International Journal of Digital Humanities
https://doi.org/10.1007/s42803-025-00107-7

ORIGINAL PAPER

From research proposal to project management. A guide from the Transkribus community on planning and executing workflows for researchers and GLAM-professionals

C. Annemieke Romein[1,2,3,4] · Süphan Kırmızıaltın[5] · Ronny Reshef[6] · Christa Schneider[2] · Giorgia Agostini[3,7] · Ash Charlton[8] · Melissa Terras[3,8] · Joseph Nockels,[9] et al. ***[full author details at the end of the article]***

Received: 5 November 2024 / Accepted: 1 August 2025

Abstract

Transitioning from research proposal to project management in Automatic Text Recognition (ATR) for cultural heritage text collections demands meticulous planning, realistic budgeting, and efficient team coordination. Insights from the Transkribus User Conference 2024 offer a roadmap for success, emphasising clear objectives, ethical considerations, and practical outcomes. Effective budgeting should include digitization, software, staffing, and long-term storage, while a comprehensive Data Management Plan (DMP) ensures efficient data handling. Key project management tasks include team formation, workflow documentation, and document layout optimization. High-quality training and validation data in ATR require clear guidelines, iterative testing, and agreed licensing. Crowdsourcing can enhance ATR projects, and sharing models and datasets through platforms like Transkribus Sites promotes collaboration and visibility.

Keywords Research Proposal · Project Management · Automatic Text Recognition (ATR) · Data Management Plan (DMP) · Transkribus · Digitization · Budgeting · Crowdsourcing · Volunteers · Digitalization · Collaboration · Document Layout Optimization · Cultural Heritage Sector · Digital Humanities

This article is the result of a workshop/roundtable and consequent writing sprint organised during a workshop at the Transkribus User Conference (TUC) 2024 on 'from Project Idea to Project Management' by Süphan Kırmızıaltın, Annemieke Romein, Christa Schneider, Ronny Reshef, Melissa Terras, Heleen Wilbrink, Achim Rabus, Mirjam El Attal, Günter Mühlberger.

Published online: 01 September 2025

1 Foreword

While a foreword is uncommon in academic articles, we find it necessary to contextualise this work within our lived experiences. The true value of culture to society often becomes evident only when it is inaccessible. Current political developments and financial constraints, including (looming) budget cuts, pose significant challenges to those in academia and the GLAM sector (Galleries, Libraries, Archives, and Museums). These difficulties are particularly apparent in the research funding application process, where the field's challenges are acutely felt.

This paper does not aim to provide evidence of success or a formula for achieving it. Instead, it synthesises its contributors'collective knowledge and experiences, many of whom have faced repeated rejections of grant applications and research proposals. These experiences require reflection, resilience, and, as some have admitted, occasional indulgence in (Swiss) chocolate. We empathise with others encountering similar challenges.

Despite these setbacks, the knowledge gained, and research conducted during the preparation of grant proposals is not wasted. Much of this work can be repurposed for future proposals or integrated into ongoing workflows. Reviewers may not always recognize the long-term value of technical innovations. For example, one author received a review questioning the necessity of Automatic Text Recognition, suggesting it was feasible to analyse approximately 200,000 documents manually. While such feedback can be discouraging, perseverance is crucial as the value of innovations may be better appreciated by future reviewers.

2 Introduction

This article synthesises shared insights on transitioning from research proposals to effective project management for automated transcription of historical text collections using AI-driven text recognition models. It offers a comprehensive roadmap for researchers and project managers planning projects that utilise Automatic Text Recognition (ATR) tools for historical and cultural heritage corpora.

Research projects, especially those involving extensive digitization and transcription, demand meticulous planning and management. Effective data management is essential for handling and preserving large ATR datasets, while good project management ensures coordinated teams and well-defined workflows for processing text collections in ATR software such as Transkribus. Maintaining high-quality, consistent training data and transcription output is vital, as is the potential use of crowdsourcing for ATR and Named Entity Recognition (NER). Ultimately, integrating digitised materials into mainstream workflows and sharing ATR models enhances the availability and impact of the project's outputs.

A key theme highlighted throughout the Transkribus User Conference 2024 was the collaborative nature of research and project management. Speakers

emphasised the importance of leveraging collective knowledge and the value of engaging volunteers in preserving cultural heritage. This collaborative approach not only enhances the quality of individual projects but also contributes to the broader research community and society at large. The themes which emerged from our discussion include Procedures, Collaboration, Documentation and Transparency, High efficiency and quality outputs, and teamwork.

3 Workshop and collaborative writing process

On 15 February 2024, around 60 enthusiastic participants attended a vibrant workshop. Participants documented their diverse perspectives and insights using a Padlet (Padlet 2024), facilitating comprehensive note-taking. These notes were later synthesised through a Research-in-Action methodology, fostering an iterative and collaborative process of understanding and organising information (McDermott et al., 2008). This approach involved active engagement with relevant literature and field experiences, ongoing reflection on findings, and continuous analysis refinement to ensure accuracy and relevance. All contributions have been anonymized.

Subsequently, the co-writing process commenced, employing writing sprints by multiple first authors to maintain focus and momentum. These intensive sessions, interspersed with feedback and discussion, facilitated the rapid development of the article. Employing community editing, we invited contributions from all workshop participants to ensure the final document accurately reflected the collective knowledge and perspectives. This collaborative approach enriched the content and fostered inclusivity, ownership, and engagement, underscoring the value of shared effort and mutual support in achieving a common goal (Blau & Caspi, 2009; Connelly, 2023). Figure 1 illustrates/shows the themes discussed in the paper.

4 Project idea

Successful ATR projects begin with a clear objective, often identified by pinpointing research gaps where tools like Transkribus can be applied (Colutto et al., 2019; Muehlberger et al., 2019). Technological advancements have made ATR vital for improving the accessibility of cultural and historical documents, thereby supporting heritage management. For example, ATR can address legibility issues in archival materials with unique fonts or handwriting. Hence, identifying collections that would benefit from ATR research is essential.

Initially, researchers must define their project goals. It is essential to differentiate between projects focused solely on digitization and those where digitization facilitates datafication for advanced analysis. In the former, resources can be dedicated to transcription and training data for the ATR system. In the latter, careful allocation of time and resources is necessary to support subsequent analyses. Long-term objectives may include creating a digital edition, preserving endangered records, enabling large-scale text mining, or addressing research questions that require a machine-readable corpus.

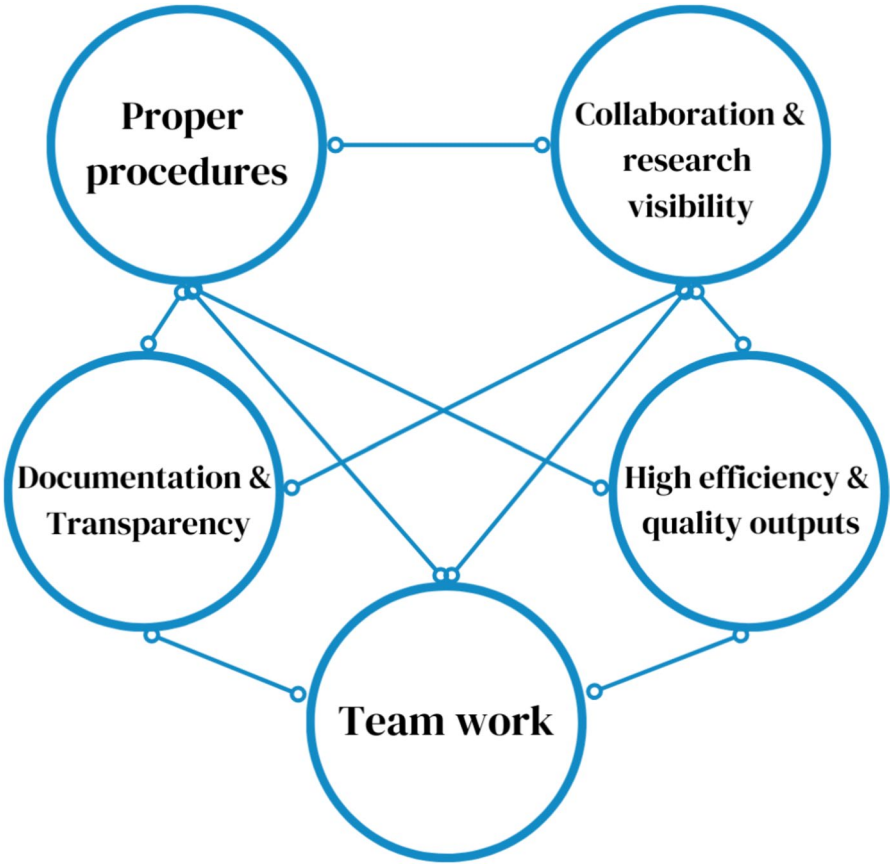

Fig. 1 Main themes for ATR-projects of historical texts as discussed in this paper (created with canva)

Assessing the suitability of ATR tools is critical and should align with the project's scope and objectives. Researchers must match their goals with the capabilities of the platforms and available technical support. For instance, projects that primarily utilise ATR outputs may favour platforms with graphical user interfaces like Transkribus, while those needing customisation might choose open-source options such as eScriptorium. Projects focused on development may opt to implement their own Optical Character Recognition or Handwritten Text Recognition tools, including, but not limited to, Kraken, Teklia, Loghi, tranSkriptorium, Calamari, Tesseract, OCR4all, ABBYY FineReader, and Calfa HTR. Although many of these tools are open-source and free to download, there may be additional costs for programming, server maintenance, and ongoing support (Nockels et al., 2024).

Envisioning the project's outcomes is crucial for developing a focused strategy. Given the workshop's focus, this paper will specifically address the use of Transkribus for transcription projects, rather than bespoke ATR tools.

