specified (e.g. national) repositories. Consequently, a solution to the decentralised distribution of sources is discussed below.

2.2 Publishing Data in a Repository

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 structured format. Images and XML files should reside in sub-folders, with descriptive names for folders and 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 web pages. Furthermore, adding more detailed information in a README file is good practice and helps to navigate the data dump in the case of reuse [Sicilia et al., 2017]. Alternatively, data can be provided using publicly available Git repositories such as GitHub (owned by Microsoft) or Gitlab, but these do not provide DOIs. 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 is then used for handling the versioning and creation of releases. At the same time, Zenodo provides the user with a DOI, making the repository findable in the Zenodo search engine (see 3). If set in place, this allows different versions of transcriptions and documents to become available online, based on different parameters or (underlying) ATR 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.

At the same time, it is helpful to provide transcription guidelines or manuals to inform users about rules guiding the process and characteristics of the transcription of documents. In connection with particular Ground Truth, this information will allow potential users to search for data sets adhering to transcriptions that fit the criteria they are interested in [Sahle, 2016]. For example, the textual output could include larger or smaller character sets or only parts of a 'document' could be published.

Using a repository to share data is 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.[7]

2.3 HTR-United: Sharing Your Data

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

Alix Chagué and Thibault Clérice [Chagué and Clérice, 2022a] developed the HTR-United initiative to bring together different Ground Truth sets (see 4). HTR-United consists of three

[7] 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.

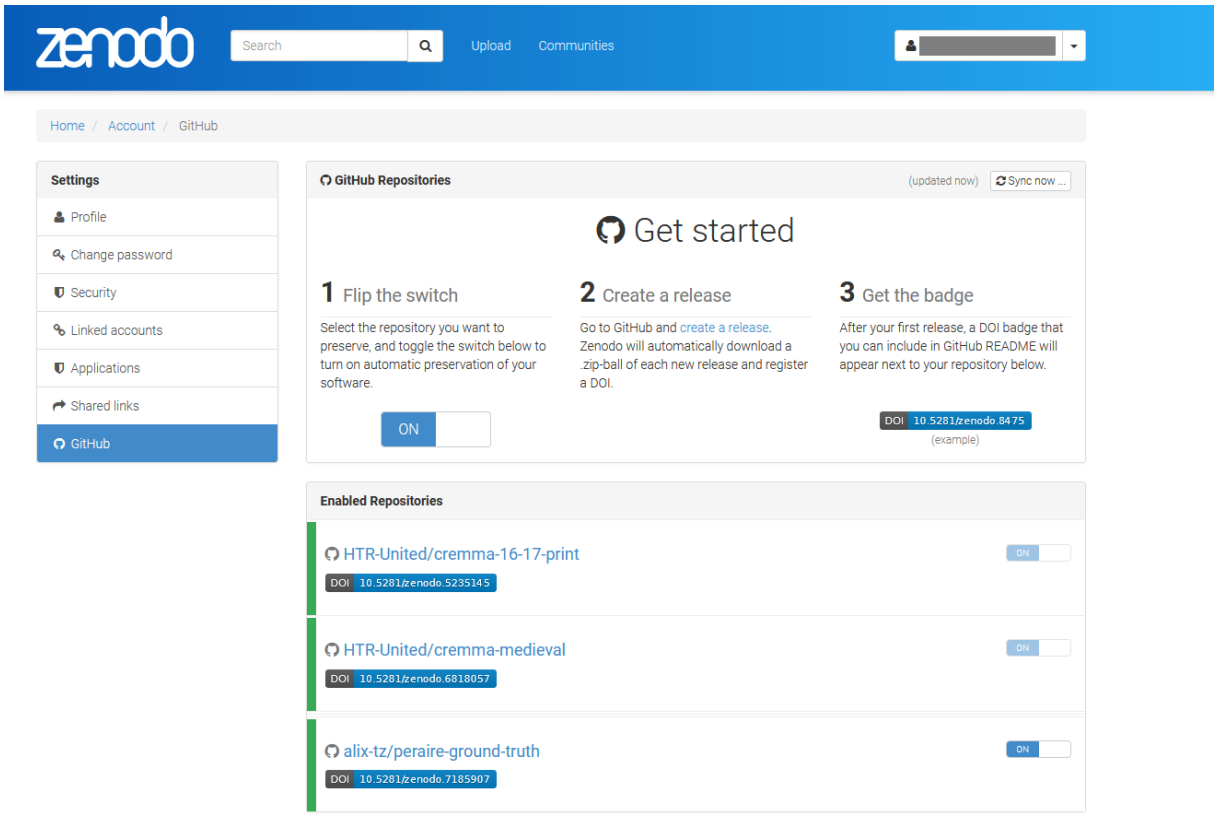

Figure 3: GitHub synchronisation and DOI generation on Zenodo. [17 October 2022]

imperatives: 'a collaborative enterprise for the community; friendly to consumers and data producers; as low tech as possible (because $$)' [Chagué and Clérice, 2022b]. Furthermore, according to Risam and Gil [2022] '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.'

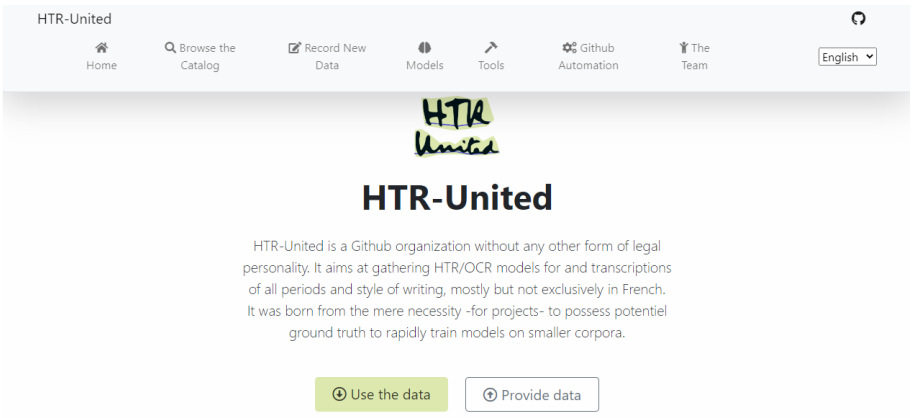

Figure 4: Website of HTR-United. https://htr-united.github.io/index.html [30 September 2022]

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 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 [Chagué and Clérice, 2022b, slide 15]. The developers (and at the same time, initiators) know that the schema with questions gains complexity However, they think it is worth the effort as it provides a uniform overview of the digital environment.

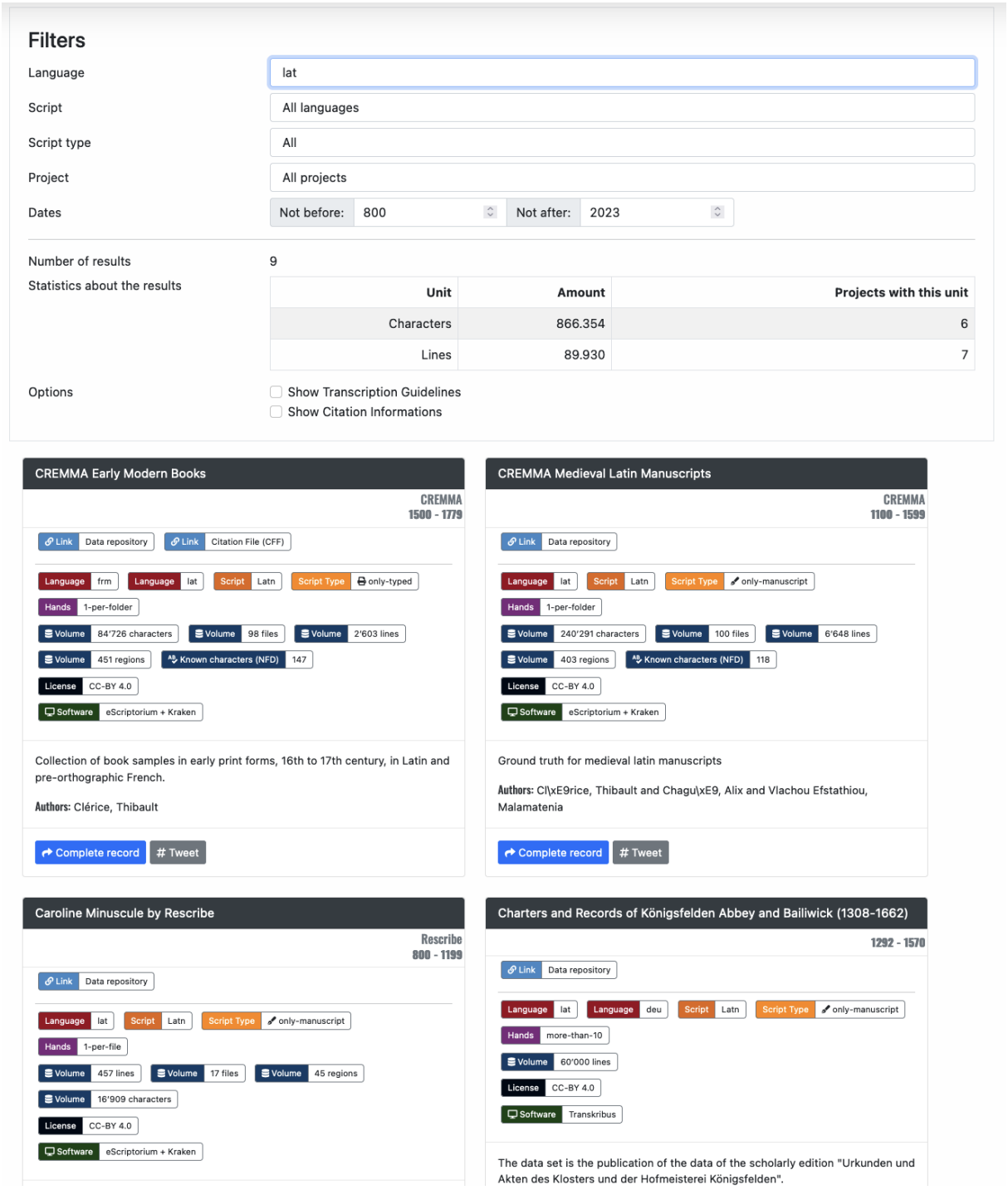

Figure 5: Catalogue of HTR-United. https://htr-united.github.io/catalog.html [30 September 2022]

HTR-United limits itself, 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. Also, it supports searches across the catalogue [Chagué and Clérice, 2022b, slide 10].