5 Project elements and planning

Developing a comprehensive project plan and a compelling proposal is crucial for securing approval and support from academic departments, collaborators, or grant agencies and for successful project execution. Effective planning involves clearly defining the project's scope, pipeline, and expected outcomes.

5.1 Establishing a workflow

A well-defined workflow is essential for managing project complexities. It should cover all stages, from digitization to final transcription, metadata management, and publication. Documenting each step in detail is crucial to ensure team familiarity with the pipeline and adherence to tools and standards. The workflow should be continuously reviewed and refined based on feedback, with regular updates to documentation. This promotes transparency, consistency, and effective collaboration. Key workflow steps include preparation, digitization, cataloguing, metadata creation, and transcription scheme development.

5.1.1 Document preparation

When digitising fragile or deteriorated documents, an initial preparation and stabilisation stage following institutional standards may be required, this includes archival preservation techniques such as cleaning, repairing physical damage, and stabilising materials to prevent further deterioration. These measures ensure documents are in optimal condition for subsequent tasks like cataloguing, digitization, and storage. While this stage is vital for maintaining document integrity, it may not be necessary for all projects. Researchers working with pre-digitised collections or scanning documents from archives typically do not need to perform these preservation tasks.

5.1.2 Document digitization

Scanning and imaging should use high-quality equipment to ensure optimal resolution for accurate ATR transcription. For Transkribus, the Transkribus ScanTent is recommended for capturing high-quality, preferably high-resolution photos with a smartphone or tablet (Terras, 2022a). In large-scale projects, it is important to distinguish between master files, which should be high-resolution and uncompressed for long-term storage, and ATR-optimised files, which can be lower resolution for faster processing. Smaller research groups may opt for lower-resolution files to balance quality and processing efficiency based on their project needs and budget.

Preprocessing When required, image enhancement techniques should be applied to optimise material legibility and augment transcription accuracy. These may

encompass binarization, deskewing, and other refinement methods, thereby ensuring optimal input quality for ATR systems.

5.1.3 Cataloguing and metadata creation

Depending on the type of project and the documents involved, this stage might occur before or after the digitization/scanning of the documents.

Cataloguing Essential for substantial archival projects, cataloguing should utilise appropriate software (e.g., AtoM, Omeka) and adhere to international standards such as ISAD(G) and other International Council on Archives (ICA) descriptive standards (ISAAR(CPF), ISDF, ISDIAH). This approach ensures consistency, accessibility, and interoperability of archival descriptions, facilitating efficient organisation and retrieval of records.

Metadata creation Whether for institutional cataloguing purposes or small-scale document management on personal devices, a comprehensive metadata schema should be developed to capture essential information about each document, such as title, author, date, language, and source. This schema should be based on widely recognized metadata standards like the Dublin Core or MARC. Some metadata can be created within the Transkribus user interface, allowing for streamlined data entry and management (see Fig. 2). Alternatively, Microsoft Excel sheets may be utilised for smaller projects to create and maintain metadata records. Ensuring consistency and adherence to one of the aforementioned metadata standards is crucial for maintaining the quality and usability of the metadata.

5.1.4 Transcription scheme preparation

Transcribing historical texts demands astute interpretation and well-defined conventions. The project's objectives—whether scholarly publication, enhanced accessibility, or research support—should dictate the transcription methodology. Thus, a bespoke transcription scheme aligned with these aims is imperative. Romein et al., (2023: p. 3) stated, 'Two frequently used approaches are diplomatic and semi-diplomatic transcriptions.' While diplomatic transcriptions might be easier to merge into larger models, it may not be in a project's interest (or budget) to spend many additional hours to create these. Transcription schemes should align with project goals—and not necessarily the goals of potential future/other projects. For facsimile publication, detailed representation is crucial, capturing elements like typos, strikethroughs, abbreviations, and marginalia. This may require special characters and keyboards (Romein et al., 2023). Accurate palaeographic ground truth production is essential, particularly for early modern and medieval documents with varied scripts (see Fig. 3 and Fig. 4).

Conversely, projects prioritising accessibility may adopt less stringent schemes, potentially normalising words and abbreviations. Regardless of the objective, it is advisable to consult peers'work and adapt existing schemes to project-specific needs.

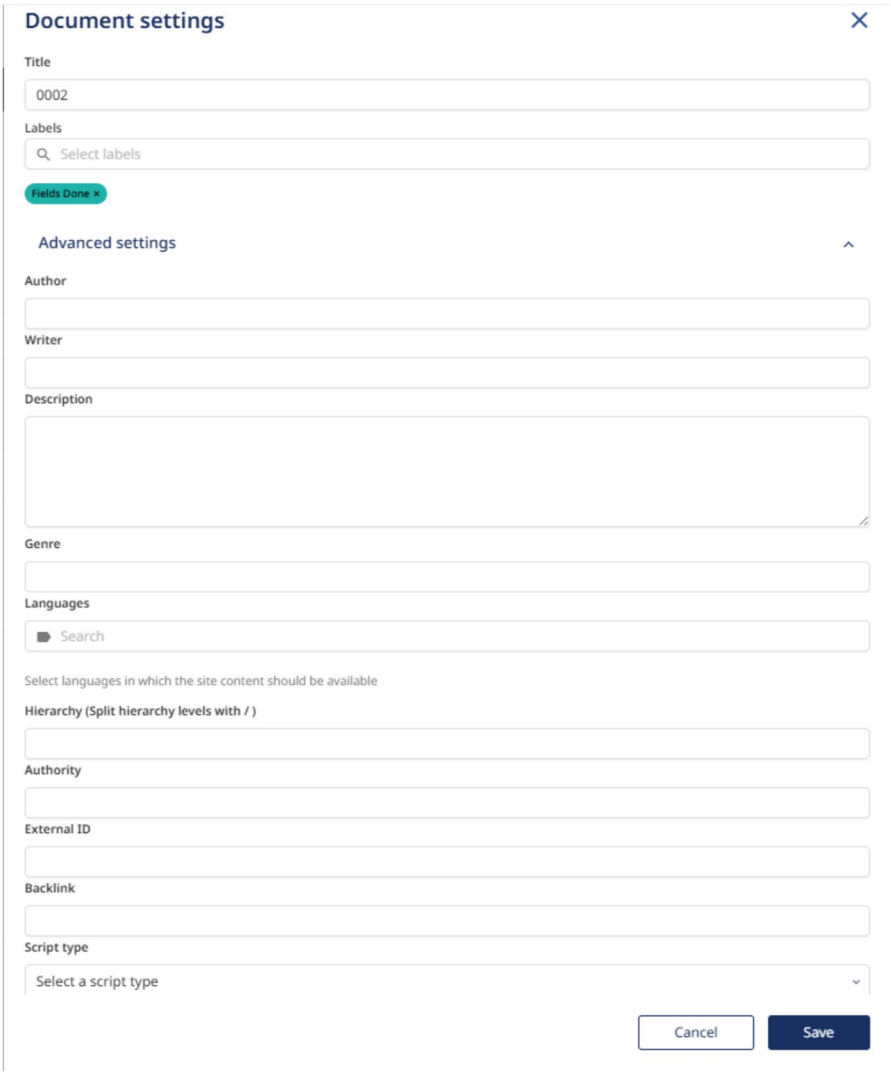

Fig. 2 Transkribus document metadata visual. Screenshot taken on 22.05.2025

Establishing, documenting, and disseminating transcription conventions at the project's outset is paramount. Given the iterative nature of such endeavours, early evaluation of initial outputs is crucial to ensure alignment with project objectives. This could be done by testing out existing models and refining them or by building one's own model. Regular assessment maintains transcription quality and consistency. Effective management of these complexities necessitates consistent workflow supervision, a topic revisited in Sect. 7 regarding adherence to transcription schemes and ensuring consistent ATR output.

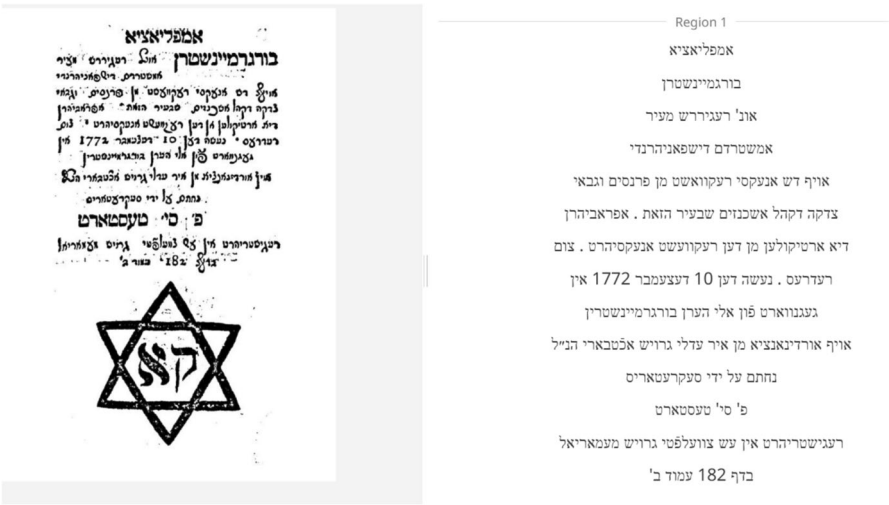

Fig. 3 Diplomatic transcription of p.16 from Kehal Kodesh Ashkenazim (Amsterdam. Netherlands). (1772). https://www.hebrewbooks.org/46840 (טראנסלאט, אמפליאציען על תקנות קהלתינו (8 אקטאבר))