From the catalogue on the HTR-United website (see 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 6).

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, another research group ran a text-to-image process that resulted in an annotated dataset suitable for training an ATR 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 emphasise four key approaches to processed textual data for future text recognition.

- Export your data (including images, if possible);
- Upload it online, using services compatible with versioning like GitHub or better in repositories;

- Get a DOI, make it a publication;
- 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 ATR-produced transcriptions would also prove helpful, even though they are not perfect due to errors. 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) [Hodel et al. 2021, 13; Cordell 2017; Cordell 2020]. In the case of such machine-generated transcriptions, it could be advisable also to indicate the model used. Although CER is often used to measure quality, it needs to be mentioned that the calculation varies in the different tools, as recent studies show [Neudecker et al., 2021], so it is necessary to mention the used tool. Also, ideally, CER is only measured on test sets that an engine did not use in the training process.[8]

III REFERENCING DIGITISED RESOURCES AND DIGITAL OUTPUT

For certain objects in the humanities, such as physically published books, it is obvious how to cite them clear what questions need to be answered in a citation. It needs to state who wrote the text, who contributed, and what the source was. An exact structure must be followed, depending on the citation style. For this section, we focus on referencing digital objects, whether they are resources (digitised texts), datasets (recognized texts) or even ATR models (algorithms that allow the processing of digitised texts). Compared to manuscripts, prints, and other forms of written documentation referenced for centuries and even millennia, approaches to dealing with digital (ephemeral) objects are in their infancy [Föhr, 2018].

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. Zotero is a free, open-source referencing tool provided by the Corporation for Digital Scholarship that can adapt to various referencing styles. As it is a free and open-source tool that has been programmed by and for humanities scholars [Takats, 2010], we use Zotero as a point of reference for suggesting how to reference and acknowledge digital (re)sources and contributors.[9] We have combined experiences, suggestions, and guidelines in this section. As above, we focus on FAIR and CARE principles (see section 3.3), 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, and what aspects of a digital resource should be acknowledged.

3.1 Referencing Datasets

Data models are only starting to be cited at this point in time in research[10], especially in the humanities which results in a lack of standards within the field.[11] However, in computer sciences and machine learning, guidelines on how to cite datasets and software exist and are mostly adhered to Gebru et al. [2021].

[8] In sense of a set of pages that has not been part of the used training and validation pages. This is i.e. currently not the case in Transkribus models.

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

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

[11] Only in the Natural Language Processing field we encounter references to hubs like Hugging Face (https://huggingface.co/) and language models, taggers, etc. stored on the platform.

Several kinds of datasets could and should be cited. First, 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 and anything in between, such as machine-generated but manually corrected Ground Truth. Second, in addition to 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 like 'datasets' or 'data/ATR models', though they state that the category 'datasets' will soon be added.[12] In this way, the ability to cite these kinds of scholarly and scientific contributions will be easier and hopefully, part of future releases of large data sets that acknowledge such contributions accordingly.[13]

Since data [Gitelman, 2013], models [Speer, 2017], and even concrete objects [Woolgar and Cooper, 1999] are seldom neutral, we need to think about metadata and data publications not only in terms of citation technologies but as a mean to an end in itself. Over the last few years, the potentially egregious effects of using skewed or biased training data have been more coherently acknowledged in computer science, machine learning, Natural Language Processing, and other data-intensive fields [Mehrabi et al., 2022]. 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 ensuring sufficient pre-analysis documentation exists to allow for the responsible use of data. As Gebru et al. [2021] 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'. Thus, responsible metadata does not just encompass the application of FAIR principles [Wilkinson et al., 2016] and sufficient provenance information; it also details why the data was gathered, and for what research purposes and to what end the research was conducted, which relevant tools and technologies were used in the collection process, and if and how it underwent possible transformation processes (selection and 'cleaning') 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 let alone data 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 [Riley and Becker, 2010]. 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 (meta)data is adhering.

[12] See 'DataSets', Zotero Forums, accessed 20 October 2022: 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*.

[13] 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.

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' [Druskat, 2022]. Proper citing of datasets facilitates research transparency and ensures credit and accountability on the part of the dataset producers [Ball and Duke, 2015]. Metadata fields that should be part of any data set citation, if known, include author, publication date, title, version, resource type, publisher, identifier, and location.

3.2 Referencing ATR Models

In parallel to data sets, the 'Item Type: Software' could be used for referencing ATR models, as is suggested on the Zotero forum.[14] 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 ATR models or whether other disciplines might offer more fitting approaches. Let us first make an inventory of elements useful for citing an ATR 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 ATR 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(s) of an ATR model. 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).

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 highly informative data to judge the quality of the ATR model and its tendencies to overfit [Hodel, 2020].

Then, to complicate matters, it is possible to create new models based on existing models as a so-called 'base model' (called 'fine-tuning' in machine learning terms). 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.