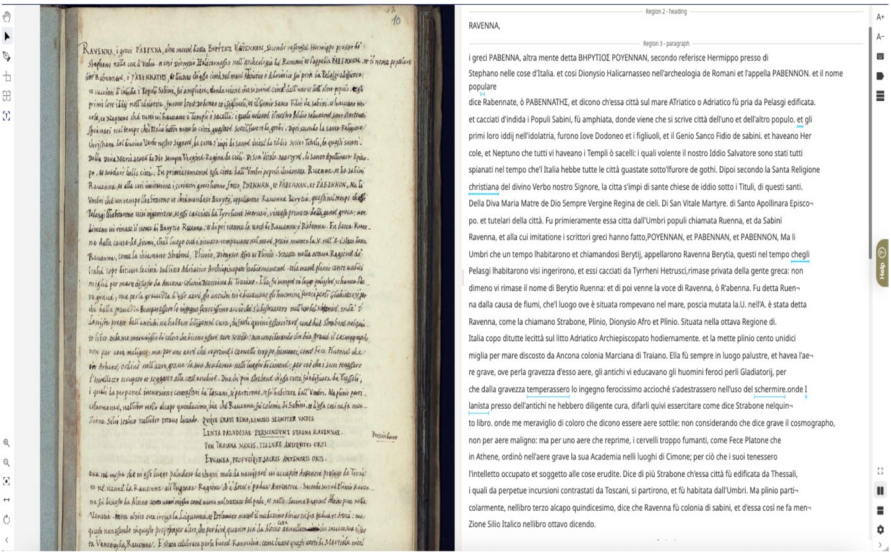

Fig. 4 Detail of Ground Truth transcription from folio 10r, Pirro Ligorio, Delle Antichità, Vol.15 Lib. XVII: Enciclopedia del mondo antico (R). Archivio di stato, Turin (Italy)

5.2 Team and roles

Establishing a well-defined team structure with clear roles is crucial in project planning. Determine the required number of team members and volunteers, identifying necessary skills for each role.[1] The CRediT (CRediT 2022) taxonomy can effectively delineate responsibilities, enhancing collaboration and data sharing by attributing specific contributions (Allen et al., 2014: 312). This approach clarifies individual tasks, their alignment with project objectives, and mitigates potential authorship disputes.

Team composition is contingent on project type, budget, and anticipated outcomes. Larger institutional projects may require specialists in Digital Humanities, digitization, palaeography, and volunteer management. Training team members in ATR tools and, if needed, hiring experts like language specialists or palaeographers is crucial. Regular workshops,[2] webinars, and meetings should be held to foster (remote) collaboration and keep the team updated on technological advancements. These practices ensure inclusive participation and support, especially for newcomers.

5.3 Grant writing

Ensuring that the selected grant aligns with the project's objectives, whether for digitization or research, is essential. Beyond this fundamental requirement, clear communication and realistic expectations for your topic are crucial for persuading committees and external reviewers.

A strong proposal should clearly define the project's problem, providing context and demonstrating its relevance. It should outline the approach and methodology, detailing the steps, tools, and processes that will be employed. Expected outputs, such as digitised texts, searchable databases, or enhanced accessibility of historical data, should be specified to illustrate the tangible outcomes. Additionally, the proposal should describe the envisaged solutions and benefits, addressing potential challenges and highlighting how the project might advance current practices. Finally, a realistic budget is crucial, with a detailed explanation of the project's costs that reflects a comprehensive understanding of the overall organisational process.

It is crucial to thoroughly plan the workflow before beginning to write the proposal. The Pareto principle applies here: around 80% of the effort should be devoted to preparation, including thinking, discussing, and refining the proposal, while only about 20% is spent on the actual writing. When describing the project to collaborators, team members, or for grant applications, use clear and concise language with a strategic structure. Start each paragraph with a summary sentence and ensure that the first paragraph of each section encapsulates the main points for clarity and coherence. Highlight the project's potential impact by addressing unexplored issues

[1] For a more detailed discussion on crowdsourcing, see below.

[2] An onboarding session is included in a Transkribus Organisation plan. A Transkribus online beginners'workshop or a dedicated Transkribus workshop can be purchased s. Free introductory webinars are frequently offered and can be found at https://www.transkribus.org/events.

and advancing existing methods. Define 3–5 clear goals and expected outcomes, focusing on quality over quantity.

Your grant proposal can benefit from addressing three specific elements that might not fit neatly under the standard proposal headers but are important to include: ethical considerations, estimated costs specific to ATR projects, and the potential gains of an ATR project. These themes will be discussed below.

5.3.1 Ethical considerations

Ethical considerations are crucial in both project planning and grant application stages, particularly regarding document handling protocols and respecting personal rights. A strong proposal outlines the project's goals and impact and addresses how it will contribute to the field and improve the accessibility and understanding of historical data. Ethical considerations can be divided into two key themes: FAIR/CARE principles and environmental costs.

From an ethical standpoint, it is important to account for the roles of creators, curators, and descendants of the materials involved. This is especially significant when dealing with historical materials originating from colonial contexts. In such cases, it is vital to consider the history of a document, including how it became part of an institution, and to critically assess the implications of making these materials publicly accessible as data. This includes acknowledging the perspectives and practices of non-Western communities, which may have different concepts of ownership and respect for historical documents.

Given these ethical complexities, it is essential to go beyond the widely recognized FAIR (Findable, Accessible, Interoperable, and Reusable) principles by also incorporating the CARE (Collective Benefit, Authority to Control, Responsibility, and Ethics) principles proposed by the Global Indigenous Data Alliance (Wilkinson et al., 2016; Carroll et al., 2020; 'CARE Principles of Indigenous Data Governance', 2022). The CARE principles emphasise the collective benefits of data production and sharing, while advocating for communities'rights to maintain authority over their data. While not yet as widely adopted as the FAIR principles, CARE principles introduce a crucial ethical dimension to data management, urging all involved to act responsibly and respectfully towards the communities whose data is used (Carroll et al., 2020).

Environmental considerations of using ATR technology include the potential impact of training and running machine learning models (Nockels et al., 2024). To minimise these costs, it is advisable to first explore existing models that can be customised for your needs, rather than developing new models from scratch. A key strategy is to avoid frequent retraining of models each time a small amount of new Ground Truth (GT) data is generated, as this can significantly increase carbon emissions. Where possible, training should be scheduled during periods of lower carbon-intensive energy use. Furthermore, the choice between models, such as Pylaia and TrOCR, should consider their differing environmental footprints. Adopting genuinely sustainable practices requires more than superficial compliance (DHCC, 2024); it involves an awareness of the ecological impact of model selection and usage. Collaboration is also crucial, as sharing models efficiently can reduce the overall need for model training and further lessen the environmental burden.

5.3.2 Budgeting: ATR-cost estimations

ATR projects are costly and demand substantial budgets, particularly when covering all phases from digitization to publication. It is crucial to conduct realistic tests with the material during the planning stage to ensure cost estimates are based on practical experience rather than overly optimistic projections. Hence, co-authors mention that they first test existing models on a subset and identify accuracy, errors, and issues. This practice allows for estimating how much labour will go into the process: both timewise and money-wise.

For accurate cost estimates, consider key areas such as image capture, Transkribus usage, workforce, optional crowdsourcing, and other essential resources. Additionally, organisational budgeting should include expenses for publishing and long-term storage, which may extend over 5, 10, or more years.

Cost estimations:

- Digitization: Account for both the quality and quantity of materials to be digitised, image capture equipment, and future resource needs, such as hardware.
- Using ATR-software: Include costs for the software, particularly any advanced features required. Also, consider expenses for other ATR approaches and providers, including staffing for programming and maintenance. If there are legal constraints on processing certain files (e.g. sensitive documents), consider solutions that can be used at local servers for example.
- Staffing: Estimate labour costs for transcription, validation, and correcting ATR outputs. This should include both paid staff and, if applicable, compensated or thanked volunteers.
- Crowdsourcing: If applicable, budget for engaging and managing a crowdsourced workforce.
- Other Resources: Include costs for additional essential resources, such as software tools, equipment, or specialised services.
- Publishing: Consider optional costs for publishing results to make resources widely accessible.
- Storage: Plan for long-term storage costs, covering periods of 5, 10, or more years. This could include visible results and separate storage of Ground Truth Data for future reuse.
- Outreach and visibility: Budget for at least one event—whether small or large, public or scholarly—to promote the project, foster collaboration, and enhance research visibility.

5.3.3 Gains of ATR-projects

While costs are an important consideration, the gains from ATR projects should not be overlooked. Human-led transcription is traditionally labour-intensive, with experienced paleographers achieving a maximum output of about 5 pages per hour. In contrast, advanced ATR models can significantly enhance efficiency. For instance, a robust Pylaia model can process a page every 20–30 s, whereas a TrOCR model typically takes 35 s.

ATR projects offer several significant benefits beyond their costs. One major advantage is preservation, as digitization protects original sources from physical handling and deterioration. Enhanced ATR models also improve efficiency by drastically reducing transcription time, which boosts productivity and enables the processing of larger datasets more rapidly. Moreover, digitised and transcribed texts become searchable, greatly enhancing their utility for researchers and users compared to manually reading through originals.

Additionally, transcriptions make texts reusable for various purposes, such as NER, and making sources and models openly accessible provides substantial benefits to the community and academic fields. This open access democratises information and supports a wide range of scholarly and public interests (Leonelli, 2023). Publishing GT data sets and models further reduces the time and resources required for similar projects, potentially decreasing the need for transcription and enabling the development of more robust models. This approach also helps lower the environmental impact of server capacity usage.

In conclusion, while crafting a comprehensive and realistic budget for an ATR project is essential, it is equally important to emphasise the substantial benefits and efficiencies these projects offer. By enhancing data availability, searchability, and transcription efficiency, ATR projects are crucial in advancing knowledge and improving access to historical data.

5.4 ATR-methodologies: techniques, engines, and platform

Addressing technical issues is vital in project formulation and proposal detailing. A clear grasp of training methodologies, validation procedures, and Character Error Rate (CER) metrics is essential for setting realistic expectations and demonstrating the capabilities and potential of the ATR methodologies used.