As mentioned above, 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 ATR models. Congruent with what has been said about metadata and datasets, we need a quite granular schema for describing ATR models. Mitchell et al. propose a 'model card' to inscribe sufficient metadata and context about a model [Mitchell et al., 2019]. Such model cards have been implemented in the Hugging Face 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.[15] ATR models

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

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

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. ATR models should be cited on the same basis as any other research product [Ball and Duke, 2015]. Consequently, the metadata fields to include for ATR 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.

3.3 Ethics and Limitations of Sharing

Those sharing data must be aware of the 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 an ATR 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 ATR models to indicate various points of view without siding with either.

Without going deep into discussions about business models of services and platforms, different trajectories to guarantee sustainability can be taken. READ-COOP SCE does 'share as much as possible, and retain as much control as necessary' in order to sustain its business and maintain its infrastructure. While eScriptorium (as a second example) provides its software open source but no or only limited server space and power to train and use models. In both approaches, the sharing of Ground Truth and recognized text is foreseen and possible. Allowing to switch between systems and make vital data available.[16]

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 [e.g. Ortolja-Baird and Nyhan [2022]]. 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, the consequences of working with and sharing data must be kept in mind. For this reason, in addition to 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.[17] In short, CARE stands for "Collective benefit", "authority to control", "responsibil-

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

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

ity", and "ethics", making us aware of the necessity to think about people and cultures that are being treated as "data" and to give those affected a voice to consider Carroll et al. [2020].

An example from the NIOD Institute for War, Holocaust, and Genocide Studies illustrates the challenges that sources can bring to the surface. In the ATR-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 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 [Keijzer et al., 2022]. By considering ethics as one important part of data publication, CARE has partially been accounted for in this case.

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.[18] The strategies relevant to the case at hand require anonymisation, pseudonymisation, data minimisation, retention period and timely deletion, privacy 'by default', honouring the rights of whom the data concerns, and information security.[19]

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.

IV ACKNOWLEDGING CONTRIBUTIONS

When we consider the proper acknowledgement of datasets and ATR models we should not forget that their creation was a joint effort. As Ground Truth and transcriptions that underlie ATR models are often supported by 'the crowd', volunteers, or citizen scientists as a joint effort, and digitisation is often the result of institutional activities, we would like to address issues that come up when acknowledging these contributions in this penultimate part.

4.1 Acknowledging the Crowd/ Citizen Scientist

In an increasing number of digitisation projects, 'the crowd' is essential in generating Ground Truth data by transcribing or correcting transcriptions which are then used for the training of

[18] 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, 33-34]. See: https://kia.pleio.nl/attachment/entity/a8e1caa5-0d59-4267-bbc0-4cd288b2a56c.

[19] Put into a more narrative form this means, 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; 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.

new ATR models. Acknowledging the crowd is important not only due to their hard work but to provide insight into *how* and *with what resources* Ground Truth data was produced. Properly citing the crowd contributes to a more transparent data production process. However, there are no clear standards yet for how this should be done. The following section deals with the question of how to acknowledge the crowd sustainably and fairly. 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.

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/ [31-10-2022]