Using advanced features of Transkribus eXpert, such as page sample selection and comparison of test sets, enhances testing robustness and provides a clearer view of the model's performance.[3] Conducting preliminary tests, such as evaluating public models against manual transcriptions or training a small model, helps identify potential issues and areas for improvement early in the project.

Showcasing'low-quality'recognition results from existing models can highlight the need for the proposed model. However, set realistic expectations for the new model's performance and avoid over-promising. Depending on project goals, less-than-perfect transcriptions might be sufficient. Clearly define acceptable accuracy levels and CER standards.

Including case studies of successful applications and lessons learned adds value to the proposal. Use practical examples and avoid excessive technical jargon. For instance, replace'We trained ATR models using the CER metric'with'We used ATR software to train the system on old manuscripts and measured its accuracy using CER (Character Error Rate), a standard metric for evaluating transcription

[3] Several are expected to be included into the App in the future.

quality.'Clear, straightforward language aids reviewers'understanding of the methodology and its impact.

Advanced features like sample selection and comparison of test sets in case studies showcase a sophisticated ATR approach. Meticulous documentation and clear communication of these processes enhance credibility, demonstrate expertise, and contribute valuable insights to the research community.

6 Project approval and next steps

This section will explore project implementation, covering tasks such as drafting a management plan, conducting layout recognition, performing text recognition, and evaluating performance.

6.1 Data management plan: ensuring effective data handling and preservation

A comprehensive Data Management Plan (DMP) is vital for successfully executing any ATR project. The DMP outlines data collection, management, sharing, and preservation strategies, ensuring that all aspects of data handling are meticulously planned, executed, and documented. For data collection, it is necessary to describe the type of data that will be collected, including both the input and output data produced by the project. Mentioning the tools and methods used for data creation and supplying details on the expected size and format of the data, is essential for understanding the scope and scale of the data management requirements. Again, one should consider environmental factors when devising a DMP (DHCC Information 2022). A DMP is designed to be a dynamic document, allowing for detailed information to be added or refined through updates as the project develops and when major changes occur. Consequently, DMPs should clearly indicate their version number and include an update schedule. At a minimum, the European Research Council advises that the DMP should be revised during the periodic evaluation or assessment of the project (Data Management—H2020 Online Manual, 2024).

The DMP should specify data access protocols, including authorised users, timing, and locations. To ensure transparent and secure practices, it should delineate data-sharing methodologies and address openness limitations, such as anonymising sensitive information or redacting names. Comprehensive metadata and lucid explanations are essential to render the data intelligible and accessible to a wider audience (DCC (2024); DCC (2024); LIBER Europe (2024)).

Explicit policies must delineate data reuse and redistribution, addressing permission constraints and anticipated applications. Creative Commons licences effectively communicate usage terms, fostering ethical data utilisation whilst maximising research impact.

Designating individuals responsible for the data (data stewards) and specifying preservation durations is advisable. Detailing archival locations and long-term strategies ensures data integrity and accessibility, safeguarding research outputs'longevity

and relevance. Addressing these aspects in the DMP ensures meticulous data handling, supports immediate project requirements and long-term research objectives, and enhances the ATR project's success and sustainability.

6.2 Project management

Project management orchestrates the strategic planning, coordination, and execution of activities and resources to achieve specified objectives within defined temporal and material constraints (Cremer et al., 2024).

6.2.1 Ensuring high-quality and consistent transcription output

Having established the necessity of selecting a project-specific transcription scheme—whether for digital editions, collection accessibility, or research purposes—we now focus on maintaining transcription quality for ATR projects. This necessitates ongoing evaluation of conventions and adherence to best practices throughout the project lifecycle. The iterative process encompasses key project management steps:

Training and testing Implement iterative transcription and model training trials to identify challenges and refine the scheme. This ongoing process ensures continuous alignment with project objectives and enhances accuracy.

Element tracking Maintain a project spreadsheet documenting novel codicological features, punctuation, abbreviations, and unique elements. Codify these in team meetings to evolve the transcription scheme systematically.

Consistency Ensure uniformity when integrating existing transcriptions or ATR models by aligning conventions across ground truth and output.

Volunteer quality control Implement a multi-tiered approach for volunteer contributions:

1. Double-checking or comparing independent transcriptions.
2. Assigning distinct transcription and proofreading roles.
3. Expert validation of outputs to meet quality standards.

6.2.2 Version control

After uploading documents to Transkribus, utilizing the app's version control features to monitor changes throughout the project is helpful. Regularly updating the status of each page and document ensures transparency and consistency within the team as it becomes clear whether a page is *new*, *in progress, done, final* or *ground truth* (see Fig. 5). Depending on agreements within your team, one would

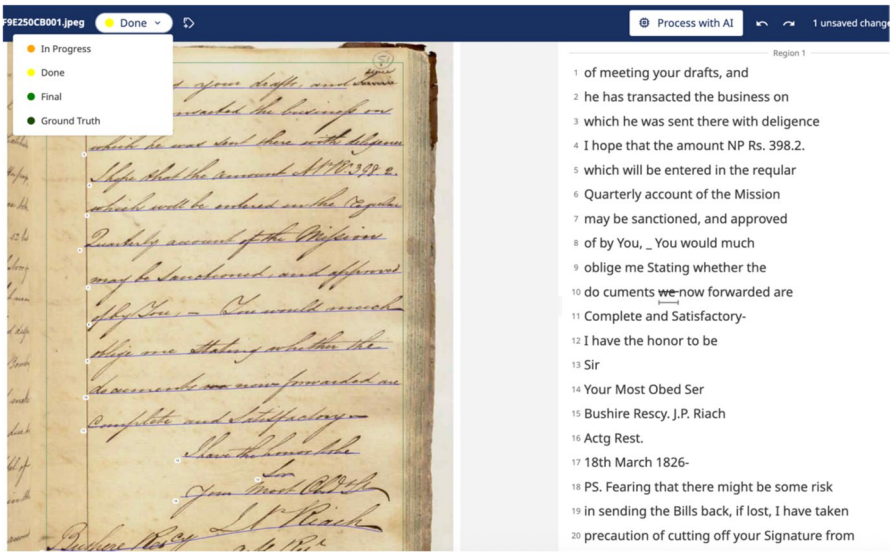

Fig. 5 Transcription text status control in Transkribus web application. India Office Records IOR/R/15/1/36, ff 50–51 via Qatar Digital Library

be able to assume that those pages that are new—and thus untouched—require some work, while those that are with a *final* or *ground truth* status concern finished work.

Another sort of version control, or at least it could be regarded as such, is the version history, which saves earlier versions of work on a particular page (see Fig. 6). If something goes wrong with your present version, you can always return to a previous version.

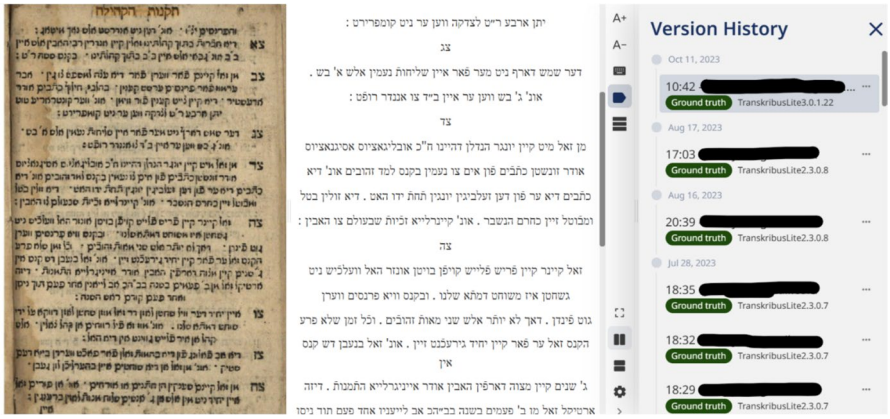

Fig. 6 Version control of p.24, Ḳehal Ḳodesh Ashkenazim (Amsterdam, Netherlands). (1711). תקנות הקהלה דקהל אשכנזים דק"ק אמשטילרדם. bi-defus uve-vet shel Anshel ben Eliʿezer Ḥazan. bi-defus uve-vet shel Anshel ben Eliʿezer Ḥazan. https://mss.huc.edu/portfolio/rbr-o-128/

6.2.3 Layout analysis

In ATR projects, layout recognition is as critical as text recognition. Moreover, within documents with a more complex layout, layout analysis errors can significantly degrade text recognition accuracy in the processing chain. Consider developing a baseline or field model tailored to the document layout to streamline training and enhance efficiency. While existing models may handle basic layouts, complex structures like tables, multi-columns, or embedded images often require further training or fine-tuning. Accurate layout recognition is helpful for the precise interpretation of these elements.

Optimising document layout is crucial for efficient post-processing and long-term usability. A well-structured layout enhances transcription and data management and ensures future accessibility by aligning with the FAIR principles: *F*indable, *A*ccessible, *I*nteroperable, and *R*eusable (Wilkinson et al., 2016). A structured layout recognition simplifies the subsequent data handling process. Structural tags, such as tagging text regions or fields, facilitate navigation and processing, especially when using multiple models for transcription. Key elements include clear headings, consistent formatting, and logical content organisation such as standardised formats that were used for e.g. letters or formal documents. A streamlined layout reduces the time and effort needed for post-processing, enhancing overall project efficiency. When designing document layouts, it is essential to consider the needs of future users by adhering to the FAIR principles. Organising documents according to these principles ensures their value and accessibility beyond the project's completion. This involves implementing precise metadata, standardised formats, and comprehensive documentation, thereby enhancing the longevity and utility of the data.

A fundamental principle of document management is purposeful organisation: each element should serve a clear function and be logically arranged. This approach results in an intuitive and functional layout, benefiting both immediate transcription tasks and future research. Optimising document layout is crucial in ATR projects, as it facilitates post-processing and supports long-term data availability. By prioritising a structured and efficient workflow, projects can achieve their immediate goals while ensuring data remains accessible for future use.

6.2.4 Utilising and improving existing models in ATR projects

In ATR projects, choosing between using, fine-tuning, or training models from scratch is pivotal to the project's efficiency, accuracy, and scope. Pre-trained ATR models offer a strong foundation, having been trained on extensive datasets, and are ideal when the models align well with the document types and when the goal is to achieve reasonable accuracy quickly without extensive initial training.

Fine-tuning adjusts an existing public model to meet project requirements better, enhancing its ability to handle document-specific nuances. This approach is ideal when the base model is generally suitable but needs refinement for slight variations in text types or layouts, aiming to improve performance on distinct document characteristics.

Training an ATR model from scratch, though labour and time-intensive, provides the highest level of customisation. This approach is essential when existing models fail to meet the project's unique demands, such as rare scripts or highly specialised document types. It is particularly valuable when there is access to a substantial dataset, such as 10,000 words, and when the project requires specific recognition capabilities that fine-tuning cannot address.

Leveraging existing resources and networks can greatly enhance a project by facilitating the reuse of GT data and fostering collaboration. Collaborating with colleagues who have previously generated GT data, such as for printed editions, saves time and ensures consistency and reliability in training data. This involves identifying and accessing relevant datasets or transcriptions from these sources and integrating them into the training and validation processes. By strategically utilising these resources, one can optimise the efficiency and effectiveness of the ATR project. Whether using, fine-tuning, or training models, it is crucial to consider the document's specific requirements and available resources to achieve the best outcomes.

6.2.5 Creating and recycling training data for ATR model generation

Recycling training data A methodical approach is crucial when developing training and validation data for an ATR project. Rather than rushing into transcription, thorough preparation—including identifying potential challenges and familiarising oneself with source materials—ensures more effective project execution. Key considerations may include text orientation, language, and document structure. This foundational knowledge facilitates navigation of transcription complexities. Moreover, identifying potential remote colleagues to leverage their expertise and resources can prove invaluable.

To maximise efficiency, prioritise recycling existing materials and transcriptions where feasible. While reusing related datasets can save time, aligning them with the project's agreed transcription scheme is crucial for consistency. The optimal size of training data should be determined based on the ATR platform, research material characteristics (e.g. languages, number of hands, image quality, page layout complexity), and project goals. Transkribus's recommended initial training set is approximately 30 pages or 10,000 words, depending on quality and complexity (Reshef & Gutschow, 2024). This benchmark establishes a solid foundation for model training.

ATR model generation Developing a detailed work plan is necessary for efficient training data creation and model generation. Pair work, with one transcribing and the other checking, enhances accuracy and speed. Test interim models using 5,000–7,000 word datasets and use these to refine new transcriptions. This iterative process helps identify and address issues early, improving overall model quality. If available, utilise a strong base model as the foundation for training efforts. This can significantly enhance transcription efficiency and speed when producing the initial training data.

Consider whether the goal is a highly specific model for particular document types or a generalised model for various typefaces and text sorts (e.g. legal, religious, or theatre texts). Specific models offer precision for certain documents, while generalised models require more initial effort but provide broader applicability.

Document any limitations encountered during model development, such as limited vocabulary or uniform document layout. Keep a detailed record of these limitations and strategies used to overcome them, as well as recurring letter or word identification errors. Address these systematically to enhance model performance.

This approach develops a resilient, efficient workflow for creating high-quality transcription models, optimising resource use and establishing a foundation for ongoing improvements and collaboration in the ATR project.

6.2.6 CER assessment

Error rates are an essential metric in evaluating the accuracy of ATR models. CER and Word Error Rate (WER) are commonly used accuracy measurements in written and spoken text recognition practices. These indicate what percentage of characters or words respectively have been incorrectly recognised and transcribed in the text output.

CER measures are currently prioritised in Transkribus, although WER and other metrics were previously available in Transkribus eXpert Client (now deprecated) (‘5. Computing Accuracy’, n.d.). WER metrics do not account for variations in word length, and a singular wrongly recognised character will designate the entire word as incorrect. As such, CER is a better indicator of model accuracy and value (Alvermann, 2019).

Error rates should aim for the lowest percentage possible, though achieving a CER of zero is unfeasible (see Fig. 7). Instead, setting a practical threshold for CER that balances precision and feasibility is essential. The following detailed approach outlines assessing CER and ensuring high-quality ATR output.

Old Czech Handwriting (with spaces)	CZE	121 340	4.92%
Chuhaister	UKR	140 435	0.62%
AHP Handwritten Portuguese 19th-20th Centuries	POR	186 342	2.53%
Dabbas 1706-1711	ARA	54 519	6.73%
Scottish Custom Books V0.8	ENG	116 003	10.42%
Danish Newspapers 1750-1850	DAN	420 266	0.56%
Middeleeuws Amsterdam	DUT	1 562 527	5.11%
Notaire de la Nouvelle-France : Adhémar dit St-Martin, Antoine (1668-1714)	FRE	54 513	6.15%
1700-tallets administrativ, gotisk håndskrift	DAN	893 395	3.27%
Notaire de la Nouvelle-France : Rageot de Beaurivage, François (1709-1753)	FRE	52 153	5.72%
Notaire de la Nouvelle-France : Dionne, Joseph (1741-1779)	FRE	55 835	3.85%
Agapet13	ARA	59 103	7.48%
Notaire de la Nouvelle-France : Genaple, François (1682-1709)	FRE	62 311	3.61%

Fig. 7 Columns representing public Pylaia- model, the number of words it was trained on, language and Character Error Rate within Transkribus

6.2.7 Setting realistic CER goals

Defining an acceptable CER is crucial for ATR projects. While a zero CER is unattainable, a range of 2–5% is generally considered excellent (Hodel, 2020; Hodel et al., 2021; Mühlberger et al., 2014; Muehlberger et al., 2019). The target CER should align with project requirements and the intended use of the ATR output. Scholarly editions may demand higher accuracy than projects focused on text analysis or increasing collection access. Document type also influences acceptable CER; legal texts, for instance, may require greater precision than literary works. To determine an appropriate CER threshold, consult with stakeholders such as historians, archivists, and end-users. Their input can help balance expectations with practical limitations. Additionally, conduct random page testing to identify potential weaknesses in the model, particularly in transcribing uncommon characters or complex layouts. This comprehensive approach ensures that the CER target effectively guides the project towards its specific goals.

Real-world performance Evaluate the model's performance on new documents beyond the training and validation sets to assess its ability to generalise. Consider output usability alongside CER; even with a low error rate, critical errors like frequent character misinterpretations may necessitate further refinement. This comprehensive evaluation ensures the model's practical effectiveness in real-world scenarios.

Iterative improvement Insights from random page testing can be used to iteratively enhance the model, including further training on problematic areas or refining the transcription scheme (Reshef & Gutschow, 2024). Documenting these iterations and their effects on the CER provides a clear record of progress and facilitates future optimisation of workflows.

6.2.8 Practical steps for CER assessment

Begin by evaluating the CER across the dataset to establish the model's baseline accuracy. Conduct random sampling by selecting a subset of pages for detailed review, noting recurring errors and specific challenges. To ensure a comprehensive assessment, focus on diverse sections of the dataset that present various challenges, such as differing handwriting styles and complex layouts.

Establish a feedback loop, using insights from random and targeted tests to guide further model improvements, which may include additional training data, preprocessing adjustments, or modifications to the transcription scheme. Implementing cross-validation techniques—by splitting the data into subsets and iteratively training and validating the model—can help ensure improvements are consistent and not overfitted to specific data subsets, providing a more reliable estimate of the model's overall performance (James et al., 2021; Kohavi, 1995).

A thorough assessment of an ATR model's quality should extend beyond error rate calculations and include practical, real-world testing. To ensure the transcriptions meet the required quality standards, it is crucial to set realistic error rate

targets, understand the intended use of the output, and perform comprehensive random page testing. This approach enhances the precision and reliability of the ATR model, ensuring that its output is suitable for its intended applications, whether for scholarly work, text analysis, or other purposes.

6.2.9 Leveraging crowdsourcing for ATR and NER

Crowdsourcing can greatly benefit ATR projects by leveraging volunteer contributions to expand workforce and expertise. Volunteers, or citizen scientists, can significantly enhance ATR and NER by contributing to various tasks. A key area is the creation of GT data, which is essential for training accurate ATR models. Volunteers can manually transcribe handwritten documents to generate high-quality GT data, crucial for model training and validation. Additionally, they can assist in quality control by reviewing and correcting transcriptions, thereby ensuring the accuracy and reliability of GT data (Funk & Sayers, 2022; Nilsson-Fernàndez & Dombrowski, 2022; Terras, 2022a).

After the ATR model processes documents, the output often requires further refinement. Volunteers can enhance transcription accuracy by identifying and correcting errors in the ATR output and providing feedback on common mistakes or challenging sections, helping to improve model performance. In NER volunteers can annotate texts by identifying and categorising entities such as names, locations, and dates (Romein et al., 2020; Aguilar & Tannier & Chastang, 2016; Cardellino et al., 2017; Dozier et al., 2010). They also play a vital role in verifying and validating the accuracy of automated entity recognition, ensuring corrections are made as needed (Terras, 2022b: 257).

Involving volunteers offers multiple benefits, including an additional review layer that helps catch errors and enhance output quality (Kasperowski & Johansson & Karsvall, 2024; Ridge, 2016). Their participation expands the project's capacity, enabling more data to be processed and transcribed, thus accelerating the project timeline. Moreover, engaging volunteers fosters a sense of community and interest, often increasing support for the project.