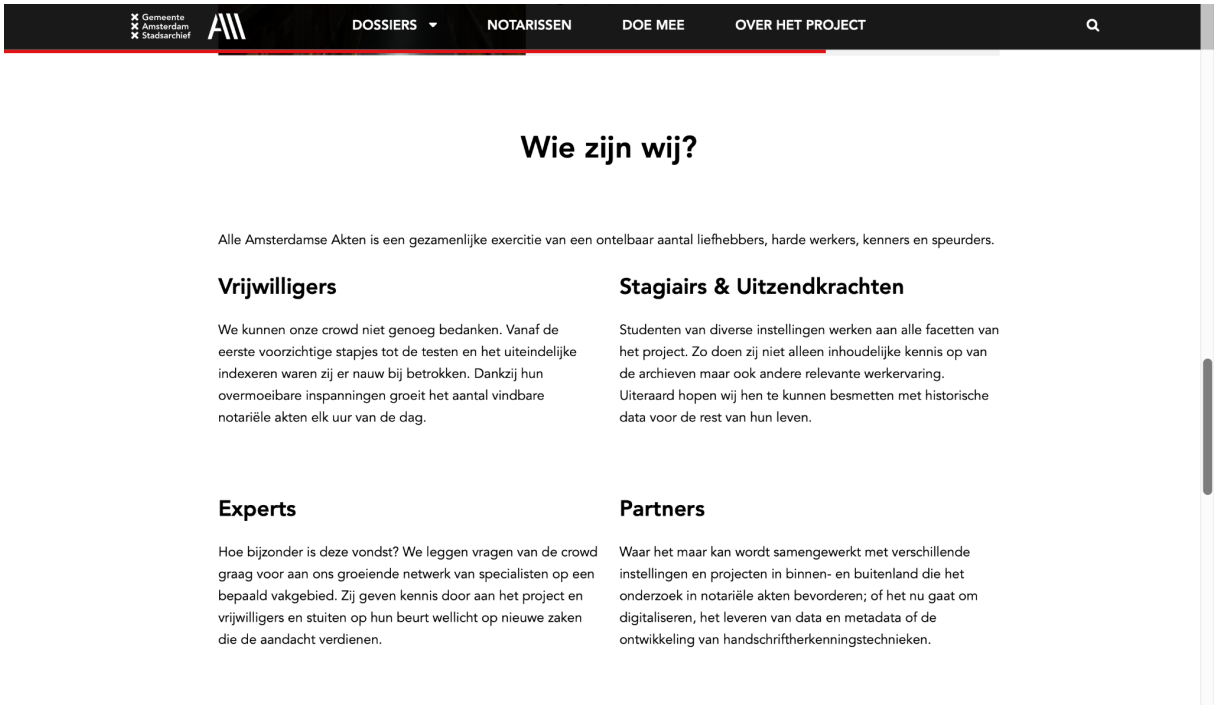

Figure 8: Part of Alle Amsterdamse Akten (All Amsterdam Notarial Deeds) Website, acknowledging the volunteers (crowd) as one group. [31-10-2022].

4.1.1 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 7 and 8).[20] Some do so for practical reasons,

[20]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/.

others to intentionally emphasise the collective effort instead of the individual. Second, there are projects, especially smaller ones, that acknowledge their volunteers by listing them with their full credentials in recognition of their work (see 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 in 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 section 4.1.2).

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, one 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.

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.

4.1.2 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 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.

Future complications could be avoided by presenting the citizen scientists with a digital form

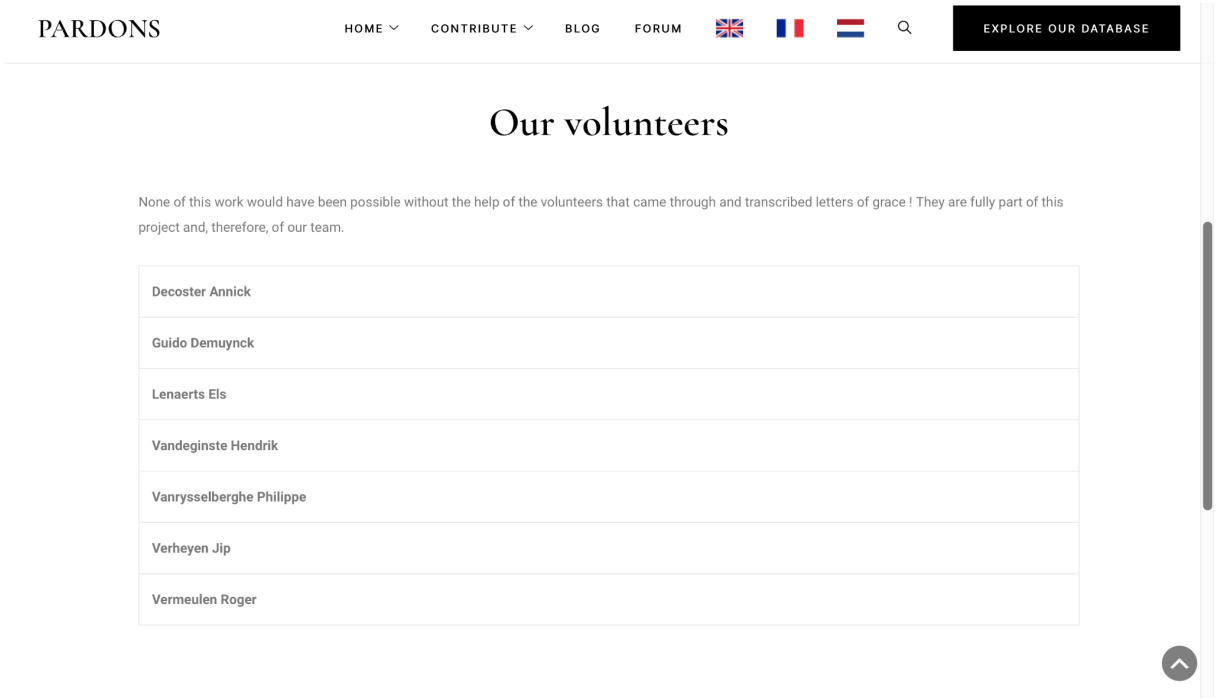

Figure 9: Part of the project website of Pardons; here, the names of all volunteers are listed (https://pardons.eu/the-team/). [31-10-2022]

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

However, one can imagine that, especially for larger projects which have already started, asking every individual member of the crowd for their consent can result in an administrative nightmare. There is an 'opt-out' method for these cases to deal with the GDPR. 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 time frame. 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 options, the option should remain for volunteers and their heirs to withdraw their names 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.

As shown, acknowledging involved people is not a simple task and requires action on many

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). [31-10-2022]

levels. A feasible and widespread approach for acknowledging has been provided within the frame of the CRediT taxonomy [Allen et al., 2014]. Some journals, such as *Science*, already work with this model and add an acknowledgement section to their article [Kestemont et al., 2022]. 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.'[21]

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

While this overview might look complicated, work on Ground Truth, datasets, or databases generally fits within the frame of *data curation* or *resources*. 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.[22]

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

[22] 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. However, courtesy could and should require a proper acknowledgement of work put into creating files. ECLI:NL:RBDHA:2022:8828, Rechtbank Den Haag, C/09/586380 / HA ZA 20-36,