Effective crowdsourcing relies on strategic recruitment, training, tools, and quality assurance. Recruiting volunteers with expertise in historical documents or transcription can be facilitated through platforms like Zooniverse (Hanson & Simenstad, 2018), FromthePage, and VeleHanden.nl.[4] Comprehensive training, including tutorials, example tasks, and ongoing support, ensures volunteers understand transcription and annotation guidelines. Utilising user-friendly tools that simplify transcription, correction, and annotation will boost volunteer efficiency. Lastly, implementing robust quality assurance processes, such as peer reviews, spot checks, and feedback loops, is crucial for maintaining high standards in volunteer contributions.

Involving volunteers requires substantial time investment, ideally with a dedicated team member for management. Volunteers should not be viewed merely as free labour but recognised as active contributors. Their contributions should be

[4] It is important to keep in mind that these tools can come with a significant cost, depending on the volume of the data that needs to be processed and stored.

valued, and the benefits of participation—such as recognition, access to exclusive collections, or joining a community of citizen scholars—should be clearly communicated (Owens, 2014; 'Collective Wisdom—the State of the Art in Crowdsourcing in Cultural Heritage', 2024).

7 Project outputs

7.1 Integration of digitalised materials and data into the GLAM-workflow

Integrating ATR-digitised materials into mainstream workflows is indispensable for ensuring accessibility, usability, and longevity of project outputs. Key considerations include implementing a transparent version control system to manage data versions effectively. This can be achieved through systematic file naming, timestamps, and detailed change logs. Comprehensive documentation for each version, outlining changes and their rationale, is also crucial. This approach helps users track data evolution and select the most suitable version for their needs.

To enhance the academic value and citation of data, each dataset version should be assigned a Persistent Identifier (PID), such as a Digital Object Identifier (DOI) (Romein et al., 2023: 6). Additionally, detailed metadata—including authorship, creation date, changes made, and relevant context—should be provided to ensure proper attribution and understanding of the data.

Implementing robust backup and redundancy measures is crucial to prevent data loss. Regular backups should be conducted, and the adequacy of the current system's backup capabilities must be evaluated. Additionally, consider using external storage solutions, such as cloud services, institutional repositories, or dedicated preservation platforms, ensuring they are secure and compliant with data protection standards.

Effective data preservation strategies are vital for ensuring long-term storage and access. This involves choosing file formats that are likely to remain accessible, such as non-proprietary formats like XML or plain text. Additionally, maintaining multiple copies in various locations—both physical (e.g. external hard drives) and cloud-based—is essential for safeguarding against data loss. It is also important to establish and document policies for data openness and sharing, balancing controlled access for sensitive information with promoting open access where appropriate.

Effective implementation of data management and preservation strategies involves adhering to best practices and guidelines from organisations such as the Digital Preservation Coalition or the Open Archival Information System (OAIS). Regular audits should be conducted to identify and address potential weaknesses, including verifying backups, updating metadata, and reviewing access policies. Additionally, providing user training through workshops, tutorials, and detailed guides is crucial for ensuring proper access, citation, and use of the data.

7.2 Sharing ATR models

Researchers and practitioners in ATR share their models for several reasons. They often take pride in their work and want to make it accessible to a broader community, thereby enhancing the collective capabilities of others in similar fields. Additionally, sharing ATR models serves as a form of publication, contributing to the researcher's academic portfolio. This trend is exemplified by publicly available models on platforms such as Transkribus Public Models.

Sharing models also creates opportunities for collaboration with scholars working on related projects, leading to significant advancements and innovations. Researchers may be motivated by a genuine desire to support the community, offering resources for those dealing with similar scripts and languages. This altruistic approach contributes to a robust body of shared knowledge and tools. Additionally, sharing models can be especially valuable when training data cannot be disclosed due to privacy concerns.

To share an ATR model, researchers should follow a structured process. First, they must notify Transkribus of their intention to make the model public.[5] The model should ideally be trained on a dataset of around 50,000 words, achieving a CER of 7% or lower. For scripts or languages not currently supported by Transkribus, different criteria may apply.

A succinct model description is crucial, elucidating training content, transcription conventions, scope and pertinent information for users. A representative image or snippet can effectively illustrate the model's capabilities. Researchers must determine data visibility, balancing privacy concerns with potential utility. Proper attribution of creators, whether individuals or research groups, is essential for academic and professional recognition.

This structured approach facilitates effective dissemination of ATR models within scholarly and professional spheres, advancing the field and catalysing further research and development in Automatic Text Recognition.

7.3 Using Transkribus sites

A dedicated website enhances ATR result accessibility and utility. Transkribus Sites offers an exemplary platform, featuring advanced search capabilities for efficient document and transcription retrieval (see Fig. 8). Its intuitive Content Management System (CMS) facilitates implementation and maintenance, accommodating users with varied technical proficiencies.

Transkribus Sites also allows integration of multiple ATR outputs on a single platform, managing various datasets and transcriptions together. It offers a cost-effective solution for publishing and maintaining materials. Additional details and setup guidance are available in the Transkribus Sites page on the READ-COOP website, which provides resources to optimise the use of Transkribus Sites for ATR projects ('Transkribus Sites', 2024).

[5] This can be done by emailing them at info@readcoop.eu or through the contact form on their help centre webpage.

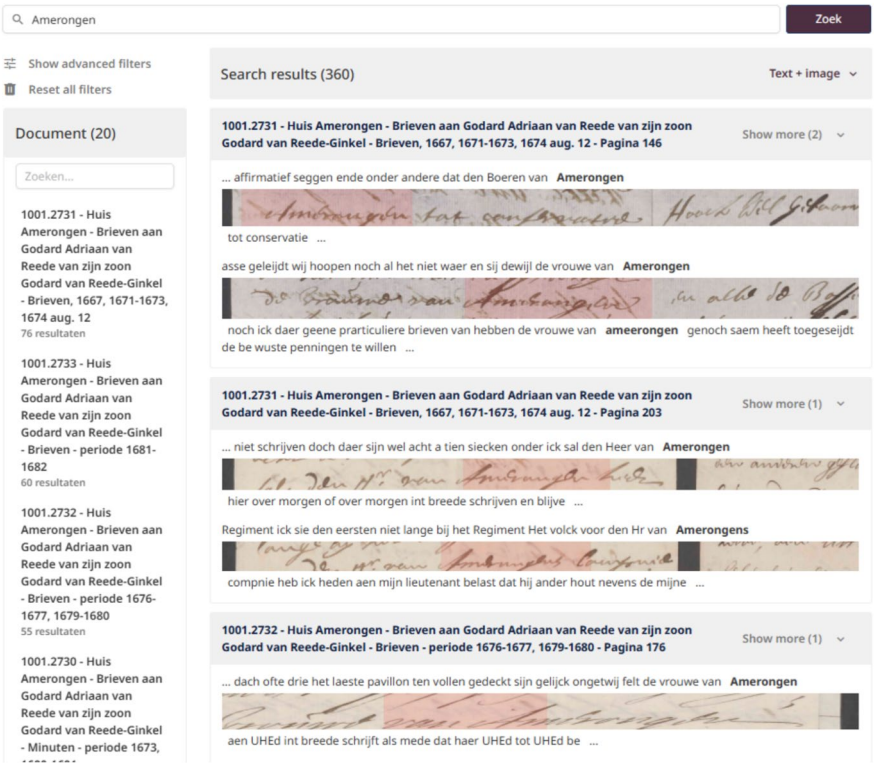

Fig. 8 Search results on Transkribus Sites—with fuzziness

In conclusion, Transkribus Sites excels as an ATR results publication platform. Its robust search functionality, intuitive CMS, integration capabilities, and cost-effectiveness render it optimal for open-access transcription project websites. This platform augments research visibility and impact, substantially benefiting the academic community.

8 Conclusion

Transitioning from proposal to project management demands careful planning, realistic budgeting, and effective team coordination. This comprehensive guide, outlining indispensable steps and best practices to ensure the optimisation of projects, focusses on extensive digitization and transcription projects such as those facilitated by Transkribus. The steps and insights mentioned are meant to result in well-structured, impactful, and sustainable projects (see Fig. 9).

It is of the utmost importance to begin with a well-defined project idea. Researchers must identify research gaps, align project goals with existing tools

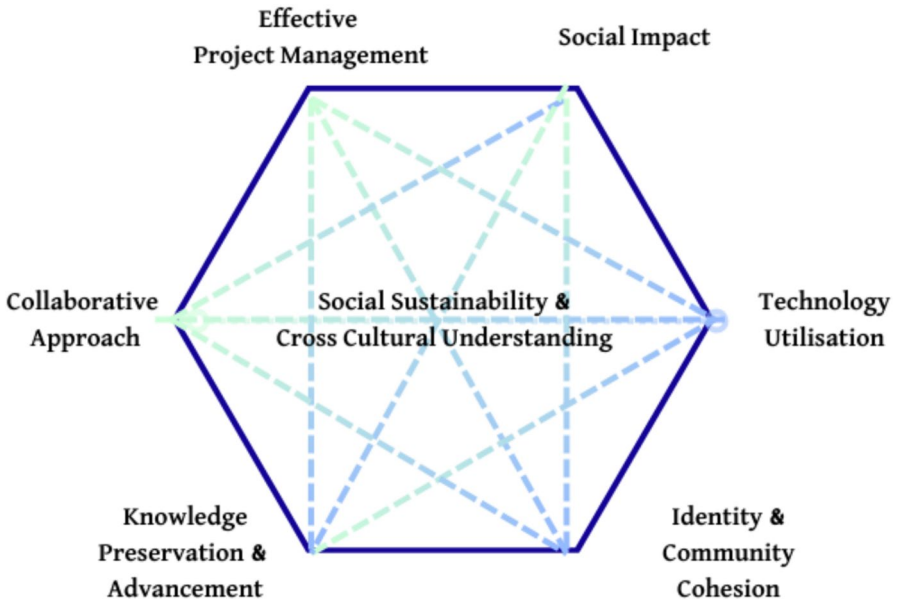

Fig. 9 Project management, or project balancing (Created with Canva)

such as Transkribus, and envision the short and long-term outcomes of the project. The crafting of a compelling project proposal is the next critical step. A successful proposal should clearly state the problem, describe the methodology, outline the expected solution, and specify the tangible outcomes that may, or may not be reached (sometimes technical innovations go faster than expected, or not). In addition, it is paramount to consider ethical considerations, particularly in document handling and data privacy, to gain approval and support.