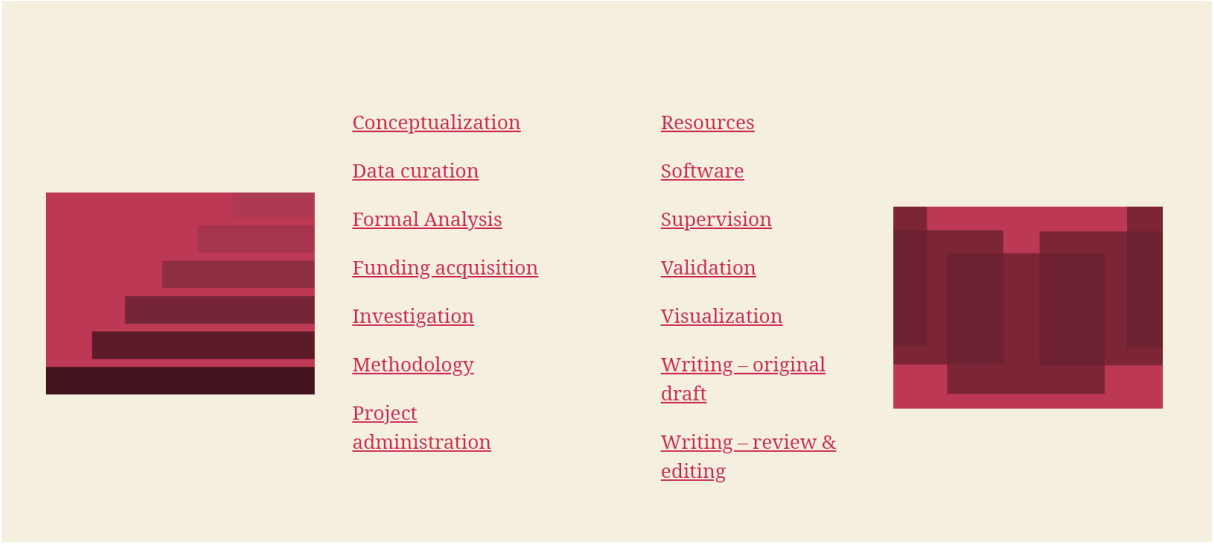

Figure 11: CRediT taxonomy. https://credit.niso.org/

4.2 Acknowledging Institutional Activities: Digitisation Activity and Contextualisation

GLAM-sector institutions, but of course also private institutions, digitise their collections. Digitisation is a time-consuming and costly process that is, by now, part of their core business.[23] It takes time, and this steadily paced process is only occasionally 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.

4.2.1 Reflections, exports, and clarifying documentation

Researchers active with digital resources have developed a critical perspective on collections and their provenance from archives, covering questions such as the selection of digitized data, physical aspects and others shown in 12 below.

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.[24] 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.

No. ECLI:NL:RBDHA:2022:8828 (Rb. Den Haag 3 August 2022).

[23] Although not explicitly covered in this article, it is polite to acknowledge institutions and funders. Their efforts and/or financial support allowed for creating 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.

[24] 'What Are Archives? | International Council on Archives', accessed 2 October 2022, https://www.ica.org/en/what-archive.

- 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 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 [Hoekstra and Koolen, 2019; Engelhardt et al, 2022; Hauswedell et al, 2020].

This also raises 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 digitise. Hauswedell et al. suggests that the institutional choices that went into choosing items for digitisation should be made clear to users [Hauswedell et al., 2020]. Jensen suggests that digital archives could be encouraged to demonstrate the extent and content of their digitisation efforts [Jensen, 2021, 256]. 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. 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 [Jensen, 2021, 258]. 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.

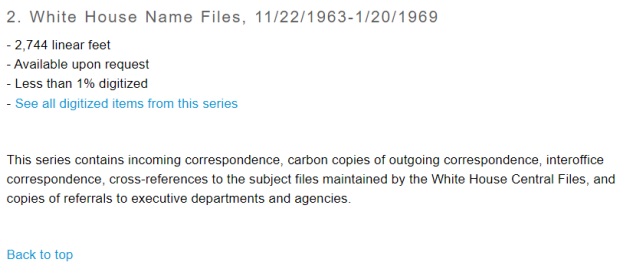

Figure 13: Section from the White House Central Files (WHCF) created or collected by President Lyndon B. Johnson and his staff, with an indication how much of the collection has been digitised. https://www.discoverlbj.org/exhibits/show/loh/pres/whcf

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 multi-

ple entries that allow for differences and similarities between conceptual models found in the archive and researchers' (changing) conceptual models [Jensen, 2021, 257]. 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. At the same time she highlights problems of bias, the historicity of the language as well as standardization that can cause problems for future historians [Jensen, 2021, 258-9].

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' [Jensen, 2021, 258-9]. This concern could go two ways. On the one hand, one could argue that a selection bias based on the interests of funding individuals/institutions has been, and still is, also a problem of analogue archives. In other words, traditional archives require funding too, and the ones paying for them will necessarily have an influence on the archive's contents. One could spin this thought 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 [Jensen, 2021, 258-9], we cannot agree with that stance, noting in particular political and infrastructural decisions. 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 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.