The technical aspects of ATR methodologies require careful handling. It is fundamental to understand the distinctions between training, validation, and test sets and to conduct preliminary tests to set realistic expectations and demonstrate the process and potential of ATR methodologies. The inclusion of illustrative examples of recognition challenges and an emphasis on the necessity for specific models serves to enhance the credibility of the proposal. Budgeting is another critical component. Cost estimations should be based on practical experience, considering factors such as digitization, software costs, staffing, crowdsourcing, publishing, and long-term storage. The presentation of the significant gains of ATR projects, such as increased efficiency and improved availability, is also equally important. A robust Pylaia model, for instance, can significantly enhance transcription efficiency, thereby facilitating the management of extensive datasets.

A comprehensive DMP is fundamental to successful project execution. The DMP should delineate data collection, management, sharing, and preservation strategies, ensuring that all aspects of data handling are meticulously planned. Detailed documentation and clear explanations facilitate the openness and usability of data by a

broader audience, supporting both immediate project requirements and long-term research goals.

Effective project management necessitates establishing a well-defined team with clearly delineated roles and responsibilities, training team members, and convening regular workshops and meetings. Documenting the transcription scheme, workflow, metadata, and version management ensures transparency and consistency. Optimising document layout to facilitate post-processing and long-term usability adheres to the FAIR principles, enhancing the project's efficiency and future availability.

A systematic approach is required to create high-quality training and validation data for ATR projects. Other valuable steps can be recycling existing materials, working in pairs or small groups for improved accuracy, and testing interim models iteratively. When fine-tuning existing models or training new ones, it is imperative to consider the specific needs of the documents and available resources.

Crowdsourcing can significantly enhance ATR projects by affording an additional workforce and expertise. Volunteers can create ground-truth data, assist in quality control, and participate in NER. Implementing productive recruitment, training, and quality assurance procedures is of the utmost importance to fully harness the potential of volunteer contributions.

Integrating digitalised materials into a mainstream workflow ensures the availability, usability, and longevity of project outputs. Implementing transparent version control, assigning persistent identifiers, providing extensive metadata, and establishing solid backup and storage solutions are critical practices. Regular audits and user training further support the long-term sustainability of the project. Users should adhere to a structured process for sharing models, including descriptions, representative images, and the specification of credits for the appropriate attribution. Transkribus Sites offers a valuable platform for publishing ATR results, combining robust search functionality, a user-friendly content management system, and affordability.

9 Summary

A key takeaway from the Transkribus User Conference 2024 is the importance of collaboration and leveraging of collective knowledge. The collaborative approach fosters a sense of community, promotes cultural diversity and contributes significantly to advancing methodologies and disseminating knowledge (Snyder & Omoto, 2008; Walcher et al., 2023). The optimal outcomes of such a successful project can benefit not only the academia and heritage sector, but society at large. By studying and preserving cultural, historical and linguistic knowledge, we can foster a sense of identity, enhance community cohesion, promote cross-cultural understanding and stimulate collaboration (Alvarez-Espinar, 2023). The objective of this paper is to provide a comprehensive guide on navigating the transition from research proposals to effective project management in Automatic Text Recognition (ATR) projects, with a specific focus on historical text collections. The discussion has highlighted

the importance of key themes, including meticulous project planning, realistic budgeting, and the essential role of collaboration.

One of the principal conclusions to emerge is that a transparent and well-defined project objective is a crucial determinant of success. This necessitates the identification of research gaps where ATR can be applied and the assurance that the project's objectives are aligned with the capabilities of the selected platforms. The utilisation of tools such as Transkribus has been demonstrated to be of significant benefit in these projects, not only due to their advanced transcription capabilities, but also in terms of fostering community engagement through the sharing of models and the undertaking of collaborative efforts.

It is also imperative to emphasise the significance of generating accurate training and validation data. The recycling of existing materials, collaboration with colleagues, and the utilisation of pre-existing models wherever feasible can facilitate the process and enhance efficiency. Furthermore, this emphasises the value of crowdsourcing, which has the potential to considerably enhance project capacity by harnessing the contributions of volunteers in the creation of ground truth data, quality control, and named entity recognition tasks.

Effective project management necessitates the maintenance of comprehensive documentation at each stage of the process, encompassing aspects such as workflow and transcription schemes, data management protocols, and long-term storage strategies. These practices guarantee that the project outputs remain accessible, usable, and sustainable, in accordance with principles such as FAIR (Findable, Accessible, Interoperable, and Reusable) and CARE (Collective Benefit, Authority to Control, Responsibility, and Ethics). Moreover, the implementation of robust version control systems, the provision of comprehensive metadata, and the establishment of secure, long-term storage solutions are vital for ensuring that digitised materials continue to serve the academic community in the future.

Ultimately, the dissemination of ATR models and the accessibility of project outputs through platforms such as Transkribus Sites serve to enhance the visibility and impact of the research. Such an approach not only contributes to the broader academic field but also ensures that the project's benefits extend beyond its immediate goals, thereby supporting a wide range of scholarly and public interests.

In short, this paper has sought to highlight the intricate yet gratifying nature of ATR project management. By adhering to best practices in project planning, collaboration, data management, and transparency, researchers can successfully navigate the transition from proposal to implementation, ensuring that their work has a lasting and meaningful impact on the field of cultural heritage preservation.

Acknowledgements We thank the participants of the Transkribus User Conference 2024 and the organisers of this event for the opportunity to discuss this topic there. This project received research ethics scrutiny from Koninklijke Nederlandse Academie voor Wetenschappen (KNAW). Annemieke Romein's research was funded by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (grant VI.Veni.191H.035).

Author contributions Conceptualisation: SK, CAR, CS Data Curation: CAR, SK, RR Formal Analysis: CAR, SK Investigation: all Methodology: CAR, SK, CS, RR Resources:RR, CAR, NaA, LDV, CC, EK, AK-S Supervision:MT, CAR Visualisation:RR Writing original draft: CAR, SK, RR, all Writing—review & editing:AC, RR, CAR, CS, SK, MT, JN.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

Board affiliations disclosure Three authors on the Board of Directors of Transkribus/the READ-COOP SCE—each honorary, thus without any benefits out of their positions:

Melissa Terras is Research Director;
Günter Mühlberger was the Board's Chairperson (until May 19th, 2025);
Annemieke Romein is Community Director, and the Board's Chairperson (since May 19th, 2025).
All of the authors of this paper have served as authors from their position as *researchers*, not in any other capacity.

References

Aguilar, S. T., Tannier, X., & Chastang, P. (2016). Named entity recognition applied on a data base of medieval Latin charters. The case of Chartae Burgundiae. In *Proceedings of the 3rd HistoInformatics Workshop* (pp. 1–5). CEUR Workshop Proceedings. http://ceur-ws.org/Vol-1632/paper_9.pdf. Accessed 11 June 2024.

Allen, L., Scott, J., Brand, A., Hlava, M., & Altman, M. (2014). Publishing: Credit where credit is due. *Nature, 508*(7496), 312–313. https://doi.org/10.1038/508312a

Alvarez-Espinar, M. (2023). Smart heritage preservation through open data, crowdsourcing, and contextual light apps. In *2023 IEEE International Smart Cities Conference (ISC2)* (pp. 1–4). IEEE. https://doi.org/10.1109/ISC257844.2023.10293430

Alvermann, D. (2019). Word error rate & character error rate – How to evaluate a model. Retrieved October 23, 2024, from https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/word-error-rate-character-error-rate-how-to-evaluate-a-model/

Blau, I., & Caspi, A. (2009). What type of collaboration helps? Psychological ownership, perceived learning and outcome quality of collaboration using Google Docs. In *Learning in the Technological Era* (pp. 48–55). The Open University of Israel. https://cris.openu.ac.il/en/publications/what-type-of-collaboration-helps-psychological-ownership-perceive. Accessed 11 June 2024.

Cardellino, C., Teruel, M., Alemany, L. A., & Villata, S. (2017). Legal NERC with ontologies, Wikipedia and curriculum learning. In *15th European Chapter of the Association for Computational Linguistics (EACL 2017)* (pp. 254–259). Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-2041

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., et al. (2020). The CARE principles for indigenous data governance. *Data Science Journal, 19*(1), Article 43. https://doi.org/10.5334/dsj-2020-043

Colutto, S., Kahle, P., Hackl, G., & Mühlberger, G. (2019). Transkribus: A platform for automated text recognition and searching of historical documents. In *2019 15th International Conference on eScience (eScience)* (pp. 463–466). IEEE. https://doi.org/10.1109/eScience.2019.00060

Connelly, M. A. (2023). "A process that works for everybody": Prioritizing intentionality and conversation to facilitate inclusion in a digitally-mediated collaborative writing process [Doctoral

dissertation, Indiana University of Pennsylvania]. ProQuest Dissertations and Theses Global. https://www.proquest.com/docview/2896719016

CRediT. (2022). *CRediT*. https://credit.niso.org/. Accessed 11 June 2024.

Cremer, F., Dogunke, S., Neubert, A. M., spsampsps Wübbena, T. (Eds.). (2024). *Projektmanagement und Digital Humanities*. Bielefeld University Press. https://www.bielefeld-university-press.de/978-3-8376-6967-1/projektmanagement-und-digital-humanities/. Accessed 11 June 2024.

Data Management - H2020 Online Manual. (2024). European Commission. Retrieved September 16, 2024, from https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm#A1-template

DHCC. (2024). Introduction. Retrieved June 15, 2024, from https://sas-dhrh.github.io/

DHCC Information, Measurement and Practice Action Group. (2022). *A researcher guide to writing a climate justice oriented data management plan*. https://doi.org/10.5281/zenodo.6451499

Digital Curation Centre. (2024). Example DMPs and guidance. Retrieved September 16, 2024, from https://www.dcc.ac.uk/resources/data-management-plans/guidance-examples

Digital Curation Centre. (2024). How to develop a data management and sharing plan. Retrieved September 16, 2024, from https://www.dcc.ac.uk/guidance/how-guides/develop-data-plan

Dozier, C., Kondadadi, R., Light, M., Vachher, A., Veeramachaneni, S., & Wudali, R. (2010). Named entity recognition and resolution in legal text. In *Semantic Processing of Legal Texts* (pp. 27–43). Springer. https://doi.org/10.1007/978-3-642-12837-0_2

Funk, J., & Sayers, J. (2022). Autoethnographies of mediation. In J. O'Sullivan (Ed.), *The Bloomsbury Handbook to the Digital Humanities* (pp. 101–110). Bloomsbury Publishing.

Global Indigenous Data Alliance. (2022). CARE principles of indigenous data governance. https://www.gida-global.org/care. Accessed 11 June 2024.

Hanson, D., & Simenstad, A. (2018). Combining human and machine transcriptions on the Zooniverse platform. In W. Xu, A. Ritter, T. Baldwin, & A. Rahimi (Eds.), *Proceedings of the 2018 EMNLP workshop W-NUT: the 4th workshop on noisy user-generated text* (pp. 215–216). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6129

Hodel, T. (2020). Best-practices zur Erkennung alter Drucke und Handschriften – Die Nutzung von Transkribus large- und small-scale. In C. Schöch (Ed.), *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts*. Zenodo. https://doi.org/10.5281/zenodo.3666690

Hodel, T., Schoch, D., Schneider, C., & Purcell, J. (2021). General models for handwritten text recognition: Feasibility and state-of-the art. German Kurrent as an example. *Journal of Open Humanities Data, 7*, 13. https://doi.org/10.5334/johd.46

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). Statistical learning. In *An introduction to statistical learning (Springer texts in statistics)* (pp. 15–57). Springer. https://doi.org/10.1007/978-1-0716-1418-1_2

Kasperowski, D., Johansson, K.-M., & Karsvall, O. (2024). Temporalities and values in an epistemic culture: Citizen humanities, local knowledge, and AI-supported transcription of archives. *Archives & Manuscripts*, *51*, e10937. https://doi.org/10.37683/asa.v51.10937

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on artificial intelligence - volume 2* (pp. 1137–1143). Morgan Kaufmann Publishers.

Leonelli, S. (2023). Philosophy of open science. Elements in the Philosophy of Science. https://doi.org/10.1017/9781009416368

LIBER Europe. (2024). DMP catalogue. Retrieved September 16, 2024, from https://libereurope.eu/working-group/research-data-management/plans/

McDermott, A., Coghlan, D., & Keating, M. A. (2008). Research for action and research in action: Processual and action research in dialogue. *Irish Journal of Management, 29*(1), 1–18.

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., & Fiel, S. (2019). Transforming scholarship in the archives through handwritten text recognition. *Journal of Documentation, 75*(5), 954–976.

Mühlberger, G., Zelger, J., & Sagmeister, D. (2014). User-driven correction of OCR errors: Combining crowdsourcing and information retrieval technology. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage* (pp. 53–56). Association for Computing Machinery. https://doi.org/10.1145/2595188.2595212

Nilsson-Fernàndez, P., & Dombrowski, Q. (2022). Multilingual digital humanities. In J. O'Sullivan (Ed.), *The Bloomsbury Handbook to the Digital Humanities* (pp. 83–92). Bloomsbury Publishing.

Nockels, J., Gooding, P., & Terras, M. (2024). The implications of handwritten text recognition for accessing the past at scale. *Journal of Documentation, 80*(7), 148–167. https://doi.org/10.1108/JD-09-2023-0183

Owens, T. (2014). Making crowdsourcing compatible with the missions and values of cultural heritage organisations. In *Crowdsourcing our cultural heritage*. Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315575162-15/makingcrowdsourcing-compatible-missions-values-cultural-heritage-organisations-trevor-owens

Padlet. (2024). Padlet: Beauty will save the work. Retrieved June 15, 2024, from https://padlet.com/

Reshef, R., & Gutschow, M. (2024). Text recognition model for Yiddish in Vaybertaytsh typeface, based on community regulations. *Journal of Open Humanities Data, 10*(1), 35. https://doi.org/10.5334/johd.194

Ridge, M. (2016). Making digital history: The impact of digitality on public participation and scholarly practices in historical research [Doctoral dissertation]. ProQuest Dissertations and Theses Global. https://www.proquest.com/docview/1896365285/abstract/18A17DC1ED234171PQ/1. Accessed 11 June 2024.

Romein, C. A., Hodel, T., Gordijn, F., van Zundert, J. J., Chagué, A., van Lange, M., Jensen, H. S., et al. (2023). 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. *Journal of Data Mining and Digital Humanities*. https://doi.org/10.5281/zenodo.8116009

Romein, C. A., Kemman, M., Birkholz, J. M., Baker, J., De Gruijter, M., Meroño-Peñuela, A., Ries, T., Ros, R., & Scagliola, S. (2020). State of the field: Digital history. *History, 105*(365), 291–312. https://doi.org/10.1111/1468-229X.12969

Snyder, M., & Omoto, A. M. (2008). Volunteerism: Social issues perspectives and social policy implications. *Social Issues & Policy Review,2*(1), 1–36. https://doi.org/10.1111/j.1751-2409.2008.00009.x

Terras, M. (2022a). Chapter 7: Inviting AI into the archives: The reception of handwritten recognition technology into historical manuscript transcription. In *AI in the Archives* (pp. 179–204). Bielefeld University Press. https://doi.org/10.1515/9783839455845-008

Terras, M. (2022b). Digital humanities and digitized cultural heritage. In J. O'Sullivan (Ed.), *The Bloomsbury Handbook to the Digital Humanities* (pp. 255–266). Bloomsbury Publishing.

The collective wisdom handbook: Perspectives on crowdsourcing in cultural heritage - Community review version. (2024). PubPub. Retrieved October 23, 2024, from https://britishlibrary.pubpub.org/the-collective-wisdom-handbook-perspectives-on-crowdsourcing-incultural-heritage%2D%2D-community-review-version

Transkribus. (2024). Transkribus sites. Retrieved June 15, 2024, from https://www.transkribus.org/sites

Walcher, J., Abel, A., Andresen, J., Brasolin, P., Dissertori, I., Eberwein, E., Franzini, G., et al. (2023). On a digital journey into yesterday's future: Zeit.Shift – Preserving Tyrolians cultural text heritage. In *Proceedings of Austrian Citizen Science Conference 2022 — PoS(ACSC2022)* (407:019). SISSA Medialab. https://doi.org/10.22323/1.407.0019

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, Article 160018. https://doi.org/10.1038/sdata.2016.18

Authors and Affiliations

C. Annemieke Romein[1,2,3,4] · Süphan Kırmızıaltın[5] · Ronny Reshef[6] · Christa Schneider[2] · Giorgia Agostini[3,7] · Ash Charlton[8] · Melissa Terras[3,8] · Joseph Nockels[9] · Seoyeong Ahn[10] · Fatma Aladağ[11] · Carlotta Capurro[12] · Levi Damsma[13] · Lorena De Vita[12] · Laura Fahnenbruck[12] · Gert Gielis[14] · Rachael Griffiths[15] · Adi Keinan-Schoonbaert[16] · Erika Kuijpers[17] · Mirjam El Attal[18] · Günter Mühlberger[3,18] · Juma Noah Omollo[19] · Achim Rabus[20] · Salvatore Spina[21] · Nour al Assali[17] · Eugenio Torres Flawiá[3,22,23] · David Joseph Wrisley[5] · Heleen Wilbrink[24]

✉ C. Annemieke Romein
c.a.romein@utwente.nl

1 Huygens Institute for the History and Culture of the Netherlands, Amsterdam, the Netherlands

2 University of Bern, Bern, Switzerland

3 Transkribus/READ-COOP SCE, Innsbruck, Austria

4 Universiteit Twente, Enschede, Netherlands

5 NYU Abu Dhabi, Abu Dhabi, United Arab Emirates

6 Department of Business Society Management, Rotterdam School of Management, Erasmus University Rotterdam, Rotterdam, the Netherlands

7 Universities of Florence, Pisa and Siena, Italy

8 University of Edinburgh, Edinburgh, UK

9 University of Sheffield, Sheffield, UK

10 University of Amsterdam, Amsterdam, the Netherlands

11 Marmara University, Istanbul, Turkey

12 Utrecht University, Utrecht, the Netherlands

13 University Library, Radboud University Nijmegen, Nijmegen, the Netherlands

14 State Archives of Belgium, Brussels, Belgium

15 École Pratique Des Hautes Études (EPHE), Paris, France

16 British Library, London, UK

17 Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

18 Universität Innsbruck, Innsbruck, Austria

19 Max Planck Institute for Legal History and Legal Theory, Frankfurt am Main, Germany

20 Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau, Germany

21 Università Degli Studi Di Catania, Catania, Italy

22 Humanities Department, Universidad de San Andrés (UDESA), Victoria, Buenos Aires, Argentina

23 Archivo Bunge y Born, Buenos Aires, Argentina

24 Utrecht City Archive, Utrecht, the Netherlands