4.2.2 Digital images as proper 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 accidentally or intentionally not digitised, and (more or less) deliberate decisions on colouring and lighting, all leading to specific representations of objects that require critical approaches [Cordell, 2022]. To differentiate between digital facsimiles and their physical objects, digitising institutions should provide explicit guidelines for how they want their digitised facsimiles to be referenced [Rueda et al., 2017].[25]

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. Jensen, in the above-mentioned piece remarks that historians rarely disclose whether they accessed a physical or digitised version of their sources making us aware of the notion to discuss digital archives, as part of GLAM institutions [Jensen, 2021, 260].

[25] 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/.

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 sub-optimal/skewed way in the digitised version. Nevertheless, the obvious 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 datafied version and not adequately in the physical object.

The digital turn in the humanities thus requires that researchers become more aware of their data's source and 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.[26] 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 sustainability, since they can be changed at any time, and so do not provide the stability academic users seek [Padfield et al., 2022]. Furthermore, several GLAM institutions even offer references to the exact locations of words within their digitised resources.[27]

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 [Featherstone, 2006], the preservation of digital derivatives should – in our opinion – follow the same principles: *authenticity, reliability, integrity* and *usability*.[28] 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.[29] 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 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 un-

[26] 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.

[27] 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. 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].

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

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

der threat or suffering damage. If and where possible, such a structure could be used to provide such versioned images within an IIIF-manifest.[30]

V CONCLUSION 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 an 'umbrella' solution allowing 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.[31]

Proper referencing of datasets and ATR models requires an overview of not only the underlying sources, but also adequate acknowledgement of contributors. In addition, in the case of ATR 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' 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/citizen scientists properly, if they agree to be mentioned.

Although this is one example of how machine learning is being rolled out in the humanities, but in parallel 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.

VI REFERENCES

References

Liz Allen, Jo Scott, Amy Brand, Marjorie Hlava, and Micah Altman. Publishing: Credit where credit is due. *Nature*, 508(7496):312–313, April 2014. ISSN 1476-4687. doi: 10.1038/508312a. URL https://www.nature.com/articles/508312a. Number: 7496 Publisher: Nature Publishing Group.

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. doi: 10.1007/1-4020-5340-1. URL http://www.dcc.ac.uk/resources/how-guides.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. The CARE Principles for Indigenous Data Governance. 19(1):43, November 2020. ISSN 1683-1470. doi: 10.5334/dsj-2020-043. URL https://datascience.codata.org/articles/10.5334/dsj-2020-043. Number: 1 Publisher: Ubiquity Press.

[30] Especially the archival specification for IIIF is to be mentioned in this regard. See online: https://archival-iiif.github.io/.

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

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

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, June 2022b. CREMMALab. URL https://hal.inria.fr/hal-03703989.

Dalmeet Singh Chawla. A new 'accelerator' aims to bring big science to psychology, November 2017. URL https://www.science.org/content/article/new-accelerator-aims-bring-big-science-psychology.

Ryan Cordell. "Q i-jtb the Raven": Taking Dirty OCR Seriously. *Book History*, 20(1):188–225, 2017. ISSN 1529-1499. doi: 10.1353/bh.2017.0006. URL https://muse.jhu.edu/article/674968. Publisher: Johns Hopkins University Press.

Ryan C. Cordell. Talking about Viral Texts Failures, June 2020. URL https://ryancordell.org/research/VT-database-fail/.

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

Stephan Druskat. Research software citation for researchers, October 2022. URL https://cite.research-software.org/researchers/.

Mike Featherstone. Archive. *Theory, Culture & Society*, 23(2-3):591–596, May 2006. ISSN 0263-2764. doi: 10.1177/0263276406023002106. URL https://doi.org/10.1177/0263276406023002106. Number: 2-3 Publisher: SAGE Publications Ltd.

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

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets, December 2021. URL http://arxiv.org/abs/1803.09010. arXiv:1803.09010 [cs].

Lisa Gitelman, editor. *Raw Data Is an Oxymoron*. MIT Press, January 2013. doi: 10.7551/mitpress/9302.001.0001. URL https://direct.mit.edu/books/book/3992/Raw-Data-Is-an-Oxymoron.

Tessa Hauswedell, 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(2):139–165, June 2020. ISSN 1573-7519. doi: 10.1007/s10502-020-09332-1. URL https://doi.org/10.1007/s10502-020-09332-1.

Tobias Hodel. Best-Practices Zur Erkennung Alter Drucke Und Handschriften – Die Nutzung von Transkribus Large- Und Small-Scale. In Christof Schöch, editor, *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts*. Christof Schöch, Paderborn, February 2020. ISBN 978-3-945437-07-0. doi: 10.5281/zenodo.3666690. URL https://zenodo.org/record/3666690.

Tobias Hodel, 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(0):13, July 2021. ISSN 2059-481X. doi: 10.5334/johd.46. URL http://openhumanitiesdata.metajnl.com/articles/10.5334/johd.46/. Number: 0 Publisher: Ubiquity Press.

Helle Strandgaard Jensen. Digital Archival Literacy for (All) Historians. *Media History*, 27(2):251–265, April 2021. ISSN 1368-8804. doi: 10.1080/13688804.2020.1779047. URL https://doi.org/10.1080/13688804.2020.1779047. Publisher: Routledge _eprint: https://doi.org/10.1080/13688804.2020.1779047.

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

Mike Kestemont, 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(6582):765–769, February 2022. doi: 10.1126/science.abl7655. URL https://www.science.org/doi/10.1126/science.abl7655. Publisher: American Association for the Advancement of Science.

Benjamin Kiessling, 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)*, volume 2, pages 19–19, September 2019. doi: 10.1109/ICDARW.2019.10032.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning, January 2022. URL http://arxiv.org/abs/1908.09635. arXiv:1908.09635 [cs].

Margaret Mitchell, 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*, FAT* '19, pages 220–229, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596.

Guenter Muehlberger, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, Hervé Déjean, Markus Diem, Stefan Fiel, Basilis Gatos, Albert Greinoecker, Tobias Grüning, Guenter Hackl, Vili Haukkovaara, Gerhard Heyer, Lauri Hirvonen, Tobias Hodel, Matti Jokinen, Philip Kahle, Mario Kallio, Frederic Kaplan, Florian Kleber, Roger Labahn, Eva Maria Lang, Sören Laube, Gundram Leifert, Georgios Louloudis, Rory McNicholl, Jean-Luc Meunier, Johannes Michael, Elena Mühlbauer, Nathanael Philipp, Ioannis Pratikakis, Joan Puigcerver Pérez, Hannelore Putz, George Retsinas, Verónica Romero, Robert Sablatnig, Joan Andreu Sánchez, Philip Schofield, Giorgos Sfikas, Christian Sieber, Nikolaos Stamatopoulos, Tobias Strauß, Tamara Terbul, Alejandro Héctor Toselli, Berthold Ulreich, Mauricio Villegas, Enrique Vidal, Johanna Walcher, Max Weidemann, Herbert Wurster, and Konstantinos Zagoris. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 75(5):954–976, January 2019. ISSN 0022-0418. doi: 10.1108/JD-07-2018-0114. URL https://doi.org/10.1108/JD-07-2018-0114. Publisher: Emerald Publishing Limited.

Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stephan Pletschacher. A survey of OCR evaluation tools and metrics | The 6th International Workshop on Historical Document Imaging and Processing. (HIP '21: The 6th International Workshop on Historical Document Imaging and Processing):13–18, 2021. doi: 10.1145/3476887.3476888. URL https://dl.acm.org/doi/10.1145/3476887.3476888.

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 archive1. *Digital Scholarship in the Humanities*, 37(3):844–867, September 2022. ISSN 2055-7671. doi: 10.1093/llc/fqab065. URL https://doi.org/10.1093/llc/fqab065.

Joseph Padfield, Charlotte Bolland, Neil Fitzgerald, Anne McLaughlin, Glen Robson, and Melissa Terras. Practical applications of IIIF as a building block towards a digital National Collection. Technical report, Zenodo, July 2022. URL https://zenodo.org/record/6884885.

Ariane Pinche, Kelly Christensen, and Simon Gabay. Between automatic and manual encoding Towards a generic TEI model for historical prints and manuscripts, September 2022. URL https://zenodo.org/record/7092214.

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

Roopika Risam and Alex Gil. Introduction: The Questions of Minimal Computing. *Digital Humanities Quarterly*, 16(2), 2022. ISSN 1938-4122.

Laura Rueda, Martin Fenner, and Patricia Cruse. DataCite: Lessons Learned on Persistent Identifiers for Research Data. *International Journal of Digital Curation*, 11(2):39–47, July 2017. ISSN 1746-8256. doi: 10.2218/ijdc.v11i2.421. URL http://ijdc.net/index.php/ijdc/article/view/11.2.39. Number: 2.

Patrick Sahle. What is a Scholarly Digital Edition? In Matthew James Driscoll and Elena Pierazzo, editors, *Digital Scholarly Editing: Theories and Practices*, pages 19–40. Open Book Publishers, August 2016. ISBN 978-1-78374-238-7 978-1-78374-239-4 978-1-78374-240-0 978-1-80064-514-1 978-1-78374-627-9 978-1-78374-241-7 978-1-78374-242-4. doi: 10.11647/obp.0095. URL https://doi.org/10.11647/obp.0095.02.

Miguel-Angel Sicilia, Elena García-Barriocanal, and Salvador Sánchez-Alonso. Community Curation in Open Dataset Repositories: Insights from Zenodo. *Procedia Computer Science*, 106:54–60, January 2017. ISSN 1877-0509. doi: 10.1016/j.procs.2017.03.009. URL https://www.sciencedirect.com/science/article/pii/S1877050917302776.

Robyn Speer. How to make a racist AI without really trying, July 2017. URL http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/.

Phillip Benjamin Ströbel, Simon Clematide, Martin Volk, and Tobias Hodel. Transformer-based HTR for Historical Documents, March 2022. URL http://arxiv.org/abs/2203.11008. arXiv:2203.11008 [cs].

Sean Takats. Facing Abundance: Zotero as an Enlightenment Tool, March 2010. URL https://orbilu.uni.lu/handle/10993/50339.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E.

Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, March 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.18. URL https://www.nature.com/articles/sdata201618. Number: 1 Publisher: Nature Publishing Group.

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(3):433–449, 1999. ISSN 0306-3127. URL https://www.jstor.org/stable/285412. Publisher: Sage Publications, Ltd.

A 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.